Scalability Enhancement and Data-Heterogeneity Awareness in Gradient Tracking based Decentralized Bayesian Learning

Kinjal Bhar KBHAR@OKSTATE.EDU
MAE, Oklahoma State University, Stillwater, OK 74078

He Bai HE.BAI@OKSTATE.EDU
MAE, Oklahoma State University, Stillwater, OK 74078

Jemin George JGEORGE@NSF.GOV
U.S. National Science Foundation, Alexandria, VA 22314

Carl Busart CARL.E.BUSART.CIV@ARMY.MIL
DEVCOM Army Research Laboratory, Adelphi, MD 20783

Editors: N. Ozay, L. Balzano, D. Panagou, A. Abate

Abstract

This paper proposes a Gradient Tracking Decentralized Unadjusted Langevin Algorithm (GT-DULA) to perform Bayesian learning via MCMC sampling. GT-DULA enhances the scalability of the process when compared with the conventional DULA as it reduces the dependence of the convergence bias on the network size by an order of magnitude for constant gradient step size. GT-DULA uses an estimate of the global gradient as a substitute for local gradients which is shared among neighbors in the network. Our theoretical analysis shows that the proposed GT-DULA successfully tracks the global gradient within a certain neighborhood, which leads to a two-fold benefit. First, the optimal mixing of the gradient estimates leads to a lower bias in convergence. Second, the successful tracking of the global gradient implies robustness towards data heterogeneity which is a major concern in decentralized learning.

Keywords: decentralized machine learning, data heterogeneity, Bayesian learning, gradient tracking, Langevin algorithm

1. Introduction

Data is an indispensable asset in the modern age of big data to exploit data-driven learning methods and extract information. However, a multitude of factors, ranging from storage hardware constraints and time constraints to unreliable communication and data privacy, often prohibit centralized data collection and processing. Consequently, a push for economic methods of data collection and storage has led to distributed data storage across different devices. To bridge the gap between data collection and data mining in such situations, distributed learning has gained significant attention in the machine learning community. Distributed learning can also be helpful to train models online with real-time data distributed across agents, instead of waiting for data centralization and subsequent offline training. The goal of distributed learning is to simultaneously train a common model on multiple agents based on distributed data without sharing the raw data.

In this paper, we focus on *decentralized* learning (without a central coordinating server), although the presented results can be extended to master-slave situations. Our approach is based on Bayesian statistics, which learns a target posterior distribution and is more resilient to overfitting than optimization methods that rely on point estimates. However, the posterior distribution cannot

be analytically computed except in a few simple cases. Hence, numerical approaches are relied upon, of which Markov Chain Monte Carlo (MCMC) Tierney and Mira (1999); Lye et al. (2019); Qian et al. (2003); Chib (2001) and Variational Inference (VI) Fox and Roberts (2012); Seeger and Wipf (2010); Grimmer (2011); Blei et al. (2017); Tzikas et al. (2008) are the common ones. Unlike VI, MCMC methods can produce samples of the exact posterior distribution asymptotically. MCMC involves initializing a random sample and iteratively converging its distribution to a posterior target via some algorithm. The algorithm used in this paper is the Unadjusted Langevin Algorithm (ULA) which is well studied in literature Ma et al. (2019); Vempala and Wibisono (2019); Geng (2024).

Conventional decentralized learning approaches often result in biases with constant learning step sizes. Annealing step sizes can circumvent this issue but involve devising additional strategies for the step sizes. In contrast, a constant step size is simple to implement, though reducing asymptotic biases is of significance. Another practically significant aspect of decentralized learning is *data heterogeneity*. In existing literature, data heterogeneity is often quantified by the variance in the gradients across the agents and its effect on convergence Tang et al. (2018); Lu and De Sa (2021); Dandi et al. (2022); Sun et al. (2023); Lin et al. (2021); Wu and Sun (2024); Le Bars et al. (2023). Failure to account for heterogeneity can lead to slower convergence or higher asymptotic biases and thus unsatisfactory learning performance.

