Journal of Machine Learning Research 22 (2021) 1-69 Submitted 3/21; Revised 7/21; Published 8/21

Decentralized Stochastic Gradient Langevin Dynamics and Hamiltonian Monte Carlo

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

Xuefeng Gao* XFGAO@SE.CUHK.EDU.HK
Department of Systems Engineering and Engineering Management
The Chinese University of Hong Kong
Shatin, N.T., Hong Kong, China

Yuanhan Hu* YUANHAN.HU@RUTGERS.EDU
Department of Management Science and Information Systems
Rutgers Business School
Piscataway, NJ 08854, United States of America

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

* *The authors are in alphabetical order.*

Editor: Marc Peter Deisenroth

Abstract

Stochastic gradient Langevin dynamics (SGLD) and stochastic gradient Hamiltonian Monte Carlo (SGHMC) are two popular Markov Chain Monte Carlo (MCMC) algorithms for Bayesian inference that can scale to large datasets, allowing to sample from the posterior distribution of the parameters of a statistical model given the input data and the prior distribution over the model parameters. However, these algorithms do not apply to the decentralized learning setting, when a network of agents are working collaboratively to learn the parameters of a statistical model without sharing their individual data due to privacy reasons or communication constraints. We study two algorithms: Decentralized SGLD (DE-SGLD) and Decentralized SGHMC (DE-SGHMC) which are adaptations of SGLD and SGHMC methods that allow scaleable Bayesian inference in the decentralized setting for large datasets. We show that when the posterior distribution is strongly log-concave and smooth, the iterates of these algorithms converge linearly to a neighborhood of the target distribution in the 2-Wasserstein distance if their parameters are selected appropriately. We illustrate the efficiency of our algorithms on decentralized Bayesian linear regression and Bayesian logistic regression problems.

Keywords: Langevin dynamics, Hamiltonian Monte Carlo, decentralized algorithms, decentralized Bayesian inference, stochastic gradient, momentum acceleration, Heavy-ball method, convergence rate, Wasserstein distance.

1. Introduction

Recent decades have witnessed the era of big data, and there has been an exponential growth in the amount of data collected and stored with ever-increasing rates. Since the rate at which data is generated is often outpacing our ability to analyze it in terms of computational resources at hand, there has been a lot of recent interests for developing scaleable machine learning algorithms which are efficient on large datasets.

In the modern world, digital devices such as smart phones, tablets, wearables, sensors or video cameras are major sources of data generation. Often these devices are connected over a communication network (such as a wireless network or a sensor network) that has a high latency or a limited bandwidth. 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. Arjevani et al. (2020); He et al. (2018); Hendrikx et al. (2019); Kungurtsev (2020).

A number of approaches for scaleable decentralized learning have been proposed in the literature such as decentralized stochastic approximation and optimization algorithms (Gorbunov et al., 2019; Nedic, 2020; Scaman et al., 2019; Uribe et al., 2017) or decentralized maximum-likelihood estimation approaches (Blatt and Hero, 2004; Rabbat and Nowak, 2004). However, these approaches are optimization-based or in the maximum-likelihood settings, and therefore lead to point estimates for the model parameters to be learned. On the other hand, Bayesian methods allow a characterization of the full posterior distribution over the parameters, and therefore can provide a more detailed grasp of uncertainties that are part of the learning process and offer robustness to overfitting. There are a number of scaleable Bayesian methods in the literature based on variational inference methods (Sato, 2001; Hoffman et al., 2010; Lin, 2013), Bayesian coreset methods (Huggins et al., 2016; Campbell and Broderick, 2019) and Markov Chain Monte Carlo (MCMC) based methods including Stochastic Gradient Langevin Dynamics (SGLD) (Welling and Teh, 2011), Stochastic Gradient Hamiltonian Monte Carlo (SGHMC) (Chen et al., 2014; Zou et al., 2018a) and their variants that can handle streaming data (Broderick et al., 2013). There are also versions of these methods such as consensus Monte Carlo methods which distribute and parallelize the computations needed for Monte Carlo sampling across many computational nodes on a cluster (Ahn et al., 2014; Xu et al., 2014; Rabinovich et al., 2015; Broderick et al., 2013), however none of these methods are applicable to the decentralized setting either because they need to move the data to a centralized location or because they require a global computational unit with which each learning agent is in communication or is the main thread on a multi-threaded computer which is not applicable to decentralized learning applications. In this paper, we consider two algorithms DE-SGLD and DE-SGHMC which are adaptations of the SGLD and SGHMC algorithms to the decentralized setting and show that they can be both theoretically and practically efficient for sampling from the posterior distribution when the density of the target distribution $\pi(x) \propto e^{-f(x)}$ is strongly log-concave (i.e. f is strongly convex) and f is smooth.

Before introducing the DE-SGLD algorithm, we consider the problem of decentralized Bayesian inference: We have N agents connected over a network $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ where $\mathcal{V} = \{1, 2, \ldots, N\}$ represents the agents and $\mathcal{E} \subseteq \mathcal{V} \times \mathcal{V}$ is 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 $A = [a_1, \ldots, a_n]$ be a dataset consisting of n independent and identically distributed (i.i.d.) data vectors sampled from a parametrized distribution $p(A|x)$ 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 A_i of the data where $A_i = \{a_1^i, a_2^i, \ldots, a_{n_i}^i\}$ and n_i is the number of samples of the agent i. The data is held disjointly over agents; i.e. $A = \cup_i A_i$ with $A_i \cap A_j = \emptyset$ for $j \neq i$. The goal is to sample from the posterior distribution $p(x|A) \propto p(A|x)p(x)$. Since the data points are independent, the log-likelihood function will be additive; $\log p(A|x) = \sum_{i=1}^{N} \sum_{j=1}^{n_i} \log p(a_j^i|x)$. Thus, if we set

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

the aim is to sample from the posterior distribution with density $\pi(x) := p(x|A) \propto e^{-f(x)}$. The functions $f_i(x)$ are called "component functions" where $f_i(x)$ is associated to the local data of agent i and is only accessible by the agent i. Clearly, different choices of the log-likelihood function and therefore the component functions result in different problems. In particular, this framework covers many Bayesian inference problems such as Bayesian linear regression (Hoff, 2009), Bayesian logistic regression (Hoff, 2009), Bayesian principal component analysis (Dubey et al., 2016) or Bayesian deep learning (Wang and Yeung, 2016; Polson and Sokolov, 2017).

Let $x_i^{(k)}$ denote the local variable of node i at iteration k. The decentralized SGLD (DE-SGLD) algorithm (previously considered in Swenson et al. (2020) in the non-convex global optimization setting) consists of a weighted averaging with the local variables $x_j^{(k)}$ of node i's immediate neighbors $j \in \Omega_i := \{j : (i, j) \in \mathcal{G}\}$ as well as a stochastic gradient step over the node's component function $f_i(x)$, i.e.

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

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 (see (12) for more details). 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 minibatch 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. When gradients are deterministic, DE-SGLD algorithm reduces to the decentralized Langevin algorithm previously considered and studied in Kungurtsev (2020). Without the Gaussian noise, the iterations are also

equivalent to the decentralized stochastic gradient algorithm (Swenson et al., 2020a; Fallah et al., 2019) which has its origins in the decentralized gradient descent (DGD) methods introduced in Nedic and Ozdaglar (2009).

Contributions. In this paper, our contributions can be summarized as follows:

First, we give non-asymptotic performance guarantees for DE-SGLD when each of the components $f_i(x)$ is smooth and strongly convex in which case the target distribution has density $\pi(x) \propto e^{-f(x)}$ that is strongly log-concave (i.e. f is strongly convex) and f is smooth. More specifically, we provide an explicit upper bound on the Wasserstein distance between the target distribution $\pi(x)$ and the distribution of the iterate $x_i^{(k)}$ of node i. Our results 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 Wasserstein metric with a properly chosen stepsize. We also provide explicit bounds on the size of this neighborhood as a function of the noise level σ^2 in the stochastic gradients, the number of agents N and the dimension d. 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 DE-SGLD as a perturbed version of the Euler-Maruyama discretization of the overdamped Langevin diffusion (properly defined in Section 2) and use the fact that this diffusion admits the posterior distribution with density $\pi(x) \propto e^{-f(x)}$ as the stationary distribution where the perturbation effect is due to the stochasticity of the gradients and due to the "network effect" where agents are only able to communicate with their immediate neighbours. For achieving the results, we first derive a uniform L_2 bound on the gradients (Lemma 6) as well as a uniform L_2 bound on the deviation of the iterates $\bar{x}_i^{(k)}$ from their mean $\bar{x}^{(k)}$ over the agents (Lemma 7). Then, we derive an L^2 bound on the error between the average of gradients $\frac{1}{N}\sum_{i=1}^N \nabla f_i\left(x_i^{(k)}\right)$ and the scaled gradient of the average $\frac{1}{N}\nabla f\left(\bar{x}^{(k)}\right)$ (Lemma 8). Finally, we control the error between the mean iterates and the discretization of the overdamped diffusion (Lemma 9) and build on the existing results which characterizes the Wasserstein distance between the overdamped diffusion and its discretization. Putting everything together, we obtain our main result for DE-SGLD (Theorem 2).

Second, we propose a new algorithm decentralized SGHMC (DE-SGHMC) which can be viewed as the decentralized version of the SGHMC algorithm. In centralized settings, it is known that SGHMC algorithm can be faster than the SGLD algorithm both in practice and in theory (Gao et al., 2018; Chen et al., 2014). The underlying reason is that SGHMC is based on a discretization of the (underdamped) inertial Langevin diffusion which can converge to its equilibrium faster than the overdamped diffusion due to a momentum-based acceleration effect (Gao et al., 2018; Eberle et al., 2019). This effect is analogous to the fact that momentum-based optimization methods can accelerate gradient descent (Polyak, 1987; Nesterov, 1983; Su et al., 2016). We show that with proper choice of the stepsize and momentum parameters, the distribution of the DE-SGHMC iterates $x_i^{(k)}$ will converge to a neighborhood of the posterior distribution $\pi(x)$ linearly (in k) in the Wasserstein metric (Theorem 12). To our knowledge, these are first non-asymptotic performance guarantees for SGHMC methods in the decentralized setting. The approach we take is analogous to our analysis of the DE-SGLD however obtaining stability (uniform L_2) bounds on the iterates requires significantly more work. For this purpose, we develop a novel analysis where we show that the DE-SGHMC iterates can be viewed as a noisy version of Polyak's

(deterministic) heavy-ball method (Polyak, 1987); the noise comes from stochasticity of the gradients (which is proportional to the stepsize η), the injected Gaussian noise (which is proportional to $\sqrt{\eta}$) and the network effect where iterates can only access information from their neighbors. When the stepsize η is sufficiently small, the Gaussian noise dominates the stochastic gradient noise and therefore existing analysis for stochastic heavy-ball methods (Can et al., 2019b; Kuru et al., 2020) are not directly applicable to obtain stability estimates (i.e. uniform L_2 bounds) when η is sufficiently small. Our analysis relies on a careful choice of the Lyapunov function and obtaining sufficient conditions on the stepsize and the momentum parameters of the DE-SGHMC algorithm to guarantee stability (Lemma 14). As a by-product, our results contribute to the growing literature about the stability of the heavy-ball methods where it has been observed repeatedly that optimization methods such as heavy-ball and Nesterov's accelerated gradient methods are more sensitive to noise in the iterations compared to gradient descent methods (Aybat et al., 2020; Fallah et al., 2019; Flammarion and Bach, 2015; Devolder et al., 2014; Can et al., 2019b). Recent literature focused on the amount of noise heavy-ball methods can tolerate before they diverge and on their convergence rate subject to noise and perturbations (Can et al., 2019b; Flammarion and Bach, 2015; Liu et al., 2020; Kuru et al., 2020). Our analysis in the proof of Lemma 14 provides sufficient conditions for noisy heavy-ball iterations to be stable when subject to noise that is on the order of the square root of the stepsize.

Finally, we provide numerical experiments that illustrate our theory and showcase the practical performance of the DE-SGLD and DE-SGHMC algorithms: We show on Bayesian linear regression and Bayesian logistic regression tasks that our method allows each agent to sample from the posterior distribution efficiently without communicating local data.

Related literature. Decentralized optimization has been studied in the literature in the last few decades, at least going back to the seminal works of Bertsekas and Tsitsiklis (1989); Tsitsiklis (1984) which studied minimization of objective functions when the parameter vector can be decentralized. There has also been a growing literature and a lot of recent interest on decentralized optimization with first-order methods for both deterministic and stochastic optimization; See e.g. Swenson et al. (2020b); Fallah et al. (2019); Arjevani et al. (2020); Can et al. (2019a); Pu et al. (2020) and also the surveys Nedic (2020); Yang et al. (2019). Among the papers published in this area, Swenson et al. (2020); Swenson et al. (2020a) are most relevant to our paper, where the authors study a class of algorithms including DE-SGLD and show that DE-SGLD iterates with a particular decaying stepsize schedule converge in probability to the set of global minima for non-convex objectives under some assumptions. Momentum-based acceleration techniques based on heavy-ball method (Xin and Khan, 2020) and Nesterov's accelerated gradient method have also been studied for solving optimization problems in the decentralized setting (Fallah et al., 2019; Arjevani et al., 2020; Qu and Li, 2016; Xu et al., 2020), we refer the readers to Nedic (2020) for a survey in decentralized optimization. However, these papers are focused on solving optimization problems and the results do not apply to our setting where we are interested in sampling from the posterior distribution.

There are also a number of papers for distributed Bayesian inference based on data-parallel MCMC algorithms (Ge et al., 2015; Neiswanger et al., 2014; Xu et al., 2014; Scott et al., 2016; Rabinovich et al., 2015; Scott, 2017; Rendell et al., 2020; Ahn et al., 2015) where the computations are parallelized in a distributed computing environment, however

these papers are not applicable to the decentralized setting either. The variational inference methods which approximate the posterior distribution with a tractable distribution in the exponential family can be applied in the decentralized setting (Campbell and How, 2014; Lalitha et al., 2019) where agents average the parameters of their local parametrized distribution that estimates the posterior distribution, however to our knowledge, convergence rate guarantees to a posterior distribution for such approaches in the decentralized setting are not provided except the special case when the posterior distribution is in the exponential family (Lalitha et al., 2019). There are also other parallel MCMC techniques (Wang and Dunson, 2013; Neiswanger et al., 2014; Wang et al., 2015; Chowdhury and Jermaine, 2018; Nishihara et al., 2014) which require a central node to aggregate the samples generated at each computational node to estimate the posterior distribution; these methods are also not directly applicable to the decentralized setting.

Finally, very recently Kungurtsev (2020) showed that in the special case when the gradients are deterministic (i.e. when $\sigma = 0$), DE-SGLD algorithm converges to the target distribution $\pi(x)$ with rate $\mathcal{O}(\frac{1}{\sqrt{k}})$ for decaying stepsize $\alpha_k = \frac{1}{k}$ in the Wasserstein metric for strongly convex and smooth f with bounded gradients. Since strongly convex functions on $\mathbb{R}^d$ cannot have bounded gradients, these results are not applicable to problems we consider in this paper. In a concurrent work, Parayil et al. (2020) studied a Bayesian learning algorithm based on the decentralized Langevin dynamics in a non-convex setting. They obtained theoretical convergence guarantees in KL-divergence and evaluated the proposed algorithm on a wide variety of machine learning tasks. In another recent work, Cadena et al. (2021) proposed a modified Langevin dynamics algorithm for sensor networks. This algorithm can be implemented in a decentralized manner, where each sensor communicates with a randomly selected subset of sensors either via direct links or via multi-hop mechanism. The authors also show that when the gradient of the logarithm of the target density is bounded and Lipschitz, the proposed algorithm converges to the true centralized posterior distribution for networks where the communication delays are bounded.

2. Preliminaries and Background

Langevin algorithms. *Langevin algorithms* are core MCMC methods in statistics that allow one to sample from a given density $\pi(x)$ of interest. The classical Langevin Monte Carlo algorithm is based on the *overdamped (or first-order) Langevin diffusion*; see e.g. Dalalyan (2017); Durmus and Moulines (2019, 2017); Dalalyan and Karagulyan (2019):

$$dX(t) = -\nabla f(X(t))dt + \sqrt{2}dW_t, \tag{3}$$

where $f : \mathbb{R}^d \to \mathbb{R}$ 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 (3) admits a unique stationary distribution with the density $\pi(x) \propto e^{-f(x)}$, also known as the *Gibbs distribution* (Pavliotis, 2014). For computational purposes, this diffusion is simulated by considering its discretization. Although various discretization schemes are proposed, Euler-Maruyama discretization is the simplest one:

$$x_{k+1} = x_k - \eta\nabla f(x_k) + \sqrt{2\eta}w_k\,, \tag{4}$$

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 (4) does not converge to the target π and has a bias that needs to be properly characterized to provide performance guarantees (Dalalyan and Karagulyan, 2019).[1] There has been growing recent interest in the non-asymptotic analysis of discretized Langevin diffusions (4), motivated by applications to large-scale data analysis and Bayesian inference. The discretized Langevin diffusions admit convergence guarantees to a stationary distribution in a variety of metrics and under various assumptions on f; see e.g. Dalalyan (2017); Durmus and Moulines (2017, 2019); Bubeck et al. (2015); Cheng and Bartlett (2018); Erdogdu and Hosseinzadeh (2020); Dalalyan and Karagulyan (2019); Barkhagen et al. (2021); Raginsky et al. (2017); Xu et al. (2018); Chau et al. (2019); Zhang et al. (2019).

On the other hand, one can also design sampling algorithms based on the underdamped (a.k.a. inertial or kinetic) Langevin diffusion given by the SDE; see e.g. Cheng et al. (2018); Cheng et al. (2018); Dalalyan and Riou-Durand (2020); Gao et al. (2018, 2020); Ma et al. (2021); Akyildiz and Sabanis (2020); Cao et al. (2019); Zou et al. (2019):

$$dV(t) = -\gamma V(t)dt - \nabla f(X(t))dt + \sqrt{2\gamma}dW_t, \tag{5}$$

$$dX(t) = V(t)dt, \tag{6}$$

where $\gamma > 0$ is the friction coefficient, $X(t), V(t) \in \mathbb{R}^d$ models the position and the momentum of a particle moving in a field of force (described by the gradient of f) plus a random (thermal) force described by the Brownian noise, and W_t is a standard d-dimensional Brownian motion that starts at zero at time zero. It is known that under some mild assumptions on f, the Markov process $(X(t), V(t))_{t\geq 0}$ is ergodic and admits a unique stationary distribution π with density $\pi(x, v) \propto \exp\big(-\big(\frac{1}{2}\|v\|^2 + f(x)\big)\big)$ (Pavliotis, 2014). Hence, the x-marginal distribution of the stationary distribution with the density $\pi(x, v)$ is exactly the invariant distribution of the overdamped Langevin diffusion. For approximate sampling, various discretization schemes of (5)-(6) have been used in the literature; see e.g. Cheng et al. (2018); Teh et al. (2016); Chen et al. (2016a, 2015).

Decentralized setting. Agents are connected over a network $\mathcal{G} = (V, E)$ where $f_i : \mathbb{R}^d \to \mathbb{R}$ is the local objective of the agent i and we assume $\mathcal{G}$ is connected. Agents can only communicate with immediate neighbors using links defined by the edge set $\mathcal{E}$. We associate this network with an $N \times N$ symmetric, doubly stochastic[2] weight matrix W. We have $W_{ij} = W_{ji} > 0$ if $\{i, j\} \in E$ and $i \neq j$, and $W_{ij} = W_{ji} = 0$ if $\{i, j\} \notin E$ and $i \neq j$, and finally $W_{ii} = 1 - \sum_{j\neq i} W_{ij} > 0$ for every $1 \leq i \leq N$. The eigenvalues of W ordered in a descending manner satisfy:

$$1 = \lambda_1^W > \lambda_2^W \geq \cdots \geq \lambda_N^W > -1\,, \tag{7}$$

1. In principle, Metropolis-Hasting correction step can be employed to correct for the discretization errors, however for large-scale datasets, this correction step is computationally expensive and thus it is often not employed (Dalalyan and Riou-Durand, 2020; Dalalyan and Karagulyan, 2019; Teh et al., 2016). For this reason, we will not consider Metropolis-Hasting steps in our algorithms and analyses.
2. A square matrix $A \in \mathbb{R}^{N\times N}$ is called *doubly stochastic* if its entries A_{ij} are non-negative and if its rows and columns all sum up to 1, i.e. if $\sum_{j=1}^N A_{ij} = 1$ for all $i = 1, 2, \ldots, N$ and $\sum_{i=1}^N A_{ij} = 1$ for all $j = 1, 2, \ldots, N$.

with $W\mathbf{1} = \mathbf{1}$ where $\mathbf{1}$ is a vector of length N with each entry equal to one. For any connected $\mathcal{G}$, there is always a choice of W that satisfies the eigenvalue conditions (Can et al., 2019b; Boyd et al., 2006). A possible choice is the Metropolis weights (Xiao et al., 2006; Olshevsky, 2017) where $W_{ij} = \frac{1}{\max(d_i,d_j)}$ if $(i,j) \in \mathcal{E}$ where d_i is the degree (number of neighbors) of the node i. The mixing matrix W can also be chosen in many other ways (Can et al., 2019a; Boyd et al., 2006). In this paper, we will assume that W is given and fixed.

Our objective is to sample from a target distribution with density $\pi(x) \propto e^{-f(x)}$ on $\mathbb{R}^d$ where

$$f(x) := \sum_{i=1}^{N} f_i(x). \tag{8}$$

The agents can only pass vectors between their neighbors (not matrices) as the communication is typically more expensive than local computations in modern applications (Woodruff and Zhang, 2017). Throughout this paper, we assume $f_i \in \mathcal{S}_{\mu,L}(\mathbb{R}^d)$ for every $i = 1, 2, \ldots, N$,[3] where $\mathcal{S}_{\mu,L}(\mathbb{R}^d)$ denotes the set of functions from $\mathbb{R}^d$ to $\mathbb{R}$ that are μ-strongly convex and L-smooth, that is, for any $g \in \mathcal{S}_{\mu,L}(\mathbb{R}^d)$, for every $x, y \in \mathbb{R}^d$,

$$\frac{L}{2}\|x-y\|^2 \geq g(x) - g(y) - \nabla g(y)^T(x-y) \geq \frac{\mu}{2}\|x-y\|^2. \tag{9}$$

Wasserstein distance. Define $\mathcal{P}_2(\mathbb{R}^d)$ as the space consisting of all the Borel probability measures ν on $\mathbb{R}^d$ with the finite 2nd moment (based on the Euclidean norm). For any two Borel probability measures $\nu_1, \nu_2 \in \mathcal{P}_2(\mathbb{R}^d)$, the 2-Wasserstein distance $\mathcal{W}_2$ (see e.g. Villani (2009)) is defined as: $\mathcal{W}_2(\nu_1,\nu_2) := \left(\inf \mathbb{E}\left[\|Z_1 - Z_2\|^2\right]\right)^{1/2}$, where the infimum is taken over all joint distributions of the random variables Z_1, Z_2 with marginal distributions ν_1, ν_2 respectively.

Notations. For two matrices $A \in \mathbb{R}^{m\times n}$ and $B \in \mathbb{R}^{p\times q}$, we denote their Kronecker product by $A \otimes B$. We use I_d to denote the $d \times d$ identity matrix; if the dimension d is clear from the context we will also use I to denote the identity matrix. We denote $x_* \in \mathbb{R}^d$ as the (unique) minimizer of $f \in \mathcal{S}_{\mu,L}$ defined in (8). Moreover, we also denote

$$x^* = \left[x_*^T, x_*^T, \ldots, x_*^T\right]^T \in \mathbb{R}^{Nd}. \tag{10}$$

For any random variable X, we use $\mathcal{L}(X)$ to denote the probability distribution of X. We say that the distribution $\pi(x) \propto e^{-f(x)}$ is strongly log-concave if $f(x)$ is μ-strongly convex for some $\mu > 0$. Given two functions $g(x)$ and $h(x)$ defined on a subset $\mathcal{D}$ of real numbers, we say $h(x) = \mathcal{O}(g(x))$ as $x \to a$ if there exist positive numbers δ and M such that for all $x \in \mathcal{D}$ with $0 < |x-a| < \delta$, we have $|f(x)| \leq Mg(x)$ whereas we say $h(x) = \Theta(g(x))$ if there exist positive numbers δ and M_1, M_2 such that for all $x \in \mathcal{D}$ with $0 < |x-a| < \delta$, we have $M_1 g(x) \leq |f(x)| \leq M_2 g(x)$. The dependency to the point a will be omitted if it is clear from the context. Given real scalars x, y, we consider the ratio $h(x,y) := \frac{x^k - y^k}{x-y}$ with the convention that $h(y,y) := \lim_{x\to y} h(x,y) = ky^{k-1}$.

3. Our results in this paper would also hold if $f_i \in \mathcal{S}_{\mu_i,L_i}(\mathbb{R}^d)$ and one considers $\mu = \min_i \mu_i$ and $L = \max_i L_i$ in our main theorems.

3. Decentralized Stochastic Gradient Langevin Dynamics

We recall from (2) that decentralized stochastic gradient Langevin dynamics (DE-SGLD) are based on stochastic estimates $\tilde{\nabla} f_i(x)$ of the actual gradients $\nabla f_i(x)$. We make the following assumption throughout this paper regarding the stochastic estimates $\tilde{\nabla} f_i(x)$ which basically says that the gradient error is unbiased with a finite variance. This is a common assumption in the literature for analyzing stochastic optimization and stochastic-gradient MCMC algorithms; see e.g. Dalalyan and Riou-Durand (2020); Chen et al. (2016b); Liu et al. (2020).

Assumption 1 *Let $x_i^{(k)}$ denote the local variable of node i at iteration k. At iteration k, node i has access to $\tilde{\nabla} f_i\left(x_i^{(k)}, z_i^{(k)}\right)$ where $z_i^{(k)}$ is a random variable independent of $\{z_j^{(t)}\}_{j=1,\dots,N,t=1,\dots,k-1}$ and $\{z_j^{(k)}\}_{j\neq i}$. To simplify the notation, we suppress the $z_i^{(k)}$ dependency and let $\tilde{\nabla} f_i\left(x_i^{(k)}\right)$ denote $\tilde{\nabla} f_i\left(x_i^{(k)}, z_i^{(k)}\right)$. We assume the* gradient noise *defined as*

$$\xi_i^{(k+1)} := \tilde{\nabla} f_i\left(x_i^{(k)}\right) - \nabla f_i\left(x_i^{(k)}\right), \tag{11}$$

is unbiased with a finite second moment, i.e.,

$$\mathbb{E}\left[\xi_i^{(k+1)} \middle| \mathcal{F}_k\right] = 0, \qquad \mathbb{E}\left\|\xi_i^{(k+1)}\right\|^2 \leq \sigma^2, \tag{12}$$

where $\mathcal{F}_k$ is the natural filtration of the iterates $x_i^{(k)}$ up to (and including) time k.

Based on (11), we rewrite the DE-SGLD iterations (2) in terms of the gradient noise $\xi_i^{(k+1)}$ as

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

where $\eta > 0$ is the stepsize, $w_i^{(k)}$ are i.i.d. Gaussian noise with mean 0 and covariance being identity matrices and $\Omega_i = \{j : (i,j) \in \mathcal{G}\}$ are the neighbors of the node i.[4] By defining the column vector

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

which concetenates the local decision variables into a single vector, we can express the DE-SGLD iterations further as

$$x^{(k+1)} = \mathcal{W} x^{(k)} - \eta \nabla F\left(x^{(k)}\right) - \eta \xi^{(k+1)} + \sqrt{2\eta} w^{(k+1)}, \quad \text{with} \quad \mathcal{W} = W \otimes I_d, \tag{13}$$

where we recall that $\otimes$ denotes the Kronecker product, $F : \mathbb{R}^{Nd} \to \mathbb{R}$ is defined as

$$F(x) := F(x_1, \dots, x_N) = \sum_{i=1}^N f_i(x_i), \tag{14}$$

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

and

$$w^{(k+1)} := \left[\left(w_1^{(k+1)}\right)^T, \left(w_2^{(k+1)}\right)^T, \ldots, \left(w_N^{(k+1)}\right)^T\right]^T$$

are i.i.d. Gaussian noise with mean 0 and with a covariance matrix given by the identity matrix. The vectors

$$\xi^{(k+1)} := \left[\left(\xi_1^{(k+1)}\right)^T, \left(\xi_2^{(k+1)}\right)^T, \ldots, \left(\xi_N^{(k+1)}\right)^T\right]^T$$

are the gradient noise so that

$$\mathbb{E}\left[\xi^{(k+1)}\middle|\mathcal{F}_k\right] = 0, \qquad \mathbb{E}\left\|\xi^{(k+1)}\right\|^2 \leq \sigma^2 N. \tag{15}$$

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

$$\bar{x}^{(k+1)} = \bar{x}^{(k)} - \eta\frac{1}{N}\sum_{i=1}^N \nabla f_i\left(x_i^{(k)}\right) - \eta\bar{\xi}^{(k+1)} + \sqrt{2\eta}\bar{w}^{(k+1)}, \tag{16}$$

where

$$\bar{w}^{(k+1)} := \frac{1}{N}\sum_{i=1}^N w_i^{(k+1)} \sim \frac{1}{\sqrt{N}}\mathcal{N}(0, I_d), \qquad \bar{\xi}^{(k+1)} := \frac{1}{N}\sum_{i=1}^N \xi_i^{(k+1)}, \tag{17}$$

that satisfies

$$\mathbb{E}\left[\bar{\xi}^{(k+1)}\middle|\mathcal{F}_k\right] = 0, \qquad \mathbb{E}\left\|\bar{\xi}^{(k+1)}\right\|^2 \leq \frac{\sigma^2}{N}. \tag{18}$$

We now state the main result of this section, which bounds the average of $\mathcal{W}_2$ distance between the distribution of $x_i^{(k)}$ and the target distribution π (that has a density proportional to $\exp(-f(x))$) over $1 \leq i \leq N$. This result provides also a bound on the $\mathcal{W}_2$ distance of the node averages $\bar{x}^{(k)}$ and the target distribution π. To facilitate the presentation, we define the second largest magnitude of the eigenvalues of W as

$$\bar{\gamma} := \max\left\{\left|\lambda_2^W\right|, \left|\lambda_N^W\right|\right\} \in [0, 1), \tag{19}$$

which is related to the connectivity of the graph $\mathcal{G}$. For instance, consider Metropolis weights where $W_{ij} = \frac{1}{\max(d_i, d_j)}$ if $(i, j) \in \mathcal{E}$ where d_i is the degree (number of neighbors) of the node i with the convention that each node is a neighbor of itself. In this case, for complete graphs with N nodes where each node is connected to all the other nodes, we have $\bar{\gamma} = 0$ whereas for a circular graph with N nodes we have $d_i = 3$ for every i and $\bar{\gamma} = \frac{1}{3} + \frac{2}{3}\cos(\frac{2\pi}{N}) = 1 - \mathcal{O}(\frac{1}{N})$ (see Chung and Graham (1997, Example 1.1 and Example 1.5)).

