

DIGing stochastic gradient Langevin dynamics

Student¹, Waheed Bajwa², Mert Gürbüzbalaban³, Lingjiong Zhu⁴

April 18, 2025

Abstract

TBD.

1 Introduction

MG: Rewrite the intro, as it is currently too similar to a previous paper on EXTRA. Check if we need to cite <https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=9186344>

In our era of big data, the amount of data collected and stored has seen exponential growth with ever-increasing rates. Since the rate at which data is generated is often beyond our capability to analyze, for example, due to the constraint of the available computational resources, there has been a growing interests for developing scaleable machine learning algorithms that are efficient on large datasets. Very often, because of communication constraints and privacy constraints, gathering all these data for centralized processing is often impractical or infeasible. Decentralized machine learning algorithms have received a lot of attention for such applications where agents can collaboratively learn a predictive model without sharing their own data but sharing only their local models with their immediate neighbors at some frequency to generate a global model; see e.g. [HBJ18, HBM19, ABC⁺20].

Although there is a large body of literature on scaleable first-order decentralized learning methods have been proposed in the literature such as decentralized stochastic approximation and optimization algorithms (see e.g. [ULGN17, GDG19, SBB⁺19, Ned20]), very few of them deal with decentralized Bayesian learning (inference) [PBG20, GGHZ21]. Let us introduce the problem of decentralized Bayesian inference. Assume there are N agents connected over a network $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ with $\mathcal{V} = \{1, 2, \dots, N\}$ representing the agents and $\mathcal{E} \subseteq \mathcal{V} \times \mathcal{V}$ being the set of edges; i.e. i and j are connected if $(i, j) \in \mathcal{E}$ where the network is undirected, i.e. $(i, j) \in \mathcal{E}$ then $(j, i) \in \mathcal{E}$. **do we need to introduce \mathcal{E} here? it seems this problem is independent of \mathcal{E} , only later the algorithm depends on \mathcal{E} . pls double check.** Let $Z = [z_1, \dots, z_n]$ be a dataset consisting of n independent and identically distributed (i.i.d.) data vectors sampled from a parametrized distribution $p(Z|x)$ where the parameter $x \in \mathbb{R}^d$ has a common prior distribution $p(x)$. Due to the decentralization in the data collection, each agent i possesses a subset Z_i of the data where $Z_i = \{z_1^i, z_2^i, \dots, z_{n_i}^i\}$ and n_i is the number of samples of the agent i . The data is held disjointly over agents; i.e. $Z = \bigcup_{i=1}^N Z_i$ with $Z_i \cap Z_j = \emptyset$ for any $j \neq i$. The goal is to sample from the posterior distribution

¹Department of Electrical and Computer Engineering, Rutgers University, Piscataway, New Jersey, United States of America; yahya.ayach@rutgers.edu

²Department of Electrical and Computer Engineering, Rutgers University, Piscataway, New Jersey, United States of America; waheed.bajwa@rutgers.edu

³Department of Management Science and Information Systems, Rutgers Business School, Piscataway, New Jersey, United States of America; mg1366@rutgers.edu

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

$p(x|Z) \propto p(Z|x)p(x)$. Since the data points are independent, the log-likelihood function will be additive; $\log p(Z|x) = \sum_{i=1}^N \sum_{j=1}^{n_i} \log p(z_j^i|x)$. Thus, if we set

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

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

Decentralized Langevin algorithms have been proposed in the recent literature that can be used in the large-scale decentralized sampling problems [PBG20, GGHZ21]. However, these works have two restrictions. First, these DE-SGLD algorithms induce a bias at every agent that can negatively impact performance. Such a bias is attributable to network effects and it persists even when using full batches. Second, in these DE-SGLD algorithms, the communication matrix is time-independent such that the network effect is time-homogeneous. Motivated by the EXTRA algorithm and its generalizations for decentralized optimization, generalized EXTRA SGLD was proposed and studied in [GIWZ24], which eliminates this bias in the full-batch setting and addresses the first challenge. However, it does not address the second challenge. To overcome both of these two drawbacks, in this paper, we propose *DIGing stochastic gradient Langevin dynamics* (DIGing SGLD) algorithm that can eliminate the bias in the full-batch setting, and moreover, its the communication matrix is time-dependent. Our algorithm is inspired by the **DIGing** algorithm for distributed optimization problem, which uses **D**istributed **I**nexact **G**radients and gradient tracking [NOS17].

2 Preliminaries and Background

Notations. Let $\mathbf{1}$ denote a column all-one vector. For any matrix v , its average across all agents is defined as $\bar{v} = \frac{1}{N} v^\top \mathbf{1}$ and its consensus violation is denoted as $\tilde{v} = v - \mathbf{1} \bar{v}^\top = v - \frac{1}{N} \mathbf{1} \mathbf{1}^\top v = L_N v$, where $L_N := I - \frac{1}{N} \mathbf{1} \mathbf{1}^\top$ is a symmetric matrix. We denote $\|a\|_{L_N} := \sqrt{\langle a, L_N a \rangle}$ so that $\|a\|_{L_N} = \|L_N a\|_F$. For any vector x , let $\|x\|$ denote its Euclidean norm, and let $\|x\|_{L_2} := (\mathbb{E}\|x\|^2)^{1/2}$ denote its L_2 norm. We need to be careful about the notation. I noticed that in [GGHZ21], $x^{(k)}$ is a big vector in \mathbb{R}^{Nd} , whereas in [NOS17], their $x^{(k)}$ is $N \times d$ matrix. I think it is more convenient to stick with the notation in [GGHZ21].

Decentralized setting. We consider decentralized algorithm where the agent is connected over a connected network by N nodes. We aim to sample from a target distribution with density $\pi(x) \propto e^{-f(x)}$ on \mathbb{R}^d with

$$f(x) := \sum_{i=1}^N f_i(x). \quad (2.1)$$

We make the first assumption on the objective function.

Assumption 1. We assume the objective functions f_i are non-negative and $f_i \in \mathcal{S}_{\mu,L}(\mathbb{R}^d)$ for every $i = 1, 2, \dots, N$, where $\mathcal{S}_{\mu,L}(\mathbb{R}^d)$ denotes the set of functions from \mathbb{R}^d to \mathbb{R} that are μ -strongly convex and L -smooth, that is, for any $g \in \mathcal{S}_{\mu,L}(\mathbb{R}^d)$, for every $x, y \in \mathbb{R}^d$ such that

$$\frac{L}{2} \|x - y\|^2 \geq g(x) - g(y) - \nabla g(y)^\top (x - y) \geq \frac{\mu}{2} \|x - y\|^2, \quad x, y \in \mathbb{R}^d. \quad (2.2)$$

Next, we define the function $F(x) : \mathbb{R}^{Nd} \rightarrow \mathbb{R}$:

$$F(x) := \sum_{i=1}^N f_i(x_i), \quad \text{for any } x := \left(x_1^\top, \dots, x_N^\top\right)^\top \in \mathbb{R}^{Nd}. \quad (2.3)$$

Let $\mathbf{x}_* = [x_*^\top, \dots, x_*^\top]^\top$ be the minimizer, it satisfies conditions: (1). consensus: $\mathbf{x}_* = \mathcal{W}\mathbf{x}_*$, (2). optimality: $\mathbf{1}^\top \nabla F(\mathbf{x}_*) = \sum_{i=1}^N \nabla f_i(x_*) = 0$. For example, decentralized gradient descent (DGD) carries the following iterative algorithm

$$x^{(k+1)} = \mathcal{W}x^{(k)} - \eta \nabla F(x^{(k)}), \quad \mathcal{W} = W \otimes I_d, \quad (2.4)$$

with $x^{(k)} = \left[\left(x_1^{(k)}\right)^\top, \dots, \left(x_N^{(k)}\right)^\top \right]^\top \in \mathbb{R}^{Nd}$ and $k = 0, 1, 2, \dots$, where $x_i^{(k)} \in \mathbb{R}^d$ is the local copy of x by the agent i at the iteration k .

