

# Generalized EXTRA stochastic gradient Langevin dynamics

Mert Gürbüzbalaban <sup>1</sup>, Mohammad Rafiqul Islam <sup>2</sup>, Xiaoyu Wang <sup>3</sup>, Lingjiong Zhu <sup>4</sup>

December 4, 2024

## Abstract

Langevin algorithms are popular Markov Chain Monte Carlo methods for Bayesian learning, particularly when the aim is to sample from the posterior distribution of a parametric model, given the input data and the prior distribution over the model parameters. Their stochastic versions such as stochastic gradient Langevin dynamics (SGLD) allow iterative learning based on randomly sampled mini-batches of large datasets and are scalable to large datasets. However, when data is decentralized across a network of agents subject to communication and privacy constraints, standard SGLD algorithms cannot be applied. Instead, we employ decentralized SGLD (DE-SGLD) algorithms, where Bayesian learning is performed collaboratively by a network of agents without sharing individual data. Nonetheless, existing DE-SGLD algorithms induce a bias at every agent that can negatively impact performance; this bias persists even when using full batches and is attributable to network effects. Motivated by the EXTRA algorithm and its generalizations for decentralized optimization, we propose the generalized EXTRA stochastic gradient Langevin dynamics, which eliminates this bias in the full-batch setting. Moreover, we show that, in the mini-batch setting, our algorithm provides performance bounds that significantly improve upon those of standard DE-SGLD algorithms in the literature. Our numerical results also demonstrate the efficiency of the proposed approach.

## 1 Introduction

In our era of big data, the amount of data collected and stored has seen exponential growth with ever-increasing rates. Given the rapid pace at which data are generated, often exceeding our ability to analyze it—particularly due to limitations in computational resources—there is a growing interest in developing scalable machine learning algorithms that can efficiently handle large datasets. Very often, because of communication constraints and privacy constraints, gathering all these data for centralized processing is often impractical or infeasible. Decentralized machine learning algorithms have received a lot of attention for such applications where agents can collaboratively learn a predictive model without sharing their own data but sharing only their local models with their immediate neighbors at some frequency to generate a global model; see e.g. [HBJ18, HBM19, ABC<sup>+</sup>20].

Although there is a large body of literature on scaleable first-order decentralized learning methods have been proposed in the literature such as decentralized stochastic approximation and optimization algorithms (see e.g. [ULGN17, GDG19, SBB<sup>+</sup>19, Ned20]), very few of them deal with decentralized Bayesian learning (inference) [PBGG20, GGHZ21]. With this context, we now introduce the problem of decentralized Bayesian inference. Assume there are  $N$  agents connected over

---

<sup>1</sup>Department of Management Science and Information Systems, Rutgers Business School, Piscataway, New Jersey, United States of America; mg1366@rutgers.edu

<sup>2</sup>Department of Mathematics, Florida State University, Tallahassee, Florida, United States of America; rislam@fsu.edu

<sup>3</sup>FinTech Thrust, Hong Kong University of Science and Technology (Guangzhou), Guangzhou, Guangdong, People's Republic of China; xiaoyuwang@hkust-gz.edu.cn

<sup>4</sup>Department of Mathematics, Florida State University, Tallahassee, Florida, United States of America; zhu@math.fsu.edu

a network  $\mathcal{G} = (\mathcal{V}, \mathcal{E})$  with  $\mathcal{V} = \{1, 2, \dots, N\}$  representing the agents and  $\mathcal{E} \subseteq \mathcal{V} \times \mathcal{V}$  being the set of edges; i.e.  $i$  and  $j$  are connected if  $(i, j) \in \mathcal{E}$  where the network is undirected, i.e.  $(i, j) \in \mathcal{E}$  then  $(j, i) \in \mathcal{E}$ . Let  $Z = [z_1, \dots, z_n]$  be a dataset consisting of  $n$  independent and identically distributed (i.i.d.) data vectors sampled from a parameterized distribution  $p(Z|x)$  where the parameter  $x \in \mathbb{R}^d$  has a common prior distribution  $p(x)$ . Due to the decentralization in the data collection, each agent  $i$  possesses a subset  $Z_i$  of the data where  $Z_i = \{z_1^i, z_2^i, \dots, z_{n_i}^i\}$  and  $n_i$  is the number of samples of the agent  $i$ . The data is held disjointly over agents; i.e.  $Z = \cup_i Z_i$  with  $Z_i \cap Z_j = \emptyset$  for  $j \neq i$ . The goal is to sample from the posterior distribution  $p(x|Z) \propto p(Z|x)p(x)$ . Since the data points are independent, the log-likelihood function will be additive;  $\log p(Z|x) = \sum_{i=1}^N \sum_{j=1}^{n_i} \log p(z_j^i|x)$ . Thus, if we set

$$f(x) := \sum_{i=1}^N f_i(x), \quad f_i(x) := - \sum_{j=1}^{n_i} \log p(z_j^i|x) - \frac{1}{N} \log p(x), \quad (1.1)$$

the aim is to sample from the posterior distribution with density  $\pi(x) := p(x|Z) \propto e^{-f(x)}$ , where the functions  $f_i(x)$  are called *component functions* with  $f_i(x)$  being associated to the local data of agent  $i$  that is only accessible by the agent  $i$ . Different choices of the log-likelihood function and therefore the component functions result in different problems, including for example Bayesian linear regression [Hof09], Bayesian logistic regression [Hof09], Bayesian principal component analysis [DRW<sup>+</sup>16] and Bayesian deep learning [WY20, PS17].

Decentralized Langevin algorithms have been proposed in the recent literature that can be used in the large-scale decentralized sampling problems [PBGG20, GGHZ21]. In this paper, we propose and study a new class of Langevin algorithms for decentralized Bayesian inference. For these algorithms, we provide a non-asymptotic convergence analysis alongside numerical experiments. More specifically, our contributions are as follows:

First, inspired by the EXTRA algorithm and its extensions in the decentralized optimization literature [SLWY15, Jak18], we propose a new algorithm, termed the *generalized EXTRA stochastic gradient Langevin dynamics*, enabling collaborative Bayesian learning across a network of agents without requiring them to share individual data. Our algorithm eliminates the network-induced bias present in existing DE-SGLD algorithms that rely on full-batch processing [GGHZ21]. We provide non-asymptotic performance guarantees for generalized EXTRA SGLD when each of the components  $f_i(x)$  is strongly convex and smooth in which case the target distribution has density  $\pi(x) \propto e^{-f(x)}$  where  $f$  is strongly convex and smooth. More specifically, we show that the distribution of the iterates  $x_i^{(k)}$  converges to a neighborhood of the posterior distribution  $\pi(x)$  linearly (geometrically fast in  $k$ ) in the 2-Wasserstein distance with a properly chosen stepsize and communication matrices (Theorem 4). We can also show similar results for the averaged iterates  $\bar{x}^{(k)} = \frac{1}{N} \sum_{i=1}^N x_i^{(k)}$ . Our proof technique relies on analyzing generalized EXTRA SGLD as a perturbed version of the Euler-Maruyama discretization of an overdamped Langevin diffusion where the perturbation effect is due to the stochastic nature of the gradients and the network effect where agents are only able to communicate with their immediate neighbors. The proof technique relies on developing novel bounds on the  $L^2$  distance between  $x_i^{(k)}$  and their average  $\bar{x}^{(k)}$ , as well as the  $L^2$  distance between the average iterate  $\bar{x}^{(k)}$  and iterates based on the Euler-Maruyama discretization of overdamped diffusion.

Second, we rigorously compare the iteration complexity of our generalized EXTRA SGLD algorithm to the existing iteration complexity results for the DE-SGLD algorithm, and show improvement by a factor of at least  $L$ , where  $L$  is the smoothness coefficient of  $f_i$ 's (Proposition 5).

Finally, we provide numerical experiments that illustrate our theory and showcase the practical performance of the EXTRA SGLD algorithm: We show on Bayesian linear regression with synthetic data and Bayesian logistic regression tasks with both synthetic and real data that our method allows each agent to sample from the posterior distribution efficiently without communicating local data. We compare the numerical results of the EXTRA SGLD with DE-SGLD in the literature and show superior performance.

## 2 Preliminaries and Background

**Langevin algorithms.** One of the most widely used Markov Chain Monte Carlo methods in statistics are *Langevin algorithms*, that allow one to sample from a given density  $\pi(x)$  of interest. The classical one is based on the *overdamped Langevin SDE*; see e.g. [Dal17, DM19, DM17, DK19]:

$$dX(t) = -\nabla f(X(t))dt + \sqrt{2}dW_t, \quad (2.1)$$

where  $f : \mathbb{R}^d \rightarrow \mathbb{R}$  and  $W_t$  is a standard  $d$ -dimensional Brownian motion that starts at zero at time zero. Under some mild assumptions on  $f$ , the diffusion (2.1) admits a unique stationary distribution with the density  $\pi(x) \propto e^{-f(x)}$ , also known as the *Gibbs distribution* [Pav14]. For computational purposes, this diffusion is simulated by considering its discretization. Although various discretization schemes are proposed, Euler-Maruyama discretization is the simplest one and is known as the unadjusted Langevin algorithm in the literature [DM17, DM19]:

$$x_{k+1} = x_k - \eta \nabla f(x_k) + \sqrt{2\eta} w_{k+1}, \quad (2.2)$$

where  $\eta > 0$  is the stepsize parameter, and  $w_k \in \mathbb{R}^d$  is a sequence of i.i.d. standard Gaussian random vectors  $\mathcal{N}(0, I_d)$ . But then the discretized chain (2.2) does not converge to the target  $\pi$  and has a bias that needs to be properly characterized to provide performance guarantees [DK19]. The unadjusted Langevin algorithm (2.2) assumes availability of the gradient  $\nabla f$ . On the other hand, in many settings in machine learning, computing the full gradient  $\nabla f$  is either infeasible or impractical. For example, in Bayesian regression or classification problems,  $f$  can have a finite-sum form as the sum of many component functions over all the data points and the number of data points can be large (see, e.g., [GGHZ21, XCZG18]). In such settings, algorithms that rely on *stochastic gradients*, i.e., unbiased stochastic estimates of the gradient obtained by a randomized sampling of the data points, is often more efficient [Bot10]. This fact motivated the development of Langevin algorithms that can support stochastic gradients. In particular, if one replaces the full gradient  $\nabla f$  in (2.2) by a stochastic gradient, the resulting algorithm is known as the *stochastic gradient Langevin dynamics* (SGLD) (see, e.g., [WT11]). There has been growing recent interest in the non-asymptotic analysis of Langevin algorithms, motivated by applications to large-scale data analysis and Bayesian inference. The Langevin algorithms admit convergence guarantees to a stationary distribution in a variety of metrics and under various assumptions on  $f$ ; see e.g. [Dal17, DM17, DM19, CB18, EHZ22, DK19, BCM<sup>+</sup>21, RRT17, XCZG18, CMR<sup>+</sup>21, ZADS23].

**Decentralized setting.** We consider decentralized algorithm where the agent is connected over a connected network by  $N$  nodes, and  $W = [W_{ij}] \in \mathbb{R}^{N \times N}$  is a symmetric, doubly stochastic matrix such that, for  $i \neq j$ ,  $W_{ij} = W_{ji} > 0$  if  $\{i, j\} \in \mathcal{E}$ , and  $W_{ij} = W_{ji} = 0$  if  $\{i, j\} \notin \mathcal{E}$ , and  $W_{ii} = 1 - \sum_{j \neq i} W_{ij} > 0$ . Moreover, we have  $\sigma_{\max}(W - \frac{1}{n} \mathbf{1}_N \mathbf{1}_N^T) < 1$ , where  $\sigma_{\max}$  denotes the

largest singular value and  $1_N \in \mathbb{R}^N$  is a column vector of ones. We aim to sample from a target distribution with density  $\pi(x) \propto e^{-f(x)}$  on  $\mathbb{R}^d$  with  $f(x) := \sum_{i=1}^N f_i(x)$ .

In decentralized optimization, decentralized gradient descent (DGD) [NO09] carries the following iterative algorithm

$$x^{(k+1)} = \mathcal{W}x^{(k)} - \eta \nabla F\left(x^{(k)}\right), \quad \mathcal{W} = W \otimes I_d, \quad (2.3)$$

where  $x^{(k)} = \left[\left(x_1^{(k)}\right)^T, \dots, \left(x_N^{(k)}\right)^T\right]^T \in \mathbb{R}^{Nd}$  with  $\mathcal{W} = W \otimes I_d$  as the Kronecker product of matrices  $W$  and  $I_d$  where  $I_d$  is a  $d \times d$  identity matrix and  $F(x) : \mathbb{R}^{Nd} \rightarrow \mathbb{R}$  is defined as:

$$F(x) = F(x_1, \dots, x_N) := \sum_{i=1}^N f_i(x_i), \quad \text{for any } x = (x_1, \dots, x_N) \in \mathbb{R}^{Nd}. \quad (2.4)$$

Let  $\mathbf{x}_* := [x_*^T, \dots, x_*^T]^T \in \mathbb{R}^{Nd}$ , where  $x_*$  is the minimizer of  $f(x)$ . It satisfies the conditions (1)  $\mathbf{x}_* = \mathcal{W}\mathbf{x}_*$ , (2)  $1_{Nd}^T \nabla F(\mathbf{x}_*) = \sum_{i=1}^N \nabla f_i(x_*) = 0$ ; these are referred to as *consensus* and *optimality* conditions respectively.

**Inexactness and exact algorithms.** If we take the limit over  $k$  in DGD iterations (2.3), we get

$$x^\infty = \mathcal{W}x^\infty - \eta \nabla F(x^\infty), \quad (2.5)$$

if  $x^\infty = \mathbf{x}_*$ , then we must have  $\mathcal{W}x^\infty = x^\infty$  by consensus, then we can get

$$x^\infty = x^\infty - \eta \nabla F(x^\infty), \quad (2.6)$$

which means  $\nabla F(x^\infty) = \mathbf{0}$ , i.e.  $\nabla f_i(x_i^\infty) = 0$  for every  $i$ , that implies for any agent  $i$ ,  $x_i^\infty$  simultaneously minimizes the objective function  $f_i$ , which is impossible in general. Hence,  $x^\infty \neq \mathbf{x}_*$  in general and DGD is inexact. Although it is inexact, it is shown  $\|x^\infty - \mathbf{x}_*\| \leq \mathcal{O}(\eta\sqrt{N})$ , see [GGHZ21], [YLY16], [AFGO19] and [FGO<sup>+</sup>22]. A decentralized exact first-order algorithm (EXTRA) proposed by [SLWY15] can solve the consensus optimization problem and converges to the exact solution. [Jak18] unified and generalized this exact distributed first-order algorithm.

**Decentralized Langevin algorithms.** The *decentralized stochastic gradient Langevin dynamics* (DE-SGLD) algorithm [SSP20, GGHZ21] consists of a weighted averaging with the local variables  $x_j^{(k)}$  of node  $i$ 's immediate neighbors  $j \in \Omega_i := \{j : (i, j) \in \mathcal{E}\}$ , where  $x_i^{(k)}$  denotes the local variable of node  $i$  at iteration  $k$ , as well as a stochastic gradient step over the node's component function  $f_i(x)$ , i.e.

$$x_i^{(k+1)} = \sum_{j \in \Omega_i} W_{ij} x_j^{(k)} - \eta \tilde{\nabla} f_i\left(x_i^{(k)}\right) + \sqrt{2\eta} w_i^{(k+1)}, \quad (2.7)$$

where  $\eta > 0$  is the stepsize,  $W_{ij}$  are the entries of a doubly stochastic weight matrix  $W$  with  $W_{ij} > 0$  only if  $i$  is connected to  $j$ ,  $w_i^{(k)}$  are independent and identically distributed (i.i.d.) Gaussian random variables with zero mean and identity covariance matrix for every  $i$  and  $k$ , and  $\tilde{\nabla} f_i\left(x_i^{(k)}\right)$  is an unbiased stochastic estimate of the deterministic gradient  $\nabla f_i\left(x_i^{(k)}\right)$  with a bounded variance.

When the number of data points  $n_i$  is large, stochastic estimates  $\tilde{\nabla} f_i(x)$  are cheaper to compute compared to actual gradients  $\nabla f_i(x)$  and can for instance be estimated from a mini-batch of data, i.e. from randomly selected smaller subsets of data. This allows the DE-SGLD method to be scaleable to big data settings when  $n_i$  can be large. Without Gaussian noise, iterations are also equivalent to the decentralized stochastic gradient algorithm [SKP<sup>+</sup>20, FGO<sup>+</sup>22] which has its origins in the decentralized gradient descent (DGD) methods introduced in [NO09].

**Strong convexity and smoothness.** Let  $\mathcal{S}_{\mu,L}(\mathbb{R}^d)$  denote the set of functions from  $\mathbb{R}^d$  to  $\mathbb{R}$  that are  $\mu$ -strongly convex and  $L$ -smooth, that is, for any  $g \in \mathcal{S}_{\mu,L}(\mathbb{R}^d)$ , it holds that

$$\frac{L}{2} \|x - y\|^2 \geq g(x) - g(y) - \nabla g(y)^T (x - y) \geq \frac{\mu}{2} \|x - y\|^2, \quad \text{for every } x, y \in \mathbb{R}^d. \quad (2.8)$$

**Notations.** Define  $\mathcal{P}_2(\mathbb{R}^d)$  as the space consisting of all the Borel probability measures  $\mu$  on  $\mathbb{R}^d$  with the finite second moment (based on the Euclidean norm). For any  $\mu_1, \mu_2 \in \mathcal{P}_2(\mathbb{R}^d)$ , the 2-Wasserstein distance  $\mathcal{W}_2$  (see e.g. [Vil09]) between  $\mu_1$  and  $\mu_2$  is defined as:  $\mathcal{W}_2(\mu_1, \mu_2) := (\inf \mathbb{E} [\|Y_1 - Y_2\|^2])^{1/2}$ , where the infimum is taken over all joint distributions of the random variables  $Y_1, Y_2$  with marginal distributions  $\mu_1, \mu_2$  respectively. For any  $x, y \in \mathbb{R}$ , denote  $x \vee y := \max\{x, y\}$  and  $x \wedge y := \min\{x, y\}$ . For any random variable  $X$ , denote  $\mathcal{L}(X)$  the law of  $X$ .

### 3 EXTRA Langevin Algorithms

We aim to sample from a target distribution with density  $\pi(x) \propto e^{-f(x)}$  on  $\mathbb{R}^d$  with  $f(x) := \sum_{i=1}^N f_i(x)$ . Now we make the first assumption on the objective function.

**Assumption 1.** We assume the component functions  $f_i$  are  $\mu$ -strongly convex and  $L$ -smooth with  $L > \mu$ , i.e.  $f_i \in \mathcal{S}_{\mu,L}(\mathbb{R}^d)$  for every  $i = 1, 2, \dots, N$ .

Under Assumption 1, it follows that  $F$  is also  $\mu$ -strongly convex and  $L$ -smooth where we recall from the definition in (2.4) such that  $F : \mathbb{R}^{Nd} \rightarrow \mathbb{R}$  with  $F(x_1, x_2, \dots, x_N) = \sum_{i=1}^N f_i(x_i)$  for any  $x = (x_1, \dots, x_N) \in \mathbb{R}^{Nd}$ .

We propose *EXTRA stochastic gradient Langevin dynamics* (EXTRA SGLD) to target  $\pi$  that is defined as follows

$$x_i^{(k+2)} = \sum_{j \in \Omega_i} W_{ij} x_j^{(k+1)} - \eta \tilde{\nabla} f_i(x_i^{(k+1)}) + \sqrt{2\eta} w_i^{(k+2)}, \quad (3.1)$$

$$x_i^{(k+1)} = \sum_{j \in \Omega_i} \widetilde{W}_{ij} x_j^{(k)} - \eta \tilde{\nabla} f_i(x_i^{(k)}) + \sqrt{2\eta} w_i^{(k+1)}, \quad (3.2)$$

where  $w_i^{(k)}$  are standard  $d$ -dimensional Gaussian random vectors that are i.i.d. in both  $i = 1, 2, \dots, N$  and  $k = 1, 2, 3, \dots$ . In this algorithm,  $x_i^{(k)}$  denotes the local variable of node  $i$  at iteration  $k$  for every node  $i = 1, 2, \dots, N$  and iteration  $k = 0, 1, 2, \dots$ . At the iteration  $k$ , node  $i$  accesses  $\tilde{\nabla} f_i(x_i^{(k)}, z_i^{(k)})$  where  $z_i^{(k)}$  is a random variable independent of  $\{z_j^{(t)}\}_{j=1,2,\dots,i-1,i+1,\dots,N; t=1,\dots,k-1}$ .

We let  $\tilde{\nabla} f_i(x_i^{(k)})$  denote  $\tilde{\nabla} f_i(x_i^{(k)}, z_i^{(k)})$  and define the *gradient noise* as

$$\xi_i^{(k)} := \tilde{\nabla} f_i(x_i^{(k)}) - \nabla f_i(x_i^{(k)}), \quad i = 1, 2, \dots, N, \quad (3.3)$$

and we assume the stochastic gradient noise satisfies the following assumption.

**Assumption 2.** *For every  $i = 1, 2, \dots, N$  and  $k = 0, 1, 2, \dots$ , the gradient noise defined in (3.3) is conditionally unbiased with a finite second moment such that*

$$\mathbb{E} \left[ \xi_i^{(k+1)} \middle| \mathcal{F}_k \right] = 0, \quad \mathbb{E} \left\| \xi_i^{(k+1)} \right\|^2 \leq \sigma^2, \quad (3.4)$$

where  $\mathcal{F}_k$  is the natural filtration of the iterates  $\left( x_i^{(k)} \right)_{i=1}^N, \left( z_i^{(k)} \right)_{i=1}^N$  up to (and including) time  $k$ .

Then, we re-formulate EXTRA stochastic gradient Langevin dynamics as follows.

$$x_i^{(k+2)} = \sum_{j \in \Omega_i} W_{ij} x_j^{(k+1)} - \eta \nabla f_i \left( x_i^{(k+1)} \right) - \eta \xi_i^{(k+1)} + \sqrt{2\eta} w_i^{(k+2)}, \quad (3.5)$$

$$x_i^{(k+1)} = \sum_{j \in \Omega_i} \widetilde{W}_{ij} x_j^{(k)} - \eta \nabla f_i \left( x_i^{(k)} \right) - \eta \xi_i^{(k)} + \sqrt{2\eta} w_i^{(k+1)}. \quad (3.6)$$

These updates for  $N$  agents can also be expressed as

$$x^{(k+2)} = \mathcal{W} x^{(k+1)} - \eta \nabla F \left( x^{(k+1)} \right) - \eta \xi^{(k+1)} + \sqrt{2\eta} w^{(k+2)}, \quad (3.7)$$

$$x^{(k+1)} = \widetilde{\mathcal{W}} x^{(k)} - \eta \nabla F \left( x^{(k)} \right) - \eta \xi^{(k)} + \sqrt{2\eta} w^{(k+1)}, \quad (3.8)$$

where  $\mathcal{W} = W \otimes I_d$ ,  $\widetilde{\mathcal{W}} = \widetilde{W} \otimes I_d$ ,  $x^{(k)} = \left[ \left( x_1^{(k)} \right)^T, \dots, \left( x_N^{(k)} \right)^T \right]^T \in \mathbb{R}^{Nd}$  and

$$w^{(k)} = \left[ \left( w_1^{(k)} \right)^T, \dots, \left( w_N^{(k)} \right)^T \right]^T, \quad k = 0, 1, 2, \dots,$$

and we assume that the mixing matrices  $W, \widetilde{W}$  satisfy the following assumption. Such assumptions are made for analyzing the EXTRA methods and its generalizations [SLWY15, Jak18].

**Assumption 3.** *Consider a connected network  $\mathcal{G} = (\mathcal{V}, \mathcal{E})$  consisting of a set of agents  $\mathcal{V} = \{1, 2, \dots, n\}$  and a set of undirected edges  $\mathcal{E}$ . The doubly stochastic matrices  $W = [W_{ij}] \in \mathbb{R}^{N \times N}$  and  $\widetilde{W} = [\widetilde{W}_{ij}] \in \mathbb{R}^{N \times N}$  satisfy*

(1) *Null space property:*

$$\text{null} \left\{ W - \widetilde{W} \right\} = \text{span} \{ \mathbf{1}_N \}, \quad \text{null} \left\{ I_N - \widetilde{W} \right\} \supseteq \text{span} \{ \mathbf{1}_N \},$$

where  $\text{span} \{ \mathbf{1}_N \}$  is the span of the vector space supported by all-one vector  $[1_N^T, 1_N^T, \dots, 1_N^T]$ .

(2) *Spectral property:*

$$\widetilde{W} \succ 0, \quad \frac{I_N + W}{2} \succcurlyeq \widetilde{W} \succcurlyeq W.$$

The assumption implies  $W \succ -I_N$  and  $\frac{I_N + W}{2} \succcurlyeq W$ , so the eigenvalues of  $W$  lie in  $(-1, 1]$  and the eigenvalues of  $\widetilde{W}$  lie in  $(0, 1]$ . We will only consider  $\widetilde{W} \neq W$ ; if  $\widetilde{W} = W$ , then the EXTRA SGLD iterate in (3.11) reduces to DE-SGLD algorithm studied by [GGHZ21]. We assume that

$$\widetilde{W} = h I_N + (1-h) W, \quad h \in (0, 1/2]. \quad (3.9)$$

Note that the definition of  $\widetilde{W}$  satisfies Assumption 3, where we can compute that  $h(I_N - W) \succcurlyeq 0$ , which implies  $hI_N + (1-h)W \succcurlyeq W$ , and it is clear that  $\frac{I_N+W}{2} \succcurlyeq hI_N + (1-h)W$  with  $h \leq 1/2$ .

In the noiseless case, EXTRA has a primal-dual interpretation as a gradient descent ascent on a particular energy function [Jak18]. EXTRA was proposed by [SLWY15], and its unification and generalization was studied by [Jak18] to solve the dual optimization problem, where the author showed the iterates converges to the exact optimal solution if the parameters are chosen appropriately. Motivated by this work for decentralized optimization, we aim to propose a generalized EXTRA stochastic gradient Langevin dynamics which can produce the exact target distribution. We can use some algebraic transformation to generalize EXTRA SGLD algorithm in (3.7)-(3.8). By subtracting (3.8) from (3.7), the updating iterates follow

$$\begin{aligned} x^{(k+2)} - x^{(k+1)} &= \mathcal{W}x^{(k+1)} - \widetilde{\mathcal{W}}x^{(k)} - \eta \left( \nabla F(x^{(k+1)}) - \nabla F(x^{(k)}) \right) \\ &\quad - \eta (\xi^{(k+1)} - \xi^{(k)}) + \sqrt{2\eta} (w^{(k+2)} - w^{(k+1)}), \end{aligned} \quad (3.10)$$

where  $\xi^{(k)} = \left[ \left( \xi_1^{(k)} \right)^T, \left( \xi_2^{(k)} \right)^T, \dots, \left( \xi_N^{(k)} \right)^T \right]^T$  for every  $k$ . Next, we sum up the subtraction terms  $(x^{(2)} - x^{(1)}), (x^{(3)} - x^{(2)}), \dots, (x^{(k+2)} - x^{(k+1)})$  in (3.10), and by telescopic cancellation, we obtain

$$x^{(k+2)} = \mathcal{W}x^{(k+1)} - \eta \nabla F(x^{(k+1)}) - \eta \xi^{(k+1)} + \sum_{h=0}^k (\mathcal{W} - \widetilde{\mathcal{W}}) x^{(h)} + \sqrt{2\eta} w^{(k+1)}, \quad (3.11)$$

and equivalently,

$$x^{(k+2)} = \widetilde{\mathcal{W}}x^{(k+1)} - \eta \nabla F(x^{(k+1)}) - \eta \xi^{(k+1)} - \sum_{h=0}^{k+1} (\widetilde{\mathcal{W}} - \mathcal{W}) x^{(h)} + \sqrt{2\eta} w^{(k+1)}, \quad (3.12)$$

provided  $x^{(1)} = \mathcal{W}x^{(0)} - \eta \nabla f(x^{(0)}) - \eta \xi^{(0)} + \sqrt{2\eta} w^{(1)}$ . Since  $U = \widetilde{W} - W$  is positive semi-definite, we are able to have the following matrix decomposition  $U^{1/2} = PD^{1/2}P^T$ , where  $D$  is diagonal with non-negative diagonal entries and  $P$  is an orthogonal matrix. Thus, we can introduce an auxiliary sequence as follows:

$$q^{(k)} = \mathcal{U}^{1/2} \sum_{h=0}^{(k)} x^{(h)}, \quad \mathcal{U} = U \otimes I_d, \quad U = \widetilde{W} - W \in \mathbb{R}^{N \times N}. \quad (3.13)$$

Moreover, it is easy to observe that

$$q^{(k+1)} = q^{(k)} + \mathcal{U}^{1/2} x^{(k+1)}. \quad (3.14)$$

Thus, we obtain the following  $2(Nd)$ -dimensional recursive expression:

$$x^{(k+1)} = x^{(k)} - \eta \left( \frac{1}{\eta} (I_{Nd} - \widetilde{\mathcal{W}}) x^{(k)} + \nabla F(x^{(k)}) + \xi^{(k)} + \frac{1}{\eta} \mathcal{U}^{1/2} q^{(k)} \right) + \sqrt{2\eta} w^{(k+1)}, \quad (3.15)$$

$$q^{(k+1)} = q^{(k)} + \mathcal{U}^{1/2} x^{(k+1)}, \quad \mathcal{U} = \widetilde{\mathcal{W}} - \mathcal{W}. \quad (3.16)$$

By denoting

$$v^{(k)} = \frac{1}{\eta} \mathcal{U}^{1/2} q^{(k)}, \quad k = 0, 1, 2, \dots, \quad (3.17)$$

we get from (3.15) that

$$v^{(k)} + \nabla F(x^{(k)}) + \xi^{(k)} - \frac{1}{\eta} \widetilde{\mathcal{W}} x^{(k)} - \sqrt{\frac{2}{\eta}} w^{(k+1)} = -\frac{1}{\eta} x^{(k+1)}. \quad (3.18)$$

Moreover, we can compute from (3.16) and (3.17) that

$$v^{(k+1)} - v^{(k)} = \frac{1}{\eta} \mathcal{U}^{1/2} q^{(k+1)} - \frac{1}{\eta} \mathcal{U}^{1/2} q^{(k)} = \frac{1}{\eta} \mathcal{U}^{1/2} (q^{(k)} + \mathcal{U}^{1/2} x^{(k+1)}) - \frac{1}{\eta} \mathcal{U}^{1/2} q^{(k)} = \frac{1}{\eta} \mathcal{U} x^{(k+1)}. \quad (3.19)$$

Hence, we can re-write (3.16) as the following update:

$$v^{(k+1)} = v^{(k)} + \frac{1}{\eta} \mathcal{U} x^{(k+1)}, = v^{(k)} - \mathcal{U} \left( v^{(k)} + \nabla F(x^{(k)}) + \xi^{(k)} - \frac{1}{\eta} \widetilde{\mathcal{W}} x^{(k)} - \sqrt{\frac{2}{\eta}} w^{(k+1)} \right). \quad (3.20)$$

We introduce the *generalized EXTRA stochastic gradient Langevin dynamics* as follows

$$x^{(k+1)} = \widetilde{\mathcal{W}} x^{(k)} - \eta \left( \nabla F(x^{(k)}) + v^{(k)} \right) - \eta \xi^{(k)} + \sqrt{2\eta} w^{(k+1)}, \quad (3.21)$$

$$v^{(k+1)} = v^{(k)} - \mathcal{U} \left( v^{(k)} + \nabla F(x^{(k)}) - \mathcal{B} x^{(k)} \right) - \mathcal{U} \xi^{(k)} + \mathcal{U} \sqrt{\frac{2}{\eta}} w^{(k+1)}, \quad \mathcal{U} = \widetilde{\mathcal{W}} - \mathcal{W}. \quad (3.22)$$

We can observe from the iterates (3.21)-(3.22) that if we choose  $\mathcal{U} = 0_N \otimes I_d$ , then it reduces iterative updates to the DE-SGLD algorithm in [GGHZ21]. We will study the algorithm with the matrix  $\widetilde{\mathcal{W}} = \widetilde{W} \otimes I_d$  defined in (3.9), and the matrix  $\mathcal{B} = B \otimes I_d$  has the property

$$1_N^T B = c \quad \text{with } c \in \mathbb{R}. \quad (3.23)$$

The corresponding deterministic optimization algorithm without gradient noise was studied in [SLWY15] and [Jak18]<sup>5</sup> with  $\widetilde{W} = \frac{I_N + W}{2}$ , which corresponds to our choice on  $\widetilde{W}$  when  $h = 1/2$ . [GGHZ21] studied decentralized SGLD, it corresponds to let  $h = 0$  in our algorithm, that is  $U = 0_N$ . We also note that by taking  $\mathcal{B} = \widetilde{\mathcal{W}}/\eta$ , the algorithm (3.21)-(3.22) reduces to EXTRA SGLD algorithm in (3.1)-(3.2). In particular, [Jak18] considered the case  $\mathcal{B} = bI_{Nd}$  with  $b > 0$ .

## 4 Convergence Analysis

In this section, we provide the main results of the paper. Our non-asymptotic convergence analysis provides the convergence guarantees for the 2-Wasserstein distance between the law of the average of the iterates  $\bar{x}^{(K)}$  and the target distribution  $\pi$ , as well as the average of the 2-Wasserstein distance between the law of the individual iterates  $x_i^{(K)}$  and the target distribution  $\pi$ .

