

DIGing–SGLD: Decentralized and Scalable Langevin Sampling over Time–Varying Networks

Waheed U. Bajwa[†], Mert Gürbüzbalaban^{†,‡}, Mustafa Ali Kutbay[‡],
Lingjiong Zhu[§], Muhammad Zulqarnain[†]

[†]Department of Electrical and Computer Engineering, Rutgers University, Piscataway, NJ, USA

[‡]Department of Management Science and Information Systems, Rutgers University, Piscataway, NJ, USA

[§]Department of Mathematics, Florida State University, Tallahassee, FL, USA

Abstract

Sampling from a target distribution induced by training data is central to Bayesian learning, with Stochastic Gradient Langevin Dynamics (SGLD) serving as a key tool for scalable posterior sampling and decentralized variants enabling learning when data are distributed across a network of agents. This paper introduces DIGing-SGLD, a decentralized SGLD algorithm designed for scalable Bayesian learning in multi-agent systems operating over time-varying networks. Existing decentralized SGLD methods are restricted to static network topologies, and many exhibit steady-state sampling bias caused by network effects, even when full batches are used. DIGing-SGLD overcomes these limitations by integrating Langevin-based sampling with the gradient-tracking mechanism of the DIGing algorithm, originally developed for decentralized optimization over time-varying networks, thereby enabling efficient and bias-free sampling without a central coordinator. To our knowledge, we provide the first finite-time non-asymptotic Wasserstein convergence guarantees for decentralized SGLD-based sampling over time-varying networks, with explicit constants. Under standard strong convexity and smoothness assumptions, DIGing-SGLD achieves geometric convergence to an $O(\sqrt{\eta})$ neighborhood of the target distribution, where η is the stepsize, with dependence on the target accuracy matching the best-known rates for centralized and static-network SGLD algorithms. Numerical experiments on Bayesian linear and logistic regression validate the theoretical results and demonstrate the strong empirical performance of DIGing-SGLD under dynamically evolving network conditions.

Keywords: Decentralized Bayesian learning, decentralized sampling, gradient tracking, stochastic gradient Langevin dynamics, time-varying networks

1 Introduction

Consider a random vector $X \in \mathbb{R}^d$ with density $\pi(x)$ and, without loss of generality, write $\pi(x) \propto e^{-f(x)}$. The objective is to generate samples from π given access to f . A canonical instance arises in Bayesian machine learning: given n independent and identically distributed (i.i.d.) observations $Z = \{z_i\}_{i=1}^n$ with likelihood $p(z | x)$ and prior $p(x)$, the posterior satisfies

$$\pi(x) = p(x | Z) \propto p(x) \prod_{i=1}^n p(z_i | x) \iff f(x) = -\log p(x) - \sum_{i=1}^n \log p(z_i | x). \quad (1.1)$$

Posterior sampling enables principled Bayesian estimation and uncertainty quantification in a wide range of models and tasks, including logistic regression, linear and nonlinear regression, principal component analysis, and neural-network training (see, e.g., [32, 60] for representative applications).

*The authors are listed in alphabetical order. The authors can be reached at the following email addresses: waheed.bajwa@rutgers.edu, mg1366@rutgers.edu, mustafa.kutbay@newark.rutgers.edu, zhu@math.fsu.edu, m.zulqarnain@rutgers.edu.

A widely used route to sampling is Markov chain Monte Carlo (MCMC) [9, 28]. Among MCMC methods, Langevin-type algorithms discretize the overdamped Langevin diffusion and exploit gradient information about f [19]. The Unadjusted Langevin Algorithm (ULA) requires exact gradients $\nabla f(x)$ at every step [22]; when f aggregates many data terms, computing $\nabla f(x)$ entails a full pass over the dataset and becomes computationally expensive. Stochastic Gradient Langevin Dynamics (SGLD) addresses this by replacing the full gradient with unbiased stochastic estimates built from mini-batches and by avoiding Metropolis–Hastings corrections [10, 47, 56, 61]; this leads to scalable sampling procedures that remain effective on large datasets and high-dimensional models. These centralized methods, however, presuppose that data (or stochastic gradients) can be accessed and aggregated at a single location each iteration.

In many modern systems, data and computation are distributed across a network of devices, such as sensors, Internet-of-Things (IoT) platforms, autonomous fleets, and multi-robot systems, where privacy, bandwidth, and energy constraints preclude raw-data aggregation [44, 64]. In such settings, it is natural to consider the decomposition

$$f(x) = \sum_{j=1}^N f_j(x), \quad (1.2)$$

over N *agents*, where the term *agent* generically refers to any computational entity (e.g., a device, node, or processor) capable of performing local computation and communication. Each agent j can compute (stochastic) gradients of its local component f_j but can only exchange limited information with its immediate neighbors on a communication graph. This setting encompasses, for instance, cases in which n data samples are partitioned across N agents, each holding (for simplicity) n/N samples. Because neither the full gradient $\nabla f(x)$ nor a centralized stochastic-gradient estimator is accessible, classical SGLD cannot be applied directly. Instead, one must design sampling dynamics that rely solely on local gradient information and constrained inter-agent communication. While decentralized and federated optimization have been extensively studied in both deterministic and stochastic regimes (see, e.g., [1, 2, 21, 39, 41, 43, 51]), decentralized *sampling* introduces additional challenges arising from the interplay between gradient noise, network mixing, and sampling bias.

Although decentralized variants of ULA and SGLD have been developed in recent years [6, 14, 30, 31, 45], these approaches are restricted to *static* communication graphs, where the connectivity pattern between agents remains fixed throughout the sampling process. This assumption is limiting, as in many realistic multi-agent systems the communication topology is inherently *time-varying*: links may appear or disappear due to agent mobility, wireless interference, packet drops, or asynchronous operation, and networks may also be reconfigured to enhance privacy, alleviate congestion, or improve robustness against failures and attacks. Consequently, modeling the network as time-varying is more realistic than assuming a fixed topology [27, 40–42, 50, 55]. The variability of connectivity introduces new challenges to maintaining stability and sampling accuracy, and the development of decentralized SGLD algorithms that can operate reliably under such dynamic conditions remains an open problem. To our knowledge, the only work to date that has investigated decentralized Langevin sampling with time-varying connectivity is [35]. However, that work focuses on exact (deterministic) gradients, as in ULA, rather than stochastic gradients, as in SGLD, which limits scalability with the number of samples. In addition, it builds upon the decentralized optimization framework of [40], which is known to exhibit network-induced bias and slow convergence, resulting in significantly weaker theoretical rates.

In this work, we propose *DIGing Stochastic Gradient Langevin Dynamics* (DIGing-SGLD), a decentralized sampling algorithm designed to operate over *undirected time-varying* networks while relying only on local stochastic gradients and neighbor-to-neighbor communication. DIGing-SGLD

integrates stochastic gradient Langevin dynamics with the distributed inexact gradient-tracking mechanism originally introduced in the DIGing algorithm for decentralized optimization over time-varying graphs [42]. Each agent maintains an auxiliary variable that tracks the evolving average of local stochastic gradients across the network, thereby compensating for the drift caused by time-dependent communication weights. This mechanism enables agents to collaboratively approximate the global gradient and perform sampling without a central coordinator, extending decentralized Bayesian inference to dynamically changing network topologies without sacrificing convergence speed in the sense that the total number of iterations required to achieve a target accuracy ϵ matches the $\mathcal{O}\left(\frac{\log(1/\epsilon)}{\epsilon^2}\right)$ convergence rates established for SGLD-type algorithms on static graphs [31] and in centralized settings [20, Theorem 4].

1.1 Relation to Prior Work

On static graphs, *decentralized SGLD* (DE-SGLD) extends SGLD by interleaving local stochastic-gradient updates with consensus steps among neighboring agents [11, 30, 45]. Convergence guarantees for DE-SGLD have been established under strong convexity and smoothness assumptions for constant stepsizes, and in certain nonconvex regimes with decaying stepsizes [11, 45]. Momentum-based and event-triggered variants have also been proposed to enhance convergence behavior and reduce communication costs, respectively [30]. Within these approaches, using decaying stepsizes for strongly convex problems can lead to slow convergence, motivating constant-stepsize schemes in the literature. However, with constant stepsizes a key limitation is the emergence of a steady-state bias—even in the full-batch limit—arising from network-induced discrepancies among agents’ local gradients. EXTRA-based decentralized Langevin methods [31], including *EXTRA-SGLD* and *generalized EXTRA Langevin dynamics*, mitigate this issue on *static* graphs by incorporating bias-correction techniques from deterministic decentralized optimization, specifically, the EXTRA algorithm and its generalizations [34, 51]. The recent work [6] also employs a gradient-tracking mechanism similar in spirit to EXTRA but focuses on decentralized ULA with deterministic gradients, without addressing the stochastic-gradient setting. Another line of research considers *Metropolis-adjusted decentralized Hamiltonian Monte Carlo*, which achieves asymptotically exact sampling when local gradients are deterministic [36].

Unlike the case of static networks, the focus of this paper is on decentralized Langevin-based sampling using stochastic gradients over *time-varying* networks. For such networks, while there exists a substantial body of work on decentralized optimization (see, e.g., [40–42, 50, 55, 64]), these studies primarily address optimization or consensus problems rather than sampling from a target distribution. The closest related work on decentralized Langevin sampling over time-varying networks is [35], which analyzes deterministic-gradient Langevin dynamics on directed time-varying graphs and establishes convergence guarantees for the network average. However, the rates in [35] are based on the decentralized optimization framework of [40], which is known to exhibit network-induced bias and slow convergence, and the analysis assumes access to exact (deterministic) gradients, excluding the stochastic-gradient setting relevant to large-scale datasets.

More precisely, [35] considers decentralized Langevin methods with deterministic gradients over directed, time-varying graphs. Assuming each local function f_i is μ -strongly convex and L -smooth, and employing a decaying stepsize $\alpha_k = c_0/(1+k)$ with $c_0 = \min\{1/(2L), \mu/(4L^2)\}$, they show that the 2-Wasserstein distance \mathcal{W}_2 between the network average and the target distribution $\pi(x)$ satisfies $\mathcal{O}(1/((\mu c_0 - c)K^c))$ for any $c < \mu c_0$ after K iterations. Although the constants are not explicit, this bound implies that to achieve an accuracy ϵ , one requires $\Omega(\epsilon^{-1/\kappa^2})$ iterations, where $\kappa = L/\mu$ is the condition number. Consequently, for large κ , the admissible exponent $c = \mathcal{O}(1/\kappa^2)$ becomes arbitrarily small, leading to potentially conservative theoretical rates, that is, only weak

polynomial decay guarantees are available even when the gradients are deterministic. The design of decentralized sampling algorithms that admit favorable convergence guarantees when using mini-batches (stochastic gradients) over time-varying networks therefore remains a fundamental open problem from both theoretical and practical perspectives.

Before presenting our main contributions, we note that several other studies consider related but distinct problems. Some works focus on decentralized maximum-likelihood estimation rather than full Bayesian inference [4, 7, 49], while others address decentralized Bayesian inference by requiring agents to share local posterior distributions [12], an approach that is typically communication-intensive and computationally costly. We also note that distributed sampling methods such as Consensus Monte Carlo [48] and other ADMM-style distributed MCMC algorithms [58, 59], as well as parallel data-partitioning approaches [5, 17, 18], all rely on a central coordinator to aggregate information. Similarly, MCMC and Langevin algorithms developed for federated learning settings depend on a central server to orchestrate parameter updates [24, 38, 52]. In contrast, our setting considers an *ad hoc* network without any central coordinator capable of aggregating information. Consequently, these methods are not directly applicable to the decentralized, coordinator-free environment studied in this paper.

1.2 Our Contributions

We develop and analyze DIGing-SGLD, a decentralized sampling algorithm that extends SGLD to undirected *time-varying* networks. DIGing-SGLD integrates stochastic gradient Langevin dynamics with distributed inexact gradient tracking [42] to correct for network-induced drift and maintain consensus among agents under dynamic connectivity. From a theoretical standpoint, we establish the first *non-asymptotic convergence guarantees* with explicit constants for decentralized SGLD in time-varying networks. Under standard assumptions that each local function f_i is μ -strongly convex and L -smooth, and that the sequence of mixing matrices satisfies a joint spectral condition over bounded time intervals, we show that the marginal distribution of each agent’s iterate converges in the 2-Wasserstein distance at a *geometric rate* to an $\mathcal{O}(\sqrt{\eta})$ neighborhood of the target distribution, where η is the constant stepsize. In particular, we prove that with an appropriate choice of the stepsize $\eta = \mathcal{O}(\epsilon^2)$, after $K = \mathcal{O}\left(\frac{\log(1/\epsilon)}{\epsilon^2}\right)$ iterations, every agent can sample from a distribution that lies within ϵ of the target in \mathcal{W}_2 distance. The resulting bounds make explicit the dependence on the stepsize, network connectivity, gradient-noise variance, problem dimension, and number of agents. Notably, the ϵ -dependence of our rates matches the best known results for SGLD-type methods in both centralized and decentralized static-graph settings [20, 31], despite the additional challenges posed by time-varying networks.

Beyond the theoretical analysis, DIGing-SGLD addresses sampling over time-varying networks by unifying and extending the frameworks of DE-SGLD [30] and DIGing-based optimization [42], thereby unifying decentralized stochastic sampling and gradient-tracking-based optimization over time-varying networks within one framework. Experiments on Bayesian linear and logistic regression with both synthetic and real datasets corroborate the theory and demonstrate that, under time-varying network topologies, DIGing-SGLD outperforms DE-SGLD. Overall, this work provides the first mathematically rigorous foundation for *decentralized Bayesian sampling using stochastic gradients over time-varying networks*, addressing a key theoretical and practical gap in scalable, decentralized Bayesian inference.

1.3 Notation

We let $\mathbf{1}_m$ denote the m -dimensional all-ones column vector and I_n the $n \times n$ identity matrix. For $v = [v_1^\top, \dots, v_N^\top]^\top \in \mathbb{R}^{Nd}$ with $v_i \in \mathbb{R}^d$, its average is defined as $\bar{v} := \frac{1}{N} \sum_{i=1}^N v_i \in \mathbb{R}^d$, and the *replicated/stacked average vector* as $\bar{v} := [\bar{v}^\top, \dots, \bar{v}^\top]^\top = \frac{1}{N} ((\mathbf{1}_N \mathbf{1}_N^\top) \otimes I_d) v \in \mathbb{R}^{Nd}$. The *consensus error* of v is $\tilde{v} := v - \bar{v} = \mathcal{L}_N v$, where $\mathcal{L}_N := I_{Nd} - \frac{1}{N} ((\mathbf{1}_N \mathbf{1}_N^\top) \otimes I_d)$ is a symmetric $Nd \times Nd$ matrix. For $a \in \mathbb{R}^{Nd}$, we define $\|a\|_{\mathcal{L}_N} := \sqrt{\langle a, \mathcal{L}_N a \rangle}$. For a vector x , $\|x\|$ denotes the Euclidean norm, while for a random vector X , we write $\|X\|_{L_2} := (\mathbb{E}\|X\|^2)^{1/2}$.

We denote by $\mathcal{S}_{\mu,L}(\mathbb{R}^d)$ the class of functions $g : \mathbb{R}^d \rightarrow \mathbb{R}$ that are μ -strongly convex and L -smooth, i.e.,

$$\frac{\mu}{2} \|x - y\|^2 \leq g(x) - g(y) - \nabla g(y)^\top (x - y) \leq \frac{L}{2} \|x - y\|^2, \quad \forall x, y \in \mathbb{R}^d. \quad (1.3)$$

Let $\mathcal{P}_2(\mathbb{R}^d)$ denote the space of Borel probability measures on \mathbb{R}^d with finite second moment. For $\mu_1, \mu_2 \in \mathcal{P}_2(\mathbb{R}^d)$, the 2-Wasserstein distance is defined as $\mathcal{W}_2(\mu_1, \mu_2) := (\inf \mathbb{E}[\|Z_1 - Z_2\|^2])^{1/2}$, where the infimum is taken over all pairs of random variables (Z_1, Z_2) defined on a common probability space with marginal distributions μ_1 and μ_2 , respectively; see [57] for further details.

2 Preliminaries and Problem Formulation

Langevin algorithms are popular MCMC methods for obtaining samples from a given target density $\pi(x)$ of interest. If we consider $f(x) := -\log(\pi(x))$, classic first-order Langevin algorithms are based on discretizing the *overdamped Langevin diffusion*:

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

(see e.g. [3, 15, 19, 20, 22, 23, 25]). This diffusion admits $\pi(x)$ as the unique stationary distribution under some smoothness and growth assumptions on f [46]. Here, W_t is a standard d -dimensional Brownian motion for $t \geq 0$ initialized as $W_0 = 0$. ULA [22, 23] is a fundamental Langevin algorithm, based on the Euler-Maruyama discretization of (2.1), and results in the dynamics

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

where $\eta > 0$ is the stepsize parameter, and $w_k \in \mathbb{R}^d$ is a sequence of i.i.d. standard Gaussian random vectors $\mathcal{N}(0, I_d)$. Then, it is known that for strongly convex and smooth functions f , the iterates x_k converge to a neighborhood of the target distribution in the 2-Wasserstein distance where the size of the neighborhood goes to zero as $\eta \rightarrow 0$ [19, 20].

The ULA algorithm works with deterministic gradients which is often expensive to compute for Bayesian inference problems involving large data which motivated the development of Langevin algorithms that can support stochastic gradients. In particular, if one replaces the full gradient ∇f in (2.2) by a stochastic unbiased estimate of the gradient $\tilde{\nabla} f$, the resulting algorithm is known as the *stochastic gradient Langevin dynamics* (SGLD) (see, e.g., [61]) that found many applications to machine learning and large-scale Bayesian data analysis due to their scalability properties. As an example, when f has the finite-sum form $f(x) = \sum_{j=1}^N f_j(x)$ where the number of data points N is large, computing the gradient $\nabla f(x)$ requires going over all the data points and is often computationally expensive. However, if we consider estimating the gradients based on b randomly sampled data points, i.e. the estimator $\tilde{\nabla} f(x) := \sum_{\ell=1}^b \nabla f_{j_\ell}(x)$ where the index j_ℓ is sampled with replacement uniformly over the data indices $\{1, 2, \dots, N\}$ and b is small compared to N , then $\tilde{\nabla} f(x)$

is a stochastic unbiased estimate of the actual gradient $\nabla f(x)$ and is cheaper to compute. SGLD and its variants admit various performance guarantees in a variety of metrics and under various assumptions on f , see e.g. [19, 20, 47]. However, SGLD is still a centralized Langevin algorithm and is not applicable to decentralized sampling problems, which we discuss next.

2.1 Decentralized sampling over a network

In the context of decentralized sampling, the aim is to sample from a target distribution with density $\pi(x) \propto e^{-f(x)}$ on \mathbb{R}^d where the potential (or objective) function f admits a decomposition over the network: $f(x) := \sum_{i=1}^N f_i(x)$. In this context, the component function f_i and the estimates of the gradient ∇f_i is only available to agent $i \in \{1, 2, \dots, N\}$.

Decentralized stochastic gradient Langevin dynamics (DE-SGLD) is a decentralized version of the SGLD algorithm for undirected static graphs. Given an undirected graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ where $\mathcal{V} = \{1, 2, \dots, N\}$ is the set of nodes, and \mathcal{E} is the set of links/edges between the nodes. Each node i at step k owns a local variable $x_i^{(k)}$ and a component function $f_i(x)$ that contribute to the sum $f(x) = \sum_{i=1}^N f_i(x)$ and updates its local variable $x_i^{(k)}$ by taking weighted averages with the local variables $x_j^{(k)}$ of node i 's immediate neighbors $j \in \Omega_i := \{j : (i, j) \in \mathcal{E}\}$ and takes stochastic gradient steps with respect to their own component function $f_i(x)$ subject to additive Gaussian noise [30, 54]:

$$x_i^{(k+1)} = \sum_{j \in \Omega_i} W_{ij} x_j^{(k)} - \eta \tilde{\nabla} f_i(x_i^{(k)}) + \sqrt{2\eta} w_i^{(k+1)}. \quad (2.3)$$

Here, as before, $\eta > 0$ is the stepsize, $w_i^{(k)}$ are i.i.d. standard Gaussian random vectors with zero mean and identity covariance matrix for every i and k and $\tilde{\nabla} f_i(x_i^{(k)})$ is an unbiased estimator for the gradient $\nabla f_i(x_i^{(k)})$. Here W_{ij} is a double stochastic matrix, that determines the weights for the averaging and it respects the network structure, i.e. $W_{ij} > 0$ when $(i, j) \in \mathcal{E}$. If the Gaussian term is omitted, the iterations reduce to the decentralized stochastic gradient algorithm [26, 53] which itself builds on the decentralized gradient descent (DGD) methods [43]. There are other decentralized Langevin algorithms that can improve upon DE-SGLD in terms of their asymptotic bias behavior for static graphs [31].

A limitation of existing decentralized Langevin algorithms is the lack of theoretical guarantees when the underlying communication network is time-varying. In contrast, decentralized deterministic optimization methods such as DIGing [42] provide rigorous convergence guarantees for minimizing $f(x) = \sum_{i=1}^n f_i(x)$ over time-varying networks. To the best of our knowledge, however, no Langevin-based decentralized algorithms with rigorous performance guarantees have been established. To bridge this gap, and motivated by the DIGing framework, we introduce in the next section the *DIGing-SGLD* algorithm.

3 DIGing-SGLD for Time-Varying Graphs

In this section, we propose the *DIGing stochastic gradient Langevin dynamics* (DIGing-SGLD) for time-varying graphs. Consider a time-varying undirected graph sequence $\{\mathcal{G}(k)\}_{k=0}^\infty$. For every k , $\mathcal{G}(k)$ consists of a time-invariant set of agents $\mathcal{V} = \{1, 2, \dots, N\}$ and a set of time-varying edges $\mathcal{E}(k)$. The unordered pair of vertices $(j, i) \in \mathcal{E}(k)$ if and only if agents j and i can communicate at time k . By undirectedness, if $(j, i) \in \mathcal{E}(k)$ then $(i, j) \in \mathcal{E}(k)$. The set of neighbors of agents i —including agent i itself— at time k is defined as $\Omega_i(k) := \{j | (j, i) \in \mathcal{E}(k)\}$.

We will denote the local iterate of node i at iteration k by $x_i^{(k)}$ and make the following standard assumption about the stochasticity of the gradient noise, which basically says that the gradient noise is independent from the past iterates and is centered with a finite variance.

Assumption 3.1. *We assume that at iteration k , node i has access to $\tilde{\nabla} f_i(x_i^{(k)}, v_i^{(k+1)})$, which is an estimate of $\nabla f_i(x_i^{(k)})$, where $v_i^{(k+1)}$ is a random variable independent of the natural filtration \mathcal{F}_k generated by the iterates $\{x_j^{(t)}\}_{j=1, \dots, N, t=1, \dots, k}$. Moreover, the stochastic gradient noise at node i at iteration k defined as*

$$\xi_i^{(k+1)} := \tilde{\nabla} f_i(x_i^{(k)}, v_i^{(k+1)}) - \nabla f_i(x_i^{(k)}), \quad i = 1, 2, \dots, N, \quad (3.1)$$

is centered with a finite variance, i.e. it satisfies $\mathbb{E}[\xi_i^{(k+1)} | \mathcal{F}_k] = 0$ and $\mathbb{E} \|\xi_i^{(k+1)}\|^2 \leq \sigma^2$ for every $i = 1, 2, \dots, N$ and $k = 0, 1, 2, \dots$. To simplify notation, we suppress the dependence on $v_i^{(k+1)}$ and write $\tilde{\nabla} f_i(x_i^{(k)})$ for $\tilde{\nabla} f_i(x_i^{(k)}, v_i^{(k+1)})$.