Langevin algorithms. One of the most widely used Markov Chain Monte Carlo methods in statistics are *Langevin algorithms*, that allow one to sample from a given density $\pi(x)$ of interest. The classical one is based on the *overdamped Langevin diffusion*; see e.g. [Dal17, DM19, DM17, DK19, EH21, BCE⁺22, CEL⁺24]:

$$dX(t) = -\nabla f(X(t))dt + \sqrt{2}dW_t, \quad (2.5)$$

where $f : \mathbb{R}^d \rightarrow \mathbb{R}$ and W_t is a standard d -dimensional Brownian motion that starts at zero at time zero. Under some mild assumptions on f , the diffusion (2.5) admits a unique stationary distribution with the density $\pi(x) \propto e^{-f(x)}$, also known as the *Gibbs distribution* [Pav14]. For computational purposes, this diffusion is simulated by considering its discretization. Although various discretization schemes are proposed, Euler-Maruyama discretization is the simplest one and is known as the *unadjusted Langevin algorithm* in the literature [DM17, DM19]:

$$x_{k+1} = x_k - \eta \nabla f(x_k) + \sqrt{2\eta} w_k, \quad (2.6)$$

where $\eta > 0$ is the stepsize parameter, and $w_k \in \mathbb{R}^d$ is a sequence of i.i.d. standard Gaussian random vectors $\mathcal{N}(0, I_d)$. But then the discretized chain (2.6) does not converge to the target π and has a bias that needs to be properly characterized to provide performance guarantees [DK19]. The unadjusted Langevin algorithm (2.6) assumes availability of the gradient ∇f . On the other hand, in many settings in machine learning, computing the full gradient ∇f is either infeasible or impractical. For example, in Bayesian regression or classification problems, f can have a finite-sum form as the sum of many component functions over all the data points and the number of data points can be large (see, e.g., [GGHZ21, XCZG18]). In such settings, algorithms that rely on *stochastic gradients*, i.e., unbiased stochastic estimates of the gradient obtained by a randomized

sampling of the data points, is often more efficient [Bot10]. This fact motivated the development of Langevin algorithms that can support stochastic gradients. In particular, if one replaces the full gradient ∇f in (2.6) by a stochastic gradient, the resulting algorithm is known as the *stochastic gradient Langevin dynamics* (SGLD) (see, e.g., [WT11]). There has been growing recent interest in the non-asymptotic analysis of discretized Langevin diffusions, motivated by applications to large-scale data analysis and Bayesian inference. The discretized Langevin diffusions admit convergence guarantees to a stationary distribution in a variety of metrics and under various assumptions on f ; see e.g. [Dal17, DM17, DM19, CB18, EH22, DK19, BCM⁺21, RRT17, XCZG18, CMR⁺21, VW19, ZADS23, EH21, LZT22, BCE⁺22, CEL⁺24].

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

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

where $\eta > 0$ is the stepsize, W_{ij} are the entries of a doubly stochastic weight matrix W with $W_{ij} > 0$ only if i is connected to j , $w_i^{(k)}$ are independent and identically distributed (i.i.d.) Gaussian random variables with zero mean and identity covariance matrix for every i and k , and $\tilde{\nabla} f_i(x_i^{(k)})$ is an unbiased stochastic estimate of the deterministic gradient $\nabla f_i(x_i^{(k)})$ with a bounded variance. When the number of data points n_i is large, stochastic estimates $\tilde{\nabla} f_i(x)$ are cheaper to compute compared to actual gradients $\nabla f_i(x)$ and can for instance be estimated from a mini-batch of data, i.e. from randomly selected smaller subsets of data. This allows the DE-SGLD method to be scalable to big data settings when n_i can be large. Without the Gaussian noise, the iterations are also equivalent to the decentralized stochastic gradient algorithm [SKP⁺20, FGO⁺22] which has its origins in the decentralized gradient descent (DGD) methods introduced in [NO09].

EXTRA Langevin algorithms. Inspired by the *exact* decentralized optimization literature [SLWY15, Jak18], EXTRA Langevin algorithms are proposed in [GIWZ24]. *EXTRA stochastic gradient Langevin dynamics* (EXTRA SGLD) is defined in [GIWZ24] as:

$$x_i^{(k+2)} = \sum_{j \in \Omega_i} W_{ij} x_j^{(k+1)} - \eta \tilde{\nabla} f_i(x_i^{(k+1)}) + \sqrt{2\eta} w_i^{(k+2)}, \quad (2.8)$$

$$x_i^{(k+1)} = \sum_{j \in \Omega_i} \widetilde{W}_{ij} x_j^{(k+1)} - \eta \tilde{\nabla} f_i(x_i^{(k)}) + \sqrt{2\eta} w_i^{(k+1)}, \quad (2.9)$$

where W_{ij} are the entries of a doubly stochastic weight matrix W , \widetilde{W}_{ij} are the entries of another doubly stochastic weight matrix \widetilde{W} , $w_i^{(k)}$ are standard d -dimensional Gaussian random vectors that are i.i.d. in both $i = 1, 2, \dots, N$ and $k = 1, 2, 3, \dots$, and $\tilde{\nabla} f_i$ are stochastic gradients. Non-asymptotic convergence analysis are obtained in [GIWZ24].

2-Wasserstein distance. Define $\mathcal{P}_2(\mathbb{R}^d)$ as the space consisting of all the Borel probability measures μ on \mathbb{R}^d with the finite second moment (based on the Euclidean norm). For any two Borel probability measures $\mu_1, \mu_2 \in \mathcal{P}_2(\mathbb{R}^d)$, the 2-Wasserstein distance \mathcal{W}_2 (see e.g. [Vil09]) is defined as: $\mathcal{W}_2(\mu_1, \mu_2) := (\inf_{\mathbb{E}} [\|X_1 - X_2\|^2])^{1/2}$, where the infimum is taken over all joint distributions of the random variables X_1, X_2 with marginal distributions μ_1, μ_2 respectively.

3 DIGing Langevin Algorithms for Undirected Graphs

In this section, we propose the *DIGing stochastic gradient Langevin dynamics* (DIGing SGLD) as follows:

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

$$y_i^{(k+1)} = \sum_{j \in \Omega_i(k)} W_{ij}^{(k)} y_j^{(k)} + \tilde{\nabla} f_i(x_i^{(k+1)}) - \tilde{\nabla} f_i(x_i^{(k)}), \quad (3.2)$$

where $w_i^{(k)}$ are standard d -dimensional Gaussian random vectors that are i.i.d. in both $i = 1, 2, \dots, N$ and $k = 1, 2, 3, \dots$, and $\tilde{\nabla} f_i$ are stochastic gradients, and we initialize with $x_i^{(0)}$ and $y_i^{(0)} = \nabla f_i(x_i^{(0)})$. Next, we define the stochastic gradient noise as:

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

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

Assumption 2. For every $i = 1, 2, \dots, N$ and $k = 0, 1, 2, \dots$,

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

Then we can re-write DIGing stochastic gradient Langevin dynamics as follows.

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

$$y_i^{(k+1)} = \sum_{j \in \Omega_i(k)} W_{ij}^{(k)} y_j^{(k)} + \nabla f_i(x_i^{(k+1)}) - \nabla f_i(x_i^{(k)}) + \xi_i^{(k+1)} - \xi_i^{(k)}, \quad (3.6)$$

We can also write the algorithm (3.5)-(3.6) in the form of matrix format for N agents as follows.

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

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

where

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

Consider a time-varying undirected graph sequence $\{\mathcal{G}^{\text{un}}(k)\}_{k=0}^\infty$. For every k , $\mathcal{G}^{\text{un}}(k)$ 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 . The set of neighbors of agents i (including agent i him/herself) at time k is defined as $\Omega_i(k) := \{j | (j, i) \in \mathcal{E}(k)\}$.

Next, we introduce the assumption on the mixing matrices $W^{(k)}$. First, we introduce the notation that

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

for any $k = 0, 1, 2, \dots$ and any $B = 0, 1, 2, \dots$ with the convention that $W_B^{(k)} = I$ for any $k < 0$ and $W_0^{(k)} = I$ for any k .

Assumption 3. For any $k = 0, 1, 2, \dots$, the mixing matrix $W^{(k)} = (W_{ij}^{(k)}) \in \mathbb{R}^{N \times N}$ 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) $W^{(k)} \mathbf{1} = \mathbf{1}$, $\mathbf{1}^\top W^{(k)} = \mathbf{1}^\top$.
- (iii) (joint spectral property) There exists some $B \in \mathbb{N}$ such that for every $k = 0, 1, 2, \dots$, $\delta := \sup_{k \geq B-1} \delta(k) < 1$, where $\delta(k) := \sigma_{\max} \left\{ W_B^{(k)} - \frac{1}{N} \mathbf{1} \mathbf{1}^\top \right\}$.

I am thinking in the numerical examples, we will probably try a few examples of $W^{(k)}$. maybe it would be interesting to compute out δ for these examples? MG: Indeed.

3.1 Main Results

Theorem 4. Assume $\|x^{(0)}\|_{L_2}$ is finite, and the stepsize η satisfies $\eta < \frac{1}{\mu+L}$, and $\eta\mu(1 - \frac{\eta L}{2}) \leq 1$. Also assume that $\delta < \lambda^B < 1$ with $\delta(k)$ defined in Assumption 3 and (3.24) holds. Moreover, we assume that $0 < \gamma_1 \gamma_2 \gamma_3 \gamma_4 < 1$, where $\gamma_1, \gamma_2, \gamma_3, \gamma_4$ are defined in (3.19)-(3.20). Then, for every k ,