---

<sup>5</sup>We note [Jak18] exchanged the notations  $\widetilde{\mathcal{W}}$  and  $\mathcal{W}$  to get their Equations (16)-(17) and Lemma 3. In their notation, they have  $\mathcal{L} := I_{Nd} - \mathcal{W} = \mathcal{W} - \widetilde{\mathcal{W}}$  by using  $W = \frac{I_N + \widetilde{W}}{2}$  in their algorithm, it corresponds to  $\mathcal{U} = \widetilde{\mathcal{W}} - \mathcal{W}$  in ours. In their algorithm, they only have the notation  $\mathcal{W}$ , so they further denote  $\widetilde{\mathcal{W}} = \mathcal{W} - \mathcal{J}$  in their proofs which is different from our choice on  $\widetilde{W}$  in (3.9).

**Theorem 4.** Consider the generalized EXTRA Langevin dynamics with the network averaging matrix  $\widetilde{W} = hI_N + (1-h)W$  where

$$0 < h \leq \frac{1 - \bar{\gamma}_w}{4\bar{\gamma}_{I_N-W}^2} \wedge \frac{1}{2} \wedge \frac{1}{\gamma_1 \gamma_2}, \quad (4.1)$$

and assume that the stepsize  $\eta$  is chosen satisfying

$$0 < \eta < \frac{1}{h\gamma_1 \gamma_2} \wedge \frac{\bar{\gamma}_{\widetilde{W}}}{6(L+\mu) \vee 2A} \wedge 1 \wedge \frac{1}{L+\mu} \wedge \frac{\bar{\gamma}_{\widetilde{W}}}{6(L+\mu)}, \quad (4.2)$$

where  $\gamma_1, \gamma_2, \bar{\gamma}_{\widetilde{W}}, \bar{\gamma}_{I_N-W}^2$  are constants defined in Table 1. Then, for any  $K \geq K_0$ , the following bound holds:

$$\begin{aligned} \mathcal{W}_2 \left( \mathcal{L} \left( \bar{x}^{(K)} \right), \pi \right) &\leq \left( \frac{\bar{\gamma}_{\widetilde{W}}^{2K} - \left( 1 - \eta \mu \left( 1 - \frac{\eta L}{2} \right) \right)^K}{\bar{\gamma}_{\widetilde{W}}^2 - 1 + \eta \mu \left( 1 - \frac{\eta L}{2} \right)} \right)^{1/2} \frac{2L\bar{\gamma}_{\widetilde{W}}}{\sqrt{N}} \left( \mathbb{E} \|x^{(0)}\|^2 \right)^{1/2} \\ &\quad + (1 - \mu\eta)^K \mathcal{W}_2 \left( \mathcal{L} \left( x_0 \right), \pi \right) + \sqrt{\eta} \mathcal{E}_1, \end{aligned} \quad (4.3)$$

where we have

$$\begin{aligned} \mathcal{E}_1 := &\left( \frac{\eta}{\mu \left( 1 - \frac{\eta L}{2} \right)} + \frac{(1 + \eta L)^2}{\mu^2 \left( 1 - \frac{\eta L}{2} \right)^2} \right)^{1/2} \cdot \left( \frac{4L^2(R_h + R'_h)\eta}{N(1 - \bar{\gamma}_{\widetilde{W}})^2} + \frac{4L^2\sigma^2\eta}{1 - \bar{\gamma}_{\widetilde{W}}^2} + \frac{8L^2d}{1 - \bar{\gamma}_{\widetilde{W}}^2} \right)^{1/2} \\ &+ \frac{\sigma}{\sqrt{\mu \left( 1 - \frac{\eta L}{2} \right) N}} + \frac{1.65L}{\mu} \sqrt{dN^{-1}}. \end{aligned} \quad (4.4)$$

Moreover,

$$\begin{aligned} &\frac{1}{N} \sum_{i=1}^N \mathcal{W}_2 \left( \mathcal{L} \left( x_i^{(K)} \right), \pi \right) \\ &\leq \eta \cdot \frac{D_1}{\sqrt{N}} + \sqrt{\eta} \cdot (D_2 + \mathcal{E}_1) + \left( \frac{\bar{\gamma}_{\widetilde{W}}^{2K} - \left( 1 - \eta \mu \left( 1 - \frac{\eta L}{2} \right) \right)^K}{\bar{\gamma}_{\widetilde{W}}^2 - 1 + \eta \mu \left( 1 - \frac{\eta L}{2} \right)} \right)^{1/2} \frac{2L\bar{\gamma}_{\widetilde{W}}}{\sqrt{N}} \left( \mathbb{E} \|x^{(0)}\|^2 \right)^{1/2} \\ &\quad + (1 - \mu\eta)^K \mathcal{W}_2 \left( \mathcal{L} \left( x_0 \right), \pi \right) + \frac{2(\bar{\gamma}_{\widetilde{W}})^K}{\sqrt{N}} \sqrt{\mathbb{E} [\|x^{(0)}\|^2]}, \end{aligned} \quad (4.5)$$

where the constants  $R_h$  and  $R'_h$  are made explicit and given in Table 1.

## 5 Comparison with DE-SGLD

In this section, we are interested in comparing our generalized EXTRA SGLD method with the DE-SGLD method in the literature [GGHZ21]. In particular, we highlight the dependence on strong-convexity constant  $\mu$ , the smoothness constant  $L$ , the dimension  $d$  and the accuracy level  $\varepsilon$ . We have the following proposition.

**Proposition 5.** For DE-SGLD, under the assumptions in Theorem 1 in [GGHZ21], as  $\varepsilon \rightarrow 0$ ,

$$\frac{1}{N} \sum_{i=1}^N \mathcal{W}_2^{\text{de-sgld}} \left( \mathcal{L} \left( x_i^{(K)} \right), \pi \right) \leq \mathcal{O}(\varepsilon), \quad (5.1)$$

provided that

$$K \geq K^{\text{de-sgld}} = \tilde{\mathcal{O}} \left( \frac{L^4 d}{\varepsilon^2 \mu^3} \right) \quad (5.2)$$

where  $\tilde{\mathcal{O}}$  hides the logarithmic dependence on  $\varepsilon$ .

For generalized EXTRA SGLD, under the assumptions in Theorem 4, as  $\varepsilon \rightarrow 0$ , by taking  $h \geq \Omega(\eta\mu)$  and  $h < \frac{1}{(L/\mu)^4(L+\|B\|^2)} \wedge \frac{1}{2\gamma_1\gamma_2}$ , it holds that

$$\frac{1}{N} \sum_{i=1}^N \mathcal{W}_2^{\text{extra-sgld}} \left( \mathcal{L} \left( x_i^{(K)} \right), \pi \right) \leq \mathcal{O}(\varepsilon), \quad (5.3)$$

provided that

$$K \geq K^{\text{extra-sgld}} = \tilde{\mathcal{O}} \left( \frac{L^2 d}{\varepsilon^2 \mu^3} \right), \quad (5.4)$$

where  $\tilde{\mathcal{O}}$  hides the logarithmic dependence on  $\varepsilon$  and the constants  $\gamma_1, \gamma_2$  are provided in Table 1.

By comparing the complexities of DE-SGLD in (5.2) to generalized EXTRA SGLD in (5.4), we find that generalized EXTRA SGLD achieves an improvement on the order of  $\tilde{\mathcal{O}}(L^2)$ . When  $\mu$  is large (and therefore  $L$  is large), the Gibbs distribution  $\pi \propto e^{-f}$  becomes concentrated around the minimizer of  $f$ , making the sampling problem approximately equivalent to the global optimization problem  $\min_{x \in \mathbb{R}^d} f(x)$ . In the decentralized optimization setting, EXTRA is known to improve upon decentralized gradient descent, and the improvement achieved with generalized EXTRA SGLD in the sampling setting is analogous. Intuitively, DE-SGLD introduces a bias, which is evident in the constant  $D$  in (6.100) that results from the agents' gradient updates. The generalized EXTRA SGLD algorithm corrects this bias in the agents' gradient updates, leading to improved performance.

## 6 Proof of the Main Results

In this section, we provide the proofs of Theorem 4 and Proposition 5.

### 6.1 Proof of Theorem 4

In this section, we provide the proof of Theorem 4 via establishing a sequence of key technical results whose proofs will be provided in Appendix A. In order to derive Theorem 4, based on the triangle inequality, we consider the following decomposition:

$$\frac{1}{N} \sum_{i=1}^N \mathcal{W}_2 \left( \mathcal{L} \left( x_i^{(k)} \right), \pi \right) \leq \frac{1}{N} \sum_{i=1}^N \mathcal{W}_2 \left( \mathcal{L} \left( x_i^{(k)} \right), \mathcal{L} \left( \bar{x}^{(k)} \right) \right) + \mathcal{W}_2 \left( \mathcal{L} \left( \bar{x}^{(k)} \right), \pi \right), \quad (6.1)$$

Constants	Source
$\gamma_{\bar{w}} =  \lambda_2^{\bar{w}} ^2 1_{0 <  \lambda_2^{\bar{w}} ^2 < \frac{1}{2}} + \frac{ \lambda_2^{\bar{w}} ^2 ( \lambda_2^{\bar{w}} ^2 - \frac{1}{2})}{1 -  \lambda_2^{\bar{w}} ^2} 1_{\frac{1}{2} \leq  \lambda_2^{\bar{w}} ^2 \leq \frac{2}{3}} + \frac{5 \lambda_2^{\bar{w}} ^2 - 3 \lambda_2^{\bar{w}} ^4 - 2}{3 \lambda_2^{\bar{w}} ^2 - 1} 1_{\frac{2}{3} \leq  \lambda_2^{\bar{w}} ^2 < 1}$	(6.20)
$A = \left( \frac{L}{\mu} - 1 + \frac{\gamma_{\bar{w}}}{2(1 + \mu/L)} \right) \cdot \frac{4L^2}{N^2} \left( 1 + \frac{2+2L}{\mu} \right)$	(6.23)
$\bar{\gamma}_{I_N-W} = \max \{ 1 -  \lambda_2^W , 1 -  \lambda_N^W  \}, \quad \bar{\gamma}_w := \max \{  \lambda_2^W ,  \lambda_N^W  \}$	(6.25)
$\gamma_1 = \frac{1}{\gamma_{\bar{w}}} \left( \frac{1}{L} + 2 + \frac{1}{L\mu} \right), \quad \gamma_2 = \frac{12(L^2 + L\ B\ ^2)}{(1 - \bar{\gamma}_w)(1 - \bar{\gamma}_{I_N-W}^2)} \left( 1 + \frac{4L^2 \left( 1 + \frac{2+2L}{\mu} \right)}{N^2\mu} \right)$	(6.30)
$w_1 = 2 \left( \frac{N^2 + 1}{\gamma_{\bar{w}}} + \frac{4}{\gamma_{\bar{w}}} \cdot \left( \frac{L}{\mu} + 3\eta L - 1 \right) \right), \quad w_2 = \frac{8 \left( 6(L^2 + L\ B\ ^2) + N^2\mu \right)}{N\mu(1 - \bar{\gamma}_w)(1 - \bar{\gamma}_{I_N-W}^2)}$	(6.31), (6.32)
$E_1 = \frac{8}{\gamma_{\bar{w}}} (L/\mu + 3\eta L - 1), \quad E_2 = \frac{2}{\gamma_{\bar{w}}}$	(6.33)
$E_3 = \frac{12(L^2 + L\ B\ ^2)}{\mu(1 - \bar{\gamma}_w)(1 - \bar{\gamma}_{I_N-W}^2)}, \quad E_4 = \frac{4}{(1 - \bar{\gamma}_w)(1 - \bar{\gamma}_{I_N-W}^2)}$	(6.34)
$R_h = h\delta^2 \left( \frac{C_1\gamma_2}{2L^2} + \frac{C_0\gamma_1\gamma_2}{2L^2} \right) + (h/\eta)\delta^2 \left( \gamma_2 D_0 + \frac{w_2}{N} (\eta\sigma^2 + 2d) \right) + \ \nabla F(\mathbf{x}_*)\ ^2$	(6.51)
$R'_h = \eta\delta^2 (C_1 + C_3 + \gamma_1 C_0 + D_0 C_2) + \delta^2\eta^2 \left( \frac{C_1 C_2}{2L^2} + \frac{\gamma_1 C_0 C_2}{2L^2} \right) + 3 \ \nabla F(\mathbf{x}_*)\ ^2$	(6.52)
$K_0 = \frac{\delta^2}{1 - \delta^2} \left[ \left( 1 - \frac{\ \nabla F(\mathbf{x}_*)\ ^2}{D_0 + C_4} \right) \vee \left( 1 - \frac{\ \nabla F(\mathbf{x}_*)\ ^2}{C_0} \right) \right] \vee 0$	(6.53)
$C_0 = \left( (h/\eta)E_3 \mathbb{E} \left[ \left\  \bar{e}_x^{(0)} \right\ ^2 \right] + E_4 \mathbb{E} \left[ \left\  \tilde{v}^{(0)} \right\ ^2 \right] \right) \cdot \frac{2L^2}{1 - \eta\gamma_1\gamma_2}$	(6.45)
$C_1 = \frac{2L^2(\eta\sigma^2 + 2d)}{N} \cdot \frac{w_2\gamma_1(h/\eta) + w_1}{1 - h\gamma_1\gamma_2}, \quad C_2 = \frac{2L^4 \left( \eta + \frac{1+\eta L}{\mu(1-\frac{\eta L}{2})} \right)}{N^2 \left( \delta^2 + \eta\mu \left( 1 - \frac{\eta L}{2} \right) - 1 \right)}$	(6.46), (6.47)
$C_3 = \frac{2L^2}{N} \cdot \frac{\eta\sigma^2 + 2d}{\delta^2 + \eta\mu \left( 1 - \frac{\eta L}{2} \right) - 1}, \quad C_4 = \frac{2L^2}{\delta^2 + \eta\mu \left( 1 - \frac{\eta L}{2} \right) - 1} \mathbb{E} \left[ \left\  \bar{e}_x^{(0)} \right\ ^2 \right]$	(6.48), (6.49)
$D_0 = \frac{1}{1 - h\gamma_1\gamma_2} \left( E_1 \mathbb{E} \left[ \left\  \tilde{x}^{(0)} \right\ ^2 \right] + E_2 \mathbb{E} \left[ \left\  \bar{e}_x^{(0)} \right\ ^2 \right] \right)$	(6.42)
$D_1 = 2 \frac{\sqrt{2(R_h + R'_h)}}{1 - \bar{\gamma}_{\bar{w}}} + \frac{2\sigma}{\sqrt{1 - \bar{\gamma}_{\bar{w}}^2}}, \quad D_2 = 2 \sqrt{\frac{2d}{1 - \bar{\gamma}_{\bar{w}}^2}}$	(6.42), (6.58)
$\delta^2 \in \left[ \left( 1 - \frac{\eta\mu}{2} \left( 1 - \frac{\eta L}{2} \right) \right) \vee \left( 1 - h \frac{1 - \bar{\gamma}_w}{4} (1 - \bar{\gamma}_{I_N-W}) \right), 1 \right)$	(6.39)

Table 1: Summary of the constants and where they are defined in the text.

where

$$\mathcal{W}_2 \left( \mathcal{L} \left( \bar{x}^{(k)} \right), \pi \right) \leq \mathcal{W}_2 \left( \mathcal{L} \left( \bar{x}^{(k)} \right), \mathcal{L}(x_k) \right) + \mathcal{W}_2 \left( \mathcal{L}(x_k), \pi \right), \quad (6.2)$$

$\bar{x}^{(k)} := \frac{1}{N} \sum_{i=1}^N x_i(k)$  is the average iterates and  $x_k$  has the iterates

$$x_{k+1} = x_k - \frac{\eta}{N} \nabla f(x_k) + \sqrt{2\eta} \bar{w}^{(k+1)}. \quad (6.3)$$

These iterates correspond to the Euler-Maruyama discretization of overdamped Langevin diffusion

$$dX_t = -\frac{1}{N} \nabla f(X_t) dt + \sqrt{2N^{-1}} dW_t, \quad (6.4)$$

where  $W_t$  is a standard  $d$ -dimensional Brownian motion,  $\bar{w}^{(k)} := \frac{1}{N} \sum_{i=1}^N w_i^{(k)}$ , and  $w_i^{(k)}$  are  $\mathcal{N}(0, I_d)$  distributed that are i.i.d. in both  $k \in \mathbb{N}$  and  $i = 1, 2, \dots, N$ .

The main idea of our proof technique is to bound the following three terms: (1) the  $L^2$  distance between  $x_i^{(k)}$  and their average  $\bar{x}^{(k)}$ ; (2) the  $L^2$  distance between the average iterate  $\bar{x}^{(k)}$  and iterates  $x_k$  in (6.3) obtained from Euler-Maruyama discretization of overdamped SDE (6.4); and (3) the  $\mathcal{W}_2$  distance between the law of  $x_k$  in (6.3) and the Gibbs distribution  $\pi$ . First, we upper bound the  $L^2$  distance between  $x_i^{(k)}$  and their average.

### 6.1.1 Uniform $L^2$ bounds between $x_i^{(k)}$ and their average $\bar{x}^{(k)}$

Denoting by  $\mathbf{a} = \{a^{(0)}, a^{(1)}, \dots, a^{(k)}, \dots\}$  an infinite sequence of vectors, where  $a^{(k)} \in \mathbb{R}^p$ ,  $k = 0, 1, \dots$  for some  $p \in \mathbb{N}$ . For a fixed  $\delta \in (0, 1)$ , we define the following quantity

$$\|\mathbf{a}\|_2^{\delta, K} := \max_{k=0,1,\dots,K} \mathbb{E} \left[ \left\| \frac{1}{\delta^k} a^{(k)} \right\|^2 \right]. \quad (6.5)$$

We first state a preliminary lemma that will be frequently used in the following analysis, and this lemma is a modification of Lemma 6 in [Jak18], and its proof will be provided in Appendix B.

**Lemma 6.** Consider two infinite random sequences  $\mathbf{a} = \{a^{(0)}, a^{(1)}, \dots, a^{(k)}, \dots\}$  and  $\mathbf{b} = \{b^{(0)}, b^{(1)}, \dots, b^{(k)}, \dots\}$ , with  $a^{(k)}, b^{(k)} \in \mathbb{R}^p$  for some  $p \in \mathbb{N}$  such that  $\mathbb{E} \|a^{(k)}\|^2 < \infty$  and  $\mathbb{E} \|b^{(k)}\|^2 < \infty$  for every  $k \geq 0$ . Suppose that, for all  $k = 0, 1, \dots$ , there holds:

$$\mathbb{E} \|a^{(k+1)}\|^2 \leq c_1 \mathbb{E} \|a^{(k)}\|^2 + c_2 \mathbb{E} \|b^{(k)}\|^2 + c_0. \quad (6.6)$$

where  $c_i \geq 0$ ,  $i = 0, 1, 2$ . Then, for all  $K = 0, 1, \dots$ , for any  $\delta \in (0, 1)$ , we have:

$$\|\mathbf{a}\|_2^{\delta, K} \leq \frac{c_1}{\delta^2} \|\mathbf{a}\|_2^{\delta, K} + \frac{c_2}{\delta^2} \|\mathbf{b}\|_2^{\delta, K} + \frac{c_0}{\delta^{2K}} + \mathbb{E} \|a^{(0)}\|^2. \quad (6.7)$$

Next, we introduce the following technical lemma, which is an extension of Lemma 6 and this extension will be used in the proof of Lemma 10.

**Lemma 7.** Given any  $n \in \mathbb{N}$  with  $n \geq 2$ , if

$$\mathbb{E} \|a^{(k+1)}\|^2 \leq c_1 \mathbb{E} \|a^{(k)}\|^2 + \sum_{i=2}^n c_i \mathbb{E} \|b_i^{(k)}\|^2 + c_0, \quad (6.8)$$

for every  $k = 0, 1, \dots, K$ , then

$$\|\mathbf{a}\|_2^{\delta, K} \leq \frac{c_1}{\delta^2} \|\mathbf{a}\|_2^{\delta, K} + \sum_{i=2}^n \frac{c_i}{\delta^2} \|\mathbf{b}_i\|_2^{\delta, K} + \frac{c_0}{\delta^{2K}} + \mathbb{E} \|a^{(0)}\|^2. \quad (6.9)$$

Next, we define the error vectors:

$$e_x^{(k)} := x^{(k)} - \mathbf{x}_*, \quad e_v^{(k)} := v^{(k)} + \nabla F(\mathbf{x}_*), \quad e^{(k)} := \left( \left( e_x^{(k)} \right)^T, \left( e_v^{(k)} \right)^T \right)^T, \quad (6.10)$$

where  $\mathbf{x}_* = \left[ (x_*)^T, \dots, (x_*)^T \right]^T \in \mathbb{R}^{Nd}$  is the vector of minimizer of the objective from the target Gibbs distribution. For any  $k = 0, 1, 2, \dots$ , let us further define

$$\bar{\mathbf{x}}^{(k)} := \left( \left( \bar{x}^{(k)} \right)^T, \left( \bar{x}^{(k)} \right)^T, \dots, \left( \bar{x}^{(k)} \right)^T \right)^T \in \mathbb{R}^{Nd}, \quad (6.11)$$

$$\bar{\mathbf{v}}^{(k)} := \left( \left( \bar{v}^{(k)} \right)^T, \left( \bar{v}^{(k)} \right)^T, \dots, \left( \bar{v}^{(k)} \right)^T \right)^T \in \mathbb{R}^{Nd}, \quad (6.12)$$

where  $\bar{x}^{(k)} := \frac{1}{N} \sum_{i=1}^N x_i^{(k)}$  and  $\bar{v}^{(k)} := \frac{1}{N} \sum_{i=1}^N v_i^{(k)}$ . By introducing the following quantities:

$$\tilde{x}^{(k)} := x^{(k)} - \bar{\mathbf{x}}^{(k)}, \quad \tilde{v}^{(k)} := v^{(k)} - \bar{\mathbf{v}}^{(k)}, \quad (6.13)$$

we define the average errors as follows.

$$\bar{e}_x^{(k)} := \frac{1}{N} \sum_{i=1}^N (x_i^{(k)} - x_*), \quad \bar{e}_v^{(k)} := \frac{1}{N} \sum_{i=1}^N (v_i^{(k)} + \nabla f_i(x_*)). \quad (6.14)$$

Now we can decompose the error terms  $e_x^{(k)}$  and  $e_v^{(k)}$  in (6.10) and get the following lemma.

**Lemma 8.** For all  $k = 0, 1, 2, \dots$ , the error terms  $e_x^{(k)}$  and  $e_v^{(k)}$  have the decomposition:

$$e_x^{(k)} = \tilde{x}^{(k)} + 1_N \otimes \bar{e}_x^{(k)}, \quad e_v^{(k)} = \tilde{v}^{(k)} + 1_N \otimes \bar{e}_v^{(k)}, \quad (6.15)$$

with

$$\tilde{v}^{(k)} = e_v^{(k)}, \quad \bar{e}_v^{(k)} = \bar{v}^{(k)} = 0. \quad (6.16)$$

Next, to facilitate the presentations, let us define two sequences  $\tilde{\mathbf{x}} := \{\tilde{x}^{(0)}, \tilde{x}^{(1)}, \dots, \tilde{x}^{(k)}, \dots\}$  and  $\tilde{\mathbf{v}} := \{\tilde{v}^{(0)}, \tilde{v}^{(1)}, \dots, \tilde{v}^{(k)}, \dots\}$  where  $\tilde{x}^{(k)}, \tilde{v}^{(k)} \in \mathbb{R}^{Nd}$  are given in (6.13). By following the notation in (6.5), we denote

$$\|\tilde{\mathbf{x}}\|_2^{\delta, K} := \max_{k=0,1,\dots,K} \mathbb{E} \left[ \left\| \frac{1}{\delta^k} \tilde{x}^{(k)} \right\|^2 \right], \quad \|\tilde{\mathbf{v}}\|_2^{\delta, K} = \max_{k=0,1,\dots,K} \mathbb{E} \left[ \left\| \frac{1}{\delta^k} \tilde{v}^{(k)} \right\|^2 \right]. \quad (6.17)$$

Similarly, we also define the sequence  $\bar{\mathbf{e}}_x = \left\{ \bar{e}_x^{(0)}, \bar{e}_x^{(1)}, \dots, \bar{e}_x^{(k)}, \dots \right\}$  where  $\bar{e}_x^{(k)} \in \mathbb{R}^d$  is defined in (6.14) and moreover, we denote

$$\|\bar{\mathbf{e}}_x\|_2^{\delta, K} := \max_{k=0,1,\dots,K} \mathbb{E} \left[ \left\| \frac{1}{\delta^k} \bar{e}_x^{(k)} \right\|^2 \right]. \quad (6.18)$$

Now we present a sequence of technical lemmas. First, we provide an upper bound on  $\|\bar{\mathbf{e}}_x\|_2^{\delta, K}$  by using  $\|\tilde{\mathbf{x}}\|_2^{\delta, K}$ .

**Lemma 9.** *Suppose Assumptions 1, 2, and 3 hold. Taking the stepsize  $0 < \eta < \frac{2}{L} \wedge 1$ , then for any  $\delta^2 \in \left(1 - \eta\mu \left(1 - \frac{\eta L}{2}\right), 1\right)$ , the following inequality holds for every  $K \geq 0$ :*

$$\begin{aligned} \|\bar{\mathbf{e}}_x\|_2^{\delta, K} &\leq \eta \cdot \frac{L^2}{N^2 \left(\delta^2 + \eta\mu \left(1 - \frac{\eta L}{2}\right) - 1\right)} \left( \eta + \frac{1 + \eta L}{\mu \left(1 - \frac{\eta L}{2}\right)} \right) \|\tilde{\mathbf{x}}\|_2^{\delta, K} \\ &\quad + \frac{\eta}{N\delta^{2K-2}} \cdot \frac{\eta\sigma^2 + 2d}{\delta^2 + \eta\mu \left(1 - \frac{\eta L}{2}\right) - 1} + \frac{\delta^2}{\delta^2 + \eta\mu \left(1 - \frac{\eta L}{2}\right) - 1} \mathbb{E} \left[ \left\| \bar{e}_x^{(0)} \right\|^2 \right]. \end{aligned} \quad (6.19)$$

Next, we provide an upper bound on  $\|\tilde{\mathbf{x}}\|_2^{\delta, K}$  in terms of  $\|\bar{\mathbf{e}}_x\|_2^{\delta, K}$  and  $\|\tilde{\mathbf{v}}\|_2^{\delta, K}$ .

**Lemma 10.** *Under the assumptions in Lemma 9, in addition, let the stepsize  $\eta \leq \frac{1}{L}$ . Denoting the eigenvalues of matrix  $\widetilde{W}$  such that  $1 = \lambda_1^{\widetilde{W}} > \lambda_2^{\widetilde{W}} \geq \dots \geq \lambda_N^{\widetilde{W}} > 0$ . Define the positive constant*

$$\gamma_{\widetilde{W}} := \begin{cases} \left| \lambda_2^{\widetilde{W}} \right|^2 & \text{if } 0 < \left| \lambda_2^{\widetilde{W}} \right|^2 < \frac{1}{2}, \\ \frac{\left| \lambda_2^{\widetilde{W}} \right|^2 \left( \left| \lambda_2^{\widetilde{W}} \right|^2 - \frac{1}{2} \right)}{1 - \left| \lambda_2^{\widetilde{W}} \right|^2} & \text{if } \frac{1}{2} \leq \left| \lambda_2^{\widetilde{W}} \right|^2 \leq \frac{2}{3}, \\ \frac{5\left| \lambda_2^{\widetilde{W}} \right|^2 - 3\left| \lambda_2^{\widetilde{W}} \right|^4 - 2}{3\left| \lambda_2^{\widetilde{W}} \right|^2 - 1} & \text{if } \frac{2}{3} \leq \left| \lambda_2^{\widetilde{W}} \right|^2 < 1. \end{cases} \quad (6.20)$$

By taking

$$0 < \eta \leq \frac{\gamma_{\widetilde{W}}}{6(L + \mu)}, \quad (6.21)$$

it holds that:

$$\begin{aligned} \|\tilde{\mathbf{x}}\|_2^{\delta, K} &\leq \frac{\eta\mu}{\gamma_{\widetilde{W}}} (L/\mu + 3\eta L - 1) \|\bar{\mathbf{e}}_x\|_2^{\delta, K} + \frac{\eta}{\gamma_{\widetilde{W}}} \left( \frac{1}{2L} + \eta + \frac{\eta}{2L\mu} \right) \|\tilde{\mathbf{v}}\|_2^{\delta, K} \\ &\quad + \frac{\eta}{\gamma_{\widetilde{W}}\delta^{2K-2}} \left( N + \frac{1}{N} \right) (\eta\sigma^2 + 2d) + \frac{\delta^2}{\gamma_{\widetilde{W}}} \mathbb{E} \left[ \left\| \tilde{x}^{(0)} \right\|^2 \right], \end{aligned} \quad (6.22)$$

where the constant  $\delta^2$  depends on  $\left| \lambda_2^{\widetilde{W}} \right|^2$  in three regimes:

- (1). If  $\left| \lambda_2^{\widetilde{W}} \right|^2 < \frac{1}{2}$ , then (6.22) holds for all  $\delta$  such that  $1 > \delta^2 \geq 2\left| \lambda_2^{\widetilde{W}} \right|^2$ ;
- (2). If  $\frac{2}{3} \geq \left| \lambda_2^{\widetilde{W}} \right|^2 \geq \frac{1}{2}$ , then (6.22) holds for all  $\delta$  such that  $1 > \delta^2 \geq \frac{\left| \lambda_2^{\widetilde{W}} \right|^2}{2(1 - \left| \lambda_2^{\widetilde{W}} \right|^2)} \geq \frac{1}{2}$ ;
- (3). If  $1 > \left| \lambda_2^{\widetilde{W}} \right|^2 > \frac{2}{3}$ , then (6.22) holds for all  $\delta$  such that  $1 > \delta^2 \geq \frac{4\left| \lambda_2^{\widetilde{W}} \right|^2 - 2}{3\left| \lambda_2^{\widetilde{W}} \right|^2 - 1} > \frac{1}{2}$ .

As a direct result from Lemmas 9 and 10, we obtain the following lemma that provides an upper bound on  $\|\tilde{\mathbf{x}}\|_2^{\delta,K}$  in terms of  $\|\tilde{\mathbf{v}}\|_2^{\delta,K}$ .

**Lemma 11.** Denote

$$A := \left( \frac{L}{\mu} - 1 + \frac{\gamma_{\tilde{w}}}{2(1 + \mu/L)} \right) \cdot \frac{4L^2}{N^2} \left( 1 + \frac{2+2L}{\mu} \right). \quad (6.23)$$

Given any  $\delta^2 \in \left[ 1 - \frac{\eta\mu}{2} \left( 1 - \frac{\eta L}{2} \right), 1 \right]$  under the conditions for  $\delta$  in Lemma 10, and suppose  $\eta \leq \frac{\gamma_{\tilde{w}}}{6(L+\mu)} \wedge \frac{\gamma_{\tilde{w}}}{2A}$ , there holds,

$$\begin{aligned} \|\tilde{\mathbf{x}}\|_2^{\delta,K} &\leq \eta \cdot \frac{2}{\gamma_{\tilde{w}}} \left( \frac{1}{2L} + \eta + \frac{\eta}{2L\mu} \right) \|\tilde{\mathbf{v}}\|_2^{\delta,K} \\ &\quad + \eta \cdot \frac{2}{\delta^{2K-2}} (\eta\sigma^2 + 2d) \left( \frac{N + \frac{1}{N}}{\gamma_{\tilde{w}}} + \frac{4}{N\gamma_{\tilde{w}}} \cdot (L/\mu + 3\eta L - 1) \right) \\ &\quad + \frac{8\delta^2}{\gamma_{\tilde{w}}} (L/\mu + 3\eta L - 1) \mathbb{E} \left[ \left\| \bar{e}_x^{(0)} \right\|^2 \right] + \frac{2\delta^2}{\gamma_{\tilde{w}}} \mathbb{E} \left[ \left\| \tilde{x}^{(0)} \right\|^2 \right], \end{aligned} \quad (6.24)$$

where  $\gamma_{\tilde{w}}$  defined in (6.20) depending on three regimes in Lemma 10.

Next, we define the quantities

$$\bar{\gamma}_{I_N-W} := \max \{1 - |\lambda_2^W|, 1 - |\lambda_N^W|\}, \quad \bar{\gamma}_w := \max \{|\lambda_2^W|, |\lambda_N^W|\}, \quad (6.25)$$

so that  $1 > \bar{\gamma}_{I_N-W} \geq 1 - \bar{\gamma}_w > 0$ . In the following lemma, we derive an upper bound on  $\|\tilde{\mathbf{v}}\|_2^{\delta,K}$  in terms of  $\|\tilde{\mathbf{x}}\|_2^{\delta,K}$ .