Now, we are ready to introduce the iterates of DIGing-SGLD as follows:

$$x_i^{(k+1)} = \sum_{j \in \Omega_i(k)} W_{ij}^{(k)} x_j^{(k)} - \eta y_i^{(k)} + \sqrt{2\eta} w_i^{(k+1)}, \quad (3.2)$$

$$y_i^{(k+1)} = \sum_{j \in \Omega_i(k)} W_{ij}^{(k)} y_j^{(k)} + \tilde{\nabla} f_i(x_i^{(k+1)}) - \tilde{\nabla} f_i(x_i^{(k)}), \quad (3.3)$$

where $\tilde{\nabla} f_i(x_i^{(k+1)})$ are stochastic gradients and $w_i^{(k+1)}$ are standard d -dimensional Gaussian random vectors that are independent from the stochastic gradient vector $\tilde{\nabla} f_i(x_i^{(k)})$ as well as the natural filtration \mathcal{F}_k and are i.i.d. in both $i = 1, 2, \dots, N$ and $k = 0, 1, 2, \dots$. Here, $W^{(k)}$ denotes the mixing matrix at iteration k , and we initialize with $x_i^{(0)}$ and $y_i^{(0)} = \nabla f_i(x_i^{(0)})$. Since the underlying graph $\mathcal{G}(k)$ is time-varying, $W^{(k)}$ is also time-dependent. The precise structural assumptions on $W^{(k)}$ will be specified later.

Note that we can re-write DIGing stochastic gradient Langevin dynamics as follows

$$x_i^{(k+1)} = \sum_{j \in \Omega_i(k)} W_{ij}^{(k)} x_j^{(k)} - \eta y_i^{(k)} + \sqrt{2\eta} w_i^{(k+1)}, \quad (3.4)$$

$$y_i^{(k+1)} = \sum_{j \in \Omega_i(k)} W_{ij}^{(k)} y_j^{(k)} + \nabla f_i(x_i^{(k+1)}) - \nabla f_i(x_i^{(k)}) + \xi_i^{(k+2)} - \xi_i^{(k+1)}. \quad (3.5)$$

By introducing the function $F(x) : \mathbb{R}^{Nd} \rightarrow \mathbb{R}$:

$$F(x) := \sum_{i=1}^N f_i(x_i), \quad \text{for any } x := (x_1^\top, \dots, x_N^\top)^\top \in \mathbb{R}^{Nd}, \quad (3.6)$$

and by stacking the local variables $x_i^{(k)}, y_i^{(k)}$ into single vectors $x^{(k)} = \left[(x_1^{(k)})^\top, \dots, (x_N^{(k)})^\top \right]^\top \in \mathbb{R}^{Nd}$ and $y^{(k)} = \left[(y_1^{(k)})^\top, \dots, (y_N^{(k)})^\top \right]^\top \in \mathbb{R}^{Nd}$, we can also rewrite the algorithm (3.4)–(3.5) in the following form for N agents:

$$x^{(k+1)} = \mathcal{W}^{(k)} x^{(k)} - \eta y^{(k)} + \sqrt{2\eta} w^{(k+1)}, \quad (3.7)$$

$$y^{(k+1)} = \mathcal{W}^{(k)} y^{(k)} + \nabla F(x^{(k+1)}) - \nabla F(x^{(k)}) + \xi^{(k+2)} - \xi^{(k+1)}, \quad (3.8)$$

where $w^{(k)} := \left[(w_1^{(k)})^\top, \dots, (w_N^{(k)})^\top \right]^\top \in \mathbb{R}^{Nd}$, $\xi^{(k)} := \left[(\xi_1^{(k)})^\top, \dots, (\xi_N^{(k)})^\top \right]^\top \in \mathbb{R}^{Nd}$, and $\mathcal{W}^{(k)} := W^{(k)} \otimes I_d$ for any $k = 0, 1, 2, \dots$. Next, we consider the mixing matrices $W^{(k)}$ and introduce the notation

$$W_B^{(k)} := W^{(k)} W^{(k-1)} \dots W^{(k-B+1)}, \quad (3.9)$$

for any $k = 0, 1, 2, \dots$ and any $B = 1, 2, \dots$ with the convention that $W_B^{(k)} = I_N$ for any $k < 0$ and $W_0^{(k)} = I_N$ for any k . The product $W_B^{(k)}$ captures the connectivity of the graph over the time interval from $k - B + 1$ to k . Moreover, we introduce the notation,

$$\mathcal{W}_B^{(k)} := W_B^{(k)} \otimes I_d, \quad (3.10)$$

for any $k = 0, 1, 2, \dots$ and any $B = 1, 2, \dots$. We make the following assumption on the mixing matrices.

Assumption 3.2. *For any $k = 0, 1, 2, \dots$, the mixing matrix $W^{(k)} = (W_{ij}^{(k)}) \in \mathbb{R}^{N \times N}$ is symmetric and satisfies the following properties:*

- (i) (decentralized property) *If $i \neq j$ and the edge $(j, i) \notin \mathcal{E}(k)$, then $W_{ij}^{(k)} = 0$.*
- (ii) (double stochasticity) $W^{(k)} \mathbf{1}_N = \mathbf{1}_N$, $\mathbf{1}_N^\top W^{(k)} = \mathbf{1}_N^\top$.
- (iii) (joint spectral property) *There exists a positive integer B such that for every $k = 0, 1, 2, \dots$, $\delta := \sup_{k \geq B-1} \delta(k) < 1$, where $\delta(k) := \sigma_{\max} \left\{ W_B^{(k)} - \frac{1}{N} \mathbf{1}_N \mathbf{1}_N^\top \right\}$.*

Parts (i) and (ii) of Assumption 3.2 are standard and are imposed even when the underlying graph is static [30, 31, 65]. In particular, part (i) requires that the averaging operation respects the network's connectivity pattern, while part (ii) guarantees that if all nodes converge to the same vector, the mixing process preserves this consensus. Part (iii) of Assumption 3.2 controls the spectral gap $\Delta := 1 - \delta$ uniformly over iterations and ensures that the connectivity observed by the iterates across any B consecutive iterations is sufficient (i.e. $\Delta > 0$); this assumption, along with closely related variants, appears frequently in the literature; see e.g. [21, 42] and the references therein. Also, for static graphs, it is standard to assume part (iii) with $B = 1$ in which case $\delta = \delta(k)$ for every k [30, 31, 65].

For analysis purposes, we will also make the following assumption throughout on the local objectives.

Assumption 3.3. *We assume that each local objective function $f_i \in \mathcal{S}_{\mu, L}(\mathbb{R}^d)$ for every $i = 1, 2, \dots, N$; that is, each $f_i(x)$ is μ -strongly convex and L -smooth.*

Under this assumption, the global objective $f(x) = \sum_{i=1}^N f_i(x)$ admits a unique minimizer x_* , and the target distribution $\pi(x) \propto e^{-f(x)}$ is strongly log-concave. This assumption has been employed in the literature to analyze the DIGing algorithm in the literature [37, 42] for distributed optimization. In this work, we adopt the same assumption but pursue a fundamentally different goal – we focus on sampling rather than optimization. Moreover, while existing guarantees for decentralized Langevin algorithms such as DE-SGLD and EXTRA-SGLD also rely on this assumption, they are restricted to static communication graphs. In contrast, our analysis accommodates time-varying graphs.

3.1 Main Results

Before we proceed to the main result of the paper, we first introduce some notations. Let α, β, λ be all positive scalars and $\lambda \in (\delta^{1/B}, 1)$ with $B \geq 1$ so that $\delta < \lambda^B < 1$. Consider the non-negative quantities

$$\tilde{\omega}_1 := \frac{\lambda^B}{\lambda^B - \delta} \sum_{t=1}^B \lambda^{1-t} \left\| \tilde{y}^{(t-1)} \right\|_{L_2}, \quad \hat{\omega}_1 := \frac{\lambda^B}{\lambda^B - \delta} \cdot 2B\sigma\sqrt{N}, \quad (3.11)$$

$$\tilde{\omega}_3 := 2\sqrt{N} \left\| \bar{x}^{(0)} - x_* \right\|, \quad \hat{\omega}_3 := \frac{\sqrt{N}}{\lambda} \left(\sqrt{\frac{L(1+\alpha)}{\mu\alpha}} + \beta \right) \frac{1}{\mu} \left(\sigma + \sqrt{\frac{2d}{\eta}} \right), \quad (3.12)$$

$$\tilde{\omega}_4 := \frac{\lambda^B}{\lambda^B - \delta} \sum_{t=1}^B \lambda^{1-t} \left\| \tilde{x}^{(t-1)} \right\|_{L_2}, \quad \hat{\omega}_4 := \frac{\lambda^B}{\lambda^B - \delta} \cdot B\sqrt{2\eta Nd}, \quad (3.13)$$

where $\tilde{x}^{(t-1)} = x^{(t-1)} - \bar{x}^{(t-1)}$ and $\tilde{y}^{(t-1)} = y^{(t-1)} - \bar{y}^{(t-1)}$ for every $t = 1, \dots, B$. We also introduce:

$$\gamma_1 := \frac{\lambda(1 - \lambda^B)}{(\lambda^B - \delta)(1 - \lambda)}, \quad \gamma_2 := L \left(1 + \frac{1}{\lambda} \right), \quad (3.14)$$

$$\gamma_3 := \left(1 + \frac{\sqrt{N}}{\lambda} \left(\sqrt{\frac{L(1+\alpha)}{\mu\alpha}} + \beta \right) \right), \quad \gamma_4 := \frac{\eta(1 - \lambda^B)}{(\lambda^B - \delta)(1 - \lambda)}. \quad (3.15)$$

Note that when the stepsize $\eta > 0$ is sufficiently small we have $\gamma_1\gamma_2\gamma_3\gamma_4 \in (0, 1)$.

We now present our main results (Theorem 3.4 and Theorem 3.5). It shows that if the stepsize η is chosen sufficiently small—satisfying certain inequalities that guarantee the stability of our algorithm (in particular, ensuring that the L_2 norm of the iterates remains bounded)—then the distribution of the nodes converges linearly (i.e., at a geometric rate) in the 2-Wasserstein distance to a neighborhood of the target distribution. Moreover, the radius of this neighborhood scales as $\mathcal{O}(\sqrt{\eta})$ as $\eta \rightarrow 0$. We first provide the 2-Wasserstein convergence guarantees for the distribution of the average of iterates to the Gibbs distribution.

Theorem 3.4. *Consider the DIGing-SGLD algorithm with constant stepsize $\eta > 0$. Assume that $\left\| x^{(0)} \right\|_{L_2}$ is finite. Let $\alpha, \beta > 0$ be fixed scalars, and let $\lambda \in (\delta^{1/B}, 1)$, where $\delta \in (0, 1)$ is as given in Assumption 3.2. The stepsize $\eta > 0$ is chosen such that the following conditions are satisfied:*

$$\sqrt{1 - \frac{\eta\mu\beta}{\beta + 1}} \leq \lambda < 1, \quad \eta \leq \frac{1}{(1 + \alpha)L}, \quad \gamma_1\gamma_2\gamma_3\gamma_4 \in (0, 1), \quad (3.16)$$

where $\gamma_1, \gamma_2, \gamma_3, \gamma_4$ are defined in (3.14)–(3.15). Then, for every iteration k , we have

$$\mathcal{W}_2 \left(\text{Law} \left(\bar{x}^{(k)} \right), \pi \right) \leq E_1(k, \eta) + E_2(k, \eta, \delta), \quad (3.17)$$

where $E_1 = E_1(k, \eta)$ and $E_2 = E_2(k, \eta, \delta)$ are given by:

$$E_1 := (1 - \mu\eta)^k \left(\left(\mathbb{E} \left\| \bar{x}^{(0)} - x_* \right\|^2 \right)^{1/2} + \sqrt{2\mu^{-1}dN^{-1}} \right) + \frac{1.65L}{\mu} \sqrt{\eta dN^{-1}}, \quad (3.18)$$

$$E_2 := \eta^{1/2} \left(\frac{\eta}{\mu(1 - \frac{\eta L}{2})} + \frac{(1 + \eta L)^2}{\mu^2(1 - \frac{\eta L}{2})^2} \right)^{1/2} \cdot \left(\frac{3L^2 D^2 \eta \delta^{-2}}{N(1 - \delta^{\frac{1}{B}})^2} + \frac{6dL^2 \cdot \delta^{-2}}{1 - \delta^{\frac{2}{B}}} \right)^{1/2}$$

$$+ \frac{\sqrt{\eta}\sigma}{\sqrt{\mu(1-\frac{\eta L}{2})N}} + \left(\frac{\left(\delta^{\frac{2}{B}}\right)^k - \left(1 - \eta\mu\left(1 - \frac{\eta L}{2}\right)\right)^k}{\delta^{\frac{2}{B}} - 1 + \eta\mu\left(1 - \frac{\eta L}{2}\right)} \right)^{1/2} \cdot \frac{\sqrt{3}L \cdot \delta^{-1}}{\sqrt{N}} \delta^{\frac{1}{B}} \cdot \|x^{(0)}\|_{L_2}, \quad (3.19)$$

where x_* is the minimizer of f , $\bar{x}^{(0)} = \frac{1}{N} \sum_{i=1}^N x_i^{(0)}$,

$$D := \left[2 \left(\frac{\gamma_1 \gamma_2 \gamma_3 (\tilde{\omega}_4 + \hat{\omega}_4) + \gamma_1 \gamma_2 (\tilde{\omega}_3 + \hat{\omega}_3) + \tilde{\omega}_1 + \hat{\omega}_1}{1 - \gamma_1 \gamma_2 \gamma_3 \gamma_4} \right)^2 + \frac{4L^2}{N} \left(\frac{\gamma_3 \gamma_4 (\tilde{\omega}_1 + \hat{\omega}_1) + \gamma_3 (\tilde{\omega}_4 + \hat{\omega}_4) + \tilde{\omega}_3 + \hat{\omega}_3}{1 - \gamma_1 \gamma_2 \gamma_3 \gamma_4} \right)^2 + \frac{4}{N} \sigma^2 \right]^{1/2}, \quad (3.20)$$

with $\tilde{\omega}_i$ and $\hat{\omega}_i$ defined by (3.11)–(3.13) for $i = 1, 2, 3, 4$ and π is the Gibbs distribution with probability density function proportional to $\exp(-f(x))$.

Next, we present the non-asymptotic convergence guarantees for average of the 2-Wasserstein distance between the distribution of the iterates of each node to the Gibbs distribution.

Theorem 3.5. *Under the same setting as in Theorem 3.4, for every iteration k , we have*

$$\frac{1}{N} \sum_{i=1}^N \mathcal{W}_2 \left(\text{Law} \left(x_i^{(k)} \right), \pi \right) \leq E_1(k, \eta) + E_2(k, \eta, \delta) + E_3(k, \eta, \delta), \quad (3.21)$$

where $E_1 = E_1(k, \eta)$ and $E_2 = E_2(k, \eta, \delta)$ are given in (3.18)–(3.19) and $E_3 = E_3(k, \eta, \delta)$ is defined as:

$$E_3 := \frac{\sqrt{3}\delta^{-1}\delta^{\frac{k}{B}}}{\sqrt{N}} \|x^{(0)}\|_{L_2} + \frac{\sqrt{3}D\eta\delta^{-1}}{\sqrt{N}(1-\delta^{\frac{1}{B}})} + \frac{\sqrt{6d\eta}\delta^{-1}}{\sqrt{1-\delta^{\frac{2}{B}}}}, \quad (3.22)$$

where x_* is the minimizer of f , $\bar{x}^{(0)} = \frac{1}{N} \sum_{i=1}^N x_i^{(0)}$, D is given in (3.20), and π is the Gibbs distribution with the probability density function proportional to $\exp(-f(x))$.

Remark 1 (Interpretations of E_1, E_2 and E_3). In Theorem 3.4 and Theorem 3.5, E_1 serves as an upper bound on the 2-Wasserstein distance between the distribution of x_k , the k -th iterate of the (centralized) unadjusted Langevin algorithm to the Gibbs distribution. Moreover, E_2 is an upper bound on the 2-Wasserstein distance between the distribution of the average of iterates $\bar{x}^{(k)}$ and that of x_k . Finally, E_3 is an upper bound on the averaged 2-Wasserstein distance between the distribution of each node $x_i^{(k)}$ and that of the average of iterates $\bar{x}^{(k)}$.

Remark 2 (Feasibility of parameter choices). Theorem 3.4 and Theorem 3.5 ensures that the constraints in (3.16) can be satisfied simultaneously. Indeed, fix any $\alpha, \beta > 0$. Choose λ such that $\frac{\delta+1}{2} \leq \lambda^B < 1$. Under this choice, the quantities $\tilde{\omega}_i, \hat{\omega}_i, \gamma_i$ ($i = 1, 2, 3, 4$) defined in (3.13) remain uniformly bounded, i.e., all of order $\mathcal{O}(1)$ except $\hat{\omega}_3$ which is of order $\mathcal{O}(1/\sqrt{\eta})$. Next, select $\eta > 0$ sufficiently small (independently of λ) so that $\gamma_1 \gamma_2 \gamma_3 \gamma_4 \leq \frac{1}{2}$ and $\eta \leq \frac{1}{(1+\alpha)L}$. With this choice, D in (3.20) is of order $\mathcal{O}(1/\sqrt{\eta})$, $D\sqrt{\eta} = \mathcal{O}(1)$ as $\eta \rightarrow 0$ and the upper bound $E_1 + E_2 + E_3$ on the Wasserstein distance given in (3.21) scales like

$$E_1 + E_2 + E_3 = \mathcal{O}\left((1 - \mu\eta)^k\right) + \mathcal{O}\left(\left(1 - \eta\mu\left(1 - \frac{\eta L}{2}\right)\right)^k\right) + \mathcal{O}\left(\left(\delta^{1/B}\right)^k\right) + \mathcal{O}(\sqrt{\eta}),$$

which can be made arbitrarily small by choosing the stepsize $\eta > 0$ small enough. Finally, once η is fixed, we can always choose λ so that $\sqrt{1 - \frac{\eta\mu\beta}{\beta+1}} \leq \lambda < 1$, which is possible since the square root term lies strictly below 1 for any $\eta > 0$. Therefore, feasible pairs (η, λ) that satisfy the conditions (3.16) always exist. Moreover, as $k \rightarrow \infty$, the upper bound $E_1 + E_2 + E_3 = \mathcal{O}(\eta)$. This shows given any target accuracy $\epsilon > 0$, by choosing the stepsize η sufficiently small (as a function of ϵ) and the number of iterates k large enough, we can ensure each node's iterates are close to the target distribution, i.e. $\frac{1}{N} \sum_{i=1}^N \mathcal{W}_2(\text{Law}(x_i^{(k)}), \pi) \leq \epsilon$.

In the following, we will present some explicit particular choice of stepsize η , and the parameters λ, α, β . Other possible choices of $\eta, \lambda, \alpha, \beta$ exist and can be obtained based on optimizing the upper bound $E_1 + E_2 + E_3$ numerically, but here our aim is to give the parameters explicitly. We first present a lemma which suggests some explicit parameter choices within DIGing-SGLD and provides an explicit bound on the quantity $D\sqrt{\eta}$ showing it is $\mathcal{O}(1)$ as $\eta \rightarrow 0$. The main challenge in deriving an explicit bound on D and explicit parameter choices is that D depends on all four parameters $\eta, \lambda, \alpha, \beta$, which are constrained by the nonlinear conditions in (3.16). This bound will later be crucial for deriving the iteration complexity of DIGing-SGLD.

Lemma 3.6. *In the setting of Theorem 3.4, consider DIGing-SGLD with stepsize $\eta \in (0, \bar{\eta}]$, where*

$$\bar{\eta} := \frac{3(1 - \delta^2)}{\mu J_1} \quad \text{with} \quad J_1 := 3\kappa B^2 \left(1 + 4\sqrt{N}\sqrt{\kappa}\right) \quad \text{with} \quad \kappa := \frac{L}{\mu}, \quad (3.23)$$

and take $\alpha = 1, \beta = 2L/\mu$ and

$$\lambda(\eta) = \begin{cases} \sqrt[2B]{1 - \frac{\eta\mu}{1.5}}, & \text{if } \eta \in (0, \check{\eta}); \\ \sqrt[B]{\sqrt{\frac{\eta\mu J_1}{1.5}} + \delta}, & \text{if } \eta \in (\check{\eta}, \bar{\eta}]. \end{cases} \quad (3.24)$$

Then, conditions (3.16) are satisfied and Theorems 3.4 and 3.5 are applicable. Furthermore, $D\sqrt{\eta}$ where D is defined by (3.20) admits the bound

$$D\sqrt{\eta} \leq \bar{D} := \left[2 \left(\frac{\bar{\gamma}_1 \bar{\gamma}_2 \bar{\gamma}_3 (\bar{\omega}_4 + \bar{\omega}_1) + \bar{\gamma}_1 \bar{\gamma}_2 (\bar{\omega}_3 + \bar{\omega}_3) + \bar{\omega}_1 + \bar{\omega}_1}{1 - \bar{\gamma}_1 \bar{\gamma}_2 \bar{\gamma}_3 \bar{\gamma}_4} \right)^2 + \frac{4L^2}{N} \left(\frac{\bar{\gamma}_3 \bar{\gamma}_4 (\bar{\omega}_1 + \bar{\omega}_1) + \bar{\gamma}_3 (\bar{\omega}_4 + \bar{\omega}_4) + \bar{\omega}_3 + \bar{\omega}_3}{1 - \bar{\gamma}_1 \bar{\gamma}_2 \bar{\gamma}_3 \bar{\gamma}_4} \right)^2 + \frac{4}{N} \sigma^2 \right]^{1/2} \sqrt{\eta}, \quad (3.25)$$

where

$$\bar{\omega}_1 := \frac{\lambda^B}{\lambda^B - \delta} \sum_{t=1}^B \lambda^{1-t} \|\tilde{y}^{(t-1)}\|_{L_2}, \quad \bar{\omega}_1 := \frac{\lambda^B}{\lambda^B - \delta} \cdot 2B\sigma\sqrt{N}, \quad (3.26)$$

$$\bar{\omega}_3 := 2\sqrt{N} \|\bar{x}^{(0)} - x_*\|, \quad \bar{\omega}_3 := \frac{\sqrt{N}}{\lambda} \left(\sqrt{\frac{L(1+\alpha)}{\mu\alpha}} + \beta \right) \frac{1}{\mu} \left(\sigma + \sqrt{\frac{2d}{\eta}} \right), \quad (3.27)$$

$$\bar{\omega}_4 := \frac{\lambda^B}{\lambda^B - \delta} \sum_{t=1}^B \lambda^{1-t} \|\tilde{x}^{(t-1)}\|_{L_2}, \quad \bar{\omega}_4 := \frac{\lambda^B}{\lambda^B - \delta} \cdot B\sqrt{2\eta Nd}, \quad (3.28)$$

with

$$\lambda := \left(\frac{\sqrt{J_1^2 + (1 - \delta^2)J_1} + \delta}{J_1 + 1} \right)^{\frac{1}{B}}, \quad \bar{\gamma}_1 := \frac{\lambda \cdot (1 - \lambda^B)}{(\lambda^B - \delta)(1 - \lambda)}, \quad \bar{\gamma}_2 := L \left(1 + \frac{1}{\lambda} \right), \quad (3.29)$$

$$\bar{\gamma}_3 := \left(1 + \frac{\sqrt{N}}{\underline{\lambda}} \left(\sqrt{\frac{L(1+\alpha)}{\mu\alpha}} + \beta \right) \right), \quad \bar{\gamma}_4 := \frac{\bar{\eta} (1 - \underline{\lambda}^B)}{(\underline{\lambda}^B - \delta) (1 - \underline{\lambda})}. \quad (3.30)$$

Proof. The proof is deferred to Appendix A.1. \square

In the next result, we establish an iteration complexity bound that quantifies how many iterations of DIGing-SGLD are required to ensure the 2-Wasserstein error is at most ϵ . To achieve this complexity, we propose a stepsize that adapts explicitly to the target accuracy ϵ and build on the previous lemma.