Theorem 2 *Assume* $\mathbb{E}\|x^{(0)}\|^2 < \infty$ *and* $\eta \in \left(0, \bar{\eta}\right)$ *where* $\bar{\eta} := \min(\frac{1+\lambda_N^W}{L}, \frac{1}{L+\mu})$. *Then, for every* k, *DE-SGLD iterates* $x_i^{(k)}$ *given by* (2) *and their average* $\bar{x}^{(k)}$ *satisfy*

$$\mathcal{W}_2\left(\mathcal{L}\left(\bar{x}^{(k)}\right), \pi\right) \leq (1-\mu\eta)^k \left(\left(\mathbb{E}\|\bar{x}^{(0)} - x_*\|^2\right)^{1/2} + \sqrt{2\mu^{-1}dN^{-1}}\right) + \left(\bar{\gamma}^2 \frac{\left(1-\eta\mu\left(1-\frac{\eta L}{2}\right)\right)^k - \bar{\gamma}^{2k}}{\left(1-\eta\mu\left(1-\frac{\eta L}{2}\right)\right) - \bar{\gamma}^2}\right)^{1/2} \frac{2L}{\sqrt{N}} \left(\mathbb{E}\left\|x^{(0)}\right\|^2\right)^{1/2} + \sqrt{\eta} E_1,$$

and

$$E_1 := \frac{1.65L}{\mu}\sqrt{dN^{-1}} + \frac{\sigma}{\sqrt{\mu(1-\frac{\eta L}{2})N}} + \left(\frac{\eta}{\mu(1-\frac{\eta L}{2})} + \frac{(1+\eta L)^2}{\mu^2(1-\frac{\eta L}{2})^2}\right)^{1/2} \cdot \left(\frac{4L^2D^2\eta}{N(1-\bar{\gamma})^2} + \frac{4L^2\sigma^2\eta}{(1-\bar{\gamma}^2)} + \frac{8L^2d}{(1-\bar{\gamma}^2)}\right)^{1/2}.$$

Furthermore,

$$\begin{aligned}
&\frac{1}{N}\sum_{i=1}^N \mathcal{W}_2\left(\mathcal{L}\left(x_i^{(k)}\right), \pi\right) \\
&\leq (1-\mu\eta)^k \left(\left(\mathbb{E}\|\bar{x}^{(0)} - x_*\|^2\right)^{1/2} + \sqrt{2\mu^{-1}dN^{-1}}\right) + \frac{2\bar{\gamma}^k}{\sqrt{N}}\left(\mathbb{E}\left\|x^{(0)}\right\|^2\right)^{1/2} \\
&\quad + \left(\bar{\gamma}^2 \frac{\left(1-\eta\mu\left(1-\frac{\eta L}{2}\right)\right)^k - \bar{\gamma}^{2k}}{\left(1-\eta\mu\left(1-\frac{\eta L}{2}\right)\right) - \bar{\gamma}^2}\right)^{1/2} \frac{2L}{\sqrt{N}} \left(\mathbb{E}\left\|x^{(0)}\right\|^2\right)^{1/2} + \sqrt{\eta}E_2 + \eta E_3, \qquad (20)
\end{aligned}$$

with $E_2 := E_1 + \frac{2\sqrt{2d}}{\sqrt{1-\bar{\gamma}^2}}$ *and* $E_3 := \frac{2D}{\sqrt{N}(1-\bar{\gamma})} + \frac{2\sigma}{\sqrt{1-\bar{\gamma}^2}}$, *where* x_* *is the minimizer of* f, $\bar{x}^{(0)} = \frac{1}{N}\sum_{i=1}^N x_i^{(0)}$, D *is defined in* (26), $\mathcal{L}\left(x_i^{(k)}\right)$ *denotes the law of* $x_i^{(k)}$ *and* π *is the Gibbs distribution with probability density function proportional to* $\exp(-f(x))$.

Remark 3 *We observe that in the setting of Theorem 2, the asymptotic error with respect to the target distribution in 2-Wasserstein satisfies* $\limsup_{k\to\infty} \mathcal{W}_2\left(\mathcal{L}\left(\bar{x}^{(k)}\right), \pi\right) = \mathcal{O}\left(\sqrt{\eta}\right)$, *where* $\mathcal{O}(\cdot)$ *hides other constants (*d, μ, L, σ, N *and* $\bar{\gamma}$*). This shows that the asymptotic error can be made arbitrarily smaller by choosing* $\eta > 0$ *small enough. In particular, for sufficiently small* η, *it is easy to check that* $\left(1-\eta\mu\left(1-\frac{\eta L}{2}\right)\right) \geq \bar{\gamma}^2$ *and consequently from Theorem 2,*

$$\mathcal{W}_2\left(\mathcal{L}\left(\bar{x}^{(2K)}\right), \pi\right) \leq \left(1-\eta\mu\left(1-\frac{\eta L}{2}\right)\right)^K \psi_0 + \mathcal{O}\left(\sqrt{\eta}\right) \qquad (21)$$

$$\leq e^{-\eta\mu\left(1-\frac{\eta L}{2}\right)K} a_0 + \mathcal{O}\left(\sqrt{\eta}\right) \qquad (22)$$

for some a_0 (that depends on the initialization $x^{(0)}, d, \mu, L, \sigma, \bar{\gamma}$ and N) where K is the iteration budget. Given K, if we choose $\eta = \frac{c}{\mu K}$ for some constant c, then the right-hand side of (21) becomes $\Theta(1)$ as $K \to \infty$; this is because $(1 - \frac{c}{K})^K \to e^{-c} = \Theta(1)$ as $K \to \infty$ for any constant $c > 0$. This is not desirable, as ideally, we want the Wasserstein error bound (right-hand side of (21)) go to zero if the iteration budget $K \to \infty$. This can be achieved by choosing a stepsize such as $\eta = \frac{c \log \sqrt{K}}{\mu K}$ for a constant $c > 1$. Then, given $c > 1$ fixed, if K is large enough satisfying $K \geq \bar{K}$ with $\bar{K} = \max(e, \frac{a^2}{e})$ where $a := \frac{c(L+\mu)}{2\mu(1+\lambda)}$, then the stepsize $\eta = \frac{c \log \sqrt{K}}{\mu K}$ satisfies the assumptions of Theorem 2 (this follows simply from the inequality $\log(K) \leq 1 + \frac{K-e}{\sqrt{eK}}$ for $K \geq e$). Consequently, from (22), we obtain

$$\mathcal{W}_2\left(\mathcal{L}\left(\bar{x}^{(2K)}\right), \pi\right) = \mathcal{O}\left(\frac{1}{(\sqrt{K})^c} + \frac{\sqrt{c\log(K)}}{\sqrt{K}}\right) = \mathcal{O}\left(\frac{\sqrt{\log(K)}}{\sqrt{K}}\right), \tag{23}$$

where the last $\mathcal{O}(\cdot)$ term hides constants that depends on $x^{(0)}, d, \mu, L, \sigma, \bar{\gamma}, N$ and c. This shows that to sample from a distribution that is ε close to the target in the 2-Wasserstein distance, it suffices to have $\mathcal{O}(\frac{1}{\varepsilon^2})$ iterations of DE-SGLD, ignoring logarithmic factors. The appearance of logarithmic factors in the iteration complexity as well as in (23) is related to the fact that constant stepsize is used, and similar logarithmic factors also appear even in centralized SGLD methods with constant stepsize (see Dalalyan and Karagulyan (2019, Theorem 1 and Section 2)). It is possible to avoid the logarithmic terms by employing a time-varying stepsize similar to Dalalyan and Karagulyan (2019, Theorem 2).

Remark 4 *The upper bound given for the 2-Wasserstein distances to the target π in Theorem 2 is monotonically increasing in the parameter $\bar{\gamma}$. To see this, consider the function $H(x,y) := \frac{x^k - y^k}{x-y} y = \sum_{i=0}^{k-1} x^i y^{k-i}$ for $x \in (0,1), y \in (0,1)$ with the convention that $H(y,y) := ky^k$. For given x fixed, the partial derivative $\partial_y H(x,y) = \sum_{i=0}^{k-1}(k-i)x^i y^{k-1-i} > 0$. Therefore H is monotonically decreasing in y, so is the function $\sqrt{H}$. If we set $y = \bar{\gamma}^2$ and $x = 1 - \eta\mu(1 - \frac{\eta L}{2})$, the third term that appears in the bound (20) is an affine function of $\sqrt{H}$ and hence monotonically increasing in $\bar{\gamma}^2$ and in $\bar{\gamma}$. Finally, after a straightforward computation it can be seen that the remaining terms E_1, E_2 and E_3 that appear in the bound (20) are also monotonically increasing in $\bar{\gamma}$. It follows from this argument that closer $\bar{\gamma}$ to zero, better connectivity properties the network has (with $\bar{\gamma} = 0$ for complete graphs that are fully-connected) and the Wasserstein distance to the target becomes (smaller) better. Hence, roughly speaking, the parameter $\bar{\gamma}$ determines the additional cost of the distributed algorithm (i.e. increased bias and variance) when there is not full connectivity among the nodes.*

Remark 5 *In Assumption 1, we assumed that the variance of the gradient noise is bounded. It is a reasonable assumption in many applications including linear regressions with stochastic gradients estimated using minibatches, since one can show that if the stepsize $\eta > 0$ is small enough the variance of the gradients for DE-SGLD will stay bounded and satisfy our assumptions on the gradient noise (Assumption 1) with an analysis similar to Aybat et al. (2019, Section K). We will illustrate this point in detail in Appendix D.*

3.1 Proof of Theorem 2

To facilitate the analysis, let us define x_k from the iterates:

$$x_{k+1} = x_k - \eta \frac{1}{N} \nabla f(x_k) + \sqrt{2\eta} \bar{w}^{(k+1)}, \tag{24}$$

where $x_0 = \bar{x}_0 = \frac{1}{N}\sum_{i=1}^N x_i^{(0)}$ and $\bar{w}^{(k+1)}$ is defined in (17). This is an Euler-Maruyama discretization (with stepsize η) of the continuous-time overdamped Langevin diffusion:

$$dX_t = -\frac{1}{N}\nabla f(X_t)dt + \sqrt{2N^{-1}}dW_t, \tag{25}$$

where W_t is a standard d-dimensional Brownian motion.

To bound the average of $\mathcal{W}_2$ distance between $\mathcal{L}\left(x_i^{(k)}\right)$ and π over $1 \le i \le N$, 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 (mean) $\bar{x}^{(k)} = \frac{\sum_{i=1}^N x_i^{(k)}}{N}$ for $1 \le i \le N$; (2) the L^2 distance between the average iterate $\bar{x}^{(k)}$ and iterates x_k obtained from Euler-Maruyama discretization of overdamped SDE; and (3) the $\mathcal{W}_2$ distance between between $\mathcal{L}(x_k)$ and π, i.e. the convergence of Euler-Maruyama discretization of the overdamped SDE. The next subsections are devoted to controlling each of these three terms.

3.1.1 Uniform L^2 bounds between $x_i^{(k)}$ and their average

We first state a key lemma which provides L^2 bounds on the gradients $\nabla F\left(x^{(k)}\right)$ that are uniform in k, where F is defined in (14). Recall from (10) that $x_* \in \mathbb{R}^d$ denotes the unique minimizer of $f(x)$, and $x^* = \left[x_*^T, x_*^T, \ldots, x_*^T\right]^T$ is an Nd-dimensional vector. We view DE-SGLD as a decentralized gradient descent (DGD) method subject to stochastic gradient and Gaussian noise, and our analysis is inspired by the proof techniques of Yuan et al. (2016) for analyzing DGD methods. The proof of this lemma is provided in the Appendix.

Lemma 6 *Under the assumptions of Theorem 2, we have,*

$$\mathbb{E}\left\|\nabla F\left(x^{(k)}\right)\right\|^2 \le D^2, \quad \textit{for any } k,$$

where

$$D^2 := 4L^2\mathbb{E}\left\|x^{(0)} - x^*\right\|^2 + 8L^2\frac{C_1^2\eta^2 N}{(1-\bar{\gamma})^2} + \frac{2L^2(\eta\sigma^2 N + 2dN)}{\mu(1+\lambda_N^W - \eta L)} + 4\left\|\nabla F\left(x^*\right)\right\|^2. \tag{26}$$

Here, $x^* \in \mathbb{R}^{Nd}$ is given in (10)*, $\bar{\gamma}$ is defined by* (19) *and*

$$C_1 := \bar{C}_1 \cdot \left(1 + \frac{2(L+\mu)}{\mu}\right), \textit{ where } \quad \bar{C}_1 := \sqrt{2L\sum_{i=1}^N \left(f_i(0) - f_i^*\right)}, \quad f_i^* := \min_{x\in\mathbb{R}^d} f_i(x). \tag{27}$$

It is clear from the DE-SGLD iterations that the deviations between the iterates $x_i^{(k)}$ and their means $\bar{x}^{(k)}$ depend on the magnitude of the gradients $\nabla F(x^{(k)})$, the stepsize as well as the magnitude of the injected Gaussian noise. Building on Lemma 6 which gives us a control over the second moment of the gradients, in the next result we provide uniform L_2 bounds between the iterates $x_i^{(k)}$ and their means. The proof can be found in the Appendix.

Lemma 7 *Under the assumptions of Theorem 2, for any k, we have*

$$\sum_{i=1}^{N} \mathbb{E}\left\|x_i^{(k)} - \bar{x}^{(k)}\right\|^2 \leq 4\bar{\gamma}^{2k}\mathbb{E}\left\|x^{(0)}\right\|^2 + \frac{4D^2\eta^2}{(1-\bar{\gamma})^2} + \frac{4\sigma^2 N\eta^2}{(1-\bar{\gamma}^2)} + \frac{8dN\eta}{(1-\bar{\gamma}^2)},$$

where D is defined in (26) *and $\bar{\gamma}$ is given in* (19).

Note that we can deduce from (16) that

$$\bar{x}^{(k+1)} = \bar{x}^{(k)} - \eta\frac{1}{N}\nabla f\left(\bar{x}^{(k)}\right) + \eta\mathcal{E}_{k+1} - \eta\bar{\xi}^{(k+1)} + \sqrt{2\eta}\bar{w}^{(k+1)}, \tag{28}$$

where

$$\mathcal{E}_{k+1} := \frac{1}{N}\sum_{i=1}^{N}\left[\nabla f_i\left(\bar{x}^{(k)}\right) - \nabla f_i\left(x_i^{(k)}\right)\right]. \tag{29}$$

We observe that the average iterate $\bar{x}^{(k)}$ in (28) follows a gradient descent dynamics subject to gradient errors and Gaussian noise, if we view $\mathcal{E}_k$ as a gradient error term. Since ∇f_i is Lipschitz by our assumptions, the gradient error (29) can be controlled based on Lemma 7. In particular, as a corollary of Lemma 7, we obtain the following result; the proof is given in the Appendix for the sake of completeness.

Lemma 8 *Under the assumptions of Theorem 2, for any k, we have*

$$\mathbb{E}\left\|\mathcal{E}_{k+1}\right\|^2 \leq \frac{4L^2\bar{\gamma}^{2k}}{N}\mathbb{E}\left\|x^{(0)}\right\|^2 + \frac{4L^2D^2\eta^2}{N(1-\bar{\gamma})^2} + \frac{4L^2\sigma^2\eta^2}{(1-\bar{\gamma}^2)} + \frac{8L^2d\eta}{(1-\bar{\gamma}^2)},$$

where $\mathcal{E}_{k+1}$ is defined in (29).

3.1.2 L^2 distance between the mean and the discretized overdamped SDE

Recall the iterates x_k defined in (24) which is an Euler-Maruyama discretization of the continuous-time overdamped Langevin SDE in (25) with stepsize η, and the mean $\bar{x}^{(k)}$ in (28). Since the L^2 bound of the error term $\mathcal{E}_{k+1}$ can be controlled as in Lemma 8, we will show that the mean $\bar{x}^{(k)}$ and x_k are close to each other in L^2 distance. Indeed, we have the following estimate:

Lemma 9 *Under the assumptions of Theorem 2, for every k,*

$$\mathbb{E}\left\|\bar{x}^{(k)} - x_k\right\|^2 \leq \eta\left(\frac{\eta}{\mu(1-\frac{\eta L}{2})} + \frac{(1+\eta L)^2}{\mu^2(1-\frac{\eta L}{2})^2}\right)\left(\frac{4L^2D^2\eta}{N(1-\bar{\gamma})^2} + \frac{4L^2\sigma^2\eta}{(1-\bar{\gamma}^2)} + \frac{8L^2d}{(1-\bar{\gamma}^2)}\right)$$
$$+ \frac{\eta\sigma^2}{\mu(1-\frac{\eta L}{2})N} + \frac{\bar{\gamma}^{2k} - \left(1-\eta\mu\left(1-\frac{\eta L}{2}\right)\right)^k}{\bar{\gamma}^2 - 1 + \eta\mu\left(1-\frac{\eta L}{2}\right)}\frac{4L^2\bar{\gamma}^2}{N}\mathbb{E}\left\|x^{(0)}\right\|^2.$$

3.1.3 $\mathcal{W}_2$ distance between the iterates and the Gibbs distribution

Bounds on the $\mathcal{W}_2$ distance between the Euler-Maruyama discretization x_k of the over-damped Langevin diffusion and Gibbs distribution π has been established in the literature. We note that the function $\frac{1}{N}f$ is $\frac{\mu}{N}$-strongly convex and $\frac{L}{N}$-smooth, and we state Theorem 4 in Dalalyan and Karagulyan (2019) as follows.

Lemma 10 (Theorem 4 in Dalalyan and Karagulyan (2019)) *For any $\eta \in (0, \frac{2N}{L+\mu}]$, we have*

$$\mathcal{W}_2\left(\mathcal{L}(x_k), \pi\right) \leq (1-\mu\eta)^k \mathcal{W}_2\left(\mathcal{L}(x_0), \pi\right) + \frac{1.65L}{\mu}\sqrt{\eta d N^{-1}}.$$

The proof of this lemma is based on the so-called "synchronous coupling" technique to control the $\mathcal{W}_2$ distances, see Dalalyan and Karagulyan (2019) for details. Next, we bound the L^2 distance between the minimizer of f and Gibbs distribution π; the proof is provided in Appendix C.

Lemma 11 *Let x_* be the unique minimizer of $f(x)$. Then, we have $\mathbb{E}_{X\sim\pi}\|X - x_*\|^2 \leq \frac{2dN^{-1}}{\mu}$.*

Putting all the pieces together, the stage is set for the proof of Theorem 2.

3.1.4 Proof of Theorem 2

Since $x_0 = \frac{1}{N}\sum_{i=1}^N x_i^{(0)}$, we have $\mathbb{E}\|x_0\|^2 < \infty$. By Lemma 11,

$$\begin{aligned}\mathcal{W}_2\left(\mathcal{L}(x_0), \pi\right) &\leq \left(\mathbb{E}\|x_0 - x_*\|^2\right)^{1/2} + \left(\mathbb{E}_{X\sim\pi}\|X - x_*\|^2\right)^{1/2} \\ &\leq \left(\mathbb{E}\|x_0 - x_*\|^2\right)^{1/2} + \sqrt{2\mu^{-1}dN^{-1}}.\end{aligned}$$

Under our assumptions on the stepsize η, we have clearly $\eta \in (0, \frac{2N}{L+\mu}]$ as $N \geq 1$. Therefore Lemma 10 is applicable. More specifically, it follows from Lemma 10 that,

$$\mathcal{W}_2\left(\mathcal{L}(x_k), \pi\right) \leq (1-\mu\eta)^k \left(\left(\mathbb{E}\|x_0 - x_*\|^2\right)^{1/2} + \sqrt{2\mu^{-1}dN^{-1}}\right) + \frac{1.65L}{\mu}\sqrt{\eta d N^{-1}}.$$

Moreover, it follows from Lemma 9 that

$$\begin{aligned}&\mathcal{W}_2\left(\mathcal{L}\left(\bar{x}^{(k)}\right), \mathcal{L}(x_k)\right) \\ &\leq \left(\mathbb{E}\left\|\bar{x}^{(k)} - x_k\right\|^2\right)^{1/2} \\ &\leq \eta^{1/2}\left(\frac{\eta}{\mu(1-\frac{\eta L}{2})} + \frac{(1+\eta L)^2}{\mu^2(1-\frac{\eta L}{2})^2}\right)^{1/2} \cdot \left(\frac{4L^2D^2\eta}{N(1-\bar{\gamma})^2} + \frac{4L^2\sigma^2\eta}{(1-\bar{\gamma}^2)} + \frac{8L^2d}{(1-\bar{\gamma}^2)}\right)^{1/2} \\ &\quad + \frac{\sqrt{\eta}\sigma}{\sqrt{\mu(1-\frac{\eta L}{2})N}} + \left(\frac{\bar{\gamma}^{2k} - \left(1-\eta\mu\left(1-\frac{\eta L}{2}\right)\right)^k}{\bar{\gamma}^2 - 1 + \eta\mu\left(1-\frac{\eta L}{2}\right)}\right)^{1/2} \frac{2L\bar{\gamma}}{\sqrt{N}}\left(\mathbb{E}\left\|x^{(0)}\right\|^2\right)^{1/2}.\end{aligned}$$

Hence, we conclude that

$$\mathcal{W}_2\left(\mathcal{L}\left(\bar{x}^{(k)}\right),\pi\right) \leq (1-\mu\eta)^k\left(\left(\mathbb{E}\|\bar{x}^{(0)}-x_*\|^2\right)^{1/2}+\sqrt{2\mu^{-1}dN^{-1}}\right) + \left(\frac{\left(1-\eta\mu\left(1-\frac{\eta L}{2}\right)\right)^k-\bar{\gamma}^{2k}}{\left(1-\eta\mu\left(1-\frac{\eta L}{2}\right)\right)-\bar{\gamma}^2}\right)^{1/2}\frac{2L\bar{\gamma}}{\sqrt{N}}\left(\mathbb{E}\left\|x^{(0)}\right\|^2\right)^{1/2}+\sqrt{\eta}E_1, \tag{30}$$

with

$$E_1 := \frac{1.65L}{\mu}\sqrt{dN^{-1}}+\frac{\sigma}{\sqrt{\mu(1-\frac{\eta L}{2})N}} + \left(\frac{\eta}{\mu(1-\frac{\eta L}{2})}+\frac{(1+\eta L)^2}{\mu^2(1-\frac{\eta L}{2})^2}\right)^{1/2}\cdot\left(\frac{4L^2D^2\eta}{N(1-\bar{\gamma})^2}+\frac{4L^2\sigma^2\eta}{(1-\bar{\gamma}^2)}+\frac{8L^2d}{(1-\bar{\gamma}^2)}\right)^{1/2}.$$

Finally, by the Cauchy-Schwarz inequality,

$$\begin{aligned}\frac{1}{N}\sum_{i=1}^N\mathcal{W}_2\left(\mathcal{L}\left(x_i^{(k)}\right),\mathcal{L}\left(\bar{x}^{(k)}\right)\right) &\leq \sqrt{\frac{1}{N}\sum_{i=1}^N\mathcal{W}_2^2\left(\mathcal{L}\left(x_i^{(k)}\right),\mathcal{L}\left(\bar{x}^{(k)}\right)\right)}\\ &\leq \sqrt{\frac{1}{N}\sum_{i=1}^N\mathbb{E}\left\|x_i^{(k)}-\bar{x}^{(k)}\right\|^2}.\end{aligned} \tag{31}$$

Also, by Lemma 7, we have

$$\begin{aligned}\sqrt{\frac{1}{N}\sum_{i=1}^N\mathbb{E}\left\|x_i^{(k)}-\bar{x}^{(k)}\right\|^2} &\leq \left(\frac{4\bar{\gamma}^{2k}}{N}\mathbb{E}\left\|x^{(0)}\right\|^2+\frac{4D^2\eta^2}{N(1-\bar{\gamma})^2}+\frac{4\sigma^2\eta^2}{(1-\bar{\gamma}^2)}+\frac{8d\eta}{(1-\bar{\gamma}^2)}\right)^{1/2}\\ &\leq \frac{2\bar{\gamma}^k}{\sqrt{N}}\left(\mathbb{E}\left\|x^{(0)}\right\|^2\right)^{1/2}+\frac{2D\eta}{\sqrt{N}(1-\bar{\gamma})}+\frac{2\sigma\eta}{\sqrt{1-\bar{\gamma}^2}}+\frac{2\sqrt{2d\eta}}{\sqrt{1-\bar{\gamma}^2}}.\end{aligned}$$

The inequality (20) then follows from the triangular inequality for the 2-Wasserstein distance. This completes the proof. □

4. Decentralized Stochastic Gradient Hamiltonian Monte Carlo

We introduce the following algorithm which we call decentralized stochastic gradient Hamiltonian Monte Carlo (DE-SGHMC): For each agent $i = 1, \ldots, N$,

$$v_i^{(k+1)} = v_i^{(k)} - \eta\left[\gamma v_i^{(k)}+\tilde{\nabla}f_i\left(x_i^{(k)}\right)\right]+\sqrt{2\gamma\eta}w_i^{(k+1)}, \tag{32}$$

$$x_i^{(k+1)} = \sum_{j\in\Omega_i}W_{ij}x_j^{(k)}+\eta v_i^{(k+1)}, \tag{33}$$

starting from the initializations $x_i^{(0)}, v_i^{(0)} \in \mathbb{R}^d$, where $\eta > 0$ is the stepsize, $w_i^{(k+1)}$ are i.i.d. Gaussian noise with mean 0 and covariance being $d-$dimensional identity matrices. We note that in this section, we are abusing the notation for simplicity of the presentation and using $x_i^{(k)}$ to denote the DE-SGHMC iterates instead of DE-SGLD iterates. This algorithm is a natural adaptation of the SGHMC algorithm to the decentralized setting: If the term $\sum_{j\in\Omega_i} W_{ij} x_j^{(k)}$ is replaced by $x_i^{(k)}$, then the resulting dynamics at each node reduces to SGHMC which is a discretization of the underdamped Langevin diffusion given in (5)-(6) (see e.g. Gao et al. (2018)).

Note that the gradient noise $\xi_i^{(k+1)} := \tilde{\nabla} f_i(x_i^{(k)}) - \nabla f_i(x_i^{(k)})$ satisfies Assumption 1 so that $\xi^{(k+1)} := \left[(\xi_1^{(k+1)})^T, \ldots, (\xi_N^{(k+1)})^T\right]^T$ satisfies (15) and $\bar{\xi}^{(k+1)} := \frac{1}{N}\sum_{i=1}^N \xi_i^{(k+1)}$ satisfies (18). By defining the column vectors

$$x^{(k)} := \left[\left(x_1^{(k)}\right)^T, \left(x_2^{(k)}\right)^T, \ldots, \left(x_N^{(k)}\right)^T\right]^T \in \mathbb{R}^{Nd},$$
$$v^{(k)} := \left[\left(v_1^{(k)}\right)^T, \left(v_2^{(k)}\right)^T, \ldots, \left(v_N^{(k)}\right)^T\right]^T \in \mathbb{R}^{Nd},$$

where $v_i^{(k)}$ and $x_i^{(k)}$ satisfy (32)–(33), we can rewrite the DE-SGHMC as follows:

$$v^{(k+1)} = v^{(k)} - \eta\left[\gamma v^{(k)} + \nabla F\left(x^{(k)}\right) + \xi^{(k+1)}\right] + \sqrt{2\gamma\eta} w^{(k+1)}, \tag{34}$$
$$x^{(k+1)} = \mathcal{W} x^{(k)} + \eta v^{(k+1)}, \tag{35}$$

where $\mathcal{W} = W \otimes I_d$ and $F : \mathbb{R}^{Nd} \to \mathbb{R}$ is defined as $F(x) := F(x_1, \ldots, x_N) = \sum_{i=1}^N f_i(x_i)$, $w^{(k+1)}$ are i.i.d. Gaussian noise with mean 0 and covariance being $Nd-$dimensional identity matrix. Let us define the average at k-th iteration as:

$$\bar{x}^{(k)} := \frac{1}{N}\sum_{i=1}^N x_i^{(k)}, \qquad \bar{v}^{(k)} := \frac{1}{N}\sum_{i=1}^N v_i^{(k)}. \tag{36}$$

Since $\mathcal{W}$ is doubly stochastic, we get

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

where $\bar{\xi}^{(k+1)} := \frac{1}{N}\sum_{i=1}^N \xi_i^{(k+1)}$ and $\bar{w}^{(k+1)} := \frac{1}{N}\sum_{i=1}^N w_i^{(k+1)} \sim \frac{1}{\sqrt{N}}\mathcal{N}(0, I_d)$.