**Lemma 12.** By taking

$$0 < h \leq \frac{1 - \bar{\gamma}_w}{4\bar{\gamma}_{I_N-W}^2} \wedge \frac{1}{2}, \quad (6.26)$$

and  $\delta^2 \geq 1 - h \frac{1 - \bar{\gamma}_w}{4} (1 - \bar{\gamma}_{I_N-W}) > 0$  in three regimes defined in Lemma 10, the following bound holds:

$$\begin{aligned} \|\tilde{\mathbf{v}}\|_2^{\delta,K} &\leq \frac{12(h/\eta) \left( L^2 + L \|B\|^2 \right)}{(1 - \bar{\gamma}_w) (1 - \bar{\gamma}_{I_N-W}^2)} \left( 1 + \frac{4L^2 \left( 1 + \frac{2+2L}{\mu} \right)}{N^2 \mu} \right) \|\tilde{\mathbf{x}}\|_2^{\delta,K} \\ &\quad + \left( \frac{6 \left( L^2 + L \|B\|^2 \right)}{N\mu} + N \right) \cdot \frac{8(h/\eta)}{(1 - \bar{\gamma}_w) (1 - \bar{\gamma}_{I_N-W}^2)} \cdot \frac{\eta\sigma^2 + 2d}{\delta^{2K-2}} \\ &\quad + \frac{12\delta^2(h/\eta) \left( L^2 + L \|B\|^2 \right)}{\eta\mu(1 - \bar{\gamma}_w) (1 - \bar{\gamma}_{I_N-W}^2)} \mathbb{E} \left[ \left\| \bar{e}_x^{(0)} \right\|^2 \right] + \frac{4\delta^2}{h(1 - \bar{\gamma}_w) (1 - \bar{\gamma}_{I_N-W}^2)} \left\| \tilde{v}^{(0)} \right\|^2. \end{aligned} \quad (6.27)$$

Now one can immediately derive from Lemma 11 and Lemma 12 that

$$\|\tilde{\mathbf{x}}\|_2^{\delta,K} \leq \eta\gamma_1 \|\tilde{\mathbf{v}}\|_2^{\delta,K} + \eta \frac{w_1(\eta\sigma^2 + 2d)}{N\delta^{2K-2}} + \delta^2 E_1 \mathbb{E} \left[ \left\| \bar{e}_x^{(0)} \right\|^2 \right] + \delta^2 E_2 \mathbb{E} \left[ \left\| \tilde{x}^{(0)} \right\|^2 \right], \quad (6.28)$$

$$\begin{aligned} \|\tilde{\mathbf{v}}\|_2^{\delta,K} &\leq (h/\eta)\gamma_2 \|\tilde{\mathbf{x}}\|_2^{\delta,K} + (h/\eta) \frac{w_2(\eta\sigma^2 + 2d)}{N\delta^{2K-2}} \\ &\quad + \delta^2(h/\eta)(E_3/\eta) \mathbb{E} \left[ \left\| \bar{e}_x^{(0)} \right\|^2 \right] + \delta^2(E_4/h) \mathbb{E} \left[ \left\| \tilde{v}^{(0)} \right\|^2 \right], \end{aligned} \quad (6.29)$$

where the constants are defined as:

$$\gamma_1 := \frac{1}{\gamma_{\bar{w}}} \left( \frac{1}{L} + 2 + \frac{1}{L\mu} \right), \quad \gamma_2 := \frac{12 \left( L^2 + L \|B\|^2 \right)}{(1 - \bar{\gamma}_w) (1 - \bar{\gamma}_{I_N-W}^2)} \left( 1 + \frac{4L^2 \left( 1 + \frac{2+2L}{\mu} \right)}{N^2 \mu} \right), \quad (6.30)$$

and

$$w_1 := 2 \left( \frac{N^2 + 1}{\gamma_{\bar{w}}} + \frac{4}{\gamma_{\bar{w}}} \cdot (L/\mu + 3\eta L - 1) \right), \quad (6.31)$$

$$w_2 := \left( \frac{6 \left( L^2 + L \|B\|^2 \right)}{N\mu} + N \right) \cdot \frac{8}{(1 - \bar{\gamma}_w) (1 - \bar{\gamma}_{I_N-W}^2)}, \quad (6.32)$$

$$E_1 := \frac{8}{\gamma_{\bar{w}}} (L/\mu + 3\eta L - 1), \quad E_2 := \frac{2}{\gamma_{\bar{w}}}, \quad (6.33)$$

$$E_3 := \frac{12 \left( L^2 + L \|B\|^2 \right)}{\mu(1 - \bar{\gamma}_w) (1 - \bar{\gamma}_{I_N-W}^2)}, \quad E_4 := \frac{4}{(1 - \bar{\gamma}_w) (1 - \bar{\gamma}_{I_N-W}^2)}, \quad (6.34)$$

where  $h \leq \frac{1 - \bar{\gamma}_w}{4\bar{\gamma}_{I_N-W}^2} \wedge \frac{1}{2}$  from Lemma 12.

We note that if  $h = 0$ , then  $U = \widetilde{W} - W = 0$ , and by (3.16) and (6.13), we have  $\tilde{v}^{(0)} = v^{(0)} = \mathbf{0}$ , and we observe from (B.60) in the proof,  $\|\tilde{v}^{(k+1)}\|^2 = \|\tilde{v}^{(k)}\|^2 = \dots = \|\tilde{v}^{(0)}\|^2 = 0$ ; hence, we have  $\|\tilde{\mathbf{v}}\|_2^{\delta, K} = 0$ . In the case  $h = 0$ , we can get from (6.29) that

$$\|\tilde{\mathbf{x}}\|_2^{\delta, K} \leq \eta \frac{w_1(\eta\sigma^2 + 2d)}{N\delta^{2K-2}} + \delta^2 E_1 \mathbb{E} \left[ \left\| \tilde{e}_x^{(0)} \right\|^2 \right] + \delta^2 E_2 \mathbb{E} \left[ \left\| \tilde{x}^{(0)} \right\|^2 \right], \quad (6.35)$$

and moreover, EXTRA SGLD algorithm reduces to DE-SGLD when  $h = 0$ , and our result implies:

$$\begin{aligned} \mathbb{E} \left[ \left\| x^{(K)} - \bar{x}^{(K)} \right\|^2 \right] &\leq \delta^{2K} \|\tilde{\mathbf{x}}\|_2^{\delta, K} \\ &\leq \eta^2 \delta^2 \frac{w_1 \sigma^2}{N} + \eta \delta^2 \frac{2dw_1}{N} + \delta^{2K+2} E_1 \mathbb{E} \left[ \left\| \tilde{e}_x^{(0)} \right\|^2 \right] + \delta^{2K+2} E_2 \mathbb{E} \left[ \left\| \tilde{x}^{(0)} \right\|^2 \right]. \end{aligned} \quad (6.36)$$

This bound is of the same order as the one shown by Lemma 6 from [GGHZ21]. Suppose  $h \neq 0$ , we present upper bounds on  $\|\tilde{\mathbf{x}}\|_2^{\delta, K}$  and  $\|\tilde{\mathbf{v}}\|_2^{\delta, K}$  for EXTRA SGLD algorithm as follows.

**Theorem 13.** *Assume that the stepsize  $\eta$  satisfies*

$$0 < \eta < \frac{1}{h\gamma_1\gamma_2} \wedge \frac{\gamma_{\bar{w}}}{6(L + \mu) \vee 2A} \wedge 1, \quad (6.37)$$

where

$$0 < h \leq \frac{1 - \bar{\gamma}_w}{4\bar{\gamma}_{I_N-W}^2} \wedge \frac{1}{2} \wedge \frac{1}{\gamma_1\gamma_2}, \quad (6.38)$$

so that the condition in (6.26) is satisfied. Moreover, the constant  $A$  is defined in (6.23) and  $\gamma_{\bar{W}}$  defined by  $|\lambda_2^{\bar{W}}|^2$  in (6.20). Under the conditions in Lemma 12, for any constant  $\delta$  in three regimes depending on  $|\lambda_2^{\bar{W}}|^2$  from Lemma 10, and furthermore, it satisfies:

$$\delta^2 \in \left[ \left( 1 - \frac{\eta\mu}{2} \left( 1 - \frac{\eta L}{2} \right) \right) \vee \left( 1 - h \frac{1 - \bar{\gamma}_w}{4} (1 - \bar{\gamma}_{I_N-W}) \right), 1 \right], \quad (6.39)$$

then it holds that:

$$\begin{aligned} \|\tilde{\mathbf{x}}\|_2^{\delta,K} &\leq \frac{\eta^2}{\delta^{2K-2}} \cdot \frac{(w_2\gamma_1(h/\eta) + w_1)\sigma^2/N}{1 - h\gamma_1\gamma_2} + \frac{\eta}{\delta^{2K-2}} \cdot \left[ \frac{2d(w_2\gamma_1(h/\eta) + w_1)/N}{1 - h\gamma_1\gamma_2} \right. \\ &\quad \left. + \frac{\gamma_1}{1 - h\gamma_1\gamma_2} \delta^{2K} \left( (h/\eta)(E_3/\eta) \mathbb{E} \left[ \left\| \bar{e}_x^{(0)} \right\|^2 \right] + (E_4/h) \mathbb{E} \left[ \left\| \tilde{v}^{(0)} \right\|^2 \right] \right) \right] + \delta^2 D_0, \end{aligned} \quad (6.40)$$

and

$$\begin{aligned} \|\tilde{\mathbf{v}}\|_2^{\delta,K} &\leq \frac{h\eta}{\delta^{2K-2}} \cdot \frac{\gamma_2(w_2\gamma_1(h/\eta) + w_1)\sigma^2/N}{1 - h\gamma_1\gamma_2} + \frac{h}{\delta^{2K-2}} \cdot \left[ \frac{2\gamma_2 d(w_2\gamma_1(h/\eta) + w_1)/N}{1 - h\gamma_1\gamma_2} + \frac{w_2\sigma^2}{N} \right. \\ &\quad \left. + \frac{\gamma_1\gamma_2}{1 - h\gamma_1\gamma_2} \delta^{2K} \left( (h/\eta)(E_3/\eta) \mathbb{E} \left[ \left\| \bar{e}_x^{(0)} \right\|^2 \right] + (E_4/h) \mathbb{E} \left[ \left\| \tilde{v}^{(0)} \right\|^2 \right] \right) \right] + (h/\eta)\delta^2\gamma_2 D_0 \\ &\quad + (h/\eta) \frac{2dw_2}{N\delta^{2K-2}} + \delta^2(h/\eta)(E_3/\eta) \mathbb{E} \left[ \left\| \bar{e}_x^{(0)} \right\|^2 \right] + \delta^2(E_4/h) \mathbb{E} \left[ \left\| \tilde{v}^{(0)} \right\|^2 \right], \end{aligned} \quad (6.41)$$

where

$$D_0 := \frac{1}{1 - h\gamma_1\gamma_2} \left( E_1 \mathbb{E} \left[ \left\| \tilde{x}^{(0)} \right\|^2 \right] + E_2 \mathbb{E} \left[ \left\| \bar{e}_x^{(0)} \right\|^2 \right] \right), \quad (6.42)$$

and  $\gamma_1, \gamma_2, w_1, w_2, E_1, E_2, E_3, E_4$  are defined in (6.30), (6.31), (6.32), (6.33) and (6.34).

Next, we provide the uniform bounds on  $\mathbb{E} \left[ \left\| \tilde{v}^{(k)} \right\|^2 \right]$  and  $\mathbb{E} \left[ \left\| \nabla F(x^{(k)}) \right\|^2 \right]$ . We first use the upper bounds on  $\|\tilde{\mathbf{x}}\|_2^{\delta,K}$  and  $\|\tilde{\mathbf{v}}\|_2^{\delta,K}$  from Theorem 13 to get the next lemma.

**Lemma 14.** *Under the assumptions for Theorem 13, the following bounds hold for  $\mathbb{E} \left[ \left\| \tilde{v}^{(k)} \right\|^2 \right]$  and  $\mathbb{E} \left[ \left\| \nabla F(x^{(k)}) \right\|^2 \right]$  uniformly for every  $k = 1, 2, 3, \dots$ :*

$$\begin{aligned} \mathbb{E} \left[ \left\| \tilde{v}^{(k)} \right\|^2 \right] &\leq h\delta^2 \cdot \frac{C_1\gamma_2}{2L^2} + \delta^{2k+2}h \cdot \frac{C_0\gamma_1\gamma_2}{2L^2} \\ &\quad + \delta^{2k+2} \cdot ((h/\eta)\gamma_2 D_0 + C_0) + (h/\eta)\delta^2 \cdot \frac{w_2}{N} (\eta\sigma^2 + 2d), \end{aligned} \quad (6.43)$$

$$\begin{aligned} \mathbb{E} \left[ \left\| \nabla F(x^{(k)}) \right\|^2 \right] &\leq \eta\delta^2(C_1 + C_3) + \delta^2\eta^2 \left( \frac{C_1C_2}{2L^2} \right) + \delta^{2K_0}\eta(\gamma_1C_0 + D_0C_2) \\ &\quad + \delta^{2K_0}\eta^2 \left( \frac{\gamma_1C_0C_2}{2L^2} \right) + \delta^{2K_0}(D_0 + C_4) + 2\|\nabla F(\mathbf{x}_*)\|^2, \end{aligned} \quad (6.44)$$

where the constants are given by:

$$C_0 := \frac{2L^2}{1-h\gamma_1\gamma_2} \left( (h/\eta)(E_3/\eta)\mathbb{E} \left[ \left\| \bar{e}_x^{(0)} \right\|^2 \right] + (E_4/h)\mathbb{E} \left[ \left\| \tilde{v}^{(0)} \right\|^2 \right] \right), \quad (6.45)$$

$$C_1 := \frac{2L^2(\eta\sigma^2 + 2d)}{N} \cdot \frac{w_2\gamma_1(h/\eta) + w_1}{1-h\gamma_1\gamma_2}, \quad (6.46)$$

$$C_2 := \frac{2L^4}{N^2 \left( \delta^2 + \eta\mu \left( 1 - \frac{\eta L}{2} \right) - 1 \right)} \left( \eta + \frac{1+\eta L}{\mu \left( 1 - \frac{\eta L}{2} \right)} \right), \quad (6.47)$$

$$C_3 := \frac{2L^2}{N} \cdot \frac{\eta\sigma^2 + 2d}{\delta^2 + \eta\mu \left( 1 - \frac{\eta L}{2} \right) - 1}, \quad (6.48)$$

$$C_4 := \frac{2L^2}{\delta^2 + \eta\mu \left( 1 - \frac{\eta L}{2} \right) - 1} \mathbb{E} \left[ \left\| \bar{e}_x^{(0)} \right\|^2 \right], \quad (6.49)$$

and  $D_0$  is defined in (6.42),  $E_3, E_4$  are defined in (6.34) and  $w_1, w_2$  are defined in (6.31)-(6.32).

As an immediate corollary of Lemma 14, we obtain the following upper bounds for  $\mathbb{E} \left[ \left\| \tilde{v}^{(k)} \right\|^2 \right]$  and  $\mathbb{E} \left[ \left\| \nabla F(x^{(k)}) \right\|^2 \right]$  that are uniform in  $k$  when  $k$  is larger than a specific lower bound.

**Corollary 15.** Under the assumptions for Theorem 13, for any  $k \geq K_0$ , we have

$$\mathbb{E} \left[ \left\| \tilde{v}^{(k)} \right\|^2 \right] \leq R_h, \quad \mathbb{E} \left[ \left\| \nabla F(x^{(k)}) \right\|^2 \right] \leq R'_h, \quad (6.50)$$

$$R_h := h\delta^2 \left( \frac{C_1\gamma_2}{2L^2} + \frac{C_0\gamma_1\gamma_2}{2L^2} \right) + (h/\eta)\delta^2 \left( \gamma_2 D_0 + \frac{w_2}{N} (\eta\sigma^2 + 2d) \right) + \|\nabla F(\mathbf{x}_*)\|^2, \quad (6.51)$$

$$R'_h := \eta\delta^2 (C_1 + C_3 + \gamma_1 C_0 + D_0 C_2) + \delta^2\eta^2 \left( \frac{C_1 C_2}{2L^2} + \frac{\gamma_1 C_0 C_2}{2L^2} \right) + 3 \|\nabla F(\mathbf{x}_*)\|^2, \quad (6.52)$$

and

$$K_0 := \frac{\delta^2}{1-\delta^2} \left[ \left( 1 - \frac{\|\nabla F(\mathbf{x}_*)\|^2}{D_0 + C_4} \right) \vee \left( 1 - \frac{\|\nabla F(\mathbf{x}_*)\|^2}{C_0} \right) \right] \vee 0. \quad (6.53)$$

Now we are ready to present our main technical result on the error between the iterate  $x^{(k)}$  and their average (taken over  $N$  agents)  $\bar{\mathbf{x}}^{(k)}$  after  $k$  iterations:

$$\bar{x}^{(k)} = \frac{1}{N} \sum_{i=0}^N x_i^{(k)} \in \mathbb{R}^d, \quad \bar{\mathbf{x}}^{(k)} = \left[ \left( \bar{x}^{(k)} \right)^T, \left( \bar{x}^{(k)} \right)^T, \dots, \left( \bar{x}^{(k)} \right)^T \right]^T \in \mathbb{R}^{Nd}. \quad (6.54)$$

We have the next corollary.

**Corollary 16.** With the assumptions for Theorem 13, it holds for  $k \geq K_0$ , with  $K_0$  given in (6.53),

$$\sum_{i=1}^N \mathbb{E} \left[ \left\| x_i^{(k)} - \bar{\mathbf{x}}^{(k)} \right\|^2 \right] \leq 4(\bar{\gamma}_{\tilde{W}})^{2k} \mathbb{E} \left[ \left\| x^{(0)} \right\|^2 \right] + 8\eta^2 \cdot \frac{R_h + R'_h}{(1-\bar{\gamma}_{\tilde{W}})^2} + \frac{4\eta^2\sigma^2 N}{1-\bar{\gamma}_{\tilde{W}}^2} + \frac{8\eta d N}{1-\bar{\gamma}_{\tilde{W}}^2}, \quad (6.55)$$

where  $R_h, R'_h$  are defined in (6.51)-(6.52) and  $\bar{\gamma}_{\tilde{W}} := \max \{ |\lambda_2^{\tilde{W}}|, |\lambda_N^{\tilde{W}}| \} \in [0, 1)$ .

In order to sample from the Gibbs distribution, we recall the decomposition (6.1):

$$\frac{1}{N} \sum_{i=1}^N \mathcal{W}_2 \left( \mathcal{L} \left( x_i^{(k)} \right), \pi \right) \leq \frac{1}{N} \sum_{i=1}^N \mathcal{W}_2 \left( \mathcal{L} \left( x_i^{(k)} \right), \mathcal{L} \left( \bar{x}^{(k)} \right) \right) + \mathcal{W}_2 \left( \mathcal{L} \left( \bar{x}^{(k)} \right), \pi \right). \quad (6.56)$$

The first term in (6.56) can be bounded by Corollary 16 as follows:

$$\begin{aligned} \frac{1}{N} \sum_{i=1}^N \mathcal{W}_2 \left( \mathcal{L} \left( x_i^{(K)} \right), \mathcal{L} \left( \bar{x}^{(K)} \right) \right) &\leq \left( \frac{1}{N} \sum_{i=1}^N \mathbb{E} \left[ \left\| x_i^{(K)} - \bar{x}^{(K)} \right\|^2 \right] \right)^{1/2} \\ &\leq \eta \cdot \frac{D_1}{\sqrt{N}} + \sqrt{\eta} \cdot D_2 + \frac{2(\bar{\gamma}_{\tilde{w}})^K}{\sqrt{N}} \sqrt{\mathbb{E} \left[ \|x^{(0)}\|^2 \right]}, \end{aligned} \quad (6.57)$$

where

$$D_1 := 2 \frac{\sqrt{2(R_h + R'_h)}}{1 - \bar{\gamma}_{\tilde{w}}} + \frac{2\sigma}{\sqrt{1 - \bar{\gamma}_{\tilde{w}}^2}}, \quad D_2 := 2 \sqrt{\frac{2d}{1 - \bar{\gamma}_{\tilde{w}}^2}}. \quad (6.58)$$

### 6.1.2 $L^2$ distance between $\bar{x}^{(k)}$ and $x_k$

Next, we upper bound the second term in (6.56), which, according to (6.2), can be bounded as

$$\mathcal{W}_2 \left( \mathcal{L} \left( \bar{x}^{(k)} \right), \pi \right) \leq \mathcal{W}_2 \left( \mathcal{L} \left( \bar{x}^{(k)} \right), \mathcal{L}(x_k) \right) + \mathcal{W}_2 \left( \mathcal{L}(x_k), \pi \right), \quad (6.59)$$

where  $x_k$  given in (6.3) is the Euler-Maruyama discretization of the overdamped Langevin SDE (6.4). First, we bound the first term in (6.59) by providing an upper bound on the  $L^2$  distance between the average iterate  $\bar{x}^{(k)}$  and  $x_k$ .

Since  $W$  is doubly stochastic, we can compute  $\mathcal{W}\bar{x}^{(k)} = \bar{x}^{(k)}$ , that is  $\bar{x}^{(k)}$  is consensual, similarly, we also have  $\frac{1}{N}(1_N 1_N^T)^T \bar{x}^{(k)} = \bar{x}^{(k)}$  for  $k = 1, 2, \dots$ . The following mean iterates can be found by taking average of (3.11).

$$\bar{x}^{(k+1)} = \bar{x}^{(k)} - \eta \frac{1}{N} \sum_{i=1}^N \nabla f_i \left( x_i^{(k)} \right) - \eta \bar{\xi}^{(k)} + \sqrt{2\eta w^{(k+1)}}, \quad (6.60)$$

we can find the mean iterates have the same format as the one of decentralized stochastic gradient Langevin dynamics in [GHZ21]. Then, we can get

$$\bar{x}^{(k+1)} = \bar{x}^{(k)} - \frac{\eta}{N} \nabla f \left( \bar{x}^{(k)} \right) + \eta \hat{\mathcal{E}}_{k+1} - \eta \bar{\xi}^{(k)} + \sqrt{2\eta w^{(k+1)}}, \quad (6.61)$$

where the error term is

$$\hat{\mathcal{E}}_{k+1} := \frac{1}{N} \sum_{i=1}^N \left( \nabla f_i \left( \bar{x}^{(k)} \right) - \nabla f_i \left( x_i^{(k)} \right) \right). \quad (6.62)$$

On the other hand, we recall from (6.3) that  $x_k$  is the Euler-Maruyama discretization of overdamped Langevin diffusion (6.4) with the iterates:

$$x_{k+1} = x_k - \frac{\eta}{N} \nabla f(x_k) + \sqrt{2\eta w^{(k+1)}}. \quad (6.63)$$

Hence, we get

$$\bar{x}^{(k+1)} - x_{k+1} = \bar{x}^{(k)} - x_k - \frac{\eta}{N} \left( \nabla f \left( \bar{x}^{(k)} \right) - \nabla f \left( x_k \right) \right) + \eta \hat{\mathcal{E}}_{k+1} - \eta \bar{\xi}^{(k)}. \quad (6.64)$$

Now, we are ready to state the next corollary to bound  $L^2$  distance between the mean  $\bar{x}^{(k)}$  in (6.61) and discretized overdamped Langevin iterate  $x_k$  in (6.63).

**Corollary 17.** *Suppose  $\eta < \frac{2}{L} \wedge 1$ , under assumptions in Corollary 16, then there holds*

$$\begin{aligned} \mathbb{E} \left[ \left\| \bar{x}^{(k)} - x_k \right\|^2 \right] &\leq \frac{\eta \left( \eta + \frac{(1+\eta L)^2}{\mu(1-\frac{\eta L}{2})} \right) \left( \eta^2 \cdot \frac{4L^2}{N} \left( \frac{2(R_h+R'_h)}{(1-\bar{\gamma}_{\bar{W}})^2} + \frac{\sigma^2 N}{1-\bar{\gamma}_{\bar{W}}^2} \right) + \eta \cdot \frac{8L^2 d}{1-\bar{\gamma}_{\bar{W}}^2} \right) + \eta^2 \frac{\sigma^2}{N}}{\eta \mu \left( 1 - \frac{\eta L}{2} \right)} \\ &\quad + \frac{\bar{\gamma}_{\bar{W}}^{2k} - \left( 1 - \eta \mu \left( 1 - \frac{\eta L}{2} \right) \right)^k}{(\bar{\gamma}_{\bar{W}})^2 - 1 + \eta \mu \left( 1 - \frac{\eta L}{2} \right)} \frac{4L^2 (\bar{\gamma}_{\bar{W}})^2}{N} \mathbb{E} \left\| x^{(0)} \right\|^2, \end{aligned} \quad (6.65)$$

where the constants  $R_h, R'_h$  and  $\bar{\gamma}_{\bar{W}}$  are defined in Lemma 14 and Corollary 16.

### 6.1.3 $\mathcal{W}_2$ distance between the law of $x_k$ and the Gibbs distribution $\pi$

Next, we provide the 2-Wasserstein distance between the law of  $x_k$  in (6.63), which is the Euler-Maruyama discretization of (6.4) and the Gibbs distribution  $\pi$ . We note that the function  $\frac{1}{N}f$  is  $\mu$ -strongly convex and  $L$ -smooth. We simply quote an existing result, that is, Theorem 4 from [DK19] restated in the next lemma.

**Lemma 18** (Theorem 4 in [DK19]). *For any  $\eta \in \left( 0, \frac{2N}{L+\mu} \right]$ , we have*

$$\mathcal{W}_2(\mathcal{L}(x_K), \pi) \leq (1 - \mu\eta)^K \mathcal{W}_2(\mathcal{L}(x_0), \pi) + \frac{1.65L}{\mu} \sqrt{dN^{-1}} \sqrt{\eta}. \quad (6.66)$$

Now, we are finally ready to complete the proof of Theorem 4.

### 6.1.4 Completing the proof of Theorem 4

Under our assumptions, the conditions in Theorem 13 are satisfied, so that one can apply Corollary 16, Corollary 17 and Lemma 18. Note that Corollary 17 and Lemma 18 give the bound on the second term in decomposition (6.56) such that it follows that:

$$\begin{aligned} \mathcal{W}_2 \left( \mathcal{L} \left( \bar{x}^{(k)} \right), \pi \right) &\leq \left( \mathbb{E} \left[ \left\| \bar{x}^{(k)} - x_k \right\|^2 \right] \right)^{1/2} + \mathcal{W}_2(\mathcal{L}(x_k), \pi) \\ &\leq \left( \frac{\bar{\gamma}_{\bar{W}}^{2k} - \left( 1 - \eta \mu \left( 1 - \frac{\eta L}{2} \right) \right)^k}{\bar{\gamma}_{\bar{W}}^2 - 1 + \eta \mu \left( 1 - \frac{\eta L}{2} \right)} \right)^{1/2} \frac{2L\bar{\gamma}_{\bar{W}}}{\sqrt{N}} \left( \mathbb{E} \left\| x^{(0)} \right\|^2 \right)^{1/2} + (1 - \mu\eta)^K \mathcal{W}_2(\mathcal{L}(x_0), \pi) + \sqrt{\eta} \mathcal{E}_1, \end{aligned} \quad (6.67)$$

where we have

$$\begin{aligned} \mathcal{E}_1 := & \left( \frac{\eta}{\mu \left( 1 - \frac{\eta L}{2} \right)} + \frac{(1 + \eta L)^2}{\mu^2 \left( 1 - \frac{\eta L}{2} \right)^2} \right)^{1/2} \cdot \left( \frac{4L^2(R_h + R'_h)\eta}{N(1 - \bar{\gamma}_{\tilde{w}})^2} + \frac{4L^2\sigma^2\eta}{1 - \bar{\gamma}_{\tilde{w}}^2} + \frac{8L^2d}{1 - \bar{\gamma}_{\tilde{w}}^2} \right)^{1/2} \\ & + \frac{\sigma}{\sqrt{\mu \left( 1 - \frac{\eta L}{2} \right) N}} + \frac{1.65L}{\mu} \sqrt{dN^{-1}}, \end{aligned} \quad (6.68)$$

and this proves (4.3). Finally, by (6.56), (6.57) (which follows from Corollary 16) and (6.67), we can derive that

$$\begin{aligned} & \frac{1}{N} \sum_{i=1}^N \mathcal{W}_2 \left( \mathcal{L} \left( x_i^{(K)} \right), \pi \right) \\ & \leq \eta \cdot \frac{D_1}{\sqrt{N}} + \sqrt{\eta} \cdot (D_2 + \mathcal{E}_1) + \left( \frac{\bar{\gamma}_{\tilde{w}}^{2K} - \left( 1 - \eta \mu \left( 1 - \frac{\eta L}{2} \right) \right)^K}{\bar{\gamma}_{\tilde{w}}^2 - 1 + \eta \mu \left( 1 - \frac{\eta L}{2} \right)} \right)^{1/2} \frac{2L\bar{\gamma}_{\tilde{w}}}{\sqrt{N}} \left( \mathbb{E} \|x^{(0)}\|^2 \right)^{1/2} \\ & \quad + (1 - \mu\eta)^K \mathcal{W}_2(\mathcal{L}(x_0), \pi) + \frac{2(\bar{\gamma}_{\tilde{w}})^K}{\sqrt{N}} \sqrt{\mathbb{E} [\|x^{(0)}\|^2]}, \end{aligned} \quad (6.69)$$

and this completes the proof.

## 6.2 Proof of Proposition 5

First, we consider generalized EXTRA SGLD. We recall that one can take  $1 > \delta^2 = 1 - \frac{\eta\mu}{2} \left( 1 - \frac{\eta L}{2} \right) \geq 1 - \mu\eta \geq \bar{\gamma}_{\tilde{w}}$  where  $\delta^2$  satisfies the constraint in (6.39). Therefore, for any sufficiently large  $K$ , the term  $\left( 1 - \eta \mu \left( 1 - \frac{\eta L}{2} \right) \right)^K$  dominates both terms  $(1 - \mu\eta)^K$  and  $(\bar{\gamma}_{\tilde{w}})^{2K}$ . Hence, we can get the order of the last three terms in (6.69) in Theorem 4 is  $\mathcal{O}((1 - \eta \mu (1 - \eta L/2))^{K/2})$ .

Since  $1 - x \leq e^{-x}$  for any  $0 \leq x \leq 1$ , we conclude from Theorem 4, for any  $K \geq K_0$ , with  $K_0$  given in (6.53),

$$\frac{1}{N} \sum_{i=1}^N \mathcal{W}_2^{\text{extra-sgld}} \left( \mathcal{L} \left( x_i^{(K)} \right), \pi \right) \leq \mathcal{O} \left( \sqrt{\eta} \left( \sqrt{d} + \mathcal{E}_1 \right) \right) + \mathcal{O} \left( e^{-\frac{\eta\mu}{2} (1 - \frac{\eta L}{2}) K} \right), \quad (6.70)$$

where  $\mathcal{E}_1$  is defined in (4.4). Hence, for any  $\varepsilon \rightarrow 0$ , we have

$$\frac{1}{N} \sum_{i=1}^N \mathcal{W}_2^{\text{extra-sgld}} \left( \mathcal{L} \left( x_i^{(K)} \right), \pi \right) \leq \mathcal{O}(\varepsilon), \quad (6.71)$$

provided that

$$\eta \leq \mathcal{O} \left( \frac{\varepsilon^2}{(\sqrt{d} + \mathcal{E}_1)^2} \right), \quad (6.72)$$

and

$$K \geq K^{\text{extra-sgld}} := K_0 \vee K_1 := K_0 \vee \mathcal{O} \left( \frac{\log(1/\varepsilon)}{\eta\mu} \right) = K_0 \vee \mathcal{O} \left( \frac{\log(1/\varepsilon)(\sqrt{d} + \mathcal{E}_1)^2}{\varepsilon^2\mu} \right), \quad (6.73)$$

where  $K_0$  is given in Table 1. We recall the definition of  $\mathcal{E}_1$  from (4.4):

$$\begin{aligned} \mathcal{E}_1 = & \left( \frac{\eta}{\mu \left( 1 - \frac{\eta L}{2} \right)} + \frac{(1+\eta L)^2}{\mu^2 \left( 1 - \frac{\eta L}{2} \right)^2} \right)^{1/2} \cdot \left( \frac{4L^2(R_h + R'_h)\eta}{N(1-\bar{\gamma}_{\bar{w}})^2} + \frac{4L^2\sigma^2\eta}{1-\bar{\gamma}_{\bar{w}}^2} + \frac{8L^2d}{1-\bar{\gamma}_{\bar{w}}^2} \right)^{1/2} \\ & + \frac{\sigma}{\sqrt{\mu \left( 1 - \frac{\eta L}{2} \right) N}} + \frac{1.65L}{\mu} \sqrt{dN^{-1}}, \end{aligned}$$

where the constants are given in Table 1.

$$\begin{aligned} \text{Under our assumption } \eta L = \mathcal{O}(1) \text{ such that } \mathcal{O} \left( \left( \frac{\eta}{\mu(1-\frac{\eta L}{2})} + \frac{(1+\eta L)^2}{\mu^2(1-\frac{\eta L}{2})^2} \right)^{1/2} \right) = \mathcal{O}(1/\mu). \text{ Thus} \\ \mathcal{E}_1 = \mathcal{O} \left( \frac{1}{\mu} \left( L\sqrt{\eta} \sqrt{R_h + R'_h} + L\sqrt{d} \right) \right). \end{aligned} \quad (6.74)$$

Therefore,  $\frac{1}{N} \sum_{i=1}^N \mathcal{W}_2^{\text{extra-sgld}} \left( \mathcal{L} \left( x_i^{(K)} \right), \pi \right) \leq \mathcal{O}(\varepsilon)$  provided that

$$K \geq K^{\text{extra-sgld}} = K_0 \vee K_1 = K_0 \vee \tilde{\mathcal{O}} \left( \frac{L^2(\eta(R_h + R'_h) + d)}{\varepsilon^2\mu^3} \right), \quad (6.75)$$

where  $\tilde{\mathcal{O}}$  hides the logarithmic dependence on  $\varepsilon$ .

We will show that under our assumptions  $K_0 = \mathcal{O} \left( \frac{1}{\eta\mu} \vee \frac{1}{h} \right)$  and under the assumption  $h \leq \frac{1}{(L/\mu)^4(L+\|B\|^2)}$ , we will have  $\eta(R_h + R'_h) \leq \mathcal{O}(d)$  and under the assumption  $h \geq \Omega(\eta\mu)$ , we have  $K_0 = \mathcal{O} \left( \frac{1}{\eta\mu} \right)$  such that  $K \geq K^{\text{extra-sgld}} = K_0 \vee K_1 = \tilde{\mathcal{O}} \left( \frac{1}{\eta\mu} \vee \frac{L^2d}{\varepsilon^2\mu^3} \right) = \tilde{\mathcal{O}} \left( \frac{L^2d}{\varepsilon^2\mu^3} \right)$ .