Corollary 3.7. *In the setting of Theorem 3.5, let the target accuracy $\epsilon > 0$ be given. Consider DIGing-SGLD with stepsize*

$$\eta_* := \min(\bar{\eta}, \eta_{\text{noise}}(\epsilon)), \quad \text{with} \quad \eta_{\text{noise}}(\epsilon) := \min \left(\frac{\epsilon^2}{9 \cdot \bar{C}_3^2}, \frac{\epsilon}{3 \cdot \bar{C}_4} \right),$$

where $\bar{\eta}$ is defined by (3.23),

$$\begin{aligned} \bar{C}_3 &:= \frac{1.65L}{\mu} \sqrt{dN^{-1}} + \frac{\sqrt{6d}\delta^{-1}}{\sqrt{1 - \delta^{\frac{2}{B}}}} + \frac{2\sigma}{\sqrt{3\mu N}} + \frac{2}{\mu} \left(\frac{6dL^2\delta^{-2}}{1 - \delta^{\frac{2}{B}}} \right)^{1/2} + \frac{\sqrt{3}\delta^{-1}\bar{D}}{\sqrt{N}(1 - \delta^{\frac{1}{B}})} + \frac{2\bar{D}}{\mu} \left(\frac{3L^2\delta^{-2}}{N(1 - \delta^{\frac{1}{B}})^2} \right)^{1/2}, \\ \bar{C}_4 &:= \frac{2}{\sqrt{3\mu}} \cdot \left(\frac{6dL^2 \cdot \delta^{-2}}{1 - \delta^{\frac{2}{B}}} \right)^{1/2} + \frac{2}{\sqrt{3\mu}} \cdot \left(\frac{3L^2\delta^{-2}}{N(1 - \delta^{\frac{1}{B}})^2} \right)^{1/2} \bar{D}, \end{aligned}$$

with \bar{D} given in (3.25). Then, DIGing-SGLD satisfies the error bound

$$\frac{1}{N} \sum_{i=1}^N \mathcal{W}_2 \left(\text{Law} \left(x_i^{(k)} \right), \pi \right) \leq \epsilon \quad \text{after} \quad k \geq k_*(\epsilon) := \frac{3}{\mu\eta_*} \log \left(\frac{4(C_1 + C_2)}{\epsilon} \right)$$

iterations where

$$\bar{C}_1 = \left(\left(\mathbb{E} \left\| \bar{x}^{(0)} - x_* \right\|^2 \right)^{1/2} + \sqrt{2\mu^{-1}dN^{-1}} \right), \quad (3.31)$$

$$\bar{C}_2 = \frac{1}{\sqrt{1 - \frac{\bar{\eta}\mu}{1.5} - \delta^{\frac{2}{B}}}} \frac{\sqrt{3}L \cdot \delta^{-1}}{\sqrt{N}} \delta^{\frac{1}{B}} \cdot \left\| x^{(0)} \right\|_{L_2} + \frac{\sqrt{3}\delta^{-1}}{\sqrt{N}} \left\| x^{(0)} \right\|_{L_2}. \quad (3.32)$$

Proof. The proof is deferred to Appendix B. \square

Remark 3. We observe that in the setting of Corollary 3.7, the constants \bar{C}_i for $i = 1, 2, 3, 4$ are all independent from the target accuracy $\epsilon > 0$ and the stepsize $\eta_* = \Theta(\epsilon^2)$ as $\epsilon \rightarrow 0$. Hence, the iteration complexity of DIGing-SGLD satisfies

$$k_*(\epsilon) = \Theta \left(\frac{\log(1/\epsilon)}{\epsilon^2} \right).$$

3.2 Proofs of the Main Results

In this section, we present the proof of Theorem 3.4 and Theorem 3.5 by establishing a sequence of technical lemmas whose proofs will be provided in Appendix A. To prove Theorem 3.4 and Theorem 3.5, based on the triangle inequality for the 2-Wasserstein distance, we consider the following decomposition:

$$\frac{1}{N} \sum_{i=1}^N \mathcal{W}_2 \left(\text{Law} \left(x_i^{(k)} \right), \pi \right) \leq \frac{1}{N} \sum_{i=1}^N \mathcal{W}_2 \left(\text{Law} \left(x_i^{(k)} \right), \text{Law} \left(\bar{x}^{(k)} \right) \right) + \mathcal{W}_2 \left(\text{Law} \left(\bar{x}^{(k)} \right), \pi \right), \quad (3.33)$$

where

$$\mathcal{W}_2 \left(\text{Law} \left(\bar{x}^{(k)} \right), \pi \right) \leq \mathcal{W}_2 \left(\text{Law} \left(\bar{x}^{(k)} \right), \text{Law}(x_k) \right) + \mathcal{W}_2 \left(\text{Law}(x_k), \pi \right). \quad (3.34)$$

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

$$x_{k+1} = x_k - \frac{\eta}{N} \nabla f(x_k) + \sqrt{2\eta} \bar{w}^{(k+1)}, \quad (3.35)$$

which correspond to the Euler-Maruyama discretization of overdamped Langevin diffusion

$$dX_t = -\frac{1}{N} \nabla f(X_t) dt + \sqrt{2N^{-1}} dW_t, \quad (3.36)$$

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

The main idea of our proof technique is to bound the following three terms: (1) the L_2 distance between $x_i^{(k)}$ and their average $\bar{x}^{(k)}$; (2) the L_2 distance between the average iterate $\bar{x}^{(k)}$ and iterates x_k in (3.35) obtained from Euler-Maruyama discretization of overdamped diffusion (3.36); and (3) the \mathcal{W}_2 distance between the law of x_k in (3.35) and the Gibbs distribution π . First, we upper bound the L_2 distance between $x_i^{(k)}$ and their average.

3.2.1 Uniform L_2 bounds between $x_i^{(k)}$ and their average $\bar{x}^{(k)}$

In this section, we derive uniform L_2 bounds between $x_i^{(k)}$ and their average $\bar{x}^{(k)}$, where “uniform” refers to validity for all iterations k . As a first step, we derive a uniform L_2 bound for $y^{(k)}$, which is a key ingredient. First, we recall from the notations we introduced in Section 2 that

$$\tilde{x}^{(k)} = x^{(k)} - \bar{x}^{(k)}, \quad \tilde{y}^{(k)} = y^{(k)} - \bar{y}^{(k)}, \quad (3.37)$$

where we recall from (3.7)-(3.8) that $x^{(k)}, y^{(k)}$ satisfy the iterates:

$$x^{(k+1)} = \mathcal{W}^{(k)} x^{(k)} - \eta y^{(k)} + \sqrt{2\eta} w^{(k+1)}, \quad (3.38)$$

$$y^{(k+1)} = \mathcal{W}^{(k)} y^{(k)} + \nabla F \left(x^{(k+1)} \right) - \nabla F \left(x^{(k)} \right) + \xi^{(k+2)} - \xi^{(k+1)}. \quad (3.39)$$

Lemma 3.8. *Let $\alpha, \beta > 0$ and $\lambda \in (\delta^{1/B}, 1)$ be given and fixed, where $\delta = \sup_{k \geq B-1} \delta(k)$ with $\delta(k)$ defined in Assumption 3.2. Assume the following conditions hold,*

$$\sqrt{1 - \frac{\eta\mu\beta}{\beta+1}} \leq \lambda < 1, \quad \text{and} \quad \eta \leq \frac{1}{(1+\alpha)L}, \quad \text{and} \quad \gamma_1 \gamma_2 \gamma_3 \gamma_4 \in (0, 1), \quad (3.40)$$

where $\gamma_1, \gamma_2, \gamma_3$, and γ_4 are defined by (3.14)–(3.15). Then, for every k ,

$$\mathbb{E} \left\| y^{(k)} \right\|^2 \leq D^2, \quad (3.41)$$

where D is as in (3.20).

The proof of Lemma 3.8, which is deferred to Appendix A.2, relies on a sequence of technical lemmas, that we will introduce next. For any $k = 0, 1, 2, \dots$, we define:

$$q^{(k)} := x^{(k)} - \mathbf{x}_*, \quad z^{(k)} := \nabla F \left(x^{(k)} \right) - \nabla F \left(x^{(k-1)} \right), \quad (3.42)$$

where $\mathbf{x}_* = [x_*^\top, x_*^\top, \dots, x_*^\top]^\top$. Inspired by [42], we introduce the weighted L_2 norms

$$\|q\|_{L_2}^{\lambda, K} := \max_{0, 1, \dots, K} \frac{1}{\lambda^k} \left(\mathbb{E} \left\| q^{(k)} \right\|^2 \right)^{1/2}, \quad \|z\|_{L_2}^{\lambda, K} := \max_{0, 1, \dots, K} \frac{1}{\lambda^k} \left(\mathbb{E} \left\| z^{(k)} \right\|^2 \right)^{1/2}, \quad (3.43)$$

and the following loop

$$q \rightarrow z \rightarrow \tilde{y} \rightarrow \tilde{x} \rightarrow q, \quad (3.44)$$

as a proof technique. Here, each arrow means that the (weighted) L_2 norm of the sequence at the head of the arrow, can be controlled by the (weighted) L_2 norm of the sequence at the tail of the arrow; we will explain below what exactly we mean by this. For example, the arrow $q \rightarrow z$ in (3.44) means that we would like to establish an upper bound on $\|z\|_{L_2}^{\lambda, K}$ using $\|q\|_{L_2}^{\lambda, K}$. As we shall discuss next, treating each arrow separately, will allow us to complete the loop and control the boundedness of the $\tilde{y} = \{\tilde{y}_k\}_{k \geq 0}$ sequence, and this in return will allow us to ensure the boundedness of the y sequence in L_2 . Let us first study the first arrow $q \rightarrow z$ in (3.44). We have the following technical lemma.

Lemma 3.9. *For any $K = 0, 1, 2, \dots$, and $\lambda \in (0, 1)$, we have*

$$\|z\|_{L_2}^{\lambda, K} \leq L \left(1 + \frac{1}{\lambda} \right) \|q\|_{L_2}^{\lambda, K}. \quad (3.45)$$

Proof. The proof is deferred to Appendix A.3. \square

Next, let us consider the second arrow $z \rightarrow \tilde{y}$ in (3.44). Recall from (3.8) and (3.42) that

$$y^{(k+1)} = \mathcal{W}^{(k)} y^{(k)} + z^{(k+1)} + \xi^{(k+2)} - \xi^{(k+1)}. \quad (3.46)$$

In a similar manner as before, we define:

$$\|\tilde{y}\|_{L_2}^{\lambda, K} := \max_{0, 1, \dots, K} \frac{1}{\lambda^k} \left(\mathbb{E} \left\| \tilde{y}^{(k)} \right\|^2 \right)^{1/2}. \quad (3.47)$$

We now consider the second arrow $z \rightarrow \tilde{y}$ in (3.44) where we will establish an upper bound on $\|\tilde{y}\|_{L_2}^{\lambda, K}$ using $\|z\|_{L_2}^{\lambda, K}$ and $\|\tilde{y}^{(t-1)}\|_{L_2}$ for $t = 1, 2, \dots, B$. We have the following technical lemma.

Lemma 3.10. *Let $\delta := \sup_{k \geq B-1} \delta(k)$, where $\delta(k)$ is defined in Assumption 3.2. Let λ be such that $\delta < \lambda^B < 1$. Then for any $K = 0, 1, 2, \dots$, we have*

$$\|\tilde{y}\|_{L_2}^{\lambda, K} \leq \frac{\lambda(1 - \lambda^B)}{(\lambda^B - \delta)(1 - \lambda)} \|z\|_{L_2}^{\lambda, K} + \frac{\lambda^B}{\lambda^B - \delta} \frac{2B\sigma\sqrt{N}}{\lambda^K} + \frac{\lambda^B}{\lambda^B - \delta} \sum_{t=1}^B \lambda^{1-t} \|\tilde{y}^{(t-1)}\|_{L_2}. \quad (3.48)$$

Proof. The proof is deferred to Appendix A.4. \square

Next, we consider the third arrow $\tilde{y} \rightarrow \tilde{x}$ in (3.44), where our aim will be to obtain an upper bound on $\|\tilde{x}\|_{L_2}^{\lambda,K}$ using $\|\tilde{y}\|_{L_2}^{\lambda,K}$ and $\|\tilde{x}^{(t-1)}\|_{L_2}$ for $t = 1, 2, \dots, B$. We have the following technical lemma.

Lemma 3.11. *Let $\delta := \sup_{k \geq B-1} \delta(k)$, where $\delta(k)$ is defined in Assumption 3.2. Let λ be such that $\delta < \lambda^B < 1$. Then for any $K = 0, 1, 2, \dots$, we have*

$$\|\tilde{x}\|_{L_2}^{\lambda,K} \leq \frac{\eta(1-\lambda^B)}{(\lambda^B-\delta)(1-\lambda)} \|\tilde{y}\|_{L_2}^{\lambda,K} + \frac{\lambda^B}{\lambda^B-\delta} \frac{B\sqrt{2\eta Nd}}{\lambda^K} + \frac{\lambda^B}{\lambda^B-\delta} \sum_{t=1}^B \lambda^{1-t} \|\tilde{x}^{(t-1)}\|_{L_2}. \quad (3.49)$$

Proof. The proof is given in Appendix A.5. \square

Finally, let us consider the last arrow $\tilde{x} \rightarrow q$ in (3.44), for which we would like to establish an upper bound on $\|q\|_{L_2}^{\lambda,K}$ by using $\|\tilde{x}\|_{L_2}^{\lambda,K}$. We have the following result.

Lemma 3.12. *Assume that the parameters $\alpha, \beta > 0$ satisfy*

$$\sqrt{1 - \frac{\eta\mu\beta}{\beta+1}} \leq \lambda < 1, \quad \text{and} \quad \eta \leq \frac{1}{(1+\alpha)L}. \quad (3.50)$$

Then, for every $K = 0, 1, 2, \dots$, we have

$$\begin{aligned} \|q\|_{L_2}^{\lambda,K} &\leq 2\sqrt{N} \|\tilde{x}^{(0)} - x_*\| + \left(1 + \frac{\sqrt{N}}{\lambda} \left(\sqrt{\frac{L(1+\alpha)}{\mu\alpha}} + \beta\right)\right) \|\tilde{x}\|_{L_2}^{\lambda,K} \\ &\quad + \frac{\sqrt{N}}{\lambda} \left(\sqrt{\frac{L(1+\alpha)}{\mu\alpha}} + \beta\right) \frac{1}{\mu} \left(\sigma + \sqrt{\frac{2d}{\eta}}\right) \frac{1}{\lambda^K}. \end{aligned}$$

Proof. The proof is given in Appendix A.6. \square

It follows from Lemma 3.9, Lemma 3.10, Lemma 3.11 and Lemma 3.12 that

$$\|\tilde{y}\|_{L_2}^{\lambda,K} \leq \gamma_1 \|z\|_{L_2}^{\lambda,K} + \omega_1(K), \quad (3.51)$$

$$\|z\|_{L_2}^{\lambda,K} \leq \gamma_2 \|q\|_{L_2}^{\lambda,K} + \omega_2(K), \quad (3.52)$$

$$\|q\|_{L_2}^{\lambda,K} \leq \gamma_3 \|\tilde{x}\|_{L_2}^{\lambda,K} + \omega_3(K), \quad (3.53)$$

$$\|\tilde{x}\|_{L_2}^{\lambda,K} \leq \gamma_4 \|\tilde{y}\|_{L_2}^{\lambda,K} + \omega_4(K), \quad (3.54)$$

where $\gamma_1, \gamma_2, \gamma_3, \gamma_4$ are defined in (3.14)–(3.15) and

$$\omega_1(K) := \frac{\lambda^B}{\lambda^B-\delta} \frac{2B\sigma\sqrt{N}}{\lambda^K} + \frac{\lambda^B}{\lambda^B-\delta} \sum_{t=1}^B \lambda^{1-t} \|\tilde{y}^{(t-1)}\|_{L_2}, \quad \omega_2(K) := 0, \quad (3.55)$$

$$\omega_3(K) := 2\sqrt{N} \|\tilde{x}^{(0)} - x_*\| + \frac{\sqrt{N}}{\lambda} \left(\sqrt{\frac{L(1+\alpha)}{\mu\alpha}} + \beta\right) \frac{1}{\mu} \left(\sigma + \sqrt{\frac{2d}{\eta}}\right) \frac{1}{\lambda^K}, \quad (3.56)$$

$$\omega_4(K) := \frac{\lambda^B}{\lambda^B-\delta} \frac{B\sqrt{2\eta Nd}}{\lambda^K} + \frac{\lambda^B}{\lambda^B-\delta} \sum_{t=1}^B \lambda^{1-t} \|\tilde{x}^{(t-1)}\|_{L_2}. \quad (3.57)$$

As an immediate consequence of (3.51)–(3.54), we obtain the following technical lemma.

Lemma 3.13. Assume that $\delta < \lambda^B < 1$ with $\delta(k)$ defined in Assumption 3.2 and (3.40) holds. Then, for every $K = 0, 1, 2, \dots$,

$$\|\tilde{y}\|_{L_2}^{\lambda, K} \leq \frac{\gamma_1 \gamma_2 \gamma_3 \omega_4(K) + \gamma_1 \gamma_2 \omega_3(K) + \gamma_1 \omega_2(K) + \omega_1(K)}{1 - \gamma_1 \gamma_2 \gamma_3 \gamma_4}, \quad (3.58)$$

$$\|q\|_{L_2}^{\lambda, K} \leq \frac{\gamma_3 \gamma_4 \gamma_1 \omega_2(K) + \gamma_3 \gamma_4 \omega_1(K) + \gamma_3 \omega_4(K) + \omega_3(K)}{1 - \gamma_1 \gamma_2 \gamma_3 \gamma_4}, \quad (3.59)$$

where $\gamma_1, \gamma_2, \gamma_3, \gamma_4$ are defined in (3.14)–(3.15) and $\omega_1(K), \omega_2(K), \omega_3(K), \omega_4(K)$ are defined in (3.55), (3.56) and (3.57).

Proof. The proof is provided in Appendix A.7. \square

Next, we present a technical lemma that upper bounds the averaged L_2 distance between the iterates $x_i^{(k)}$ and the average $\bar{x}^{(k)}$.

Lemma 3.14. Assume that $\delta < \lambda^B < 1$ with $\delta(k)$ defined in Assumption 3.2 and (3.40) holds. Then, for any $k \geq 1$, we have

$$\sum_{i=1}^N \mathbb{E} \|x_i^{(k)} - \bar{x}^{(k)}\|^2 \leq 3 \left(\bar{\gamma}_k^{(k-1)} \right)^2 \mathbb{E} \|x^{(0)}\|^2 + 3D^2 \eta^2 \left(\sum_{s=0}^{k-1} \bar{\gamma}_{k-1-s}^{(k-1)} \right)^2 + 6dN\eta \sum_{s=0}^{k-1} \left(\bar{\gamma}_{k-1-s}^{(k-1)} \right)^2,$$

where D is defined in (3.20) and

$$\bar{\gamma}_j^{(k-1)} := \left\| W_j^{(k-1)} - \frac{1}{N} 1_N 1_N^\top \right\| \quad \text{for } j = 0, 1, \dots, k-1. \quad (3.60)$$

Proof. The proof is given in Appendix A.8. \square

Next, we aim to provide an upper bound for $\bar{\gamma}_{k-1-s}^{(k)}$ in Lemma 3.14 under Assumption 3.2 for the mixing matrices $W^{(k)}$. The following corollary of Lemma 3.14 establishes this and shows that the iterates $x_i^{(k)}$ are close to the average $\bar{x}^{(k)}$ on average.

Lemma 3.15. Assume that $\delta < \lambda^B < 1$ with $\delta(k)$ defined in Assumption 3.2 and (3.40) holds. Then, for any k , we have

$$\sum_{i=1}^N \mathbb{E} \|x_i^{(k)} - \bar{x}^{(k)}\|^2 \leq 3 \cdot \delta^{-2} \left(\delta^{\frac{2}{B}} \right)^k \mathbb{E} \|x^{(0)}\|^2 + \frac{3D^2 \eta^2 \delta^{-2}}{(1 - \delta^{\frac{1}{B}})^2} + \frac{6dN\eta \cdot \delta^{-2}}{1 - \delta^{\frac{2}{B}}}, \quad (3.61)$$

where D is defined in (3.20).

Proof. The proof is provided in Appendix A.9. \square

3.2.2 L_2 distance between $\bar{x}^{(k)}$ and x_k

In this section, we derive bounds on the L_2 distance between $\bar{x}^{(k)}$ and x_k , which is the k -th iterate of the Euler discretization of an overdamped Langevin diffusion given in (3.35).

First, by taking the average of N nodes in (3.4)–(3.5), and using the fact that $W^{(k)}$ is doubly stochastic, we obtain:

$$\bar{x}^{(k+1)} = \bar{x}^{(k)} - \eta \bar{y}^{(k)} + \sqrt{2\eta} \bar{w}^{(k+1)}, \quad (3.62)$$

where for any $k = 0, 1, 2, \dots$,

$$\bar{y}^{(k+1)} = \bar{y}^{(k)} + \frac{1}{N} \sum_{i=1}^N \nabla f_i(x_i^{(k+1)}) - \frac{1}{N} \sum_{i=1}^N \nabla f_i(x_i^{(k)}) + \bar{\xi}^{(k+2)} - \bar{\xi}^{(k+1)}, \quad (3.63)$$

which implies that for any $k = 0, 1, 2, \dots$,

$$\bar{y}^{(k)} = \frac{1}{N} \sum_{i=1}^N \nabla f_i(x_i^{(k)}) + \bar{\xi}^{(k+1)}. \quad (3.64)$$

Therefore, we have

$$\bar{x}^{(k+1)} = \bar{x}^{(k)} - \eta \frac{1}{N} \sum_{i=1}^N \nabla f_i(x_i^{(k)}) - \eta \bar{\xi}^{(k+1)} + \sqrt{2\eta} \bar{w}^{(k+1)}, \quad (3.65)$$

which can be re-written as

$$\bar{x}^{(k+1)} = \bar{x}^{(k)} - \eta \frac{1}{N} \nabla f(\bar{x}^{(k)}) + \eta \mathcal{E}_k - \eta \bar{\xi}^{(k+1)} + \sqrt{2\eta} \bar{w}^{(k+1)}, \quad (3.66)$$

where

$$\mathcal{E}_k := \frac{1}{N} \sum_{i=1}^N \left[\nabla f_i(\bar{x}^{(k)}) - \nabla f_i(x_i^{(k)}) \right]. \quad (3.67)$$

In the next lemma, we provide an explicit upper bound on the L_2 norm of the error term \mathcal{E}_k .