We now state the main result of this section which bounds the average of $\mathcal{W}_2$ distance between the distribution of the node iterates $x_i^{(k)}$ and the target distribution π. The result shows that if the parameters η and γ are suitably chosen, then this distance decays geometrically fast (in k) to a level of $\mathcal{O}(\eta)$. This result also bounds the $\mathcal{W}_2$ distance of the node averages $\bar{x}^{(k)}$ and the target distribution π. The main idea of the proof is to analyze DE-SGHMC as a perturbed heavy-ball method (see Section 4.1 and the proof of Lemma 14) which appears to be a new technique to analyze SGHMC methods. Recall $\bar{\gamma} = \max\left\{\left|\lambda_2^W\right|, \left|\lambda_N^W\right|\right\} \in [0, 1)$ from (19), and x_* is the minimizer of $f(x)$.

Theorem 12 *Assume $\mathbb{E}\|x^{(0)}\|^2$ and $\mathbb{E}\|v^{(0)}\|^2$ are finite. Let η be given satisfying*

$$\eta^2 \in \left(0, \frac{1+\lambda_N^W}{2(L+\mu)}\right]. \tag{37}$$

Then, we can can choose $\gamma \in (0, \frac{1}{\eta}]$ such that $\beta := 1-\gamma\eta \in [0,1)$ and satisfies the inequality

$$\beta \leq \bar{\beta} := \min\left(\frac{1+\lambda_N^W - 4\eta^2\mu}{4}, \eta^3\sqrt{c_1\mu^3\frac{(1+\lambda_N^W)}{64}}\right), \tag{38}$$

where

$$c_1 := \frac{1}{2}\frac{\eta^2\mu}{(1+\beta)+(1-\beta)\left(\frac{\eta^2\mu}{1-\lambda_N^W+\eta^2 L}\right)},$$

and for every k, DE-SGHMC iterates $x_i^{(k)}$ given by (33) and their average $\bar{x}^{(k)}$ satisfy

$$\begin{aligned}\mathcal{W}_2\left(\mathcal{L}\left(\bar{x}^{(k)}\right), \pi\right) \leq \left(1-\mu\eta^2\right)^k \left(\left(\mathbb{E}\left\|\bar{x}^{(0)} - x_*\right\|^2\right)^{1/2} + \sqrt{2\mu^{-1}dN^{-1}}\right) \\ + \left(\bar{\gamma}^2 \frac{\left(1-\eta^2\mu\left(1-\frac{\eta^2 L}{2}\right)\right)^k - \bar{\gamma}^{2k}}{\left(1-\eta^2\mu\left(1-\frac{\eta^2 L}{2}\right)\right) - \bar{\gamma}^2}\right)^{1/2} \frac{2L}{\sqrt{N}}\left(\mathbb{E}\left\|x^{(0)}\right\|^2\right)^{1/2} + \eta E_4,\end{aligned} \tag{39}$$

with

$$\begin{aligned}E_4 := \sqrt{2}\left(\frac{\eta^2}{\mu(1-\frac{\eta^2 L}{2})} + \frac{(1+\eta^2 L)^2}{\mu^2(1-\frac{\eta^2 L}{2})^2}\right)^{1/2} \\ \cdot \left[\left(\frac{\beta^2 c_5}{\eta^4 N} + \frac{2L^2 c_5}{N(1-\bar{\gamma})^2}\right)^{1/2} + \left(\frac{(\sqrt{1-\beta}-1)^2}{\eta^4}\frac{d}{N}\right)^{1/2}\right] \\ + \frac{1.65L}{\mu}\sqrt{dN^{-1}} + \frac{\sigma}{\sqrt{\mu(1-\frac{\eta^2 L}{2})N}} = \mathcal{O}(1),\end{aligned}$$

and

$$\begin{aligned}&\frac{1}{N}\sum_{i=1}^N \mathcal{W}_2\left(\mathcal{L}\left(x_i^{(k)}\right), \pi\right) \\ &\leq \left(1-\mu\eta^2\right)^k \left(\left(\mathbb{E}\left\|\bar{x}^{(0)} - x_*\right\|^2\right)^{1/2} + \sqrt{2\mu^{-1}dN^{-1}}\right) + \frac{\sqrt{2}\bar{\gamma}^k}{\sqrt{N}}\left(\mathbb{E}\left\|x^{(0)}\right\|^2\right)^{1/2} \\ &\quad + \left(\frac{\left(1-\eta^2\mu\left(1-\frac{\eta^2 L}{2}\right)\right)^k - \bar{\gamma}^{2k}}{\left(1-\eta^2\mu\left(1-\frac{\eta^2 L}{2}\right)\right) - \bar{\gamma}^2}\right)^{1/2} \frac{2L\bar{\gamma}}{\sqrt{N}}\left(\mathbb{E}\left\|x^{(0)}\right\|^2\right)^{1/2} + \eta E_5,\end{aligned} \tag{40}$$

with $E_5 := E_4 + \frac{\sqrt{2c_5}}{\sqrt{N}(1-\bar{\gamma})} = \mathcal{O}(1)$, *and* $\beta = \mathcal{O}(\eta^4)$ *where* $\mathcal{O}(\cdot)$ *hides the constants that depend on* d, μ, L, σ *and* $\bar{\gamma}$ *and* N, $\mathcal{L}\left(\bar{x}^{(k)}\right)$ *denotes the law of* $\bar{x}^{(k)}$ *and* π *denotes the Gibbs distribution with probability density function proportional to* $e^{-f(x)}$, *and* c_5 *is defined in Lemma 14.*

Remark 13 *We observe that in the setting of Theorem 12, the asymptotic error with respect to the target distribution satisfies* $\limsup_{k\to\infty} \frac{1}{N}\sum_{i=1}^{N} \mathcal{W}_2\left(\mathcal{L}\left(x_i^{(k)}\right), \pi\right) = \mathcal{O}(\eta)$, *where* $\mathcal{O}(\cdot)$ *hides other constants (*d, μ, L, σ, N *and* $\bar{\gamma}$*). In particular, for* η *small enough, it is easy to check that* $\left(1 - \eta^2\mu\left(1 - \frac{\eta^2 L}{2}\right)\right) \geq \bar{\gamma}^2$ *and consequently from Theorem 12, defining* $\alpha := \eta^2$, *we obtain*

$$\mathcal{W}_2\left(\mathcal{L}\left(\bar{x}^{(2K)}\right), \pi\right) \leq \left(1 - \alpha\mu\left(1 - \frac{\alpha L}{2}\right)\right)^K b_0 + \mathcal{O}\left(\sqrt{\alpha}\right) \tag{41}$$

$$\leq e^{-\alpha\mu\left(1-\frac{\alpha L}{2}\right)K} b_0 + \mathcal{O}\left(\sqrt{\alpha}\right) \tag{42}$$

for some b_0 *(that depends on the initialization* $x^{(0)}, d, \mu, L, \sigma, \bar{\gamma}$ *and* N*) where* K *is the iteration budget. We observe that this bound in* α *is similar to that of DE-SGLD case analyzed in* (21)–(22) *if we were to replace* η *in* (21)–(22) *by* α. *By following the same argument as in Remark 3, if we choose* $\alpha = \eta^2 = \frac{c\log\sqrt{K}}{\mu K}$, *where the constant* $c > 1$, *then we obtain*

$$\mathcal{W}_2\left(\mathcal{L}\left(\bar{x}^{(2K)}\right), \pi\right) = \mathcal{O}\left(\frac{\sqrt{\log(K)}}{\sqrt{K}}\right).$$

We conclude that in order to sample fwerom a distribution that is ε *close to the target in the 2-Wasserstein distance, it suffices to have* $\mathcal{O}(\frac{1}{\varepsilon^2})$ *iterations of DE-SGHMC, ignoring logarithmic factors. This iteration complexity bound is of the same order with that we obtained for DE-SGLD (see Remark 3). However, in practice we have seen that DE-SGHMC outperformed DE-SGLD in some cases (see Section 5.3). We also note that, with a similar analysis to that in Remark 4, it can be shown that all the terms appearing in the performance bounds is monotonically increasing as a function of* $\bar{\gamma}$ *in the setting of Theorem 12 except the constant* c_5 *whose dependency to* $\bar{\gamma}$ *is more complicated to determine within our analysis.*

4.1 Proof of Theorem 12

To facilitate the analysis, we introduce the iterates (x_k) (with slight abuse of notations):

$$x_{k+1} = x_k - \eta^2 \frac{1}{N}\nabla f(x_k) + \sqrt{2}\eta\bar{w}^{(k+1)}, \tag{43}$$

where $\bar{w}^{(k+1)}$ is the Gaussian noise given in (17) and $x_0 = \bar{x}^{(0)}$. This is an Euler-Maruyama discretization (with stepsize η^2) of the continuous-time overdamped Langevin diffusion:

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

where W_t is a standard d-dimensional Brownian motion. We also define iterates $(\tilde{x}_k)$:

$$\tilde{x}_{k+1} = \tilde{x}_k - \eta^2 \frac{1}{N}\nabla f(\tilde{x}_k) + \sqrt{2\gamma\eta}\eta\bar{w}^{(k+1)}, \tag{44}$$

where $\tilde{x}_0 = x_0 = \bar{x}^{(0)}$.

We recall that SGHMC can be viewed as a discretization of the kinetic (inertial) Langevin SDE (5)–(6). It is also known that as the friction coefficient $\gamma \to \infty$, the paths of this SDE becomes more and more similar to the paths of the overdamped Langevin SDE (see e.g. Leimkuhler et al. (2016)). However, this is not the case when $\gamma > 0$ is small enough. Therefore, the stepsize η is small enough and the friction coefficient γ is large enough, it is reasonable to expect that the node averages $\bar{x}^{(k)}$ of DE-SGHMC given by (36) track the overdamped SDE dynamics. In the setting of Theorem 12, we consider such a case when the stepsize η is small enough and $\gamma \approx \frac{1}{\eta}$ (see (37)-(38)). We will next show that node averages $\bar{x}^{(k)}$ and the discretized overdamped SDE iterates will be close to each other in the 2-Wasserstein metric and that the iterates $x_i^{(k)}$ will remain close to their average $\bar{x}^{(k)}$ in L_2 distance. We note that in general, the optimal choice of γ is not known in the decentralized setting; and it is only known in the centralized setting for special cases: For centralized Langevin dynamics with deterministic gradients and μ-strongly convex quadratic objectives, it is recently shown that the choice of $\gamma = 2\sqrt{\mu}$ optimizes the convergence rate to the stationary distribution in the 2-Wasserstein distance (Gao et al., 2020). Studying the convergence of DE-SGHMC iterates for other choices of the friction coefficient γ will be left as a future work, as our current proof techniques do not allow arbitrary choice of γ.

In the proof of Theorem 12, to bound the $\mathcal{W}_2$ distance between the average of $\mathcal{L}\left(x_i^{(k)}\right)$ and π over $1 \le i \le N$, we follow a similar approach to the analysis of DE-SGLD where the idea is to bound the following four terms: (1) the L^2 distance between $x_i^{(k)}$ and the average iterate $\bar{x}^{(k)}$; (2) the L^2 distance between the average iterate $\bar{x}^{(k)}$ and iterates $\tilde{x}_k$ in (44); (3) the L^2 distance between the iterates $\tilde{x}_k$ and the iterates x_k in (43); (4) the $\mathcal{W}_2$ distance between $\mathcal{L}(x_k)$ and π, i.e. the convergence of overdamped Langevin dynamics. For analyzing the first term, we first present a technical lemma (Lemma 14) on uniform L^2 bounds on the iterates $v^{(k)}, x^{(k)}$ in (34)–(35). The result will be used in the proof of Lemma 15. The proof idea is to analyze DE-SGHMC as a perturbed heavy-ball method. Momentum-based first order methods such as heavy-ball methods are less robust to noise compared to gradient descent methods (see e.g. Can et al. (2019b); Kuru et al. (2020); Flammarion and Bach (2015); Mohammadi et al. (2021); Devolder et al. (2014)), and achieving this result requires significantly more work compared to the analogous result we obtained for DE-SGLD. The proof of this result (and all the other lemmas) are given in the Appendix.

We first provide uniform L^2 bounds on the iterates $v^{(k)}, x^{(k)}$ in (34)–(35) in the following lemma.

Lemma 14 *Under the assumptions of Theorem 12, there exist constants c_4 and c_5 (that do not depend on η or γ) that can be made explicit such that*

$$\sup_{k\ge 1} \mathbb{E}\left[\left\| x^{(k)} + \frac{\beta}{1-\beta}\left(x^{(k)} - x^{(k-1)}\right)\right\|^2\right] \le c_4\,, \tag{45}$$

$$\sup_{k\ge 1} \max\left(\mathbb{E}\left\|v^{(k)}\right\|^2, \mathbb{E}\left\|x^{(k)}\right\|^2\right) \le c_5\,. \tag{46}$$

With this lemma, we can bound the deviation of $x_i^{(k)}$ in (33) from the mean $\bar{x}^{(k)}$ in (36). We state the result in the next subsection.

4.1.1 Uniform L^2 bounds on the deviation from the mean

Lemma 15 *Under the assumptions of Theorem 12, for any k, we have*

$$\sum_{i=1}^{N} \mathbb{E}\left\|x_i^{(k)} - \bar{x}^{(k)}\right\|^2 \leq 2\bar{\gamma}^{2k}\mathbb{E}\left\|x^{(0)}\right\|^2 + \frac{2c_5\eta^2}{(1-\bar{\gamma})^2},$$

where c_5 is defined in Lemma 14 and $\bar{\gamma} = \max\left\{\left|\lambda_2^W\right|, \left|\lambda_N^W\right|\right\} \in [0,1)$.

Note that we have

$$\bar{v}^{(k+1)} = \bar{v}^{(k)} - \gamma\eta\bar{v}^{(k)} - \eta\frac{1}{N}\nabla f\left(\bar{x}^{(k)}\right) + \eta\mathcal{E}_{k+1} - \eta\bar{\xi}^{(k+1)} + \sqrt{2\gamma\eta}\bar{w}^{(k+1)}, \tag{47}$$

$$\bar{x}^{(k+1)} = \bar{x}^{(k)} + \eta\bar{v}^{(k+1)}, \tag{48}$$

where $\mathcal{E}_{k+1} := \frac{1}{N}\nabla f\left(\bar{x}^{(k)}\right) - \frac{1}{N}\sum_{i=1}^{N}\nabla f_i\left(x_i^{(k)}\right)$. As a corollary of Lemma 15, we have the following estimate.

Lemma 16 *Under the assumptions of Theorem 12, for any k, we have*

$$\mathbb{E}\left\|\mathcal{E}_{k+1}\right\|^2 \leq \frac{2L^2\bar{\gamma}^{2k}}{N}\mathbb{E}\left\|x^{(0)}\right\|^2 + \frac{2L^2c_5\eta^2}{N(1-\bar{\gamma})^2}. \tag{49}$$

4.1.2 L^2 distance between the mean and discretized overdamped SDE

Given the dynamics of the average iterate $(\bar{v}^{(k)}, \bar{x}^{(k)})$ in (47)–(48), we next show $\bar{x}^{(k)}$ is close to the iterates $\tilde{x}_k$ in (44), which is close to the iterates x_k in (43) obtained from an Euler-Maruyama discretization of an overdamped Langevin SDE. By plugging (47) into (48), we get

$$\bar{x}^{(k+1)} = \bar{x}^{(k)} + \eta\bar{v}^{(k)} - \gamma\eta^2\bar{v}^{(k)} - \eta^2\frac{1}{N}\nabla f\left(\bar{x}^{(k)}\right) + \eta^2\mathcal{E}_{k+1} - \eta^2\bar{\xi}^{(k+1)} + \sqrt{2\gamma\eta}\eta\bar{w}^{(k+1)}. \tag{50}$$

By (47), we get $\bar{v}^{(k)} = \frac{\bar{x}^{(k)} - \bar{x}^{(k-1)}}{\eta}$, so that (50) becomes:

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

where we recall that $\beta = 1 - \gamma\eta$. Also recall that we define $\tilde{x}_k$ from the iterates:

$$\tilde{x}_{k+1} = \tilde{x}_k - \eta^2\frac{1}{N}\nabla f(\tilde{x}_k) + \sqrt{2(1-\beta)}\eta\bar{w}^{(k+1)}, \tag{51}$$

where $\tilde{x}_0 = \frac{1}{N}\sum_{i=1}^{N} x_i^{(0)}$. We have the following estimate.

Lemma 17 *Under the assumptions of Theorem 12, we have for every k,*

$$\mathbb{E}\left\|\bar{x}^{(k)} - \tilde{x}_k\right\|^2 \leq 2\left(\frac{\eta^2}{\mu(1-\frac{\eta^2 L}{2})} + \frac{(1+\eta^2 L)^2}{\mu^2(1-\frac{\eta^2 L}{2})^2}\right)\left(\frac{\beta^2 c_5}{\eta^2 N} + \frac{2L^2c_5\eta^2}{N(1-\bar{\gamma})^2}\right) + \frac{\eta^2\sigma^2}{\mu(1-\frac{\eta^2 L}{2})N}$$
$$+ \frac{\bar{\gamma}^{2k} - \left(1-\eta^2\mu\left(1-\frac{\eta^2 L}{2}\right)\right)^k}{\bar{\gamma}^2 - 1 + \eta^2\mu\left(1-\frac{\eta^2 L}{2}\right)}\frac{4L^2\bar{\gamma}^2}{N}\mathbb{E}\left\|x^{(0)}\right\|^2,$$

where the constant c_5 is as in Lemma 14.

Next, recall the iterates x_k defined in (43):

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

where $x_0 = \tilde{x}_0 = \bar{x}^{(0)}$. This is an Euler-Maruyama discretized version of the continuous-time overdamped Langevin diffusion with stepsize η^2. Since $\beta = 1 - \gamma\eta$ is small (see (38)), we will then show that $\tilde{x}_k$ and x_k are close to each other in L^2 distance. Indeed, we have the following estimate.

Lemma 18 *Under the assumptions of Theorem 12, we have for every k,*

$$\mathbb{E}\left\|\tilde{x}_k - x_k\right\|^2 \leq 2\left(\frac{\eta^2}{\mu(1-\frac{\eta^2 L}{2})} + \frac{(1+\eta^2 L)^2}{\mu^2(1-\frac{\eta^2 L}{2})^2}\right)\left(\frac{(\sqrt{1-\beta}-1)^2}{\eta^2}\frac{d}{N}\right).$$

4.1.3 Proof of Theorem 12

Since $x_0 = \frac{1}{N}\sum_{i=1}^N x_i^{(0)}$, we have $\mathbb{E}\|x_0\|^2 < \infty$. By assumption we have also $\eta^2 \leq \frac{2N}{\mu+L}$. Then, it follows from Lemma 10 and Lemma 11 that for x_k defined in (43) we have

$$\mathcal{W}_2\left(\mathcal{L}(x_k), \pi\right) \leq \left(1-\mu\eta^2\right)^k \left(\left(\mathbb{E}\|x_0 - x_*\|^2\right)^{1/2} + \sqrt{2\mu^{-1}dN^{-1}}\right) + \frac{1.65L}{\mu}\sqrt{\eta^2 dN^{-1}}.$$

Moreover, it follows from Lemma 17 that

$$\begin{aligned}
&\mathcal{W}_2\left(\mathcal{L}\left(\bar{x}^{(k)}\right), \mathcal{L}(\tilde{x}_k)\right) \\
&\leq \left(\mathbb{E}\left\|\bar{x}^{(k)} - \tilde{x}_k\right\|^2\right)^{1/2} \\
&\leq \sqrt{2}\left(\frac{\eta^2}{\mu(1-\frac{\eta^2 L}{2})} + \frac{(1+\eta^2 L)^2}{\mu^2(1-\frac{\eta^2 L}{2})^2}\right)^{1/2}\left(\frac{\beta^2 c_5}{\eta^2 N} + \frac{2L^2 c_5 \eta^2}{N(1-\bar{\gamma})^2}\right)^{1/2} \\
&\quad + \frac{\eta\sigma}{\sqrt{\mu(1-\frac{\eta^2 L}{2})N}} + \left(\frac{\bar{\gamma}^{2k} - \left(1-\eta^2\mu\left(1-\frac{\eta^2 L}{2}\right)\right)^k}{\bar{\gamma}^2 - 1 + \eta^2\mu\left(1-\frac{\eta^2 L}{2}\right)}\right)^{1/2} \frac{2L\bar{\gamma}}{\sqrt{N}}\left(\mathbb{E}\left\|x^{(0)}\right\|^2\right)^{1/2},
\end{aligned}$$

whereas it follows from Lemma 18 that

$$\begin{aligned}
\mathcal{W}_2\left(\mathcal{L}(\tilde{x}_k), \mathcal{L}(x_k)\right) &\leq \left(\mathbb{E}\left\|\tilde{x}_k - x_k\right\|^2\right)^{1/2} \\
&\leq \sqrt{2}\left(\frac{\eta^2}{\mu(1-\frac{\eta^2 L}{2})} + \frac{(1+\eta^2 L)^2}{\mu^2(1-\frac{\eta^2 L}{2})^2}\right)^{1/2}\left(\frac{(\sqrt{1-\beta}-1)^2}{\eta^2}\frac{d}{N}\right)^{1/2}.
\end{aligned}$$

Hence, we conclude that

$$\mathcal{W}_2\left(\mathcal{L}\left(\bar{x}^{(k)}\right),\pi\right) \leq \left(1-\mu\eta^2\right)^k \left(\left(\mathbb{E}\left\|\bar{x}^{(0)}-x_*\right\|^2\right)^{1/2} + \sqrt{2\mu^{-1}dN^{-1}}\right) + \left(\frac{\left(1-\eta^2\mu\left(1-\frac{\eta^2 L}{2}\right)\right)^k - \bar{\gamma}^{2k}}{\left(1-\eta^2\mu\left(1-\frac{\eta^2 L}{2}\right)\right)-\bar{\gamma}^2}\right)^{1/2} \frac{2L\bar{\gamma}}{\sqrt{N}}\left(\mathbb{E}\left\|x^{(0)}\right\|^2\right)^{1/2} + \eta E_4, \tag{52}$$

with

$$E_4 := \sqrt{2}\left(\frac{\eta^2}{\mu(1-\frac{\eta^2 L}{2})} + \frac{(1+\eta^2 L)^2}{\mu^2(1-\frac{\eta^2 L}{2})^2}\right)^{1/2} \cdot \left[\left(\frac{\beta^2 c_5}{\eta^4 N} + \frac{2L^2 c_5}{N(1-\bar{\gamma})^2}\right)^{1/2} + \left(\frac{(\sqrt{1-\beta}-1)^2}{\eta^4}\frac{d}{N}\right)^{1/2}\right] + \frac{1.65L}{\mu}\sqrt{dN^{-1}} + \frac{\sigma}{\sqrt{\mu(1-\frac{\eta^2 L}{2})N}}.$$

Finally, by (31), we have

$$\frac{1}{N}\sum_{i=1}^N \mathcal{W}_2\left(\mathcal{L}\left(x_i^{(k)}\right),\mathcal{L}\left(\bar{x}^{(k)}\right)\right) \leq \sqrt{\frac{1}{N}\sum_{i=1}^N \mathbb{E}\left\|x_i^{(k)}-\bar{x}^{(k)}\right\|^2}.$$

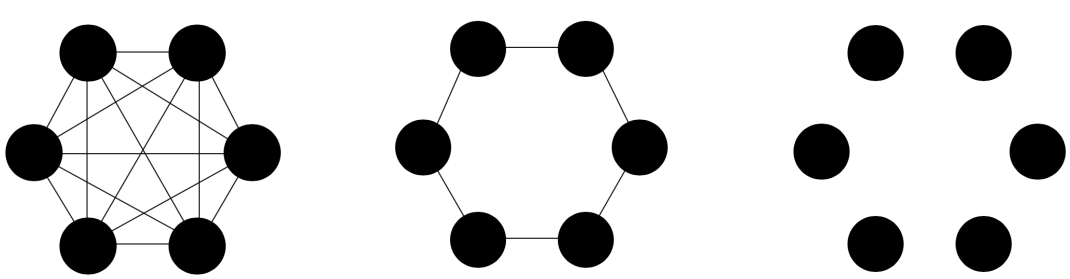
On the other hand, by Lemma 15,

$$\sqrt{\frac{1}{N}\sum_{i=1}^N \mathbb{E}\left\|x_i^{(k)}-\bar{x}^{(k)}\right\|^2} \leq \left(\frac{2\bar{\gamma}^{2k}}{N}\mathbb{E}\left\|x^{(0)}\right\|^2 + \frac{2c_5\eta^2}{N(1-\bar{\gamma})^2}\right)^{1/2} \leq \frac{\sqrt{2}\bar{\gamma}^k}{\sqrt{N}}\left(\mathbb{E}\left\|x^{(0)}\right\|^2\right)^{1/2} + \frac{\sqrt{2c_5}\eta}{\sqrt{N}(1-\bar{\gamma})}.$$

We then obtain (40) by applying the triangular inequality for the 2-Wasserstein distance. Finally, since β satisfies the inequality (38), we have $\beta = \mathcal{O}(\eta^4)$ as $\eta \to 0$ and this implies that $E_4 = \mathcal{O}(1)$ and $E_5 = \mathcal{O}(1)$ as claimed. This completes the proof. □

5. Numerical Experiments

We present our numerical results in this section. We conduct several experiments to validate our theory and investigate the performance of DE-SGLD and DE-SGHMC. We focus on applying our methods to Bayesian linear regression and Bayesian logistic regression problems. In our experiments, each agent has its own data in the form of i.i.d. samples. We will consider three different network architectures: (a) Fully-connected network (b) Circular network (c) A disconnected network with no edges as illustrated in Figure 1. Fully-connected



(a) Fully-connected (b) Circular (c) Disconnected

Figure 1: Illustration of the network architectures.

network structure corresponds to the complete graph where all the nodes are connected to each other whereas for the circular graph, each node can communicate with only "left" and "right" neighbors. Disconnected graph corresponds to the case when nodes do not communicate at all with each other. The disconnected network is considered as a baseline case for comparison purposes to see how the individual agents would perform without sharing any information among themselves.

Before we proceed to the numerical experiments, we remark that the following examples all satisfy the assumptions in our paper. We will have a discussion on this in the Appendix in detail. In particular, Appendix D shows that the variance of the gradient noise is bounded, and Appendix E shows that the gradient of the component functions are Lipschitz.

5.1 Bayesian linear regression

In this section, we present our experiments on the Bayesian linear regression problem, where our main goal is to validate Theorems 2 and 12 in a basic setting and show that each agent can sample from the posterior distribution up to an error tolerance with constant stepsize. In this set of experiments, we first generate data for each agent by simulating the model:

$$\delta_j \sim \mathcal{N}\left(0, \xi^2\right), \quad X_j \sim \mathcal{N}(0, I), \quad y_j = x^T X_j + \delta_j, \tag{53}$$

where the noise term δ_j are i.i.d. scalars with $\xi = 1$, $x \in \mathbb{R}^2$, and the prior distribution of x follows $\mathcal{N}(0, \lambda I)$ where we take $\lambda = 10$ in the experiments. For the Bayesian linear regression, we can derive the posterior distribution as:

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

where $\Sigma = \lambda I$ is the covariance matrix of the prior distribution of x, $X = [X_1^T, X_2^T, \dots]^T$ and $Y = [y_1, y_2, \dots]^T$ are the matrices containing all data points. We simulate 5,000 data points and partition them randomly among the $N = 100$ agents so that each agent will have the same number of data points. Each agent has access to its own data but not to other agents' data. The posterior distribution $\pi(x) \propto e^{-f(x)}$ is of the form $f(x) = \sum_{i=1}^N f_i(x)$ with

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

where

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

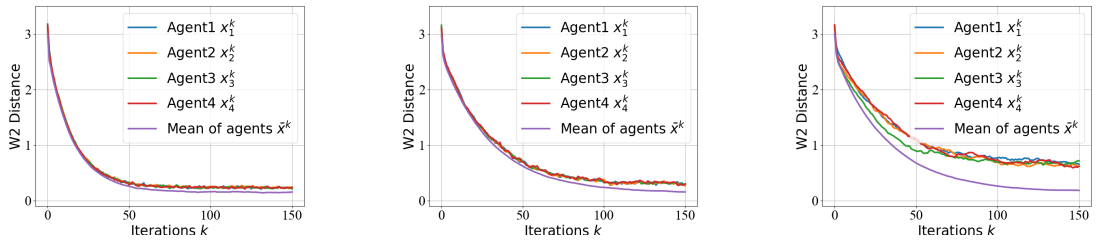
and agent i has $n_i = 50$ data points $\{(X_j^i, y_j^i)\}_{j=1}^{n_i}$.