As a first step of the proof, we spell out the dependence of the constants in Table 1 on  $L, \mu, d, h, \eta, \|B\|^2$ , and we summarize the results in Table 2. Next, we compute from (6.51)-(6.52):

$$\eta R_h = h\eta\delta^2 \left( \frac{C_1\gamma_2}{2L^2} + \frac{C_0\gamma_1\gamma_2}{2L^2} \right) + h\delta^2 \left( \gamma_2 D_0 + \frac{w_2}{N} (\eta\sigma^2 + 2d) \right) + \eta \|\nabla F(\mathbf{x}_*)\|^2, \quad (6.76)$$

$$\eta R'_h = \eta^2\delta^2 (C_1 + C_3 + \gamma_1 C_0 + D_0 C_2) + \delta^2\eta^3 \left( \frac{C_1 C_2}{2L^2} + \frac{\gamma_1 C_0 C_2}{2L^2} \right) + 3\eta \|\nabla F(\mathbf{x}_*)\|^2. \quad (6.77)$$

Under the assumption  $h \geq \Omega(\eta\mu)$ , in particular, for  $h/(\eta\mu) \geq \frac{1}{L+\|B\|^2}$ , we get

$$C_1 = \mathcal{O} \left( d(L + \|B\|^2)(L^3/\mu^2)(h/\eta) + (L^2d)(L/\mu) \right) \leq \mathcal{O} \left( d(L + \|B\|^2)L(L/\mu)^2(h/\eta) \right). \quad (6.78)$$

Moreover, from Table 2, we have:

$$\frac{\gamma_2}{2L^2} (C_1 + C_0\gamma_1) = \mathcal{O}(dL(L + \|B\|^2)^2(L/\mu)^4(h/\eta)), \quad (6.79)$$

$$\gamma_2 D_0 + \frac{w_2}{N} (\eta\sigma^2 + 2d) = \mathcal{O}((L/\mu)L^2(L + \|B\|^2)((L/\mu)^2 \vee d)), \quad (6.80)$$

$$\begin{aligned} C_1 + C_3 + \gamma_1 C_0 + D_0 C_2 &= \mathcal{O}((h/\eta)dL(L/\mu)^2(L + \|B\|^2)) + \mathcal{O}(Ld(L/\mu)(1/\eta)) \\ &\quad + \mathcal{O}((L/\mu)^2L(h/\eta)(L + \|B\|^2)) + \mathcal{O}(L^2(L/\mu)^3(1/\eta)) \\ &= \mathcal{O}((L/\mu)(h/\eta)(L + \|B\|^2)Ld \vee L^2d(L/\mu)^3(1/\eta)) \end{aligned} \quad (6.81)$$

$$\frac{C_2}{2L^2} (C_1 + C_0\gamma_1) = \mathcal{O}(L(L/\mu)^4(L + \|B\|^2)(h/\eta)(d/\eta)). \quad (6.82)$$

Now we can compute (6.76) and (6.77) as follows. We first use (6.79) and (6.80) to get

$$\begin{aligned} \eta R_h &= \mathcal{O}(h\eta \cdot dL(L + \|B\|^2)^2(L/\mu)^4(h/\eta) + h \cdot (L/\mu)L^2(L + \|B\|^2)((L/\mu)^2 \vee d)) \\ &= \mathcal{O}(h^2 \cdot dL(L + \|B\|^2)^2(L/\mu)^4 + h \cdot (L/\mu)L^2(L + \|B\|^2)((L/\mu)^2 \vee d)) \\ &\leq \mathcal{O}\left(hdL(L + \|B\|^2)\left(1 + (L/\mu)\left(\frac{(L/\mu)^2}{d} \vee 1\right)\right)\right) \end{aligned} \quad (6.83)$$

$$\begin{aligned} &\leq \mathcal{O}\left(hdL(L/\mu)(L + \|B\|^2)\left(1 + \left(\frac{(L/\mu)^2 + d}{d}\right)\right)\right) \\ &\leq \mathcal{O}\left(hdL(L/\mu)(L + \|B\|^2) \cdot \frac{(L/\mu)^2 d}{d}\right) = \mathcal{O}(hdL(L/\mu)^3(L + \|B\|^2)), \end{aligned} \quad (6.84)$$

where we used the assumption

$$h \leq \frac{1}{(L/\mu)^4(L + \|B\|^2)}, \quad (6.85)$$

to get  $h^2dL(L + \|B\|^2)^2(L/\mu)^4 \leq hdL(L + \|B\|^2)$  in (6.83). Next, we use (6.81) and (6.82) to compute that

$$\begin{aligned} \eta R'_h &= \mathcal{O}(\eta^2 \cdot ((L/\mu)(h/\eta)(L + \|B\|^2)Ld \vee dL^2(L/\mu)^3(1/\eta)) + \eta^3 \cdot L(L/\mu)^4(L + \|B\|^2)(h/\eta)(d/\eta)) \\ &= \mathcal{O}(\eta \cdot ((L/\mu)h(L + \|B\|^2)Ld \vee dL^2(L/\mu)^3) + (L/\mu)^4(L + \|B\|^2)hd) \end{aligned} \quad (6.86)$$

$$\leq \mathcal{O}(\eta dL^2(L/\mu)^3 + (L/\mu)^4(L + \|B\|^2)hd) \quad (6.87)$$

$$\leq \mathcal{O}(h(d/\mu)(L/\mu)^3 + (L/\mu)^4(L + \|B\|^2)hd) \quad (6.88)$$

$$= \mathcal{O}(hd(L/\mu) \cdot (L/\mu)^3(L + \|B\|^2)), \quad (6.89)$$

where we used  $\eta L \leq 1$  to get (6.86), and moreover, we used the assumption (6.85) to get (6.87) and then used the assumption  $h \geq \Omega(\eta\mu)$  again to get (6.88). As a consequence, we can compute that

$$\eta(R_h + R'_h) = \mathcal{O}(hd(L/\mu)^4(L + \|B\|^2)) \leq \mathcal{O}(d), \quad (6.90)$$

where we use (6.85) again in the last inequality. Now we can compute the term in (6.75) such that

$$\tilde{\mathcal{O}}\left(\frac{L^2(\eta(R_h + R'_h) + d)}{\varepsilon^2\mu^3}\right) = \tilde{\mathcal{O}}\left(\frac{L^2d}{\varepsilon^2\mu^3}\right). \quad (6.91)$$

To compute the term  $K_0$  in (6.75), we use Table 2 to get

$$D_0 + C_4 = \mathcal{O}(L(L/\mu)(1/\eta)), \quad C_0 = \mathcal{O}(L^2(L/\mu)(h/\eta)(L + \|B\|^2)), \quad (6.92)$$

$$\delta^2 = \mathcal{O}(1), \quad \frac{1}{1 - \delta^2} = \mathcal{O}\left(\frac{1}{\eta\mu} \vee \frac{1}{h}\right). \quad (6.93)$$

Under the setting in (6.85), we have  $h \leq \frac{1}{L(L + \|B\|^2)}$ , then we get

$$\left(1 - \frac{\|\nabla F(\mathbf{x}_*)\|^2}{D_0 + C_4}\right) \vee \left(1 - \frac{\|\nabla F(\mathbf{x}_*)\|^2}{C_0}\right) = \mathcal{O}\left(1 - \frac{1}{L(L/\mu)(1/\eta)}\right) = \mathcal{O}(1). \quad (6.94)$$

Hence, we can compute from (6.53) such that

$$K_0 = \frac{\delta^2}{1 - \delta^2} \left[ \left(1 - \frac{\|\nabla F(\mathbf{x}_*)\|^2}{D_0 + C_4}\right) \vee \left(1 - \frac{\|\nabla F(\mathbf{x}_*)\|^2}{C_0}\right) \right] \vee 0 = \mathcal{O}\left(\frac{1}{\eta\mu} \vee \frac{1}{h}\right) = \mathcal{O}\left(\frac{1}{\eta\mu}\right), \quad (6.95)$$

where we used the assumption  $h \geq \Omega(\eta\mu)$ . We recall from (6.70) and (6.75) that for generalized EXTRA SGLD:

$$\frac{1}{N} \sum_{i=1}^N \mathcal{W}_2^{\text{extra-sgld}} \left( \mathcal{L} \left( x_i^{(K)} \right), \pi \right) \leq \mathcal{O} \left( \sqrt{\eta} \left( \sqrt{d} + \mathcal{E}_1 \right) \right) + \mathcal{O} \left( e^{-\frac{\eta\mu}{2} \left( 1 - \frac{\eta L}{2} \right) K} \right). \quad (6.96)$$

Hence, we conclude that for any  $\varepsilon \rightarrow 0$ ,

$$\frac{1}{N} \sum_{i=1}^N \mathcal{W}_2^{\text{extra-sgld}} \left( \mathcal{L} \left( x_i^{(K)} \right), \pi \right) \leq \mathcal{O}(\varepsilon), \quad (6.97)$$

provided that

$$\eta \leq \mathcal{O} \left( \frac{\varepsilon^2}{(\sqrt{d} + \mathcal{E}_1)^2} \right), \quad \text{and} \quad K \geq K^{\text{extra-sgld}} = K_0 \vee K_1 = \tilde{\mathcal{O}} \left( \frac{L^2 d}{\varepsilon^2 \mu^3} \right), \quad (6.98)$$

where we used (6.75), (6.91), (6.95) and  $K_1 = \mathcal{O} \left( \frac{\log(1/\varepsilon)}{\eta\mu} \right) = \tilde{\mathcal{O}} \left( \frac{\log(1/\varepsilon)}{\eta\mu} \right)$  from (6.73).

Next, we consider DE-SGLD. For DE-SGLD, it follows from Theorem 1 in [GGHZ21] that:

$$\frac{1}{N} \sum_{i=1}^N \mathcal{W}_2^{\text{de-sgld}} \left( \mathcal{L} \left( x_i^{(K)} \right), \pi \right) \leq \mathcal{O} \left( e^{-\frac{\eta\mu}{2} \left( 1 - \frac{\eta L}{2} \right) K} \right) + \mathcal{O} \left( \sqrt{\eta} \left( \sqrt{d} + \mathcal{E}'_1 \right) \right), \quad (6.99)$$

where

$$\begin{aligned} \mathcal{E}'_1 := & \frac{1.65L}{\mu} \sqrt{dN^{-1}} + \frac{\sigma}{\sqrt{\mu \left( 1 - \frac{\eta L}{2} \right) N}} \\ & + \left( \frac{\eta}{\mu \left( 1 - \frac{\eta L}{2} \right)} + \frac{(1 + \eta L)^2}{\mu^2 \left( 1 - \frac{\eta L}{2} \right)^2} \right)^{1/2} \cdot \left( \frac{4L^2 D^2 \eta}{N(1 - \bar{\gamma}_{\widetilde{W}})^2} + \frac{4L^2 \sigma^2 \eta}{1 - \bar{\gamma}_{\widetilde{W}}^2} + \frac{8L^2 d}{1 - \bar{\gamma}_{\widetilde{W}}^2} \right)^{1/2}, \end{aligned}$$

with  $\bar{\gamma}_{\tilde{w}} := \max \{ |\lambda_2^{\tilde{w}}|, |\lambda_N^{\tilde{w}}| \} \in [0, 1)$  and the constant (see Lemma 6 in [GGHZ21])

$$D^2 := 4L^2 \mathbb{E} \|x^{(0)} - x_*\|^2 + 8L^2 \frac{\hat{C}_1^2 \eta^2 N}{(1 - \bar{\gamma}_{\tilde{w}})^2} + \frac{2L^2 (\eta \sigma^2 N + 2dN)}{\mu (1 + \lambda_N^{\tilde{w}} - \eta L)} + 4 \|\nabla F(\mathbf{x}_*)\|^2, \quad (6.100)$$

where

$$\hat{C}_1 := \bar{C}_1 \cdot \left( 1 + \frac{2(L + \mu)}{\mu} \right), \quad \text{with} \quad \bar{C}_1 := \sqrt{2L \sum_{i=1}^N (f_i(0) - f_i^*)}, \quad f_i^* := \min_{x \in \mathbb{R}^d} f_i(x).$$

Hence, by  $\eta L = \mathcal{O}(1)$  and  $\hat{C}_1^2 = \mathcal{O}(L^3/\mu^2)$ , we can compute that

$$D^2 = \mathcal{O}(L^2 + L^3/\mu^2 + Ld(L/\mu)) = \mathcal{O}(L^3 d/\mu). \quad (6.101)$$

Therefore, we have

$$\begin{aligned} \mathcal{E}'_1 &= \mathcal{O}\left(\frac{1}{\mu} (L\sqrt{\eta}D + L\sqrt{d})\right) = \mathcal{O}\left(\frac{1}{\mu} (D\sqrt{L} + L\sqrt{d})\right) \\ &= \mathcal{O}\left(\sqrt{Ld}(L/\mu)\sqrt{L/\mu} + (L/\mu)\sqrt{d}\right) = \mathcal{O}\left(\sqrt{Ld}(L/\mu)\sqrt{L/\mu}\right), \end{aligned} \quad (6.102)$$

where we use  $\eta \leq 1/L$  to get the second equality in the first line. By following (6.70) and (6.75), we conclude that

$$\begin{aligned} \frac{1}{N} \sum_{i=1}^N \mathcal{W}_2^{\text{de-sgld}} \left( \mathcal{L} \left( x_i^{(K)} \right), \pi \right) &\leq \mathcal{O}\left(\sqrt{\eta} \left( \sqrt{d} + \mathcal{E}'_1 \right)\right) + \mathcal{O}\left(e^{-\frac{\eta\mu}{2}(1-\frac{\eta L}{2})K}\right) \\ &\leq \mathcal{O}\left(\sqrt{\eta} \left( \sqrt{d} + \sqrt{Ld}(L/\mu)\sqrt{L/\mu}\right)\right) + \mathcal{O}\left(e^{-\frac{\eta\mu}{2}(1-\frac{\eta L}{2})K}\right). \end{aligned} \quad (6.103)$$

Hence, we conclude that for any  $\varepsilon \rightarrow 0$ , we have

$$\frac{1}{N} \sum_{i=1}^N \mathcal{W}_2^{\text{de-sgld}} \left( \mathcal{L} \left( x_i^{(K)} \right), \pi \right) \leq \mathcal{O}(\varepsilon), \quad (6.104)$$

provided that

$$\eta \leq \mathcal{O}\left(\frac{\varepsilon^2}{(\sqrt{d} + \sqrt{Ld}(L/\mu)\sqrt{L/\mu})^2}\right) \quad \text{and} \quad K \geq K^{\text{de-sgld}} := \tilde{\mathcal{O}}\left(\frac{L^4 d}{\varepsilon^2 \mu^3}\right). \quad (6.105)$$

The proof is complete.

## 7 Numerical Experiments

In this section, we present some results from the numerical experiments based on our algorithms and investigate the relative performance of DE-SGLD and EXTRA SGLD. We mainly perform Bayesian linear regression and Bayesian logistic regression by distributing the sample data evenly

<i>Constants</i>	<i>Source</i>
$\gamma_1 = \frac{1}{\gamma_{\bar{w}}} \left( \frac{1}{L} + 2 + \frac{1}{L\mu} \right) = \mathcal{O}(1/\mu)$	
$\gamma_2 = \frac{12 \left( L^2 + L \ B\ ^2 \right)}{(1 - \bar{\gamma}_w) (1 - \bar{\gamma}_{I_N-W}^2)} \left( 1 + \frac{4L^2 \left( 1 + \frac{2+2L}{\mu} \right)}{N^2 \mu} \right) = \mathcal{O}(L^2(L/\mu)^2(L + \ B\ ^2))$	
$w_1 = 2 \left( \frac{N^2 + 1}{\gamma_{\bar{w}}} + \frac{4}{\gamma_{\bar{w}}} \cdot \left( \frac{L}{\mu} + 3\eta L - 1 \right) \right) = \mathcal{O}(L/\mu)$	
$w_2 = \frac{8 \left( 6 \left( L^2 + L \ B\ ^2 \right) + N^2 \mu \right)}{N \mu (1 - \bar{\gamma}_w) (1 - \bar{\gamma}_{I_N-W}^2)} = \mathcal{O}((L/\mu)(L + \ B\ ^2))$	
$E_1 = \frac{8}{\gamma_{\bar{w}}} (L/\mu + 3\eta L - 1) = \mathcal{O}(L/\mu), \quad E_2 = \frac{2}{\gamma_{\bar{w}}} = \mathcal{O}(1)$	
$E_3 = \frac{12 \left( L^2 + L \ B\ ^2 \right)}{\mu (1 - \bar{\gamma}_w) (1 - \bar{\gamma}_{I_N-W}^2)} = \mathcal{O}((L/\mu)(L + \ B\ ^2)), \quad E_4 = \frac{4}{(1 - \bar{\gamma}_w) (1 - \bar{\gamma}_{I_N-W}^2)} = \mathcal{O}(1)$	
$C_0 = \left( (h/\eta) E_3 \mathbb{E} \left[ \left\  \bar{e}_x^{(0)} \right\ ^2 \right] + E_4 \mathbb{E} \left[ \left\  \tilde{v}^{(0)} \right\ ^2 \right] \right) \cdot \frac{2L^2}{1 - \eta \gamma_1 \gamma_2} = \mathcal{O}(L^2(L/\mu)(h/\eta)(L + \ B\ ^2))$	
$C_1 = \frac{2L^2 (\eta \sigma^2 + 2d)}{N} \cdot \frac{w_2 \gamma_1 (h/\eta) + w_1}{1 - h \gamma_1 \gamma_2} = \mathcal{O}(Ld(L + \ B\ ^2)(L/\mu)^2(h/\eta) + (L^2 d)(L/\mu))$	
$C_2 = \frac{2L^4 \left( \eta + \frac{1+\eta L}{\mu(1-\frac{\eta L}{2})} \right)}{N^2 \left( \delta^2 + \eta \mu \left( 1 - \frac{\eta L}{2} \right) - 1 \right)} = \mathcal{O}(L^2(L/\mu)^2(1/\eta))$	
$C_3 = \frac{2L^2}{N} \cdot \frac{\eta \sigma^2 + 2d}{\delta^2 + \eta \mu \left( 1 - \frac{\eta L}{2} \right) - 1} = \mathcal{O}(Ld(L/\mu)(1/\eta))$	
$C_4 = \frac{2L^2}{\delta^2 + \eta \mu \left( 1 - \frac{\eta L}{2} \right) - 1} \mathbb{E} \left[ \left\  \bar{e}_x^{(0)} \right\ ^2 \right] = \mathcal{O}(L(L/\mu)(1/\eta))$	
$D_0 = \frac{1}{1 - h \gamma_1 \gamma_2} \left( E_1 \mathbb{E} \left[ \left\  \tilde{x}^{(0)} \right\ ^2 \right] + E_2 \mathbb{E} \left[ \left\  \bar{e}_x^{(0)} \right\ ^2 \right] \right) = \mathcal{O}(L/\mu)$	
$\delta^2 \in \left[ \left( 1 - \frac{\eta \mu}{2} \left( 1 - \frac{\eta L}{2} \right) \right) \vee \left( 1 - h \frac{1 - \bar{\gamma}_w}{4} (1 - \bar{\gamma}_{I_N-W}) \right), 1 \right], \delta^2 = \mathcal{O}(1), \frac{\delta^2}{1 - \delta^2} = \mathcal{O} \left( \frac{1}{\eta \mu} \vee \frac{1}{h} \right)$	

Table 2: Summary of the constants in the proof of Proposition 5.

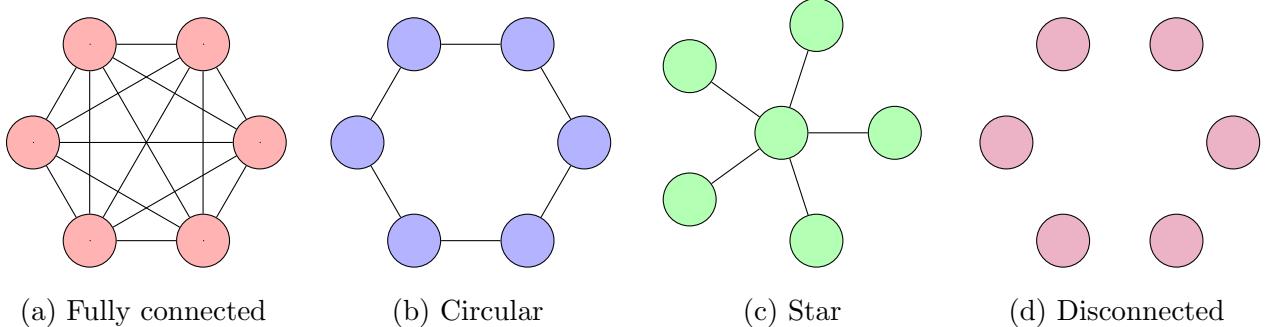


Figure 1: Different types of network structures

among the agents or nodes of different network structures. We ensure that each agent receives randomly distributed independent and identically distributed (i.i.d.) sample data.

Figure 1 represents four different types of networks: (a) fully connected network, (b) circular network, (c) star network, and (d) fully disconnected network where no agents are connected. A fully connected network is a structure in which all nodes are connected to each other. In contrast, a circular network is one in which each node is only connected to its immediate left and right neighbors. Additionally, in a star-shaped structure, the central node is connected to all other nodes, but those nodes are not connected to each other.

## 7.1 Network architecture

We follow the common approach to select the communication matrix  $W = I_N - \delta L$  where  $I_N$  is the  $N \times N$  identity matrix,  $L$  is the graph Laplacian, and  $\delta > 0$  is a small number [Chu97]. In our experiments, we select  $\delta$  in the following way. First, we compute the graph Laplacian  $L = (D_{deg} - A)$  from the degree matrix  $D_{deg}$  and the adjacency matrix  $A$ . The degree  $D_{deg}$  is a diagonal matrix with the entries in the main diagonal representing the degree of connections of each node and the adjacent matrix  $A = (a_{ij})_{1 \leq i,j \leq N}$  is the matrix with  $a_{ij} = 1$  if there is an edge between the nodes  $i$  and  $j$  otherwise  $a_{ij} = 0$ . Then we choose  $\delta$  at random so that  $0 < \delta < \frac{2}{\lambda_N^L}$ , where  $\lambda_N^L$  is the largest real eigenvalue of  $L$ . For example, a star-like graph with  $N$  vertices is given by

$$W = I_N - \delta L = \begin{bmatrix} 1 - \delta(N-1) & \delta & \delta & \cdots & \delta & \delta \\ \delta & 1 - \delta & 0 & 0 & \cdots & 0 \\ \delta & 0 & 1 - \delta & 0 & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ \delta & 0 & \cdots & 0 & 1 - \delta & 0 \\ \delta & 0 & \cdots & \cdots & 0 & 1 - \delta \end{bmatrix}.$$

For the EXTRA SGLD algorithm, we compute that  $\widetilde{W} = hI_N - (1-h)W$  for  $h \in (0, 1/2]$ .

## 7.2 Bayesian linear regression

First, we present the Bayesian linear regression with the synthetic data that we generate by simulating the following model:

$$\delta_i \sim \mathcal{N}(0, \xi^2), \quad X_i \sim \mathcal{N}(0, I_2), \quad y_i = \beta^T X_i + \delta_i, \quad (7.1)$$

where the white noise  $\delta_i$ 's are i.i.d. scalars with  $\xi = 1$ ,  $\beta \in \mathbb{R}^2$ , and  $I_2$  is the  $2 \times 2$  identity matrix. The prior distribution of  $\beta$  follows  $\mathcal{N}(0, \lambda I_2)$ , and we set  $\lambda = 10$  for this set of experiments. The posterior distribution can be derived from the following model

$$\pi(\beta) \sim \mathcal{N}(m, V), \quad m := \left( \Sigma^{-1} + \frac{X^T X}{\xi^2} \right)^{-1} \left( \frac{X^T y}{\xi^2} \right), \quad V := \left( \frac{X^T X}{\xi^2} + \Sigma^{-1} \right)^{-1},$$

where  $\Sigma = \lambda I_2$  is the covariance matrix of the prior of  $\beta$ , and  $X = [X_1^T, X_2^T, \dots]^T$  and  $Y = [y_1, y_2, \dots]^T$  are the input and output matrices, respectively. For this experiment, we simulate 5000 data points using the model (7.1) and then we distribute these data points randomly among the  $N = 20$  agents. All agents have an equal amount of data exclusively, and share only the parameter estimates. The posterior distribution  $\pi(\beta) \propto e^{-f(\beta)}$  where  $f(\beta) = \sum_{i=1}^N f_i(\beta)$  with

$$f_i(\beta) := - \sum_{j=1}^{n_i} \log p(y_j^i | \beta, X_j^i) - \frac{1}{N} \log p(\beta) = \sum_{j=1}^{n_i} (y_j^i - \beta^T X_j^i)^2 + \frac{1}{2\lambda N} \|\beta\|^2,$$

where

$$p(y_j^i | \beta, X_j^i) = \frac{1}{\sqrt{2\pi\xi^2}} e^{-\frac{1}{2\xi^2}(y_j^i - \beta^T X_j^i)^2}, \quad p(\beta) \propto e^{-\frac{1}{2\lambda}\|\beta\|^2},$$

and each agent  $i$  has an equal number of  $n_i = 50$  data points  $\{(X_j^i, y_j^i)\}_{j=1}^{n_i}$ .

We report the results of the EXTRA SGLD algorithm as follows. Figure 2 presents the results of the four networks. We restrict the experiments with a deterministic gradient, i.e.  $\sigma = 0$  and a fixed step size  $\eta = 0.009$ . The doubly stochastic mixing matrix  $\widetilde{W} = hI_N - (1-h)W$  is calculated for different values of the parameter  $h$ . We consider 5 linearly spaced  $h$  values with the minimum being 0.001 and the maximum being 0.5 and we tune up the parameter  $h$  to the network. For the fully connected network  $h = 0.50$ , circular network  $h = 0.38$ , star network  $h = 0.13$ , and for the disconnected network  $h = 0.38$ . In this setup, the iterations  $\beta_i^{(k)} \sim \mathcal{N}(m_i^{(k)}, \Sigma_i^{(k)})$  for some mean vector  $m_i^{(k)}$  and covariance matrix  $\Sigma_i^{(k)}$ , by using the formula from [GS84], we can compute the 2-Wasserstein distance,  $\mathcal{W}_2$  with the posterior distribution  $\pi(\beta) \sim \mathcal{N}(m, V)$ . From 200 independent runs, we can estimate  $m_i^{(k)}$  and  $\Sigma_i^{(k)}$  and then plot the  $\mathcal{W}_2$  distance of the stationary distribution for each agent and the distribution of the average  $\bar{\beta}^{(k)} = \frac{1}{N} \sum_i \beta_i^{(k)}$ . From the plot, we see that for the first three network types, all the agents converge to the posterior distribution up to some error level. However, the convergence for the star-type network is not as good as compared to a fully connected and circular-type network. In the case of a disconnected network, individual agents perform relatively worse compared to other scenarios where the network is connected, as they are unable to leverage information from their neighbors' data points. We also notice that the convergence is better for strongly connected networks, i.e. as the agent in a network loses its connectivity, the convergence becomes slower.

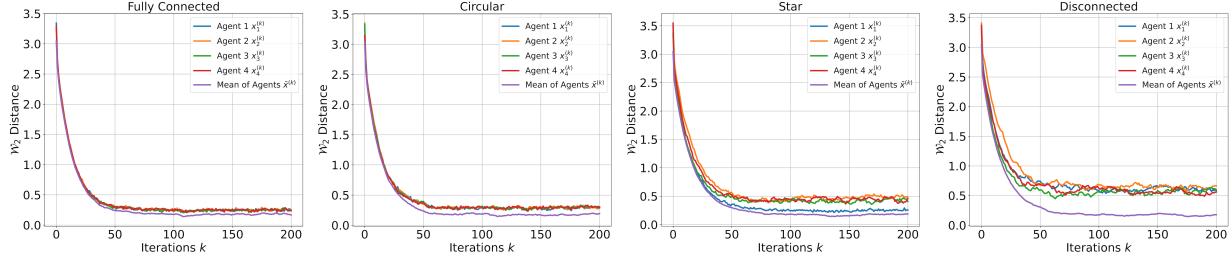


Figure 2: Performance of the EXTRA SGLD for Bayesian linear regression on four different network structures. Out of 20 agents, we report only the first 4 agents and the mean of the nodes  $\bar{\beta}^{(k)} = \frac{1}{N} \sum_{i=1}^N \beta_i^{(k)}$ .

Next, we present the comparative analysis of the performances of the DE-SGLD and EXTRA SGLD algorithms. In this case, instead of computing the  $\mathcal{W}_2$  distances for each agent, we compute the  $\mathcal{W}_2$  distances of the mean of the agents from the posterior distribution  $\pi(\beta) \sim \mathcal{N}(m, V)$  for all four networks. Then we compute the minimum of these distances for each network and plot them against the mean of the nodes from the DE-SGLD algorithm in the same plot which is represented in Figure 3.

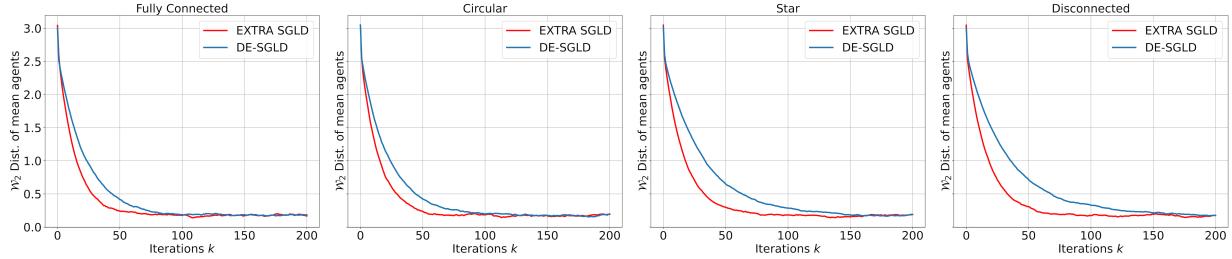


Figure 3: Comparative performance of the DE-SGLD and EXTRA SGLD for Bayesian linear regression on four different network structures in terms of the  $\mathcal{W}_2$  distance of mean agents

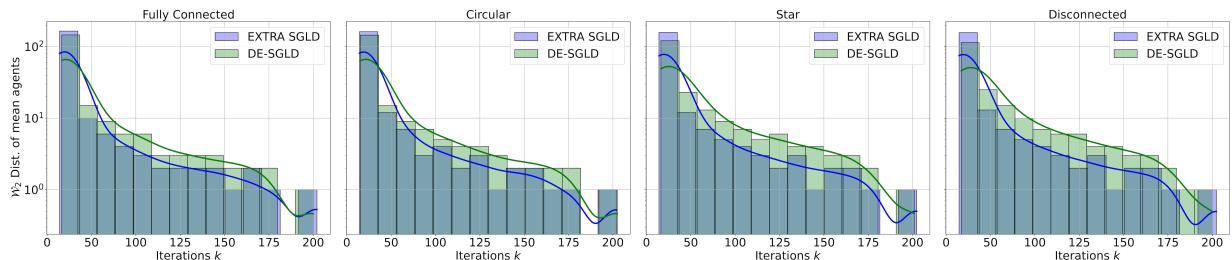


Figure 4: Histogram of the comparative performances of the DE-SGLD and EXTRA SGLD for Bayesian linear regression on four different network structures.

In the comparative analysis of EXTRA SGLD and DE-SGLD, the performance is evaluated across various network structures, as shown in Figures 3 and 4. Figure 3 illustrates the  $\mathcal{W}_2$  distance of the mean agents over 200 iterations for fully connected, circular, star, and disconnected networks, revealing that EXTRA SGLD consistently achieves faster and more stable convergence than DE-SGLD for any  $h \in (0, 1/2]$ . Figure 4, which depicts histograms of distance distributions on a logarithmic scale at specific iterations, shows that EXTRA SGLD achieves a more concentrated distribution near zero, indicating better convergence. Asymptotically, both algorithms stabilize, but EXTRA SGLD reaches a smaller quantity in  $\mathcal{W}_2$  distance, highlighting its efficiency in attaining consensus, especially in the more challenging disconnected setting.