Lemma 3.16. *Assume that $\delta < \lambda^B < 1$ with $\delta(k)$ defined in Assumption 3.2 and (3.40) holds. Then, for any k , we have*

$$\mathbb{E} \|\mathcal{E}_k\|^2 \leq \frac{3L^2 \delta^{-2}}{N} \left(\delta^{\frac{2}{B}} \right)^k \mathbb{E} \|x^{(0)}\|^2 + \frac{3L^2 D^2 \eta^2 \delta^{-2}}{N(1 - \delta^{\frac{1}{B}})^2} + \frac{6dL^2 \eta \cdot \delta^{-2}}{1 - \delta^{\frac{2}{B}}}.$$

Proof. The proof is given in Appendix A.10. □

Next, we recall from (3.35) that the iterates x_k are given by:

$$x_{k+1} = x_k - \eta \frac{1}{N} \nabla f(x_k) + \sqrt{2\eta} \bar{w}^{(k+1)}, \quad (3.68)$$

where we take $x_0 = \bar{x}_0 = \frac{1}{N} \sum_{i=1}^N x_i^{(0)}$. This is a Euler-Maruyama discretization (with stepsize η) of the continuous-time overdamped Langevin diffusion (3.36). Since the L_2 bound of the error term \mathcal{E}_k can be controlled as in Lemma 3.16, we will show that the average $\bar{x}^{(k)}$ and x_k are close to each other in L_2 distance. Indeed, we have the following estimate.

Lemma 3.17. *Assume that $\delta < \lambda^B < 1$ with $\delta(k)$ defined in Assumption 3.2 and (3.40) holds. We also assume $\mathbb{E} \|x^{(0)}\|^2 < \infty$. For any stepsize $\eta \in (0, 2/L)$, we have for every k ,*

$$\begin{aligned} \mathbb{E} \|\bar{x}^{(k)} - x_k\|^2 &\leq \eta \left(\frac{\eta}{\mu(1 - \frac{\eta L}{2})} + \frac{(1 + \eta L)^2}{\mu^2(1 - \frac{\eta L}{2})^2} \right) \left(\frac{3L^2 D^2 \eta \delta^{-2}}{N(1 - \delta^{\frac{1}{B}})^2} + \frac{6dL^2 \cdot \delta^{-2}}{1 - \delta^{\frac{2}{B}}} \right) + \frac{\eta \sigma^2}{\mu(1 - \frac{\eta L}{2})N} \\ &\quad + \frac{\left(\delta^{\frac{2}{B}} \right)^k - \left(1 - \eta \mu \left(1 - \frac{\eta L}{2} \right) \right)^k}{\delta^{\frac{2}{B}} - 1 + \eta \mu \left(1 - \frac{\eta L}{2} \right)} \frac{3L^2 \delta^{-2}}{N} \delta^{\frac{2}{B}} \mathbb{E} \|x^{(0)}\|^2. \end{aligned}$$

Proof. The proof is provided in Appendix A.11. □

3.2.3 \mathcal{W}_2 distance between the law of x_k and the Gibbs distribution π

The \mathcal{W}_2 distance between the Euler-Mariyama discretization x_k in (3.35) of the overdamped Langevin diffusion (3.36) and the Gibbs distribution $\pi \propto e^{-f}$ has been established in the literature. 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 [20] as follows.

Lemma 3.18 (Theorem 4 in [20]). *For any $\eta \leq \frac{2N}{\mu+L}$, we have*

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

Now, we are finally ready to prove Theorem 3.4 and Theorem 3.5.

3.2.4 Completing the Proofs of Theorem 3.4 and Theorem 3.5

Proof of Theorem 3.4. The L_2 distance between the minimizer of f and Gibbs distribution π has been studied in the literature; see e.g. [30]. More precisely, we have

$$\mathbb{E}_{X \sim \pi} \|X - x_*\|^2 \leq \frac{2dN^{-1}}{\mu}, \quad (3.69)$$

where x_* is the unique minimizer of $f(x)$; see Lemma 11 in [30]. Since $x_0 = \frac{1}{N} \sum_{i=1}^N x_i^{(0)}$, we have $\mathbb{E} \|x_0\|^2 < \infty$. By (3.69), we get

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

It then follows from Lemma 3.18 that for any $\eta \leq \frac{2N}{\mu+L}$, we have

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

Moreover, it follows from Lemma 3.17 that

$$\begin{aligned} \mathcal{W}_2(\text{Law}(\bar{x}^{(k)}), \text{Law}(x_k)) &\leq \left(\mathbb{E} \|\bar{x}^{(k)} - x_k\|^2 \right)^{1/2} \\ &\leq \eta^{1/2} \left(\frac{\eta}{\mu(1 - \frac{\eta L}{2})} + \frac{(1 + \eta L)^2}{\mu^2(1 - \frac{\eta L}{2})^2} \right)^{1/2} \cdot \left(\frac{3L^2 D^2 \eta \delta^{-2}}{N(1 - \delta^{\frac{1}{B}})^2} + \frac{6dL^2 \cdot \delta^{-2}}{1 - \delta^{\frac{2}{B}}} \right)^{1/2} \\ &\quad + \frac{\sqrt{\eta}\sigma}{\sqrt{\mu(1 - \frac{\eta L}{2})N}} + \left(\frac{\left(\delta^{\frac{2}{B}} \right)^k - \left(1 - \eta\mu \left(1 - \frac{\eta L}{2} \right) \right)^k}{\delta^{\frac{2}{B}} - 1 + \eta\mu \left(1 - \frac{\eta L}{2} \right)} \right)^{1/2} \frac{\sqrt{3}L \cdot \delta^{-1}}{\sqrt{N}} \delta^{\frac{1}{B}} \left(\mathbb{E} \|x^{(0)}\|^2 \right)^{1/2}. \end{aligned}$$

The result then follows from the triangular inequality for the 2-Wasserstein distance. The proof is complete. \square

Proof of Theorem 3.5. By the Cauchy-Schwarz inequality,

$$\frac{1}{N} \sum_{i=1}^N \mathcal{W}_2(\text{Law}(x_i^{(k)}), \text{Law}(\bar{x}^{(k)})) \leq \sqrt{\frac{1}{N} \sum_{i=1}^N \mathcal{W}_2^2(\text{Law}(x_i^{(k)}), \text{Law}(\bar{x}^{(k)}))}$$

$$\leq \sqrt{\frac{1}{N} \sum_{i=1}^N \mathbb{E} \|x_i^{(k)} - \bar{x}^{(k)}\|^2}. \quad (3.70)$$

By Lemma 3.15, we have

$$\begin{aligned} \sqrt{\frac{1}{N} \sum_{i=1}^N \mathbb{E} \|x_i^{(k)} - \bar{x}^{(k)}\|^2} &\leq \left(\frac{3 \cdot \delta^{-2} \left(\delta^{\frac{2}{B}}\right)^k}{N} \mathbb{E} \|x^{(0)}\|^2 + \frac{3D^2\eta^2\delta^{-2}}{N(1 - \delta^{\frac{1}{B}})^2} + \frac{6d\eta \cdot \delta^{-2}}{1 - \delta^{\frac{2}{B}}} \right)^{1/2} \\ &\leq \frac{\sqrt{3}\delta^{-1}\delta^{\frac{k}{B}}}{\sqrt{N}} \left(\mathbb{E} \|x^{(0)}\|^2 \right)^{1/2} + \frac{\sqrt{3}D\eta\delta^{-1}}{\sqrt{N}(1 - \delta^{\frac{1}{B}})} + \frac{\sqrt{6d\eta}\delta^{-1}}{\sqrt{1 - \delta^{\frac{2}{B}}}}. \end{aligned}$$

The result then follows from Theorem 3.4 and the triangular inequality for the 2-Wasserstein distance. The proof is complete. \square

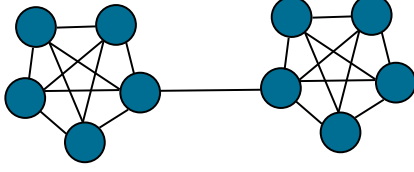
4 Numerical Experiments

In this section, we present numerical experiments evaluating the sampling performance of DIGing-SGLD, whose iterates are given in (3.2)–(3.3), in comparison to the DE-SGLD iterations described in [30]. All experiments are conducted under a constant step-size and over time-varying network topologies. We evaluate both methods on Bayesian linear and logistic regression using synthetic datasets, and additionally on a real-world dataset [62] for Bayesian logistic regression.

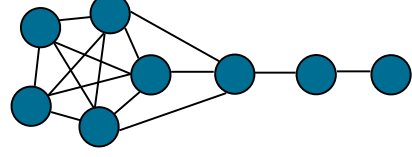
For Bayesian linear regression, we examine how closely the samples generated by each agent approximate the target posterior distribution in the \mathcal{W}_2 distance, showing that each agent converges to a distribution lying within an ϵ -neighborhood of the true posterior. For Bayesian logistic regression, we assess convergence in terms of classification accuracy on a held-out test set. Across both problems, we demonstrate that incorporating the DIGing-based gradient-tracking mechanism [42] in DIGing-SGLD effectively corrects the network-induced drift observed in DE-SGLD over time-varying networks. This advantage is particularly pronounced for Bayesian logistic regression in classification accuracy, although the \mathcal{W}_2 metric for Bayesian linear regression also exhibits a consistent performance gap between DIGing-SGLD and DE-SGLD.

For the experimental setup, we consider two classes of undirected time-varying networks composed of N agents. The first class is the barbell graph consisting of two cliques connected by a single edge, and the second is a generalized lollipop graph $L_{m,r}^s$ comprising a clique K_m with m nodes and a path P_r of length r joined by s distinct edges to the path's terminal node. Graphs such as the barbell and lollipop graphs and their generalizations are widely used in the literature to evaluate decentralized algorithms, as they represent near-worst-case scenarios for information propagation over a network [2, 13, 33]. For the barbell topology, we use $N \in \{20, 30\}$ depending on the problem. The time-varying nature is introduced by first constructing two undirected complete graphs, each with $N/2$ agents. At every algorithmic iteration k , two random agents, one from each complete graph, are connected to form a single connected network. Figure 1a illustrates an example of this topology for $N = 10$.

For the time-varying generalized lollipop network, we use $N = 20$ agents. At each iteration k , we sample $N' \sim \mathcal{U}[3, 4]$ from the discrete uniform distribution to set the length of the lollipop's branch (the path subgraph), while the remaining $N - N'$ agents form a clique. Three random agents from this complete subgraph are then connected to the branch through its terminal node, i.e., the agent with the smallest index. This corresponds to the generalized lollipop graph $L_{m,r}^s$



(a) Example of a barbell topology for $N = 10$ agents.



(b) Example of a generalized lollipop topology for $N = 8$ agents.

Figure 1: Illustrations of the two undirected time-varying network structures used in our experiments.

with $m = N - N'$, $r = N'$ and $s = 3$. Figure 1b shows an example of this topology for $N = 8$, where $N' = 3$ and the remaining $N - N' = 5$ agents form the clique.

After generating the network topology in each iteration for both classes of time-varying networks, we construct the corresponding weight matrix $W^{(k)}$ using the Metropolis constant edge-weight rule [8, 63], defined as

$$W_{ij}^{(k)} = \begin{cases} \frac{1}{\max\{\deg_i^{(k)}, \deg_j^{(k)}\} + \hat{\varepsilon}}, & \text{if } (i, j) \in \mathcal{E}(k), \\ 0, & \text{if } (i, j) \notin \mathcal{E}(k) \text{ and } i \neq j, \\ 1 - \sum_{\ell \in \Omega_i(k)} W_{i\ell}^{(k)}, & \text{if } i = j, \end{cases} \quad (4.1)$$

where $\deg_i^{(k)}$ denotes the degree of agent i at iteration k , and $\hat{\varepsilon} > 0$ (set to 10^{-6} in our experiments) ensures that the Markov chain represented by $W^{(k)}$ is aperiodic. In all experiments, the time-varying networks are generated such that each topology repeats every 50 algorithmic iterations in a cyclic manner. While this repetition is an approximation, it can be interpreted as controlling the parameter B in our theoretical model, with larger repetition intervals corresponding to higher values of B .

4.1 Bayesian Linear Regression

We now present the results of DIGing-SGLD and DE-SGLD for Bayesian linear regression over the time-varying barbell and lollipop topologies. In linear regression, each data sample consists of a response variable $y \in \mathbb{R}$ and a feature vector $\hat{z} \in \mathbb{R}^d$. To include an intercept term, we use the augmented feature vector $z = [\hat{z}^\top \ 1]^\top \in \mathbb{R}^{d+1}$. In our experiments, we generate $n = 100$ samples $\{(y_i, z_i)\}_{i=1}^n$ with $d = 5$, and distribute them uniformly across $N = 20$ agents, so that each agent holds $\bar{n} := n/N = 5$ samples.

To enable reporting of performance in terms of the \mathcal{W}_2 distance, we work with synthetic data generated under a Gaussian linear model. The samples are generated according to $y_i = z_i^\top x + \delta_i$, where the true parameter $x \in \mathbb{R}^{d+1}$ is drawn once from $x \sim \mathcal{N}(0, \lambda^{-1} I_{d+1})$ with $\lambda = 0.1$ and then held fixed. The noise variables satisfy $\delta_i \sim \mathcal{N}(0, \sigma^2)$ with $\sigma^2 = 1$, and the underlying feature vectors are drawn independently as $\hat{z}_i \sim \mathcal{N}(0, I_d)$. Under this model, the posterior distribution for x given all n samples is $\pi(x) \propto \exp(-\sum_{j=1}^n f_j(x))$, where the local function at agent j is the strongly convex and smooth function

$$f_j(x) := \frac{\|Z_j x - y_j\|^2}{2} + \frac{\lambda \|x\|^2}{2N}, \quad (4.2)$$

with $Z_j \in \mathbb{R}^{\bar{n} \times (d+1)}$ denoting the matrix of local feature vectors and $y_j \in \mathbb{R}^{\bar{n}}$ the corresponding responses. The posterior in this case also admits a closed-form expression, given by $\pi(x) = \mathcal{N}(\hat{m}, \hat{\Sigma})$, where $\hat{\Sigma} = (Z^\top Z + \lambda I_{d+1})^{-1}$ and $\hat{m} = \hat{\Sigma} Z^\top y$ for the full dataset matrix $Z \in \mathbb{R}^{n \times (d+1)}$ and full response vector $y \in \mathbb{R}^n$.

We conduct three experiments using this synthetic-data construction for both DIGing-SGLD and DE-SGLD. In the first experiment (Figure 2a), each agent uses its full batch of $b = \bar{n} = 5$ samples to compute the gradient of (4.2) at every iteration, and the DIGing-SGLD updates are run over a time-varying barbell network. The second experiment (Figure 2b) uses the same data-generation procedure and barbell topology, but each agent now samples a mini-batch of $b = 3$ out of its $\bar{n} = 5$ local samples at each iteration. Since this results in a stochastic gradient of $f_j(x)$, and we require this gradient to be an unbiased estimate of the local gradient, we scale the mini-batch gradient by the factor \bar{n}/b . In the third experiment (Figure 2c), we evaluate DIGing-SGLD and DE-SGLD over a time-varying lollipop network, with each agent again using its full batch of $b = \bar{n} = 5$ samples.

In all three experiments, both methods are run independently for 200 trials with different initializations. These independent runs are used to estimate the mean vector $m_j^{(k)}$ and covariance matrix $\Sigma_j^{(k)}$ of the approximate posterior at each agent j and iteration k . Using the closed-form expression for the \mathcal{W}_2 distance between Gaussian distributions [29], we compute the gap between the empirical and true posteriors via

$$\mathcal{W}_{2,j}^{(k)} = \left(\left\| \hat{m} - m_j^{(k)} \right\|^2 + \text{Tr} \left(\Sigma_j^{(k)} \right) + \text{Tr} \left(\hat{\Sigma} \right) - 2 \text{Tr} \left(\left(\Sigma_j^{(k)} \right)^{1/2} \hat{\Sigma}^{1/2} \left(\Sigma_j^{(k)} \right)^{1/2} \right)^{1/2} \right)^{1/2}.$$

Figure 2 reports the results across the three experimental setups. Each plot shows, for both DIGing-SGLD and DE-SGLD, the average Wasserstein distance over all agents, with one standard deviation of the agent-wise distances indicated as a shaded region. In each case, we fix the number of iterations to 100 and tune the step-size to minimize the \mathcal{W}_2 error at iteration 100.

The results illustrate that DIGing-SGLD outperforms DE-SGLD across both network structures. Both algorithms converge more slowly on graphs with small spectral gaps, such as the barbell topology, as expected, and using full local batches leads to a faster reduction in \mathcal{W}_2 distance compared to mini-batch updates. Most importantly, in all scenarios, DIGing-SGLD demonstrates a persistent advantage over DE-SGLD, confirming the benefit of incorporating DIGing-based gradient tracking in dynamic network environments. We also note that DE-SGLD lacks convergence guarantees for time-varying networks, further underscoring the importance of the guarantees established here for DIGing-SGLD.

4.2 Bayesian Logistic Regression

We next evaluate DIGing-SGLD and DE-SGLD for Bayesian logistic regression under time-varying barbell and lollipop topologies, using both synthetic datasets ($N = 20$ agents) and a real-world dataset ($N = 30$ agents; barbell topology only). In Bayesian logistic regression, each data sample consists of a binary class label $y \in \{0, 1\}$ and a feature vector $\hat{z} \in \mathbb{R}^d$. As in the linear regression setting, we work with the augmented feature vector $z = [\hat{z}^\top \ 1]^\top \in \mathbb{R}^{d+1}$ to incorporate an intercept term. Given z and a parameter vector (model) $x \in \mathbb{R}^{d+1}$, the likelihood of the class label is

$$\mathbb{P}(y = 1 \mid z, x) = \frac{1}{1 + \exp(-z^\top x)} = \sigma(z^\top x),$$

and we adopt the Gaussian prior for the model x : $p(x) = \mathcal{N}(0, \lambda^{-1} I_{d+1})$ with $\lambda > 0$.

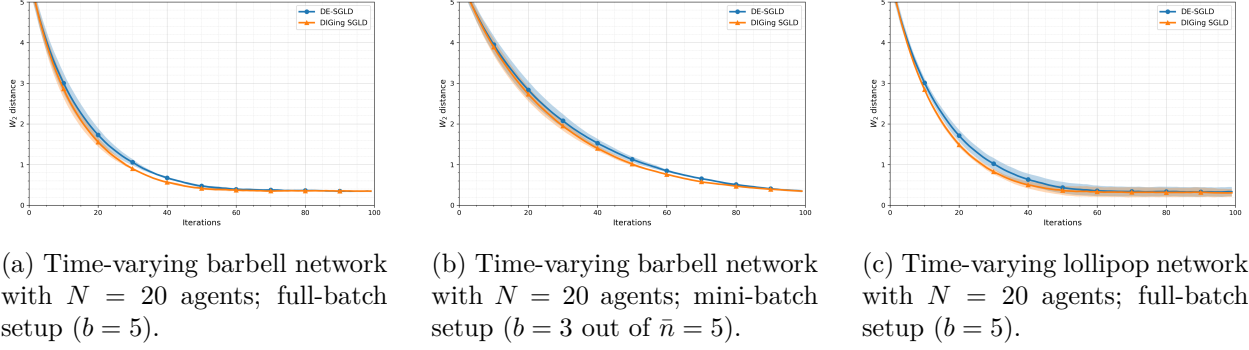


Figure 2: Comparison of DIGing-SGLD and DE-SGLD for Bayesian linear regression on synthetic data under time-varying barbell and lollipop network structures. Each plot displays the average Wasserstein distance across agents, with one standard deviation shown as a shaded region.

Given n labeled training samples $\{(y_i, z_i)\}_{i=1}^n$, the posterior of x is of the form $\pi(x) \propto \exp(-f(x))$. When the data are evenly distributed across N agents, so that each agent holds $\bar{n} = n/N$ samples, the global function $f(x)$ decomposes as $f(x) = \sum_{j=1}^N f_j(x)$, where each local function f_j is smooth and strongly convex and is given by

$$f_j(x) = \sum_{i=1}^{\bar{n}} \left[-\log(1 - \sigma(z_{i,j}^\top x)) + y_{i,j} \log\left(\frac{1 - \sigma(z_{i,j}^\top x)}{\sigma(z_{i,j}^\top x)}\right) \right] + \frac{\lambda}{2N} \|x\|^2, \quad (4.3)$$

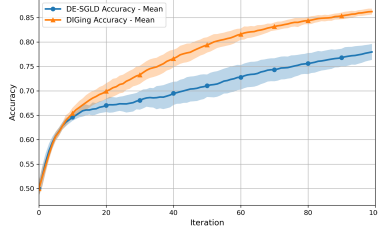
with $(z_{i,j}, y_{i,j})$ denoting the i th sample at agent j .

Because the posterior $\pi(x)$ does not admit a closed-form expression in this model, in contrast to the Bayesian linear regression setting, we compare DIGing-SGLD and DE-SGLD using classification accuracy on a separate test dataset, defined as the proportion of correctly predicted labels. As in the linear regression experiments, the accuracy at each agent and iteration is computed by averaging the per-agent accuracy over multiple independent runs, and step-sizes are hand-tuned to optimize accuracy at the final iteration (here, iteration 100).

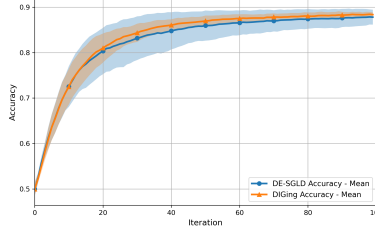
4.2.1 Synthetic Data Experiments

For the synthetic experiments, the parameter $x \in \mathbb{R}^{d+1}$ is drawn once from the prior $p(x) = \mathcal{N}(0, \lambda^{-1} I_{d+1})$ with $\lambda = 0.1$, and feature vectors are generated by sampling $\hat{z}_i \sim \mathcal{N}(0, I_d)$. Each class label is assigned by drawing $p_i \sim \mathcal{U}(0, 1)$ and setting $y_i = 1$ if $p_i \leq \sigma(\hat{z}_i^\top x)$ and $y_i = 0$ otherwise. We set $d = 5$ and generate 600 samples, which are then split into training and test sets using a 70–30 train–test ratio. This yields 420 training samples, and we discard 40 to obtain a class-balanced training set of size $n = 380$. These samples are distributed evenly across $N = 20$ agents, giving each agent $\bar{n} = 19$ samples.

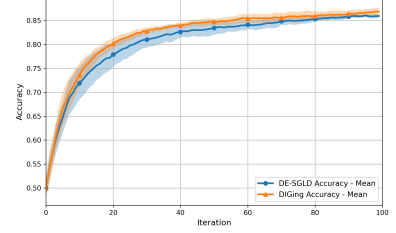
We conduct two experiments under this setup, one using a time-varying barbell topology and one using a time-varying lollipop topology. In both experiments, at each iteration, every agent uses a mini-batch of $b = 1$ sample to compute a stochastic gradient of the logistic loss. Since this yields a stochastic gradient of $f_j(x)$, we scale the gradient by $\bar{n}/b = 19$ to obtain an unbiased local gradient estimator. Each experiment is repeated independently 200 times; accuracy is evaluated for each estimate $x_j^{(k)}$ at each agent j , and the plots in Figure 3 report the average accuracy computed across both agents and independent trials, with one standard deviation across agents shown as a shaded region in the corresponding plots.



(a) Time-varying barbell network ($N = 20$); synthetic dataset; mini-batch setup ($b = 1$ out of $\bar{n} = 19$).



(b) Time-varying lollipop network ($N = 20$); synthetic dataset; mini-batch setup ($b = 1$ out of $\bar{n} = 19$).