In the first experiment, we report the performance of the DE-SGLD method on the fully-connected, circular and disconnected networks in Figure 2. We tune the stepsize η to the dataset where we take $\eta = 0.009$. We consider the case when gradients are deterministic (i.e. $\sigma = 0$). In this case, it can be seen that $x_i^{(k)}$ follows a Gaussian distribution, i.e. $x_i^{(k)} \sim \mathcal{N}\left(m_i^{(k)}, \Sigma_i^{(k)}\right)$ for some mean vector $m_i^{(k)}$ and covariance matrix $\Sigma_i^{(k)}$. Based on 100 independent runs, we estimate the parameters $m_i^{(k)}$ and $\Sigma_i^{(k)}$ and then compute the 2-Wasserstein distance with respect to the posterior distribution $\pi(x) \sim \mathcal{N}(m, V)$ based on the explicit formula (Givens and Shortt, 1984) which characterizes the 2-Wasserstein distance between any two Gaussian distributions. This allows us to plot the 2-Wasserstein distance to the stationary distribution for each agent and for the distribution of the average $\bar{x}^{(k)} = \sum_{i=1}^N x_i^{(k)}/N$ over the iterations in Figure 2. We observe that for both complete and circular graphs all the agents will converge to the posterior distribution up to an error tolerance. In the case of the disconnected network, we observe that individual agents do relatively worse compared to the fully-connected and circular network cases; as they do not leverage any information about their neighbors' data points. For the disconnected case, the nodes averages $\bar{x}^{(k)}$ (which is neither computed nor accessible by agents) is closer to the target distribution than the individual iterates $x_i^{(k)}$ as it contains information from each agent; however the performance of the node averages $\bar{x}^{(k)}$ is still worse compared to the performance of node averages for the fully-connected case as expected. We observe that the experiments in the fully-connected network converges faster than the circular network. This behavior is predicted by Theorem 2. Since the fully-connected network has a larger *spectral gap* $1-\bar{\gamma}$ compared to the circular network (see the paragraph after (19)), our performance bounds for the fully-connected network is better compared to the circular network case.[5]

In the next experiment, we investigate the performance of the DE-SGHMC method on the same data set with (the same) three network structures. The stepsize η and the friction coefficient γ are tuned to the dataset where we take $\eta = 0.1$ and $\gamma = 7$. The results are displayed in Figure 3. The results are qualitatively similar to the DE-SGLD case. The convergence of DE-SGHMC is fastest for the fully-connected case and is the slowest for the disconnected case.

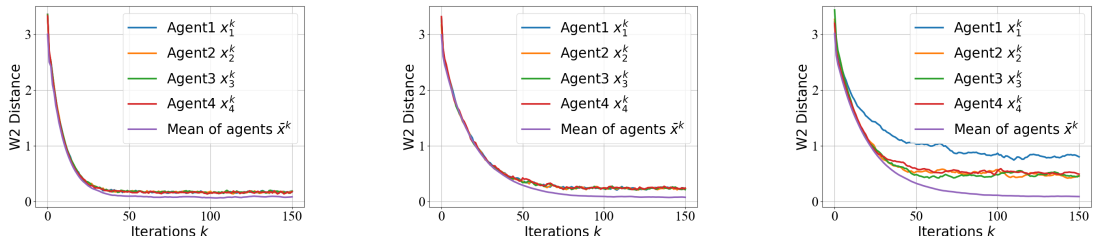
In the next set of experiments, we investigate the effect of changing stepsize, batch size and the network structure on the speed of convergence where we stick to the DE-SGLD method for this set of experiments. We measure the 2-Wasserstein distance to the target π with a similar approach as before by fitting a Gaussian distribution $\mathcal{N}(m_i^{(k)}, \Sigma_i^{(k)})$ to the empirical distribution of $x_i^{(k)}$ over 100 independent runs. The results are shown in Figure 4. Both Figure 4(a) and Figure 4(b) are based on the fully-connected network architecture. In Figure 4(a), we fix the stepsize to $\eta = 0.009$ and vary the batch sizes (the

5. This is a consequence of the fact that our upper bounds given for the 2-Wasserstein distances to the target π in Theorem 2 and Theorem 12 are both monotonically increasing in $\bar{\gamma}$ (see Remark 4 and Remark 13).



(a) Fully-connected (b) Circular (c) Disconnected

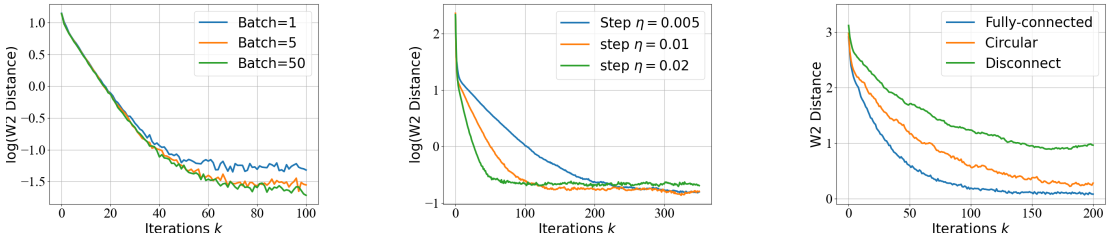
Figure 2: Performance of DE-SGLD for Bayesian regression on different network structures with $N = 100$ agents. The results of the first 4 agents x_i^k and the node averages $\bar{x}^k = \sum_{i=1}^N x_i^{(k)}/N$ are reported.



(a) Fully-connected (b) Circular (c) Disconnected

Figure 3: Performance of DE-SGHMC method for Bayesian regression on different network structures. The stepsize η and the friction coefficient γ are tuned to the dataset where we take $\eta = 0.1$ and $\gamma = 7$.

number of data points sampled with replacement to estimate the gradient). We conclude that different batch sizes affect the asymptotic error the iterates have with respect to the 2-Wasserstein distance. Larger batch sizes reduce the amount of noise (i.e. the upper bound σ^2 on the gradient noise) and therefore lead to smaller asymptotic error as predicted by Theorem 2. In Figure 4(b), we used stochastic gradients with batch size $b = 25$ while we varied the stepsize. The result clearly demonstrates the trade-off between the convergence rate and the asymptotic accuracy; for larger stepsize the algorithm converges faster to an asymptotic error region but the accuracy becomes worse as predicted by Theorem 2 (see also Remark 3). In Figure 4(c) we report the effect of network structure with a constant stepsize $\eta = 0.008$ and batch size $b = 25$ where we report the performance of a randomly picked agent. The fastest convergence is observed for the fully-connected network. For the disconnected network, each agent will converge to a stationary distribution based on its own data rather than the posterior distribution based on the whole data set; therefore the asymptotic error in 2-Wasserstein distance will be bounded away from zero.



(a) Batch size (b) Stepsize (c) Network structure

Figure 4: Performance of DE-SGLD method for Bayesian regression under different settings. Figures are based on one randomly picked agent. The y-axis is presented in a logarithmic scale in (a) and (b) .

5.2 Bayesian logistic regression

In Bayesian logistic regression, we are given a dataset of input-output pairs $A = \{a_j\}_{j=1}^n$ where $a_j = (X_j, y_j)$, $X_j \in R^d$ are the features and $y_j \in \{0, 1\}$ are the binary labels. We assume that X_j are independent, and that the probability distribution of the output y_j given features X_j and the regression coefficients $x \in \mathbb{R}^d$ is given by

$$\mathbb{P}(y_j = 1 | X_j, x) = \frac{1}{1 + e^{-x^T X_j}}. \tag{54}$$

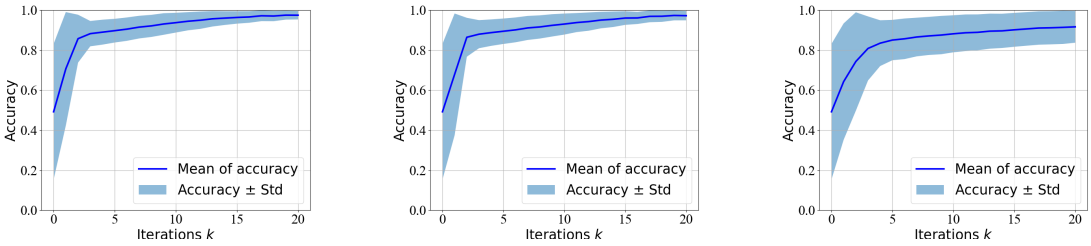
The prior distribution $p(x)$ is often taken as a Gaussian distribution $\mathcal{N}(0, \lambda I)$ for some $\lambda > 0$ (see e.g. Chatterji et al. (2018); Dubey et al. (2016); Zou et al. (2018b)). If each agent i possesses a subset A_i of the data where $A_i = \{(X_j^i, y_j^i)\}_{i=1}^{n_i}$, then the goal in Bayesian logistic regression is to sample from $\pi(x) \propto e^{-f(x)}$ with $f(x) = \sum_i f_i(x)$ where

$$f_i(x) := -\sum_{j=1}^{n_i} \log p\left(y_j^i = 1 | X_j^i, x\right) - \frac{1}{N} \log p(x) = \sum_{j=1}^{n_i} \log\left(1 + e^{-x^T X_j^i}\right) + \frac{1}{2N\lambda}\|x\|^2 \tag{55}$$

is strongly convex and smooth. We first test our algorithms on synthetic data where we simulate (54) by

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

where $\mathcal{U}(0, 1)$ is the uniform distribution on $[0, 1]$, $x = [x_1, x_2, x_3]^T \in \mathbb{R}^3$ and the prior distribution of x follows $\mathcal{N}(0, \lambda I)$ where we take $\lambda = 10$ in the experiments. Similar to the case of Bayesian linear regression, we separate the data points approximately equally among all the agents, where we take $N = 6$. Each agent can access to one part of the data set. Unlike Bayesian linear regression, where the posterior distribution admits an explicit formula, the posterior distribution $\pi(x)$ of Bayesian logistic regression does not admit an explicit formula. In principle, one can approximate the stationary distribution

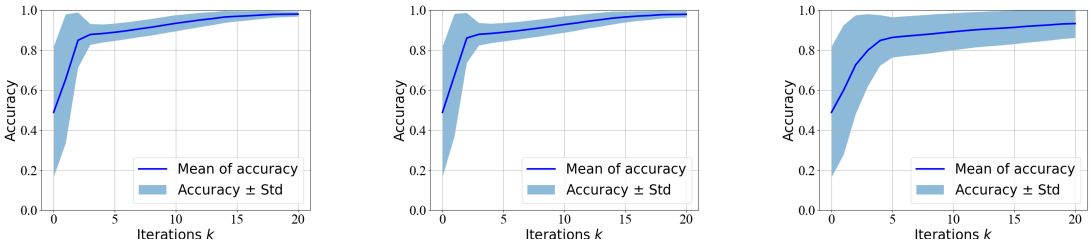


(a) Fully-connected (b) Circular (c) Disconnected

Figure 5: The plots show the the accuracy over the data set versus number of iterations for the DE-SGLD method on different network structures. Figures are based on one randomly picked agent. Here, the stepsizes are tuned to the dataset where we take $\eta = 0.0003$. We use the stochastic gradient with batch size $b = 32$ in the experiments.

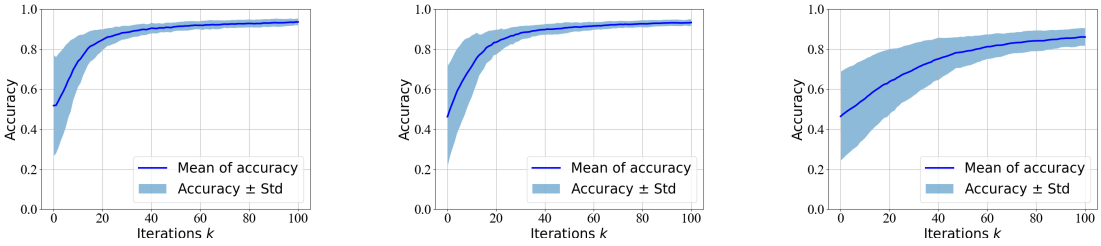
by running the algorithm over many runs and compute the Wasserstein distance between this approximate distribution and the empirical distribution of the iterates and report this distance as a performance measure. However, this is not practical. Instead, we resort to another performance measure for each agent, which is the distribution of the accuracy over the whole data set where accuracy is defined as the ratio of the correctly predicted labels. This ratio is relatively simpler to compute and serves as a measure correlated with the goodness of fit to the training data. For this purpose, we run the DE-SGLD method multiple times and for each realization of the k-th iterate $x_i^{(k)}$ at node i, we classify the whole data set based on $x_i^{(k)}$ and calculate the accuracy over $n = 1,000$ data points. Over 100 independent runs of the DE-SGLD algorithm with batch size $b = 32$, we estimate the distribution of the accuracy for each agent at step k. We report the mean and the standard deviation of the accuracy in Figure 5. We can clearly observe that the DE-SGLD method works well for both fully-connected and circular networks for Bayesian logistic regression, which supports our theory. In the right panel of Figure 5, we show the results of the DE-SGLD method for the disconnected network. The performance on the disconnected network is worse compared to the fully-connected and circular network settings as expected.

In our next set of experiments, we investigate the DE-SGHMC method in Figure 6 where we take $\eta = 0.02$ and $\gamma = 30$ after tuning these parameters to the dataset. We use the batch size $b = 32$ in this set of experiments. We see that the performance of DE-SGHMC for fully-connected and circular networks is also better compared to the disconnected setting as expected.



(a) Fully-connected (b) Circular (c) Disconnected

Figure 6: The plots show the accuracy over the data set versus number of iterations for the DE-SGHMC method on different network structures. Figures are based on one randomly picked agent. Here, the stepsize η and the friction coefficient γ are tuned to the dataset where we take $\eta = 0.02$ and $\gamma = 30$. We use the stochastic gradient with batch size $b = 32$ in the experiments.



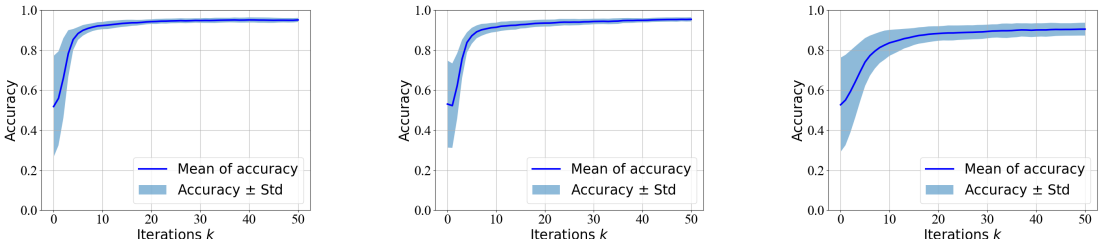
(a) Fully-connected (b) Circular (c) Disconnected

Figure 7: The plots show the accuracy over the data set versus number of iterations for the DE-SGLD method on different network structures over Breast Cancer data set. Figures are based on one randomly picked agent. Here, the stepsizes are chosen as $\eta = 0.0008$. We use batch size $b = 32$ in the experiments.

5.3 Bayesian logistic regression with real data

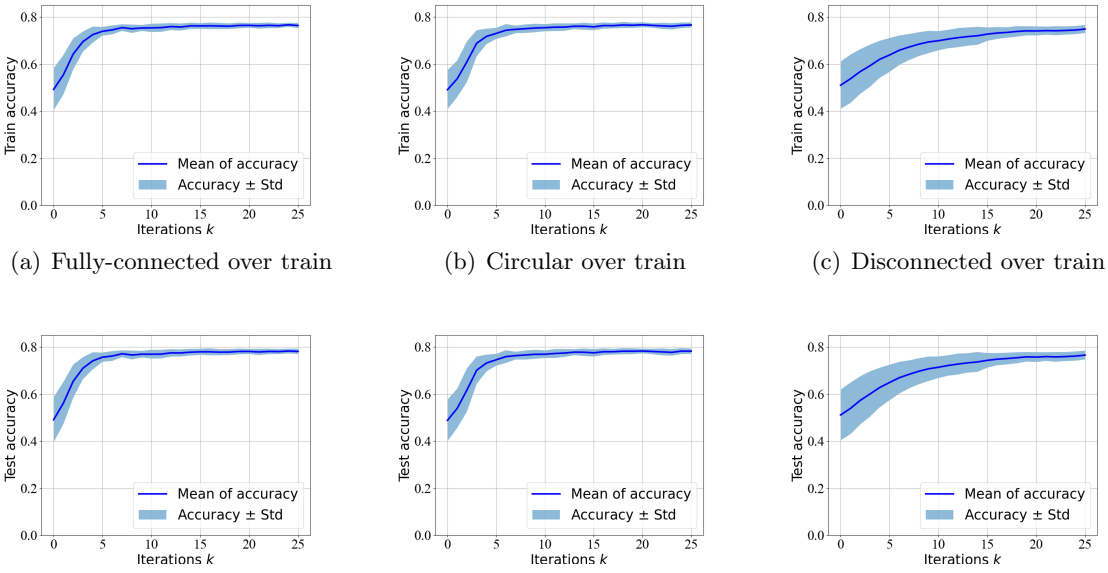
In this section, we consider the Bayesian logistic regression problem on the UCI ML Breast Cancer Wisconsin (Diagnostic) data set[6] and MAGIC Gamma Telescope data set[7]. The Breast Cancer data set contains 569 samples with dimension 31 and each sample describes characteristics of the cell nuclei present in a digitized image of a fine needle aspirate (FNA) of a breast mass. The Telescope data set contains 19,020 samples with dimension 11 and each sample describes the registration of high energy gamma particles in a ground-based atmospheric Cherenkov gamma telescope using the imaging technique.

6. The corresponding data set is available online at `https://archive.ics.uci.edu/ml/datasets/Breast+Cancer+Wisconsin+(Diagnostic)`.
7. The data set is available at `https://archive.ics.uci.edu/ml/datasets/magic+gamma+telescope`.



(a) Fully-connected (b) Circular (c) Disconnected

Figure 8: The plots show the accuracy over the data set versus number of iterations for the DE-SGHMC method on different network structures over Breast Cancer data set. Figures are based on one randomly picked agent. Here, the stepsize η and the friction coefficient γ are well tuned to the data set so that we take $\eta = 0.05$, $\gamma = 10$. We use batch size $b = 32$ in the experiments.

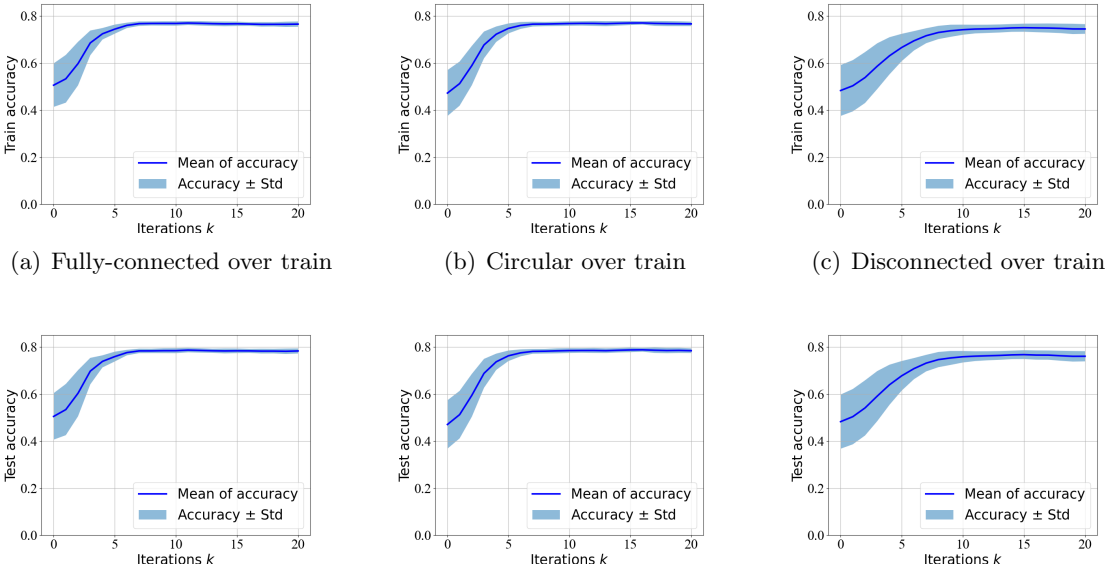


(a) Fully-connected over train (b) Circular over train (c) Disconnected over train

(d) Fully-connected over test (e) Circular over test (f) Disconnected over test

Figure 9: The plots show the "distribution of the accuracy over the data set" versus number of iterations for the DE-SGLD method on different network structures over the training data and the test data of Telescope data set. Figures are based on one randomly picked agent. Here, the stepsizes are chosen as $\eta = 0.008$. We use batch size $b = 100$ in the experiments.

Figure 7 and Figure 8 illustrate the results of using DE-SGLD and DE-SGHMC methods applied to the classification problem over the Breast Cancer data set. For Breast Cancer



(a) Fully-connected over train (b) Circular over train (c) Disconnected over train

(d) Fully-connected over test (e) Circular over test (f) Disconnected over test

Figure 10: The plots show the accuracy over the data set versus number of iterations for the DE-SGHMC method on different network structures over training data and test data from the Telescope data set. Figures are based on one randomly picked agent. Here, the stepsize η and the friction coefficient γ are tuned to the data set where we take $\eta = 0.07$, $\gamma = 5$. We use batch size $b = 100$ in the experiments.

data set, we separate the data set into $N = 6$ parts with approximately equal sizes and each agent can access to only one part of the whole data set. Similar to the previous section, we use the distribution of accuracy over the whole data set as the performance measure and report the performance of a randomly picked agent. The performance in the disconnected network setting is worse compared to the connected setting. We observe that the convergence of DE-SGHMC method displayed in Figure 8 compared to DE-SGLD is slightly faster with a smaller standard deviation of accuracy in general.

Figure 9 and Figure 10 illustrate the results of using DE-SGLD and DE-SGHMC methods for classification over the Telescope data set. For Telescope data set, we separate the data set into training data and test data, where test data has 10% data points. Then we separate the training data into 6 parts same as before. We report the accuracy over both training data and test data in the figures. We get similar results that illustrate that both methods perform better for fully-connected and circular networks compared to the disconnected network setting. These results illustrate our theoretical results and show the performance of our methods for decentralized Bayesian logistic regression problems.

6. Conclusion

In this paper, we studied DE-SGLD and DE-SGHMC methods which allow scalable Bayesian inference for decentralized learning settings. For both methods, we show that the distribution of the iterate $x_i^{(k)}$ of node i converges linearly (in k) to a neighborhood of the target distribution in the 2-Wasserstein metric when the target density $\pi(x) \propto e^{-f(x)}$ is strongly log-concave (i.e. f is strongly convex) and f is smooth. Our results are non-asymptotic and provide performance bounds for any finite k. We also illustrated the efficiency of our methods on the Bayesian linear regression and Bayesian logistic regression problems.

Acknowledgements

The authors are indebted to Umut Şimşekli for fruitful discussions and for his help with the experiments. The authors are also grateful to the Associate Editor and three anonymous referees for helpful suggestions and comments. Mert Gürbüzbalaban and Yuanhan Hu's research are supported in part by the grants Office of Naval Research Award Number N00014-21-1-2244, National Science Foundation (NSF) CCF-1814888, NSF DMS-2053485, NSF DMS-1723085. Xuefeng Gao acknowledges support from Hong Kong RGC GRF Grants 14201117, 14201520 and 14201421. Lingjiong Zhu is grateful to the partial support from a Simons Foundation Collaboration Grant and the grant NSF DMS-2053454 from the National Science Foundation.

Appendix A. Proofs of Technical Results in Section 3.1

A.1 Proof of Lemma 6

In this proof, we aim to provide an L^2 bound on the gradients $\nabla F\left(x^{(k)}\right)$ that is uniform in k, where F is defined in (14). Let us define

$$F_{\mathcal{W},\eta}(x) := \frac{1}{2\eta} x^T (I - \mathcal{W}) x + F(x). \tag{56}$$

Then $F_{\mathcal{W},\eta}$ is μ-strongly convex and L_η-smooth with $L_\eta = \frac{1-\lambda_N^W}{\eta} + L$, and we can re-write the DE-SGLD iterates as

$$x^{(k+1)} = x^{(k)} - \eta \nabla F_{\mathcal{W},\eta}\left(x^{(k)}\right) - \eta \xi^{(k+1)} + \sqrt{2\eta} w^{(k+1)}, \quad \text{with} \quad \mathcal{W} = W \otimes I_d. \tag{57}$$

Define x_η^* as the minimizer of $F_{\mathcal{W},\eta}$. Since $\nabla F(x)$ is L-Lipschitz, we have

$$\begin{aligned}
\mathbb{E}\left\|\nabla F\left(x^{(k)}\right)\right\|^2 &\leq 2\mathbb{E}\left\|\nabla F\left(x^{(k)}\right) - \nabla F\left(x_\eta^*\right)\right\|^2 + 2\left\|\nabla F\left(x_\eta^*\right)\right\|^2 \\
&\leq 2L^2\mathbb{E}\left[\left\|x^{(k)} - x_\eta^*\right\|^2\right] + 2\left\|\nabla F\left(x_\eta^*\right)\right\|^2 \\
&\leq 2L^2\mathbb{E}\left[\left\|x^{(k)} - x_\eta^*\right\|^2\right] + 4\left\|\nabla F\left(x_\eta^*\right) - \nabla F\left(x^*\right)\right\|^2 + 4\left\|\nabla F\left(x^*\right)\right\|^2 \\
&\leq 2L^2\mathbb{E}\left[\left\|x^{(k)} - x_\eta^*\right\|^2\right] + 4L^2\left\|x_\eta^* - x^*\right\|^2 + 4\left\|\nabla F\left(x^*\right)\right\|^2,
\end{aligned} \tag{58}$$

where we recall from (10) that $x^* = \left[x_*^T, x_*^T, \ldots, x_*^T\right]^T$, where x_* is the minimizer of $f(x)$. Therefore, in order to derive Lemma 6, we need to have a control on $\|x_\eta^* - x^*\|$. This is provided in the following lemma, which follows from Corollary 9 in Yuan et al. (2016).

Lemma 19 *If* $\eta \leq \min(\frac{1+\lambda_N^W}{L}, \frac{1}{L+\mu})$, *then*

$$\|x_\eta^* - x^*\| \leq C_1 \frac{\eta\sqrt{N}}{1-\bar{\gamma}}, \quad \textit{with} \quad \bar{\gamma} := \max\left\{\left|\lambda_2^W\right|, \left|\lambda_N^W\right|\right\},$$

where C_1 *is defined in* (27).