### 7.3 Bayesian logistic regression with synthetic data

To test the performance of our algorithm, we first implement the Bayesian logistic regression on synthetic data. Ideally, we have a dataset  $Z = \{z_j\}_{j=1}^n$  where  $z_j = (X_j, y_j)$ ,  $X_j \in \mathbb{R}^d$  are the features and  $y_j \in \{0, 1\}$  are the labels with the assumption that  $X_j$  are independent and the probability distribution of  $y_j$  given  $X_j$  and regression coefficients  $\beta \in \mathbb{R}^d$  is given by

$$\mathbb{P}(y_j = 1|X_j, \beta) = \frac{1}{1 + e^{-\beta^T X_j}}. \quad (7.2)$$

The prior distribution  $p(\beta) \sim \mathcal{N}(0, \lambda I_3)$  for some  $\lambda > 0$ , where  $I_3$  is the  $3 \times 3$  identity matrix [CFM<sup>+</sup>18, DRW<sup>+</sup>16, ZXG18]. In a distributed network system, if each agent  $i$  contains a subset  $Z_i$  of data, then the goal of the Bayesian logistic regression is to sample from  $\pi(\beta) \propto e^{-f(\beta)}$  with  $f(\beta) = \sum_{i=1}^N f_i(\beta)$  where

$$f_i(\beta) := - \sum_{j=1}^{n_i} \log p(y_j^i = 1|X_j^i, \beta) - \frac{1}{N} \log p(\beta) = \sum_{j=1}^{n_i} \log \left( 1 + e^{-\beta^T X_j} \right) + \frac{1}{2N\lambda} \|\beta\|^2 \quad (7.3)$$

is strongly convex and smooth. We generate the synthetic data from the following model

$$X_j \sim \mathcal{N}(0, 20I_3), \quad p_j \sim \mathcal{U}(0, 1), \quad y_j = \begin{cases} 1 & \text{if } p_j \leq \frac{1}{1+e^{-\beta^T X_j}}, \\ 0 & \text{otherwise,} \end{cases}$$

where  $\mathcal{U}(0, 1)$  is the uniform distribution on  $[0, 1]$ ,  $\beta = [\beta_1, \beta_2, \beta_3]^T \in \mathbb{R}^3$  and the prior distribution  $\beta \sim \mathcal{N}(0, \lambda I_3)$ , where  $I_3$  is the  $3 \times 3$  identity matrix. For this experiment, we take  $\lambda = 10$  just like the linear regression, but we limit the number of nodes to  $N = 6$  and distribute the data points equally to each node. For Bayesian logistic regression, we consider 10 linearly spaced  $h$  values and tune the parameter to the network style. For the fully connected network, we take  $h = 0.111$ , circular network  $h = 0.056$ , star network  $h = 0.001$  and for the disconnected network  $h = 0.445$ . For each node  $i$ , we calculate their accuracy over  $n = 1000$  data points and 20 runs with batch size  $b = 32$  and step size  $\eta = 0.005$ . However, unlike Bayesian linear regression, Bayesian logistic regression does not have a closed-form solution for the posterior distribution  $\pi(\beta)$ . Therefore, in order to compute  $\mathcal{W}_2$  distance between the prior and posterior distributions for Bayesian logistic regression, one may need to run the algorithm over many iterations which is not practical. For these shortcomings, we apply a different technique to measure the performance of the algorithm which is the distribution of the accuracy over the whole data set. This accuracy measure is defined as the ratio of the correctly predicted labels over the whole data set. Since our experiments are

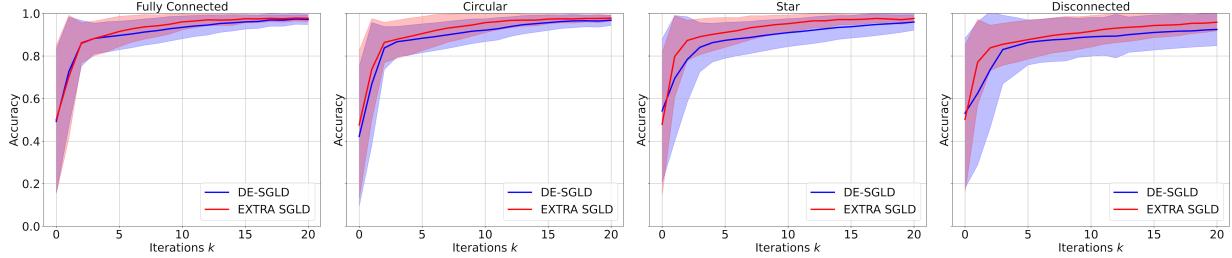


Figure 5: Accuracy distribution of the EXTRA SGLD method across different network structures at a randomly selected node.

identical except for the  $h$  parameters to that of the DE-SGLD.

Figure 5 shows the mean and standard deviation of the accuracy distribution of the agent  $i$  using the DE-SGLD and EXTRA SGLD algorithms. It is clearly noticeable from Figure 5 that, from left to right, all three networks provide somewhat better accuracy than the disconnected one. We also see that for any  $h \in (0, 1/2]$ , EXTRA SGLD performs slightly better than DE-SGLD in general in terms of the accuracy distribution irrespective of network structures.

#### 7.4 Bayesian logistic regression with real data

At this point, we implement our algorithms for real data. In this case, we consider the UCI ML Breast Cancer Wisconsin (Diagnostic) data set [WMSS95]. The data set contains 569 instances with 30 features which are computed from a digitized image of a fine needle aspirate (FNA) of a breast mass. For this experiment, we keep the other parameters same as the logistic regression with synthetic data except for the EXTRA parameter  $h$ . In this case, we take  $h = 0.278, 0.389, 0.167$ , and  $0.278$  for fully connected network, circular network, star network, and disconnected network, respectively.

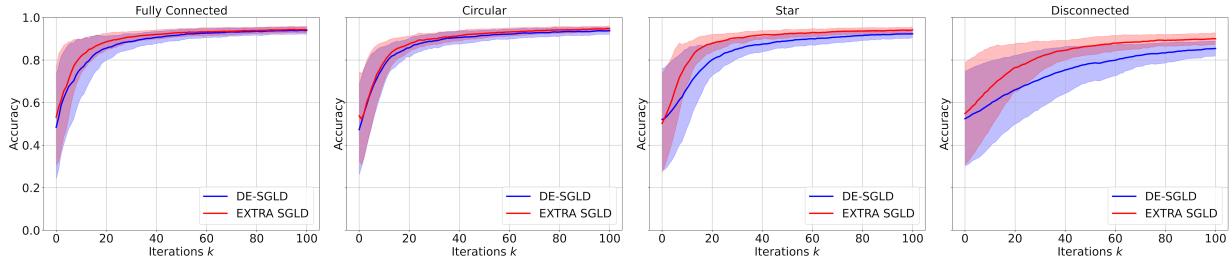


Figure 6: Comparative accuracy distribution of the DE-SGLD and EXTRA SGLD method across different network structures on Breast Cancer data set. The plots are from a randomly selected node.

Figure 6 represents the comparative accuracy distribution of the DE-SGLD and EXTRA SGLD algorithms for Bayesian logistic regression type problems. Both algorithms exhibit a rapid increase in accuracy during the initial iterations. EXTRA SGLD and DE-SGLD behave similarly in the

early stages; but when the number of iterations increases, EXTRA SGLD achieves a higher accuracy. The shaded regions around the accuracy curves indicate the variance, with EXTRA SGLD demonstrating less variability in general, which implies a more stable performance. As iterations progress, both algorithms converge towards their maximum accuracy. Nonetheless, EXTRA SGLD maintains a slight edge, reaching a higher asymptotic accuracy and showcasing a reliable convergence behavior across fully connected, circular, star, and disconnected network structures.

## 8 Conclusion

Langevin algorithms are widely used Markov Chain Monte Carlo methods in Bayesian learning, particularly for sampling from a parametric model’s posterior distribution based on input data and prior parameter distributions. Their stochastic variants, such as stochastic gradient Langevin dynamics (SGLD), facilitate iterative learning using mini-batches from large datasets, making them scalable. However, in scenarios where data are decentralized across a network with communication and privacy restrictions, standard SGLD approaches are unsuitable. To address this, we utilize decentralized SGLD (DE-SGLD) algorithms, which enable collaborative Bayesian learning across a network of agents without sharing individual data points. Despite their advantages, existing DE-SGLD algorithms introduce a bias at each agent that can degrade performance. This bias persists even with full-batch processing and stems from network-related effects. Inspired by the EXTRA algorithm and its generalizations for decentralized optimization, we introduce a generalized EXTRA SGLD that eliminates this bias in full-batch scenarios. Additionally, we demonstrate that, in the mini-batch context, our algorithm offers performance bounds that significantly surpass those of conventional DE-SGLD algorithms. Our empirical results further validate the effectiveness of our approach.

## Acknowledgments

Mert Gürbüzbalaban’s research is supported in part by the grants Office of Naval Research Award Number N00014-21-1-2244, National Science Foundation (NSF) CCF-1814888, NSF DMS-2053485. Mohammad Rafiqul Islam is partially supported by the grant NSF DMS-2053454. Xiaoyu Wang is supported by the Guangzhou-HKUST(GZ) Joint Funding Program (No.2024A03J0630), Guangzhou Municipal Key Laboratory of Financial Technology Cutting-Edge Research. Lingjiong Zhu is partially supported by the grants NSF DMS-2053454 and DMS-2208303.

## References

- [ABC<sup>+</sup>20] Yossi Arjevani, Joan Bruna, Bugra Can, Mert Gürbüzbalaban, Stefanie Jegelka, and Hongzhou Lin. IDEAL: Inexact DEcentralized accelerated augmented Lagrangian method. In *Advances in Neural Information Processing Systems*, volume 33, 2020.
- [AFGO19] Necdet Serhat Aybat, Alireza Fallah, Mert Gurbuzbalaban, and Asuman Ozdaglar. A universally optimal multistage accelerated stochastic gradient method. In *Advances in Neural Information Processing Systems*, volume 32, 2019.

- [BCM<sup>+</sup>21] Mathias Barkhagen, Ngoc Huy Chau, Éric Moulines, Miklós Rásonyi, Sotirios Sabanis, and Ying Zhang. On stochastic gradient Langevin dynamics with dependent data streams in the logconcave case. *Bernoulli*, 27(1):1–33, 2021.
- [Bot10] Léon Bottou. Large-scale machine learning with stochastic gradient descent. In *Proceedings of COMPSTAT’2010*, pages 177–186. Springer, 2010.
- [CB18] Xiang Cheng and Peter L. Bartlett. Convergence of Langevin MCMC in KL-divergence. In *Proceedings of the 29th International Conference on Algorithmic Learning Theory (ALT)*, volume 83, pages 186–211. PMLR, 2018.
- [CFM<sup>+</sup>18] Niladri Chatterji, Nicolas Flammarion, Yian Ma, Peter Bartlett, and Michael Jordan. On the theory of variance reduction for stochastic gradient Monte Carlo. In *International Conference on Machine Learning*, volume 80, pages 764–773. PMLR, 2018.
- [Chu97] Fan RK Chung. *Spectral Graph Theory*, volume 92. American Mathematical Society, 1997.
- [CMR<sup>+</sup>21] Ngoc Huy Chau, Éric Moulines, Miklos Rásonyi, Sotirios Sabanis, and Ying Zhang. On stochastic gradient Langevin dynamics with dependent data streams: the fully non-convex case. *SIAM Journal of Mathematics of Data Science*, 3(3):959–986, 2021.
- [Dal17] Arnak S Dalalyan. Theoretical guarantees for approximate sampling from smooth and log-concave densities. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 79(3):651–676, 2017.
- [DK19] Arnak S Dalalyan and Avetik Karagulyan. User-friendly guarantees for the Langevin Monte Carlo with inaccurate gradient. *Stochastic Processes and their Applications*, 129(12):5278–5311, 2019.
- [DM17] Alain Durmus and Eric Moulines. Non-asymptotic convergence analysis for the Unadjusted Langevin Algorithm. *Annals of Applied Probability*, 27(3):1551–1587, 2017.
- [DM19] Alain Durmus and Eric Moulines. High-dimensional Bayesian inference via the Unadjusted Langevin Algorithm. *Bernoulli*, 25(4A):2854–2882, 2019.
- [DRW<sup>+</sup>16] Kumar Avinava Dubey, Sashank J Reddi, Sinead A Williamson, Barnabas Poczos, Alexander J Smola, and Eric P Xing. Variance reduction in stochastic gradient Langevin dynamics. In *Advances in Neural Information Processing Systems*, pages 1154–1162, 2016.
- [EHZ22] Murat A. Erdogdu, Rasa Hosseinzadeh, and Matthew S. Zhang. Convergence analysis of Langevin Monte Carlo in chi-square and Rényi divergence. In *Proceedings of the 25th International Conference on Artificial Intelligence and Statistics (AISTATS)*, volume 151, pages 8151–8175. PMLR, 2022.
- [FGO<sup>+</sup>22] Alireza Fallah, Mert Gürbüzbalaban, Asuman Ozdaglar, Umut Şimşekli, and Lingjiong Zhu. Robust distributed accelerated stochastic gradient methods for multi-agent networks. *Journal of Machine Learning Research*, 23(220):1–96, 2022.

- [GDG19] Eduard Gorbunov, Darina Dvinskikh, and Alexander Gasnikov. Optimal decentralized distributed algorithms for stochastic convex optimization. *arXiv:1911.07363*, 2019.
- [GGHZ21] Mert Gürbüzbalaban, Xuefeng Gao, Yuanhan Hu, and Lingjiong Zhu. Decentralized stochastic gradient Langevin dynamics and Hamiltonian Monte Carlo. *Journal of Machine Learning Research*, 22(1):10804–10872, 2021.
- [GS84] Clark R Givens and Rae Michael Shortt. A class of Wasserstein metrics for probability distributions. *Michigan Mathematical Journal*, 31(2):231–240, 1984.
- [HBJ18] Lie He, An Bian, and Martin Jaggi. COLA: Decentralized linear learning. In *Advances in Neural Information Processing Systems*, pages 4536–4546, 2018.
- [HBM19] Hadrien Hendrikx, Francis Bach, and Laurent Massoulié. An accelerated decentralized stochastic proximal algorithm for finite sums. In *Advances in Neural Information Processing Systems*, volume 32, pages 954–964, 2019.
- [Hof09] Peter D Hoff. *A First Course in Bayesian Statistical Methods*, volume 580. Springer, 2009.
- [Jak18] Dušan Jakovetić. A unification and generalization of exact distributed first-order methods. *IEEE Transactions on Signal and Information Processing over Networks*, 5(1):31–46, 2018.
- [Ned20] Angelia Nedic. Distributed gradient methods for convex machine learning problems in networks: Distributed optimization. *IEEE Signal Processing Magazine*, 37(3):92–101, 2020.
- [NO09] Angelia Nedic and Asuman Ozdaglar. Distributed subgradient methods for multi-agent optimization. *IEEE Transactions on Automatic Control*, 54(1):48–61, 2009.
- [Pav14] Grigorios A Pavliotis. *Stochastic Processes and Applications: Diffusion processes, the Fokker-Planck and Langevin Equations*, volume 60. Springer, 2014.
- [PBGG20] Anjaly Parayil, He Bai, Jemin George, and Prudhvi Gurram. Decentralized Langevin dynamics for Bayesian learning. In *Advances in Neural Information Processing Systems*, volume 33, 2020.
- [PS17] Nicholas G. Polson and Vadim Sokolov. Deep learning: A Bayesian perspective. *Bayesian Analysis*, 12(4):1275–1304, 2017.
- [RRT17] Maxim Raginsky, Alexander Rakhlin, and Matus Telgarsky. Non-convex learning via stochastic gradient Langevin dynamics: a nonasymptotic analysis. In *Conference on Learning Theory*, volume 65, pages 1674–1703. PMLR, 2017.
- [SBB<sup>+</sup>19] Kevin Scaman, Francis Bach, Sébastien Bubeck, Yin Lee, and Laurent Massoulié. Optimal convergence rates for convex distributed optimization in networks. *Journal of Machine Learning Research*, 20:1–31, 2019.

- [SKP<sup>+</sup>20] Brian Swenson, Soummya Kar, H. Vincent Poor, José M. F. Moura, and Aaron Jaech. Distributed gradient methods for nonconvex optimization: Local and global convergence guarantees. *arXiv e-prints*, page arXiv:2003.10309, March 2020.
- [SLWY15] 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.
- [SSP20] Brian Swenson, Anirudh Sridhar, and H Vincent Poor. On distributed stochastic gradient algorithms for global optimization. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 8594–8598. IEEE, 2020.
- [ULGN17] César A Uribe, Soomin Lee, Alexander Gasnikov, and Angelia Nedić. Optimal algorithms for distributed optimization. *arXiv preprint arXiv:1712.00232*, 2017.
- [Vil09] Cédric Villani. *Optimal Transport: Old and New*. Springer, Berlin, 2009.
- [WMSS95] William Wolberg, Olvi Mangasarian, Nick Street, and W. Street. Breast Cancer Wisconsin (Diagnostic). UCI Machine Learning Repository, 1995. DOI: <https://doi.org/10.24432/C5DW2B>.
- [WT11] Max Welling and Yee Whye Teh. Bayesian learning via stochastic gradient Langevin dynamics. In *Proceedings of the 28th International Conference on Machine Learning (ICML-11)*, pages 681–688, 2011.
- [WY20] Hao Wang and Dit-Yan Yeung. A survey on Bayesian deep learning. *ACM Computing Surveys*, 52(5):1–37, 2020.
- [XCZG18] Pan Xu, Jinghui Chen, Difan Zou, and Quanquan Gu. Global convergence of Langevin dynamics based algorithms for nonconvex optimization. In *Advances in Neural Information Processing Systems*, pages 3122–3133, 2018.
- [YLY16] Kun Yuan, Qing Ling, and Wotao Yin. On the convergence of decentralized gradient descent. *SIAM Journal on Optimization*, 26(3):1835–1854, 2016.
- [ZADS23] Ying Zhang, Ömer Deniz Akyildiz, Theodoros Damoulas, and Sotirios Sabanis. Nonasymptotic estimates for Stochastic Gradient Langevin Dynamics under local conditions in nonconvex optimization. *Applied Mathematics & Optimization*, 87:25, 2023.
- [ZXG18] Difan Zou, Pan Xu, and Quanquan Gu. Subsampled stochastic variance-reduced gradient Langevin dynamics. In *34th Conference on Uncertainty in Artificial Intelligence 2018, UAI 2018*, 2018.

## A Proofs of the Key Technical Results

### A.1 Proof of Theorem 13

By substituting the upper bound for  $\|\tilde{\mathbf{v}}\|_2^{\delta, K}$  to the upper bound for  $\|\tilde{\mathbf{x}}\|_2^{\delta, K}$  in (6.29), we get

$$\begin{aligned} \|\tilde{\mathbf{x}}\|_2^{\delta, K} &\leq \eta\gamma_1 \left( (h/\eta)\gamma_2 \|\tilde{\mathbf{x}}\|_2^{\delta, K} + (h/\eta) \frac{w_2(\eta\sigma^2 + 2d)}{N\delta^{2K-2}} \right. \\ &\quad \left. + \delta^2(h/\eta)(E_3/\eta)\mathbb{E}\left[\left\|\bar{e}_x^{(0)}\right\|^2\right] + \delta^2(E_4/h)\mathbb{E}\left[\left\|\tilde{v}^{(0)}\right\|^2\right] \right) \\ &\quad + \eta \frac{w_1(\eta\sigma^2 + 2d)}{N\delta^{2K-2}} + \delta^2 E_1 \mathbb{E}\left[\left\|\tilde{x}^{(0)}\right\|^2\right] + \delta^2 E_2 \mathbb{E}\left[\left\|\bar{e}_x^{(0)}\right\|^2\right]. \end{aligned} \quad (\text{A.1})$$

Under the assumption  $h < \frac{1}{\gamma_1\gamma_2}$ , where  $h$  is defined in (3.9), we have  $h\gamma_1\gamma_2 < 1$  and the constants  $\gamma_1, \gamma_2$  are constants independent of  $\eta$  and  $\delta$  in (6.30). We can compute that

$$\begin{aligned} \|\tilde{\mathbf{x}}\|_2^{\delta, K} &\leq \frac{1}{1-h\gamma_1\gamma_2} \left\{ \eta^2 \cdot \frac{(w_2\gamma_1(h/\eta) + w_1)\sigma^2}{N\delta^{2K-2}} + \eta \cdot \left[ \frac{2d(w_2\gamma_1(h/\eta) + w_1)}{N\delta^{2K-2}} \right. \right. \\ &\quad \left. \left. + \gamma_1\delta^2 \left( (h/\eta)(E_3/\eta)\mathbb{E}\left[\left\|\bar{e}_x^{(0)}\right\|^2\right] + (E_4/h)\mathbb{E}\left[\left\|\tilde{v}^{(0)}\right\|^2\right] \right) \right] \right. \\ &\quad \left. + \delta^2 \left( E_1 \mathbb{E}\left[\left\|\tilde{x}^{(0)}\right\|^2\right] + E_2 \mathbb{E}\left[\left\|\bar{e}_x^{(0)}\right\|^2\right] \right) \right\} \\ &:= \frac{\eta^2}{\delta^{2K-2}} \cdot \frac{(w_2\gamma_1(h/\eta) + w_1)\sigma^2/N}{1-h\gamma_1\gamma_2} + \frac{\eta}{\delta^{2K-2}} \cdot \left[ \frac{2d(w_2\gamma_1(h/\eta) + w_1)/N}{1-h\gamma_1\gamma_2} \right. \\ &\quad \left. + \frac{\gamma_1}{1-h\gamma_1\gamma_2}\delta^{2K} \left( (h/\eta)(E_3/\eta)\mathbb{E}\left[\left\|\bar{e}_x^{(0)}\right\|^2\right] + (E_4/h)\mathbb{E}\left[\left\|\tilde{v}^{(0)}\right\|^2\right] \right) \right] + \delta^2 D_0, \end{aligned} \quad (\text{A.2})$$

with  $D_0 := \frac{1}{1-h\gamma_1\gamma_2} \left( E_1 \mathbb{E}\left[\left\|\tilde{x}^{(0)}\right\|^2\right] + E_2 \mathbb{E}\left[\left\|\bar{e}_x^{(0)}\right\|^2\right] \right)$ .

Finally, by Lemma 12 (or equivalently (6.29)) for the bound of  $\|\tilde{\mathbf{v}}\|_2^{\delta, K}$  and the bound of  $\|\tilde{\mathbf{x}}\|_2^{\delta, K}$

in (A.2), we get

$$\begin{aligned}
\|\tilde{\mathbf{v}}\|_2^{\delta,K} &\leq (h/\eta)\gamma_2 \|\tilde{\mathbf{x}}\|_2^{\delta,K} + (h/\eta) \frac{w_2(\eta\sigma^2 + 2d)}{N\delta^{2K-2}} + \delta^2(h/\eta)(E_3/\eta)\mathbb{E}\left[\left\|\bar{e}_x^{(0)}\right\|^2\right] + \frac{\delta^2}{h}E_4\mathbb{E}\left[\left\|\tilde{v}^{(0)}\right\|^2\right] \\
&\leq (h/\eta)\gamma_2 \left[ \frac{\eta^2}{\delta^{2K-2}} \cdot \frac{(w_2\gamma_1(h/\eta) + w_1)\sigma^2/N}{1 - h\gamma_1\gamma_2} + \frac{\eta}{\delta^{2K-2}} \cdot \left[ \frac{2d(w_2\gamma_1(h/\eta) + w_1)/N}{1 - h\gamma_1\gamma_2} \right. \right. \\
&\quad \left. \left. + \frac{\gamma_1}{1 - h\gamma_1\gamma_2}\delta^{2K}\left((h/\eta)(E_3/\eta)\mathbb{E}\left[\left\|\bar{e}_x^{(0)}\right\|^2\right] + (E_4/h)\mathbb{E}\left[\left\|\tilde{v}^{(0)}\right\|^2\right]\right)\right] + \delta^2D_0 \right] \\
&\quad + (h/\eta) \frac{w_2(\eta\sigma^2 + 2d)}{N\delta^{2K-2}} + \delta^2(h/\eta)(E_3/\eta)\mathbb{E}\left[\left\|\bar{e}_x^{(0)}\right\|^2\right] + \delta^2(E_4/h)\mathbb{E}\left[\left\|\tilde{v}^{(0)}\right\|^2\right] \\
&= \frac{h\eta}{\delta^{2K-2}} \cdot \frac{\gamma_2(w_2\gamma_1(h/\eta) + w_1)\sigma^2/N}{1 - h\gamma_1\gamma_2} + \frac{h}{\delta^{2K-2}} \cdot \left[ \frac{2\gamma_2d(w_2\gamma_1(h/\eta) + w_1)/N}{1 - h\gamma_1\gamma_2} + \frac{w_2\sigma^2}{N} \right. \\
&\quad \left. + \frac{\gamma_1\gamma_2}{1 - h\gamma_1\gamma_2}\delta^{2K}\left((h/\eta)(E_3/\eta)\mathbb{E}\left[\left\|\bar{e}_x^{(0)}\right\|^2\right] + (E_4/h)\mathbb{E}\left[\left\|\tilde{v}^{(0)}\right\|^2\right]\right)\right] + (h/\eta)\delta^2\gamma_2D_0 \\
&\quad + (h/\eta) \frac{2dw_2}{N\delta^{2K-2}} + \delta^2(h/\eta)(E_3/\eta)\mathbb{E}\left[\left\|\bar{e}_x^{(0)}\right\|^2\right] + \delta^2(E_4/h)\mathbb{E}\left[\left\|\tilde{v}^{(0)}\right\|^2\right]. \quad (\text{A.3})
\end{aligned}$$

The proof is complete.

## A.2 Proof of Corollary 16

By (3.21), we can compute that

$$\begin{aligned}
x^{(k+1)} &= \left(\widetilde{W}^{k+1} \otimes I_d\right)x^{(0)} - \eta \sum_{s=0}^k \left(\widetilde{W}^{k-s} \otimes I_d\right) \left(\nabla F\left(x^{(s)}\right) + v^{(s)}\right) - \eta\xi^{(k)} + \sqrt{2\eta}w^{(k+1)} \\
&\quad - \eta \sum_{s=0}^k \left(\widetilde{W}^{k-s} \otimes I_d\right) \xi^{(s+1)} + \sqrt{2\eta} \sum_{s=0}^k \left(\widetilde{W}^{k-s} \otimes I_d\right) w^{(s+1)}. \quad (\text{A.4})
\end{aligned}$$

Then, by using the definition of  $\bar{\mathbf{x}}^{(k)}$  in (6.11), we can compute that

$$x^{(k+1)} - \bar{\mathbf{x}}^{(k+1)} = x^{(k+1)} - \frac{1}{N} ((1_N 1_N^T) \otimes I_d) x^{(k+1)} \quad (\text{A.5})$$

$$\begin{aligned} &= \left( \widetilde{W}^{k+1} \otimes I_d \right) x^{(0)} - \frac{1}{N} \left( (1_N 1_N^T \widetilde{W}^{k+1}) \otimes I_d \right) x^{(0)} \\ &\quad - \eta \sum_{s=0}^k \left( \widetilde{W}^{k-s} \otimes I_d \right) \left( \nabla F(x^{(s)}) + v^{(s)} \right) \\ &\quad + \eta \sum_{s=0}^k \frac{1}{N} \left( (1_N 1_N^T \widetilde{W}^{k-s}) \otimes I_d \right) \left( \nabla F(x^{(s)}) + v^{(s)} \right) \\ &\quad - \eta \sum_{s=0}^k \left( \widetilde{W}^{k-s} \otimes I_d \right) \xi^{(s+1)} + \eta \sum_{s=0}^k \frac{1}{N} \left( (1_N 1_N^T \widetilde{W}^{k-s}) \otimes I_d \right) \xi^{(s+1)} \\ &\quad + \sqrt{2\eta} \sum_{s=0}^k \left( \widetilde{W}^{k-s} \otimes I_d \right) w^{(s+1)} - \sqrt{2\eta} \sum_{s=0}^k \frac{1}{N} \left( (1_N 1_N^T \widetilde{W}^{k-s}) \otimes I_d \right) w^{(s+1)}. \end{aligned} \quad (\text{A.6})$$

It follows that

$$\begin{aligned} \left\| x^{(k+1)} - \frac{1}{N} ((1_N 1_N^T) \otimes I_d) x^{(k+1)} \right\|^2 &\leq 4 \left\| \left( \left( \widetilde{W}^{k+1} - \frac{1}{N} 1_N 1_N^T \right) \otimes I_d \right) x^{(0)} \right\|^2 \\ &\quad + 4\eta^2 \left\| \sum_{s=0}^k \left( \left( \widetilde{W}^{k-s} - \frac{1}{N} 1_N 1_N^T \right) \otimes I_d \right) \left( \nabla F(x^{(s)}) + v^{(s)} \right) \right\|^2 \\ &\quad + 4\eta^2 \left\| \sum_{s=0}^k \left( \left( \widetilde{W}^{k-s} - \frac{1}{N} 1_N 1_N^T \right) \otimes I_d \right) \xi^{(s+1)} \right\|^2 \\ &\quad + 8\eta \left\| \sum_{s=0}^k \left( \left( \widetilde{W}^{k-s} - \frac{1}{N} 1_N 1_N^T \right) \otimes I_d \right) w^{(s+1)} \right\|^2, \end{aligned} \quad (\text{A.7})$$

where we can further compute

$$\begin{aligned} 4 \left\| \left( \left( \widetilde{W}^{k+1} - \frac{1}{N} 1_N 1_N^T \right) \otimes I_d \right) x^{(0)} \right\|^2 &\leq 4 \left\| \left( \left( \widetilde{W}^{k+1} - \frac{1}{N} 1_N 1_N^T \right) \otimes I_d \right) \right\|^2 \mathbb{E} \left[ \|x^{(0)}\|^2 \right] \\ &\leq 4\bar{\gamma}_{\widetilde{W}}^{2k+2} \mathbb{E} \left[ \|x^{(0)}\|^2 \right]. \end{aligned} \quad (\text{A.8})$$

Then, it follows that

$$\begin{aligned}
& 4\eta^2 \left\| \sum_{s=0}^k \left( \left( \widetilde{W}^{k-s} - \frac{1}{N} \mathbf{1}_N \mathbf{1}_N^T \right) \otimes I_d \right) \left( \nabla F(x^{(s)}) + v^{(s)} \right) \right\|^2 \\
& \leq 4\eta^2 \left( \sum_{s=0}^k \left\| \widetilde{W}^{k-s} - \frac{1}{N} \mathbf{1}_N \mathbf{1}_N^T \right\| \cdot \left\| \nabla F(x^{(s)}) + v^{(s)} \right\| \right)^2 \\
& = 4\eta^2 \left( \sum_{s=0}^k \bar{\gamma}_{\widetilde{W}}^{k-s} \cdot \left( \left\| \nabla F(x^{(s)}) \right\| + \left\| v^{(s)} \right\| \right) \right)^2 \\
& = 4\eta^2 \left( \sum_{s=0}^k \bar{\gamma}_{\widetilde{W}}^{k-s} \right)^2 \left( \frac{\sum_{s=0}^k \bar{\gamma}_{\widetilde{W}}^{k-s} \cdot (\left\| \nabla F(x^{(s)}) \right\| + \left\| v^{(s)} \right\|)}{\sum_{s=0}^k \bar{\gamma}_{\widetilde{W}}^{k-s}} \right)^2 \\
& \leq 8\eta^2 \left( \sum_{s=0}^k \bar{\gamma}_{\widetilde{W}}^{k-s} \right)^2 \sum_{s=0}^k \frac{\bar{\gamma}_{\widetilde{W}}^{k-s}}{\sum_{s=0}^k \bar{\gamma}_{\widetilde{W}}^{k-s}} \left( \left\| \nabla F(x^{(s)}) \right\|^2 + \left\| v^{(s)} \right\|^2 \right). \tag{A.9}
\end{aligned}$$

Therefore, we can obtain from Lemma 14 that

$$4\eta^2 \mathbb{E} \left[ \left\| \sum_{s=0}^k \left( \left( \widetilde{W}^{k-s} - \frac{1}{N} \mathbf{1}_N \mathbf{1}_N^T \right) \otimes I_d \right) \left( \nabla F(x^{(s)}) + v^{(s)} \right) \right\|^2 \right] \leq 8\eta^2 \cdot \frac{R_h + R'_h}{(1 - \bar{\gamma}_{\widetilde{W}})^2}. \tag{A.10}$$

Finally, we can also compute that:

$$\begin{aligned}
& 4\eta^2 \mathbb{E} \left[ \left\| \sum_{s=0}^k \left( \left( \widetilde{W}^{k-s} - \frac{1}{N} \mathbf{1}_N \mathbf{1}_N^T \right) \otimes I_d \right) \xi^{(s+1)} \right\|^2 \right] \\
& \quad + 8\eta \mathbb{E} \left[ \left\| \sum_{s=0}^k \left( \left( \widetilde{W}^{k-s} - \frac{1}{N} \mathbf{1}_N \mathbf{1}_N^T \right) \otimes I_d \right) w^{(s+1)} \right\|^2 \right] \\
& \leq 4\eta^2 \sum_{s=0}^k \bar{\gamma}_{\widetilde{W}}^{2(k-s)} \mathbb{E} \left\| \xi^{(s+1)} \right\|^2 + 8\eta \sum_{s=0}^k \bar{\gamma}_{\widetilde{W}}^{2(k-s)} \mathbb{E} \left\| w^{(s+1)} \right\|^2 \\
& \leq 4\eta^2 \sigma^2 N \sum_{s=0}^k \bar{\gamma}_{\widetilde{W}}^{2(k-s)} + 8\eta d N \sum_{s=0}^k \bar{\gamma}_{\widetilde{W}}^{2(k-s)} \\
& \leq \frac{4\eta^2 \sigma^2 N}{1 - \bar{\gamma}_{\widetilde{W}}^2} + \frac{8\eta d N}{1 - \bar{\gamma}_{\widetilde{W}}^2}. \tag{A.11}
\end{aligned}$$

As a result, for every  $k = 1, 2, 3, \dots$ , we have

$$\mathbb{E} \left[ \left\| x^{(k)} - \frac{1}{N} ((\mathbf{1}_N \mathbf{1}_N^T) \otimes I_d) x^{(k)} \right\|^2 \right] \leq 4(\bar{\gamma}_{\widetilde{W}})^{2k} \mathbb{E} \left[ \left\| x^{(0)} \right\|^2 \right] + 8\eta^2 \frac{R_h + R'_h}{(1 - \bar{\gamma}_{\widetilde{W}})^2} + \frac{4\eta^2 \sigma^2 N}{1 - \bar{\gamma}_{\widetilde{W}}^2} + \frac{8\eta d N}{1 - \bar{\gamma}_{\widetilde{W}}^2}. \tag{A.12}$$

The proof is complete.