(c) Time-varying barbell network ($N = 30$); real dataset; full-batch setup ($b = \bar{n} = 1$).

Figure 3: Performance comparison of DIGing-SGLD and DE-SGLD for Bayesian logistic regression under time-varying barbell and lollipop network structures. Each plot displays the average classification accuracy across agents and independent trials, with one standard deviation across agents shown as a shaded region.

The first synthetic experiment uses the time-varying barbell topology (Figure 3a); the second uses the time-varying lollipop topology (Figure 3b). In both cases, DIGing-SGLD consistently outperforms DE-SGLD, achieving higher accuracy throughout the training horizon. The performance gap is more pronounced for the barbell topology, which has a smaller spectral gap and therefore induces a more severe information-flow bottleneck than the lollipop topology. These results parallel our findings for Bayesian linear regression: DIGing-SGLD maintains stable performance under time variations, whereas DE-SGLD exhibits larger variability and reduced accuracy. We again note that DE-SGLD does not admit any convergence guarantees for time-varying networks, and its weaker empirical performance in these experiments further highlights the robustness of DIGing-SGLD in dynamic network environments.

4.2.2 Real Data Experiments

We finally compare DIGing-SGLD and DE-SGLD on Bayesian logistic regression using real data under a time-varying barbell topology with $N = 30$ agents. We use the UCI ML Breast Cancer Wisconsin (Diagnostic) dataset [62], which contains 569 samples and $d = 30$ features derived from digitized images of fine needle aspirate biopsies. Class labels are encoded as 1 for benign and 0 for malignant. We split the dataset using a 10–90 train–test ratio, yielding 51 training samples. After standardizing and class-balancing the training data, we obtain $n = 30$ samples that are then evenly distributed across the 30 agents, so that each agent has $\bar{n} = 1$ sample. At each iteration, each agent processes its single sample ($b = \bar{n} = 1$) to compute a stochastic gradient. For a fixed iteration budget, both the step-size and the regularization parameter λ are tuned, with the final choice of the regularization parameter being $\lambda = 0.3$.

Results for this experiment are shown in Figure 3c. The curves display the average accuracy computed across both independent trials (200 repetitions) and agents, with one standard deviation across agents shown as a shaded region. Consistent with our synthetic experiments, DIGing-SGLD converges reliably in the time-varying setting, while DE-SGLD converges more slowly and, importantly, has no known convergence guarantees for time-varying networks.

5 Conclusion

In this work, we introduced *DIGing-SGLD*, a decentralized Langevin-based sampling algorithm that operates over time-varying networks. The method integrates distributed inexact gradient tracking (DIGing)—originating in decentralized optimization—into stochastic gradient Langevin dynamics, thereby removing the sampling bias and convergence degradation characteristic of vanilla decentralized SGLD methods on static graphs. Under strong convexity and smoothness assumptions on the component functions f_i and a connectivity condition of the underlying time-varying graph, we establish finite-time, non-asymptotic guarantees in the 2-Wasserstein distance with explicit constants: each agent’s marginal distribution converges geometrically to an $\mathcal{O}(\sqrt{\eta})$ neighborhood of the target distribution $\pi(x) \propto e^{-f(x)}$ with $f(x) = \sum_{i=1}^n f_i(x)$. Choosing a stepsize $\eta = \mathcal{O}(\epsilon^2)$ yields $\mathcal{O}\left(\frac{\log(1/\epsilon)}{\epsilon^2}\right)$ iteration complexity to reach ϵ -accuracy, matching the best known rates for centralized and static-graph SGLD while extending them to the more realistic setting of dynamic, decentralized (coordinator-free) networks. Numerical experiments on Bayesian linear and logistic regression corroborate the theory, demonstrating robust performance under changing topologies and inexact (stochastic) gradient updates. Looking forward, promising research directions include extending DIGing-SGLD to non-convex objectives or directed network settings, thereby strengthening its theoretical guarantees and enhancing its applicability to large-scale decentralized Bayesian inference.

A Proofs of Technical Lemmas

A.1 Proof of Lemma 3.6

Proof. We first note that by Lemma C.4, $\underline{\lambda} \leq \lambda(\eta)$ and conditions (3.16) are satisfied and that 3.5 is applicable. Second, we note that

$$D\sqrt{\eta} = \left[2 \left(\frac{\gamma_1\gamma_2\gamma_3(\tilde{\omega}_4^{(\eta)} + \hat{\omega}_4^{(\eta)}) + \gamma_1\gamma_2(\tilde{\omega}_3^{(\eta)} + \hat{\omega}_3^{(\eta)}) + \tilde{\omega}_1^{(\eta)} + \hat{\omega}_1^{(\eta)}}{1 - \gamma_1\gamma_2\gamma_3\gamma_4} \right)^2 + \frac{4L^2}{N} \left(\frac{\gamma_3\gamma_4(\tilde{\omega}_1^{(\eta)} + \hat{\omega}_1^{(\eta)}) + \gamma_3(\tilde{\omega}_4^{(\eta)} + \hat{\omega}_4^{(\eta)}) + \tilde{\omega}_3^{(\eta)} + \hat{\omega}_3^{(\eta)}}{1 - \gamma_1\gamma_2\gamma_3\gamma_4} \right)^2 + \frac{4}{N}\sigma^2\eta \right]^{1/2}, \quad (\text{A.1})$$

where

$$\hat{\omega}_i^{(\eta)} := \hat{\omega}_i\sqrt{\eta} \quad \text{and} \quad \tilde{\omega}_i^{(\eta)} := \tilde{\omega}_i\sqrt{\eta} \quad i \in \{1, 3, 4\}.$$

Since conditions (3.16) are satisfied, the denominator term $1 - \gamma_1\gamma_2\gamma_3\gamma_4 > 0$ and the quantity $D\sqrt{\eta}$ is well-defined. It is straightforward to verify from (A.1) that $D\sqrt{\eta}$ is a non-decreasing function of γ_i , $\hat{\omega}_i^{(\eta)}$ and $\tilde{\omega}_i^{(\eta)}$ for every $i \in \{1, 3, 4\}$. That is, if we replace any of these variables with their upper bounds, we can obtain an upper bound for $D\sqrt{\eta}$; which is the main proof technique we will use.

It is straightforward to check that both $\hat{\omega}_i^{(\eta)}$ and $\tilde{\omega}_i^{(\eta)}$ are non-increasing functions of λ for $\lambda \in (\delta^{1/B}, 1)$ for every $i \in \{1, 3, 4\}$, when $\eta \in (0, \bar{\eta}]$ is fixed. Similarly, both $\hat{\omega}_i^{(\eta)}$ and $\tilde{\omega}_i^{(\eta)}$ are non-decreasing functions of η , when λ is fixed for every $i \in \{1, 3, 4\}$. Therefore, by replacing $\lambda = \lambda(\eta)$ with its lower bound $\underline{\lambda}$ and by replacing η with its upper bound $\bar{\eta}$, in the definition of $\hat{\omega}_i^{(\eta)}$ and $\tilde{\omega}_i^{(\eta)}$, we obtain the bounds:

$$\tilde{\omega}_i^{(\eta)} \leq \bar{\omega}_i\sqrt{\bar{\eta}}, \quad \text{and} \quad \hat{\omega}_i^{(\eta)} \leq \bar{\omega}_i\sqrt{\bar{\eta}}. \quad (\text{A.2})$$

Similarly, with some straightforward computations, it can be seen that γ_i are all non-increasing functions of λ and non-decreasing functions of η for $i = 1, 2, 3, 4$. Therefore, we can have the following analogous bounds

$$\gamma_1 \leq \bar{\gamma}_1 = \frac{\lambda \cdot (1 - \lambda^B)}{(\lambda^B - \delta)(1 - \lambda)}, \quad \gamma_2 \leq \bar{\gamma}_2 = L \left(1 + \frac{1}{\lambda}\right), \quad (\text{A.3})$$

$$\gamma_3 \leq \bar{\gamma}_3 = \left(1 + \frac{\sqrt{N}}{\lambda} \left(\sqrt{\frac{L(1+\alpha)}{\mu\alpha}} + \beta\right)\right), \quad \gamma_4 \leq \bar{\gamma}_4 = \frac{\bar{\eta} (1 - \lambda^B)}{(\lambda^B - \delta)(1 - \lambda)}. \quad (\text{A.4})$$

Putting everything together; and replacing $\tilde{\omega}_i^{(\eta)}, \hat{\omega}_i^{(\eta)}$ and γ_i with their corresponding upper bounds in the formula (A.1) based on (A.2), (A.3) and (A.4), proves the bound (3.25). \square

A.2 Proof of Lemma 3.8

Proof. It follows from (3.58) and (3.59) that

$$\begin{aligned} \|\tilde{y}\|_{L_2}^{\lambda, K} &\leq \frac{\gamma_1 \gamma_2 \gamma_3 \omega_4(K) + \gamma_1 \gamma_2 \omega_3(K) + \gamma_1 \omega_2(K) + \omega_1(K)}{1 - \gamma_1 \gamma_2 \gamma_3 \gamma_4} \\ &= \frac{\gamma_1 \gamma_2 \gamma_3 \tilde{\omega}_4 + \gamma_1 \gamma_2 \tilde{\omega}_3 + \tilde{\omega}_1}{1 - \gamma_1 \gamma_2 \gamma_3 \gamma_4} + \frac{\gamma_1 \gamma_2 \gamma_3 \hat{\omega}_4 + \gamma_1 \gamma_2 \hat{\omega}_3 + \hat{\omega}_1}{1 - \gamma_1 \gamma_2 \gamma_3 \gamma_4} \frac{1}{\lambda^K}, \end{aligned}$$

and

$$\begin{aligned} \|q\|_{L_2}^{\lambda, K} &\leq \frac{\gamma_3 \gamma_4 \gamma_1 \omega_2(K) + \gamma_3 \gamma_4 \omega_1(K) + \gamma_3 \omega_4(K) + \omega_3(K)}{1 - \gamma_1 \gamma_2 \gamma_3 \gamma_4} \\ &= \frac{\gamma_3 \gamma_4 \tilde{\omega}_1 + \gamma_3 \tilde{\omega}_4 + \tilde{\omega}_3}{1 - \gamma_1 \gamma_2 \gamma_3 \gamma_4} + \frac{\gamma_3 \gamma_4 \hat{\omega}_1 + \gamma_3 \hat{\omega}_4 + \hat{\omega}_3}{1 - \gamma_1 \gamma_2 \gamma_3 \gamma_4} \frac{1}{\lambda^K}, \end{aligned}$$

where $\gamma_1, \gamma_2, \gamma_3, \gamma_4$ are defined in (3.14)-(3.15) and $\omega_1(K), \omega_2(K), \omega_3(K), \omega_4(K)$ are defined in (3.55), (3.56) and (3.57) and we recall from (3.11), (3.12) and (3.13) that

$$\tilde{\omega}_1 := \frac{\lambda^B}{\lambda^B - \delta} \sum_{t=1}^B \lambda^{1-t} \left\| \tilde{y}^{(t-1)} \right\|_{L_2}, \quad \hat{\omega}_1 := \frac{\lambda^B}{\lambda^B - \delta} \cdot 2B\sigma\sqrt{N}, \quad (\text{A.5})$$

$$\tilde{\omega}_3 := 2\sqrt{N} \left\| \tilde{x}^{(0)} - x_* \right\|, \quad \hat{\omega}_3 := \frac{\sqrt{N}}{\lambda} \left(\sqrt{\frac{L(1+\alpha)}{\mu\alpha}} + \beta \right) \frac{1}{\mu} \left(\sigma + \sqrt{\frac{2d}{\eta}} \right), \quad (\text{A.6})$$

$$\tilde{\omega}_4 := \frac{\lambda^B}{\lambda^B - \delta} \sum_{t=1}^B \lambda^{1-t} \left\| \tilde{x}^{(t-1)} \right\|_{L_2}, \quad \hat{\omega}_4 := \frac{\lambda^B}{\lambda^B - \delta} \cdot \sqrt{2\eta Nd}. \quad (\text{A.7})$$

Hence, for every k , we have

$$\begin{aligned} \left\| \tilde{y}^{(k)} \right\|_{L_2} &\leq \frac{\gamma_1 \gamma_2 \gamma_3 \tilde{\omega}_4 + \gamma_1 \gamma_2 \tilde{\omega}_3 + \tilde{\omega}_1}{1 - \gamma_1 \gamma_2 \gamma_3 \gamma_4} \lambda^k + \frac{\gamma_1 \gamma_2 \gamma_3 \hat{\omega}_4 + \gamma_1 \gamma_2 \hat{\omega}_3 + \hat{\omega}_1}{1 - \gamma_1 \gamma_2 \gamma_3 \gamma_4} \\ &\leq \frac{\gamma_1 \gamma_2 \gamma_3 \tilde{\omega}_4 + \gamma_1 \gamma_2 \tilde{\omega}_3 + \tilde{\omega}_1}{1 - \gamma_1 \gamma_2 \gamma_3 \gamma_4} + \frac{\gamma_1 \gamma_2 \gamma_3 \hat{\omega}_4 + \gamma_1 \gamma_2 \hat{\omega}_3 + \hat{\omega}_1}{1 - \gamma_1 \gamma_2 \gamma_3 \gamma_4}, \end{aligned}$$

and

$$\left\| q^{(k)} \right\|_{L_2} \leq \frac{\gamma_3 \gamma_4 \tilde{\omega}_1 + \gamma_3 \tilde{\omega}_4 + \tilde{\omega}_3}{1 - \gamma_1 \gamma_2 \gamma_3 \gamma_4} \lambda^k + \frac{\gamma_3 \gamma_4 \hat{\omega}_1 + \gamma_3 \hat{\omega}_4 + \hat{\omega}_3}{1 - \gamma_1 \gamma_2 \gamma_3 \gamma_4}$$

$$\leq \frac{\gamma_3\gamma_4\tilde{\omega}_1 + \gamma_3\tilde{\omega}_4 + \tilde{\omega}_3}{1 - \gamma_1\gamma_2\gamma_3\gamma_4} + \frac{\gamma_3\gamma_4\hat{\omega}_1 + \gamma_3\hat{\omega}_4 + \hat{\omega}_3}{1 - \gamma_1\gamma_2\gamma_3\gamma_4},$$

where we used $0 < \lambda < 1$. Next, we can compute that

$$\mathbb{E} \|y^{(k)}\|^2 \leq 2\mathbb{E} \|\tilde{y}^{(k)}\|^2 + 2\mathbb{E} \|\bar{y}^{(k)}\|^2 = 2\mathbb{E} \|\tilde{y}^{(k)}\|^2 + 2N\mathbb{E} \|\bar{y}^{(k)}\|^2.$$

Moreover,

$$\begin{aligned} 2N\mathbb{E} \|\bar{y}^{(k)}\|^2 &= 2N\mathbb{E} \left\| \frac{1}{N} \sum_{i=1}^N \nabla f_i(x_i^{(k)}) + \bar{\xi}^{(k+1)} \right\|^2 \\ &= 2N\mathbb{E} \left\| \frac{1}{N} \sum_{i=1}^N (\nabla f_i(x_i^{(k)}) - f_i(x_*)) + \bar{\xi}^{(k+1)} \right\|^2 \\ &\leq 4N\mathbb{E} \left\| \frac{1}{N} \sum_{i=1}^N (\nabla f_i(x_i^{(k)}) - f_i(x_*)) \right\|^2 + 4N\mathbb{E} \|\bar{\xi}^{(k+1)}\|^2 \\ &\leq \frac{4L^2}{N} \mathbb{E} \sum_{i=1}^N \|x_i^{(k)} - x_*\|^2 + \frac{4}{N} \sigma^2 = \frac{4L^2}{N} \mathbb{E} \|q^{(k)}\|^2 + \frac{4}{N} \sigma^2. \end{aligned}$$

Hence, we conclude that

$$\begin{aligned} \mathbb{E} \|y^{(k)}\|^2 &\leq 2\mathbb{E} \|\tilde{y}^{(k)}\|^2 + \frac{4L^2}{N} \mathbb{E} \|q^{(k)}\|^2 + \frac{4}{N} \sigma^2 \\ &\leq 2 \left(\frac{\gamma_1\gamma_2\gamma_3(\tilde{\omega}_4 + \hat{\omega}_4) + \gamma_1\gamma_2(\tilde{\omega}_3 + \hat{\omega}_3) + \tilde{\omega}_1 + \hat{\omega}_1}{1 - \gamma_1\gamma_2\gamma_3\gamma_4} \right)^2 \\ &\quad + \frac{4L^2}{N} \left(\frac{\gamma_3\gamma_4(\tilde{\omega}_1 + \hat{\omega}_1) + \gamma_3(\tilde{\omega}_4 + \hat{\omega}_4) + \tilde{\omega}_3 + \hat{\omega}_3}{1 - \gamma_1\gamma_2\gamma_3\gamma_4} \right)^2 + \frac{4}{N} \sigma^2. \end{aligned} \quad (\text{A.8})$$

This completes the proof. \square

A.3 Proof of Lemma 3.9

Proof. The proof can be directly adapted from the proof of Lemma 3.9 in [42] by replacing the matrix notation in [42] by the vector notation in our paper. Since $\nabla F(x)$ is L -Lipschitz, we have

$$\|\nabla F(x^{(k+1)}) - \nabla F(x^{(k)})\| \leq L \|x^{(k+1)} - x^{(k)}\| \leq L \|x^{(k+1)} - \mathbf{x}_*\| + L \|x^{(k+1)} - \mathbf{x}_*\|. \quad (\text{A.9})$$

By the definition of z and q , it follows from (A.9) that

$$\lambda^{-(k+1)} \|z^{(k+1)}\| \leq L\lambda^{-(k+1)} \|q^{(k+1)}\| + \frac{L}{\lambda} \lambda^{-k} \|q^{(k)}\|. \quad (\text{A.10})$$

By taking the maximum on both hand sides of (A.10) over $k = 0, 1, \dots, K-1$, we conclude that

$$\|z\|_{L_2}^{\lambda, K} \leq L\|q\|_{L_2}^{\lambda, K} + \frac{L}{\lambda} \|q\|_{L_2}^{\lambda, K-1} \leq L \left(1 + \frac{1}{\lambda}\right) \|q\|_{L_2}^{\lambda, K}. \quad (\text{A.11})$$

This completes the proof. \square

A.4 Proof of Lemma 3.10

Proof. First, we recall from (3.8) and (3.42) that

$$y^{(k+1)} = \mathcal{W}^{(k)} y^{(k)} + z^{(k+1)} + \xi^{(k+2)} - \xi^{(k+1)}. \quad (\text{A.12})$$

Therefore, for any $k \geq B - 1$, we have

$$\begin{aligned} \|\tilde{y}^{(k+1)}\|_{L_2} &= \|\mathcal{L}_N y^{(k+1)}\|_{L_2} \\ &\leq \|\mathcal{L}_N \mathcal{W}_B^{(k)} y^{(k+1-B)}\|_{L_2} + \|\mathcal{L}_N \mathcal{W}_{B-1}^{(k)} z^{(k+2-B)}\|_{L_2} + \dots + \|\mathcal{L}_N \mathcal{W}_1^{(k)} z^{(k)}\|_{L_2} + \|\mathcal{L}_N \mathcal{W}_0^{(k)} z^{(k+1)}\|_{L_2} \\ &\quad + \|\mathcal{L}_N \mathcal{W}_{B-1}^{(k)} \xi^{(k+3-B)}\|_{L_2} + \dots + \|\mathcal{L}_N \mathcal{W}_1^{(k)} \xi^{(k+1)}\|_{L_2} + \|\mathcal{L}_N \mathcal{W}_0^{(k)} \xi^{(k+2)}\|_{L_2} \\ &\quad + \|\mathcal{L}_N \mathcal{W}_{B-1}^{(k)} \xi^{(k+2-B)}\|_{L_2} + \dots + \|\mathcal{L}_N \mathcal{W}_1^{(k)} \xi^{(k)}\|_{L_2} + \|\mathcal{L}_N \mathcal{W}_0^{(k)} \xi^{(k+1)}\|_{L_2}, \end{aligned}$$

where $\mathcal{L}_N = I_{Nd} - \frac{1}{N} \left((1_N 1_N^\top) \otimes I_d \right)$. By applying Lemma C.1 and Assumption 3.1, we get

$$\|\tilde{y}^{(k+1)}\|_{L_2} \leq \delta \|\tilde{y}^{(k+1-B)}\|_{L_2} + \sum_{t=1}^B \|z^{(k+2-t)}\|_{L_2} + 2B\sigma\sqrt{N}. \quad (\text{A.13})$$

Therefore, for any $k = B - 1, B, \dots$, we have

$$\lambda^{-(k+1)} \|\tilde{y}^{(k+1)}\|_{L_2} \leq \frac{\delta}{\lambda^B} \lambda^{-(k+1-B)} \|\tilde{y}^{(k+1-B)}\|_{L_2} + \sum_{t=1}^B \frac{1}{\lambda^{t-1}} \lambda^{-(k+2-t)} \|z^{(k+2-t)}\|_{L_2} + 2B\sigma\sqrt{N}. \quad (\text{A.14})$$

By following the similar argument as in the proof of Lemma 3.10 in [42], we obtain that for every K :

$$\|\tilde{y}\|_{L_2}^{\lambda, K} \leq \frac{\delta}{\lambda^B} \|\tilde{y}\|_{L_2}^{\lambda, K} + \sum_{t=1}^B \frac{1}{\lambda^{t-1}} \|z\|_{L_2}^{\lambda, K} + \frac{2B\sigma\sqrt{N}}{\lambda^K} + \sum_{t=1}^B \lambda^{1-t} \|\tilde{y}^{(t-1)}\|_{L_2}. \quad (\text{A.15})$$

This implies that

$$\|\tilde{y}\|_{L_2}^{\lambda, K} \leq \frac{\lambda(1 - \lambda^B)}{(\lambda^B - \delta)(1 - \lambda)} \|z\|_{L_2}^{\lambda, K} + \frac{\lambda^B}{\lambda^B - \delta} \frac{2B\sigma\sqrt{N}}{\lambda^K} + \frac{\lambda^B}{\lambda^B - \delta} \sum_{t=1}^B \lambda^{1-t} \|\tilde{y}^{(t-1)}\|_{L_2}. \quad (\text{A.16})$$

This completes the proof. \square

A.5 Proof of Lemma 3.11

Proof. Recall from (3.7) that $x^{(k+1)} = \mathcal{W}^{(k)} x^{(k)} - \eta y^{(k)} + \sqrt{2\eta} w^{(k+1)}$. This recursion is structurally similar to the one in (A.12), which appears in the proof of Lemma 3.10. The proof therefore proceeds by following the same steps as in the proof of Lemma 3.10 and the fact that $\mathbb{E} \left\| \sqrt{2\eta} w^{(k+1)} \right\|^2 = 2\eta Nd$. \square

A.6 Proof of Lemma 3.12

Proof. First, let us recall from (3.42) that $q^{(k)} = x^{(k)} - \mathbf{x}_* = x^{(k)} - \bar{\mathbf{x}}^{(k)} + \bar{\mathbf{x}}^{(k)} - \mathbf{x}_*$. Next, by taking the average of N nodes in (3.4)–(3.5), and using the fact that $W^{(k)}$ is doubly stochastic, we obtain:

$$\bar{x}^{(k+1)} = \bar{x}^{(k)} - \eta \bar{y}^{(k)} + \sqrt{2\eta} \bar{w}^{(k+1)}, \quad (\text{A.17})$$

where for any $k = 0, 1, 2, \dots$,

$$\bar{y}^{(k+1)} = \bar{y}^{(k)} + \frac{1}{N} \sum_{i=1}^N \nabla f_i(x_i^{(k+1)}) - \frac{1}{N} \sum_{i=1}^N \nabla f_i(x_i^{(k)}) + \bar{\xi}^{(k+2)} - \bar{\xi}^{(k+1)}, \quad (\text{A.18})$$

which implies that for any $k = 0, 1, 2, \dots$,

$$\bar{y}^{(k)} = \frac{1}{N} \sum_{i=1}^N \nabla f_i(x_i^{(k)}) + \bar{\xi}^{(k+1)}. \quad (\text{A.19})$$

Therefore, we have

$$\bar{x}^{(k+1)} = \bar{x}^{(k)} - \eta \frac{1}{N} \sum_{i=1}^N \nabla f_i(x_i^{(k)}) - \eta \bar{\xi}^{(k+1)} + \sqrt{2\eta} \bar{w}^{(k+1)}. \quad (\text{A.20})$$

By Lemma C.3, we can re-write the equation (A.20) as:

$$\bar{x}^{(k+1)} = \bar{x}^{(k)} - \eta \frac{1}{N} \sum_{i=1}^N \nabla f_i(s_i^{(k)}), \quad (\text{A.21})$$

where $s_i^{(k)}$ is defined implicitly via:

$$\nabla f_i(s_i^{(k)}) = \nabla f_i(x_i^{(k)}) + \bar{\xi}^{(k+1)} - \sqrt{\frac{2}{\eta}} \bar{w}^{(k+1)}. \quad (\text{A.22})$$

Since f_i is μ -strongly convex, we have $\|s_i^{(k)} - x_i^{(k)}\| \leq \frac{1}{\mu} \|\bar{\xi}^{(k+1)} - \sqrt{\frac{2}{\eta}} \bar{w}^{(k+1)}\|$, which implies that

$$\|s_i^{(k)} - x_i^{(k)}\|_{L_2} \leq \frac{1}{\mu} \left(\frac{\sigma}{\sqrt{N}} + \sqrt{\frac{2d}{\eta N}} \right). \quad (\text{A.23})$$

By applying Lemma C.2, under the assumption that $\sqrt{1 - \frac{\eta\mu\beta}{\beta+1}} \leq \lambda < 1$ and $\eta \leq \frac{1}{(1+\alpha)L}$, where $\alpha, \beta > 0$, we have

$$\|\bar{x} - x_*\|_{L_2}^{\lambda, K} \leq 2 \|\bar{x}^{(0)} - x_*\| + (\lambda\sqrt{N})^{-1} \left(\sqrt{\frac{L(1+\alpha)}{\mu\alpha}} + \beta \right) \sum_{i=1}^N \|\bar{x} - s_i\|_{L_2}^{\lambda, K}, \quad (\text{A.24})$$

for any $K = 0, 1, 2, \dots$ where x_* is the minimizer of f and $\|\bar{x} - s_i\|_{L_2}^{\lambda, K} = \max_{0,1,\dots,K} \frac{1}{\lambda^k} \left(\mathbb{E} \|\bar{x} - s_i^{(k)}\|^2 \right)^{1/2}$.