$$\begin{aligned} & \frac{1}{N} \sum_{i=1}^N \mathcal{W}_2 \left(\text{Law} \left(x_i^{(k)} \right), \pi \right) \\ & \leq (1 - \mu\eta)^k \left(\left(\mathbb{E} \|\bar{x}^{(0)} - x_*\|^2 \right)^{1/2} + \sqrt{2\mu^{-1}dN^{-1}} \right) + \frac{1.65L}{\mu} \sqrt{\eta d N^{-1}} \\ & \quad + \eta^{1/2} \left(\frac{\eta}{\mu(1 - \frac{\eta L}{2})} + \frac{(1 + \eta L)^2}{\mu^2(1 - \frac{\eta L}{2})^2} \right)^{1/2} \cdot \left(\frac{3L^2 D^2 \eta 2^{2(B-1)} \delta^{-2}}{N(1 - \delta^{\frac{1}{B}})^2} + \frac{6dL^2 \cdot 2^{2(B-1)} \delta^{-2}}{1 - \delta^{\frac{2}{B}}} \right)^{1/2} \\ & \quad + \frac{\sqrt{\eta}\sigma}{\sqrt{\mu(1 - \frac{\eta L}{2})N}} + \left(\frac{\left(\delta^{\frac{2}{B}} \right)^k - \left(1 - \eta\mu \left(1 - \frac{\eta L}{2} \right) \right)^k}{\delta^{\frac{2}{B}} - 1 + \eta\mu \left(1 - \frac{\eta L}{2} \right)} \right)^{1/2} \cdot \frac{\sqrt{3}L \cdot 2^{B-1} \delta^{-1}}{\sqrt{N}} \delta^{\frac{1}{B}} \cdot \|x^{(0)}\|_{L_2} \\ & \quad + \frac{\sqrt{3}2^{B-1} \delta^{-1} \delta^{\frac{k}{B}}}{\sqrt{N}} \|x^{(0)}\|_{L_2} + \frac{\sqrt{3}D\eta 2^{B-1} \delta^{-1}}{\sqrt{N}(1 - \delta^{\frac{1}{B}})} + \frac{\sqrt{6d\eta} 2^{B-1} \delta^{-1}}{\sqrt{1 - \delta^{\frac{2}{B}}}}, \end{aligned}$$

where x_* is the minimizer of f , $\bar{x}^{(0)} = \frac{1}{N} \sum_{i=1}^N x_i^{(0)}$, D is defined in (3.18) and π is the Gibbs distribution with probability density function proportional to $\exp(-f(x))$.

3.2 Proofs of the Main Results

In this section, we present the proof of Theorem 4 by establishing a sequence of technical lemmas whose proofs will be provided in Appendix A. To prove Theorem 4, based on the triangle inequality for the 2-Wasserstein distance, we consider the following decomposition:

$$\frac{1}{N} \sum_{i=1}^N \mathcal{W}_2 \left(\text{Law} \left(x_i^{(k)} \right), \pi \right) \leq \frac{1}{N} \sum_{i=1}^N \mathcal{W}_2 \left(\text{Law} \left(x_i^{(k)} \right), \text{Law} \left(\bar{x}^{(k)} \right) \right) + \mathcal{W}_2 \left(\text{Law} \left(\bar{x}^{(k)} \right), \pi \right), \quad (3.10)$$

where

$$\mathcal{W}_2 \left(\text{Law} \left(\bar{x}^{(k)} \right), \pi \right) \leq \mathcal{W}_2 \left(\text{Law} \left(\bar{x}^{(k)} \right), \text{Law}(x_k) \right) + \mathcal{W}_2 \left(\text{Law}(x_k), \pi \right), \quad (3.11)$$

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

$$x_{k+1} = x_k - \frac{\eta}{N} \nabla f(x_k) + \sqrt{2\eta \bar{w}^{(k+1)}} W_t, \quad (3.12)$$

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

$$dX_t = -\frac{1}{N} \nabla f(X_t) dt + \sqrt{2N^{-1}} dW_t, \quad (3.13)$$

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 (3.12) obtained from Euler-Maruyama discretization of overdamped diffusion (3.13); and (3) the \mathcal{W}_2 distance between the law of x_k in (3.12) 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 the uniform L_2 bounds between $x_i^{(k)}$ and their average $\bar{x}^{(k)}$. As a first step, we derive a uniform L_2 bound for $y^{(k)}$, which is a key ingredient. First, we recall that [We need to be careful about the notation. I noticed that in \[GGHZ21\], \$x^{\(k\)}\$ is a big vector in \$\mathbb{R}^{Nd}\$, whereas in \[NOS17\], their \$x^{\(k\)}\$ is \$N \times d\$ matrix.](#)

$$\tilde{x}^{(k)} = x^{(k)} - \mathbf{1} \left(\bar{x}^{(k)} \right)^\top, \quad \tilde{y}^{(k)} = y^{(k)} - \mathbf{1} \left(\bar{y}^{(k)} \right)^\top, \quad (3.14)$$

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

$$x^{(k+1)} = \mathcal{W}^{(k)} x^{(k)} - \eta y^{(k)} + \sqrt{2\eta} w^{(k+1)}, \quad (3.15)$$

$$y^{(k+1)} = \mathcal{W}^{(k)} y^{(k)} + \nabla F \left(x^{(k+1)} \right) - \nabla F \left(x^{(k)} \right) + \xi^{(k+1)} - \xi^{(k)}. \quad (3.16)$$

Lemma 5. *For every k ,*

$$\mathbb{E} \left\| y^{(k)} \right\|^2 \leq D^2, \quad (3.17)$$

where

$$D^2 := 2 \left(\frac{\gamma_1 \gamma_2 \gamma_3 (\tilde{\omega}_4 + \hat{\omega}_4) + \gamma_1 \gamma_2 (\tilde{\omega}_3 + \hat{\omega}_3) + \tilde{\omega}_1 + \hat{\omega}_1}{1 - \gamma_1 \gamma_2 \gamma_3 \gamma_4} \right)^2 + \frac{4L^2}{N} \left(\frac{\gamma_3 \gamma_4 (\tilde{\omega}_1 + \hat{\omega}_1) + \gamma_3 (\tilde{\omega}_4 + \hat{\omega}_4) + \tilde{\omega}_3 + \hat{\omega}_3}{1 - \gamma_1 \gamma_2 \gamma_3 \gamma_4} \right)^2 + \frac{4}{N} \sigma^2, \quad (3.18)$$

provided that $\gamma_1 \gamma_2 \gamma_3 \gamma_4 \in (0, 1)$, where

$$\gamma_1 := \frac{\lambda(1 - \lambda^B)}{(\lambda^B - \delta)(1 - \lambda)}, \quad \gamma_2 := L \left(1 + \frac{1}{\lambda} \right), \quad (3.19)$$

$$\gamma_3 := \left(1 + \sqrt{N}(\lambda)^{-1} \left(\sqrt{\frac{L(1 + \alpha)}{\mu\alpha}} + \beta \right) \right), \quad \gamma_4 := \frac{\eta(1 - \lambda^B)}{(\lambda^B - \delta)(1 - \lambda)}, \quad (3.20)$$

and

$$\tilde{\omega}_1 := \frac{\lambda^B}{\lambda^B - \delta} \sum_{t=1}^B \lambda^{1-t} \|\tilde{y}^{(t-1)}\|_{L_2}, \quad \hat{\omega}_1 := \frac{\lambda^B}{\lambda^B - \delta} \cdot 2B\sigma\sqrt{N}, \quad (3.21)$$

$$\tilde{\omega}_3 := 2\sqrt{N} \|\tilde{x}^{(0)} - x_*\|, \quad \hat{\omega}_3 := \sqrt{N}(\lambda)^{-1} \left(\sqrt{\frac{L(1 + \alpha)}{\mu\alpha}} + \beta \right) \frac{1}{\mu} \left(\sigma + \sqrt{\frac{2d}{\eta}} \right), \quad (3.22)$$

$$\tilde{\omega}_4 := \frac{\lambda^B}{\lambda^B - \delta} \sum_{t=1}^B \lambda^{1-t} \|\tilde{x}^{(t-1)}\|_{L_2}, \quad \hat{\omega}_4 := \frac{\lambda^B}{\lambda^B - \delta} \cdot \sqrt{2\eta Nd}, \quad (3.23)$$

where $\delta = \sup_{k \geq B-1} \delta(k)$, with $\delta(k)$ defined in Assumption 3, and α, β, λ are tunable parameters such that $\delta < \lambda^B < 1$ and

$$\sqrt{1 - \frac{\eta\mu\beta}{\beta + 1}} \leq \lambda < 1, \quad \text{and} \quad \eta \leq \frac{1}{(1 + \alpha)L}. \quad (3.24)$$

The proof of Lemma 5 relies on a sequence of technical lemmas, that we will introduce next. For any $k = 0, 1, 2, \dots$, let us define:

$$q^{(k)} := x^{(k)} - x^*, \quad (3.25)$$

where $x^* = [x_*^\top, x_*^\top, \dots, x_*^\top]^\top$ and for any $k = 1, 2, \dots$, let us define:

$$z^{(k)} := \nabla F(x^{(k)}) - \nabla F(x^{(k-1)}). \quad (3.26)$$

Inspired by [NOS17], we introduce the following loop:

$$q \rightarrow z \rightarrow \tilde{y} \rightarrow \tilde{x} \rightarrow q, \quad (3.27)$$

and we will explain the meaning of the loop in details later. The main idea to show Lemma 5 is to establish each arrow in (3.27), and as a result, one can show that $\mathbb{E} \|\tilde{y}^{(k)}\|^2$ is uniformly bounded in k and hence one can show that $\mathbb{E} \|y^{(k)}\|^2$ is uniformly bounded in k , which will establish Lemma 5.

Let us first study the first arrow $q \rightarrow z$ in (3.27). Let us introduce the notations:

$$\|z\|_{L_2}^{\lambda,K} := \max_{0,1,\dots,K} \frac{1}{\lambda^k} \left(\mathbb{E} \|z^{(k)}\|^2 \right)^{1/2}, \quad (3.28)$$

$$\|q\|_{L_2}^{\lambda,K} := \max_{0,1,\dots,K} \frac{1}{\lambda^k} \left(\mathbb{E} \|q^{(k)}\|^2 \right)^{1/2}. \quad (3.29)$$

The arrow $q \rightarrow z$ means that we would like to establish an upper bound on $\|z\|_{L_2}^{\lambda,K}$ using $\|q\|_{L_2}^{\lambda,K}$. We have the following technical lemma.