1.1. Related Work

There exists extensive literature on the study of decentralized learning over a graph where the learning process involves optimization and results in point-estimates of the optimization variables. Earlier studies on decentralized convex optimization can be found in Tsitsiklis (1984); Nedic and Ozdaglar (2009); Wei and Ozdaglar (2012); Agarwal et al. (2010). Decentralized learning via ADMM optimization has been presented in Shi et al. (2014); Ling et al. (2016); Li et al. (2022) where decentralized optimization is treated as a constraint optimization problem with consensus achieved by enforcing an appropriate constraint. An alternative approach via dual averaging can be found in Agarwal et al. (2010); Duchi et al. (2011); Tsianos et al. (2012); Hosseini et al. (2013); Colin et al. (2016). The advantages of gradient tracking in convex optimization has been shown in Pu and Nedić (2021); Koloskova et al. (2021) while Lu et al. (2019) uses it for non-convex optimization. Gower et al. (2020); Sun et al. (2020); Jiang et al. (2022); Xin et al. (2019a) present other convergence benefits by gradient tracking in optimization. Discussion of the effect of gradient tracking on decentralized optimization with heterogeneous data can be found in Di Lorenzo and Scutari (2016); Nedic et al. (2017); Koloskova et al. (2021). More recently Huang et al. (2022); Takezawa et al. (2022); Yan et al. (2023) have explored gradient tracking in handling data heterogeneity by assuming bounds on the variance of the stochastic gradients across agents. Gradient tracking in the context of decentralized Bayesian learning has not been extensively explored. Reference Li et al. (2024) utilizes gradient tracking for variational inference while gradient tracking for Metropolis-adjusted Hamiltonian MCMC is proposed in Kungurtsev et al. (2023). However, convergence in probability of the chain is not established in Kungurtsev et al. (2023).

1.2. Contribution

The major contributions of this paper are as follows. We investigate the effect of gradient tracking on decentralized ULA without making convexity assumptions as is common in optimization literature Scaman et al. (2017); Nedic (2020); Yang et al. (2019); Xin et al. (2019b); Pu and Nedić

(2021); Koloskova et al. (2021); Liu et al. (2024); Dandi et al. (2022); Wu and Sun (2024). We instead use the log-Sobolev Inequality (LSI) assumption on the posterior distribution which is satisfied by a broader class of distributions than the log-concave assumption (which corresponds to convexity in the space of distributions) Cheng et al. (2018); Cheng and Bartlett (2018); Durmus and Moulines (2016, 2019); Dalalyan (2017b,a); Durmus and Moulines (2017). We show that GT-DULA reduces the asymptotic bias in convergence compared to existing decentralized ULA Parayil et al. (2020); Bhar et al. (2022) by an order of magnitude.

We conclusively address the issue of data heterogeneity in decentralized ULA. Typically, a bound on the variance of the individual gradients is considered a metric of the degree of heterogeneity and the dependency of the convergence on this bound is analyzed (see e.g., Tang et al. (2018); Lu and De Sa (2021); Dandi et al. (2022); Sun et al. (2023); Lin et al. (2021); Wu and Sun (2024); Le Bars et al. (2023)). In our analysis, we show that the individual gradient estimates of the agents converge to some neighborhood of the true global gradient which scales with the step size α. We derive appropriate bounds for the gradients along the topology of the samples generated and establish convergence of the gradient error up to a bias that is controlled by tuning α. Our results are supported by rigorous theoretical analysis and simulations on synthetic and real-world data.

2. Problem Formulation

Consider a decentralized learning scenario over a communication network with n agents, each having access to their respective collection of datasets $\{\boldsymbol{X}_i\}_{i=1}^n$ where $\boldsymbol{X}_i = \{x_i^j\}_{j=1}^{M_i}$ with $x_i^j \in \mathbb{R}^{d_w}$ for $i \in \mathcal{V}$. No agent has access to others' datasets, and sharing of raw data is prohibited, i.e., the i^{th} agent can access only $\boldsymbol{X}_i$. The entire collection of data is denoted as $\boldsymbol{X} = \{\boldsymbol{X}_i\}_{i=1}^n$. We assume throughout the rest of the paper that the inter-agent communication occurs over a connected undirected graph $\mathcal{G}(\mathcal{V}, \mathcal{E})$. The end goal is to train a common model parameterized by $\boldsymbol{w} \in \mathbb{R}^{d_w}$ across all agents. We focus on a Bayesian approach for this problem, where the global posterior distribution generated by a sample of $\boldsymbol{w}$ given $\boldsymbol{X}$ is proportional to the product of the likelihood $\prod_{i\in\mathcal{V}} p(\boldsymbol{X}_i|\boldsymbol{w})$ and prior $p(\boldsymbol{w})$, i.e.,