### A.3 Proof of Corollary 17

For any given  $k \geq 1$ , we can compute that

$$\begin{aligned}
\mathbb{E} \left[ \left\| \hat{\mathcal{E}}_{k+1} \right\|^2 \right] &= \mathbb{E} \left[ \left\| \frac{1}{N} \left( \sum_{i=1}^N \nabla f_i \left( \bar{x}^{(k)} \right) - \nabla f_i \left( x_i^{(k)} \right) \right) \right\|^2 \right] \\
&\leq \frac{1}{N^2} NL^2 \sum_{i=1}^N \mathbb{E} \left[ \left\| x_i^{(k)} - \bar{x}^{(k)} \right\|^2 \right] \\
&= \frac{L^2}{N} \left( 4 (\bar{\gamma}_{\bar{w}})^{2k} \mathbb{E} \left[ \left\| x^{(0)} \right\|^2 \right] + 8\eta^2 \cdot \frac{R_h + R'_h}{(1 - \bar{\gamma}_{\bar{w}})^2} + \frac{4\eta^2\sigma^2 N}{1 - \bar{\gamma}_{\bar{w}}^2} + \frac{8\eta d N}{1 - \bar{\gamma}_{\bar{w}}^2} \right) \\
&= \frac{4L^2 (\bar{\gamma}_{\bar{w}})^{2k}}{N} \mathbb{E} \left[ \left\| x^{(0)} \right\|^2 \right] + \eta^2 \cdot \frac{4L^2}{N} \left( \frac{2(R_h + R'_h)}{(1 - \bar{\gamma}_{\bar{w}})^2} + \frac{\sigma^2 N}{1 - \bar{\gamma}_{\bar{w}}^2} \right) + \eta \cdot \frac{8L^2 d}{1 - \bar{\gamma}_{\bar{w}}^2}, \quad (\text{A.13})
\end{aligned}$$

where the last equality is due to Corollary 16. Next, we recall the dynamics:

$$\bar{x}^{(k+1)} - x_{k+1} = \bar{x}^{(k)} - x_k - \frac{\eta}{N} \left( \nabla f \left( \bar{x}^{(k)} \right) - \nabla f \left( x_k \right) \right) + \eta \hat{\mathcal{E}}_{k+1} - \eta \bar{\xi}^{(k)}. \quad (\text{A.14})$$

Under the assumption  $\eta < 2/L$  and  $L$ -smoothness and the  $\mu$ -convexity of  $\frac{1}{N}f$ , we can compute that

$$\begin{aligned}
\left\| \bar{x}^{(k+1)} - x_{k+1} \right\|^2 &\leq \left( 1 - 2\eta\mu \left( 1 - \frac{\eta L}{2} \right) \right) \left\| \bar{x}^{(k)} - x_k \right\|^2 + \eta^2 \left\| \hat{\mathcal{E}}_{k+1} - \bar{\xi}^{(k)} \right\|^2 \\
&\quad + 2 \left\langle \bar{x}^{(k)} - x_k - \eta \frac{1}{N} \left( \nabla f \left( \bar{x}^{(k)} \right) - \nabla f \left( x_k \right) \right), \eta \hat{\mathcal{E}}_{k+1} - \eta \bar{\xi}^{(k)} \right\rangle. \quad (\text{A.15})
\end{aligned}$$

Next, we take expectations in (A.15) to get

$$\begin{aligned}
& \mathbb{E} \left[ \left\| \bar{x}^{(k+1)} - x_{k+1} \right\|^2 \right] \\
& \leq \left( 1 - 2\eta\mu \left( 1 - \frac{\eta L}{2} \right) \right) \mathbb{E} \left[ \left\| \bar{x}^{(k)} - x_k \right\|^2 \right] + \eta^2 \mathbb{E} \left[ \left\| \hat{\mathcal{E}}_{k+1} \right\|^2 \right] + \eta^2 \mathbb{E} \left[ \left\| \bar{\xi}^{(k)} \right\|^2 \right] \\
& \quad + 2\mathbb{E} \left\langle \bar{x}^{(k)} - x_k - \eta \frac{1}{N} \left[ \nabla f(\bar{x}^{(k)}) - \nabla f(x_k) \right], \eta \hat{\mathcal{E}}_{k+1} \right\rangle \\
& \leq \left( 1 - 2\eta\mu \left( 1 - \frac{\eta L}{2} \right) \right) \mathbb{E} \left[ \left\| \bar{x}^{(k)} - x_k \right\|^2 \right] + \eta^2 \mathbb{E} \left[ \left\| \hat{\mathcal{E}}_{k+1} \right\|^2 \right] + \eta^2 \mathbb{E} \left[ \left\| \bar{\xi}^{(k)} \right\|^2 \right] \\
& \quad + 2\mathbb{E} \left[ \left( \left\| \bar{x}^{(k)} - x_k \right\| + \eta \frac{1}{N} \left\| \nabla f(\bar{x}^{(k)}) - \nabla f(x_k) \right\| \right) \cdot \eta \left\| \hat{\mathcal{E}}_{k+1} \right\| \right] \\
& \leq \left( 1 - 2\eta\mu \left( 1 - \frac{\eta L}{2} \right) \right) \mathbb{E} \left[ \left\| \bar{x}^{(k)} - x_k \right\|^2 \right] + \eta^2 \mathbb{E} \left[ \left\| \hat{\mathcal{E}}_{k+1} \right\|^2 \right] + \eta^2 \mathbb{E} \left[ \left\| \bar{\xi}^{(k)} \right\|^2 \right] \\
& \quad + 2 \left( 1 + \eta \frac{L}{N} \right) \eta \mathbb{E} \left[ \left\| \bar{x}^{(k)} - x_k \right\| \cdot \left\| \hat{\mathcal{E}}_{k+1} \right\| \right] \\
& \leq \left( 1 - 2\eta\mu \left( 1 - \frac{\eta L}{2} \right) \right) \mathbb{E} \left[ \left\| \bar{x}^{(k)} - x_k \right\|^2 \right] + \eta^2 \mathbb{E} \left[ \left\| \hat{\mathcal{E}}_{k+1} \right\|^2 \right] + \eta^2 \mathbb{E} \left[ \left\| \bar{\xi}^{(k)} \right\|^2 \right] \\
& \quad + 2(1 + \eta L) \eta \mathbb{E} \left[ \left\| \bar{x}^{(k)} - x_k \right\| \cdot \left\| \hat{\mathcal{E}}_{k+1} \right\| \right] \\
& \leq \left( 1 - \eta\mu \left( 1 - \frac{\eta L}{2} \right) \right) \mathbb{E} \left[ \left\| \bar{x}^{(k)} - x_k \right\|^2 \right] + \eta \left( \eta + \frac{(1 + \eta L)^2}{\mu \left( 1 - \frac{\eta L}{2} \right)} \right) \mathbb{E} \left[ \left\| \hat{\mathcal{E}}_{k+1} \right\|^2 \right] + \eta^2 \frac{\sigma^2}{N}, \tag{A.16}
\end{aligned}$$

where we used the inequality  $2xy \leq c'x^2 + \frac{y^2}{c'}$  for any  $c' > 0$  and  $x, y \in \mathbb{R}$  where we took  $c' = \frac{\mu(1 - \frac{\eta L}{2})}{1 + \eta L}$ , and also the fact that since  $\eta < 2/L$ , we have  $\eta^2 \frac{L}{2} < \eta\mu$  so that  $\eta\mu \left( 1 - \frac{\eta L}{2} \right) \in (0, 1)$ . Now by using the bound in (A.13), we get

$$\begin{aligned}
& \mathbb{E} \left[ \left\| \bar{x}^{(k+1)} - x_{k+1} \right\|^2 \right] \\
& \leq \left( 1 - \eta\mu \left( 1 - \frac{\eta L}{2} \right) \right) \mathbb{E} \left[ \left\| \bar{x}^{(k)} - x_k \right\|^2 \right] + \eta^2 \left( 1 + \frac{(1 + \eta L)^2}{\eta\mu \left( 1 - \frac{\eta L}{2} \right)} \right) \mathbb{E} \left[ \left\| \hat{\mathcal{E}}_{k+1} \right\|^2 \right] + \eta^2 \frac{\sigma^2}{N} \\
& = \left( 1 - \eta\mu \left( 1 - \frac{\eta L}{2} \right) \right) \mathbb{E} \left[ \left\| \bar{x}^{(k)} - x_k \right\|^2 \right] \\
& \quad + \eta \left( \eta + \frac{(1 + \eta L)^2}{\mu \left( 1 - \frac{\eta L}{2} \right)} \right) \left[ \frac{4L^2 (\bar{\gamma}_{\tilde{w}})^{2k}}{N} \mathbb{E} \left[ \left\| x^{(0)} \right\|^2 \right] + \eta^2 \cdot \frac{4L^2}{N} \left( \frac{2(R_h + R'_h)}{(1 - \bar{\gamma}_{\tilde{w}})^2} + \frac{\sigma^2 N}{1 - \bar{\gamma}_{\tilde{w}}^2} \right) \right. \\
& \quad \left. + \eta \cdot \frac{8L^2 d}{1 - \bar{\gamma}_{\tilde{w}}^2} \right] + \eta^2 \frac{\sigma^2}{N}. \tag{A.17}
\end{aligned}$$

Since  $x^{(0)} = x_0$ , we finally get

$$\begin{aligned}
& \mathbb{E} \left[ \left\| \bar{x}^{(k)} - x_k \right\|^2 \right] \\
& \leq \sum_{i=0}^{k-1} \left( 1 - \eta \mu \left( 1 - \frac{\eta L}{2} \right) \right)^i \\
& \quad \cdot \left( \eta \left( \eta + \frac{(1+\eta L)^2}{\mu \left( 1 - \frac{\eta L}{2} \right)} \right) \left( \eta^2 \cdot \frac{4L^2}{N} \left( \frac{2(R_h + R'_h)}{(1-\bar{\gamma}_{\bar{w}})^2} + \frac{\sigma^2 N}{1-\bar{\gamma}_{\bar{w}}^2} \right) + \eta \cdot \frac{8L^2 d}{1-\bar{\gamma}_{\bar{w}}^2} \right) + \eta^2 \frac{\sigma^2}{N} \right) \\
& \quad + \sum_{i=0}^{k-1} \left( 1 - \eta \mu \left( 1 - \frac{\eta L}{2} \right) \right)^i \eta \left( \eta + \frac{(1+\eta L)^2}{\mu \left( 1 - \frac{\eta L}{2} \right)} \right) \frac{4L^2 (\bar{\gamma}_{\bar{w}})^{2(k-i)}}{N} \mathbb{E} \left[ \|x^{(0)}\|^2 \right] \\
& = \frac{1 - \left( 1 - \eta \mu \left( 1 - \frac{\eta L}{2} \right) \right)^k}{1 - \left( 1 - \eta \mu \left( 1 - \frac{\eta L}{2} \right) \right)} \\
& \quad \cdot \left( \eta \left( \eta + \frac{(1+\eta L)^2}{\mu \left( 1 - \frac{\eta L}{2} \right)} \right) \left( \eta^2 \cdot \frac{4L^2}{N} \left( \frac{2(R_h + R'_h)}{(1-\bar{\gamma}_{\bar{w}})^2} + \frac{\sigma^2 N}{1-\bar{\gamma}_{\bar{w}}^2} \right) + \eta \cdot \frac{8L^2 d}{1-\bar{\gamma}_{\bar{w}}^2} \right) + \eta^2 \frac{\sigma^2}{N} \right) \\
& \quad + \frac{\bar{\gamma}_{\bar{w}}^{2k} - \left( 1 - \eta \mu \left( 1 - \frac{\eta L}{2} \right) \right)^k}{1 - \left( 1 - \eta \mu \left( 1 - \frac{\eta L}{2} \right) \right) (\bar{\gamma}_{\bar{w}})^{-2}} \frac{4L^2}{N} \mathbb{E} \|x^{(0)}\|^2 \\
& \leq \frac{\eta \left( \eta + \frac{(1+\eta L)^2}{\mu \left( 1 - \frac{\eta L}{2} \right)} \right) \left( \eta^2 \cdot \frac{4L^2}{N} \left( \frac{2(R_h + R'_h)}{(1-\bar{\gamma}_{\bar{w}})^2} + \frac{\sigma^2 N}{1-\bar{\gamma}_{\bar{w}}^2} \right) + \eta \cdot \frac{8L^2 d}{1-\bar{\gamma}_{\bar{w}}^2} \right) + \eta^2 \frac{\sigma^2}{N}}{\eta \mu \left( 1 - \frac{\eta L}{2} \right)} \\
& \quad + \frac{\bar{\gamma}_{\bar{w}}^{2k} - \left( 1 - \eta \mu \left( 1 - \frac{\eta L}{2} \right) \right)^k}{(\bar{\gamma}_{\bar{w}})^2 - 1 + \eta \mu \left( 1 - \frac{\eta L}{2} \right)} \frac{4L^2 (\bar{\gamma}_{\bar{w}})^2}{N} \mathbb{E} \|x^{(0)}\|^2. \tag{A.18}
\end{aligned}$$

The proof is complete.

## B Proofs of Technical Lemmas

### B.1 Proof of Lemma 6

We can compute that

$$\begin{aligned}
& \max_{k=0,\dots,K-1} \mathbb{E} \left[ \left( \frac{1}{\delta^{k+1}} \|a^{(k+1)}\| \right)^2 \right] \\
&= \frac{1}{\delta^2} \max_{k=0,\dots,K-1} \mathbb{E} \left[ \left( \frac{1}{\delta^k} \|a^{(k+1)}\| \right)^2 \right] \\
&\leq \frac{c_1}{\delta^2} \max_{k=0,\dots,K-1} \mathbb{E} \left[ \left( \frac{1}{\delta^k} \|a^{(k)}\| \right)^2 \right] + \frac{c_2}{\delta^2} \max_{k=0,\dots,K-1} \mathbb{E} \left[ \left( \frac{1}{\delta^k} \|b^{(k)}\| \right)^2 \right] + \frac{c_0}{\delta^{2K}} \\
&= \frac{c_1}{\delta^2} \|\mathbf{a}\|_2^{\delta,K} + \frac{c_2}{\delta^2} \|\mathbf{b}\|_2^{\delta,K} + \frac{c_0}{\delta^{2K}},
\end{aligned} \tag{B.1}$$

where we used the simple inequality for maximum, that is,  $\max_k(x_k + y_k) \leq \max_k(x_k) + \max_k(y_k)$  for any real sequences  $(x_k), (y_k)$  to get (B.1) by using (6.6). Therefore, for any  $\delta \in (0, 1)$ , we have

$$\begin{aligned}
\max_{k=0,\dots,K} \mathbb{E} \left[ \frac{1}{\delta^k} \|a^{(k)}\|^2 \right] &= \max_{k=-1,\dots,K-1} \mathbb{E} \left[ \frac{1}{\delta^{k+1}} \|a^{(k+1)}\|^2 \right] \\
&\leq \max_{k=0,\dots,K-1} \mathbb{E} \left[ \frac{1}{\delta^{k+1}} \|a^{(k+1)}\|^2 \right] + \mathbb{E} \left[ \|a^{(0)}\|^2 \right] \\
&\leq \frac{c_1}{\delta^2} \|\mathbf{a}\|_2^{\delta,K} + \frac{c_2}{\delta^2} \|\mathbf{b}\|_2^{\delta,K} + \frac{c_0}{\delta^{2K}} + \mathbb{E} \left[ \|a^{(0)}\|^2 \right].
\end{aligned} \tag{B.2}$$

The proof is complete.

### B.2 Proof of Lemma 7

The proof follows from an adaption of the proof of Lemma 6 by using the assumption that  $\mathbb{E} \|a^{(k+1)}\|^2$  is bounded by the sum of finite components and applying the inequality that  $\max_k(x_{1_k} + x_{2_k} + \dots + x_{n_k}) \leq \max_k(x_{1_k}) + \max_k(x_{2_k}) + \dots + \max_k(x_{n_k})$  for any real sequences  $(x_{i_k})$ ,  $i = 1, 2, \dots, n$ .

### B.3 Proof of Lemma 8

Before proving the lemma, we give a preliminary result. With the fact in (B.4), Lemma 3.1 in [SLWY15] shows first-order optimality condition of EXTRA algorithm in decentralized optimization. Given mixing matrices  $W$  and  $\widetilde{W}$ , define  $U := \widetilde{W} - W$  by letting  $U^{1/2} := PD^{1/2}P^T \in \mathbb{R}^{N \times N}$ . Under Assumptions 1 and 3, then  $\mathbf{x}_*$  is consensual if and only if there exists  $\mathbf{q}_* = \mathcal{U}\mathbf{p}$  for some  $\mathbf{p} \in \mathbb{R}^{Nd}$  where  $\mathcal{U} := U \otimes I_d$  such that

$$\begin{cases} \mathcal{U}^{1/2}\mathbf{q}_* + \eta \nabla F(\mathbf{x}_*) = \mathbf{0}, \\ \mathcal{U}^{1/2}\mathbf{x}_* = \mathbf{0}. \end{cases} \tag{B.3}$$

According to Assumption 3 and by decomposing  $U^{1/2} = PD^{1/2}P^T$ , we get

$$\text{null} \left\{ U^{1/2} \right\} = \text{null} \left\{ P^T \right\} = \text{null} \left\{ \widetilde{W} - W \right\} = \text{span} \{ \mathbf{1}_N \}, \tag{B.4}$$

where  $\text{span}\{1_N\}$  is the span of the vector space supported by all-one vectors  $[1_N^T, 1_N^T, \dots, 1_N^T]$ , and it implies  $U^{1/2}$  is symmetric and  $1_N^T U^{1/2} = 0$ .

Now we deliver the proof as the following. We have the error such that  $e_x^{(k)} = x^{(k)} - \mathbf{x}_*$ , where  $\mathbf{x}_* = [x_*^T, \dots, x_*^T]^T$  is consensual. Thus, we can compute that

$$e_x^{(k)} = x^{(k)} - \bar{\mathbf{x}}^{(k)} + 1_N \otimes (\bar{x}^{(k)} - x_*) = \tilde{x}^{(k)} + 1_N \otimes \bar{e}_x^{(k)}, \quad (\text{B.5})$$

where  $\bar{x}^{(k)}$  is the mean of  $x^{(k)}$  in (6.11), and we used the definition of  $\tilde{x}^{(k)}$  in (6.13), and the definition of  $\bar{e}_x^{(k)}$  in (6.14) to obtain the last equality above.

Next, we notice that  $e_v^{(k)} = v^{(k)} + \nabla F(\mathbf{x}_*)$ . By Lemma B.3, we have

$$\mathbf{v}_* = \frac{1}{\eta} \mathcal{U}^{1/2} \mathbf{q}_* = -\nabla F(\mathbf{x}_*), \quad (\text{B.6})$$

where we have  $\mathbf{v}_* = -\nabla F(\mathbf{x}_*) = (\nabla f_1(x_*), \dots, \nabla f_N(x_*))^T$ , which is also consensual, and similarly, we can compute

$$e_v^{(k)} = v^{(k)} - \bar{\mathbf{v}}^{(k)} + 1_N \otimes (\bar{v}^{(k)} - v_*) = \tilde{v}^{(k)} + 1_N \otimes \bar{e}_v^{(k)}, \quad (\text{B.7})$$

where  $\bar{v}^{(k)}$  is the mean of  $v^{(k)}$  in (6.11) and we used the definition of  $\tilde{v}^{(k)}$  in (6.13), and the definition of  $\bar{e}_v^{(k)}$  in (6.14) to obtain the last equality in (B.7). Moreover, by (3.22), we can compute

$$v_i^{(k+1)} = v_i^{(k)} - \sum_{j \in \Omega_i} U_{ij} \left( v^{(k)} + \nabla F(x^{(k)}) - \mathcal{B}x^{(k)} \right)_j - \sum_{j \in \Omega_i} U_{ij} \xi_j^{(k)} + \sum_{j \in \Omega_i} U_{ij} \sqrt{\frac{2}{\eta}} w_j^{(k+1)}. \quad (\text{B.8})$$

Then by the definition,  $U = \widetilde{W} - W$  with doubly stochastic matrices  $\widetilde{W}, W$  such that  $\text{null}\{W - \widetilde{W}\} = \text{span}\{1_N\}$  under Assumption 3. Therefore, we have

$$\bar{v}^{(k+1)} = \bar{v}^{(k)} = \dots = \bar{v}^{(0)} = 0, \quad (\text{B.9})$$

where  $v^{(0)} = \frac{1}{\eta} \mathcal{U}^{1/2} q^{(0)} = 0$  from (3.17). Noticing  $\frac{1}{N} \sum_{i=1}^N \nabla f_i(x_*) = 0$  under optimality condition, we can get from (6.14) that

$$\bar{e}_v^{(k)} = \bar{v}^{(k)} = 0, \quad e_v^{(k)} = \tilde{v}^{(k)}. \quad (\text{B.10})$$

The proof is complete.

#### B.4 Proof of Lemma 9

From Lemma 8,  $\bar{v}^{(k)} = 0$ , such that we can compute the average iterates  $\bar{x}^{(k)}$  from (3.21) as follows

$$\bar{x}^{k+1} - x_* = \bar{x}^{(k)} - x_* - \frac{\eta}{N} \left( \sum_{i=1}^N \nabla f_i(\bar{x}^{(k)}) - \mathcal{E}^{(k)} \right) - \eta \bar{\xi}^{(k)} + \sqrt{2\eta} \bar{w}^{(k+1)}, \quad (\text{B.11})$$

where we recall that

$$\mathcal{E}^{(k)} := \sum_{i=1}^N \left( \nabla f_i(\bar{x}^{(k)}) - \nabla f_i(x_i^{(k)}) \right). \quad (\text{B.12})$$

Now by the optimality, we have  $1^T \nabla f(\mathbf{x}_*) = \sum_{i=1}^N \nabla f_i(x_*) = 0$  and  $\bar{e}_x^{(k)} = \frac{1}{N} \sum_{i=1}^N (x_i^{(k)} - x_*) = \bar{x}^{(k)} - x_* \in \mathbb{R}^d$ , we can compute

$$\begin{aligned} \left\| \bar{e}_x^{(k+1)} \right\|^2 &= \left\| \bar{e}_x^{(k)} - \frac{\eta}{N} \sum_{i=1}^N (\nabla f_i(\bar{x}^{(k)}) - \nabla f_i(x_*)) \right\|^2 + \left\| \frac{\eta}{N} \mathcal{E}^{(k)} - \eta \bar{\xi}^{(k)} + \sqrt{2\eta} \bar{w}^{(k+1)} \right\|^2 \\ &\quad + 2 \left\langle \bar{e}_x^{(k)} - \frac{\eta}{N} \sum_{i=1}^N (\nabla f_i(\bar{x}^{(k)}) - \nabla f_i(x_*)), \frac{\eta}{N} \mathcal{E}^{(k)} - \eta \bar{\xi}^{(k)} + \sqrt{2\eta} \bar{w}^{(k+1)} \right\rangle. \end{aligned} \quad (\text{B.13})$$

Next, we compute the first term on the right hand side of (B.13) as follows

$$\begin{aligned} &\left\| \bar{e}_x^{(k)} - \frac{\eta}{N} \sum_{i=1}^N (\nabla f_i(\bar{x}^{(k)}) - \nabla f_i(x_*)) \right\|^2 \\ &= \left\| \bar{e}_x^{(k)} \right\|^2 + \eta^2 \sum_{i=1}^N \left\| \frac{1}{N} (\nabla f_i(\bar{x}^{(k)}) - \nabla f_i(x_*)) \right\|^2 \\ &\quad - 2\eta \left\langle \bar{e}_x^{(k)}, \frac{1}{N} \sum_{i=1}^N (\nabla f_i(\bar{x}^{(k)}) - \nabla f_i(x_*)) \right\rangle \\ &\leq \left\| \bar{e}_x^{(k)} \right\|^2 - (2\eta - \eta^2 L) \left\langle \bar{e}_x^{(k)}, \frac{1}{N} \sum_{i=1}^N (\nabla f_i(\bar{x}^{(k)}) - \nabla f_i(x_*)) \right\rangle \\ &\leq \left( 1 - 2\eta\mu \left( 1 - \frac{\eta L}{2} \right) \right) \left\| \bar{e}_x^{(k)} \right\|^2, \end{aligned} \quad (\text{B.14})$$

where we used the condition  $\eta < 2/L$ , and used the assumption on  $L$ -smoothness of  $f_i$  to obtain the first term in the inequality of the second line above, and the assumption on  $\mu$ -strongly convexity of  $f_i$  to get the inequality of the last line.

By taking the expectations in (B.13), since  $\bar{\xi}^{(k)}$  and  $\bar{w}^{(k)}$  have mean zero conditional on the

natural filtration, and they are independent to  $\mathcal{E}^{(k)}$ , we can compute that

$$\begin{aligned}
& \mathbb{E} \left[ \left\| \bar{e}_x^{(k+1)} \right\|^2 \right] \\
& \leq \left( 1 - 2\eta\mu \left( 1 - \frac{\eta L}{2} \right) \right) \mathbb{E} \left[ \left\| \bar{e}_x^{(k)} \right\|^2 \right] + \frac{\eta^2}{N^2} \mathbb{E} \left[ \left\| \mathcal{E}^{(k)} \right\|^2 \right] + \eta^2 \mathbb{E} \left[ \left\| \bar{\xi}^{(k)} \right\|^2 \right] + 2\eta \mathbb{E} \left[ \left\| \bar{w}^{(k+1)} \right\|^2 \right] \\
& \quad + \frac{2\eta}{N} \mathbb{E} \left\langle \bar{e}_x^{(k)}, \mathcal{E}^{(k)} \right\rangle + \frac{2\eta^2}{N^2} \mathbb{E} \left\langle \sum_{i=1}^N \left( \nabla f_i(x_*) - \nabla f_i(\bar{x}^{(k)}) \right), \mathcal{E}^{(k)} \right\rangle \\
& \leq \left( 1 - 2\eta\mu \left( 1 - \frac{\eta L}{2} \right) \right) \mathbb{E} \left[ \left\| \bar{e}_x^{(k)} \right\|^2 \right] + \frac{\eta^2}{N^2} \mathbb{E} \left[ \left\| \mathcal{E}^{(k)} \right\|^2 \right] + \eta^2 \frac{\sigma^2}{N} + 2\eta \frac{d}{N} \\
& \quad + \frac{2\eta}{N} \mathbb{E} \left[ \left\| \bar{e}_x^{(k)} \right\| \cdot \left\| \mathcal{E}^{(k)} \right\| \right] + \frac{2\eta^2}{N^2} \mathbb{E} \left[ \left\| \sum_{i=1}^N \left( \nabla f_i(x_*) - \nabla f_i(\bar{x}^{(k)}) \right) \right\| \cdot \left\| \mathcal{E}^{(k)} \right\| \right] \quad (\text{B.15}) \\
& \leq \left( 1 - 2\eta\mu \left( 1 - \frac{\eta L}{2} \right) \right) \mathbb{E} \left[ \left\| \bar{e}_x^{(k)} \right\|^2 \right] + \frac{\eta^2}{N^2} \mathbb{E} \left[ \left\| \mathcal{E}^{(k)} \right\|^2 \right] + \eta^2 \frac{\sigma^2}{N} + 2\eta \frac{d}{N} \\
& \quad + \frac{2\eta}{N} (1 + \eta L) \mathbb{E} \left[ \left\| \bar{e}_x^{(k)} \right\| \cdot \left\| \mathcal{E}^{(k)} \right\| \right] \\
& \leq \left( 1 - \eta\mu \left( 1 - \frac{\eta L}{2} \right) \right) \mathbb{E} \left[ \left\| \bar{e}_x^{(k)} \right\|^2 \right] + \frac{\eta}{N} \left( \frac{\eta}{N} + \frac{1 + \eta L}{N\mu \left( 1 - \frac{\eta L}{2} \right)} \right) \mathbb{E} \left[ \left\| \mathcal{E}^{(k)} \right\|^2 \right] + \eta^2 \frac{\sigma^2}{N} + 2\eta \frac{d}{N}, \quad (\text{B.16})
\end{aligned}$$

where we used  $1 - \eta\mu \left( 1 - \frac{\eta L}{2} \right) \in (0, 1)$ . Note that we used Cauchy-Schwarz inequality to get (B.15) and used the inequality  $2xy \leq \frac{x^2}{c'} + c'y^2$  for  $c > 0$  to get (B.16) by taking  $c' = \eta\mu \left( 1 - \frac{\eta L}{2} \right)$ . Moreover, we can compute from (B.12) to get

$$\frac{\eta}{N} \left\| \mathcal{E}^{(k)} \right\|^2 \leq \frac{\eta}{N} \sum_{i=1}^N \left\| \nabla f_i(\bar{x}^{(k)}) - \nabla f_i(x_i^{(k)}) \right\|^2 \leq \frac{\eta L^2}{N} \left\| \tilde{x}^{(k)} \right\|^2, \quad (\text{B.17})$$

where we used  $\sum_{i=1}^N \left\| \bar{x}^{(k)} - x_i^{(k)} \right\|^2 = \sum_{i=1}^N \left\| \tilde{x}_i^{(k)} \right\|^2 = \|\tilde{x}\|^2$ . Hence, we get

$$\begin{aligned}
\mathbb{E} \left[ \left\| \bar{e}_x^{(k+1)} \right\|^2 \right] & \leq \left( 1 - \eta\mu \left( 1 - \frac{\eta L}{2} \right) \right) \mathbb{E} \left[ \left\| \bar{e}_x^{(k)} \right\|^2 \right] \\
& \quad + \frac{\eta L^2}{N} \left( \frac{\eta}{N} + \frac{1 + \eta L}{N\mu \left( 1 - \frac{\eta L}{2} \right)} \right) \mathbb{E} \left[ \left\| \tilde{x}^{(k)} \right\|^2 \right] + \eta^2 \frac{\sigma^2}{N} + 2\eta \frac{d}{N}. \quad (\text{B.18})
\end{aligned}$$

By Lemma 6, we get

$$\begin{aligned}
\|\bar{e}_x\|_2^{\delta, K} & \leq \frac{1}{\delta^2} \left( 1 - \eta\mu \left( 1 - \frac{\eta L}{2} \right) \right) \|\bar{e}_x\|_2^{\delta, K} \\
& \quad + \frac{\eta L^2}{\delta^2 N} \left( \frac{\eta}{N} + \frac{1 + \eta L}{N\mu \left( 1 - \frac{\eta L}{2} \right)} \right) \|\tilde{x}\|_2^{\delta, K} + \frac{\eta}{N\delta^{2K}} (\eta\sigma^2 + 2d) + \mathbb{E} \left[ \left\| \bar{e}_x^{(0)} \right\|^2 \right]. \quad (\text{B.19})
\end{aligned}$$

By taking  $\delta \in \left( \sqrt{1 - \eta\mu \left( 1 - \frac{\eta L}{2} \right)}, 1 \right)$ , we can compute from (B.19), it follows

$$\begin{aligned} & \left( \delta^2 + \eta\mu \left( 1 - \frac{\eta L}{2} \right) - 1 \right) \|\bar{\mathbf{e}}_x\|_2^{\delta, K} \\ & \leq \frac{\eta L^2}{N} \left( \frac{\eta}{N} + \frac{1 + \eta L}{N\mu \left( 1 - \frac{\eta L}{2} \right)} \right) \|\tilde{\mathbf{x}}\|_2^{\delta, K} + \frac{\eta(\eta\sigma^2 + 2d)}{N\delta^{2K-2}} + \delta^2 \mathbb{E} \left[ \|\bar{e}_x^{(0)}\|^2 \right]. \end{aligned} \quad (\text{B.20})$$

Therefore, we conclude that for every  $K \geq 0$ ,

$$\begin{aligned} \|\bar{\mathbf{e}}_x\|_2^{\delta, K} & \leq \eta \cdot \frac{L^2}{N^2 \left( \delta^2 + \eta\mu \left( 1 - \frac{\eta L}{2} \right) - 1 \right)} \left( \eta + \frac{1 + \eta L}{\mu \left( 1 - \frac{\eta L}{2} \right)} \right) \|\tilde{\mathbf{x}}\|_2^{\delta, K} \\ & + \frac{\eta}{N\delta^{2K-2}} \cdot \frac{\eta\sigma^2 + 2d}{\delta^2 + \eta\mu \left( 1 - \frac{\eta L}{2} \right) - 1} + \frac{\delta^2}{\delta^2 + \eta\mu \left( 1 - \frac{\eta L}{2} \right) - 1} \mathbb{E} \left[ \|\bar{e}_x^{(0)}\|^2 \right]. \end{aligned} \quad (\text{B.21})$$

This completes the proof.