Therefore, we have

$$\begin{aligned} \|\bar{x} - x_*\|_{L_2}^{\lambda, K} &\leq 2 \|\bar{x}^{(0)} - x_*\| + (\lambda\sqrt{N})^{-1} \left(\sqrt{\frac{L(1+\alpha)}{\mu\alpha}} + \beta \right) \sum_{i=1}^N \|\bar{x} - x_i\|_{L_2}^{\lambda, K} \\ &\quad + (\lambda\sqrt{N})^{-1} \left(\sqrt{\frac{L(1+\alpha)}{\mu\alpha}} + \beta \right) \frac{1}{\mu} \left(\frac{\sigma}{\sqrt{N}} + \sqrt{\frac{2d}{\eta N}} \right) \frac{N}{\lambda^K} \\ &\leq 2 \|\bar{x}^{(0)} - x_*\| + (\lambda)^{-1} \left(\sqrt{\frac{L(1+\alpha)}{\mu\alpha}} + \beta \right) \|\tilde{x}\|_{L_2}^{\lambda, K} + (\lambda)^{-1} \left(\sqrt{\frac{L(1+\alpha)}{\mu\alpha}} + \beta \right) \frac{1}{\mu} \left(\sigma + \sqrt{\frac{2d}{\eta}} \right) \frac{1}{\lambda^K}, \end{aligned}$$

where we used $\|\bar{x} - x_i\|_{L_2}^{\lambda,K} \leq \sqrt{N}\|\tilde{x}\|_{L_2}^{\lambda,K}$. Finally, $q^{(k)} = \tilde{x}^{(k)} + \bar{\mathbf{x}}^{(k)} - \mathbf{x}_*$, and it follows that

$$\|q\|_{L_2}^{\lambda,K} \leq \|\tilde{x}\|_{L_2}^{\lambda,K} + \sqrt{N}\|\bar{x} - x_*\|_{L_2}^{\lambda,K}. \quad (\text{A.25})$$

Hence, we conclude that

$$\begin{aligned} \|q\|_{L_2}^{\lambda,K} &\leq 2\sqrt{N}\|\bar{x}^{(0)} - x_*\| + \left(1 + \frac{\sqrt{N}}{\lambda} \left(\sqrt{\frac{L(1+\alpha)}{\mu\alpha}} + \beta\right)\right) \|\tilde{x}\|_{L_2}^{\lambda,K} \\ &\quad + \frac{\sqrt{N}}{\lambda} \left(\sqrt{\frac{L(1+\alpha)}{\mu\alpha}} + \beta\right) \frac{1}{\mu} \left(\sigma + \sqrt{\frac{2d}{\eta}}\right) \frac{1}{\lambda^K}. \end{aligned}$$

This completes the proof. \square

A.7 Proof of Lemma 3.13

Proof. Under the assumption $\delta < \lambda^B < 1$ with $\delta(k)$ defined in Assumption 3.2 and (3.40), Lemma 3.9, Lemma 3.10, Lemma 3.11 and Lemma 3.12 hold, and it follows from (3.51)-(3.54) that

$$\|\tilde{y}\|_{L_2}^{\lambda,K} \leq \gamma_1\gamma_2\gamma_3\gamma_4\|\tilde{y}\|_{L_2}^{\lambda,K} + \gamma_1\gamma_2\gamma_3\omega_4(K) + \gamma_1\gamma_2\omega_3(K) + \gamma_1\omega_2(K) + \omega_1(K), \quad (\text{A.26})$$

and if $0 < \gamma_1\gamma_2\gamma_3\gamma_4 < 1$, we obtain:

$$\|\tilde{y}\|_{L_2}^{\lambda,K} \leq \frac{\gamma_1\gamma_2\gamma_3\omega_4(K) + \gamma_1\gamma_2\omega_3(K) + \gamma_1\omega_2(K) + \omega_1(K)}{1 - \gamma_1\gamma_2\gamma_3\gamma_4}. \quad (\text{A.27})$$

Similarly, one can show that

$$\|q\|_{L_2}^{\lambda,K} \leq \gamma_1\gamma_2\gamma_3\gamma_4\|q\|_{L_2}^{\lambda,K} + \gamma_3\gamma_4\gamma_1\omega_2(K) + \gamma_3\gamma_4\omega_1(K) + \gamma_3\omega_4(K) + \omega_3(K), \quad (\text{A.28})$$

and if $0 < \gamma_1\gamma_2\gamma_3\gamma_4 < 1$, we obtain:

$$\|q\|_{L_2}^{\lambda,K} \leq \frac{\gamma_3\gamma_4\gamma_1\omega_2(K) + \gamma_3\gamma_4\omega_1(K) + \gamma_3\omega_4(K) + \omega_3(K)}{1 - \gamma_1\gamma_2\gamma_3\gamma_4}. \quad (\text{A.29})$$

This completes the proof. \square

A.8 Proof of Lemma 3.14

Proof. By the iterates of $x^{(k)}$ given in (3.7), we get

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

It follows that for any $k \geq 1$,

$$x^{(k)} = \left(W_k^{(k-1)} \otimes I_d\right) x^{(0)} - \eta \sum_{s=0}^{k-1} \left(W_{k-1-s}^{(k-1)} \otimes I_d\right) y^{(s)} + \sqrt{2\eta} \sum_{s=0}^{k-1} \left(W_{k-1-s}^{(k-1)} \otimes I_d\right) w^{(s+1)}. \quad (\text{A.30})$$

Let us define $\bar{\mathbf{x}}^{(k)} := \left[\left(\bar{x}^{(k)}\right)^\top, \dots, \left(\bar{x}^{(k)}\right)^\top\right]^\top \in \mathbb{R}^{Nd}$ where $\bar{x}^{(k)} = \frac{1}{N} \sum_{i=1}^N x_i^{(k)}$. Notice that $\bar{\mathbf{x}}^{(k)} = \frac{1}{N} \left(\left(1_N 1_N^\top\right) \otimes I_d\right) x^{(k)}$. Therefore, we get

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

Note that it follows from (A.30) that

$$\begin{aligned}
& x^{(k)} - \frac{1}{N} \left((1_N 1_N^\top) \otimes I_d \right) x^{(k)} \\
&= \left(W_k^{(k-1)} \otimes I_d \right) x^{(0)} - \frac{1}{N} \left((1_N 1_N^\top W_k^{(k-1)}) \otimes I_d \right) x^{(0)} \\
&\quad - \eta \sum_{s=0}^{k-1} \left(W_{k-1-s}^{(k-1)} \otimes I_d \right) y^{(s)} + \eta \sum_{s=0}^{k-1} \frac{1}{N} \left((1_N 1_N^\top W_{k-1-s}^{(k-1)}) \otimes I_d \right) y^{(s)} \\
&\quad + \sqrt{2\eta} \sum_{s=0}^{k-1} \left(W_{k-1-s}^{(k-1)} \otimes I_d \right) w^{(s+1)} - \sqrt{2\eta} \sum_{s=0}^{k-1} \frac{1}{N} \left((1_N 1_N^\top W_{k-1-s}^{(k-1)}) \otimes I_d \right) w^{(s+1)}.
\end{aligned}$$

By Cauchy-Schwarz inequality, we have

$$\begin{aligned}
& \left\| x^{(k)} - \frac{1}{N} \left((1_N 1_N^\top) \otimes I_d \right) x^{(k)} \right\|^2 \\
&\leq 3 \left\| \left(W_k^{(k-1)} \otimes I_d \right) x^{(0)} - \frac{1}{N} \left((1_N 1_N^\top W_k^{(k-1)}) \otimes I_d \right) x^{(0)} \right\|^2 \\
&\quad + 3 \left\| -\eta \sum_{s=0}^{k-1} \left(W_{k-1-s}^{(k-1)} \otimes I_d \right) y^{(s)} + \eta \sum_{s=0}^{k-1} \frac{1}{N} \left((1_N 1_N^\top W_{k-1-s}^{(k-1)}) \otimes I_d \right) y^{(s)} \right\|^2 \\
&\quad + 3 \left\| \sqrt{2\eta} \sum_{s=0}^{k-1} \left(W_{k-1-s}^{(k-1)} \otimes I_d \right) w^{(s+1)} - \sqrt{2\eta} \sum_{s=0}^{k-1} \frac{1}{N} \left((1_N 1_N^\top W_{k-1-s}^{(k-1)}) \otimes I_d \right) w^{(s+1)} \right\|^2 \\
&= 3 \left\| \left(W_k^{(k-1)} \otimes I_d \right) x^{(0)} - \frac{1}{N} \left((1_N 1_N^\top) \otimes I_d \right) x^{(0)} \right\|^2 \\
&\quad + 3 \left\| -\eta \sum_{s=0}^{k-1} \left(W_{k-1-s}^{(k-1)} \otimes I_d \right) y^{(s)} + \eta \sum_{s=0}^{k-1} \frac{1}{N} \left((1_N 1_N^\top) \otimes I_d \right) y^{(s)} \right\|^2 \\
&\quad + 3 \left\| \sqrt{2\eta} \sum_{s=0}^{k-1} \left(W_{k-1-s}^{(k-1)} \otimes I_d \right) w^{(s+1)} - \sqrt{2\eta} \sum_{s=0}^{k-1} \frac{1}{N} \left((1_N 1_N^\top) \otimes I_d \right) w^{(s+1)} \right\|^2,
\end{aligned}$$

where we used the property that $W^{(k)}$ is doubly stochastic for every k . Therefore, we get

$$\begin{aligned}
\left\| x^{(k)} - \frac{1}{N} \left((1_N 1_N^\top) \otimes I_d \right) x^{(k)} \right\|^2 &\leq 3 \left\| \left(\left(W_k^{(k-1)} - \frac{1}{N} 1_N 1_N^\top \right) \otimes I_d \right) x^{(0)} \right\|^2 \\
&\quad + 3\eta^2 \left\| \sum_{s=0}^{k-1} \left(\left(W_{k-1-s}^{(k-1)} - \frac{1}{N} 1_N 1_N^\top \right) \otimes I_d \right) y^{(s)} \right\|^2 \\
&\quad + 6\eta \left\| \sum_{s=0}^{k-1} \left(\left(W_{k-1-s}^{(k-1)} - \frac{1}{N} 1_N 1_N^\top \right) \otimes I_d \right) w^{(s+1)} \right\|^2.
\end{aligned} \tag{A.31}$$

Note that

$$\begin{aligned}
3\eta^2 \left\| \sum_{s=0}^{k-1} \left(\left(W_{k-1-s}^{(k-1)} - \frac{1}{N} 1_N 1_N^\top \right) \otimes I_d \right) y^{(s)} \right\|^2 &\leq 3\eta^2 \left(\sum_{s=0}^{k-1} \left\| \left(W_{k-1-s}^{(k-1)} - \frac{1}{N} 1_N 1_N^\top \right) \otimes I_d \right\| \cdot \|y^{(s)}\| \right)^2 \\
&\leq 3\eta^2 \left(\sum_{s=0}^{k-1} \left\| W_{k-1-s}^{(k-1)} - \frac{1}{N} 1_N 1_N^\top \right\| \cdot \|y^{(s)}\| \right)^2
\end{aligned}$$

$$\begin{aligned}
&= 3\eta^2 \left(\sum_{s=0}^{k-1} \bar{\gamma}_{k-1-s}^{(k-1)} \cdot \|y^{(s)}\| \right)^2 \\
&= 3\eta^2 \left(\sum_{s=0}^{k-1} \bar{\gamma}_{k-1-s}^{(k-1)} \right)^2 \left(\frac{\sum_{s=0}^{k-1} \bar{\gamma}_{k-1-s}^{(k-1)} \cdot \|y^{(s)}\|}{\sum_{s=0}^{k-1} \bar{\gamma}_{k-1-s}^{(k-1)}} \right)^2 \\
&\leq 3\eta^2 \left(\sum_{s=0}^{k-1} \bar{\gamma}_{k-1-s}^{(k-1)} \right)^2 \sum_{s=0}^{k-1} \frac{\bar{\gamma}_{k-1-s}^{(k-1)}}{\sum_{s=0}^{k-1} \bar{\gamma}_{k-1-s}^{(k-1)}} \|y^{(s)}\|^2,
\end{aligned}$$

where we used Jensen's inequality in the last step above. Recall from Lemma 3.8 that for every $k = 0, 1, 2, \dots$, $\mathbb{E} \left[\|y^{(k)}\|^2 \right] \leq D^2$, where D is defined in (3.20). Therefore, we have

$$\begin{aligned}
&3\eta^2 \mathbb{E} \left[\left\| \sum_{s=0}^{k-1} \left(\left(W_{k-1-s}^{(k-1)} - \frac{1}{N} 1_N 1_N^\top \right) \otimes I_d \right) y^{(s)} \right\|^2 \right] \\
&\leq 3D^2 \eta^2 \left(\sum_{s=0}^{k-1} \bar{\gamma}_{k-1-s}^{(k-1)} \right)^2 \sum_{s=0}^{k-1} \frac{\bar{\gamma}_{k-1-s}^{(k-1)}}{\sum_{s=0}^{k-1} \bar{\gamma}_{k-1-s}^{(k-1)}} \leq 3D^2 \eta^2 \left(\sum_{s=0}^{k-1} \bar{\gamma}_{k-1-s}^{(k-1)} \right)^2.
\end{aligned}$$

Similarly, we can show that

$$3 \left\| \left(\left(W_k^{(k-1)} - \frac{1}{N} 1_N 1_N^\top \right) \otimes I_d \right) x^{(0)} \right\|^2 \leq 3 \left(\bar{\gamma}_k^{(k-1)} \right)^2 \|x^{(0)}\|^2.$$

It follows from (A.31) that

$$\begin{aligned}
&\sum_{i=1}^N \mathbb{E} \|x_i^{(k)} - \bar{x}^{(k)}\|^2 \\
&\leq 3 \left(\bar{\gamma}_k^{(k-1)} \right)^2 \mathbb{E} \|x^{(0)}\|^2 + 3D^2 \eta^2 \left(\sum_{s=0}^{k-1} \bar{\gamma}_{k-1-s}^{(k-1)} \right)^2 \\
&\quad + 6\eta \sum_{s=0}^{k-1} \mathbb{E} \left\| \left(\left(W_{k-1-s}^{(k-1)} - \frac{1}{N} 1_N 1_N^\top \right) \otimes I_d \right) w^{(s+1)} \right\|^2 \\
&\leq 3 \left(\bar{\gamma}_k^{(k-1)} \right)^2 \mathbb{E} \|x^{(0)}\|^2 + 3D^2 \eta^2 \left(\sum_{s=0}^{k-1} \bar{\gamma}_{k-1-s}^{(k-1)} \right)^2 + 6\eta \sum_{s=0}^{k-1} \left\| W_{k-1-s}^{(k-1)} - \frac{1}{N} 1_N 1_N^\top \right\|^2 \mathbb{E} \|w^{(s+1)}\|^2 \\
&\leq 3 \left(\bar{\gamma}_k^{(k-1)} \right)^2 \mathbb{E} \|x^{(0)}\|^2 + 3D^2 \eta^2 \left(\sum_{s=0}^{k-1} \bar{\gamma}_{k-1-s}^{(k-1)} \right)^2 + 6dN\eta \sum_{s=0}^{k-1} \left(\bar{\gamma}_{k-1-s}^{(k-1)} \right)^2.
\end{aligned}$$

The proof is complete. \square

A.9 Proof of Lemma 3.15

Proof. For $k = 0$, the bound holds trivially. Assume $k \geq 1$. It follows from Lemma 3.14 that for any $k \geq 1$,

$$\sum_{i=1}^N \mathbb{E} \|x_i^{(k)} - \bar{x}^{(k)}\|^2 \leq 3 \left(\bar{\gamma}_k^{(k-1)} \right)^2 \mathbb{E} \|x^{(0)}\|^2 + 3D^2 \eta^2 \left(\sum_{s=0}^{k-1} \bar{\gamma}_{k-1-s}^{(k-1)} \right)^2 + 6dN\eta \sum_{s=0}^{k-1} \left(\bar{\gamma}_{k-1-s}^{(k-1)} \right)^2,$$

where D is defined in (3.20) and

$$\bar{\gamma}_{k-1-s}^{(k-1)} = \left\| W_{k-1-s}^{(k-1)} - \frac{1}{N} 1_N 1_N^\top \right\|. \quad (\text{A.32})$$

Under Assumption 3.2, there exists some positive integer B such that $\delta := \sup_{k \geq B-1} \delta(k) < 1$, where $\delta(k) := \sigma_{\max} \left\{ W_B^{(k)} - \frac{1}{N} 1_N 1_N^\top \right\}$ for every $k = 0, 1, 2, \dots$. For every k , $W^{(k)}$ is doubly stochastic. That means that for every k , we have $W^{(k)} J = J W^{(k)}$ with $J := \frac{1}{N} 1_N 1_N^\top$ where $J^2 = J$. Note that if A and M are $N \times N$ doubly stochastic matrices that are not necessarily symmetric, they satisfy $AJ = JA = J$ and $MJ = JM = J$ and we always have $(A - J)(M - J) = AM - AJ - JM + J^2 = AM - J$. Furthermore, the product AM is always double stochastic, even if it is not necessarily symmetric. Therefore, writing $j = mB + r$ with $m = \lfloor j/B \rfloor$ and $0 \leq r < B$, we obtain

$$W_j^{(k-1)} - J = \left(\prod_{\ell=0}^{m-1} \left(W_B^{(k-1-\ell B)} - J \right) \right) \left(W_r^{(k-1-mB)} - J \right),$$

where matrices $W_r^{(k-1-mB)}$ and $W_B^{(k-1-\ell B)}$ are all double stochastic as products of double stochastic matrices. By part (iii) of Assumption 3.2, $\|W_B^{(k)} - J\| \leq \delta$ for all $k \geq B-1$, and because $W^{(k)}$ is non-expansive on the orthogonal complement of 1_N (i.e. the singular values of the symmetric matrix $W^{(k)} - J$ is at most 1)¹, we also have $\|W_r^{(k)} - J\| \leq \|W^{(k)} - J\| \cdot \|W^{(k-1)} - J\| \dots \|W^{(k-r+1)} - J\| \leq 1$, for any $r \geq 0$ and k . Hence

$$\bar{\gamma}_{k-1-s}^{(k-1)} \leq \delta \lfloor \frac{k-1-s}{B} \rfloor \leq \delta^{\frac{k-1-s}{B}-1}, \quad s = 0, 1, \dots, k-1, \quad (\text{A.33})$$

and $\bar{\gamma}_k^{(k-1)} \leq \delta^{\frac{k}{B}-1}$. Therefore, we get

$$\begin{aligned} \sum_{i=1}^N \mathbb{E} \|x_i^{(k)} - \bar{x}^{(k)}\|^2 &\leq 3 \left(\delta^{\frac{k}{B}-1} \right)^2 \mathbb{E} \|x^{(0)}\|^2 + 3D^2 \eta^2 \left(\sum_{s=0}^{k-1} \delta^{\frac{k-1-s}{B}-1} \right)^2 + 6dN\eta \sum_{s=0}^{k-1} \left(\delta^{\frac{k-1-s}{B}-1} \right)^2 \\ &\leq 3 \cdot \delta^{-2} \left(\delta^{\frac{2}{B}} \right)^k \mathbb{E} \|x^{(0)}\|^2 + \frac{3D^2 \eta^2 \delta^{-2}}{\left(1 - \delta^{\frac{1}{B}} \right)^2} + \frac{6dN\eta \cdot \delta^{-2}}{1 - \delta^{\frac{2}{B}}}. \end{aligned}$$

The proof is complete. \square

A.10 Proof of Lemma 3.16

Proof. First, we can compute that

$$\mathbb{E} \|\mathcal{E}_k\|^2 = \mathbb{E} \left\| \frac{1}{N} \sum_{i=1}^N \left(\nabla f_i(x_i^{(k)}) - \nabla f_i(\bar{x}^{(k)}) \right) \right\|^2 \leq \mathbb{E} \left\| \frac{1}{N} \sum_{i=1}^N \left(\nabla f_i(x_i^{(k)}) - \nabla f_i(\bar{x}^{(k)}) \right) \right\|^2.$$

By Lemma 3.15, we can compute that

$$\mathbb{E} \left\| \frac{1}{N} \sum_{i=1}^N \left(\nabla f_i(x_i^{(k)}) - \nabla f_i(\bar{x}^{(k)}) \right) \right\|^2 \leq \frac{1}{N^2} \sum_{i=1}^N N \mathbb{E} \left\| \left(\nabla f_i(x_i^{(k)}) - \nabla f_i(\bar{x}^{(k)}) \right) \right\|^2$$