Lemma 6. *For any $K = 0, 1, 2, \dots$, and $\lambda \in (0, 1)$, we have*

$$\|z\|_{L_2}^{\lambda,K} \leq L \left(1 + \frac{1}{\lambda} \right) \|q\|_{L_2}^{\lambda,K}. \quad (3.30)$$

Next, let us consider the second arrow $z \rightarrow \tilde{y}$ in (3.27). Let us recall from (3.8) and (3.26) that

$$y^{(k+1)} = \mathcal{W}^{(k)} y^{(k)} + z^{(k+1)} + \xi^{(k+1)} - \xi^{(k)}. \quad (3.31)$$

Similar as before, let us first define:

$$\|\tilde{y}\|_{L_2}^{\lambda,K} := \max_{0,1,\dots,K} \frac{1}{\lambda^k} \left(\mathbb{E} \|\tilde{y}^{(k)}\|^2 \right)^{1/2}. \quad (3.32)$$

The second arrow $z \rightarrow \tilde{y}$ means that we would like to 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 7. *Let $\delta = \sup_{k \geq B-1} \delta(k)$, where $\delta(k)$ is defined in Assumption 3. Let λ be such that $\delta < \lambda^B < 1$. Then for any $K = 0, 1, 2, \dots$, we have*

$$\|\tilde{y}\|_{L_2}^{\lambda,K} \leq \frac{\lambda(1 - \lambda^B)}{(\lambda^B - \delta)(1 - \lambda)} \|z\|_{L_2}^{\lambda,K} + \frac{\lambda^B}{\lambda^B - \delta} \frac{2B\sigma\sqrt{N}}{\lambda^K} + \frac{\lambda^B}{\lambda^B - \delta} \sum_{t=1}^B \lambda^{1-t} \|\tilde{y}^{(t-1)}\|_{L_2}. \quad (3.33)$$

Next, let us consider the third arrow $\tilde{y} \rightarrow \tilde{x}$ in (3.27). For this arrow, we would like 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 8. *Let $\delta = \sup_{k \geq B-1} \delta(k)$, where $\delta(k)$ is defined in Assumption 3. Let λ be such that $\delta < \lambda^B < 1$. Then for any $K = 0, 1, 2, \dots$, we have*

$$\|\tilde{x}\|_{L_2}^{\lambda,K} \leq \frac{\eta(1 - \lambda^B)}{(\lambda^B - \delta)(1 - \lambda)} \|\tilde{y}\|_{L_2}^{\lambda,K} + \frac{\lambda^B}{\lambda^B - \delta} \frac{\sqrt{2\eta Nd}}{\lambda^K} + \frac{\lambda^B}{\lambda^B - \delta} \sum_{t=1}^B \lambda^{1-t} \|\tilde{x}^{(t-1)}\|_{L_2}. \quad (3.34)$$

Finally, let us consider the last arrow $\tilde{x} \rightarrow q$ in (3.27), for which we would like to establish an upper bound on $\|q^{(k)}\|_{L_2}^{\lambda,K}$ by using $\|\tilde{x}\|_{L_2}^{\lambda,K}$. We have the following result.

Lemma 9. Assume that

$$\sqrt{1 - \frac{\eta\mu\beta}{\beta+1}} \leq \lambda < 1, \quad \text{and} \quad \eta \leq \frac{1}{(1+\alpha)L}, \quad (3.35)$$

where $\alpha, \beta > 0$ are tunable parameters. Then, for every $K = 0, 1, 2, \dots$, we have

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

It follows from Lemma 6, Lemma 7, Lemma 8 and Lemma 9 that

$$\|\tilde{y}\|_{L_2}^{\lambda,K} \leq \gamma_1 \|z\|_{L_2}^K + \omega_1(K), \quad (3.36)$$

$$\|z\|_{L_2}^{\lambda,K} \leq \gamma_2 \|q\|_{L_2}^K + \omega_2(K), \quad (3.37)$$

$$\|q\|_{L_2}^{\lambda,K} \leq \gamma_3 \|\tilde{x}\|_{L_2}^K + \omega_3(K), \quad (3.38)$$

$$\|\tilde{x}\|_{L_2}^{\lambda,K} \leq \gamma_4 \|\tilde{y}\|_{L_2}^K + \omega_4(K), \quad (3.39)$$

where $\gamma_1, \gamma_2, \gamma_3, \gamma_4$ are defined in (3.19)-(3.20) and

$$\omega_1(K) := \frac{\lambda^B}{\lambda^B - \delta} \frac{2B\sigma\sqrt{N}}{\lambda^K} + \frac{\lambda^B}{\lambda^B - \delta} \sum_{t=1}^B \lambda^{1-t} \|\tilde{y}^{(t-1)}\|_{L_2}, \quad \omega_2(K) := 0, \quad (3.40)$$

$$\omega_3(K) := 2\sqrt{N}\|\bar{x}^{(0)} - x_*\| + \sqrt{N}(\lambda)^{-1} \left(\sqrt{\frac{L(1+\alpha)}{\mu\alpha}} + \beta\right) \frac{1}{\mu} \left(\sigma + \sqrt{\frac{2d}{\eta}}\right) \frac{1}{\lambda^K}, \quad (3.41)$$

$$\omega_4(K) := \frac{\lambda^B}{\lambda^B - \delta} \frac{\sqrt{2\eta Nd}}{\lambda^K} + \frac{\lambda^B}{\lambda^B - \delta} \sum_{t=1}^B \lambda^{1-t} \|\tilde{x}^{(t-1)}\|_{L_2}. \quad (3.42)$$

As an immediate consequence of (3.36)-(3.39), we obtain the following technical lemma.

Lemma 10. Assume that $\delta < \lambda^B < 1$ with $\delta(k)$ defined in Assumption 3 and (3.24) holds. Then, for every $K = 0, 1, 2, \dots$,

$$\|\tilde{y}\|_{L_2}^{\lambda,K} \leq \frac{\gamma_1\gamma_2\gamma_3\omega_4(K) + \gamma_1\gamma_2\omega_3(K) + \gamma_1\omega_2(K) + \omega_1(K)}{1 - \gamma_1\gamma_2\gamma_3\gamma_4}, \quad (3.43)$$

and

$$\|q\|_{L_2}^{\lambda,K} \leq \frac{\gamma_3\gamma_4\gamma_1\omega_2(K) + \gamma_3\gamma_4\omega_1(K) + \gamma_3\omega_4(K) + \omega_3(K)}{1 - \gamma_1\gamma_2\gamma_3\gamma_4}, \quad (3.44)$$

where $\gamma_1, \gamma_2, \gamma_3, \gamma_4$ are defined in (3.19)-(3.20) and $\omega_1(K), \omega_2(K), \omega_3(K), \omega_4(K)$ are defined in (3.40), (3.41) and (3.42).

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 11. Assume that $\delta < \lambda^B < 1$ with $\delta(k)$ defined in Assumption 3 and (3.24) holds. Then, for any k , we have

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

where D is defined in (3.18) and

$$\bar{\gamma}_{k-1-s}^{(k-1)} := \left\| W_{k-1-s}^{(k-1)} - \frac{1}{N} 1_N 1_N^\top \right\|. \quad (3.45)$$

Next, we aim to provide an upper bound for $\bar{\gamma}_{k-1-s}^{(k)}$ in Lemma 11 under Assumption 3 for the mixing matrices $W^{(k)}$ to obtain the following corollary of Lemma 11, which shows that the iterates $x_i^{(k)}$ are close to the average $\bar{x}^{(k)}$ on average.

Lemma 12. Assume that $\delta < \lambda^B < 1$ with $\delta(k)$ defined in Assumption 3 and (3.24) holds. Then, for any k , we have

$$\sum_{i=1}^N \mathbb{E} \left\| x_i^{(k)} - \bar{x}^{(k)} \right\|^2 \leq 3 \cdot 2^{2(B-1)} \delta^{-2} \left(\delta^{\frac{2}{B}} \right)^k \mathbb{E} \left\| x^{(0)} \right\|^2 + \frac{3D^2 \eta^2 2^{2(B-1)} \delta^{-2}}{(1 - \delta^{\frac{1}{B}})^2} + \frac{6dN\eta \cdot 2^{2(B-1)} \delta^{-2}}{1 - \delta^{\frac{2}{B}}}, \quad (3.46)$$

where D is defined in (3.18)

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 (3.12).