## B.5 Proof of Lemma 10

Following (3.21), we can compute that

$$x^{(k+1)} - \mathbf{x}_* = \widetilde{\mathcal{W}}x^{(k)} - \mathbf{x}_* - \eta \left( \nabla F(x^{(k)}) + v^{(k)} + \nabla F(\mathbf{x}_*) - \nabla F(\mathbf{x}_*) \right) - \eta\xi^{(k)} + \sqrt{2\eta}w^{(k+1)}. \quad (\text{B.22})$$

Next, by using  $\widetilde{\mathcal{W}}\mathbf{x}_* = \mathbf{x}_*$  and (6.16) in Lemma 8, we get

$$e_x^{(k+1)} = \widetilde{\mathcal{W}}e_x^{(k)} - \eta \left( \nabla F(x^{(k)}) - \nabla F(\mathbf{x}_*) \right) - \eta\tilde{v}^{(k)} - \eta\xi^{(k)} + \sqrt{2\eta}w^{(k+1)}. \quad (\text{B.23})$$

Moreover, we use the definition of  $e_x^{(k)}$  in (6.15) from Lemma 8 and  $\mathcal{J} = J \otimes I_d$  with  $J = \frac{1}{N}\mathbf{1}_N\mathbf{1}_N^T$  to get

$$\tilde{x}^{(k)} = e_x^{(k)} - \mathbf{1}_N \otimes \bar{e}_x^{(k)} = [(I_N - J) \otimes I_d] e_x^{(k)}, \quad (\text{B.24})$$

where the component on  $(i, j)$ -position of  $[(I_N - J) \otimes I_d] e_x^{(k)}$  is  $[e_x^{(k)}]_{ij} - \frac{1}{N} \sum_{i=1}^N [e_x^{(k)}]_{ij}$  for  $i = 1, \dots, N$ . Hence, we obtained the last equality in (B.24) which can be re-written as:

$$\tilde{x}^{(k)} = (I_{Nd} - \mathcal{J}) e_x^{(k)}. \quad (\text{B.25})$$

Next, by multiplying  $(I_{Nd} - \mathcal{J})$  on both hand sides of (B.23), we get

$$\begin{aligned} \tilde{x}^{(k+1)} & = (I_{Nd} - \mathcal{J}) \widetilde{\mathcal{W}}\tilde{x}^{(k)} - \eta(I_{Nd} - \mathcal{J}) \left( \nabla F(x^{(k)}) - \nabla F(\mathbf{x}_*) \right) \\ & \quad - \eta(I_{Nd} - \mathcal{J}) \tilde{v}^{(k)} - \eta(I_{Nd} - \mathcal{J}) \xi^{(k)} + \sqrt{2\eta}(I_{Nd} - \mathcal{J}) w^{(k+1)} \\ & = (\widetilde{\mathcal{W}} - \mathcal{J}) \tilde{x}^{(k)} - \eta(I_{Nd} - \mathcal{J}) \left( \nabla F(x^{(k)}) - \nabla F(\mathbf{x}_*) \right) \\ & \quad - \eta\tilde{v}^{(k)} - \eta \left( \xi^{(k)} - \bar{\xi}^{(k)} \right) + \sqrt{2\eta} \left( w^{(k+1)} - \bar{w}^{(k+1)} \right), \end{aligned} \quad (\text{B.26})$$

where we used  $(I_N - J)W = W - JW = W - J$  and  $\mathcal{J}\tilde{v}^{(k)} = \mathcal{J}e_v^{(k)} = \bar{e}_v^{(k)} \otimes 1_N^T = \mathbf{0}$  in Lemma 8 to get the last equality. In the following, we can compute that

$$\begin{aligned} \|\tilde{x}^{(k+1)}\|^2 &= \left\| \left( \widetilde{\mathcal{W}} - \mathcal{J} \right) \tilde{x}^{(k)} - \eta (I_{Nd} - \mathcal{J}) \left( \nabla F(x^{(k)}) - \nabla F(\mathbf{x}_*) \right) \right\|^2 \\ &\quad + 2 \left\langle \left( \widetilde{\mathcal{W}} - \mathcal{J} \right) \tilde{x}^{(k)} - \eta (I_{Nd} - \mathcal{J}) \left( \nabla F(x^{(k)}) - \nabla F(\mathbf{x}_*) \right), \right. \\ &\quad \left. - \eta \tilde{v}^{(k)} - \eta \left( \xi^{(k)} - \bar{\xi}^{(k)} \right) + \sqrt{2\eta} \left( w^{(k+1)} - \bar{w}^{(k+1)} \right) \right\rangle \\ &\quad + \left\| -\eta \tilde{v}^{(k)} - \eta \left( \xi^{(k)} - \bar{\xi}^{(k)} \right) + \sqrt{2\eta} \left( w^{(k+1)} - \bar{w}^{(k+1)} \right) \right\|^2. \end{aligned} \quad (\text{B.27})$$

By the fact that  $\mathcal{J} = J \otimes I_d$ , we have  $\lambda_{\max}(J) = 1/N$  and  $\lambda_{\min}(J) = 0$ . We can further compute

$$\left\| (I_{Nd} - \mathcal{J}) \left( \nabla F(x^{(k)}) - \nabla F(\mathbf{x}_*) \right) \right\|^2 \leq \left\| \nabla F(x^{(k)}) - \nabla F(\mathbf{x}_*) \right\|^2, \quad (\text{B.28})$$

where we used  $\lambda_{\max}^2(I_N - J) = (1 - \lambda_{\min}(J))^2$  since  $\lambda(I_N - A) = 1 - \lambda(A)$  for any  $N \times N$  matrix  $A$ .

Next, we can compute the first term in (B.27) as below.

$$\begin{aligned} &\left\| \left( \widetilde{\mathcal{W}} - \mathcal{J} \right) \tilde{x}^{(k)} - \eta (I_{Nd} - \mathcal{J}) \left( \nabla F(x^{(k)}) - \nabla F(\mathbf{x}_*) \right) \right\|^2 \\ &= |\lambda_2^{\bar{w}}|^2 \left\| \tilde{x}^{(k)} \right\|^2 + \eta^2 \left\| (I_{Nd} - \mathcal{J}) \left( \nabla F(x^{(k)}) - \nabla F(\mathbf{x}_*) \right) \right\|^2 \\ &\quad - 2\eta \left\langle \left( \widetilde{\mathcal{W}} - \mathcal{J} \right) \tilde{x}^{(k)}, (I_{Nd} - \mathcal{J}) \left( \nabla F(x^{(k)}) - \nabla F(\mathbf{x}_*) \right) \right\rangle \end{aligned} \quad (\text{B.29})$$

$$\begin{aligned} &\leq |\lambda_2^{\bar{w}}|^2 \left\| \tilde{x}^{(k)} \right\|^2 + \eta^2 \left\| \nabla F(x^{(k)}) - \nabla F(\mathbf{x}_*) \right\|^2 - 2\eta \left\langle e_x^{(k)}, \nabla F(x^{(k)}) - \nabla F(\mathbf{x}_*) \right\rangle \\ &\quad - 2\eta \left\langle \left( \widetilde{\mathcal{W}} - \mathcal{J} \right) (I_{Nd} - \mathcal{J})^2 - I_{Nd}, e_x^{(k)} \right\rangle \nabla F(x^{(k)}) - \nabla F(\mathbf{x}_*) \\ &\leq |\lambda_2^{\bar{w}}|^2 \left\| \tilde{x}^{(k)} \right\|^2 + \eta^2 \left\| \nabla F(x^{(k)}) - \nabla F(\mathbf{x}_*) \right\|^2 - 2\eta \left\langle e_x^{(k)}, \nabla F(x^{(k)}) - \nabla F(\mathbf{x}_*) \right\rangle \\ &\quad + \eta \left( \frac{1}{L} \left\| \nabla F(x^{(k)}) - \nabla F(\mathbf{x}_*) \right\|^2 + L \left\| e_x^{(k)} \right\|^2 \right) \end{aligned} \quad (\text{B.30})$$

$$\begin{aligned} &= |\lambda_2^{\bar{w}}|^2 \left\| \tilde{x}^{(k)} \right\|^2 + \left( \eta^2 + \frac{\eta}{L} \right) \left\| \nabla F(x^{(k)}) - \nabla F(\mathbf{x}_*) \right\|^2 \\ &\quad - 2\eta \left\langle e_x^{(k)}, \nabla F(x^{(k)}) - \nabla F(\mathbf{x}_*) \right\rangle + \eta L \left\| e_x^{(k)} \right\|^2 \\ &\leq |\lambda_2^{\bar{w}}|^2 \left\| \tilde{x}^{(k)} \right\|^2 - (\eta - \eta^2 L) \left\langle e_x^{(k)}, \nabla F(x^{(k)}) - \nabla F(\mathbf{x}_*) \right\rangle + \eta L \left\| e_x^{(k)} \right\|^2 \end{aligned} \quad (\text{B.31})$$

$$\leq |\lambda_2^{\bar{w}}|^2 \left\| \tilde{x}^{(k)} \right\|^2 - (\eta - \eta^2 L) \mu \left\| e_x^{(k)} \right\|^2 + \eta L \left\| e_x^{(k)} \right\|^2 \quad (\text{B.32})$$

$$\begin{aligned} &= |\lambda_2^{\bar{w}}|^2 \left\| \tilde{x}^{(k)} \right\|^2 + \eta \mu (\eta L + (L/\mu) - 1) \left\| e_x^{(k)} \right\|^2 \\ &= \left( |\lambda_2^{\bar{w}}|^2 + \eta \mu (\eta L + (L/\mu) - 1) \right) \left\| \tilde{x}^{(k)} \right\|^2 + \eta \mu (\eta L + (L/\mu) - 1) \left\| \bar{e}_x^{(k)} \right\|^2, \end{aligned} \quad (\text{B.33})$$

where we used the inequality that

$$\left\| \left( \left( \widetilde{\mathcal{W}} - J \right) \otimes I_d \right) \tilde{x}^{(k)} \right\|^2 \leq \left\| \widetilde{\mathcal{W}} - J \right\|^2 \left\| \tilde{x}^{(k)} \right\|^2 \leq |\lambda_2^{\bar{w}}|^2 \left\| \tilde{x}^{(k)} \right\|^2,$$

where  $1 = \lambda_1^{\widetilde{W}} > \lambda_2^{\widetilde{W}} \geq \dots \geq \lambda_N^{\widetilde{W}} > 0$  to get (B.29). To get (B.30), we used matrix property of  $J$ ,  $J = \frac{1}{N} \mathbf{1}_N \mathbf{1}_N^T$  is symmetric and  $(\mathbf{1}_N \mathbf{1}_N^T)A = A(\mathbf{1}_N \mathbf{1}_N^T)$ , so that  $(I_N - J)^T A = A(I_N - J)$ . We notice  $(\widetilde{W} - J)J = J$ , and thus

$$(\widetilde{W} - J)(I_N - J)^2 = (\widetilde{W} - 2J)(I_N - J) = \widetilde{W} - J,$$

where we used the fact that  $J^2 = J$ . Moreover, we can also compute

$$\left\| \left( (\widetilde{W} - J - I_N) \otimes I_d \right) e_x^{(k)} \right\|^2 \leq \left\| \widetilde{W} - I_N - J \right\|^2 \left\| e_x^{(k)} \right\|^2 \leq \left\| e_x^{(k)} \right\|^2,$$

where we have  $\lambda(\widetilde{W} - I_N) = \lambda(\widetilde{W}) - 1 \in (-1, 0]$ , and since  $J = \frac{1}{N} \mathbf{1}_N^T \mathbf{1}_N$ , we can decompose  $\widetilde{W} - I_N - J$  and  $\widetilde{W} - I_N$  in the same eigenspace, so that they have the same non-zero eigenvalues, and the largest eigenvalue  $\lambda_{\max}(\widetilde{W} - I_N - J) = 1$  corresponds to eigenvector  $\mathbf{1}_N^T$ . Hence, we obtain

$$\begin{aligned} & -2\eta \left\langle \left( (\widetilde{W} - \mathcal{J})(I_{Nd} - \mathcal{J})^2 - I_{Nd} \right) e_x^{(k)}, \nabla F(x^{(k)}) - \nabla F(\mathbf{x}_*) \right\rangle \\ & \leq 2\eta \left\| \left( (\widetilde{W} - \mathcal{J})(I_{Nd} - \mathcal{J})^2 - I_{Nd} \right) e_x^{(k)} \right\| \cdot \left\| \nabla F(x^{(k)}) - \nabla F(\mathbf{x}_*) \right\| \\ & = 2\eta \left\| (\widetilde{W} - \mathcal{J} - I_{Nd}) e_x^{(k)} \right\| \cdot \left\| \nabla F(x^{(k)}) - \nabla F(\mathbf{x}_*) \right\| \\ & \leq \eta \left( \frac{1}{L} \left\| \nabla F(x^{(k)}) - \nabla F(\mathbf{x}_*) \right\|^2 + L \left\| e_x^{(k)} \right\|^2 \right), \end{aligned} \quad (\text{B.34})$$

where we used Cauchy-Schwarz inequality for the inner product and the inequality  $2xy \leq cx^2 + \frac{y^2}{c}$  for any  $x, y \in \mathbb{R}$  by taking  $c = L > 0$ . Next, we used  $L$ -smoothness of  $F$  to get (B.31) and  $\mu$ -convexity of  $F$  to get (B.32) where we also used  $\eta \leq 1/L$ . Finally, we used triangle inequality to  $\tilde{x}^{(k)} + \mathbf{1}_N \otimes \bar{e}_x^{(k)} = e_x^{(k)}$  and get  $\left\| e_x^{(k)} \right\|^2 \leq \left\| \tilde{x}^{(k)} \right\|^2 + \left\| \bar{e}_x^{(k)} \right\|^2$ , for (B.33).

Now by using the fact that the expectation of random noise terms and their average terms are zero conditioning on the natural filtration to compute the expectation of the inner product term in (B.27), we get

$$\begin{aligned} & \mathbb{E} \left\langle (\widetilde{W} - \mathcal{J}) \tilde{x}^{(k)} - \eta (I_{Nd} - \mathcal{J}) (\nabla F(x^{(k)}) - \nabla F(\mathbf{x}_*)) , \eta \tilde{v}^{(k)} \right\rangle \\ & = \eta \mathbb{E} \left\langle \widetilde{W} \tilde{x}^{(k)}, \tilde{v}^{(k)} \right\rangle + \eta^2 \mathbb{E} \left\langle (I_{Nd} - \mathcal{J}) (\nabla F(\mathbf{x}_*) - \nabla F(x^{(k)})) , \tilde{v}^{(k)} \right\rangle \end{aligned} \quad (\text{B.35})$$

$$\begin{aligned} & \leq \eta \mathbb{E} \left[ \left\| \widetilde{W} \tilde{x}^{(k)} \right\| \cdot \left\| \tilde{v}^{(k)} \right\| \right] + \eta^2 \mathbb{E} \left[ \left\| (I_{Nd} - \mathcal{J}) (\nabla F(\mathbf{x}_*) - \nabla F(x^{(k)})) \right\| \cdot \left\| \tilde{v}^{(k)} \right\| \right] \\ & \leq \eta \mathbb{E} \left[ \left\| \tilde{x}^{(k)} \right\| \cdot \left\| \tilde{v}^{(k)} \right\| \right] + \eta^2 \mathbb{E} \left[ \left\| \nabla F(\mathbf{x}_*) - \nabla F(x^{(k)}) \right\| \cdot \left\| \tilde{v}^{(k)} \right\| \right] \end{aligned} \quad (\text{B.36})$$

$$\begin{aligned} & \leq \eta L \mathbb{E} \left[ \left\| \tilde{x}^{(k)} \right\|^2 \right] + \frac{\eta}{4L} \mathbb{E} \left[ \left\| \tilde{v}^{(k)} \right\|^2 \right] \end{aligned} \quad (\text{B.37})$$

$$+ \eta^2 L \mu \mathbb{E} \left[ \left\| e_x^{(k)} \right\|^2 \right] + \frac{\eta^2}{4L\mu} \mathbb{E} \left[ \left\| \tilde{v}^{(k)} \right\|^2 \right] \quad (\text{B.38})$$

$$\leq (\eta L + \eta^2 L \mu) \mathbb{E} \left[ \left\| \tilde{x}^{(k)} \right\|^2 \right] + \eta^2 \mu L \mathbb{E} \left[ \left\| \bar{e}_x^{(k)} \right\|^2 \right] + \left( \frac{\eta}{4L} + \frac{\eta^2}{4L\mu} \right) \mathbb{E} \left[ \left\| \tilde{v}^{(k)} \right\|^2 \right]. \quad (\text{B.39})$$

To get (B.35), we used Cauchy-Schwarz inequality and the fact that  $\langle \mathcal{J}\tilde{x}^{(k)}, \tilde{v}^{(k)} \rangle = \langle \tilde{x}^{(k)}, \mathcal{J}^T \tilde{v}^{(k)} \rangle$ , where  $\mathcal{J}^T \tilde{v}^{(k)} = \mathcal{J}\tilde{v}^{(k)} = \mathbf{0}$  by Lemma 8. Recall we have the eigenvalues of  $\widetilde{W}$  are  $1 = \lambda_1^{\widetilde{W}} > \lambda_2^{\widetilde{W}} \geq \dots \geq \lambda_N^{\widetilde{W}} > 0$  to have  $\left\| (\widetilde{W} \otimes I_d) \tilde{x}^{(k)} \right\|^2 \leq \|\tilde{x}^{(k)}\|^2$ , which yields (B.36). Then we used the inequality  $2xy \leq c'x^2 + \frac{y^2}{c'}$  for  $c' > 0$  and took  $c' = 2L$  to get (B.37). Next, we used the same inequality by taking  $c' = \frac{2}{L}$  and using  $L$ -smoothness of  $F$  to get (B.38). Finally, we used triangle inequality to  $\tilde{x}^{(k)} + 1_N \otimes \bar{e}_x^{(k)} = e_x^{(k)}$  and get  $\|e_x^{(k)}\|^2 \leq \|\tilde{x}^{(k)}\|^2 + \|\bar{e}_x^{(k)}\|^2$ , and choose  $\eta \leq 1/L$  to obtain the last inequality (B.39).

Accordingly, we can get the following inequality from (B.27) and (B.33),

$$\begin{aligned} \mathbb{E} \left[ \left\| \tilde{x}^{(k+1)} \right\|^2 \right] &\leq \left( |\lambda_2^{\widetilde{W}}|^2 + \eta\mu (3(L/\mu) + 3\eta L - 1) \right) \mathbb{E} \left[ \left\| \tilde{x}^{(k)} \right\|^2 \right] \\ &\quad + \eta\mu (L/\mu + 3\eta L - 1) \mathbb{E} \left[ \left\| \bar{e}_x^{(k)} \right\|^2 \right] + \eta \left( \frac{1}{2L} + \eta + \frac{\eta}{2L\mu} \right) \mathbb{E} \left[ \left\| \tilde{v}^{(k)} \right\|^2 \right] \\ &\quad + \eta^2 \mathbb{E} \left[ \left\| \xi^{(k)} - \bar{\xi}^{(k)} \right\|^2 \right] + 2\eta \mathbb{E} \left[ \left\| w^{(k+1)} - \bar{w}^{(k+1)} \right\|^2 \right], \end{aligned} \quad (\text{B.40})$$

where we can further compute

$$\begin{aligned} &\eta^2 \mathbb{E} \left[ \left\| \xi^{(k)} - \bar{\xi}^{(k)} \right\|^2 \right] + 2\eta \mathbb{E} \left[ \left\| w^{(k+1)} - \bar{w}^{(k+1)} \right\|^2 \right] \\ &= \eta^2 \mathbb{E} \left[ \left\| \xi^{(k)} \right\|^2 \right] + \eta^2 \mathbb{E} \left[ \left\| \bar{\xi}^{(k)} \right\|^2 \right] + 2\eta \mathbb{E} \left[ \left\| w^{(k+1)} \right\|^2 \right] + 2\eta \mathbb{E} \left[ \left\| \bar{w}^{(k+1)} \right\|^2 \right] \\ &= \eta \left( N + \frac{1}{N} \right) (\eta\sigma^2 + 2d). \end{aligned} \quad (\text{B.41})$$

Under the assumption of the stepsize in Lemma 9, Lemma 6, in particular, Lemma 7 and the inequality (B.40), it follows that

$$\begin{aligned} \delta^2 \|\tilde{\mathbf{x}}\|_2^{\delta,K} &\leq \left( |\lambda_2^{\widetilde{W}}|^2 + \eta\mu (3(L/\mu) + 3\eta L - 1) \right) \|\tilde{\mathbf{x}}\|_2^{\delta,K} \\ &\quad + \eta\mu (L/\mu + 3\eta L - 1) \|\bar{e}_x\|_2^{\delta,K} + \eta \left( \frac{1}{2L} + \eta + \frac{\eta}{2L\mu} \right) \|\tilde{\mathbf{v}}\|_2^{\delta,K} \\ &\quad + \frac{\eta}{\delta^{2K-2}} \left( N + \frac{1}{N} \right) (\eta\sigma^2 + 2d) + \delta^2 \mathbb{E} \left[ \left\| \tilde{x}^{(0)} \right\|^2 \right]. \end{aligned} \quad (\text{B.42})$$

Next, we consider the following three scenarios.

(1). If  $|\lambda_2^{\widetilde{W}}|^2 < \frac{1}{2}$ , then for all  $1 > \delta^2 \geq 2|\lambda_2^{\widetilde{W}}|^2$ , we compute that

$$\begin{aligned} \delta^2 - |\lambda_2^{\widetilde{W}}|^2 - \eta\mu (3(L/\mu) + 3\eta L - 1) &\geq |\lambda_2^{\widetilde{W}}|^2 - 3L\eta - 3\eta^2\mu L + \eta\mu \\ &\geq |\lambda_2^{\widetilde{W}}|^2 - 3(L + \mu)\eta, \end{aligned} \quad (\text{B.43})$$

where we used the assumption  $\eta L \leq 1$  to get  $3\eta^2\mu L \leq 3\mu\eta$  and obtain the last inequality. Thus, by the assumption that

$$\eta \leq \frac{|\lambda_2^{\widetilde{W}}|^2}{6(L + \mu)}, \quad (\text{B.44})$$

we obtain

$$\delta^2 - |\lambda_2^{\bar{W}}|^2 - \eta\mu(3(L/\mu) + 3\eta L - 1) > \frac{|\lambda_2^{\bar{W}}|^2}{2} > 0, \quad (\text{B.45})$$

(2). If  $\frac{2}{3} \geq |\lambda_2^{\bar{W}}|^2 \geq \frac{1}{2}$ , for all  $1 > \delta^2 \geq \frac{|\lambda_2^{\bar{W}}|^2}{2(1-|\lambda_2^{\bar{W}}|^2)} \geq \frac{1}{2}$ . We can compute that

$$\delta^2 - |\lambda_2^{\bar{W}}|^2 \geq \frac{|\lambda_2^{\bar{W}}|^2}{2(1-|\lambda_2^{\bar{W}}|^2)} - |\lambda_2^{\bar{W}}|^2 = \frac{|\lambda_2^{\bar{W}}|^2 \left( |\lambda_2^{\bar{W}}|^2 - \frac{1}{2} \right)}{1-|\lambda_2^{\bar{W}}|^2} > 0, \quad (\text{B.46})$$

and under the assumption that

$$\eta \leq \frac{|\lambda_2^{\bar{W}}|^2 \left( |\lambda_2^{\bar{W}}|^2 - \frac{1}{2} \right)}{6(L+\mu)(1-|\lambda_2^{\bar{W}}|^2)}, \quad (\text{B.47})$$

we can compute that

$$\begin{aligned} \delta^2 - |\lambda_2^{\bar{W}}|^2 - \eta\mu(3(L/\mu) + 3\eta L - 1) &\geq \delta^2 - |\lambda_2^{\bar{W}}|^2 - 3(L+\mu)\eta \\ &\geq \frac{|\lambda_2^{\bar{W}}|^2 \left( |\lambda_2^{\bar{W}}|^2 - \frac{1}{2} \right)}{1-|\lambda_2^{\bar{W}}|^2} - 3(L+\mu)\eta \\ &\geq \frac{|\lambda_2^{\bar{W}}|^2 \left( |\lambda_2^{\bar{W}}|^2 - \frac{1}{2} \right)}{2(1-|\lambda_2^{\bar{W}}|^2)} > 0. \end{aligned} \quad (\text{B.48})$$

(3). If  $1 > |\lambda_2^{\bar{W}}|^2 > \frac{2}{3}$ , we can easily find the quantity relation  $1 > \frac{4|\lambda_2^{\bar{W}}|^2-2}{3|\lambda_2^{\bar{W}}|^2-1} > \frac{1}{2}$ , then for all  $1 > \delta^2 \geq \frac{4|\lambda_2^{\bar{W}}|^2-2}{3|\lambda_2^{\bar{W}}|^2-1} > \frac{1}{2}$ , we can compute that

$$\delta^2 - |\lambda_2^{\bar{W}}|^2 \geq \frac{4|\lambda_2^{\bar{W}}|^2-2}{3|\lambda_2^{\bar{W}}|^2-1} - |\lambda_2^{\bar{W}}|^2 = \frac{5|\lambda_2^{\bar{W}}|^2 - 3|\lambda_2^{\bar{W}}|^4 - 2}{3|\lambda_2^{\bar{W}}|^2-1} > 0, \quad (\text{B.49})$$

where we can find  $2/3, 1$  are two roots for  $5x - 3x^2 - 2 = 0$ , such that  $\frac{5x-3x^2-1}{3x-1} > 0$  for any  $1 > x > \frac{2}{3}$  which implies (B.49). Under the assumption on  $\eta$  such that

$$\eta \leq \frac{5|\lambda_2^{\bar{W}}|^2 - 3|\lambda_2^{\bar{W}}|^4 - 2}{6(L+\mu)(3|\lambda_2^{\bar{W}}|^2-1)}, \quad (\text{B.50})$$

we can further compute

$$\begin{aligned} \delta^2 - |\lambda_2^{\bar{W}}|^2 - \eta\mu(3(L/\mu) + 3\eta L - 1) &\geq \delta^2 - |\lambda_2^{\bar{W}}|^2 - 3(L+\mu)\eta \\ &\geq \frac{5|\lambda_2^{\bar{W}}|^2 - 3|\lambda_2^{\bar{W}}|^4 - 2}{3|\lambda_2^{\bar{W}}|^2-1} - 3(L+\mu)\eta \\ &\geq \frac{1}{2} \cdot \frac{5|\lambda_2^{\bar{W}}|^2 - 3|\lambda_2^{\bar{W}}|^4 - 2}{3|\lambda_2^{\bar{W}}|^2-1} > 0. \end{aligned} \quad (\text{B.51})$$

Moreover, we can find from (B.44), (B.47) and (B.50) such that

$$\eta \leq \frac{\gamma_{\bar{w}}}{6(L + \mu)}, \quad (\text{B.52})$$

where we define the constant

$$\gamma_{\bar{w}} := \begin{cases} |\lambda_2^{\bar{w}}|^2 & \text{if } 0 < |\lambda_2^{\bar{w}}|^2 < \frac{1}{2}, \\ \frac{|\lambda_2^{\bar{w}}|^2 (|\lambda_2^{\bar{w}}|^2 - \frac{1}{2})}{1 - |\lambda_2^{\bar{w}}|^2} & \text{if } \frac{1}{2} \leq |\lambda_2^{\bar{w}}|^2 \leq \frac{2}{3}, \\ \frac{5|\lambda_2^{\bar{w}}|^2 - 3|\lambda_2^{\bar{w}}|^4 - 2}{3|\lambda_2^{\bar{w}}|^2 - 1} & \text{if } \frac{2}{3} \leq |\lambda_2^{\bar{w}}|^2 < 1. \end{cases} \quad (\text{B.53})$$

We note that the quantities  $|\lambda_2^{\bar{w}}|^2$ ,  $\frac{|\lambda_2^{\bar{w}}|^2 (|\lambda_2^{\bar{w}}|^2 - \frac{1}{2})}{1 - |\lambda_2^{\bar{w}}|^2}$ ,  $\frac{5|\lambda_2^{\bar{w}}|^2 - 3|\lambda_2^{\bar{w}}|^4 - 2}{3|\lambda_2^{\bar{w}}|^2 - 1}$  are positive over the regimes  $0 < |\lambda_2^{\bar{w}}|^2 < \frac{1}{2}$ ,  $\frac{1}{2} \leq |\lambda_2^{\bar{w}}|^2 \leq \frac{2}{3}$  and  $\frac{2}{3} \leq |\lambda_2^{\bar{w}}|^2 < 1$ , respectively. Therefore, the constant  $\gamma_{\bar{w}} > 0$ . Thus, we can show from (B.42) that for every  $K \geq 0$  it holds that:

$$\begin{aligned} \gamma_{\bar{w}} \|\tilde{\mathbf{x}}\|_2^{\delta, K} &\leq \left( \delta^2 - |\lambda_2^{\bar{w}}|^2 - \eta\mu(3(L/\mu) + 3\eta L - 1) \right) \|\tilde{\mathbf{x}}\|_2^{\delta, K} \\ &\leq \eta\mu(L/\mu + 3\eta L - 1) \|\bar{\mathbf{e}}_x\|_2^{\delta, K} + \eta \left( \frac{1}{2L} + \eta + \frac{\eta}{2L\mu} \right) \|\tilde{\mathbf{v}}\|_2^{\delta, K} \\ &\quad + \frac{\eta}{\delta^{2K-2}} \left( N + \frac{1}{N} \right) (\eta\sigma^2 + 2d) + \delta^2 \mathbb{E} \left[ \left\| \tilde{x}^{(0)} \right\|^2 \right]. \end{aligned} \quad (\text{B.54})$$

We can obtain the desired result by dividing  $\gamma_{\bar{w}}$  on both hand sides of (B.54). The proof is complete.

## B.6 Proof of Lemma 11

First of all, Lemma 9 implies that

$$\begin{aligned} \|\bar{\mathbf{e}}_x\|_2^{\delta, K} &\leq \eta \cdot \frac{L^2 \left( \eta + \frac{1+\eta L}{\mu(1-\frac{\eta L}{2})} \right)}{N^2 \left( \delta^2 + \eta\mu \left( 1 - \frac{\eta L}{2} \right) - 1 \right)} \|\tilde{\mathbf{x}}\|_2^{\delta, K} \\ &\quad + \frac{\eta}{N\delta^{2K-2}} \cdot \frac{\eta\sigma^2 + 2d}{\delta^2 + \eta\mu \left( 1 - \frac{\eta L}{2} \right) - 1} + \frac{\delta^2}{\delta^2 + \eta\mu \left( 1 - \frac{\eta L}{2} \right) - 1} \mathbb{E} \left[ \left\| \bar{e}_x^{(0)} \right\|^2 \right] \\ &\leq \frac{4L^2 \left( 1 + \frac{2+2L}{\mu} \right)}{N^2\mu} \|\tilde{\mathbf{x}}\|_2^{\delta, K} + \frac{4}{N\delta^{2K-2}} \cdot \frac{\eta\sigma^2 + 2d}{\mu} + \frac{4\delta^2}{\eta\mu} \mathbb{E} \left[ \left\| \bar{e}_x^{(0)} \right\|^2 \right], \end{aligned} \quad (\text{B.55})$$

where we used the assumption that  $\eta \leq 1/L$  to get  $\frac{1+\eta L}{\mu(1-\frac{\eta L}{2})} \leq \frac{1+L}{\mu/2}$  in the first term, and then we chose  $\delta^2$  such that  $\delta^2 \geq 1 - \frac{\eta\mu}{2} \left( 1 - \frac{\eta L}{2} \right)$  to get  $\delta^2 - 1 + \eta\mu \left( 1 - \frac{\eta L}{2} \right) \geq \frac{\eta\mu}{2} \left( 1 - \frac{\eta L}{2} \right) \geq \frac{\eta\mu}{4}$  where we used  $\eta \leq 1/L$  again. By substituting the upper bound of  $\|\bar{\mathbf{e}}_x\|_2^{\delta, K}$  in Lemma 9 to Lemma 10, we

can compute that

$$\begin{aligned} \|\tilde{\mathbf{x}}\|_2^{\delta,K} &\leq \frac{\eta}{\gamma_{\tilde{W}}} (L/\mu + 3\eta L - 1) \frac{4L^2 \left(1 + \frac{2+2L}{\mu}\right)}{N^2} \|\tilde{\mathbf{x}}\|_2^{\delta,K} \\ &+ \frac{\eta}{\gamma_{\tilde{W}}} \left( \frac{1}{2L} + \eta + \frac{\eta}{2L\mu} \right) \|\tilde{\mathbf{v}}\|_2^{\delta,K} + \frac{\eta}{\delta^{2K-2}} (\eta\sigma^2 + 2d) \left( \frac{N + \frac{1}{N}}{\gamma_{\tilde{W}}} + \frac{4}{N\gamma_{\tilde{W}}} \cdot (L/\mu + 3\eta L - 1) \right) \\ &+ \frac{4\delta^2}{\gamma_{\tilde{W}}} (L/\mu + 3\eta L - 1) \mathbb{E} \left[ \left\| \bar{e}_x^{(0)} \right\|^2 \right] + \frac{\delta^2}{\gamma_{\tilde{W}}} \mathbb{E} \left[ \left\| \tilde{x}^{(0)} \right\|^2 \right]. \end{aligned} \quad (\text{B.56})$$

Recall the definition of  $A$  in (6.23) such that

$$A = \left( L/\mu - 1 + \frac{\gamma_{\tilde{W}}}{2(1 + \mu/L)} \right) \cdot \frac{4L^2}{N^2} \left( 1 + \frac{2+2L}{\mu} \right) \geq (L/\mu + 3\eta L - 1) \cdot \frac{4L^2}{N^2} \left( 1 + \frac{2+2L}{\mu} \right), \quad (\text{B.57})$$

where we used  $\eta \leq \frac{\gamma_{\tilde{W}}}{6(L+\mu)}$  under the assumption in Lemma 10. Furthermore, by using the assumption that  $\eta \leq \frac{\gamma_{\tilde{W}}}{2A}$ , we can further compute  $1 - \frac{\eta A}{\gamma_{\tilde{W}}} \geq \frac{1}{2}$ . Hence, we can compute from (B.56) that

$$\begin{aligned} \frac{\|\tilde{\mathbf{x}}\|_2^{\delta,K}}{2} &\leq \left( 1 - \frac{\eta A}{\gamma_{\tilde{W}}} \right) \|\tilde{\mathbf{x}}\|_2^{\delta,K} \\ &\leq \frac{\eta}{\gamma_{\tilde{W}}} \left( \frac{1}{2L} + \eta + \frac{\eta}{2L\mu} \right) \|\tilde{\mathbf{v}}\|_2^{\delta,K} + \frac{\eta}{\delta^{2K-2}} (\eta\sigma^2 + 2d) \left( \frac{N + \frac{1}{N}}{\gamma_{\tilde{W}}} + \frac{4}{N\gamma_{\tilde{W}}} \cdot (L/\mu + 3\eta L - 1) \right) \\ &+ \frac{4\delta^2}{\gamma_{\tilde{W}}} (L/\mu + 3\eta L - 1) \mathbb{E} \left[ \left\| \bar{e}_x^{(0)} \right\|^2 \right] + \frac{\delta^2}{\gamma_{\tilde{W}}} \mathbb{E} \left[ \left\| \tilde{x}^{(0)} \right\|^2 \right]. \end{aligned} \quad (\text{B.58})$$

The proof is complete by multiplying 2 on both hand sides of (B.58).