$$p(\boldsymbol{w}|\boldsymbol{X}) \propto p(\boldsymbol{w}) \prod_{i\in\mathcal{V}} p(\boldsymbol{X}_i|\boldsymbol{w}) = \prod_{i\in\mathcal{V}} p(\boldsymbol{X}_i|\boldsymbol{w}) p(\boldsymbol{w})^{\frac{1}{n}}, \tag{1}$$

where $p(\boldsymbol{X}_i|\boldsymbol{w})p(\boldsymbol{w})^{\frac{1}{n}}$ can be considered a local pseudo posterior generated by each agent's individual data. We let $p^* = p(\boldsymbol{w}|\boldsymbol{X})$, which is the target distribution.

The DULA presented in Parayil et al. (2020) provides an efficient framework for this problem by exploiting the Langevin dynamics. The posterior distribution $p(\boldsymbol{w}|\boldsymbol{X})$ is rewritten in terms of an energy function $E(\boldsymbol{w})$ which is analogous to the objective function in optimization, i.e.,

$$p(\boldsymbol{w}|\boldsymbol{X}) \propto \exp(-E(\boldsymbol{w})). \tag{2}$$

The DULA with a constant step size adapted from Parayil et al. (2020) can be written as

$$\boldsymbol{w}_{i,k+1} = \boldsymbol{w}_{i,k} - \beta \sum_{j\in\mathcal{N}_i} (\boldsymbol{w}_{i,k} - \boldsymbol{w}_{j,k}) - \alpha n \nabla E_i(\boldsymbol{w}_{i,k}) + \sqrt{2\alpha n}\boldsymbol{v}_{i,k}, \tag{3}$$

where $\boldsymbol{w}_{i,k}$ is the i-th agent's sample of the model parameter $\boldsymbol{w}$ at the k-th iteration, β and α are the constant consensus step size and gradient step size, respectively, and $\boldsymbol{v}_{i,k} \sim \mathcal{N}(\boldsymbol{0}_{d_w}, I_{d_w})$

is the injected zero-mean unit-variance Gaussian noise. The individual gradient of each agent is given by $\nabla E_i(\boldsymbol{w}_{i,k}) = -\nabla \log p(\boldsymbol{X}_i|\boldsymbol{w}_{i,k}) - \frac{1}{n}\nabla \log p(\boldsymbol{w}_{i,k})$ which combines the information from the likelihood of the local dataset $\boldsymbol{X}_i$ and the prior of $\boldsymbol{w}$. The objective for each agent is to reach consensus on $\boldsymbol{w}_{i,k}$'s (since a common model is desired across the network) and to ensure that the stationary distribution of the mean of $\boldsymbol{w}_{i,k}$'s converges to the target global posterior p^*. The relevant results for DULA in (3) (under similar conditions with constant step sizes) are presented in Section 4.4 to compare with the results for GT-DULA.

3. Methodology

In the DULA formulation (3), each agent incorporates the local learning (from local data and the prior) via the gradient term $\nabla E_i(\cdot)$ while the consensus is achieved via $\beta \sum_{j\in\mathcal{N}_i}(\boldsymbol{w}_{i,k} - \boldsymbol{w}_{j,k})$. Drawing inspiration from the decentralized optimization literature, we propose a GT-DULA that utilizes a local estimate of the global gradient $\boldsymbol{d}_{i,k}$ instead of simply the local gradient $\nabla E_i(\boldsymbol{w}_{i,k})$. Specifically, the GT-DULA takes the following form

$$\boldsymbol{w}_{i,k+1} = \boldsymbol{w}_{i,k} - \beta \sum_{j\in\mathcal{N}_i}(\boldsymbol{w}_{i,k} - \boldsymbol{w}_{j,k}) - \alpha n \boldsymbol{d}_{i,k} + \sqrt{2\alpha n}\boldsymbol{v}_{i,k}, \tag{4}$$

$$\boldsymbol{d}_{i,k+1} = \boldsymbol{d}_{i,k} - \gamma \sum_{j\in\mathcal{N}_i}(\boldsymbol{d}_{i,k} - \boldsymbol{d}_{j,k}) + \triangle \boldsymbol{d}_{i,k}, \tag{5}$$

where $\boldsymbol{d}_{i,k}$ is the estimate of the global gradient while $\triangle \boldsymbol{d}_k$ is the corresponding local update rule for the i-th agent at k-th step and γ is a constant consensus step size corresponding to the mixing of the gradient estimates. A common approach is to initialize $\boldsymbol{d}_{i,0} = \nabla E_i(\boldsymbol{w}_{i,0}), \forall i \in \mathcal{V}$ and use

$$\triangle \boldsymbol{d}_{i,k} \triangleq \nabla E_i(\boldsymbol{w}_{i,k+1}) - \nabla E_i(\boldsymbol{w}_{i,k}). \tag{6}$$