First, by taking the average of N nodes in (3.5)-(3.6), and using the fact that $W^{(k)}$ is doubly stochastic, we obtain:

$$\bar{x}^{(k+1)} = \bar{x}^{(k)} - \eta \bar{y}^{(k)} + \sqrt{2\eta \bar{w}}^{(k+1)}, \quad (3.47)$$

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

$$\bar{y}^{(k+1)} = \bar{y}^{(k)} + \frac{1}{N} \sum_{i=1}^N \nabla f_i \left(x_i^{(k+1)} \right) - \frac{1}{N} \sum_{i=1}^N \nabla f_i \left(x_i^{(k)} \right) + \bar{\xi}^{(k+1)} - \bar{\xi}^{(k)}, \quad (3.48)$$

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

$$\bar{y}^{(k)} = \frac{1}{N} \sum_{i=1}^N \nabla f_i \left(x_i^{(k)} \right) + \bar{\xi}^{(k)}. \quad (3.49)$$

Therefore, we have

$$\bar{x}^{(k+1)} = \bar{x}^{(k)} - \eta \frac{1}{N} \sum_{i=1}^N \nabla f_i \left(x_i^{(k)} \right) - \eta \bar{\xi}^{(k)} + \sqrt{2\eta \bar{w}}^{(k+1)}, \quad (3.50)$$

which can be re-written as

$$\bar{x}^{(k+1)} = \bar{x}^{(k)} - \eta \frac{1}{N} \nabla f(\bar{x}^{(k)}) + \eta \mathcal{E}_k - \eta \bar{\xi}^{(k)} + \sqrt{2\eta} \bar{w}^{(k+1)}, \quad (3.51)$$

where

$$\mathcal{E}_k := \frac{1}{N} \sum_{i=1}^N \left[\nabla f_i(\bar{x}^{(k)}) - \nabla f_i(x_i^{(k)}) \right]. \quad (3.52)$$

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

Lemma 13. *Assume that $\delta < \lambda^B < 1$ with $\delta(k)$ defined in Assumption 3 and (3.24) holds. Then, for any k , we have*

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

Next, we recall from (3.12) that the iterates x_k are given by:

$$x_{k+1} = x_k - \eta \frac{1}{N} \nabla f(x_k) + \sqrt{2\eta} \bar{w}^{(k+1)}, \quad (3.53)$$

where $x_0 = \bar{x}_0 = \frac{1}{N} \sum_{i=1}^N x_i^{(0)}$. This is a Euler-Mariyama discretization (with stepsize η) of the continuous-time overdamped Langevin diffusion (3.13). Since the L_2 bound of the error term \mathcal{E}_k can be controlled as in Lemma 13, 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 14. *Assume that $\delta < \lambda^B < 1$ with $\delta(k)$ defined in Assumption 3 and (3.24) holds. We also assume $\mathbb{E} \|x^{(0)}\|^2 < \infty$. For any stepsize $\eta < 2/L$ and $\eta\mu(1 - \frac{\eta L}{2}) \leq 1$, we have for every k ,*

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

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 (3.12) of the overdamped Langevin diffusion (3.13) 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 [DK19] as follows.

Lemma 15 (Theorem 4 in [DK19]). *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 4.

3.2.4 Completing the Proof of Theorem 4

Proof of Theorem 4. The L_2 distance between the minimizer of f and Gibbs distribution π has been studied in the literature; see e.g. [GGHZ21]. More precisely, we have

$$\mathbb{E}_{X \sim \pi} \|X - x_*\|^2 \leq \frac{2dN^{-1}}{\mu}, \quad (3.54)$$

where x_* is the unique minimizer of $f(x)$; see Lemma 10 in [GGHZ21]. Since $x_0 = \frac{1}{N} \sum_{i=1}^N x_i^{(0)}$, we have $\mathbb{E} \|x_0\|^2 < \infty$. By (3.54), we get

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

It then follows from Lemma 15 that for any $\eta \leq \frac{2N}{\mu+L}$, we have

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

Moreover, it follows from Lemma 14 that

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

Finally, by Cauchy-Schwarz inequality,

$$\begin{aligned} \frac{1}{N} \sum_{i=1}^N \mathcal{W}_2(\text{Law}(x_i^{(k)}), \text{Law}(\bar{x}^{(k)})) &\leq \sqrt{\frac{1}{N} \sum_{i=1}^N \mathcal{W}_2^2(\text{Law}(x_i^{(k)}), \text{Law}(\bar{x}^{(k)}))} \\ &\leq \sqrt{\frac{1}{N} \sum_{i=1}^N \mathbb{E} \left\| x_i^{(k)} - \bar{x}^{(k)} \right\|^2}. \end{aligned} \quad (3.55)$$

By Lemma 12, we have

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

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

4 Numerical Experiments

5 Conclusion

Acknowledgments

Mert Gürbüzbalaban’s research is supported in part by the grants Office of Naval Research Award Number N00014-21-1-2244, National Science Foundation (NSF) CCF-1814888, NSF DMS-2053485. Lingjiong Zhu is partially supported by the grants NSF DMS-2053454, NSF DMS-2208303.

References

- [ABC⁺20] Yossi Arjevani, Joan Bruna, Bugra Can, Mert Gürbüzbalaban, Stefanie Jegelka, and Hongzhou Lin. IDEAL: Inexact DEcentralized accelerated augmented Lagrangian method. In *Advances in Neural Information Processing Systems*, 2020.
- [BCE⁺22] 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.
- [BCM⁺21] Mathias Barkhagen, Ngoc Huy Chau, Éric Moulines, Miklós Rásonyi, Sotirios Sabanis, and Ying Zhang. On stochastic gradient Langevin dynamics with dependent data streams in the logconcave case. *Bernoulli*, 27(1):1–33, 2021.
- [Bot10] Léon Bottou. Large-scale machine learning with stochastic gradient descent. In *Proceedings of COMPSTAT’2010*, pages 177–186. Springer, 2010.
- [CB18] Xiang Cheng and Peter L. Bartlett. Convergence of Langevin MCMC in KL-divergence. In *Proceedings of the 29th International Conference on Algorithmic Learning Theory (ALT)*, volume 83, pages 186–211. PMLR, 2018.

- [CE20] Raffaele Chiappinelli and David E Edmunds. Remarks on surjectivity of gradient operators. *Mathematics*, 8(9):1538, 2020.
- [CEL⁺24] 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*, 2024.
- [CMR⁺21] Ngoc Huy Chau, Éric Moulines, Miklos Rásonyi, Sotirios Sabanis, and Ying Zhang. On stochastic gradient Langevin dynamics with dependent data streams: the fully non-convex case. *SIAM Journal of Mathematics of Data Science*, 3(3):959–986, 2021.
- [Dal17] Arnak S Dalalyan. Theoretical guarantees for approximate sampling from smooth and log-concave densities. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 79(3):651–676, 2017.
- [DK19] Arnak S Dalalyan and Avetik Karagulyan. User-friendly guarantees for the Langevin Monte Carlo with inaccurate gradient. *Stochastic Processes and their Applications*, 129(12):5278–5311, 2019.
- [DM17] Alain Durmus and Eric Moulines. Non-asymptotic convergence analysis for the Unadjusted Langevin Algorithm. *Annals of Applied Probability*, 27(3):1551–1587, 2017.
- [DM19] Alain Durmus and Eric Moulines. High-dimensional Bayesian inference via the Unadjusted Langevin Algorithm. *Bernoulli*, 25(4A):2854–2882, 2019.
- [DRW⁺16] Kumar Avinava Dubey, Sashank J Reddi, Sinead A Williamson, Barnabas Póczos, Alexander J Smola, and Eric P Xing. Variance reduction in stochastic gradient Langevin dynamics. In *Advances in Neural Information Processing Systems*, pages 1154–1162, 2016.
- [EH21] 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.
- [EHZ22] Murat A. Erdogdu, Rasa Hosseinzadeh, and S. Zhang, Matthew. Convergence analysis of Langevin Monte Carlo in chi-square and Rényi divergence. In *Proceedings of the 25th International Conference on Artificial Intelligence and Statistics (AISTATS)*, volume 151, pages 8151–8175. PMLR, 2022.
- [FGO⁺22] Alireza Fallah, Mert Gürbüzbalaban, Asuman Ozdaglar, Umut Şimşekli, and Lingjiong Zhu. Robust distributed accelerated stochastic gradient methods for multi-agent networks. *Journal of Machine Learning Research*, 23(220):1–96, 2022.
- [GDG19] Eduard Gorbunov, Darina Dvinskikh, and Alexander Gasnikov. Optimal decentralized distributed algorithms for stochastic convex optimization. *arXiv:1911.07363*, 2019.
- [GGHZ21] Mert Gürbüzbalaban, Xuefeng Gao, Yuanhan Hu, and Lingjiong Zhu. Decentralized stochastic gradient Langevin dynamics and Hamiltonian Monte Carlo. *The Journal of Machine Learning Research*, 22(1):10804–10872, 2021.