## B.7 Proof of Lemma 12

Considering (3.22) and the choice of  $\widetilde{W}$  in (3.9), we can compute that

$$U = \widetilde{W} - W = hI_N + (1-h)W - W = h(I_N - W), \quad h \in (0, 1/2]. \quad (\text{B.59})$$

From Lemma 8 and (6.10), and under our assumption that  $\mathcal{B} = B \otimes I_d$  with  $1_N^T B = c$ , we have  $\mathcal{U}\mathcal{B}\mathbf{x}_* = c\mathcal{U}\mathbf{x}_* = 0$ , we can compute

$$\begin{aligned} \|\tilde{v}^{(k+1)}\|^2 &= \|e_v^{(k+1)}\|^2 = \|v^{(k+1)} + \nabla F(\mathbf{x}_*)\|^2 \\ &= \left\| \tilde{v}^{(k)} - h(I_{Nd} - \mathcal{W}) \left( \tilde{v}^{(k)} + \nabla F(x^{(k)}) - \nabla F(\mathbf{x}_*) - \mathcal{B}(x^{(k)} - \mathbf{x}_*) \right) \right. \\ &\quad \left. - h(I_{Nd} - \mathcal{W})\xi^{(k)} + (h/\eta)(I_{Nd} - \mathcal{W})\sqrt{2\eta}w^{(k+1)} \right\|^2 \end{aligned} \quad (\text{B.60})$$

$$\begin{aligned} &= \left\| ((1-h)I_{Nd} + h\mathcal{W})\tilde{v}^{(k)} - \left[ h(I_{Nd} - \mathcal{W}) \left( \nabla F(x^{(k)}) - \nabla F(\mathbf{x}_*) \right) \right. \right. \\ &\quad \left. \left. - h(I_{Nd} - \mathcal{W})\mathcal{B}(x^{(k)} - \mathbf{x}_*) + h(I_{Nd} - \mathcal{W})\xi^{(k)} - (h/\eta)(I_{Nd} - \mathcal{W})\sqrt{2\eta}w^{(k+1)} \right] \right\|^2 \\ &= \|((1-h)I_{Nd} + h\mathcal{W})\tilde{v}^{(k)}\|^2 \end{aligned} \quad (\text{B.61})$$

$$+ \left\| h(I_{Nd} - \mathcal{W}) \left( \nabla F(x^{(k)}) - \nabla F(\mathbf{x}_*) \right) - h(I_{Nd} - \mathcal{W})\mathcal{B}(x^{(k)} - \mathbf{x}_*) \right\|^2 \quad (\text{B.62})$$

$$\begin{aligned} &+ \left\| h(I_{Nd} - \mathcal{W})\xi^{(k)} - (h/\eta)(I_{Nd} - \mathcal{W})\sqrt{2\eta}w^{(k+1)} \right\|^2 \\ &+ 2 \left\langle h(I_{Nd} - \mathcal{W}) \left( \nabla F(x^{(k)}) - \nabla F(\mathbf{x}_*) \right) - h(I_{Nd} - \mathcal{W})\mathcal{B}(x^{(k)} - \mathbf{x}_*) , \right. \\ &\quad \left. h(I_{Nd} - \mathcal{W})\xi^{(k)} - (h/\eta)(I_{Nd} - \mathcal{W})\sqrt{2\eta}w^{(k+1)} \right\rangle \\ &- 2 \left\langle ((1-h)I_{Nd} + h(\mathcal{W} - \mathcal{J}))\tilde{v}^{(k)} , h(I_{Nd} - \mathcal{W}) \left( \nabla F(x^{(k)}) - \nabla F(\mathbf{x}_*) \right) \right. \\ &\quad \left. - h(I_{Nd} - \mathcal{W})\mathcal{B}(x^{(k)} - \mathbf{x}_*) \right\rangle \end{aligned} \quad (\text{B.63})$$

$$+ 2 \left\langle ((1-h)I_{Nd} + h(\mathcal{W} - \mathcal{J}))\tilde{v}^{(k)} , h(I_{Nd} - \mathcal{W})\xi^{(k)} - (h/\eta)(I_{Nd} - \mathcal{W})\sqrt{2\eta}w^{(k+1)} \right\rangle.$$

Since  $\mathcal{J}\tilde{v}^{(k)} = \mathcal{J}e_v^{(k)} = \bar{e}_v^{(k)} \otimes I_d = 0$  by Lemma 8, we compute (B.61) such that

$$\begin{aligned}
\left\| ((1-h)I_{Nd} + h\mathcal{W})\tilde{v}^{(k)} \right\|^2 &= \left\| ((1-h)I_{Nd} + h(\mathcal{W} - \mathcal{J}))\tilde{v}^{(k)} \right\|^2 \\
&\leq \|(1-h)I_N + h(W - J)\|^2 \cdot \left\| \tilde{v}^{(k)} \right\|^2 \\
&\leq (1 - 2h + h^2 + 2(1-h)h\bar{\gamma}_w + h^2\bar{\gamma}_w^2) \left\| \tilde{v}^{(k)} \right\|^2 \\
&= (1 - 2h(1 - \bar{\gamma}_w) + h^2(1 - \bar{\gamma}_w)^2) \left\| \tilde{v}^{(k)} \right\|^2 \\
&\leq (1 - h(1 - \bar{\gamma}_w)) \left\| \tilde{v}^{(k)} \right\|^2,
\end{aligned} \tag{B.64}$$

where we used the assumption  $h < 1 < \frac{1}{1-\bar{\gamma}_w}$  in the last inequality. Then by  $L$ -smoothness of  $F$ , we can bound the term (B.62) as follows:

$$\begin{aligned}
&\left\| h(I_{Nd} - \mathcal{W}) \left( \nabla F(x^{(k)}) - \nabla F(\mathbf{x}_*) \right) - h(I_{Nd} - \mathcal{W})\mathcal{B}(x^{(k)} - \mathbf{x}_*) \right\|^2 \\
&\leq 2 \left\| h(I_{Nd} - \mathcal{W}) \left( \nabla F(x^{(k)}) - \nabla F(\mathbf{x}_*) \right) \right\|^2 + 2 \left\| h(I_{Nd} - \mathcal{W})\mathcal{B}(x^{(k)} - \mathbf{x}_*) \right\|^2 \\
&\leq 2\bar{\gamma}_{I_N-W}^2 (h^2L^2 + \|hB\|^2) \left\| e_x^{(k)} \right\|^2.
\end{aligned} \tag{B.65}$$

Next, we compute the inner product term in (B.63).

$$\begin{aligned}
&2 \left\langle ((1-h)I_{Nd} + h\mathcal{W})\tilde{v}^{(k)}, h(I_{Nd} - \mathcal{W})\mathcal{B}(x^{(k)} - \mathbf{x}_*) \right\rangle \\
&\quad - 2 \left\langle ((1-h)I_{Nd} + h\mathcal{W})\tilde{v}^{(k)}, h(I_{Nd} - \mathcal{W}) \left( \nabla F(x^{(k)}) - \nabla F(\mathbf{x}_*) \right) \right\rangle \\
&\leq 2 \left\langle (I_{Nd} - \mathcal{W})\tilde{v}^{(k)}, h\mathcal{B}(x^{(k)} - \mathbf{x}_*) - h \left( \nabla F(x^{(k)}) - \nabla F(\mathbf{x}_*) \right) \right\rangle \tag{B.66}
\end{aligned}$$

$$\begin{aligned}
&\quad - 2 \left\langle h(I_{Nd} - \mathcal{W})\tilde{v}^{(k)}, h(I_{Nd} - \mathcal{W})\mathcal{B}(x^{(k)} - \mathbf{x}_*) \right\rangle \\
&\quad - 2 \left\langle h(I_{Nd} - \mathcal{W})\tilde{v}^{(k)}, h(I_{Nd} - \mathcal{W}) \left( \nabla F(x^{(k)}) - \nabla F(\mathbf{x}_*) \right) \right\rangle. \tag{B.67}
\end{aligned}$$

We can use the  $L$ -smoothness of  $F$  to bound the term (B.66). It follows that

$$\begin{aligned}
&2 \left\langle (I_{Nd} - \mathcal{W})\tilde{v}^{(k)}, h\mathcal{B}(x^{(k)} - \mathbf{x}_*) - h \left( \nabla F(x^{(k)}) - \nabla F(\mathbf{x}_*) \right) \right\rangle \\
&\leq 2 \left\| (I_{Nd} - \mathcal{W})\tilde{v}^{(k)} \right\| \cdot \left\| h\mathcal{B}(x^{(k)} - \mathbf{x}_*) - h \left( \nabla F(x^{(k)}) - \nabla F(\mathbf{x}_*) \right) \right\| \\
&\leq (1/c) \left\| (I_{Nd} - \mathcal{W})\tilde{v}^{(k)} \right\|^2 + c \left\| h\mathcal{B}(x^{(k)} - \mathbf{x}_*) - h \left( \nabla F(x^{(k)}) - \nabla F(\mathbf{x}_*) \right) \right\|^2 \\
&\leq (1/c) \left\| (I_{Nd} - \mathcal{W})\tilde{v}^{(k)} \right\|^2 + 2c \left\| h\mathcal{B}(x^{(k)} - \mathbf{x}_*) \right\|^2 + 2c \left\| h \left( \nabla F(x^{(k)}) - \nabla F(\mathbf{x}_*) \right) \right\|^2 \\
&\leq (1/c)\bar{\gamma}_{I_N-W}^2 \left\| \tilde{v}^{(k)} \right\|^2 + 2c\|hB\|^2 \left\| e_x^{(k)} \right\|^2 + 2ch^2L^2 \left\| e_x^{(k)} \right\|^2,
\end{aligned} \tag{B.68}$$

where we used the inequality  $2xy \leq cx^2 + y^2/c$  for any  $c > 0$  and  $x, y \in \mathbb{R}$ . Therefore, we can compute the inner product term (B.63) as follows.

$$\begin{aligned}
& 2 \left\langle ((1-h)I_{Nd} + h\mathcal{W}) \tilde{v}^{(k)}, h(I_{Nd} - \mathcal{W}) \mathcal{B}(x^{(k)} - \mathbf{x}_*) \right\rangle \\
& - 2 \left\langle (1-h)I_{Nd} + h\mathcal{W}, h(I_{Nd} - \mathcal{W}) (\nabla F(x^{(k)}) - \nabla F(\mathbf{x}_*)) \right\rangle \\
& \leq (1/c)\bar{\gamma}_{I_N-W}^2 \|\tilde{v}^{(k)}\|^2 + 2c\|hB\|^2 \|e_x^{(k)}\|^2 + 2ch^2L^2 \|e_x^{(k)}\|^2 \\
& + h^2\bar{\gamma}_{I_N-W}^2 \|\tilde{v}^{(k)}\|^2 + \|hB\|^2 \bar{\gamma}_{I_N-W}^2 \|e_x^{(k)}\|^2 + h^2\bar{\gamma}_{I_N-W}^2 \|\tilde{v}^{(k)}\|^2 + h^2L^2 \bar{\gamma}_{I_N-W}^2 \|e_x^{(k)}\|^2 \\
& = (1/c)\bar{\gamma}_{I_N-W}^2 \|\tilde{v}^{(k)}\|^2 + 2h^2\bar{\gamma}_{I_N-W}^2 \|\tilde{v}^{(k)}\|^2 \\
& + 2ch_\eta^2 L^2 \|e_x^{(k)}\|^2 + h^2L^2 \bar{\gamma}_{I_N-W}^2 \|e_x^{(k)}\|^2 + 2cL\|hB\|^2 \|e_x^{(k)}\|^2 + \|hB\|^2 \bar{\gamma}_{I_N-W}^2 \|e_x^{(k)}\|^2 \\
& \leq (1/c)\bar{\gamma}_{I_N-W}^2 \|\tilde{v}^{(k)}\|^2 + 2h^2\bar{\gamma}_{I_N-W}^2 \|\tilde{v}^{(k)}\|^2 + (3ch^2L^2 + 3cL\|hB\|^2) \|e_x^{(k)}\|^2, \tag{B.69}
\end{aligned}$$

where we assume that  $c \geq \bar{\gamma}_{I_N-W}^2$ . Now we take the expectation of (B.60) to get

$$\begin{aligned}
\mathbb{E} \left[ \|\tilde{v}^{(k+1)}\|^2 \right] & \leq (1-h(1-\bar{\gamma}_w) + (1/c)\bar{\gamma}_{I_N-W}^2 + 2h^2\bar{\gamma}_{I_N-W}^2) \mathbb{E} \left[ \|\tilde{v}^{(k)}\|^2 \right] \\
& + (3ch^2L^2 + 3cL\|hB\|^2) \mathbb{E} \left[ \|e_x^{(k)}\|^2 \right] + 2h^2\sigma^2N + 4\eta(h/\eta)^2dN. \\
& \leq \left( 1 - \frac{h(1-\bar{\gamma}_w)}{2} + (1/c)\bar{\gamma}_{I_N-W}^2 \right) \mathbb{E} \left[ \|\tilde{v}^{(k)}\|^2 \right] \\
& + (3ch^2L^2 + 3cL\|hB\|^2) \mathbb{E} \left[ \|e_x^{(k)}\|^2 \right] + 2h^2\sigma^2N + 4\eta(h/\eta)^2dN, \tag{B.70}
\end{aligned}$$

under the assumption such that  $h \leq \frac{1-\bar{\gamma}_w}{4\bar{\gamma}_{I_N-W}^2}$ . By taking  $c = \frac{2\bar{\gamma}_{I_N-W}}{(1-\bar{\gamma}_w)h}$  where  $h \leq 1 < \frac{1}{\bar{\gamma}_{I_N-W}^2}$  under our assumptions, we can compute

$$\begin{aligned}
1 - \frac{h(1-\bar{\gamma}_w)}{2} + (1/c)\bar{\gamma}_{I_N-W}^2 & = 1 - \frac{h(1-\bar{\gamma}_w)}{2} + \frac{(1-\bar{\gamma}_w)h}{2}\bar{\gamma}_{I_N-W} \\
& \leq 1 - h\frac{1-\bar{\gamma}_w}{2} (1-\bar{\gamma}_{I_N-W}) > 0, \tag{B.71}
\end{aligned}$$

where  $1-\bar{\gamma}_w \leq \bar{\gamma}_{I_N-W} < 1$  by definition (6.25), then we can find the constant

$$\delta^2 \geq 1 - \frac{h}{2} \frac{1-\bar{\gamma}_w}{2} (1-\bar{\gamma}_{I_N-W}) > 0. \tag{B.72}$$

Therefore, by Lemma 6, we get

$$\begin{aligned}
\frac{h}{2} \frac{1-\bar{\gamma}_w}{2} (1-\bar{\gamma}_{I_N-W}) \|\tilde{\mathbf{v}}\|_2^{\delta,K} & \leq \left( \delta^2 - \left( 1 - h\frac{1-\bar{\gamma}_w}{2} (1-\bar{\gamma}_{I_N-W}) \right) \right) \|\tilde{\mathbf{v}}\|_2^{\delta,K} \\
& \leq \left( 3h(h/\eta)L^2 + 3h(h/\eta)L\|B\|^2 \right) \|\mathbf{e}_x\|_2^{\delta,K} \\
& + h^2 \cdot \frac{2\sigma^2N}{\delta^{2K-2}} + \eta \cdot (h/\eta)^2 \frac{4dN}{\delta^{2K-2}} + \delta^2 \|\tilde{v}^{(0)}\|^2. \tag{B.73}
\end{aligned}$$

Then it follows that

$$\begin{aligned} \|\tilde{\mathbf{v}}\|_2^{\delta,K} &\leq \frac{12(h/\eta) (L^2 + L \|B\|^2)}{(1 - \bar{\gamma}_w) (1 - \bar{\gamma}_{I_N-W}^2)} \|\mathbf{e}_x\|_2^{\delta,K} + \frac{8N(h/\eta)}{(1 - \bar{\gamma}_w) (1 - \bar{\gamma}_{I_N-W}^2)} \frac{\eta\sigma^2 + 2d}{\delta^{2K-2}} \\ &\quad + \frac{4\delta^2}{h(1 - \bar{\gamma}_w) (1 - \bar{\gamma}_{I_N-W}^2)} \left\| \tilde{v}^{(0)} \right\|^2, \end{aligned} \quad (\text{B.74})$$

where we note that  $h/\eta$  is in the order of  $\eta^\alpha$  under our assumption. Moreover, by Lemma 9, see also (B.55), we have

$$\|\bar{\mathbf{e}}_x\|_2^{\delta,K} \leq \frac{4L^2 \left(1 + \frac{2+2L}{\mu}\right)}{N^2\mu} \|\tilde{\mathbf{x}}\|_2^{\delta,K} + \frac{4}{N\delta^{2K-2}} \cdot \frac{\eta\sigma^2 + 2d}{\mu} + \frac{4\delta^2}{\eta\mu} \mathbb{E} \left[ \left\| \bar{e}_x^{(0)} \right\|^2 \right], \quad (\text{B.75})$$

Hence, by  $\tilde{x}^{(k)} + 1_N \otimes \bar{e}_x^{(k)} = e_x^{(k)}$ , we can compute that

$$\begin{aligned} \|\mathbf{e}_x\|_2^{\delta,K} &= \|\tilde{\mathbf{x}}\|_2^{\delta,K} + \|\bar{\mathbf{e}}_x\|_2^{\delta,K} \\ &\leq \left(1 + \frac{4L^2 \left(1 + \frac{2+2L}{\mu}\right)}{N^2\mu}\right) \|\tilde{\mathbf{x}}\|_2^{\delta,K} + \frac{4}{N\mu} \cdot \frac{\eta\sigma^2 + 2d}{\delta^{2K-2}} + \frac{4\delta^2}{\eta\mu} \mathbb{E} \left[ \left\| \bar{e}_x^{(0)} \right\|^2 \right]. \end{aligned} \quad (\text{B.76})$$

Therefore, we can substitute the formula (B.76) into the upper bound of  $\|\tilde{\mathbf{v}}\|_2^{\delta,K}$  in (B.74) to get

$$\begin{aligned} \|\tilde{\mathbf{v}}\|_2^{\delta,K} &\leq \frac{12(h/\eta) (L^2 + L \|B\|^2)}{(1 - \bar{\gamma}_w) (1 - \bar{\gamma}_{I_N-W}^2)} \|\mathbf{e}_x\|_2^{\delta,K} + \frac{8N(h/\eta)}{(1 - \bar{\gamma}_w) (1 - \bar{\gamma}_{I_N-W}^2)} \frac{\eta\sigma^2 + 2d}{\delta^{2K-2}} \\ &\quad + \frac{4\delta^2}{h(1 - \bar{\gamma}_w) (1 - \bar{\gamma}_{I_N-W}^2)} \left\| \tilde{v}^{(0)} \right\|^2 \\ &\leq \frac{12(h/\eta) (L^2 + L \|B\|^2)}{(1 - \bar{\gamma}_w) (1 - \bar{\gamma}_{I_N-W}^2)} \left(1 + \frac{4L^2 \left(1 + \frac{2+2L}{\mu}\right)}{N^2\mu}\right) \|\tilde{\mathbf{x}}\|_2^{\delta,K} \\ &\quad + \left( \frac{6(L^2 + L \|B\|^2)}{N\mu} + N \right) \cdot \frac{8(h/\eta)}{(1 - \bar{\gamma}_w) (1 - \bar{\gamma}_{I_N-W}^2)} \cdot \frac{\eta\sigma^2 + 2d}{\delta^{2K-2}} \\ &\quad + \frac{12\delta^2(h/\eta) (L^2 + L \|B\|^2)}{\eta\mu(1 - \bar{\gamma}_w) (1 - \bar{\gamma}_{I_N-W}^2)} \mathbb{E} \left[ \left\| \bar{e}_x^{(0)} \right\|^2 \right] + \frac{4\delta^2}{h(1 - \bar{\gamma}_w) (1 - \bar{\gamma}_{I_N-W}^2)} \left\| \tilde{v}^{(0)} \right\|^2. \end{aligned} \quad (\text{B.77})$$

The proof is complete.

## B.8 Proof of Lemma 14

We first use (6.13) to get the following first inequality, and then we use the uniform upper bound for  $\|\tilde{\mathbf{v}}\|_2^{\delta,k}$  from (6.41) in Theorem 13 to derive as follows.

$$\begin{aligned} & \frac{1}{\delta^{2k}} \mathbb{E} \left[ \left\| \tilde{v}^{(k)} \right\|^2 \right] \\ & \leq \|\tilde{\mathbf{v}}\|_2^{\delta,k} \\ & \leq \frac{h\eta}{\delta^{2k-2}} \cdot \frac{\gamma_2 (w_2\gamma_1(h/\eta) + w_1)\sigma^2/N}{1 - h\gamma_1\gamma_2} + \frac{h}{\delta^{2k-2}} \cdot \left[ \frac{2\gamma_2 d (w_2\gamma_1(h/\eta) + w_1)/N}{1 - h\gamma_1\gamma_2} + \frac{w_2\sigma^2}{N} \right. \\ & \quad \left. + \frac{\gamma_1\gamma_2}{1 - h\gamma_1\gamma_2} \delta^{2k} \left( (h/\eta)(E_3/\eta) \mathbb{E} \left[ \left\| \bar{e}_x^{(0)} \right\|^2 \right] + (E_4/h) \mathbb{E} \left[ \left\| \tilde{v}^{(0)} \right\|^2 \right] \right) \right] \\ & \quad + (h/\eta)\delta^2 \left( \frac{2dw_2}{N\delta^{2k}} + \gamma_2 D_0 \right) + \delta^2 (h/\eta)(E_3/\eta) \mathbb{E} \left[ \left\| \bar{e}_x^{(0)} \right\|^2 \right] + (E_4/h) \mathbb{E} \left[ \left\| \tilde{v}^{(0)} \right\|^2 \right]. \end{aligned} \quad (\text{B.78})$$

Therefore, we obtain:

$$\begin{aligned} \mathbb{E} \left[ \left\| \tilde{v}^{(k)} \right\|^2 \right] & \leq h\delta^2 \cdot (\eta\sigma^2 + 2d) \cdot \frac{\gamma_2 (w_2\gamma_1(h/\eta) + w_1)/N}{1 - h\gamma_1\gamma_2} + h\delta^2 \cdot \frac{w_2\sigma^2}{N} + (h/\eta)\delta^2 \frac{2dw_2}{N} \\ & \quad + \delta^{2k+2} \cdot \left( \frac{h\gamma_1\gamma_2}{1 - h\gamma_1\gamma_2} \left( (h/\eta)(E_3/\eta) \mathbb{E} \left[ \left\| \bar{e}_x^{(0)} \right\|^2 \right] + (E_4/h) \mathbb{E} \left[ \left\| \tilde{v}^{(0)} \right\|^2 \right] \right) \right. \\ & \quad \left. + (h/\eta)\gamma_2 D_0 + (h/\eta)(E_3/\eta) \mathbb{E} \left[ \left\| \bar{e}_x^{(0)} \right\|^2 \right] + (E_4/h) \mathbb{E} \left[ \left\| \tilde{v}^{(0)} \right\|^2 \right] \right) \\ & =: h\delta^2 \cdot \frac{C_1\gamma_2}{2L^2} + \delta^{2k+2} h \cdot \frac{C_0\gamma_1\gamma_2}{2L^2} \\ & \quad + \delta^{2k+2} \cdot ((h/\eta)\gamma_2 D_0 + C_0) + (h/\eta)\delta^2 \cdot \frac{w_2}{N} (\eta\sigma^2 + 2d). \end{aligned} \quad (\text{B.79})$$

By uniform bound for  $\|\tilde{\mathbf{x}}\|_2^{\delta,k}$  in (6.40) in Theorem 13, we can obtain that

$$\begin{aligned} \delta^{2k} \|\tilde{\mathbf{x}}\|_2^{\delta,k} & \leq \delta^2 \eta \cdot \frac{\eta\sigma^2 + 2d}{N} \cdot \frac{w_2\gamma_1(h/\eta) + w_1}{1 - h\gamma_1\gamma_2} \\ & \quad + \delta^{2k+2} \eta \cdot \frac{\gamma_1}{1 - h\gamma_1\gamma_2} \left( (h/\eta)(E_3/\eta) \mathbb{E} \left[ \left\| \bar{e}_x^{(0)} \right\|^2 \right] + (E_4/h) \mathbb{E} \left[ \left\| \tilde{v}^{(0)} \right\|^2 \right] \right) + \delta^{2k+2} D_0 \\ & = \delta^2 \eta \cdot \frac{C_1}{2L^2} + \delta^{2k+2} \eta \cdot \frac{\gamma_1 C_0}{2L^2} + \delta^{2k+2} D_0. \end{aligned} \quad (\text{B.80})$$

Then we can further get the uniform bound for  $\mathbb{E} \left[ \|\nabla F(x^{(k)})\|^2 \right]$  in the following derivation.

$$\begin{aligned}
& \mathbb{E} \left[ \|\nabla F(x^{(k)})\|^2 \right] \\
& \leq 2\mathbb{E} \left[ \|\nabla F(x^{(k)}) - \nabla F(\mathbf{x}_*)\|^2 \right] + 2 \|\nabla F(\mathbf{x}_*)\|^2 \\
& \leq 2L^2 \mathbb{E} \left[ \|x^{(k)} - \mathbf{x}_*\|^2 \right] + 2 \|\nabla F(\mathbf{x}_*)\|^2 \\
& \leq 2L^2 \delta^{2k} \left( \|\tilde{\mathbf{x}}\|_2^{\delta,k} + \|\bar{\mathbf{e}}_x\|_2^{\delta,k} \right) + 2 \|\nabla F(\mathbf{x}_*)\|^2 \\
& \leq \delta^{2k} \cdot 2L^2 \left( \|\tilde{\mathbf{x}}\|_2^{\delta,k} + \eta \cdot \frac{L^2}{N^2 \left( \delta^2 + \eta\mu \left( 1 - \frac{\eta L}{2} \right) - 1 \right)} \left( \eta + \frac{1 + \eta L}{\mu \left( 1 - \frac{\eta L}{2} \right)} \right) \|\tilde{\mathbf{x}}\|_2^{\delta,k} \right. \\
& \quad \left. + \frac{\eta}{N\delta^{2k-2}} \cdot \frac{\eta\sigma^2 + 2d}{\delta^2 + \eta\mu \left( 1 - \frac{\eta L}{2} \right) - 1} + \frac{\delta^2}{\delta^2 + \eta\mu \left( 1 - \frac{\eta L}{2} \right) - 1} \mathbb{E} \left[ \|\bar{\mathbf{e}}_x^{(0)}\|^2 \right] \right) + 2 \|\nabla F(\mathbf{x}_*)\|^2 \\
& = 2L^2 \left( \delta^2 \eta \cdot \frac{C_1}{2L^2} + \delta^{2k+2} \eta \cdot \frac{\gamma_1 C_0}{2L^2} + \delta^{2k+2} D_0 \right) \\
& \quad + \eta \cdot \left( \delta^2 \eta \cdot \frac{C_1 C_2}{2L^2} + \delta^{2k+2} \eta \cdot \frac{\gamma_1 C_0 C_2}{2L^2} + \delta^{2k+2} D_0 C_2 \right) \\
& \quad + \delta^2 \eta \cdot C_3 + \delta^{2k+2} \cdot C_4 + 2 \|\nabla F(\mathbf{x}_*)\|^2 \\
& = \eta \delta^2 (C_1 + C_3) + \delta^2 \eta^2 \left( \frac{C_1 C_2}{2L^2} \right) + \delta^{2k+2} \eta (\gamma_1 C_0 + D_0 C_2) \\
& \quad + \delta^{2k+2} \eta^2 \left( \frac{\gamma_1 C_0 C_2}{2L^2} \right) + \delta^{2k+2} (D_0 + C_4) + 2 \|\nabla F(\mathbf{x}_*)\|^2,
\end{aligned} \tag{B.81}$$

with

$$C_2 := \frac{2L^4}{N^2 \left( \delta^2 + \eta\mu \left( 1 - \frac{\eta L}{2} \right) - 1 \right)} \left( \eta + \frac{1 + \eta L}{\mu \left( 1 - \frac{\eta L}{2} \right)} \right), \tag{B.82}$$

and

$$C_3 := \frac{2L^2}{N} \cdot \frac{\eta\sigma^2 + 2d}{\delta^2 + \eta\mu \left( 1 - \frac{\eta L}{2} \right) - 1}, \quad C_4 := \frac{2L^2}{\delta^2 + \eta\mu \left( 1 - \frac{\eta L}{2} \right) - 1} \mathbb{E} \left[ \|\bar{\mathbf{e}}_x^{(0)}\|^2 \right], \tag{B.83}$$

where we used the fact that  $x^{(k)} - \mathbf{x}_* = e_x^{(k)} = \tilde{x}^{(k)} + 1_N \otimes \bar{\mathbf{e}}_x^{(k)}$  in Lemma 8, and the last inequality in (B.81) follows bound in (B.80) above. The proof is complete.

## B.9 Proof of Corollary 15

We proved in Lemma 14 such that

$$\begin{aligned}\mathbb{E} \left[ \left\| \tilde{v}^{(k)} \right\|^2 \right] &\leq h\delta^2 \cdot \frac{C_1\gamma_2}{2L^2} + \delta^{2k+2}h \cdot \frac{C_0\gamma_1\gamma_2}{2L^2} \\ &\quad + \delta^{2k+2} \cdot ((h/\eta)\gamma_2 D_0 + C_0) + (h/\eta)\delta^2 \cdot \frac{w_2}{N} (\eta\sigma^2 + 2d),\end{aligned}\quad (\text{B.84})$$

$$\begin{aligned}\mathbb{E} \left[ \left\| \nabla F(x^{(k)}) \right\|^2 \right] &\leq \eta\delta^2 (C_1 + C_3) + \delta^2\eta^2 \left( \frac{C_1C_2}{2L^2} \right) + \delta^{2k+2}\eta (\gamma_1 C_0 + D_0 C_2) \\ &\quad + \delta^{2k+2}\eta^2 \left( \frac{\gamma_1 C_0 C_2}{2L^2} \right) + \delta^{2k+2} (D_0 + C_4) + 2 \|\nabla F(\mathbf{x}_*)\|^2.\end{aligned}\quad (\text{B.85})$$

Since  $0 < \delta < 1$ , for any  $K_0 \geq 0$  such that for every  $k \geq K_0 \geq 0$ , it holds that

$$\begin{aligned}\mathbb{E} \left[ \left\| \tilde{v}^{(k)} \right\|^2 \right] &\leq h\delta^2 \left( \frac{C_1\gamma_2}{2L^2} + \frac{C_0\gamma_1\gamma_2}{2L^2} \right) + (h/\eta)\delta^2 \left( \gamma_2 D_0 + \frac{w_2}{N} (\eta\sigma^2 + 2d) \right) + \delta^{2K_0} C_0, \\ \mathbb{E} \left[ \left\| \nabla F(x^{(k)}) \right\|^2 \right] &\leq \eta\delta^2 (C_1 + C_3 + \gamma_1 C_0 + D_0 C_2) + \delta^2\eta^2 \left( \frac{C_1C_2}{2L^2} + \frac{\gamma_1 C_0 C_2}{2L^2} \right) \\ &\quad + \delta^{2K_0} (D_0 + C_4) + 2 \|\nabla F(\mathbf{x}_*)\|^2.\end{aligned}\quad (\text{B.86})$$

In particular, we use the inequality  $1 - \frac{1}{x} \leq \log(x) \leq x - 1$ , and choose the constant  $K_0$  as follows:

$$K_0 := \frac{\delta^2}{1 - \delta^2} \left[ \left( 1 - \frac{\|\nabla F(\mathbf{x}_*)\|^2}{D_0 + C_4} \right) \vee \left( 1 - \frac{\|\nabla F(\mathbf{x}_*)\|^2}{C_0} \right) \right] \vee 0,\quad (\text{B.87})$$

which induces that

$$\delta^{2K_0} C_0 \vee \delta^{2K_0} (D_0 + C_4) \leq \|\nabla F(\mathbf{x}_*)\|^2.\quad (\text{B.88})$$

Therefore, we can obtain the uniform bounds such that

$$\mathbb{E} \left[ \left\| \tilde{v}^{(k)} \right\|^2 \right] \leq R_h, \quad \mathbb{E} \left[ \left\| \nabla F(x^{(k)}) \right\|^2 \right] \leq R'_h,\quad (\text{B.89})$$

for any  $k \geq K_0$ , where

$$R_h := h\delta^2 \left( \frac{C_1\gamma_2}{2L^2} + \frac{C_0\gamma_1\gamma_2}{2L^2} \right) + (h/\eta)\delta^2 \left( \gamma_2 D_0 + \frac{w_2}{N} (\eta\sigma^2 + 2d) \right) + \|\nabla F(\mathbf{x}_*)\|^2,\quad (\text{B.90})$$

$$R'_h := \eta\delta^2 (C_1 + C_3 + \gamma_1 C_0 + D_0 C_2) + \delta^2\eta^2 \left( \frac{C_1C_2}{2L^2} + \frac{\gamma_1 C_0 C_2}{2L^2} \right) + 3 \|\nabla F(\mathbf{x}_*)\|^2.\quad (\text{B.91})$$

This completes the proof.