Proof of Lemma 19 The proof of Lemma 19 will be provided in Appendix C. □

Next, to continue with the proof of Lemma 6, we control the term $\mathbb{E}\left[\left\|x^{(k)} - x_\eta^*\right\|^2\right]$ in (58) by deriving a recursion. From (57), we get

$$x^{(k+1)} - x_\eta^* = x^{(k)} - x_\eta^* - \eta \nabla F_{\mathcal{W},\eta}\left(x^{(k)}\right) - \eta \xi^{(k+1)} + \sqrt{2\eta} w^{(k+1)}, \quad \text{with} \quad \mathcal{W} = W \otimes I_d.$$

Since $F_{\mathcal{W},\eta}$ is L_η-smooth and μ-strongly convex, we have

$$\begin{aligned}
L_\eta \langle \nabla F_{\mathcal{W},\eta}(z) - \nabla F_{\mathcal{W},\eta}(y), z - y\rangle &\geq \|\nabla F_{\mathcal{W},\eta}(z) - \nabla F_{\mathcal{W},\eta}(y)\|^2 \quad \forall z, y \in \mathbb{R}^d, & (59)\\
\langle z - y, \nabla F_{\mathcal{W},\eta}(z) - \nabla F_{\mathcal{W},\eta}(y)\rangle &\geq \mu\|z - y\|^2 \quad \forall z, y \in \mathbb{R}^d. & (60)
\end{aligned}$$

If we choose $z = x^{(k)}$ and $y = x_\eta^*$ and use the fact that $\nabla F_{\mathcal{W},\eta}\left(x_\eta^*\right) = 0$, we obtain

$$\begin{aligned}
\mathbb{E}\left[\left\|x^{(k+1)} - x_\eta^*\right\|^2\right] &= \mathbb{E}\left[\left\|x^{(k)} - x_\eta^*\right\|^2\right] - 2\eta\mathbb{E}\left\langle x^{(k)} - x_\eta^*, \nabla F_{\mathcal{W},\eta}\left(x^{(k)}\right)\right\rangle \\
&\qquad + \eta^2\mathbb{E}\left[\left\|\nabla F_{\mathcal{W},\eta}\left(x^{(k)}\right)\right\|^2\right] + \eta^2\mathbb{E}\left\|\xi^{(k+1)}\right\|^2 + 2\eta dN \\
&\leq \mathbb{E}\left[\left\|x^{(k)} - x_\eta^*\right\|^2\right] - 2\eta\left(1 - \frac{\eta L_\eta}{2}\right)\mathbb{E}\left\langle x^{(k)} - x_\eta^*, \nabla F_{\mathcal{W},\eta}\left(x^{(k)}\right)\right\rangle \\
&\qquad + \eta^2\sigma^2 N + 2\eta dN \\
&\leq \left(1 - 2\mu\eta\left(1 - \frac{\eta L_\eta}{2}\right)\right)\mathbb{E}\left[\left\|x^{(k)} - x_\eta^*\right\|^2\right] + \eta^2\sigma^2 N + 2\eta dN \\
&= \left(1 - \mu\eta(1 + \lambda_N^W - \eta L)\right)\mathbb{E}\left[\left\|x^{(k)} - x_\eta^*\right\|^2\right] + \eta^2\sigma^2 N + 2\eta dN,
\end{aligned}$$

where we used $\frac{\eta L_\eta}{2} < 1$ and $\mu\eta(1 + \lambda_N^W - \eta L) \in (0,1)$ which follows directly from our assumptions on the stepsize. Therefore,

$$\mathbb{E}\left[\left\|x^{(k)} - x_\eta^*\right\|^2\right] \leq \left(1 - \mu\eta(1 + \lambda_N^W - \eta L)\right)^k \left\|x^{(0)} - x_\eta^*\right\|^2 + \frac{\eta\sigma^2 N + 2dN}{\mu(1 + \lambda_N^W - \eta L)}.$$

Now we are ready to bound $\mathbb{E}\left\|\nabla F\left(x^{(k)}\right)\right\|^2$. We can compute from (58) that

$$\begin{aligned}
\mathbb{E}\left\|\nabla F\left(x^{(k)}\right)\right\|^2 &\leq 2L^2\mathbb{E}\left[\left\|x^{(k)} - x_\eta^*\right\|^2\right] + 4L^2\left\|x_\eta^* - x^*\right\|^2 + 4\left\|\nabla F\left(x^*\right)\right\|^2 \\
&\leq 2L^2\left(1 - \mu\eta(1 + \lambda_N^W - \eta L)\right)^k \left\|x^{(0)} - x_\eta^*\right\|^2 + \frac{2L^2(\eta\sigma^2 N + 2dN)}{\mu(1 + \lambda_N^W - \eta L)} \\
&\qquad + 4L^2\left\|x_\eta^* - x^*\right\|^2 + 4\left\|\nabla F\left(x^*\right)\right\|^2 \\
&\leq 4L^2\left(1 - \mu\eta(1 + \lambda_N^W - \eta L)\right)^k \left\|x^{(0)} - x^*\right\|^2 \\
&\qquad + 4L^2\left(1 - \mu\eta(1 + \lambda_N^W - \eta L)\right)^k \left\|x^* - x_\eta^*\right\|^2 \\
&\qquad + \frac{2L^2(\eta\sigma^2 N + 2dN)}{\mu(1 + \lambda_N^W - \eta L)} + 4L^2\left\|x_\eta^* - x^*\right\|^2 + 4\left\|\nabla F\left(x^*\right)\right\|^2 \\
&\leq 4L^2\left(1 - \mu\eta(1 + \lambda_N^W - \eta L)\right)^k \left\|x^{(0)} - x^*\right\|^2 + 8L^2\left\|x^* - x_\eta^*\right\|^2 \\
&\qquad + \frac{2L^2(\eta\sigma^2 N + 2dN)}{\mu(1 + \lambda_N^W - \eta L)} + 4\left\|\nabla F\left(x^*\right)\right\|^2,
\end{aligned}$$

where we recall from (10) that $x^* = \left[x_*^T, x_*^T, \ldots, x_*^T\right]^T$ where x_* is the minimizer of $f(x)$. Finally, we apply Lemma 19 to complete the proof of Lemma 6. □

A.2 Proof of Lemma 7

In this proof, we aim to provide uniform L_2 bounds between the iterates $x_i^{(k)}$ and their means $\bar{x}^{(k)}$. First, by the definition of $x^{(k)}$, we get

$$x^{(k+1)} = (W \otimes I_d)x^{(k)} - \eta\nabla F\left(x^{(k)}\right) - \eta\xi^{(k+1)} + \sqrt{2\eta}w^{(k+1)}.$$

It follows that

$$x^{(k)} = (W^k \otimes I_d)x^{(0)} - \eta \sum_{s=0}^{k-1} \left(W^{k-1-s} \otimes I_d\right) \nabla F\left(x^{(s)}\right) - \eta \sum_{s=0}^{k-1} \left(W^{k-1-s} \otimes I_d\right) \xi^{(s+1)} + \sqrt{2\eta} \sum_{s=0}^{k-1} \left(W^{k-1-s} \otimes I_d\right) w^{(s+1)}. \tag{61}$$

Let us define $\bar{\mathbf{x}}^{(k)} := [\bar{x}^{(k)T}, \cdots, \bar{x}^{(k)T}]^T \in \mathbb{R}^{Nd}$. Notice that

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

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

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

Note that it follows from (61) that

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

By the Cauchy-Schwarz inequality, we have

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

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

$$
\begin{aligned}
&\left\|x^{(k)}-\frac{1}{N}\left(\left(1_N1_N^T\right)\otimes I_d\right)x^{(k)}\right\|^2\\
&\le 4\left\|\left(\left(W^k-\frac{1}{N}1_N1_N^T\right)\otimes I_d\right)x^{(0)}\right\|^2\\
&\qquad+4\eta^2\left\|\sum_{s=0}^{k-1}\left(\left(W^{k-1-s}-\frac{1}{N}1_N1_N^T\right)\otimes I_d\right)\nabla F\left(x^{(s)}\right)\right\|^2\\
&\qquad\quad+4\eta^2\left\|\sum_{s=0}^{k-1}\left(\left(W^{k-1-s}-\frac{1}{N}1_N1_N^T\right)\otimes I_d\right)\xi^{(s+1)}\right\|^2\\
&\qquad\quad\quad+8\eta\left\|\sum_{s=0}^{k-1}\left(\left(W^{k-1-s}-\frac{1}{N}1_N1_N^T\right)\otimes I_d\right)w^{(s+1)}\right\|^2. \qquad (62)
\end{aligned}
$$

Note that

$$
\begin{aligned}
&4\eta^2 \left\| \sum_{s=0}^{k-1} \left(\left(W^{k-1-s} - \frac{1}{N} 1_N 1_N^T \right) \otimes I_d \right) \nabla F\left(x^{(s)} \right) \right\|^2 \\
&\leq 4\eta^2 \left(\sum_{s=0}^{k-1} \left\| \left(W^{k-1-s} - \frac{1}{N} 1_N 1_N^T \right) \otimes I_d \right\| \cdot \left\| \nabla F\left(x^{(s)} \right) \right\| \right)^2 \\
&\leq 4\eta^2 \left(\sum_{s=0}^{k-1} \left\| W^{k-1-s} - \frac{1}{N} 1_N 1_N^T \right\| \cdot \left\| \nabla F\left(x^{(s)} \right) \right\| \right)^2 \\
&= 4\eta^2 \left(\sum_{s=0}^{k-1} \bar{\gamma}^{k-1-s} \cdot \left\| \nabla F\left(x^{(s)} \right) \right\| \right)^2 \\
&= 4\eta^2 \left(\sum_{s=0}^{k-1} \bar{\gamma}^{k-1-s} \right)^2 \left(\frac{\sum_{s=0}^{k-1} \bar{\gamma}^{k-1-s} \cdot \left\| \nabla F\left(x^{(s)} \right) \right\|}{\sum_{s=0}^{k-1} \bar{\gamma}^{k-1-s}} \right)^2 \\
&\leq 4\eta^2 \left(\sum_{s=0}^{k-1} \bar{\gamma}^{k-1-s} \right)^2 \sum_{s=0}^{k-1} \frac{\bar{\gamma}^{k-1-s}}{\sum_{s=0}^{k-1} \bar{\gamma}^{k-1-s}} \left\| \nabla F\left(x^{(s)} \right) \right\|^2, \qquad (63)
\end{aligned}
$$

where we used Jensen's inequality in the last step above, and the fact that W^{k-1-s} has eigenvalues $(\lambda_i^W)^{k-1-s}$ with $1 = \lambda_1^W > \lambda_2^W \geq \cdots \geq \lambda_N^W > -1$, and hence $\left\| W^{k-1-s} - \frac{1}{N} 1_N 1_N^T \right\| = \max\{|\lambda_2^W|^{k-1-s}, |\lambda_N^W|^{k-1-s}\} = \bar{\gamma}^{k-1-s}$. Recall from Lemma 6 that for every $k = 0, 1, 2, \ldots$, $\mathbb{E}\left[\left\| \nabla F\left(x^{(k)} \right) \right\|^2 \right] \leq D^2$, where D is defined in (26). Therefore, by (63), we have

$$
\begin{aligned}
&4\eta^2 \mathbb{E}\left[\left\| \sum_{s=0}^{k-1} \left(\left(W^{k-1-s} - \frac{1}{N} 1_N 1_N^T \right) \otimes I_d \right) \nabla F\left(x^{(s)} \right) \right\|^2 \right] \\
&\leq 4D^2\eta^2 \left(\sum_{s=0}^{k-1} \bar{\gamma}^{k-1-s} \right)^2 \sum_{s=0}^{k-1} \frac{\bar{\gamma}^{k-1-s}}{\sum_{s=0}^{k-1} \bar{\gamma}^{k-1-s}} \leq 4D^2\eta^2 \frac{1}{(1-\bar{\gamma})^2}.
\end{aligned}
$$

Similarly, we can show that

$$
\begin{aligned}
4 \left\| \left(\left(W^k - \frac{1}{N} 1_N 1_N^T \right) \otimes I_d \right) x^{(0)} \right\|^2 &\leq 4 \left\| \left(W^k - \frac{1}{N} 1_N 1_N^T \right) \otimes I_d \right\|^2 \left\| x^{(0)} \right\|^2 \\
&\leq 4\bar{\gamma}^{2k} \left\| x^{(0)} \right\|^2.
\end{aligned}
$$

It follows from (62) that

$$\begin{aligned}
&\sum_{i=1}^N \mathbb{E}\left\|x_i^{(k)} - \bar{x}^{(k)}\right\|^2 \\
&= \left\|x^{(k)} - \frac{1}{N}\left(\left(1_N 1_N^T\right)\otimes I_d\right)x^{(k)}\right\|^2 \\
&\le 4\bar{\gamma}^{2k}\mathbb{E}\left\|x^{(0)}\right\|^2 + 4D^2\eta^2\frac{1}{(1-\bar{\gamma})^2} + 4\eta^2\sum_{s=0}^{k-1}\mathbb{E}\left\|\left(\left(W^{k-1-s} - \frac{1}{N}1_N 1_N^T\right)\otimes I_d\right)\xi^{(s+1)}\right\|^2 \\
&\quad + 8\eta\sum_{s=0}^{k-1}\mathbb{E}\left\|\left(\left(W^{k-1-s} - \frac{1}{N}1_N 1_N^T\right)\otimes I_d\right)w^{(s+1)}\right\|^2 \\
&\le 4\bar{\gamma}^{2k}\mathbb{E}\left\|x^{(0)}\right\|^2 + 4D^2\eta^2\frac{1}{(1-\bar{\gamma})^2} + 4\eta^2\sum_{s=0}^{k-1}\left\|W^{k-1-s} - \frac{1}{N}1_N 1_N^T\right\|^2\mathbb{E}\left\|\xi^{(s+1)}\right\|^2 \\
&\quad + 8\eta\sum_{s=0}^{k-1}\left\|W^{k-1-s} - \frac{1}{N}1_N 1_N^T\right\|^2\mathbb{E}\left\|w^{(s+1)}\right\|^2 \\
&\le 4\bar{\gamma}^{2k}\mathbb{E}\left\|x^{(0)}\right\|^2 + 4D^2\eta^2\frac{1}{(1-\bar{\gamma})^2} + 4\sigma^2 N\eta^2\sum_{s=0}^{k-1}\bar{\gamma}^{2(k-1-s)} + 8dN\eta\sum_{s=0}^{k-1}\bar{\gamma}^{2(k-1-s)} \\
&\le 4\bar{\gamma}^{2k}\mathbb{E}\left\|x^{(0)}\right\|^2 + \frac{4D^2\eta^2}{(1-\bar{\gamma})^2} + \frac{4\sigma^2 N\eta^2}{(1-\bar{\gamma}^2)} + \frac{8dN\eta}{(1-\bar{\gamma}^2)}.
\end{aligned}$$

The proof is complete. □

A.3 Proof of Lemma 8

By Lemma 7, we can compute that

$$\begin{aligned}
\mathbb{E}\left\|\mathcal{E}_{k+1}\right\|^2 &= \mathbb{E}\left\|\frac{1}{N}\sum_{i=1}^N\left(\nabla f_i\left(x_i^{(k)}\right) - \nabla f_i\left(\bar{x}^{(k)}\right)\right)\right\|^2 \\
&\le \frac{1}{N^2}\sum_{i=1}^N N\mathbb{E}\left\|\nabla f_i\left(x_i^{(k)}\right) - \nabla f_i\left(\bar{x}^{(k)}\right)\right\|^2 \\
&\le \frac{1}{N}L^2\sum_{i=1}^N\mathbb{E}\left\|x_i^{(k)} - \bar{x}^{(k)}\right\|^2 \\
&\le \frac{4L^2\bar{\gamma}^{2k}}{N}\mathbb{E}\left\|x^{(0)}\right\|^2 + \frac{4L^2D^2\eta^2}{N(1-\bar{\gamma})^2} + \frac{4L^2\sigma^2\eta^2}{(1-\bar{\gamma}^2)} + \frac{8L^2d\eta}{(1-\bar{\gamma}^2)}.
\end{aligned}$$

The proof is complete. □

A.4 Proof of Lemma 9

In this proof, we aim to show that the mean of the iterates $\bar{x}^{(k)}$ which is defined in (28) is close to x_k in L^2 distance, where x_k is defined in (24) which is an Euler-Maruyama discretization of the continuous-time overdamped Langevin SDE in (25).

First, we can compute that

$$\bar{x}^{(k+1)} - x_{k+1} = \bar{x}^{(k)} - x_k - \frac{\eta}{N}\left[\nabla f\left(\bar{x}^{(k)}\right) - \nabla f(x_k)\right] + \eta\mathcal{E}_{k+1} - \eta\bar{\xi}^{(k+1)},$$

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

$$\begin{aligned}
&\left\|\bar{x}^{(k+1)} - x_{k+1}\right\|^2 \\
&= \left\|\bar{x}^{(k)} - x_k - \frac{\eta}{N}\left[\nabla f\left(\bar{x}^{(k)}\right) - \nabla f(x_k)\right]\right\|^2 + \eta^2\left\|\mathcal{E}_{k+1} - \bar{\xi}^{(k+1)}\right\|^2 \\
&\qquad + 2\left\langle \bar{x}^{(k)} - x_k - \frac{\eta}{N}\left[\nabla f\left(\bar{x}^{(k)}\right) - \nabla f(x_k)\right], \eta\mathcal{E}_{k+1} - \eta\bar{\xi}^{(k+1)}\right\rangle \\
&= \left\|\bar{x}^{(k)} - x_k\right\|^2 + \eta^2\left\|\frac{1}{N}\left[\nabla f\left(\bar{x}^{(k)}\right) - \nabla f(x_k)\right]\right\|^2 \\
&\qquad - 2\left\langle \bar{x}^{(k)} - x_k, \eta\frac{1}{N}\left[\nabla f\left(\bar{x}^{(k)}\right) - \nabla f(x_k)\right]\right\rangle + \eta^2\left\|\mathcal{E}_{k+1} - \bar{\xi}^{(k+1)}\right\|^2 \\
&\qquad + 2\left\langle \bar{x}^{(k)} - x_k - \eta\frac{1}{N}\left[\nabla f\left(\bar{x}^{(k)}\right) - \nabla f(x_k)\right], \eta\mathcal{E}_{k+1} - \eta\bar{\xi}^{(k+1)}\right\rangle \\
&\leq \left\|\bar{x}^{(k)} - x_k\right\|^2 + \eta^2 L\left\langle \bar{x}^{(k)} - x_k, \frac{1}{N}\left[\nabla f\left(\bar{x}^{(k)}\right) - \nabla f(x_k)\right]\right\rangle \\
&\qquad - 2\left\langle \bar{x}^{(k)} - x_k, \eta\frac{1}{N}\left[\nabla f\left(\bar{x}^{(k)}\right) - \nabla f(x_k)\right]\right\rangle + \eta^2\left\|\mathcal{E}_{k+1} - \bar{\xi}^{(k+1)}\right\|^2 \\
&\qquad + 2\left\langle \bar{x}^{(k)} - x_k - \eta\frac{1}{N}\left[\nabla f\left(\bar{x}^{(k)}\right) - \nabla f(x_k)\right], \eta\mathcal{E}_{k+1} - \eta\bar{\xi}^{(k+1)}\right\rangle \\
&\leq \left(1 - 2\eta\mu\left(1 - \frac{\eta L}{2}\right)\right)\left\|\bar{x}^{(k)} - x_k\right\|^2 + \eta^2\left\|\mathcal{E}_{k+1} - \bar{\xi}^{(k+1)}\right\|^2 \\
&\qquad + 2\left\langle \bar{x}^{(k)} - x_k - \eta\frac{1}{N}\left[\nabla f\left(\bar{x}^{(k)}\right) - \nabla f(x_k)\right], \eta\mathcal{E}_{k+1} - \eta\bar{\xi}^{(k+1)}\right\rangle, \qquad (64)
\end{aligned}$$

where we used L-smoothness of $\frac{1}{N}f$ to obtain the second term after the first inequality above and μ-strongly convexity of $\frac{1}{N}f$ (inequalities (59)-(60) apply to the function $\frac{1}{N}f$ as well if we replace the smoothness constant L_η by L) and the assumption that $\eta < 2/L$ to obtain the first term after the second inequality above. Note that according to (12), $\bar{\xi}^{(k+1)}$ has mean zero conditional on the natural filtration of the iterates till time k and by Lemma 8,

$$\mathbb{E}\left\|\mathcal{E}_{k+1}\right\|^2 \leq \frac{4L^2\bar{\gamma}^{2k}}{N}\mathbb{E}\left\|x^{(0)}\right\|^2 + \frac{4L^2D^2\eta^2}{N(1-\bar{\gamma})^2} + \frac{4L^2\sigma^2\eta^2}{(1-\bar{\gamma}^2)} + \frac{8L^2d\eta}{(1-\bar{\gamma}^2)}. \qquad (65)$$

Also, we recall from (18) that $\mathbb{E}\left\|\bar{\xi}^{(k+1)}\right\|^2 \leq \frac{\sigma^2}{N}$. By taking expectations in (64) and applying (12), we get

$$\begin{aligned}
&\mathbb{E}\left\|\bar{x}^{(k+1)} - x_{k+1}\right\|^2 \\
&\leq \left(1 - 2\eta\mu\left(1 - \frac{\eta L}{2}\right)\right)\mathbb{E}\left\|\bar{x}^{(k)} - x_k\right\|^2 + \eta^2\mathbb{E}\left\|\mathcal{E}_{k+1} - \bar{\xi}^{(k+1)}\right\|^2 \\
&\qquad + \mathbb{E}\left[2\left\langle \bar{x}^{(k)} - x_k - \eta\frac{1}{N}\left[\nabla f\left(\bar{x}^{(k)}\right) - \nabla f(x_k)\right], \eta\mathcal{E}_{k+1} - \eta\bar{\xi}^{(k+1)}\right\rangle\right] \qquad (66) \\
&= \left(1 - 2\eta\mu\left(1 - \frac{\eta L}{2}\right)\right)\mathbb{E}\left\|\bar{x}^{(k)} - x_k\right\|^2 + \eta^2\mathbb{E}\left\|\mathcal{E}_{k+1}\right\|^2 + \eta^2\mathbb{E}\left\|\bar{\xi}^{(k+1)}\right\|^2 \\
&\qquad + \mathbb{E}\left[2\left\langle \bar{x}^{(k)} - x_k - \eta\frac{1}{N}\left[\nabla f\left(\bar{x}^{(k)}\right) - \nabla f(x_k)\right], \eta\mathcal{E}_{k+1}\right\rangle\right] \qquad (67) \\
&\leq \left(1 - 2\eta\mu\left(1 - \frac{\eta L}{2}\right)\right)\mathbb{E}\left\|\bar{x}^{(k)} - x_k\right\|^2 + \eta^2\mathbb{E}\left\|\mathcal{E}_{k+1}\right\|^2 + \eta^2\frac{\sigma^2}{N} \\
&\qquad + 2(1+\eta L)\eta\mathbb{E}\left[\left\|\bar{x}^{(k)} - x_k\right\| \cdot \left\|\mathcal{E}_{k+1}\right\|\right], \qquad (68)
\end{aligned}$$

where we used L-smoothness of $\frac{1}{N}f$. For any $x, y \geq 0$ and $c > 0$, we have the inequality $2xy \leq cx^2 + \frac{y^2}{c}$. Applying this inequality with $c = \frac{\mu(1-\frac{\eta L}{2})}{1+\eta L}$ to (68), we obtain

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

where we note that the leading term $1 - \eta\mu\left(1 - \frac{\eta L}{2}\right) \in [0, 1)$ by our assumption on stepsize η. By applying (65), we get

$$\begin{aligned}
&\mathbb{E}\left\|\bar{x}^{(k+1)} - x_{k+1}\right\|^2 \\
&\leq \left(1 - \eta\mu\left(1 - \frac{\eta L}{2}\right)\right)\mathbb{E}\left\|\bar{x}^{(k)} - x_k\right\|^2 \\
&\quad + \eta\left(\eta + \frac{(1+\eta L)^2}{\mu(1-\frac{\eta L}{2})}\right)\left(\frac{4L^2\bar{\gamma}^{2k}}{N}\mathbb{E}\left\|x^{(0)}\right\|^2 + \frac{4L^2D^2\eta^2}{N(1-\bar{\gamma})^2} + \frac{4L^2\sigma^2\eta^2}{(1-\bar{\gamma}^2)} + \frac{8L^2d\eta}{(1-\bar{\gamma}^2)}\right) + \eta^2\frac{\sigma^2}{N},
\end{aligned}$$

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

$$\begin{aligned}
&\mathbb{E}\left\|\bar{x}^{(k)} - x_k\right\|^2 \\
&\leq \sum_{i=0}^{k-1}\left(1-\eta\mu\left(1-\frac{\eta L}{2}\right)\right)^i \\
&\qquad \cdot\left(\eta\left(\eta+\frac{(1+\eta L)^2}{\mu(1-\frac{\eta L}{2})}\right)\left(\frac{4L^2D^2\eta^2}{N(1-\bar{\gamma})^2}+\frac{4L^2\sigma^2\eta^2}{(1-\bar{\gamma}^2)}+\frac{8L^2d\eta}{(1-\bar{\gamma}^2)}\right)+\eta^2\frac{\sigma^2}{N}\right) \\
&\quad + \sum_{i=0}^{k-1}\left(1-\eta\mu\left(1-\frac{\eta L}{2}\right)\right)^i \eta\left(\eta+\frac{(1+\eta L)^2}{\mu(1-\frac{\eta L}{2})}\right)\frac{4L^2\bar{\gamma}^{2(k-i)}}{N}\mathbb{E}\left\|x^{(0)}\right\|^2 \\
&= \frac{1-\left(1-\eta\mu\left(1-\frac{\eta L}{2}\right)\right)^k}{1-\left(1-\eta\mu\left(1-\frac{\eta L}{2}\right)\right)} \\
&\qquad \cdot\left(\eta\left(\eta+\frac{(1+\eta L)^2}{\mu(1-\frac{\eta L}{2})}\right)\left(\frac{4L^2D^2\eta^2}{N(1-\bar{\gamma})^2}+\frac{4L^2\sigma^2\eta^2}{(1-\bar{\gamma}^2)}+\frac{8L^2d\eta}{(1-\bar{\gamma}^2)}\right)+\eta^2\frac{\sigma^2}{N}\right) \\
&\quad + \frac{\bar{\gamma}^{2k}-\left(1-\eta\mu\left(1-\frac{\eta L}{2}\right)\right)^k}{1-\left(1-\eta\mu\left(1-\frac{\eta L}{2}\right)\right)(\bar{\gamma})^{-2}}\frac{4L^2}{N}\mathbb{E}\left\|x^{(0)}\right\|^2.
\end{aligned}$$

Hence[8], for every k.

$$\begin{aligned}
\mathbb{E}\left\|\bar{x}^{(k)} - x_k\right\|^2 &\leq \frac{\eta\left(\eta+\frac{(1+\eta L)^2}{\mu(1-\frac{\eta L}{2})}\right)\left(\frac{4L^2D^2\eta^2}{N(1-\bar{\gamma})^2}+\frac{4L^2\sigma^2\eta^2}{(1-\bar{\gamma}^2)}+\frac{8L^2d\eta}{(1-\bar{\gamma}^2)}\right)+\eta^2\frac{\sigma^2}{N}}{1-\left(1-\eta\mu\left(1-\frac{\eta L}{2}\right)\right)} \\
&\qquad + \frac{\bar{\gamma}^{2k}-\left(1-\eta\mu\left(1-\frac{\eta L}{2}\right)\right)^k}{1-\left(1-\eta\mu\left(1-\frac{\eta L}{2}\right)\right)(\bar{\gamma})^{-2}}\frac{4L^2}{N}\mathbb{E}\left\|x^{(0)}\right\|^2 \\
&= \frac{\eta\left(\eta+\frac{(1+\eta L)^2}{\mu(1-\frac{\eta L}{2})}\right)\left(\frac{4L^2D^2\eta}{N(1-\bar{\gamma})^2}+\frac{4L^2\sigma^2\eta}{(1-\bar{\gamma}^2)}+\frac{8L^2d}{(1-\bar{\gamma}^2)}\right)+\eta\frac{\sigma^2}{N}}{\mu\left(1-\frac{\eta L}{2}\right)} \\
&\qquad + \frac{\bar{\gamma}^{2k}-\left(1-\eta\mu\left(1-\frac{\eta L}{2}\right)\right)^k}{\bar{\gamma}^2-1+\eta\mu\left(1-\frac{\eta L}{2}\right)}\frac{4L^2\bar{\gamma}^2}{N}\mathbb{E}\left\|x^{(0)}\right\|^2.
\end{aligned}$$

The proof is complete. □

8. We recall that the last term is proportional to the ratio $h(x,y) = \frac{x^k-y^k}{x-y}$ with $x = \bar{\gamma}^2$ and $y = \left(1-\eta\mu\left(1-\frac{\eta L}{2}\right)\right)$ and according to our notation (see Section 2), we interpret this ratio as ky^{k-1} in the special case when $x = y$.

A.5 Proof of Lemma 11

The proof of Lemma 11 will be provided in Appendix C. □

Appendix B. Proofs of Technical Results in Section 4.1

B.1 Proof of Lemma 14

In this proof, we aim to provide uniform L^2 bounds on the iterates $v^{(k)}, x^{(k)}$ in (34)–(35).