- [GIWZ24] Mert Gürbüzbalaban, Mohammad Rafiqul Islam, Xiaoyu Wang, and Lingjiong Zhu. Generalized EXTRA stochastic gradient Langevin dynamics. *arXiv:2412.01993*, 2024.
- [HBJ18] Lie He, An Bian, and Martin Jaggi. COLA: Decentralized linear learning. In *Advances in Neural Information Processing Systems*, pages 4536–4546, 2018.
- [HBM19] Hadrien Hendrikx, Francis Bach, and Laurent Massoulié. An accelerated decentralized stochastic proximal algorithm for finite sums. In *Advances in Neural Information Processing Systems*, pages 954–964, 2019.
- [Hof09] Peter D Hoff. *A First Course in Bayesian Statistical Methods*, volume 580. Springer, 2009.
- [Jak18] Dušan Jakovetić. A unification and generalization of exact distributed first-order methods. *IEEE Transactions on Signal and Information Processing over Networks*, 5(1):31–46, 2018.
- [LZT22] Ruilin Li, Hongyuan Zha, and Molei Tao. Sqrt(d) dimension dependence of Langevin Monte Carlo. In *International Conference on Learning Representations*, 2022.
- [Ned20] Angelia Nedic. Distributed gradient methods for convex machine learning problems in networks: Distributed optimization. *IEEE Signal Processing Magazine*, 37(3):92–101, 2020.
- [NO09] Angelia Nedic and Asuman Ozdaglar. Distributed subgradient methods for multi-agent optimization. *IEEE Transactions on Automatic Control*, 54(1):48–61, 2009.
- [NOS17] Angelia Nedic, Alex Olshevsky, and Wei Shi. Achieving geometric convergence for distributed optimization over time-varying graphs. *SIAM Journal on Optimization*, 27(4):2597–2633, 2017.
- [Pav14] Grigorios A Pavliotis. *Stochastic Processes and Applications: Diffusion processes, the Fokker-Planck and Langevin Equations*, volume 60. Springer, 2014.
- [PBGG20] Anjaly Parayil, He Bai, Jemin George, and Prudhvi Gurram. Decentralized Langevin dynamics for Bayesian learning. In *Advances in Neural Information Processing Systems*, 2020.
- [PS17] Nicholas G. Polson and Vadim Sokolov. Deep learning: A Bayesian perspective. *Bayesian Analysis*, 12(4):1275–1304, 12 2017.
- [RRT17] Maxim Raginsky, Alexander Rakhlin, and Matus Telgarsky. Non-convex learning via stochastic gradient Langevin dynamics: a nonasymptotic analysis. In *Conference on Learning Theory*, volume 65, pages 1674–1703. PMLR, 2017.
- [SBB⁺19] Kevin Scaman, Francis Bach, Sébastien Bubeck, Yin Lee, and Laurent Massoulié. Optimal convergence rates for convex distributed optimization in networks. *Journal of Machine Learning Research*, 20:1–31, 2019.

- [SKP⁺20] 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 e-prints*, page arXiv:2003.10309, March 2020.
- [SLWY15] Wei Shi, Qing Ling, Gang Wu, and Wotao Yin. EXTRA: An exact first-order algorithm for decentralized consensus optimization. *SIAM Journal on Optimization*, 25(2):944–966, 2015.
- [SSP20] Brian Swenson, Anirudh Sridhar, and H Vincent Poor. On distributed stochastic gradient algorithms for global optimization. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 8594–8598. IEEE, 2020.
- [ULGN17] César A Uribe, Soomin Lee, Alexander Gasnikov, and Angelia Nedić. Optimal algorithms for distributed optimization. *arXiv preprint arXiv:1712.00232*, 2017.
- [Vil09] Cédric Villani. *Optimal Transport: Old and New*. Springer, Berlin, 2009.
- [VW19] Santosh S. Vempala and Andre Wibisono. Rapid convergence of the Unadjusted Langevin Algorithm: Isoperimetry suffices. In *Advances in Neural Information Processing Systems*, 2019.
- [WT11] Max Welling and Yee Whye Teh. Bayesian learning via stochastic gradient Langevin dynamics. In *Proceedings of the 28th International Conference on Machine Learning (ICML-11)*, pages 681–688, 2011.
- [WY20] Hao Wang and Dit-Yan Yeung. A survey on Bayesian deep learning. *ACM Computing Surveys*, 52(5):1–37, 2020.
- [XCZG18] Pan Xu, Jinghui Chen, Difan Zou, and Quanquan Gu. Global convergence of Langevin dynamics based algorithms for nonconvex optimization. In *Advances in Neural Information Processing Systems*, pages 3122–3133, 2018.
- [ZADS23] Ying Zhang, Ömer Deniz Akyildiz, Theodoros Damoulas, and Sotirios Sabanis. Nonasymptotic estimates for Stochastic Gradient Langevin Dynamics under local conditions in nonconvex optimization. *Applied Mathematics & Optimization*, 87:25, 2023.

A Proofs of Technical Lemmas

A.1 Proof of Lemma 5

Proof. It follows from (3.43) and (3.44) 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 (3.19)-(3.20) and $\omega_1(K), \omega_2(K), \omega_3(K), \omega_4(K)$ are defined in (3.40), (3.41) and (3.42) and we recall from (3.21), (3.22) and (3.23) that

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

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

$$\tilde{\omega}_4 := \frac{\lambda^B}{\lambda^B - \delta} \sum_{t=1}^B \lambda^{1-t} \|\tilde{x}^{(t-1)}\|_{L_2}, \quad \hat{\omega}_4 := \frac{\lambda^B}{\lambda^B - \delta} \cdot \sqrt{2\eta Nd}. \quad (\text{A.3})$$

Hence, for every k , we have

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

and

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

where we used $0 < \lambda < 1$.

Next, we can compute that We need to be careful about the notation. I noticed that in [GGHZ21], $x^{(k)}$ is a big vector in \mathbb{R}^{Nd} , whereas in [NOS17], their $x^{(k)}$ is $N \times d$ matrix.

$$\mathbb{E} \|y^{(k)}\|^2 \leq 2\mathbb{E} \|\tilde{y}^{(k)}\|^2 + 2\mathbb{E} \left\| \frac{1}{N} \mathbf{1}_N \mathbf{1}_N^\top \tilde{y}^{(k)} \right\|^2 = 2\mathbb{E} \|\tilde{y}^{(k)}\|^2 + 2N\mathbb{E} \|\bar{y}^{(k)}\|^2.$$

Moreover,

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

Hence, we conclude that

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

This completes the proof. \square

A.2 Proof of Lemma 6

Proof. The proof is almost identical to the proof of Lemma 3.9. in [NOS17] and is hence omitted here. \square

A.3 Proof of Lemma 7

Proof. First, we recall from (3.8) and (3.26) that

$$y^{(k+1)} = W^{(k)} y^{(k)} + z^{(k+1)} + \xi^{(k+1)} - \xi^{(k)}. \tag{A.5}$$

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

$$\begin{aligned}
\left\| \bar{y}^{(k+1)} \right\|_{L_2} &= \left\| L_N y^{(k+1)} \right\|_{L_2} \\
&\leq \left\| L_N W_B^{(k)} y^{(k+1-B)} \right\|_{L_2} \\
&\quad + \left\| L_N W_{B-1}^{(k)} z^{(k+2-B)} \right\|_{L_2} + \cdots + \left\| L_N W_1^{(k)} z^{(k)} \right\|_{L_2} + \left\| L_N W_0^{(k)} z^{(k+1)} \right\|_{L_2} \\
&\quad + \left\| L_N W_{B-1}^{(k)} \xi^{(k+2-B)} \right\|_{L_2} + \cdots + \left\| L_N W_1^{(k)} \xi^{(k)} \right\|_{L_2} + \left\| L_N W_0^{(k)} \xi^{(k+1)} \right\|_{L_2} \\
&\quad + \left\| L_N W_{B-1}^{(k)} \xi^{(k+1-B)} \right\|_{L_2} + \cdots + \left\| L_N W_1^{(k)} \xi^{(k-1)} \right\|_{L_2} + \left\| L_N W_0^{(k)} \xi^{(k)} \right\|_{L_2},
\end{aligned}$$

where $L_N = I - \frac{1}{N}\mathbf{1}\mathbf{1}^\top$. since our $y^{(k)}$ is a vector not a matrix, perhaps defn of L_N needs to be changed. By applying Lemma 16 and Assumption 2, we get

$$\|\tilde{y}^{(k+1)}\|_{L_2} \leq \delta \|\tilde{y}^{(k+1-B)}\|_{L_2} + \sum_{t=1}^B \|z^{(k+2-t)}\|_{L_2} + 2B\sigma\sqrt{N}. \quad (\text{A.6})$$

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

$$\lambda^{-(k+1)} \|\tilde{y}^{(k+1)}\|_{L_2} \leq \frac{\delta}{\lambda^B} \lambda^{-(k+1-B)} \|\tilde{y}^{(k+1-B)}\|_{L_2} + \sum_{t=1}^B \frac{1}{\lambda^{t-1}} \lambda^{-(k+2-t)} \|z^{(k+2-t)}\|_{L_2} + 2B\sigma\sqrt{N}. \quad (\text{A.7})$$

By following the similar argument as in the proof of Lemma 3.10 in [NOS17], we obtain that for every K :

$$\|\tilde{y}\|_{L_2}^{\lambda,K} \leq \frac{\delta}{\lambda^B} \|\tilde{y}\|_{L_2}^{\lambda,K} + \sum_{t=1}^B \frac{1}{\lambda^{t-1}} \|z\|_{L_2}^{\lambda,K} + \frac{2B\sigma\sqrt{N}}{\lambda^K} + \sum_{t=1}^B \lambda^{1-t} \|\tilde{y}^{(t-1)}\|_{L_2}. \quad (\text{A.8})$$

This implies that

$$\|\tilde{y}\|_{L_2}^{\lambda,K} \leq \frac{\lambda(1-\lambda^B)}{(\lambda^B-\delta)(1-\lambda)} \|z\|_{L_2}^{\lambda,K} + \frac{\lambda^B}{\lambda^B-\delta} \frac{2B\sigma\sqrt{N}}{\lambda^K} + \frac{\lambda^B}{\lambda^B-\delta} \sum_{t=1}^B \lambda^{1-t} \|\tilde{y}^{(t-1)}\|_{L_2}. \quad (\text{A.9})$$