¹Indeed, $W^{(k)}$ is a symmetric double stochastic matrix with all the eigenvalues values lying in the interval $[-1, 1]$ and admits one as an eigenvalue with the eigenvector 1_N . Therefore, the norm of the eigenvalues of the matrix $W^{(k)} - J$ is at most one, which coincides with the singular values of $W^{(k)} - J$.

$$\begin{aligned}
&\leq \frac{1}{N} L^2 \sum_{i=1}^N \mathbb{E} \|x_i^{(k)} - \bar{x}^{(k)}\|^2 \\
&\leq \frac{3L^2 \delta^{-2}}{N} \left(\delta^{\frac{2}{B}}\right)^k \mathbb{E} \|x^{(0)}\|^2 + \frac{3L^2 D^2 \eta^2 \delta^{-2}}{N(1 - \delta^{\frac{1}{B}})^2} + \frac{6dL^2 \eta \cdot \delta^{-2}}{1 - \delta^{\frac{2}{B}}}.
\end{aligned}$$

The proof is complete. \square

A.11 Proof of Lemma 3.17

Proof. The proof is similar to the proof of Lemma 7 in [30] and for the sake of completeness we include all the details here. From (3.66) and (3.68), we can compute that

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

where we recall from (3.67) that $\mathcal{E}_k = \frac{1}{N} \nabla f(\bar{x}^{(k)}) - \frac{1}{N} \sum_{i=1}^N \nabla f_i(x_i^{(k)})$, and this implies that

$$\begin{aligned}
\|\bar{x}^{(k+1)} - x_{k+1}\|^2 &= \left\| \bar{x}^{(k)} - x_k - \frac{\eta}{N} [\nabla f(\bar{x}^{(k)}) - \nabla f(x_k)] \right\|^2 + \eta^2 \|\mathcal{E}_k - \bar{\xi}^{(k+1)}\|^2 \\
&\quad + 2 \left\langle \bar{x}^{(k)} - x_k - \frac{\eta}{N} [\nabla f(\bar{x}^{(k)}) - \nabla f(x_k)], \eta \mathcal{E}_k - \eta \bar{\xi}^{(k+1)} \right\rangle \\
&= \|\bar{x}^{(k)} - x_k\|^2 + \eta^2 \left\| \frac{1}{N} [\nabla f(\bar{x}^{(k)}) - \nabla f(x_k)] \right\|^2 \\
&\quad - 2 \left\langle \bar{x}^{(k)} - x_k, \eta \frac{1}{N} [\nabla f(\bar{x}^{(k)}) - \nabla f(x_k)] \right\rangle + \eta^2 \|\mathcal{E}_k - \bar{\xi}^{(k+1)}\|^2 \\
&\quad + 2 \left\langle \bar{x}^{(k)} - x_k - \eta \frac{1}{N} [\nabla f(\bar{x}^{(k)}) - \nabla f(x_k)], \eta \mathcal{E}_k - \eta \bar{\xi}^{(k+1)} \right\rangle \\
&\leq \|\bar{x}^{(k)} - x_k\|^2 + \eta^2 L \left\langle \bar{x}^{(k)} - x_k, \frac{1}{N} [\nabla f(\bar{x}^{(k)}) - \nabla f(x_k)] \right\rangle \\
&\quad - 2 \left\langle \bar{x}^{(k)} - x_k, \eta \frac{1}{N} [\nabla f(\bar{x}^{(k)}) - \nabla f(x_k)] \right\rangle + \eta^2 \|\mathcal{E}_k - \bar{\xi}^{(k+1)}\|^2 \\
&\quad + 2 \left\langle \bar{x}^{(k)} - x_k - \eta \frac{1}{N} [\nabla f(\bar{x}^{(k)}) - \nabla f(x_k)], \eta \mathcal{E}_k - \eta \bar{\xi}^{(k+1)} \right\rangle \\
&\leq \left(1 - 2\eta\mu \left(1 - \frac{\eta L}{2}\right)\right) \|\bar{x}^{(k)} - x_k\|^2 + \eta^2 \|\mathcal{E}_k - \bar{\xi}^{(k+1)}\|^2 \\
&\quad + 2 \left\langle \bar{x}^{(k)} - x_k - \eta \frac{1}{N} [\nabla f(\bar{x}^{(k)}) - \nabla f(x_k)], \eta \mathcal{E}_k - \eta \bar{\xi}^{(k+1)} \right\rangle, \quad (\text{A.34})
\end{aligned}$$

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 $\frac{1}{N}f$ and the assumption that $\eta < 2/L$ to obtain the first term after the second inequality above.

Note that $\bar{\xi}^{(k+1)}$ has mean zero and is independent of \mathcal{E}_k , and by Lemma 3.16,

$$\mathbb{E} \|\mathcal{E}_k\|^2 \leq \frac{3L^2 \delta^{-2}}{N} \left(\delta^{\frac{2}{B}}\right)^k \mathbb{E} \|x^{(0)}\|^2 + \frac{3L^2 D^2 \eta^2 \delta^{-2}}{N(1 - \delta^{\frac{1}{B}})^2} + \frac{6dL^2 \eta \cdot \delta^{-2}}{1 - \delta^{\frac{2}{B}}}, \quad (\text{A.35})$$

and we also notice that $\mathbb{E} \|\bar{\xi}^{(k+1)}\|^2 \leq \frac{\sigma^2}{N}$. By taking expectations in (A.34), we get

$$\mathbb{E} \|\bar{x}^{(k+1)} - x_{k+1}\|^2 \leq \left(1 - 2\eta\mu \left(1 - \frac{\eta L}{2}\right)\right) \mathbb{E} \|\bar{x}^{(k)} - x_k\|^2 + \eta^2 \mathbb{E} \|\mathcal{E}_k - \bar{\xi}^{(k+1)}\|^2$$

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

where we used L -smoothness of $\frac{1}{N}f$.

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

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

By applying (A.35), we get

$$\begin{aligned}
\mathbb{E} \left\| \bar{x}^{(k+1)} - x_{k+1} \right\|^2 & \leq \left(1 - \eta\mu \left(1 - \frac{\eta L}{2} \right) \right) \mathbb{E} \left\| \bar{x}^{(k)} - x_k \right\|^2 \\
& \quad + \eta \left(\eta + \frac{(1 + \eta L)^2}{\mu(1 - \frac{\eta L}{2})} \right) \left(\frac{3L^2 \delta^{-2}}{N} \left(\delta^{\frac{2}{B}} \right)^k \mathbb{E} \left\| x^{(0)} \right\|^2 + \frac{3L^2 D^2 \eta^2 \delta^{-2}}{N(1 - \delta^{\frac{1}{B}})^2} + \frac{6dL^2 \eta \cdot \delta^{-2}}{1 - \delta^{\frac{2}{B}}} \right) + \eta^2 \frac{\sigma^2}{N},
\end{aligned}$$

for every k . Note that $\mathbb{E} \left\| \bar{x}^{(0)} - x_0 \right\|^2 = 0$. By iterating the above equation, we get for $k \geq 1$,

$$\begin{aligned}
\mathbb{E} \left\| \bar{x}^{(k)} - x_k \right\|^2 & \leq \sum_{i=0}^{k-1} \left(1 - \eta\mu \left(1 - \frac{\eta L}{2} \right) \right)^i \\
& \quad \cdot \left(\eta \left(\eta + \frac{(1 + \eta L)^2}{\mu(1 - \frac{\eta L}{2})} \right) \left(\frac{3L^2 D^2 \eta^2 \delta^{-2}}{N(1 - \delta^{\frac{1}{B}})^2} + \frac{6dL^2 \eta \cdot \delta^{-2}}{1 - \delta^{\frac{2}{B}}} \right) + \eta^2 \frac{\sigma^2}{N} \right) \\
& \quad + \sum_{i=0}^{k-1} \left(1 - \eta\mu \left(1 - \frac{\eta L}{2} \right) \right)^i \eta \left(\eta + \frac{(1 + \eta L)^2}{\mu(1 - \frac{\eta L}{2})} \right) \frac{3L^2 \delta^{-2}}{N} \left(\delta^{\frac{2}{B}} \right)^{k-i} \mathbb{E} \left\| x^{(0)} \right\|^2 \\
& = \frac{1 - \left(1 - \eta\mu \left(1 - \frac{\eta L}{2} \right) \right)^k}{1 - \left(1 - \eta\mu \left(1 - \frac{\eta L}{2} \right) \right)} \cdot \left(\eta \left(\eta + \frac{(1 + \eta L)^2}{\mu(1 - \frac{\eta L}{2})} \right) \left(\frac{3L^2 D^2 \eta^2 \delta^{-2}}{N(1 - \delta^{\frac{1}{B}})^2} + \frac{6dL^2 \eta \cdot \delta^{-2}}{1 - \delta^{\frac{2}{B}}} \right) + \eta^2 \frac{\sigma^2}{N} \right) \\
& \quad + \frac{\left(\delta^{\frac{2}{B}} \right)^k - \left(1 - \eta\mu \left(1 - \frac{\eta L}{2} \right) \right)^k}{1 - \left(1 - \eta\mu \left(1 - \frac{\eta L}{2} \right) \right) \left(\delta^{\frac{2}{B}} \right)^{-1}} \frac{3L^2 \delta^{-2}}{N} \mathbb{E} \left\| x^{(0)} \right\|^2.
\end{aligned}$$

Under our assumption, the stepsize $\eta < 2/L$, such that $1 - \eta\mu \left(1 - \frac{\eta L}{2} \right) \in [0, 1)$. Hence, we conclude that for every k ,

$$\mathbb{E} \left\| \bar{x}^{(k)} - x_k \right\|^2 \leq \frac{\eta \left(\eta + \frac{(1 + \eta L)^2}{\mu(1 - \frac{\eta L}{2})} \right) \left(\frac{3L^2 D^2 \eta^2 \delta^{-2}}{N(1 - \delta^{\frac{1}{B}})^2} + \frac{6dL^2 \eta \cdot \delta^{-2}}{1 - \delta^{\frac{2}{B}}} \right) + \eta^2 \frac{\sigma^2}{N}}{1 - \left(1 - \eta\mu \left(1 - \frac{\eta L}{2} \right) \right)}$$

$$\begin{aligned}
& + \frac{\left(\delta^{\frac{2}{B}}\right)^k - \left(1 - \eta\mu\left(1 - \frac{\eta L}{2}\right)\right)^k}{1 - \left(1 - \eta\mu\left(1 - \frac{\eta L}{2}\right)\right)\left(\delta^{\frac{2}{B}}\right)^{-1}} \frac{3L^2\delta^{-2}}{N} \mathbb{E} \|x^{(0)}\|^2 \\
& = \frac{\eta\left(\eta + \frac{(1+\eta L)^2}{\mu(1-\frac{\eta L}{2})}\right)\left(\frac{3L^2D^2\eta\delta^{-2}}{N(1-\delta^{\frac{1}{B}})^2} + \frac{6dL^2\cdot\delta^{-2}}{1-\delta^{\frac{2}{B}}}\right) + \eta\frac{\sigma^2}{N}}{\mu\left(1 - \frac{\eta L}{2}\right)} + \frac{\left(\delta^{\frac{2}{B}}\right)^k - \left(1 - \eta\mu\left(1 - \frac{\eta L}{2}\right)\right)^k}{\delta^{\frac{2}{B}} - 1 + \eta\mu\left(1 - \frac{\eta L}{2}\right)} \frac{3L^2\delta^{-2}}{N} \delta^{\frac{2}{B}} \mathbb{E} \|x^{(0)}\|^2.
\end{aligned}$$

The proof is complete. \square

B Proof of Corollary 3.7

With this choice of parameters $(\alpha, \beta, \eta$ and $\lambda)$, the fact that conditions (3.16) hold is a direct consequence of Lemma C.4. Then, both Theorem 3.4 and Theorem 3.5 are applicable. Recall from (3.18), (3.19) and (3.22) that

$$E_1 := (1 - \mu\eta)^k \left(\left(\mathbb{E} \|\bar{x}^{(0)} - x_*\|^2 \right)^{1/2} + \sqrt{2\mu^{-1}dN^{-1}} \right) + \frac{1.65L}{\mu} \sqrt{\eta dN^{-1}}, \quad (\text{B.1})$$

$$\begin{aligned}
E_2 := & \eta^{1/2} \left(\frac{\eta}{\mu(1 - \frac{\eta L}{2})} + \frac{(1 + \eta L)^2}{\mu^2(1 - \frac{\eta L}{2})^2} \right)^{1/2} \cdot \left(\frac{3L^2D^2\eta\delta^{-2}}{N(1 - \delta^{\frac{1}{B}})^2} + \frac{6dL^2\cdot\delta^{-2}}{1 - \delta^{\frac{2}{B}}} \right)^{1/2} \\
& + \frac{\sqrt{\eta}\sigma}{\sqrt{\mu(1 - \frac{\eta L}{2})N}} + \left(\frac{\left(\delta^{\frac{2}{B}}\right)^k - \left(1 - \eta\mu\left(1 - \frac{\eta L}{2}\right)\right)^k}{\delta^{\frac{2}{B}} - 1 + \eta\mu\left(1 - \frac{\eta L}{2}\right)} \right)^{1/2} \cdot \frac{\sqrt{3}L\cdot\delta^{-1}}{\sqrt{N}} \delta^{\frac{1}{B}} \cdot \|x^{(0)}\|_{L_2}, \quad (\text{B.2})
\end{aligned}$$

$$E_3 := \frac{\sqrt{3}\delta^{-1}\delta^{\frac{k}{B}}}{\sqrt{N}} \|x^{(0)}\|_{L_2} + \frac{\sqrt{3}D\eta\delta^{-1}}{\sqrt{N}(1 - \delta^{\frac{1}{B}})} + \frac{\sqrt{6d\eta}\delta^{-1}}{\sqrt{1 - \delta^{\frac{2}{B}}}}, \quad (\text{B.3})$$

and we have from Theorem 3.5 that $\frac{1}{N} \sum_{i=1}^N \mathcal{W}_2(\text{Law}(x_i^{(k)}), \pi) \leq E_1 + E_2 + E_3$. Using $D\sqrt{\eta} \leq \bar{D}$ from Lemma 3.6, we can compute that

$$\begin{aligned}
E_2 \leq & \eta^{1/2} \left(\left(\frac{\eta}{\mu(1 - \frac{\eta L}{2})} \right)^{1/2} + \left(\frac{(1 + \eta L)^2}{\mu^2(1 - \frac{\eta L}{2})^2} \right)^{1/2} \right) \cdot \left(\left(\frac{3L^2\bar{D}^2\delta^{-2}}{N(1 - \delta^{\frac{1}{B}})^2} \right)^{1/2} + \left(\frac{6dL^2\cdot\delta^{-2}}{1 - \delta^{\frac{2}{B}}} \right)^{1/2} \right) \\
& + \frac{\sqrt{\eta}\sigma}{\sqrt{\mu(1 - \frac{\eta L}{2})N}} + \left(\frac{\left(\delta^{\frac{2}{B}}\right)^k - \left(1 - \eta\mu\left(1 - \frac{\eta L}{2}\right)\right)^k}{\delta^{\frac{2}{B}} - \left(1 - \eta\mu\left(1 - \frac{\eta L}{2}\right)\right)} \right)^{1/2} \cdot \frac{\sqrt{3}L\cdot\delta^{-1}}{\sqrt{N}} \delta^{\frac{1}{B}} \cdot \|x^{(0)}\|_{L_2} \quad (\text{B.4})
\end{aligned}$$

$$\begin{aligned}
\leq & \eta^{1/2} \left(\left(\frac{\eta}{\mu(1 - \frac{\eta L}{2})} \right)^{1/2} + \left(\frac{(1 + \eta L)^2}{\mu^2(1 - \frac{\eta L}{2})^2} \right)^{1/2} \right) \cdot \left(\left(\frac{3L^2\bar{D}^2\delta^{-2}}{N(1 - \delta^{\frac{1}{B}})^2} \right)^{1/2} + \left(\frac{6dL^2\cdot\delta^{-2}}{1 - \delta^{\frac{2}{B}}} \right)^{1/2} \right) \\
& + \frac{\sqrt{\eta}\sigma}{\sqrt{\mu(1 - \frac{\eta L}{2})N}} + \left(\frac{\left(\delta^{\frac{2}{B}}\right)^k - \left(1 - \frac{\eta\mu}{1.5}\right)^k}{\delta^{\frac{2}{B}} - \left(1 - \frac{\eta\mu}{1.5}\right)} \right)^{1/2} \cdot \frac{\sqrt{3}L\cdot\delta^{-1}}{\sqrt{N}} \delta^{\frac{1}{B}} \cdot \|x^{(0)}\|_{L_2}, \quad (\text{B.5})
\end{aligned}$$

where in the first inequality we used $(a+b)^{1/2} \leq a^{1/2} + b^{1/2}$ for $a, b \geq 0$ and in the second inequality we used the fact that function

$$H_k(a, b) := \sum_{i=0}^{k-1} a^{k-1-i} b^i = \begin{cases} (a^k - b^k)/(a - b) & \text{if } a \neq b, \\ \frac{1-a^k}{1-a} & \text{if } a = b, \end{cases}$$

defined for $a, b \in [0, 1]$ is non-decreasing in the variable b , i.e. $H_k(a_1, b_1) \leq H_k(a_1, b_2)$ if $b_1 \leq b_2$ for any $a_1 \in [0, 1]$ and $k \geq 0$. More specifically, we used $H_k(a_1, b_1) \leq H_k(a_1, b_2)$ with $a_1 = \delta^{2/B}$ and $b_1 = 1 - \eta\mu(1 - \frac{\eta L}{2}) < b_2 = 1 - \frac{\eta\mu}{1.5}$. The latter (strict) inequality is due to the fact that we have $\eta \leq \frac{1}{2L}$ implied by the condition (3.16) with $\alpha = 1$. Also, for this choice of a_1 and b_2 , since $\eta \leq \bar{\eta}$, we have $a_1 = \delta^{2/B} < b_2 = 1 - \frac{\eta\mu}{1.5}$ and this implies $H_k(a_1, b_2) \leq b_2^k/(b_2 - a_1)$ and (B.5) becomes

$$E_2 \leq \eta^{1/2} \left(\left(\frac{\eta}{\mu(1 - \frac{\eta L}{2})} \right)^{1/2} + \left(\frac{(1 + \eta L)^2}{\mu^2(1 - \frac{\eta L}{2})^2} \right)^{1/2} \right) \cdot \left(\left(\frac{3L^2 \bar{D}^2 \delta^{-2}}{N(1 - \delta^{\frac{1}{B}})^2} \right)^{1/2} + \left(\frac{6dL^2 \cdot \delta^{-2}}{1 - \delta^{\frac{2}{B}}} \right)^{1/2} \right) \\ + \frac{2\sqrt{\eta}\sigma}{\sqrt{3} \cdot \mu N} + \left(\frac{(1 - \frac{\eta\mu}{1.5})^k}{(1 - \frac{\eta\mu}{1.5}) - \delta^{\frac{2}{B}}} \right)^{1/2} \cdot \frac{\sqrt{3}L \cdot \delta^{-1}}{\sqrt{N}} \delta^{\frac{1}{B}} \cdot \|x^{(0)}\|_{L_2}, \quad (\text{B.6})$$

where in the last line we used $1 - \eta L/2 \geq 3/4$ due to the fact that $\eta \leq 1/(2L)$. Similarly, using $D\sqrt{\eta} \leq \bar{D}$, we have the bounds

$$E_3 \leq \frac{\sqrt{3}\delta^{-1}\delta^{\frac{k}{B}}}{\sqrt{N}} \|x^{(0)}\|_{L_2} + \frac{\sqrt{3} \bar{D} \sqrt{\eta} \delta^{-1}}{\sqrt{N}(1 - \delta^{\frac{1}{B}})} + \frac{\sqrt{6d\eta} \delta^{-1}}{\sqrt{1 - \delta^{\frac{2}{B}}}}. \quad (\text{B.7})$$

Note that $D = \Theta(1/\sqrt{\eta})$ as $\eta \rightarrow 0$; this stems from the fact that $\bar{\omega}_3 = \Theta(1/\sqrt{\eta})$ as $\eta \rightarrow 0$, whereas all the other terms that appear in the definition of \bar{D} stays bounded as $\eta \rightarrow 0$. Therefore, we have $\bar{D} = \Theta(1)$. On the other hand, based on (B.6) and (B.7), and the definitions of E_1 , we can control the total error $E_1 + E_2 + E_3$ as

$$E_1 + E_2 + E_3 \leq C_1 (1 - \eta\mu)^k + C_2 \left(\sqrt{1 - \frac{\eta\mu}{1.5}} \right)^k + C_3 \sqrt{\eta} + C_4 \eta, \quad (\text{B.8})$$

where

$$C_1 = \left(\left(\mathbb{E} \left\| \bar{x}^{(0)} - x_* \right\|^2 \right)^{1/2} + \sqrt{2\mu^{-1}dN^{-1}} \right), \quad (\text{B.9})$$

$$C_2 = \frac{1}{\sqrt{1 - \frac{\eta\mu}{1.5} - \delta^{\frac{2}{B}}}} \frac{\sqrt{3}L \cdot \delta^{-1}}{\sqrt{N}} \delta^{\frac{1}{B}} \cdot \|x^{(0)}\|_{L_2} + \frac{\sqrt{3}\delta^{-1}}{\sqrt{N}} \|x^{(0)}\|_{L_2}, \quad (\text{B.10})$$

$$C_3 = \frac{1.65L}{\mu} \sqrt{dN^{-1}} + \frac{\sqrt{6d}\delta^{-1}}{\sqrt{1 - \delta^{\frac{2}{B}}}} + \frac{2\sigma}{\sqrt{3\mu N}} + \left(\frac{(1 + \eta L)^2}{\mu^2(1 - \frac{\eta L}{2})^2} \right)^{1/2} \cdot \left(\frac{6dL^2 \cdot \delta^{-2}}{1 - \delta^{\frac{2}{B}}} \right)^{1/2} \\ + \frac{\sqrt{3}\delta^{-1}}{\sqrt{N}(1 - \delta^{\frac{1}{B}})} \bar{D} + \left(\frac{(1 + \eta L)^2}{\mu^2(1 - \frac{\eta L}{2})^2} \right)^{1/2} \cdot \left(\frac{3L^2 \delta^{-2}}{N(1 - \delta^{\frac{1}{B}})^2} \right)^{1/2} \bar{D}, \quad (\text{B.11})$$

$$C_4 = \left(\frac{1}{\mu(1 - \frac{\eta L}{2})} \right)^{1/2} \cdot \left(\frac{6dL^2 \cdot \delta^{-2}}{1 - \delta^{\frac{2}{B}}} \right)^{1/2} + \left(\frac{1}{\mu(1 - \frac{\eta L}{2})} \right)^{1/2} \cdot \left(\frac{3L^2 \delta^{-2}}{N(1 - \delta^{\frac{1}{B}})^2} \right)^{1/2} \bar{D}, \quad (\text{B.12})$$

where in deriving the constant C_2 given in (B.10) we used the fact that $\sqrt{a_1} = \delta^{1/B} < \sqrt{b_2} = \sqrt{1 - \frac{\eta\mu}{1.5}}$. Furthermore, $C_i = \mathcal{O}(1)$ as $\eta \rightarrow 0$ for $i = 1, 2, 3, 4$ because $\bar{D} = \Theta(1)$ as $\eta \rightarrow 0$. While C_1 does not depend on η , the other bounds C_2, C_3 and C_4 depend on η . To simplify this dependence, we can use the inequalities $\eta \leq 1/(2L)$ and $1 - \frac{\eta L}{2} \geq \frac{3}{4}$ again and this yields

$$C_3 \leq \bar{C}_3 = \frac{1.65L}{\mu} \sqrt{dN^{-1}} + \frac{\sqrt{6d}\delta^{-1}}{\sqrt{1 - \delta^{\frac{2}{B}}}} + \frac{2\sigma}{\sqrt{3\mu N}} + \frac{2}{\mu} \cdot \left(\frac{6dL^2 \cdot \delta^{-2}}{1 - \delta^{\frac{2}{B}}} \right)^{1/2}$$

$$+ \frac{\sqrt{3}\delta^{-1}}{\sqrt{N}(1-\delta^{\frac{1}{B}})}\bar{D} + \frac{2}{\mu} \cdot \left(\frac{3L^2\delta^{-2}}{N(1-\delta^{\frac{1}{B}})^2} \right)^{1/2} \bar{D},$$

and

$$C_4 \leq \bar{C}_4 = \frac{2}{\sqrt{3}\mu} \cdot \left(\frac{6dL^2 \cdot \delta^{-2}}{1-\delta^{\frac{2}{B}}} \right)^{1/2} + \frac{2}{\sqrt{\mu}} \cdot \left(\frac{3L^2\delta^{-2}}{N(1-\delta^{\frac{1}{B}})^2} \right)^{1/2} \bar{D}.$$

On the other hand, using $\delta^{1/B} < 1 - \frac{\eta\mu}{1.5} \leq 1 - \frac{\eta\mu}{1.5}$ for $\eta \leq \bar{\eta}$, we get $C_2 \leq \bar{C}_2$. Noting that $\bar{C}_1 = C_1$, from (B.8), we get

$$E_1 + E_2 + E_3 \leq (\bar{C}_1 + \bar{C}_2) \left(\sqrt{1 - \frac{\eta\mu}{1.5}} \right)^k + \bar{C}_3\sqrt{\eta} + \bar{C}_4\eta.$$

For the desired target error, it suffices that $(\bar{C}_1 + \bar{C}_2) \left(\sqrt{1 - \frac{\eta\mu}{1.5}} \right)^k \leq \epsilon/3$, $\bar{C}_3\sqrt{\eta} \leq \epsilon/3$, and $\bar{C}_4\eta \leq \epsilon/3$. Using $\left(\sqrt{1 - \frac{\eta\mu}{1.5}} \right)^k \leq \exp(-\eta\mu k/3)$, these conditions are satisfied whenever

$$\eta \leq \eta_{\text{noise}}(\epsilon) := \min \left(\frac{\epsilon^2}{9 \cdot \bar{C}_3^2}, \frac{\epsilon}{3 \cdot \bar{C}_4} \right) \quad \text{and} \quad \eta k \geq \frac{3}{\mu} \log \left(\frac{3(\bar{C}_1 + \bar{C}_2)}{\epsilon} \right).$$

By assumption, $\eta \leq \bar{\eta}$. Therefore, if we let $\eta = \eta_* := \min(\bar{\eta}, \eta_{\text{noise}}(\epsilon))$, then we obtain that after $k \geq \frac{3}{\mu\eta_*} \log \left(\frac{3(\bar{C}_1 + \bar{C}_2)}{\epsilon} \right)$ iterations, we have $E_1 + E_2 + E_3 \leq \epsilon$. This completes the proof.