Based on the expression (34) for $v^{(k+1)}$, first we rewrite the DE-SGHMC iterates (35) for $k \geq 1$ as

$$\begin{aligned} x^{(k+1)} &= \mathcal{W}x^{(k)} + \eta v^{(k+1)}, \\ &= \mathcal{W}x^{(k)} + \eta\left(v^{(k)} - \eta\left[\gamma v^{(k)} + \nabla F\left(x^{(k)}\right) + \xi^{(k+1)}\right] + \sqrt{2\gamma\eta}w^{(k+1)}\right) \\ &= \mathcal{W}x^{(k)} - \eta^2\nabla F\left(x^{(k)}\right) + \eta(1-\gamma\eta)v^{(k)} + \Delta^{(k+1)} \\ &= \mathcal{W}x^{(k)} - \eta^2\nabla F\left(x^{(k)}\right) + (1-\gamma\eta)\left(x^{(k)} - \mathcal{W}x^{(k-1)}\right) + \Delta^{(k+1)}, \end{aligned} \tag{69}$$

where

$$\Delta^{(k+1)} := -\eta^2\xi^{(k+1)} + \eta\sqrt{2\gamma\eta}w^{(k+1)}.$$

If we consider

$$\alpha = \eta^2, \tag{70}$$

then (69) is equivalent to

$$\begin{aligned} x^{(k+1)} &= \mathcal{W}x^{(k)} - \alpha\nabla F\left(x^{(k)}\right) + \beta\left(x^{(k)} - \mathcal{W}x^{(k-1)}\right) + \Delta^{(k+1)} \\ &= x^{(k)} - \alpha\nabla\bar{F}\left(x^{(k)}\right) + \beta\left(x^{(k)} - x^{(k-1)}\right) + \bar{\Delta}^{(k+1)}, \end{aligned} \tag{71}$$

where $\beta = 1 - \gamma\eta$ and

$$\bar{F}(x) := F(x) + \frac{1}{2\alpha}x^T(I-\mathcal{W})x, \quad \bar{\Delta}^{(k+1)} := \Delta^{(k+1)} + \beta(I-\mathcal{W})x^{(k-1)}.$$

Let x_α^* be the unique minimizer of $\bar{F}(x)$. Since $\alpha > 0$, the function $\bar{F}(x)$ is strongly convex with parameter μ and smooth with parameter

$$L_\alpha = L + \frac{1-\lambda_N^W}{\alpha}. \tag{72}$$

In the special case, $\bar{\Delta}^{(k+1)} = 0$, the iterations (71) would exactly coincide with the iterations of the heavy-ball method of Polyak applied to the function $\bar{F}(x)$ with momentum parameter β. Therefore, we can view the iterations (71) as a perturbed heavy-ball method with perturbation $\bar{\Delta}^{(k+1)}$ at iteration k. For the heavy-ball method, linear convergence to the optimum of $\bar{F}(x)$ is obtained if the parameters α and β are properly chosen. In the rest of the proof, we will extend the proof technique of Ghadimi et al. (2015) for the convergence of the heavy-ball method to allow perturbations $\bar{\Delta}^{(k+1)}$ and show that the second moments

of the iterates remain bounded. First of all, we notice that the assumptions (37)–(38) on the choice of η and β can be restated in terms of conditions on $\alpha = \eta^2$ as follows:

$$\alpha \in \left(0, \frac{1+\lambda_N^W}{2(L+\mu)}\right], \tag{73}$$

$$0 \leq \beta \leq \frac{1+\lambda_N^W - 4\alpha\mu}{4}, \tag{74}$$

$$\beta^2 \leq c_1 \mu^3 \alpha^3 \frac{(1+\lambda_N^W)}{64}, \tag{75}$$

where we see after a straightforward computation that the constants c_1 defined by (38) can be rewritten in terms of the smoothness constant L_α as

$$c_1 = \frac{1}{2} \frac{\alpha\mu L_\alpha}{(1-\beta)(L_\alpha+\mu) + 2L_\alpha\beta}. \tag{76}$$

In particular, the condition (74) implies $\beta \in [0, \frac{1}{2})$ due to the fact that $\lambda_N^W < 1$ and $\alpha, \mu > 0$. Next, we introduce

$$p^{(k)} = \frac{\beta}{1-\beta}\left(x^{(k)} - x^{(k-1)}\right), \tag{77}$$

for $k \geq 1$. From the update rule (71), it follows that

$$\begin{aligned} x^{(k+1)} + p^{(k+1)} &= \frac{1}{1-\beta} x^{(k+1)} - \frac{\beta}{1-\beta} x^{(k)} \\ &= x^{(k)} + p^{(k)} - \frac{\alpha}{1-\beta} \nabla \bar{F}\left(x^{(k)}\right) + \frac{1}{1-\beta} \bar{\Delta}^{(k+1)}. \end{aligned}$$

This implies that

$$\begin{aligned} &\left\| x^{(k+1)} + p^{(k+1)} - x_\alpha^* \right\|^2 \\ &= \left\| x^{(k)} + p^{(k)} - x_\alpha^* \right\|^2 - \frac{2\alpha}{1-\beta} \left\langle x^{(k)} - x_\alpha^*, \nabla \bar{F}\left(x^{(k)}\right) \right\rangle \\ &\qquad + \frac{\alpha^2}{(1-\beta)^2} \left\| \nabla \bar{F}\left(x^{(k)}\right) \right\|^2 - \frac{2\alpha\beta}{(1-\beta)^2} \left\langle x^{(k)} - x^{(k-1)}, \nabla \bar{F}\left(x^{(k)}\right) \right\rangle \\ &\quad + \frac{1}{(1-\beta)^2} \left\| \bar{\Delta}^{(k+1)} \right\|^2 + 2 \left\langle x^{(k)} + p^{(k)} - \frac{\alpha}{1-\beta} \nabla \bar{F}\left(x^{(k)}\right) - x_\alpha^*, \frac{1}{(1-\beta)} \bar{\Delta}^{(k+1)} \right\rangle, \end{aligned}$$

where we used the definition (77) of $p^{(k)}$. Next, we bound the last two terms by applying the Cauchy-Schwarz inequality:

$$\begin{aligned} \mathbb{E}_k \left[\frac{1}{(1-\beta)^2} \left\| \bar{\Delta}^{(k+1)} \right\|^2 \right] &\leq \mathbb{E}_k \left[\frac{1}{(1-\beta)^2} \left(2 \left\| \Delta^{(k+1)} \right\|^2 + 2 \left\| \beta (I - \mathcal{W}) x^{(k-1)} \right\|^2 \right) \right] \\ &\leq \frac{2}{(1-\beta)^2} \mathbb{E}_k \left\| \Delta^{(k+1)} \right\|^2 + \frac{2\beta^2}{(1-\beta)^2} \left(1 - \lambda_N^W\right)^2 \left\| x^{(k-1)} \right\|^2 \\ &\leq \frac{2}{(1-\beta)^2} \left(\eta^4 \sigma^2 N + \eta^3 2\gamma N d \right) + \frac{2\beta^2}{(1-\beta)^2} \left(1 - \lambda_N^W\right)^2 \left\| x^{(k-1)} \right\|^2, \end{aligned}$$

where $\mathbb{E}_k$ denotes the conditional expectation with respect to the natural filtration up to time k (which includes the history of the iterations up to (and including) $x^{(k)}$). Similarly,

$$\begin{aligned}
&\mathbb{E}_k\left[2\left\langle x^{(k)}+p^{(k)}-\frac{\alpha}{1-\beta}\nabla\bar{F}\left(x^{(k)}\right)-x_\alpha^*,\frac{1}{(1-\beta)}\bar{\Delta}^{(k+1)}\right\rangle\right]\\
&=2\left\langle x^{(k)}+p^{(k)}-\frac{\alpha}{1-\beta}\nabla\bar{F}\left(x^{(k)}\right)-x_\alpha^*,\frac{1}{(1-\beta)}\beta(I-\mathcal{W})x^{(k-1)}\right\rangle\\
&\leq c_1\left\|x^{(k)}+p^{(k)}-\frac{\alpha}{1-\beta}\nabla\bar{F}\left(x^{(k)}\right)-x_\alpha^*\right\|^2+\frac{1}{c_1}\frac{\beta^2}{(1-\beta)^2}\left(1-\lambda_N^W\right)^2\left\|x^{(k-1)}\right\|^2,
\end{aligned}$$

where we use Cauchy-Schwarz and c_1 is the constant given by (76). Combining everything,

$$\begin{aligned}
&\mathbb{E}_k\left\|x^{(k+1)}+p^{(k+1)}-x_\alpha^*\right\|^2\\
&\leq\left\|x^{(k)}+p^{(k)}-x_\alpha^*\right\|^2-\frac{2\alpha}{1-\beta}\left\langle x^{(k)}-x_\alpha^*,\nabla\bar{F}\left(x^{(k)}\right)\right\rangle\\
&\quad+\frac{\alpha^2}{(1-\beta)^2}\left\|\nabla\bar{F}\left(x^{(k)}\right)\right\|^2-\frac{2\alpha\beta}{(1-\beta)^2}\left\langle x^{(k)}-x^{(k-1)},\nabla\bar{F}\left(x^{(k)}\right)\right\rangle+E^{(k+1)},
\end{aligned}\tag{78}$$

where

$$\begin{aligned}
E^{(k+1)}&:=\frac{2}{(1-\beta)^2}(\eta^4\sigma^2N+\eta^3 2\gamma Nd)+c_1\left\|x^{(k)}+p^{(k)}-\frac{\alpha}{1-\beta}\nabla\bar{F}\left(x^{(k)}\right)-x_\alpha^*\right\|^2\\
&\qquad+\left(2+\frac{1}{c_1}\right)\frac{\beta^2}{(1-\beta)^2}\left(1-\lambda_N^W\right)^2\left\|x^{(k-1)}\right\|^2\\
&\leq\frac{2}{(1-\beta)^2}\left(\eta^4\sigma^2N+\eta^3 2\gamma Nd\right)\\
&\quad+2c_1\left\|x^{(k)}+p^{(k)}-x_\alpha^*\right\|^2+2c_1\frac{\alpha^2}{(1-\beta)^2}\left\|\nabla\bar{F}\left(x^{(k)}\right)\right\|^2\\
&\qquad+\left(2+\frac{1}{c_1}\right)\frac{\beta^2}{(1-\beta)^2}(1-\lambda_N^W)^2\left(2\left\|x^{(k-1)}-x_\alpha^*\right\|^2+2\left\|x_\alpha^*\right\|^2\right),
\end{aligned}$$

and in the last step we used the Cauchy-Schwarz inequality, i.e.

$$\left\|x^{(k-1)}\right\|^2\leq 2\left\|x^{(k-1)}-x_\alpha^*\right\|^2+2\left\|x_\alpha^*\right\|^2.$$

Since $\bar{F}$ is μ-strongly convex and L_α smooth, we have also

$$\begin{aligned}
&\frac{\mu L_\alpha}{L_\alpha+\mu}\left\|x^{(k)}-x_\alpha^*\right\|^2+\frac{1}{L_\alpha+\mu}\left\|\nabla\bar{F}\left(x^{(k)}\right)\right\|^2\leq\left\langle x^{(k)}-x_\alpha^*,\nabla\bar{F}\left(x^{(k)}\right)\right\rangle,\\
&\bar{F}\left(x^{(k)}\right)-\bar{F}\left(x^{(k-1)}\right)+\frac{\mu}{2}\left\|x^{(k)}-x^{(k-1)}\right\|^2\leq\left\langle x^{(k)}-x^{(k-1)},\nabla\bar{F}\left(x^{(k)}\right)\right\rangle,
\end{aligned}$$

(see e.g. Nesterov (2013)). These inequalities combined with (78) implies

$$\begin{aligned}
&\frac{2\alpha\beta}{(1-\beta)^2}\left(\bar{F}\left(x^{(k)}\right)-\bar{F}^*\right)+\mathbb{E}_k\left\|x^{(k+1)}+p^{(k+1)}-x_\alpha^*\right\|^2\\
&\leq \frac{2\alpha\beta}{(1-\beta)^2}\left(\bar{F}\left(x^{(k-1)}\right)-\bar{F}^*\right)+\left\|x^{(k)}+p^{(k)}-x_\alpha^*\right\|^2-\frac{2\alpha\mu L_\alpha}{(1-\beta)(L_\alpha+\mu)}\left\|x^{(k)}-x_\alpha^*\right\|^2\\
&\quad-\frac{\alpha\beta\mu}{(1-\beta)^2}\left\|x^{(k)}-x^{(k-1)}\right\|^2+\frac{\alpha}{(1-\beta)}\left(\frac{\alpha}{1-\beta}-\frac{2}{L_\alpha+\mu}\right)\left\|\nabla\bar{F}\left(x^{(k)}\right)\right\|^2+E^{(k+1)}.
\end{aligned}$$

Plugging the upper bound for $E^{(k+1)}$, we obtain

$$\begin{aligned}
&\frac{2\alpha\beta}{(1-\beta)^2}\left(\bar{F}\left(x^{(k)}\right)-\bar{F}^*\right)+\left\|x^{(k+1)}+p^{(k+1)}-x_\alpha^*\right\|^2\\
&\leq \frac{2\alpha\beta}{(1-\beta)^2}\left(\bar{F}\left(x^{(k-1)}\right)-\bar{F}^*\right)\\
&\quad+\left\|x^{(k)}+p^{(k)}-x_\alpha^*\right\|^2(1+2c_1)-\frac{2\alpha\mu L_\alpha}{(1-\beta)(L_\alpha+\mu)}\left\|x^{(k)}-x_\alpha^*\right\|^2\\
&\quad-\frac{\alpha\beta\mu}{(1-\beta)^2}\left\|x^{(k)}-x^{(k-1)}\right\|^2+\frac{\alpha}{(1-\beta)}\left(\frac{\alpha}{1-\beta}-\frac{2}{L_\alpha+\mu}+\frac{2c_1\alpha}{1-\beta}\right)\left\|\nabla\bar{F}\left(x^{(k)}\right)\right\|^2\\
&\quad+\frac{2}{(1-\beta)^2}\left(\eta^4\sigma^2N+\eta^3 2\gamma Nd\right)\\
&\quad+\frac{2\beta^2}{(1-\beta)^2}\left(1-\lambda_N^W\right)^2\left(1+\frac{1}{2c_1}\right)\left(2\left\|x^{(k-1)}-x_\alpha^*\right\|^2+2\left\|x_\alpha^*\right\|^2\right). \qquad (79)
\end{aligned}$$

By Lemma 20, the coefficient in front of $\left\|\nabla\bar{F}\left(x^{(k)}\right)\right\|^2$ is negative, i.e.

$$k_{\alpha,\beta}:=\frac{\alpha}{(1-\beta)}\left(\frac{\alpha}{1-\beta}-\frac{2}{L_\alpha+\mu}+\frac{2c_1\alpha}{1-\beta}\right)<0. \qquad (80)$$

We then move the term with $\left\|\nabla\bar{F}\left(x^{(k)}\right)\right\|^2$ to the left-hand side of (79) to obtain

$$\begin{aligned}
&\frac{2\alpha\beta}{(1-\beta)^2}\left(\bar{F}\left(x^{(k)}\right)-\bar{F}^*\right)+\left\|x^{(k+1)}+p^{(k+1)}-x_\alpha^*\right\|^2-k_{\alpha,\beta}\left\|\nabla\bar{F}(x^{(k)})\right\|^2\\
&\leq \frac{2\alpha\beta}{(1-\beta)^2}\left(\bar{F}\left(x^{(k-1)}\right)-\bar{F}^*\right)+\left\|x^{(k)}+p^{(k)}-x_\alpha^*\right\|^2(1+2c_1)\\
&\quad-\frac{2\alpha\mu L_\alpha}{(1-\beta)(L_\alpha+\mu)}\left\|x^{(k)}-x_\alpha^*\right\|^2-\frac{\alpha\beta\mu}{(1-\beta)^2}\left\|x^{(k)}-x^{(k-1)}\right\|^2\\
&\quad+\frac{2}{(1-\beta)^2}\left(\eta^4\sigma^2N+\eta^3 2\gamma Nd\right)\\
&\quad+\frac{2\beta^2}{(1-\beta)^2}\left(1-\lambda_N^W\right)^2\left(1+\frac{1}{2c_1}\right)\left(2\left\|x^{(k-1)}-x_\alpha^*\right\|^2+2\left\|x_\alpha^*\right\|^2\right). \qquad (81)
\end{aligned}$$

By standard inequalities for μ-strongly convex functions from Nesterov (2013, Section 2.1), also

$$\begin{aligned}
2\mu\left(\bar{F}\left(x^{(k)}\right)-\bar{F}^*\right) &\leq \left\|\nabla\bar{F}\left(x^{(k)}\right)\right\|^2, \qquad (82)\\
\left\|x^{(k-1)}-x_\alpha^*\right\|^2 &\leq \frac{2}{\mu}\left[\bar{F}\left(x^{(k-1)}\right)-\bar{F}\left(x_\alpha^*\right)\right], \qquad (83)
\end{aligned}$$

where $\bar{F}^* := \bar{F}(x_\alpha^*)$ is the global minimum of $\bar{F}$. In particular, by multiplying both sides of the first inequality (82) with $-k_{\alpha,\beta} > 0$ we obtain

$$-2k_{\alpha,\beta}\mu\left(\bar{F}\left(x^{(k)}\right) - \bar{F}^*\right) \leq -k_{\alpha,\beta}\left\|\nabla\bar{F}\left(x^{(k)}\right)\right\|^2. \tag{84}$$

Inserting the estimates (83) and (84) into (81), we obtain

$$\begin{aligned} &b\left(\bar{F}\left(x^{(k)}\right) - \bar{F}^*\right) + \left\|x^{(k+1)} + p^{(k+1)} - x_\alpha^*\right\|^2 \\ &\leq a\left(\bar{F}\left(x^{(k-1)}\right) - \bar{F}^*\right) + \left\|x^{(k)} + p^{(k)} - x_\alpha^*\right\|^2 (1+2c_1) - \frac{2\alpha\mu L_\alpha}{(1-\beta)(L_\alpha+\mu)}\left\|x^{(k)} - x_\alpha^*\right\|^2 \\ &\quad - \frac{\alpha\beta\mu}{(1-\beta)^2}\left\|x^{(k)} - x^{(k-1)}\right\|^2 + \frac{2}{(1-\beta)^2}\left(\eta^4\sigma^2 N + \eta^3 2\gamma N d\right) + c_2\|x_\alpha^*\|^2, \end{aligned}$$

where

$$a := \frac{2\alpha\beta}{(1-\beta)^2} + \frac{2c_2}{\mu}, \tag{85}$$

$$b := \frac{2\alpha\beta}{(1-\beta)^2} - 2k_{\alpha,\beta}\mu = \frac{2\alpha}{(1-\beta)}\left(\frac{\beta-\mu\alpha}{1-\beta} + \frac{2\mu}{L_\alpha+\mu} - \frac{2c_1\alpha\mu}{1-\beta}\right), \tag{86}$$

where $k_{\alpha,\beta}$ is defined by (80), and

$$c_2 := \frac{4\beta^2}{(1-\beta)^2}\left(1-\lambda_N^W\right)^2\left(1+\frac{1}{c_1}\right).$$

We can also write

$$\left\|x^{(k)} + p^{(k)} - x_\alpha^*\right\|^2 = \left[z^{(k)}\right]^T M z^{(k)},$$

where

$$z^{(k)} = \left[x^{(k)} - x_\alpha^*, x^{(k)} - x^{(k-1)}\right]^T, \qquad \text{and} \qquad M = \begin{bmatrix} I_d & \frac{\beta}{1-\beta}I_d \\ \frac{\beta}{1-\beta}I_d & \frac{\beta^2}{(1-\beta)^2}I_d \end{bmatrix}.$$

Therefore, we can write

$$\begin{aligned} &b\left(\bar{F}\left(x^{(k)}\right) - \bar{F}^*\right) + \mathbb{E}_k\left[\left(z^{(k+1)}\right)^T M\left(z^{(k+1)}\right)\right] \\ &\leq a\left(\bar{F}\left(x^{(k-1)}\right) - \bar{F}^*\right) + \left(z^{(k)}\right)^T Q\left(z^{(k)}\right) + \frac{2}{(1-\beta)^2}\left(\eta^4\sigma^2 N + \eta^3 2\gamma N d\right) + c_2\|x_\alpha^*\|^2, \end{aligned}$$

where

$$Q := \begin{bmatrix} \left(1+2c_1 - \frac{2\alpha\mu L_\alpha}{(1-\beta)(L+\mu)}\right) I_d & \frac{\beta}{(1-\beta)}(1+2c_1)I_d \\ \frac{\beta}{(1-\beta)}(1+2c_1)I_d & \left(\frac{(1+2c_1)\beta^2 - \alpha\beta\mu}{(1-\beta)^2}\right) I_d \end{bmatrix}. \tag{87}$$

By Lemma 21 and Lemma 22, for a given positive scalar s, we have $sM \succeq Q$ as long as $s \geq q_2$ and $a \leq sb$ as long as $s \geq q_1$ where $q_1, q_2 \in [0,1)$ and are defined by (106) and (107)

respectively. If we introduce $q := \max(q_1, q_2)$ and choose $s = q$, then we obtain

$$\begin{aligned} & b\left(\bar{F}\left(x^{(k)}\right) - \bar{F}^*\right) + \mathbb{E}_k\left[\left(z^{(k+1)}\right)^T M\left(z^{(k+1)}\right)\right] \\ & \leq q\left(b\left(\bar{F}\left(x^{(k-1)}\right) - \bar{F}^*\right) + \left(z^{(k)}\right)^T M\left(z^{(k)}\right)\right) \\ & \qquad + \frac{2}{(1-\beta)^2}\left(\eta^4\sigma^2 N + \eta^3 2\gamma N d\right) + c_2\left\|x_\alpha^*\right\|^2. \end{aligned}$$

Let us introduce the Lyapunov function

$$\begin{aligned} V_{k+1} &:= \mathbb{E}\left[b\left(\bar{F}\left(x^{(k)}\right) - \bar{F}^*\right) + \left(z^{(k+1)}\right)^T M\left(z^{(k+1)}\right)\right] \\ &= \mathbb{E}\left[b\left(\bar{F}\left(x^{(k)}\right) - \bar{F}^*\right) + \left(z^{(k+1)}\right)^T M\left(z^{(k+1)}\right)\right], \end{aligned}$$

for $k \geq 0$. Then, taking expectations, we get

$$V_{k+1} \leq qV_k + \frac{2}{(1-\beta)^2}\left(\eta^4\sigma^2 N + \eta^3 2\gamma N d\right) + c_2\left\|x_\alpha^*\right\|^2.$$

This recursion implies that

$$\begin{aligned} & b\mathbb{E}\left[\bar{F}\left(x^{(k)}\right) - \bar{F}^*\right] + \mathbb{E}\left\|x^{(k+1)} + p^{(k+1)} - x_\alpha^*\right\|^2 = V_{k+1} \\ & \leq V_1 q^k + \frac{1}{1-q}\left(\frac{2}{(1-\beta)^2}\left(\eta^4\sigma^2 N + \eta^3 2\gamma N d\right) + c_2\left\|x_\alpha^*\right\|^2\right), \end{aligned} \tag{88}$$

where we recall that $q = \max(q_1, q_2)$. From the representation, (38), and the fact that $\alpha = \eta^2$ we observe that

$$\frac{1}{2}\frac{\alpha\mu}{(1+\beta) + (1-\beta)\left(\frac{\eta^2\mu}{1-\lambda_N^W + \eta^2 L}\right)} = c_1 \leq \frac{1}{2}\frac{\alpha\mu}{(1+\beta)}. \tag{89}$$

Using the fact that the function $h(\alpha) := \frac{\alpha\mu}{1-\lambda_N^W + \alpha L}$ is monotonically increasing on the positive real line and by the assumption (37), we obtain

$$h(\alpha) \leq h\left(\frac{1+\lambda_N^W}{2(L+\mu)}\right) = \frac{(1+\lambda_N^W)\mu}{(1+\lambda_N^W)L + 2(1-\lambda_N^W)(L+\mu)} = \Theta(1).$$

Therefore, we also have the following lower bound for c_1:

$$\frac{1}{2}\frac{\alpha\mu}{(1+\beta) + (1-\beta)\Theta(1)} \leq c_1. \tag{90}$$

It follows then from (90) and (89) that

$$c_1 = \Theta(\alpha) = \Theta\left(\eta^2\right). \tag{91}$$

Consequently, by our assumption (38), we have

$$\beta = 1 - \eta\gamma = \mathcal{O}\left(\eta^3\sqrt{c_1}\right) = \mathcal{O}\left(\eta^4\right) = \mathcal{O}\left(\alpha^2\right), \quad \gamma\eta = 1 - \beta = \Theta(1). \tag{92}$$

Then, it follows from the definition of c_2 that

$$c_2 = \Theta\left(\beta^2\left(1 + \frac{1}{\alpha}\right)\right) = \mathcal{O}\left(\beta^2 + \frac{\beta^2}{\alpha}\right).$$

Since $\beta = \mathcal{O}(\alpha^2)$ by (92), we obtain

$$c_2 = \mathcal{O}\left(\alpha^3\right). \tag{93}$$

This also implies that

$$V_1 = b\mathbb{E}\left[\bar{F}\left(x^{(0)}\right) - \bar{F}^*\right] + \mathbb{E}\left\|x^{(0)} + \frac{\beta}{1-\beta}\left(x^{(1)} - x^{(0)}\right) - x_\alpha^*\right\|^2 \tag{94}$$

$$\leq bL_\alpha\left\|x^{(0)} - x_\alpha^*\right\|^2 + 2\mathbb{E}\left\|x^{(0)} - x_\alpha^*\right\|^2 + \frac{2\beta^2}{(1-\beta)^2}\mathbb{E}\left\|x^{(1)} - x^{(0)}\right\|^2, \tag{95}$$

where

$$b = \frac{2\alpha}{1-\beta}\left(\frac{\beta - \mu\alpha}{1-\beta} + \frac{2\mu}{L_\alpha + \mu} - \frac{2c_1\alpha\mu}{1-\beta}\right) = \Theta\left(\alpha^2\right), \tag{96}$$

where we used the fact that $\beta = \mathcal{O}(\alpha^2)$, $1 - \lambda_N^W = \Theta(1)$ and (91). Then,

$$bL_\alpha = b\left(\frac{1-\lambda_N^W}{\alpha} + L\right) = \mathcal{O}(\alpha).$$

From the definition of a given in (85), we have also $a = \mathcal{O}\left(\alpha\beta + c_2\right) = \mathcal{O}\left(\alpha^3\right)$, where we used $\beta = \Theta(\alpha^2)$ and $c_2 = \mathcal{O}(\alpha^3)$ obtained in (93). Then, it follows from (96) that

$$q_1 = \frac{a}{b} = \mathcal{O}(\alpha). \tag{98}$$

We also see that

$$q_2 = \max(0, 1 - 2c_1) = 1 - \Theta(\alpha), \tag{99}$$

due to (91). This estimate and (98) implies

$$q = \max(q_1, q_2) = 1 - \Theta(\alpha), \quad \frac{1}{1-q} = \frac{1}{\Theta(\alpha)} = \Theta\left(\frac{1}{\eta^2}\right). \tag{100}$$

On the other hand, a consequence of Lemma 19 is that

$$\|x_\alpha^*\| = \mathcal{O}(1). \tag{101}$$

Then, we get from (97) and (95) that $V_1 = \mathcal{O}(1)$. Combining this fact with the estimates (91), (93), (92), (98), (99), (100) we conclude that the terms on the right-hand side of (88) satisfy $V_1 q^k = \Theta\left(\left(1 - \Theta\left(\eta^2\right)\right)^k\right)$, and

$$\frac{1}{1-q}\left(\frac{2}{(1-\beta)^2}\left(\eta^4\sigma^2 N + \eta^3 2\gamma N d\right) + c_2\|x_\alpha^*\|^2\right) = \frac{\mathcal{O}(\eta^3)}{\Theta(\eta^2)} = \mathcal{O}(\eta)\,,$$
$$V_{k+1} = b\mathbb{E}\left[\bar{F}\left(x^{(k)}\right) - \bar{F}^*\right] + \mathbb{E}\left\|x^{(k+1)} + p^{(k+1)} - x_\alpha^*\right\|^2 \leq c_3\,,$$

for some constant $c_3 = \mathcal{O}(1)$ and every $k \geq 0$. Resorting to the estimate (101), we conclude that this proves (45). The last inequality also implies that

$$\mathbb{E}\left\|x^{(k)} - x_\alpha^*\right\|^2 w \leq \frac{1}{\mu}\mathbb{E}\left[\bar{F}\left(x^{(k)}\right) - \bar{F}^*\right] \leq \frac{c_3}{\mu b}\,, \tag{102}$$

as well as the inequality

$$\mathbb{E}\left\|\left(1 + \frac{\beta}{1-\beta}\right)x^{(k)} - \beta\frac{x^{(k-1)}}{1-\beta}\right\|^2 = \mathbb{E}\left\|x^{(k)} + p^{(k)}\right\|^2$$
$$\leq 2\mathbb{E}\left\|x^{(k)} + p^{(k)} - x_\alpha^*\right\|^2 + 2\|x_\alpha^*\|^2 \leq 2c_3 + 2\|x_\alpha^*\|^2\,,$$

where we applied the Cauchy-Schwarz inequality. If we apply the Cauchy-Schwarz inequality again, we obtain

$$\begin{aligned}
\mathbb{E}\left\|x^{(k)}\right\|^2 &\leq \mathbb{E}\left\|\left(1 + \frac{\beta}{1-\beta}\right)x^{(k)}\right\|^2 \\
&\leq 2\mathbb{E}\left\|\left(1 + \frac{\beta}{1-\beta}\right)x^{(k)} - \beta\frac{x^{(k-1)}}{1-\beta}\right\|^2 + 2\mathbb{E}\left\|\beta\frac{x^{(k-1)}}{1-\beta}\right\|^2 \\
&\leq 2\mathbb{E}\left\|\left(1 + \frac{\beta}{1-\beta}\right)x^{(k)} - \beta\frac{x^{(k-1)}}{1-\beta}\right\|^2 + \frac{2\beta^2}{(1-\beta)^2}\left(2\mathbb{E}\left\|x^{(k-1)} - x_\alpha^*\right\|^2 + 2\|x_\alpha^*\|^2\right) \\
&\leq 4c_3 + 4\|x_\alpha^*\|^2 + \frac{2\beta^2}{(1-\beta)^2}\left(\frac{2c_3}{\mu b} + 2\|x_\alpha^*\|^2\right)\,,
\end{aligned}$$

where we used (102).

Within our assumptions on the stepsize and momentum β, $b = \mathcal{O}(\alpha^2)$ and $\|x_\alpha^*\| = \Theta(1)$. Furthermore, we have $\beta = \mathcal{O}(\alpha^2)$ as well as $b = \Theta(\alpha^2) = \Theta(\eta^4)$. Therefore, we conclude that $\mathbb{E}\left\|x^{(k)}\right\|^2 = \mathcal{O}(1)$, which is equivalent to (46). This implies that

$$\mathbb{E}\left\|\nabla F\left(x^{(k)}\right)\right\|^2 \leq \tilde{c}_4 := L\left(\|\nabla F(x^*)\|^2 + \sup_{k\geq 0}\mathbb{E}\left\|x^{(k)} - x^*\right\|^2\right),$$

where we used L-smoothness of F. Consequently, we find from the update equation (34) that

$$\mathbb{E}\left\|v^{(k+1)}\right\|^2 = \mathbb{E}\left\|\beta v^{(k)} - \eta\nabla F\left(x^{(k)}\right)\right\|^2 + \eta^2\sigma^2 N + 2\gamma\eta N d \tag{103}$$
$$\leq 2\beta^2\mathbb{E}\left\|v^{(k)}\right\|^2 + 2\eta^2\mathbb{E}\left\|\nabla F\left(x^{(k)}\right)\right\|^2 + \eta^2\sigma^2 N + 2\gamma\eta N d, \tag{104}$$

which implies that for any k,

$$\mathbb{E}\left\|v^{(k+1)}\right\|^2 \leq c_5 := \mathbb{E}\left\|v^{(0)}\right\|^2 + \frac{2\eta^2\tilde{c}_4 + \eta^2\sigma^2 N + 2\gamma\eta N d}{1-2\beta^2} = \mathcal{O}(1), \tag{105}$$

where we used the fact that $2\beta^2 < 1$ by our assumptions. This completes the proof of Lemma 14.