This completes the proof. \square

A.4 Proof of Lemma 8

Proof. The proof follows the same steps as in the proof of Lemma 7 and the fact that $\mathbb{E}\|\sqrt{2\eta}w^{(k+1)}\|^2 = 2\eta Nd$. \square

A.5 Proof of Lemma 9

Proof. First, let us recall from (3.25) that double check the equation below. our $x^{(k)}$ is a vector, not a matrix

$$q^{(k)} = x^{(k)} - x^* = x^{(k)} - \mathbf{1} \left(\bar{x}^{(k)} \right)^\top + \mathbf{1} \left(\bar{x}^{(k)} \right)^\top - x^* = \tilde{x}^{(k)} + \mathbf{1} \left(\bar{x}^{(k)} - x_* \right)^\top. \quad (\text{A.10})$$

Next, by taking the average of N nodes in (3.5)-(3.6), and using the fact that $W^{(k)}$ is doubly stochastic, we obtain:

$$\bar{x}^{(k+1)} = \bar{x}^{(k)} - \eta \bar{y}^{(k)} + \sqrt{2\eta} \bar{w}^{(k+1)}, \quad (\text{A.11})$$

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

$$\bar{y}^{(k+1)} = \bar{y}^{(k)} + \frac{1}{N} \sum_{i=1}^N \nabla f_i \left(x_i^{(k+1)} \right) - \frac{1}{N} \sum_{i=1}^N \nabla f_i \left(x_i^{(k)} \right) + \bar{\xi}^{(k+1)} - \bar{\xi}^{(k)}, \quad (\text{A.12})$$

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

$$\bar{y}^{(k)} = \frac{1}{N} \sum_{i=1}^N \nabla f_i \left(x_i^{(k)} \right) + \bar{\xi}^{(k)}. \quad (\text{A.13})$$

Therefore, we have

$$\bar{x}^{(k+1)} = \bar{x}^{(k)} - \eta \frac{1}{N} \sum_{i=1}^N \nabla f_i \left(x_i^{(k)} \right) - \eta \bar{\xi}^{(k)} + \sqrt{2\eta} \bar{w}^{(k+1)}. \quad (\text{A.14})$$

By Lemma 18, we can re-write the equation (A.14) as:

$$\bar{x}^{(k+1)} = \bar{x}^{(k)} - \eta \frac{1}{N} \sum_{i=1}^N \nabla f_i \left(s_i^{(k)} \right), \quad (\text{A.15})$$

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

$$\nabla f_i \left(s_i^{(k)} \right) = \nabla f_i \left(x_i^{(k)} \right) + \bar{\xi}^{(k)} - \sqrt{\frac{2}{\eta}} \bar{w}^{(k+1)}. \quad (\text{A.16})$$

Since f_i is μ -strongly convex, we have

$$\left\| s_i^{(k)} - x_i^{(k)} \right\| \leq \frac{1}{\mu} \left\| \bar{\xi}^{(k)} - \sqrt{\frac{2}{\eta}} \bar{w}^{(k+1)} \right\|, \quad (\text{A.17})$$

which implies that

$$\left\| s_i^{(k)} - x_i^{(k)} \right\|_{L_2} \leq \frac{1}{\mu} \left(\frac{\sigma}{\sqrt{N}} + \sqrt{\frac{2d}{\eta N}} \right). \quad (\text{A.18})$$

By applying Lemma 17, under the assumption that

$$\sqrt{1 - \frac{\eta\mu\beta}{\beta+1}} \leq \lambda < 1, \quad \text{and} \quad \eta \leq \frac{1}{(1+\alpha)L}, \quad (\text{A.19})$$

where $\alpha, \beta > 0$, we have

$$\|\bar{x} - x_*\|_{L_2}^{\lambda, K} \leq 2\|\bar{x}^{(0)} - x_*\| + (\lambda\sqrt{N})^{-1} \left(\sqrt{\frac{L(1+\alpha)}{\mu\alpha}} + \beta \right) \sum_{i=1}^N \|\bar{x} - s_i\|_{L_2}^{\lambda, K}, \quad (\text{A.20})$$

for any $K = 0, 1, 2, \dots$ where x_* is the minimizer of f .

Therefore, we have

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

Finally, since we are using vector instead of matrix, pls double check the notation in the below eqn.

$$q^{(k)} = \tilde{x}^{(k)} + \mathbf{1}(\bar{x}^{(k)} - x_*)^\top, \quad (\text{A.21})$$

it follows that

$$\|q^{(k)}\|_{L_2}^{\lambda,K} \leq \|\tilde{x}^{(k)}\|_{L_2}^{\lambda,K} + \sqrt{N}\|\bar{x} - x_*\|_{L_2}^{\lambda,K}. \quad (\text{A.22})$$

Hence, we conclude that

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

This completes the proof. \square

A.6 Proof of Lemma 10

Proof. Under the assumption $\delta < \lambda^B < 1$ with $\delta(k)$ defined in Assumption 3 and (3.24), Lemma 6, Lemma 7, Lemma 8 and Lemma 9 hold, and it follows from (3.36)-(3.39) that

$$\|\tilde{y}\|_{L_2}^{\lambda,K} \leq \gamma_1\gamma_2\gamma_3\gamma_4\|\tilde{y}\|_{L_2}^{\lambda,K} + \gamma_1\gamma_2\gamma_3\omega_4(K) + \gamma_1\gamma_2\omega_3(K) + \gamma_1\omega_2(K) + \omega_1(K), \quad (\text{A.23})$$

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

$$\|\tilde{y}\|_{L_2}^{\lambda,K} \leq \frac{\gamma_1\gamma_2\gamma_3\omega_4(K) + \gamma_1\gamma_2\omega_3(K) + \gamma_1\omega_2(K) + \omega_1(K)}{1 - \gamma_1\gamma_2\gamma_3\gamma_4}. \quad (\text{A.24})$$

Similarly, one can show that

$$\|q\|_{L_2}^{\lambda,K} \leq \gamma_1\gamma_2\gamma_3\gamma_4\|q\|_{L_2}^{\lambda,K} + \gamma_3\gamma_4\gamma_1\omega_2(K) + \gamma_3\gamma_4\omega_1(K) + \gamma_3\omega_4(K) + \omega_3(K), \quad (\text{A.25})$$

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

$$\|q\|_{L_2}^{\lambda,K} \leq \frac{\gamma_3\gamma_4\gamma_1\omega_2(K) + \gamma_3\gamma_4\omega_1(K) + \gamma_3\omega_4(K) + \omega_3(K)}{1 - \gamma_1\gamma_2\gamma_3\gamma_4}. \quad (\text{A.26})$$

This completes the proof. \square

A.7 Proof of Lemma 11

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

$$x^{(k)} = \left(W_k^{(k-1)} \otimes I_d\right) x^{(0)} - \eta \sum_{s=0}^{k-1} \left(W_{k-1-s}^{(k-1)} \otimes I_d\right) y^{(s)} + \sqrt{2\eta} \sum_{s=0}^{k-1} \left(W_{k-1-s}^{(k-1)} \otimes I_d\right) w^{(s+1)}. \quad (\text{A.27})$$

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

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

Therefore, we get

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

Note that it follows from (A.27) that

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

By Cauchy-Schwarz inequality, we have

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

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

$$\begin{aligned} \left\| x^{(k)} - \frac{1}{N} \left((1_N 1_N^\top) \otimes I_d \right) x^{(k)} \right\|^2 &\leq 3 \left\| \left(\left(W_k^{(k-1)} - \frac{1}{N} 1_N 1_N^\top \right) \otimes I_d \right) x^{(0)} \right\|^2 \\ &\quad + 3\eta^2 \left\| \sum_{s=0}^{k-1} \left(\left(W_{k-1-s}^{(k-1)} - \frac{1}{N} 1_N 1_N^\top \right) \otimes I_d \right) y^{(s)} \right\|^2 \\ &\quad + 6\eta \left\| \sum_{s=0}^{k-1} \left(\left(W_{k-1-s}^{(k-1)} - \frac{1}{N} 1_N 1_N^\top \right) \otimes I_d \right) w^{(s+1)} \right\|^2. \end{aligned} \tag{A.28}$$

Note that

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

where we used Jensen's inequality in the last step above. Recall from Lemma 5 that for every $k = 0, 1, 2, \dots$,

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

where D is defined in (3.18). 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.28) that

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

The proof is complete. \square

A.8 Proof of Lemma 12

Proof. It follows from Lemma 11 that for any k ,

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

where D is defined in (3.18) and

$$\bar{\gamma}_{k-1-s}^{(k-1)} := \left\| W_{k-1-s}^{(k-1)} - \frac{1}{N} \mathbf{1}_N \mathbf{1}_N^\top \right\|. \quad (\text{A.29})$$