C Additional Technical Lemmas

Lemma C.1 (Lemma 3.4 in [42]). *Under Assumption 3.2, for any $k = B-1, B, \dots$ and any Nd -dimensional vector b , we have $\|\mathcal{L}_N \mathcal{W}_B^{(k)} b\| \leq \delta(k) \|\mathcal{L}_N b\|$, where $\delta(k)$ is defined in Assumption 3.2 and $\mathcal{L}_N = I_{Nd} - \frac{1}{N} \left((1_N 1_N^\top) \otimes I_d \right)$.*

Next, we consider $\min_{x \in \mathbb{R}^d} g(x) := \frac{1}{N} \sum_{i=1}^N g_i(x)$, where g_i are μ -strongly convex and L -smooth. Consider the iterates:

$$p^{(k+1)} = p^{(k)} - \eta \frac{1}{N} \sum_{i=1}^N \nabla g_i \left(s_i^{(k)} \right). \quad (\text{C.1})$$

Then, we have the following technical lemma.

Lemma C.2 (Lemma 3.12 in [42]). *Assume $\sqrt{1 - \frac{\eta\mu\beta}{\beta+1}} \leq \lambda < 1$ and $\eta \leq \frac{1}{(1+\alpha)L}$, where $\alpha, \beta > 0$. Then,*

$$\|p - p_*\|^{\lambda, K} \leq 2 \|p^{(0)} - p_*\| + \left(\lambda \sqrt{N} \right)^{-1} \left(\sqrt{\frac{L(1+\alpha)}{\mu\alpha}} + \beta \right) \sum_{i=1}^N \|p - s_i\|^{\lambda, K}, \quad (\text{C.2})$$

for any $K = 0, 1, 2, \dots$ where p_* is the minimizer of g , where $g := \frac{1}{N} \sum_{i=1}^N g_i(x)$ and the sequence $p^{(k)}$ follows the recursion (C.1).

Lemma C.3. *For any function $f \in \mathcal{S}_{\mu, L}(\mathbb{R}^d)$, the gradient operator $\nabla f : \mathbb{R}^d \rightarrow \mathbb{R}^d$ is surjective, i.e. for every $v \in \mathbb{R}^d$, there exists some $x \in \mathbb{R}^d$ such that $\nabla f(x) = v$.*

Proof. This is a direct consequence of [16, Theorem 1]. Indeed, for $f \in \mathcal{S}_{\mu,L}(\mathbb{R}^d)$, the gradient operator ∇f is strongly coercive around x_* , i.e. it satisfies $\langle \nabla f(x) - \nabla f(x_*), x - x_* \rangle \geq \mu \|x - x_*\|^p$, for $p = 2$. As a consequence, it is coercive, i.e. $\frac{\langle \nabla f(x), x \rangle}{\|x\|} \rightarrow \infty$, as $\|x\| \rightarrow \infty$. It is also a bounded operator, i.e. $\|\nabla f(x) - \nabla f(x_*)\| \leq L\|x - x_*\|$. Finally, it is a proper operator, i.e. the preimage $\nabla f^{-1}(K)$ is a compact subset of \mathbb{R}^d whenever $K \subset \mathbb{R}^d$ is compact. To see this, note that by strong convexity, $\|\nabla f(x) - \nabla f(x_*)\| \geq \mu\|x - x_*\|$; hence, the preimage $\nabla f^{-1}(K)$ of a compact set K should be bounded. In addition, because K is closed, such a preimage should also be closed by the continuity of ∇f which implies that $\nabla f^{-1}(K)$ is indeed compact. Therefore, [16, Theorem 1] is applicable and this completes the proof. \square

Lemma C.4. *In the setting of Theorem 3.4, let $\alpha = 1$ and $\beta = 2L/\mu$, and assume $\eta \in (0, \bar{\eta}]$ where*

$$\bar{\eta} := \frac{3(1 - \delta^2)}{\mu J_1} \quad \text{with} \quad J_1 := 3\kappa B^2 \left(1 + 4\sqrt{N}\sqrt{\kappa}\right) \quad \text{with} \quad \kappa := \frac{L}{\mu}. \quad (\text{C.3})$$

Then, there exists $\lambda(\eta) \in [\underline{\lambda}, 1) \subsetneq (\delta^{1/B}, 1)$ such that the conditions (3.16) hold where

$$\underline{\lambda} := \sqrt[2B]{1 - \frac{(\sqrt{J_1^2 + (1 - \delta^2)J_1} - \delta J_1)^2}{J_1(J_1 + 1)^2}} = \sqrt[B]{\sqrt{\frac{\check{\eta}\mu J_1}{1.5}} + \delta} = \left(\frac{\sqrt{J_1^2 + (1 - \delta^2)J_1} + \delta}{J_1 + 1} \right)^{\frac{1}{B}}, \quad (\text{C.4})$$

with $\check{\eta} := \frac{1.5(\sqrt{J_1^2 + (1 - \delta^2)J_1} - \delta J_1)^2}{\mu J_1(J_1 + 1)^2} > 0$. Furthermore, we can take

$$\lambda(\eta) = \begin{cases} \sqrt[2B]{1 - \frac{\eta\mu}{1.5}}, & \text{if } \eta \in (0, \check{\eta}); \\ \sqrt[B]{\sqrt{\frac{\eta\mu J_1}{1.5}} + \delta}, & \text{if } \eta \in (\check{\eta}, \bar{\eta}]. \end{cases} \quad (\text{C.5})$$

Therefore, Theorems 3.4 and 3.5 are applicable when $\eta \in (0, \bar{\eta}]$.

Proof. The conditions (3.16) also appear in the context of the DIGing algorithm proposed in the optimization setting in [42]. By following similar steps to the proof of Theorem 3.14 from [42] we can see that the following λ given in (C.5) satisfies (3.16) and $\lambda \in (0, 1)$. Note that $\lambda = \lambda(\eta)$ is piecewise defined where it is straightforward to check that it is continuous on the interval $\eta \in (0, \bar{\eta}]$ admitting a minimum at $\check{\eta}$, i.e.

$$\begin{aligned} \inf_{\eta \in (0, \bar{\eta}]} \lambda(\eta) &= \lambda(\check{\eta}) = \sqrt[2B]{1 - \frac{\check{\eta}\mu}{1.5}} = \sqrt[B]{\sqrt{\frac{\check{\eta}\mu J_1}{1.5}} + \delta} \\ &= \sqrt[2B]{1 - \frac{(\sqrt{J_1^2 + (1 - \delta^2)J_1} - \delta J_1)^2}{J_1(J_1 + 1)^2}} = \left(\frac{\sqrt{J_1^2 + (1 - \delta^2)J_1} + \delta}{J_1 + 1} \right)^{\frac{1}{B}}, \end{aligned} \quad (\text{C.6})$$

where we also see that $\underline{\lambda} = \lambda(\check{\eta}) > \delta^{1/B}$. This proves (C.4) and completes the proof. \square

References

- [1] Yossi Arjevani, Joan Bruna, Bugra Can, Mert Gürbüzbalaban, Stefanie Jegelka, and Hongzhou Lin. IDEAL: Inexact DEcentralized accelerated augmented Lagrangian method. In *Advances in Neural Information Processing Systems*, volume 33, pages 20648–20659. Curran Associates, Inc., 2020.

- [2] Necdet Serhat Aybat and Mert Gürbüzbalaban. Decentralized computation of effective resistances and acceleration of consensus algorithms. In *2017 IEEE Global Conference on Signal and Information Processing (GlobalSIP)*, pages 538–542. IEEE, 2017.
- [3] Krishna Balasubramanian, Sinho Chewi, Murat A Erdogdu, Adil Salim, and Shunshi Zhang. Towards a theory of non-log-concave sampling: First-order stationarity guarantees for Langevin Monte Carlo. In *Proceedings of Thirty Fifth Conference on Learning Theory*, volume 178, pages 2896–2923. PMLR, 2022.
- [4] Sergio Barbarossa and Gesualdo Scutari. Decentralized maximum-likelihood estimation for sensor networks composed of nonlinearly coupled dynamical systems. *IEEE Transactions on Signal Processing*, 55(7):3456–3470, 2007.
- [5] Andrei-Cristian Barbos, François Caron, Jean-François Giovannelli, and Arnaud Doucet. Clone MCMC: parallel high-dimensional Gaussian Gibbs sampling. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017.
- [6] Kinjal Bhar, He Bai, Jemin George, and Carl Busart. Scalability enhancement and data-heterogeneity awareness in gradient tracking based decentralized Bayesian learning. In *Proceedings of the 7th Annual Learning for Dynamics & Control Conference*, volume 283, pages 591–605. PMLR, 04–06 Jun 2025.
- [7] 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.
- [8] Stephen Boyd, Persi Diaconis, and Lin Xiao. Fastest mixing Markov chain on a graph. *SIAM Review*, 46(4):667–689, 2004.
- [9] Stephen P. Brooks. Markov chain Monte Carlo method and its application. *Journal of the Royal Statistical Society: Series D (The Statistician)*, 47(1):69–100, 1998.
- [10] Nicolas Brosse, Alain Durmus, and Eric Moulines. The promises and pitfalls of stochastic gradient Langevin dynamics. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc., 2018.
- [11] 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.
- [12] Trevor Campbell and Jonathan P How. Approximate decentralized Bayesian inference. In *Proceedings of the Thirtieth Conference on Uncertainty in Artificial Intelligence*, pages 102–111, 2014.
- [13] Bugra Can, Saeed Soori, Necdet Serhat Aybat, Maryam Mehri Dehnavi, and Mert Gürbüzbalaban. Randomized gossiping with effective resistance weights: Performance guarantees and applications. *IEEE Transactions on Control of Network Systems*, 9(2):524–536, 2022.
- [14] Xiang Chen, Jason D. Lee, Tongzheng Li, and Mingyi Wang. Decentralized stochastic gradient Langevin dynamics for Bayesian learning. *IEEE Transactions on Signal Processing*, 66(17):4760–4775, 2018.

- [15] Sinho Chewi, Murat A Erdogdu, Mufan (Bill) Li, Ruoqi Shen, and Matthew Zhang. Analysis of Langevin Monte Carlo from Poincaré to log-Sobolev. *Foundations of Computational Mathematics*, 25:1345–1395, 2025.
- [16] Raffaele Chiappinelli and David E Edmunds. Remarks on surjectivity of gradient operators. *Mathematics*, 8(9):1538, 2020.
- [17] Arkabandhu Chowdhury and Christopher Jermaine. Parallel and distributed MCMC via shepherding distributions. In *International Conference on Artificial Intelligence and Statistics*, volume 84, pages 1819–1827. PMLR, 2018.
- [18] Paolo Dai Pra, Benedetto Scoppola, and Elisabetta Scoppola. Sampling from a Gibbs measure with pair interaction by means of PCA. *Journal of Statistical Physics*, 149(4):722–737, 2012.
- [19] 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.
- [20] Arnak S Dalalyan and Avetik Karagulyan. User-friendly guarantees for the Langevin Monte Carlo with inaccurate gradient. *Stochastic Processes and their Applications*, 129(12):5278–5311, 2019.
- [21] Paolo Di Lorenzo and Gesualdo Scutari. Next: In-network nonconvex optimization. *IEEE Transactions on Signal and Information Processing over Networks*, 2(2):120–136, 2016.
- [22] Alain Durmus and Eric Moulines. Non-asymptotic convergence analysis for the unadjusted Langevin algorithm. *Annals of Applied Probability*, 27(3):1551–1587, 2017.
- [23] Alain Durmus and Eric Moulines. High-dimensional Bayesian inference via the Unadjusted Langevin Algorithm. *Bernoulli*, 25(4A):2854–2882, 2019.
- [24] Khaoula El Mekkaoui, Diego Mesquita, Paul Blomstedt, and Samuel Kaski. Federated stochastic gradient Langevin dynamics. In *Uncertainty in Artificial Intelligence*, volume 161, pages 1703–1712. PMLR, 2021.
- [25] Murat A. Erdogdu and Rasa Hosseinzadeh. On the convergence of Langevin Monte Carlo: The interplay between tail growth and smoothness. In *Proceedings of Thirty Fourth Conference on Learning Theory*, volume 134, pages 1776–1822. PMLR, 2021.
- [26] Alireza Fallah, Mert Gürbüzbalaban, Asuman Ozdaglar, Umut Şimşekli, and Lingjiong Zhu. Robust distributed accelerated stochastic gradient methods for multi-agent networks. *Journal of Machine Learning Research*, 23(220):1–96, 2022.
- [27] Cheng Fang, Rishabh Dixit, Waheed U Bajwa, and Mert Gurbuzbalaban. RESIST: Resilient decentralized learning using consensus gradient descent. *arXiv preprint arXiv:2502.07977*, 2025.
- [28] Andrew Gelman, John B. Carlin, Hal S. Stern, David B. Dunson, Aki Vehtari, and Donald B. Rubin. *Bayesian Data Analysis*. CRC Press, 3 edition, 2013.
- [29] Clark R Givens and Rae Michael Shortt. A class of Wasserstein metrics for probability distributions. *Michigan Mathematical Journal*, 31(2):231–240, 1984.

- [30] Mert Gürbüzbalaban, Xuefeng Gao, Yuanhan Hu, and Lingjiong Zhu. Decentralized stochastic gradient Langevin dynamics and Hamiltonian Monte Carlo. *The Journal of Machine Learning Research*, 22(1):10804–10872, 2021.
- [31] Mert Gürbüzbalaban, Mohammad Rafiqul Islam, Xiaoyu Wang, and Lingjiong Zhu. Generalized EXTRA stochastic gradient Langevin dynamics. *arXiv:2412.01993*, 2024.
- [32] Peter D Hoff. *A First Course in Bayesian Statistical Methods*, volume 580. Springer, 2009.
- [33] Yuni Iwamasa and Naoki Masuda. Networks maximizing the consensus time of voter models. *Physical Review E*, 90:012816, Jul 2014.
- [34] Dušan Jakovetić. A unification and generalization of exact distributed first-order methods. *IEEE Transactions on Signal and Information Processing over Networks*, 5(1):31–46, 2018.
- [35] Alexander Kolesov and Vyacheslav Kungurtsev. Decentralized Langevin dynamics over a directed graph. *arXiv preprint arXiv:2103.05444*, 2021.
- [36] Vyacheslav Kungurtsev, Adam Cobb, Tara Javidi, and Brian Jalaian. Decentralized Bayesian learning via Metropolis-adjusted Hamiltonian Monte Carlo. *Machine Learning*, 112(7):2695–2724, 2023.
- [37] Huaqing Li, Lifeng Zheng, Zheng Wang, Yu Yan, Liping Feng, and Jing Guo. S-DIGing: A stochastic gradient tracking algorithm for distributed optimization. *IEEE Transactions on Emerging Topics in Computational Intelligence*, 6(1):53–65, 2020.
- [38] Jiajun Liang, Qian Zhang, Wei Deng, Qifan Song, and Guang Lin. Bayesian federated learning with Hamiltonian Monte Carlo: Algorithm and theory. *Journal of Computational and Graphical Statistics*, 34(2):509–518, 2025.
- [39] H. Brendan McMahan, Eider Moore, Daniel Ramage, Seth Hampson, and Blaise Agüera y Arcas. Communication-efficient learning of deep networks from decentralized data. In *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics (AISTATS)*, volume 54, pages 1273–1282, 2017.
- [40] Angelia Nedić and Alex Olshevsky. Distributed optimization over time-varying directed graphs. *IEEE Transactions on Automatic Control*, 60(3):601–615, 2015.
- [41] Angelia Nedić, Alex Olshevsky, and Michael G Rabbat. Network topology and communication-computation tradeoffs in decentralized optimization. *Proceedings of the IEEE*, 106(5):953–976, 2018.
- [42] Angelia Nedić, Alex Olshevsky, and Wei Shi. Achieving geometric convergence for distributed optimization over time-varying graphs. *SIAM Journal on Optimization*, 27(4):2597–2633, 2017.
- [43] Angelia Nedić and Asuman Ozdaglar. Distributed subgradient methods for multi-agent optimization. *IEEE Transactions on Automatic Control*, 54(1):48–61, 2009.
- [44] Matthew Nokleby, Haroon Raja, and Waheed U. Bajwa. Scaling-up distributed processing of data streams for machine learning. *Proceedings of the IEEE*, 108(11):1984–2012, 2020.
- [45] Anjaly Parayil, He Bai, Jemin George, and Prudhvi Gurram. Decentralized Langevin dynamics for Bayesian learning. In *Advances in Neural Information Processing Systems*, volume 33, pages 15978–15989. Curran Associates, Inc., 2020.

- [46] Grigorios A Pavliotis. *Stochastic Processes and Applications: Diffusion processes, the Fokker-Planck and Langevin Equations*, volume 60. Springer, 2014.
- [47] Maxim Raginsky, Alexander Rakhlin, and Matus Telgarsky. Non-convex learning via stochastic gradient Langevin dynamics: A nonasymptotic analysis. In *Conference on Learning Theory*, volume 65, pages 1674–1703. PMLR, 2017.
- [48] Lewis J. Rendell, Adam M. Johansen, Anthony Lee, and Nick Whiteley. Global consensus Monte Carlo. *Journal of Computational and Graphical Statistics*, 30(2):249–259, 2021.
- [49] Ioannis D. Schizas, Alejandro Ribeiro, and Georgios B. Giannakis. Consensus in ad hoc WSNs with noisy links—part i: Distributed estimation of deterministic signals. *IEEE Transactions on Signal Processing*, 56(1):350–364, 2008.
- [50] Gesualdo Scutari and Ying Sun. Distributed nonconvex constrained optimization over time-varying digraphs. *Mathematical Programming*, 176(1):497–544, 2019.
- [51] Wei Shi, Qing Ling, Gang Wu, and Wotao Yin. EXTRA: An exact first-order algorithm for decentralized consensus optimization. *SIAM Journal on Optimization*, 25(2):944–966, 2015.
- [52] Lukang Sun, Adil Salim, and Peter Richtárik. Federated sampling with Langevin algorithm under isoperimetry. *Transactions on Machine Learning Research*, pages 1–29, 2024.
- [53] Brian Swenson, Soumya Kar, H. Vincent Poor, José M. F. Moura, and Aaron Jaech. Distributed gradient methods for nonconvex optimization: Local and global convergence guarantees. *arXiv:2003.10309*, 2020.
- [54] 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, Barcelona, Spain, 2020. IEEE.
- [55] Alireza Tahbaz-Salehi and Ali Jadbabaie. A necessary and sufficient condition for consensus over random networks. *IEEE Transactions on Automatic Control*, 53(3):791–795, 2008.
- [56] 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.
- [57] Cédric Villani. *Optimal Transport: Old and New*. Springer, Berlin, 2009.
- [58] Maxime Vono, Nicolas Dobigeon, and Pierre Chainais. Split-and-augmented Gibbs sampler—application to large-scale inference problems. *IEEE Transactions on Signal Processing*, 67(6):1648–1661, 2019.
- [59] Maxime Vono, Daniel Paulin, and Arnaud Doucet. Efficient MCMC sampling with dimension-free convergence rate using ADMM-type splitting. *Journal of Machine Learning Research*, 23(25):1–69, 2022.
- [60] Hao Wang and Dit-Yan Yeung. A survey on Bayesian deep learning. *ACM Computing Surveys*, 52(5):1–37, 2020.

- [61] 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.
- [62] William Wolberg, Olvi Mangasarian, Nick Street, and W. Street. Breast cancer wisconsin (diagnostic). UCI Machine Learning Repository, 1993.
- [63] Lin Xiao, Stephen Boyd, and Seung-Jean Kim. Distributed average consensus with least-mean-square deviation. *Journal of Parallel and Distributed Computing*, 67(1):33–46, 2007.
- [64] 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.
- [65] Kun Yuan, Qing Ling, and Wotao Yin. On the convergence of decentralized gradient descent. *SIAM Journal on Optimization*, 26(3):1835–1854, 2016.

Acknowledgments

This paper is dedicated to the memory of Wei (Wilbur) Shi, a co-author of [42], that introduced the DIGing algorithm in the context of distributed optimization. He left us far too soon.