□

The next three technical lemmas are used in the proof of Lemma 14.

Lemma 20 *In the setting of the proof of Lemma 14; if the parameters α and β satisfy the inequalities* (73) *and* (74)*, then $k_{\alpha,\beta} < 0$ where $k_{\alpha,\beta}$ is defined by* (80)*.*

Proof of Lemma 20 The proof of Lemma 20 will be provided in Appendix C. □

Lemma 21 *In the setting of Lemma 14, let α and β satisfy the conditions* (73)*,* (74) *and* (75) *where α is defined by* (70)*. Then, we have*

$$q_1 := \frac{a}{b} \in (0,1)\,, \tag{106}$$

where a and b are defined in (85) *and* (86)*.*

Proof of Lemma 21 The proof of Lemma 21 will be provided in Appendix C. □

Lemma 22 *In the setting of the proof of Lemma 14, we have $sM - Q \succeq 0$ if*

$$s \geq q_2 := \max\left(0, 1 - \frac{\alpha\mu L_\alpha}{(1-\beta)(L_\alpha+\mu) + 2L_\alpha\beta}\right). \tag{107}$$

Proof of Lemma 22 The proof of Lemma 22 will be provided in Appendix C. □

B.2 Proofs of Lemma 15– 18

The proofs of Lemmas 15 and 16 are inspired by the proofs of Lemmas 7 and 8 respectively, and the proofs of Lemmas 17 and 18 are inspired by the proof of Lemma 9. We will present the proofs of Lemmas 15–18 in Appendix C. □

Appendix C. Additional Technical Proofs

C.1 Proof of Lemma 11

Note that $\mathbb{E}_{X\sim\pi}\|X - x_*\|^2 = \mathbb{E}\|X_\infty - x_*\|^2$, where X_∞ is the unique stationary distribution of the overdamped Langevin diffusion:

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

where W_t is a standard $d-$dimensional Brownian motion. By Itô's formula, we have

$$\begin{aligned} e^{\mu t}\|X_t - x_*\|^2 = \|X_0 - x_*\|^2 &+ 2\sqrt{2N^{-1}}\int_0^t e^{\mu s}\langle X_s - x_*, dW_s\rangle \\ &- 2\int_0^t e^{\mu s}\left\langle X_s - x_*, \frac{1}{N}\nabla f(X_s)\right\rangle ds \\ &+ 2N^{-1}d\int_0^t e^{\mu s}ds + \mu\int_0^t e^{\mu s}\|X_s - x_*\|^2 ds \\ \leq \|X_0 - x_*\|^2 &+ 2\sqrt{2N^{-1}}\int_0^t e^{\mu s}\langle X_s - x_*, dW_s\rangle + 2dN^{-1}\int_0^t e^{\mu s}ds, \end{aligned}$$

where we used μ-strongly convex property of $x \mapsto \frac{1}{N}f(x)$. This implies that

$$\mathbb{E}\|X_t - x_*\|^2 \leq e^{-\mu t}\|X_0 - x_*\|^2 + \frac{2dN^{-1}}{\mu},$$

and therefore $\mathbb{E}\|X_\infty - x_*\|^2 \leq \frac{2dN^{-1}}{\mu}$. The proof is complete. □

C.2 Proof of Lemma 15

In this proof, we aim to provide uniform L_2 bounds between the iterates $x_i^{(k)}$ and their means $\bar{x}^{(k)}$. First, by the definitions of $x^{(k)}$, we get

$$x^{(k+1)} = (W \otimes I_d)x^{(k)} + \eta v^{(k+1)}.$$

It follows that

$$x^{(k)} = \left(W^k \otimes I_d\right)x^{(0)} + \eta\sum_{s=0}^{k-1}\left(W^{k-1-s} \otimes I_d\right)v^{(s+1)}.$$

Let us define $\bar{\mathbf{x}}^{(k)} := [\bar{x}^{(k)}, \cdots, \bar{x}^{(k)}] \in \mathbb{R}^{Nd}$. Notice that

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

Therefore, we get

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

and by the Cauchy-Schwarz inequality

$$\begin{aligned}
&\left\|x^{(k)}-\frac{1}{N}\left(\left(1_N1_N^T\right)\otimes I_d\right)x^{(k)}\right\|^2\\
&\le 2\left\|\left(W^k\otimes I_d\right)x^{(0)}-\frac{1}{N}\left(\left(1_N1_N^TW^k\right)\otimes I_d\right)x^{(0)}\right\|^2\\
&\quad+2\left\|-\eta\sum_{s=0}^{k-1}\left(W^{k-1-s}\otimes I_d\right)v^{(s+1)}+\eta\sum_{s=0}^{k-1}\frac{1}{N}\left(\left(1_N1_N^TW^{k-1-s}\right)\otimes I_d\right)v^{(s+1)}\right\|^2\\
&=2\left\|\left(W^k\otimes I_d\right)x^{(0)}-\frac{1}{N}\left(\left(1_N1_N^T\right)\otimes I_d\right)x^{(0)}\right\|^2\\
&\quad+2\left\|-\eta\sum_{s=0}^{k-1}\left(W^{k-1-s}\otimes I_d\right)v^{(s+1)}+\eta\sum_{s=0}^{k-1}\frac{1}{N}\left(\left(1_N1_N^T\right)\otimes I_d\right)v^{(s+1)}\right\|^2\\
&=2\left\|\left(\left(W^k-\frac{1}{N}1_N1_N^T\right)\otimes I_d\right)x^{(0)}\right\|^2+2\eta^2\left\|\sum_{s=0}^{k-1}\left(\left(W^{k-1-s}-\frac{1}{N}1_N1_N^T\right)\otimes I_d\right)v^{(s+1)}\right\|^2.
\end{aligned}$$

Note that

$$\begin{aligned}
&\eta^2\left\|\sum_{s=0}^{k-1}\left(\left(W^{k-1-s}-\frac{1}{N}1_N1_N^T\right)\otimes I_d\right)v^{(s+1)}\right\|^2\\
&\le\eta^2\left(\sum_{s=0}^{k-1}\left\|\left(W^{k-1-s}-\frac{1}{N}1_N1_N^T\right)\otimes I_d\right\|\cdot\left\|v^{(s+1)}\right\|\right)^2\\
&\le\eta^2\left(\sum_{s=0}^{k-1}\left\|W^{k-1-s}-\frac{1}{N}1_N1_N^T\right\|\cdot\left\|v^{(s+1)}\right\|\right)^2\\
&=\eta^2\left(\sum_{s=0}^{k-1}\bar\gamma^{k-1-s}\cdot\left\|v^{(s+1)}\right\|\right)^2\\
&=\eta^2\left(\sum_{s=0}^{k-1}\bar\gamma^{k-1-s}\right)^2\left(\frac{\sum_{s=0}^{k-1}\bar\gamma^{k-1-s}\cdot\left\|v^{(s+1)}\right\|}{\sum_{s=0}^{k-1}\bar\gamma^{k-1-s}}\right)^2\\
&\le\eta^2\left(\sum_{s=0}^{k-1}\bar\gamma^{k-1-s}\right)^2\sum_{s=0}^{k-1}\frac{\bar\gamma^{k-1-s}}{\sum_{s=0}^{k-1}\bar\gamma^{k-1-s}}\left\|v^{(s+1)}\right\|^2,
\end{aligned}$$

where we used Jensen's inequality in the last step above.

Recall from Lemma 14 that for every s, $\mathbb{E}\left[\left\|v^{(s+1)}\right\|^2\right]\le c_5$. Therefore, we have

$$\begin{aligned}
&\eta^2\mathbb{E}\left[\left\|\sum_{s=0}^{k-1}\left(\left(W^{k-1-s}-\frac{1}{N}1_N1_N^T\right)\otimes I_d\right)v^{(s+1)}\right\|^2\right]\\
&\le c_5\eta^2\left(\sum_{s=0}^{k-1}\bar\gamma^{k-1-s}\right)^2\sum_{s=0}^{k-1}\frac{\bar\gamma^{k-1-s}}{\sum_{s=0}^{k-1}\bar\gamma^{k-1-s}}\le c_5\eta^2\frac{1}{(1-\bar\gamma)^2}.
\end{aligned}$$

Similarly, we have

$$\left\|\left(\left(W^k - \frac{1}{N}1_N 1_N^T\right)\otimes I_d\right)x^{(0)}\right\|^2 \leq \bar{\gamma}^{2k}\left\|x^{(0)}\right\|^2.$$

The proof is complete. □

C.3 Proof of Lemma 16

By Lemma 15, we can compute that

$$\begin{aligned}
\mathbb{E}\left\|\mathcal{E}_{k+1}\right\|^2 &= \mathbb{E}\left\|\frac{1}{N}\sum_{i=1}^N\left(\nabla f_i\left(x_i^{(k)}\right) - \nabla f_i\left(\bar{x}^{(k)}\right)\right)\right\|^2 \\
&\leq \frac{1}{N^2}\sum_{i=1}^N N\mathbb{E}\left\|\nabla f_i\left(x_i^{(k)}\right) - \nabla f_i\left(\bar{x}^{(k)}\right)\right\|^2 \\
&\leq \frac{1}{N}L^2\sum_{i=1}^N\mathbb{E}\left\|x_i^{(k)} - \bar{x}^{(k)}\right\|^2 \\
&\leq \frac{2L^2\bar{\gamma}^{2k}}{N}\mathbb{E}\left\|x^{(0)}\right\|^2 + \frac{2L^2c_5\eta^2}{N(1-\bar{\gamma})^2}.
\end{aligned}$$

The proof is complete. □

C.4 Proof of Lemma 17

In this proof, we aim to show that the average iterates $\bar{x}^{(k)}$ are close to the iterates $\tilde{x}_k$ which is defined in (44). First, we can compute that

$$\bar{x}^{(k+1)} - \tilde{x}_{k+1} = \bar{x}^{(k)} - \tilde{x}_k - \frac{\eta^2}{N}\left[\nabla f\left(\bar{x}^{(k)}\right) - \nabla f(\tilde{x}_k)\right] + \beta\left(\bar{x}^{(k)} - \bar{x}^{(k-1)}\right) + \eta^2\mathcal{E}_{k+1} - \eta^2\bar{\xi}^{(k+1)},$$

where

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

We also observe that under our assumptions $\eta \in (0, \sqrt{2/L})$, $\eta^2\mu(1-\frac{\eta^2 L}{2}) \leq 1$. Then, it follows from the proof of Lemma 9 that we have

$$\begin{aligned}
\mathbb{E}\left\|\bar{x}^{(k+1)} - \tilde{x}_{k+1}\right\|^2 &\leq \left(1 - \eta^2\mu\left(1 - \frac{\eta^2 L}{2}\right)\right)\mathbb{E}\left\|\bar{x}^{(k)} - \tilde{x}_k\right\|^2 + \eta^4\frac{\sigma^2}{N} \\
&\quad + \eta^2\left(\eta^2 + \frac{(1+\eta^2 L)^2}{\mu(1-\frac{\eta^2 L}{2})}\right)\mathbb{E}\left\|\frac{\beta}{\eta^2}\left(\bar{x}^{(k)} - \bar{x}^{(k-1)}\right) + \mathcal{E}_{k+1}\right\|^2 \\
&\leq \left(1 - \eta^2\mu\left(1 - \frac{\eta^2 L}{2}\right)\right)\mathbb{E}\left\|\bar{x}^{(k)} - x_k\right\|^2 + \eta^4\frac{\sigma^2}{N} \\
&\quad + 2\eta^2\left(\eta^2 + \frac{(1+\eta^2 L)^2}{\mu(1-\frac{\eta^2 L}{2})}\right)\left(\mathbb{E}\left\|\frac{\beta}{\eta}\bar{v}^{(k)}\right\|^2 + \mathbb{E}\left\|\mathcal{E}_{k+1}\right\|^2\right),
\end{aligned} \tag{108}$$

where we used $\bar{x}^{(k)} - \bar{x}^{(k-1)} = \eta \bar{v}^{(k)}$ and (18). We recall from Lemma 16 that

$$\mathbb{E}\left\|\mathcal{E}_{k+1}\right\|^2 \leq \frac{2L^2\bar{\gamma}^{2k}}{N}\mathbb{E}\left\|x^{(0)}\right\|^2 + \frac{2L^2 c_5 \eta^2}{N(1-\bar{\gamma})^2}, \tag{109}$$

and by Lemma 14, we get

$$\mathbb{E}\left\|\bar{v}^{(k)}\right\|^2 \leq \frac{1}{N}\sum_{i=1}^{N}\mathbb{E}\left\|v_i^{(k)}\right\|^2 = \frac{1}{N}\mathbb{E}\left\|v^{(k)}\right\|^2 \leq \frac{c_5}{N}. \tag{110}$$

By applying (109)-(110) to (108), we get

$$\begin{aligned}
&\mathbb{E}\left\|\bar{x}^{(k+1)} - \tilde{x}_{k+1}\right\|^2 \\
&\leq \left(1 - \eta^2\mu\left(1 - \frac{\eta^2 L}{2}\right)\right)\mathbb{E}\left\|\bar{x}^{(k)} - \tilde{x}_k\right\|^2 + \eta^4\frac{\sigma^2}{N} \\
&\quad + 2\eta^2\left(\eta^2 + \frac{(1+\eta^2 L)^2}{\mu(1-\frac{\eta^2 L}{2})}\right)\left(\frac{\beta^2 c_5}{\eta^2 N} + \frac{2L^2\bar{\gamma}^{2k}}{N}\mathbb{E}\left\|x^{(0)}\right\|^2 + \frac{2L^2 c_5\eta^2}{N(1-\bar{\gamma})^2}\right),
\end{aligned}$$

for every k. Note that $\mathbb{E}\left\|\bar{x}^{(0)} - \tilde{x}_0\right\|^2 = 0$. By our assumption on stepsize η, we have $1 - \eta^2\mu\left(1 - \frac{\eta^2 L}{2}\right) \in [0, 1)$. By following the same argument as in the proof of Lemma 9, we conclude that for every k,

$$\begin{aligned}
\mathbb{E}\left\|\bar{x}^{(k)} - \tilde{x}_k\right\|^2 \leq{}& \frac{2\eta^2\left(\eta^2 + \frac{(1+\eta^2 L)^2}{\mu(1-\frac{\eta^2 L}{2})}\right)\left(\frac{\beta^2 c_5}{\eta^2 N} + \frac{2L^2 c_5\eta^2}{N(1-\bar{\gamma})^2}\right) + \eta^4\frac{\sigma^2}{N}}{1 - \left(1 - \eta^2\mu\left(1 - \frac{\eta^2 L}{2}\right)\right)} \\
&+ \frac{\bar{\gamma}^{2k} - \left(1 - \eta^2\mu\left(1 - \frac{\eta^2 L}{2}\right)\right)^k}{\bar{\gamma}^2 - 1 + \eta^2\mu\left(1 - \frac{\eta^2 L}{2}\right)}\frac{4L^2\bar{\gamma}^2}{N}\mathbb{E}\left\|x^{(0)}\right\|^2,
\end{aligned}$$

which completes the proof. □

C.5 Proof of Lemma 18

In this proof, we aim to show that the iterates $\tilde{x}_k$, which is defined in (44), is close to the iterates x_k in (43) obtained from an Euler-Maruyama discretization of an overdamped Langevin SDE. First, we can compute that

$$\tilde{x}_{k+1} - x_{k+1} = \tilde{x}_k - x_k - \frac{\eta^2}{N}\left[\nabla f(\tilde{x}_k) - \nabla f(x_k)\right] + \left(\sqrt{2(1-\beta)} - \sqrt{2}\right)\eta\bar{w}^{(k+1)}.$$

It follows from the arguments in the proof of Lemma 9 that we have

$$\begin{aligned}
\left\|\bar{x}^{(k+1)} - \tilde{x}_{k+1}\right\|^2 \leq{}& \left(1 - \eta^2\mu\left(1 - \frac{\eta^2 L}{2}\right)\right)\left\|\bar{x}^{(k)} - \tilde{x}_k\right\|^2 \\
&+ \eta^2\left(\eta^2 + \frac{(1+\eta^2 L)^2}{\mu(1-\frac{\eta^2 L}{2})}\right)\left\|\frac{1}{\eta^2}\left(\sqrt{2(1-\beta)} - \sqrt{2}\right)\eta\bar{w}^{(k+1)}\right\|^2.
\end{aligned}$$

By taking the expectations, we get

$$\mathbb{E}\left\|\bar{x}^{(k+1)}-\tilde{x}_{k+1}\right\|^2 \\ \leq \left(1-\eta^2\mu\left(1-\frac{\eta^2 L}{2}\right)\right)\mathbb{E}\left\|\bar{x}^{(k)}-\tilde{x}_k\right\|^2+2\left(\eta^2+\frac{(1+\eta^2 L)^2}{\mu(1-\frac{\eta^2 L}{2})}\right)\left(\sqrt{(1-\beta)}-1\right)^2\frac{d}{N},$$

for every k. The rest of the proof follows similarly as in the proof of Lemma 17. □

C.6 Proof of Lemma 19

Note that x_η^* by its definition coincides with the fixed point $\hat{x}^\infty$ of the decentralized gradient descent without noise:

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

i.e.

$$\hat{x}^\infty=\mathcal{W}\hat{x}^\infty-\eta\nabla F\left(\hat{x}^\infty\right),$$

and $x_\eta^*=\hat{x}^\infty$. Since x_η^* and x^* do not depend on $\hat{x}^{(0)}$, to get a bound on $\|x_\eta^*-x^*\|$, we can assume that $\hat{x}^{(0)}=0$, and apply Corollary 9 in Yuan et al. (2016) which is re-stated in Fallah et al. (2019):

$$\|\hat{x}_i^\infty-x^*\|\leq C_1\frac{\eta}{1-\bar{\gamma}},\quad \text{where}\quad \bar{\gamma}:=\max\left\{\left|\lambda_2^W\right|,\left|\lambda_N^W\right|\right\},$$

where $x^*=(x_*^T,x_*^T,\ldots,x_*^T)^T$, where x^* is the minimizer of $f(x)$, which yields that

$$\left\|x_\eta^*-x^*\right\|\leq C_1\frac{\eta\sqrt{N}}{1-\bar{\gamma}},\quad \text{where}\quad \bar{\gamma}:=\max\left\{\left|\lambda_2^W\right|,\left|\lambda_N^W\right|\right\}.$$

The proof is complete. □

C.7 Proof of Lemma 20

By the definition of L_α given by (72), we have

$$k_{\alpha,\beta}:=\frac{\alpha}{(1-\beta)}\left(\frac{\alpha}{1-\beta}-\frac{2\alpha}{1-\lambda_N^W+(L+\mu)\alpha}+\frac{2c_1\alpha}{1-\beta}\right). \tag{111}$$

Due to (73), we have $(L+\mu)\alpha\leq\frac{1+\lambda_N^W}{2}$, therefore

$$k_{\alpha,\beta}\leq\frac{\alpha}{(1-\beta)}\left(\frac{\alpha}{1-\beta}-\frac{2\alpha}{1-\lambda_N^W+(1+\lambda_N^W)/2}+\frac{2c_1\alpha}{1-\beta}\right). \tag{112}$$

Furthermore,

$$c_1=\frac{1}{2}\frac{\alpha\mu}{(1-\beta)(1+\mu/L_\alpha)+2\beta}<\frac{1}{2}\frac{\alpha\mu}{1+\beta}<\frac{1}{2}\alpha\mu, \tag{113}$$

where we used the fact that $\mu/L_\alpha > 0$. Therefore, by replacing c_1 in (112) with its upper bound (113), we obtain

$$\begin{aligned} k_{\alpha,\beta} &\leq \frac{\alpha^2}{(1-\beta)}\left(\frac{1}{1-\beta} - \frac{2}{1-\lambda_N^W + (1+\lambda_N^W)/2} + \frac{\alpha\mu}{1-\beta}\right) \\ &= \frac{\alpha^2}{(1-\beta)}\left(\frac{1}{1-\beta} - \frac{4}{3-\lambda_N^W} + \frac{\alpha\mu}{1-\beta}\right). \end{aligned}$$

Since $\alpha > 0$ and $\beta < 1/2$ by our assumptions, $\frac{\alpha^2}{1-\beta} > 0$ and $k_{\alpha,\beta} < 0$ if and only if

$$\frac{1}{1-\beta} - \frac{4}{3-\lambda_N^W} + \frac{\alpha\mu}{1-\beta} < 0\,,$$

which is equivalent to

$$\beta < \frac{1+\lambda_N^W + \alpha\mu\lambda_N^W - 3\alpha\mu}{4}\,. \tag{114}$$

By our assumption (74) on β, we have

$$\beta \leq \frac{1+\lambda_N^W - 4\alpha\mu}{4}\,,$$

and noticing that $\lambda_N^W > -1$ and $\alpha\mu\lambda_N^W > -\alpha\mu$, we conclude that the inequality (114) holds. Hence, we obtain $k_{\alpha,\beta} < 0$ and the proof is complete. □

C.8 Proof of Lemma 21

Using the definitions of a and b from (85) and (86), we have

$$q_1 = \frac{a}{b} = \frac{2\alpha\beta + 4\beta^2(1-\lambda_N^W)^2(1+1/c_1)/\mu}{2\alpha\left(\beta - \mu\alpha + 2\mu\frac{(1-\beta)}{L_\alpha+\mu} - 2c_1\alpha\mu\right)}\,,$$

where c_1 is given by (76). Therefore, the condition $q_1 \in (0,1)$ is equivalent to

$$b(1-\beta)^2 = 2\alpha\left(\beta - \mu\alpha + 2\mu\frac{(1-\beta)}{L_\alpha+\mu} - 2c_1\alpha\mu\right) > 0\,, \tag{115}$$

where b is defined by (86) and

$$2\alpha\beta + 4\beta^2\left(1-\lambda_N^W\right)^2(1+1/c_1)/\mu < 2\alpha\left(\beta - \mu\alpha + 2\mu\frac{(1-\beta)}{L_\alpha+\mu} - 2c_1\alpha\mu\right). \tag{116}$$

It suffices to show that under our assumptions on α and β, these two conditions are satisfied. The first condition (115) is satisfied because $b = \frac{2\alpha\beta}{(1-\beta)^2} - 2k_{\alpha,\beta}\mu > \frac{2\alpha\beta}{(1-\beta)^2} > 0$ by Lemma 20. We next prove that the second condition (116) holds. We re-organize (116) as

$$4\beta^2\left(1-\lambda_N^W\right)^2(c_1+1) < 2c_1\mu\alpha\left(-\mu\alpha + 2\mu\frac{(1-\beta)}{L_\alpha+\mu} - 2c_1\alpha\mu\right). \tag{117}$$

We note that

$$c_1 = \frac{1}{2}\frac{\alpha\mu}{(1-\beta)(1+\mu/L_\alpha)+2\beta} \leq \frac{\alpha\mu}{2(1+\beta)} < 1\,,$$

where in the first inequality we used the fact that $\mu/L_\alpha > 0$ whereas in the second inequality we used the assumptions (73) and (74). Therefore, $c_1+1 < 2$ and $2c_1\alpha\mu \leq 2\alpha\mu\frac{\alpha\mu}{2(1+\beta)}$. Hence it suffices to have

$$\begin{aligned} 8\beta^2\left(1-\lambda_N^W\right)^2 &\leq 2c_1\mu\alpha\left(-\mu\alpha + 2\mu\frac{(1-\beta)}{L_\alpha+\mu} - 2\alpha\mu\frac{\alpha\mu}{2(1+\beta)}\right) \\ &= 2c_1\mu\alpha\left(-\mu\alpha + 2\mu\frac{(1-\beta)\alpha}{1-\lambda_N^W+(L+\mu)\alpha} - 2\alpha\mu\frac{\alpha\mu}{2(1+\beta)}\right), \end{aligned}$$

where we used the definition of L_α given in (72). By assumption (73), we have $\alpha \leq (1+\lambda_N^W)/(2(L+\mu))$; therefore it suffices to have

$$8\beta^2\left(1-\lambda_N^W\right)^2 \leq 2c_1\mu\alpha\left(-\mu\alpha + 2\mu\frac{(1-\beta)\alpha}{1-\lambda_N^W+\frac{1+\lambda_N^W}{2}} - 2\alpha\mu\frac{\alpha\mu}{2(1+\beta)}\right) \tag{118}$$

$$= 2c_1\mu\alpha\left(-\mu\alpha + 4\mu\frac{(1-\beta)\alpha}{3-\lambda_N^W} - 2\alpha\mu\frac{\alpha\mu}{2(1+\beta)}\right) \tag{119}$$

$$= 2c_1\mu^2\alpha^2\left(-1 + 4\frac{(1-\beta)}{3-\lambda_N^W} - \frac{\alpha\mu}{(1+\beta)}\right). \tag{120}$$

By differentiating the right-hand side of (120) with respect to β, it is easy to see that the right-hand side is a decreasing function of β under our assumptions. Therefore, by plugging in the largest allowed value $\frac{1+\lambda_N-4\alpha\mu}{4}$ for β on the right-hand side of this inequality, we can relax condition (120) to

$$8\beta^2\left(1-\lambda_N^W\right)^2 \leq 2c_1\mu^2\alpha^2\left(-1 + 4\frac{1-\frac{1+\lambda_N^W-4\alpha\mu}{4}}{3-\lambda_N^W} - \alpha\mu\right) = 2c_1\mu^2\alpha^2\left(\frac{\alpha\mu(1+\lambda_N^W)}{3-\lambda_N^W}\right).$$

Since $\lambda_N^W \in (-1,1)$, it suffices to have

$$8\beta^2\left(1-\lambda_N^W\right)^2 \leq 2c_1\mu^2\alpha^2\left(\frac{\alpha\mu(1+\lambda_N^W)}{4}\right),$$

which holds if and only if

$$\beta^2 \leq c_1\mu^3\alpha^3\left(\frac{(1+\lambda_N^W)}{16(1-\lambda_N^W)^2}\right).$$

Since $\lambda_N^W \in (-1,1)$, it suffices to have

$$\beta^2 \leq c_1\mu^3\alpha^3\left(\frac{1+\lambda_N^W}{64}\right),$$

which is exactly the condition (75) we assumed in the statement of the lemma. We conclude that the inequality (116) is also satisfied. Finally, we infer from (115) and (116) that $q_1 \in (0,1)$ completing the proof.

□

C.9 Proof of Lemma 22

Consider the matrix pencil $S_s = sM - Q$ with $s \geq 0$. We have

$$S_s = \begin{bmatrix} \left(s - 1 - 2c_1 + \frac{2\alpha\mu L_\alpha}{(1-\beta)(L_\alpha+\mu)}\right) I_d & \frac{\beta}{(1-\beta)}(s-1-2c_1) I_d \\ \frac{\beta}{(1-\beta)}(s-1-2c_1) I_d & \left(\frac{(s-1-2c_1)\beta^2+\alpha\beta\mu}{(1-\beta)^2}\right) I_d \end{bmatrix} = A_s \otimes I_d \,,$$

where $\otimes$ denotes the Kronecker product of matrices and A_s is the 2×2 matrix

$$A_s = \begin{bmatrix} s - 1 - 2c_1 + \frac{2\alpha\mu L_\alpha}{(1-\beta)(L_\alpha+\mu)} & \frac{\beta}{(1-\beta)}(s-1-2c_1) \\ \frac{\beta}{(1-\beta)}(s-1-2c_1) & \frac{(s-1-2c_1)\beta^2+\alpha\beta\mu}{(1-\beta)^2} \end{bmatrix} .$$

By the properties of the Kronecker product, the symmetric matrix S_s has the same eigenvalues with the 2×2 matrix A_s and S_s is positive semi-definite if and only if as A_s is positive semi-definite. Therefore, S_s is positive definite if and only if the principal minors of A_s are non-negative, i.e.

$$s - 1 - 2c_1 + \frac{2\alpha\mu L_\alpha}{(1-\beta)(L_\alpha+\mu)} \geq 0 \,,$$

and

$$\left(s - 1 - 2c_1 + \frac{2\alpha\mu L_\alpha}{(1-\beta)(L_\alpha+\mu)}\right)\left(\frac{(s-1-2c_1)\beta^2+\alpha\mu\beta}{(1-\beta)^2}\right) \geq \left(\frac{\beta}{(1-\beta)}(s-1-2c_1)\right)^2 .$$

After some computations we observe that the last inequality is equivalent to

$$s - 1 - 2c_1 + \frac{2\alpha\mu L_\alpha}{(1-\beta)(L_\alpha+\mu) + 2L_\alpha\beta} \geq 0 \,.$$

We conclude that S_s is positive semi-definite if and only if

$$s \geq 1 + 2c_1 - \frac{2\alpha\mu L_\alpha}{(1-\beta)(L_\alpha+\mu) + 2L_\alpha\beta} = 1 - \frac{\alpha\mu L_\alpha}{(1-\beta)(L_\alpha+\mu) + 2L_\alpha\beta} \,,$$

where we used the definition (76) of c_1 in the last equality. This completes the proof. □

Appendix D. Discussions on Gradient Noise Assumptions

In our analysis, we assumed that the variance of the gradient noise is bounded (Assumption 1). This is a reasonable assumption since it can be shown that if the stepsize $\eta > 0$ is small enough the variance of the gradients will stay bounded and satisfy our assumptions on the gradient noise (Assumption 1) with an analysis similar to Aybat et al. (2019, Section K). We can illustrate this point in detail as follows.