Under Assumption 3, there exists some $B \in \mathbb{N}$ such that for every $k = 0, 1, 2, \dots, \delta := \sup_{k \geq B-1} \delta(k) < 1$, where $\delta(k) := \sigma_{\max} \left\{ W_B^{(k)} - \frac{1}{N} \mathbf{1} \mathbf{1}^\top \right\}$. For every k , $W^{(k)}$ is doubly stochastic. Therefore, it follows that

$$\bar{\gamma}_{k-1-s}^{(k-1)} \leq 2^{B-1} \delta^{\frac{k-1-s}{B}-1}, \quad (\text{A.30})$$

for every $s = 0, 1, \dots, k-1$ and

$$\bar{\gamma}_k^{(k-1)} \leq 2^{B-1} \delta^{\frac{k}{B}-1}. \quad (\text{A.31})$$

Therefore, we get

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

The proof is complete. \square

A.9 Proof of Lemma 13

Proof. First, we can compute that

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

By Lemma 12, we can compute that

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

The proof is complete. \square

A.10 Proof of Lemma 14

Proof. The proof is similar to the proof of Lemma 7 in [GGHZ21] and for the sake of completeness we include all the details here. From (3.51) and (3.53), we can compute that

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

where we recall from (3.52) that

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

and this implies that

$$\begin{aligned}
\left\| \bar{x}^{(k+1)} - x_{k+1} \right\|^2 &= \left\| \bar{x}^{(k)} - x_k - \frac{\eta}{N} \left[\nabla f \left(\bar{x}^{(k)} \right) - \nabla f(x_k) \right] \right\|^2 + \eta^2 \left\| \mathcal{E}_k - \bar{\xi}^{(k)} \right\|^2 \\
&\quad + 2 \left\langle \bar{x}^{(k)} - x_k - \frac{\eta}{N} \left[\nabla f \left(\bar{x}^{(k)} \right) - \nabla f(x_k) \right], \eta \mathcal{E}_k - \eta \bar{\xi}^{(k)} \right\rangle \\
&= \left\| \bar{x}^{(k)} - x_k \right\|^2 + \eta^2 \left\| \frac{1}{N} \left[\nabla f \left(\bar{x}^{(k)} \right) - \nabla f(x_k) \right] \right\|^2 \\
&\quad - 2 \left\langle \bar{x}^{(k)} - x_k, \eta \frac{1}{N} \left[\nabla f \left(\bar{x}^{(k)} \right) - \nabla f(x_k) \right] \right\rangle + \eta^2 \left\| \mathcal{E}_k - \bar{\xi}^{(k)} \right\|^2 \\
&\quad + 2 \left\langle \bar{x}^{(k)} - x_k - \eta \frac{1}{N} \left[\nabla f \left(\bar{x}^{(k)} \right) - \nabla f(x_k) \right], \eta \mathcal{E}_k - \eta \bar{\xi}^{(k)} \right\rangle \\
&\leq \left\| \bar{x}^{(k)} - x_k \right\|^2 + \eta^2 L \left\langle \bar{x}^{(k)} - x_k, \frac{1}{N} \left[\nabla f \left(\bar{x}^{(k)} \right) - \nabla f(x_k) \right] \right\rangle \\
&\quad - 2 \left\langle \bar{x}^{(k)} - x_k, \eta \frac{1}{N} \left[\nabla f \left(\bar{x}^{(k)} \right) - \nabla f(x_k) \right] \right\rangle + \eta^2 \left\| \mathcal{E}_k - \bar{\xi}^{(k)} \right\|^2 \\
&\quad + 2 \left\langle \bar{x}^{(k)} - x_k - \eta \frac{1}{N} \left[\nabla f \left(\bar{x}^{(k)} \right) - \nabla f(x_k) \right], \eta \mathcal{E}_k - \eta \bar{\xi}^{(k)} \right\rangle \\
&\leq \left(1 - 2\eta\mu \left(1 - \frac{\eta L}{2} \right) \right) \left\| \bar{x}^{(k)} - x_k \right\|^2 + \eta^2 \left\| \mathcal{E}_k - \bar{\xi}^{(k)} \right\|^2 \\
&\quad + 2 \left\langle \bar{x}^{(k)} - x_k - \eta \frac{1}{N} \left[\nabla f \left(\bar{x}^{(k)} \right) - \nabla f(x_k) \right], \eta \mathcal{E}_k - \eta \bar{\xi}^{(k)} \right\rangle, \quad (\text{A.32})
\end{aligned}$$

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 $\bar{\xi}^{(k)}$ has mean zero and is independent of \mathcal{E}_k , and by Lemma 13,

$$\mathbb{E} \left\| \mathcal{E}_k \right\|^2 \leq \frac{3L^2 2^{2(B-1)} \delta^{-2}}{N} \left(\delta^{\frac{2}{B}} \right)^k \mathbb{E} \left\| x^{(0)} \right\|^2 + \frac{3L^2 D^2 \eta^2 2^{2(B-1)} \delta^{-2}}{N(1 - \delta^{\frac{1}{B}})^2} + \frac{6dL^2 \eta \cdot 2^{2(B-1)} \delta^{-2}}{1 - \delta^{\frac{2}{B}}}, \quad (\text{A.33})$$

and we also notice that $\mathbb{E} \left\| \bar{\xi}^{(k)} \right\|^2 \leq \frac{\sigma^2}{N}$. By taking expectations in (A.32), we get

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

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

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

By applying (A.33), we get

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

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

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

By our assumption on stepsize η , we have $1 - \eta\mu \left(1 - \frac{\eta L}{2}\right) \in [0, 1)$. Hence, we conclude that for every k ,

$$\begin{aligned} \mathbb{E} \left\| \bar{x}^{(k)} - x_k \right\|^2 &\leq \frac{\eta \left(\eta + \frac{(1+\eta L)^2}{\mu(1-\frac{\eta L}{2})} \right) \left(\frac{3L^2 D^2 \eta^2 2^{2(B-1)} \delta^{-2}}{N(1-\delta^{\frac{1}{B}})^2} + \frac{6dL^2 \eta \cdot 2^{2(B-1)} \delta^{-2}}{1-\delta^{\frac{2}{B}}} \right) + \eta^2 \frac{\sigma^2}{N}}{1 - \left(1 - \eta\mu \left(1 - \frac{\eta L}{2}\right)\right)} \\ &\quad + \frac{\left(\delta^{\frac{2}{B}}\right)^k - \left(1 - \eta\mu \left(1 - \frac{\eta L}{2}\right)\right)^k}{1 - \left(1 - \eta\mu \left(1 - \frac{\eta L}{2}\right)\right)} \frac{4L^2}{N} \mathbb{E} \left\| x^{(0)} \right\|^2 \\ &= \frac{\eta \left(\eta + \frac{(1+\eta L)^2}{\mu(1-\frac{\eta L}{2})} \right) \left(\frac{3L^2 D^2 \eta^2 2^{2(B-1)} \delta^{-2}}{N(1-\delta^{\frac{1}{B}})^2} + \frac{6dL^2 \cdot 2^{2(B-1)} \delta^{-2}}{1-\delta^{\frac{2}{B}}} \right) + \eta \frac{\sigma^2}{N}}{\mu \left(1 - \frac{\eta L}{2}\right)} \\ &\quad + \frac{\left(\delta^{\frac{2}{B}}\right)^k - \left(1 - \eta\mu \left(1 - \frac{\eta L}{2}\right)\right)^k}{\delta^{\frac{2}{B}} - 1 + \eta\mu \left(1 - \frac{\eta L}{2}\right)} \frac{3L^2 2^{2(B-1)} \delta^{-2}}{N} \delta^{\frac{2}{B}} \mathbb{E} \left\| x^{(0)} \right\|^2. \end{aligned}$$

The proof is complete. \square

B Additional Technical Lemmas

Lemma 16 (Lemma 3.4. in [NOS17]). *Under Assumption 3, for any $k = B - 1, B, \dots$ and any matrix b (with appropriate dimensions), then, we have $\|L_N W_B^{(k)} b\| \leq \delta(k) \|L_N b\|$, where $\delta(k)$ is defined in Assumption 3 and $L_N := I - \frac{1}{N} \mathbf{1}\mathbf{1}^\top$. since we are using long vector instead of matrix, defn of L_N might need to be changed.*

Next, we consider

$$\min_{x \in \mathbb{R}^d} g(x) := \frac{1}{N} \sum_{i=1}^N g_i(x), \quad (\text{B.1})$$

where g_i are μ -strongly convex and L -smooth. Consider the iterates:

$$p^{(k+1)} = p^{(k)} - \eta \frac{1}{N} \sum_{i=1}^N \nabla g_i \left(s_i^{(k)} \right). \quad (\text{B.2})$$

Then, we have the following technical lemma.

Lemma 17 (Lemma 3.12 in [NOS17]). *Assume that*

$$\sqrt{1 - \frac{\eta\mu\beta}{\beta + 1}} \leq \lambda < 1, \quad \text{and} \quad \eta \leq \frac{1}{(1 + \alpha)L}, \quad (\text{B.3})$$

where $\alpha, \beta > 0$. Then, we have

$$\|p - p_*\|^{\lambda, K} \leq 2\|p^{(0)} - p_*\| + (\lambda\sqrt{N})^{-1} \left(\sqrt{\frac{L(1 + \alpha)}{\mu\alpha}} + \beta \right) \sum_{i=1}^N \|p - s_i\|^{\lambda, K}, \quad (\text{B.4})$$

for any $K = 0, 1, 2, \dots$ where p_* is the minimizer of g .

Lemma 18. *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 [CE20, 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, [CE20, Theorem 1] is applicable and this completes the proof. \square