Consider a more general gradient noise setting than Assumption 1:

$$\mathbb{E}\left[\tilde{\nabla} f_i(x) - \nabla f_i(x) \middle| x\right] = 0, \qquad \mathbb{E}\left[\left\|\tilde{\nabla} f_i(x) - \nabla f_i(x)\right\|^2 \middle| x\right] \leq C\left(1 + \|x\|^2\right), \tag{121}$$

(see e.g. Jain et al. (2018)) where C is a positive constant. The assumption (121) is satisfied for a wide class of f_i functions when gradients are estimated over mini-batches. Consider the

linear regression example in the empirical risk minimization setting, where the stochastic gradients $\tilde{\nabla} f_i(x)$ are estimated from mini-batches of size b at a point x, i.e.

$$\tilde{\nabla} f_i(x) = \frac{2n_i}{b} \sum_{k=1}^{b} (y_{j_k}^i - x^T X_{j_k}^i) + \frac{1}{\lambda N} x,$$

where $j_1, j_2, \ldots, j_b$ are selected uniformly random with replacement over the index set $\{1, 2. \ldots, n_i\}$ of the data points where n_i are finite and fixed. In this setting, it is well-known that the gradient error satisfies (121). The L_2-regularized logistic regression case will be similar.

In the following, we will show that for DE-SGLD, when the stepsize η is sufficiently small, the assumption (121) implies that:

$$\mathbb{E}\left[\tilde{\nabla} f_i(x) - \nabla f_i(x) \middle| x\right] = 0, \qquad \mathbb{E}\left\|\tilde{\nabla} f_i(x) - \nabla f_i(x)\right\|^2 \leq \sigma^2, \tag{122}$$

for some $\sigma > 0$, which is the precisely the assumption we had for gradient noise (Assumption 1) and hence our main result (Theorem 2) holds under the assumption (121). This is primarily because the second moments of the iterates are uniformly bounded and taking expectation with respect to x in (121) would result in a condition like (122). To see this in more detail, we recall that in the proof of Lemma 6, by assuming $\mathbb{E}\left\|\tilde{\nabla} f_i(x) - \nabla f_i(x)\right\|^2 \leq \sigma^2$, as in Assumption 1, we had

$$\mathbb{E}\left[\left\|x^{(k+1)} - x_\eta^*\right\|^2\right] \leq \left(1 - \mu\eta\left(1 + \lambda_N^W - \eta L\right)\right) \mathbb{E}\left[\left\|x^{(k)} - x_\eta^*\right\|^2\right] + \eta^2\sigma^2 N + 2\eta d N, \tag{123}$$

provided that the stepsize η is sufficiently small, where $x^{(k)} = \left[(x_1^{(k)})^T, \ldots, (x_N^{(K)})^T\right]^T$, and x_η^* is the minimizer of $F_{\mathcal{W},\eta}(x) = \frac{1}{2\eta} x^T (I - \mathcal{W}) x + F(x)$, where $F(x) = \sum_{i=1}^N f_i(x_i)$. Now, if we assume (121), instead of (122), we will obtain

$$\begin{aligned}
\mathbb{E}\left[\left\|x^{(k+1)} - x_\eta^*\right\|^2\right] &\leq \left(1 - \mu\eta\left(1 + \lambda_N^W - \eta L\right)\right) \mathbb{E}\left[\left\|x^{(k)} - x_\eta^*\right\|^2\right] \\
&\qquad + \eta^2 C N + \eta^2 C \mathbb{E}\left[\left\|x^{(k)}\right\|^2\right] + 2\eta d N \\
&\leq \left(1 - \mu\eta\left(1 + \lambda_N^W - \eta L\right)\right) \mathbb{E}\left[\left\|x^{(k)} - x_\eta^*\right\|^2\right] \\
&\qquad + \eta^2 C N + 2\eta^2 C \mathbb{E}\left[\left\|x^{(k)} - x_\eta^*\right\|^2\right] + 2\eta^2 C \left\|x_\eta^*\right\|^2 + 2\eta d N.
\end{aligned}$$

By Lemma 19, for sufficiently small stepsize $\eta > 0$, $\|x_\eta^* - x^*\| \leq C_1 \frac{\eta\sqrt{N}}{1-\bar{\gamma}}$, where $C_1, \bar{\gamma}$ are constants defined in (27) and (19), and we recall from (10) that $x^* = \left[x_*^T, \ldots, x_*^T\right]^T$ where x_* is the minimizer of $f(x) = \sum_{i=1}^N f_i(x)$ which is unique by strong convexity. Therefore,

we obtain

$$\begin{aligned}
&\mathbb{E}\left[\left\|x^{(k+1)}-x_\eta^*\right\|^2\right]\\
&\le \left(1-\mu\eta\left(1+\lambda_N^W-\eta L\right)\right)\mathbb{E}\left[\left\|x^{(k)}-x_\eta^*\right\|^2\right]\\
&\quad+\eta^2 CN+2\eta^2 C\mathbb{E}\left[\left\|x^{(k)}-x_\eta^*\right\|^2\right]+4\eta^2 C\left\|x^*\right\|^2+4\eta^4 CC_1^2\frac{N}{(1-\bar\gamma)^2}+2\eta dN\\
&\le \left(1-\frac{1}{2}\mu\eta\left(1+\lambda_N^W-\eta L\right)\right)\mathbb{E}\left[\left\|x^{(k)}-x_\eta^*\right\|^2\right]\\
&\qquad\qquad+\eta^2 CN+4\eta^2 C\left\|x^*\right\|^2+4\eta^4 CC_1^2\frac{N}{(1-\bar\gamma)^2}+2\eta dN, \qquad (124)
\end{aligned}$$

for sufficiently small η. This implies that for sufficiently small η, such that $\frac{1}{2}\mu\eta(1+\lambda_N^W-\eta L)\in(0,1)$, we have the following uniform L_2 bound:

$$\mathbb{E}\left[\left\|x^{(k)}-x_\eta^*\right\|^2\right]\le \tilde C_1, \qquad \text{for any } k\in\mathbb{N}, \qquad (125)$$

for some constant $\tilde C_1>0$. Finally, by taking the expectation w.r.t. x in (121) and applying the tower property, we conclude that (122) holds for some $\sigma>0$. Hence, the assumption (121) implies the assumption (122) which is used in Assumption 1. This argument shows that our assumption on the finiteness of gradient noise variance is satisfied when data is sampled with mini-batches such that (121) holds.

Appendix E. Discussions on the Lipschitz Gradient Assumption

In our analysis, we consider sampling from the target distribution with density $\pi(x)\propto e^{-f(x)}=e^{-\sum_{i=1}^N f_i(x)}$, where f_i is the loss function of the agent i for $i=1,2,\dots,N$. We assume that the gradients $\nabla f_i(x)$ are (uniformly) Lipschitz with some Lipschitz constant L. In the linear regression example in Section 5, we consider the empirical risk minimization setting, where the number of data points n_i that agent i possesses is finite and the dataset is given and fixed. The Lipschitz constant L will in general depend on the dataset, but will be finite as long as the number of data points n is finite. For example, in the case of linear regression, we have

$$f_i(x)=\sum\nolimits_{j=1}^{n_i}\left(y_j^i-x^TX_j^i\right)^2+\frac{1}{2\lambda N}\|x\|^2, \qquad (126)$$

where agent i possesses a dataset $\mathcal{D}_i:=\{(X_j^i,y_j^i)\}_{j=1}^{n_i}$ of n_i data points. The Hessian of f_i satisfies

$$\nabla^2 f_i(x)=2\sum\nolimits_{j=1}^{n_i}X_j^i\left(X_j^i\right)^T+\frac{1}{\lambda N}I,$$

where I is the identity matrix. Therefore, Hessian of f_i is uniformly bounded satisfying $\|\nabla^2 f_i(x)\|\le 2\sum_{j=1}^{n_i}\|X_j^i\|^2+\frac{1}{\lambda N}$. Therefore, we can take the Lipschitz constant to be

$$L=2\max_{i=1,2,\dots,N}\left(\sum\nolimits_{j=1}^{n_i}\left\|X_j^i\right\|^2\right)+\frac{1}{\lambda N},$$

and this constant is finite because the number of samples $n = \sum_{i=1}^{N} n_i$ is finite and the data points X_j^i are given and fixed. This is the setting considered in our paper, and therefore our uniformly Lipschitz assumptions are satisfied.

More generally, one can try to bound L almost surely, i.e. for almost every realization of the dataset. If we take X_j^i to be random without a compact support (i.e. when data is Gaussian), then L will not be bounded almost surely. However, if the input data is bounded (which can often hold in machine learning practice naturally after normalizing/preprocessing data if necessary), then we will have L finite almost surely and our Lipschitz assumption will hold for almost every realization of the dataset. By similar computations, we can have the same conclusions for logistic regression. In other words, in the empirical risk minimization setting that we consider when each agent has finitely many data points and the dataset is fixed, our uniform Lipschitz gradient assumption will hold although the Lipschitz constant L will depend on the dataset. If we assume further that data has compact support, our Lipschitz assumption will hold almost surely.

It is worth noting that the recent elegant approach in Barkhagen et al. (2021) applies even if the data does not have compact support and when n goes to infinity and can handle non-i.i.d. L-mixing data streams. However, Barkhagen et al. (2021) considers the centralized setting and distributed setting is not discussed. It is not clear how to apply their techniques to the distributed setting but this would definitely be an interesting future research direction. Also, Barkhagen et al. (2021) does not discuss the stochastic gradient Hamiltonian Monte Carlo case, whereas our analysis framework provides a uniform approach where we study the stochastic gradient Hamiltonian Monte Carlo as well in the distributed setting.

References

Sungjin Ahn, Babak Shahbaba, and Max Welling. Distributed stochastic gradient MCMC. In *International Conference on Machine Learning*, pages 1044–1052, 2014.

Sungjin Ahn, Anoop Korattikara, Nathan Liu, Suju Rajan, and Max Welling. Large-scale distributed Bayesian matrix factorization using stochastic gradient MCMC. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 9–18, 2015.

Ömer Deniz Akyildiz and Sotirios Sabanis. Nonasymptotic analysis of Stochastic Gradient Hamiltonian Monte Carlo under local conditions for nonconvex optimization. *arXiv:2002.05465*, 2020.

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*, 2020.

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. Curran Associates, Inc., 2019.

Necdet Serhat Aybat, Alireza Fallah, Mert Gürbüzbalaban, and Asuman Ozdaglar. Robust accelerated gradient methods for smooth strongly convex functions. *SIAM Journal on Optimization*, 30(1):717–751, 2020.

M. Barkhagen, N.H. Chau, É. Moulines, M. Rásonyi, S. Sabanis, and Y. Zhang. On stochastic gradient Langevin dynamics with dependent data streams in the logconcave case. *Bernoulli*, 27(1):1–33, 2021.

Dimitri P Bertsekas and John N Tsitsiklis. *Parallel and Distributed Computation: Numerical Methods*, volume 23. Prentice Hall, Englewood Cliffs, NJ, 1989.

Doron Blatt and Alfred Hero. Distributed maximum likelihood estimation for sensor networks. In *2004 IEEE International Conference on Acoustics, Speech, and Signal Processing*, volume 3, pages iii–929. IEEE, 2004.

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

Tamara Broderick, Nicholas Boyd, Andre Wibisono, Ashia C Wilson, and Michael I Jordan. Streaming variational Bayes. In *Advances in Neural Information Processing Systems*, pages 1727–1735, 2013.

Sébastien Bubeck, Ronen Eldan, and Joseph Lehec. Sampling from a log-concave distribution with Projected Langevin Monte Carlo. In *Advances in Neural Information Processing Systems (NIPS)*, volume 28, 2015.

Jose Cadena, Priyadip Ray, Hao Chen, Braden Soper, Deepak Rajan, Anton Yen, and Ryan Goldhahn. Stochastic gradient-based distributed Bayesian estimation in cooperative sensor networks. *IEEE Transactions on Signal Processing*, 69:1713–1724, 2021.

Trevor Campbell and Tamara Broderick. Automated scalable Bayesian inference via Hilbert coresets. *Journal of Machine Learning Research*, 20(1):551–588, 2019.

Trevor Campbell and Jonathan P. How. Approximate decentralized Bayesian inference. In *Proceedings of the Thirtieth Conference on Uncertainty in Artificial Intelligence*, pages 102–111, July 2014.

B. Can, S. Soori, N. S. Aybat, M.M. Dehvani, and M. Gürbüzbalaban. Decentralized computation of effective resistances and acceleration of distributed optimization algorithms. *arXiv e-print arXiv:1907.13110*, 2019a.

Bugra Can, Mert Gürbüzbalaban, and Lingjiong Zhu. Accelerated linear convergence of stochastic momentum methods in Wasserstein distances. In *International Conference on Machine Learning*, pages 891–901, 2019b.

Yu Cao, Jianfeng Lu, and Lihan Wang. On explicit L^2-convergence rate estimate for underdamped Langevin dynamics. *arXiv:1908.04746*, 2019.

Niladri S Chatterji, Nicolas Flammarion, Yi-An Ma, Peter L Bartlett, and Michael I Jordan. On the theory of variance reduction for stochastic gradient Monte Carlo. In *International Conference on Machine Learning*, pages 764–773, 2018.

Ngoc Huy Chau, Éric Moulines, Miklos Rásonyi, Sotirios Sabanis, and Ying Zhang. On stochastic graident Langevin dynamics with dependent data streams: the fully non-convex case. *arXiv:1905.13142*, 2019.

Changyou Chen, Nan Ding, and Lawrence Carin. On the convergence of stochastic gradient MCMC algorithms with high-order integrators. In *Advances in Neural Information Processing Systems (NIPS)*, pages 2278–2286, 2015.

Changyou Chen, David Carlson, Zhe Gan, Chunyuan Li, and Lawrence Carin. Bridging the gap between stochastic gradient MCMC and stochastic optimization. In *Proceedings of the 19th International Conference on Artificial Intelligence and Statistics*, pages 1051–1060, 2016a.

Changyou Chen, Nan Ding, Chunyuan Li, Yizhe Zhang, and Lawrence Carin. Stochastic gradient MCMC with stale gradients. In *Advances in Neural Information Processing Systems*, volume 29, pages 2937–2945, 2016b.

Tianqi Chen, Emily Fox, and Carlos Guestrin. Stochastic gradient Hamiltonian Monte Carlo. In *International Conference on Machine Learning*, pages 1683–1691, 2014.

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)*, pages 186–211, 2018.

Xiang Cheng, Niladri S. Chatterji, Yasin Abbasi-Yadkori, Peter L. Bartlett, and Michael I. Jordan. Sharp Convergence Rates for Langevin Dynamics in the Nonconvex Setting. *arXiv:1805.01648*, 2018.

Xiang Cheng, Niladri S Chatterji, Peter L Bartlett, and Michael I Jordan. Underdamped Langevin MCMC: A non-asymptotic analysis. In *Proceedings of the 31st Conference On Learning Theory*, pages 300–323, 2018.

Arkabandhu Chowdhury and Christopher Jermaine. Parallel and distributed MCMC via shepherding distributions. In *International Conference on Artificial Intelligence and Statistics*, pages 1819–1827. PMLR, 2018.

Fan RK Chung and Fan Chung Graham. *Spectral Graph Theory*, volume 92. American Mathematical Society, 1997.

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.

Arnak S. Dalalyan and Avetik G. Karagulyan. User-friendly guarantees for the Langevin Monte Carlo with inaccurate gradient. *Stochastic Processes and their Applications*, 129 (12):5278–5311, 2019.

Arnak S Dalalyan and Lionel Riou-Durand. On sampling from a log-concave density using kinetic Langevin diffusions. *Bernoulli*, 26(3):1956–1988, 2020.

Olivier Devolder, François Glineur, and Yurii Nesterov. First-order methods of smooth convex optimization with inexact oracle. *Mathematical Programming*, 146(1-2):37–75, 2014.

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.

Alain Durmus and Eric Moulines. Non-asymptotic convergence analysis for the Unadjusted Langevin Algorithm. *Annals of Applied Probability*, 27(3):1551–1587, 2017.

Alain Durmus and Eric Moulines. High-dimensional Bayesian inference via the Unadjusted Langevin Algorithm. *Bernoulli*, 25(4A):2854–2882, 2019.

Andreas Eberle, Arnaud Guillin, and Raphael Zimmer. Couplings and quantitative contraction rates for Langevin dynamics. *Annals of Probability*, 47(4):1982–2010, 2019.

Murat A. Erdogdu and Rasa Hosseinzadeh. Convergence analysis of Langevin Monte Carlo in chi-square divergence. *arXiv:2007.11612*, 2020.

Alireza Fallah, Mert Gürbüzbalaban, Asuman Ozdaglar, Umut Şimşekli, and Lingjiong Zhu. Robust Distributed Accelerated Stochastic Gradient Methods for Multi-Agent Networks. *arXiv:1910.08701*, 2019.

Nicolas Flammarion and Francis Bach. From averaging to acceleration, there is only a step-size. In *Conference on Learning Theory*, pages 658–695, 2015.

Xuefeng Gao, Mert Gürbüzbalaban, and Lingjiong Zhu. Global convergence of Stochastic Gradient Hamiltonian Monte Carlo for non-convex stochastic optimization: Non-asymptotic performance bounds and momentum-based acceleration. *arXiv:1809.04618*, 2018.

Xuefeng Gao, Mert Gürbüzbalaban, and Lingjiong Zhu. Breaking reversibility accelerates Langevin dynamics for global non-convex optimization. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 33, 2020.

Hong Ge, Yutian Chen, Moquan Wan, and Zoubin Ghahramani. Distributed inference for Dirichlet process mixture models. *International Conference on Machine Learning*, pages 2276–2284, 2015.

Euhanna Ghadimi, Hamid Reza Feyzmahdavian, and Mikael Johansson. Global convergence of the Heavy-ball method for convex optimization. In *2015 European Control Conference (ECC)*, pages 310–315, 2015.

Clark R Givens and Rae Michael Shortt. A class of Wasserstein metrics for probability distributions. *The Michigan Mathematical Journal*, 31(2):231–240, 1984.

Eduard Gorbunov, Darina Dvinskikh, and Alexander Gasnikov. Optimal decentralized distributed algorithms for stochastic convex optimization. *arXiv:1911.07363*, 2019.

Lie He, An Bian, and Martin Jaggi. COLA: Decentralized linear learning. In *Advances in Neural Information Processing Systems*, pages 4536–4546, 2018.

Hadrien Hendrikx, Francis Bach, and Laurent Massoulié. An accelerated decentralized stochastic proximal algorithm for finite sums. In *Advances in Neural Information Processing Systems*, pages 954–964, 2019.

Peter D Hoff. *A First Course in Bayesian Statistical Methods*, volume 580. Springer, 2009.

Matthew Hoffman, Francis R Bach, and David M Blei. Online learning for latent Dirichlet allocation. In *Advances in Neural Information Processing Systems*, pages 856–864, 2010.

Jonathan Huggins, Trevor Campbell, and Tamara Broderick. Coresets for scalable Bayesian logistic regression. In *Advances in Neural Information Processing Systems*, pages 4080–4088, 2016.

Prateek Jain, Sham M Kakade, Rahul Kidambi, Praneeth Netrapalli, and Aaron Sidford. Accelerating stochastic gradient descent for least squares regression. In *Conference on Learning Theory*, pages 545–604. PMLR, 2018.

Vyacheslav Kungurtsev. Stochastic gradient Langevin dynamics on a distributed network. *arXiv:2001.00665, January 2020*, January 2020.

Nurdan Kuru, Ş. İlker Birbil, Mert Gürbüzbalaban, and Sinan Yildirim. Differentially Private Accelerated Optimization Algorithms. *arXiv e-prints*, art. arXiv:2008.01989, August 2020.

Anusha Lalitha, Xinghan Wang, Osman Kilinc, Yongxi Lu, Tara Javidi, and Farinaz Koushanfar. Decentralized Bayesian learning over graphs. *arXiv preprint arXiv:1905.10466*, 2019.

Benedict Leimkuhler, Charles Matthews, and Gabriel Stoltz. The computation of averages from equilibrium and nonequilibrium Langevin molecular dynamics. *IMA Journal of Numerical Analysis*, 36(1):13–79, 2016.

Dahua Lin. Online learning of nonparametric mixture models via sequential variational approximation. In *Advances in Neural Information Processing Systems*, pages 395–403, 2013.

Yanli Liu, Yuan Gao, and Wotao Yin. An improved analysis of stochastic gradient descent with momentum. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 33, 2020.

Yi-An Ma, Niladri S. Chatterji, Xiang Cheng, Nicolas Flammarion, Peter L. Bartlett, and Michael I. Jordan. Is there an analog of Nesterov acceleration for gradient-based MCMC? *Bernoulli*, 27(3):1942–1992, 2021.

Hesameddin Mohammadi, Meisam Razaviyayn, and Mihailo R Jovanović. Robustness of accelerated first-order algorithms for strongly convex optimization problems. *IEEE Transactions on Automatic Control*, 66(6):2480–2495, 2021.

Angelia Nedic. Distributed gradient methods for convex machine learning problems in networks: Distributed optimization. *IEEE Signal Processing Magazine*, 37(3):92–101, 2020.

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

Willie Neiswanger, Chong Wang, and Eric P Xing. Asymptotically exact, embarrassingly parallel MCMC. In *Proceedings of the Thirtieth Conference on Uncertainty in Artificial Intelligence*, pages 623–632, 2014.

Yurii Nesterov. *Introductory Lectures on Convex Optimization: A Basic Course*, volume 87. Springer Science & Business Media, 2013.

Yurii E Nesterov. A method for solving the convex programming problem with convergence rate $o(1/k^2)$. In *Dokl. Akad. Nauk SSSR*, volume 269, pages 543–547, 1983.

Robert Nishihara, Iain Murray, and Ryan P Adams. Parallel MCMC with generalized elliptical slice sampling. *Journal of Machine Learning Research*, 15(1):2087–2112, 2014.

Alex Olshevsky. Linear time average consensus and distributed optimization on fixed graphs. *SIAM Journal on Control and Optimization*, 55(6):3990–4014, 2017.

Anjaly Parayil, He Bai, Jemin George, and Prudhvi Gurram. Decentralized Langevin dynamics for Bayesian learning. In *Advances in Neural Information Processing Systems*, 2020.

Grigorios A Pavliotis. *Stochastic Processes and Applications: Diffusion processes, the Fokker-Planck and Langevin Equations*, volume 60. Springer, 2014.

Nicholas G. Polson and Vadim Sokolov. Deep learning: A Bayesian perspective. *Bayesian Analysis*, 12(4):1275–1304, 12 2017.

Boris T. Polyak. Introduction to Optimization. Translations Series in Mathematics and Engineering. *Optimization Software*, 1987.

Shi Pu, Alex Olshevsky, and Ioannis Ch Paschalidis. Asymptotic network independence in distributed stochastic optimization for machine learning: Examining distributed and centralized stochastic gradient descent. *IEEE Signal Processing Magazine*, 37(3):114–122, 2020.

Guannan Qu and Na Li. Accelerated distributed Nesterov gradient descent for smooth and strongly convex functions. In *2016 54th Annual Allerton Conference on Communication, Control, and Computing (Allerton)*, pages 209–216. IEEE, 2016.

Michael G Rabbat and Robert D Nowak. Decentralized source localization and tracking wireless sensor networks. In *2004 IEEE International Conference on Acoustics, Speech, and Signal Processing*, volume 3, pages iii–921. IEEE, 2004.

Maxim Rabinovich, Elaine Angelino, and Michael I. Jordan. Variational consensus Monte Carlo. In *Advances in Neural Information Processing Systems*, volume 28, pages 1207–1215, 2015.

Maxim Raginsky, Alexander Rakhlin, and Matus Telgarsky. Non-convex learning via stochastic gradient Langevin dynamics: a nonasymptotic analysis. In *Conference on Learning Theory*, pages 1674–1703, 2017.

Lewis J Rendell, Adam M Johansen, Anthony Lee, and Nick Whiteley. Global consensus Monte Carlo. *Journal of Computational and Graphical Statistics*, pages 1–11, 2020.

Masa-Aki Sato. Online model selection based on the variational Bayes. *Neural Computation*, 13(7):1649–1681, 2001.

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.

Steven L Scott. Comparing consensus Monte Carlo strategies for distributed Bayesian computation. *Brazilian Journal of Probability and Statistics*, 31(4):668–685, 2017.

Steven L Scott, Alexander W Blocker, Fernando V Bonassi, Hugh A Chipman, Edward I George, and Robert E McCulloch. Bayes and big data: The consensus Monte Carlo algorithm. *International Journal of Management Science and Engineering Management*, 11(2):78–88, 2016.

Weijie Su, Stephen Boyd, and Emmanuel J Candes. A differential equation for modeling Nesterov's accelerated gradient method: Theory and insights. *Journal of Machine Learning Research*, 17(1):5312–5354, 2016.

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*, art. arXiv:2003.10309, March 2020a.

Brian Swenson, Ryan Murray, Soummya Kar, and H. Vincent Poor. Distributed Stochastic Gradient Descent and Convergence to Local Minima. *arXiv e-prints*, art. arXiv:2003.02818, March 2020b.

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.

Yee Whye Teh, Alexandre H Thiery, and Sebastian J Vollmer. Consistency and fluctuations for stochastic gradient Langevin dynamics. *Journal of Machine Learning Research*, 17 (1):193–225, 2016.

John Nikolas Tsitsiklis. Problems in decentralized decision making and computation. Technical report, Massachusetts Inst of Tech Cambridge Lab for Information and Decision Systems, 1984.

César A Uribe, Soomin Lee, Alexander Gasnikov, and Angelia Nedić. Optimal algorithms for distributed optimization. *arXiv preprint arXiv:1712.00232*, 2017.

Cédric Villani. *Optimal Transport: Old and New*. Springer, Berlin, 2009.

Hao Wang and Dit-Yan Yeung. Towards Bayesian deep learning: A survey. *arXiv e-prints arXiv:1604.01662*, 2016.

Xiangyu Wang and David B Dunson. Parallelizing MCMC via Weierstrass sampler. *arXiv preprint arXiv:1312.4605*, 2013.

Xiangyu Wang, Fangjian Guo, Katherine A Heller, and David B Dunson. Parallelizing MCMC with random partition trees. *Advances in Neural Information Processing Systems*, 28:451–459, 2015.

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.

David P Woodruff and Qin Zhang. When distributed computation is communication expensive. *Distributed Computing*, 30(5):309–323, 2017.

Lin Xiao, Stephen Boyd, and Sanjay Lall. A space-time diffusion scheme for peer-to-peer least-squares estimation. In *Proceedings of the 5th International Conference on Information Processing in Sensor Networks*, pages 168–176, 2006.

Ran Xin and Usman A Khan. Distributed heavy-ball: A generalization and acceleration of first-order methods with gradient tracking. *IEEE Transactions on Automatic Control*, 65 (6):2627–2633, 2020.

Jinming Xu, Ye Tian, Ying Sun, and Gesualdo Scutari. Accelerated primal-dual algorithms for distributed smooth convex optimization over networks. In Silvia Chiappa and Roberto Calandra, editors, *Proceedings of the Twenty Third International Conference on Artificial Intelligence and Statistics*, volume 108 of *Proceedings of Machine Learning Research*, pages 2381–2391, Online, 26–28 Aug 2020. PMLR.

Minjie Xu, Balaji Lakshminarayanan, Yee Whye Teh, Jun Zhu, and Bo Zhang. Distributed Bayesian posterior sampling via moment sharing. In *Advances in Neural Information Processing Systems*, pages 3356–3364, 2014.

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.

Tao Yang, Xinlei Yi, Junfeng Wu, Ye Yuan, Di Wu, Ziyang Meng, Yiguang Hong, Hong Wang, Zongli Lin, and Karl H. Johansson. A survey of distributed optimization. *Annual Reviews in Control*, 47:278 – 305, 2019.

Kun Yuan, Qing Ling, and Wotao Yin. On the convergence of decentralized gradient descent. *SIAM Journal on Optimization*, 26:1835–1854, 2016.

Ying Zhang, Ömer Deniz Akyildiz, Theodoros Damoulas, and Sotirios Sabanis. Nonasymptotic estimates for Stochastic Gradient Langevin Dynamics under local conditions in nonconvex optimization. *arXiv:1910.02008*, 2019.

Difan Zou, Pan Xu, and Quanquan Gu. Stochastic variance-reduced Hamilton Monte Carlo methods. In *International Conference on Machine Learning*, pages 6028–6037, 2018a.

Difan Zou, Pan Xu, and Quanquan Gu. Subsampled stochastic variance-reduced gradient Langevin dynamics. In *International Conference on Uncertainty in Artificial Intelligence*, 2018b.

Difan Zou, Pan Xu, and Quanquan Gu. Stochastic gradient Hamiltonian Monte Carlo methods with recursive variance reduction. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 32, 2019.