In essence, GT-DULA has two sources of mixing, one in the values of $\{\boldsymbol{w}_{i,k}\}$ directly and the other is the global gradient estimates $\{\boldsymbol{d}_{i,k}\}$. The local gradient estimates are expected to track the global gradient (which cannot be computed by any agent directly) via the mixing across the graph and some update rule (one of them given in (6)). Thus, GT-DULA is expected to make better gradient updates compared to DULA and to generalize well to heterogeneously distributed data.

Let $\mathbf{w}_k \triangleq [\boldsymbol{w}_{1,k}^\top, \ldots, \boldsymbol{w}_{n,k}^\top]^\top$, $\mathbf{d}_k \triangleq [\boldsymbol{d}_{1,k}^\top, \ldots, \boldsymbol{d}_{n,k}^\top]^\top$, $\mathbf{g}_k \triangleq [\nabla E_1(\boldsymbol{w}_{1,k})^\top, \ldots, \nabla E_n(\boldsymbol{w}_{n,k})^\top]^\top$ and $\mathbf{v}_k \triangleq [\boldsymbol{v}_{1,k}^\top, \ldots, \boldsymbol{v}_{n,k}^\top]^\top$. Denote $\mathcal{W}_\beta \triangleq I_n - \beta\mathcal{L}$ and $\mathcal{W}_\gamma \triangleq I_n - \gamma\mathcal{L}$ where $\mathcal{L}$ is the Laplacian corresponding to the graph $\mathcal{G}(\mathcal{V}, \mathcal{E})$. The GT-DULA proposed in (4) and (5) combined with the update rule in (6) can be concisely written as

$$\mathbf{w}_{k+1} = (\mathcal{W}_\beta \otimes I_{d_w})\mathbf{w}_k - \alpha n \mathbf{d}_k + \sqrt{2\alpha n}\mathbf{v}_k, \tag{7}$$

$$\mathbf{d}_{k+1} = (\mathcal{W}_\gamma \otimes I_{d_w})\mathbf{d}_k + \mathbf{g}_{k+1} - \mathbf{g}_k. \tag{8}$$

4. Theoretical Results

We introduce the assumptions used to establish the results for GT-DULA.

***Assumption* 1** *The gradients $\nabla E_i(\cdot), \forall i \in \mathcal{V}$, are* Lipschitz *continuous, i.e., for any $\boldsymbol{w}_a, \boldsymbol{w}_b \in \mathbb{R}^{d_w}$ there exists $L_i > 0$ such that*

$$\|\nabla E_i(\boldsymbol{w}_a) - \nabla E_i(\boldsymbol{w}_b)\| \leq L_i \|\boldsymbol{w}_a - \boldsymbol{w}_b\|. \tag{9}$$

We let $L^2 = \max_{i\in\mathcal{V}}\{L_i^2\}$ and $\bar{L}^2 = \sum_{i\in\mathcal{V}} L_i^2$.

***Assumption* 2** *The communication topology of the n networked agents is a connected undirected graph $\mathcal{G}(\mathcal{V},\mathcal{E})$.*

***Assumption* 3** *The target distribution p^* satisfies the log-Sobolev inequality (LSI) condition, i.e., for any function $g(\bar{\boldsymbol{w}})$ with $\mathbb{E}_{p^*}[g(\bar{\boldsymbol{w}})] = 1$, there exists a constant $\rho_U > 0$ such that the following condition is satisfied:*

$$\mathbb{E}_{p^*}\left[g(\bar{\boldsymbol{w}})\log g(\bar{\boldsymbol{w}})\right] \leq \frac{1}{2\rho_U}\mathbb{E}_{p^*}\left[\frac{\|\nabla g(\bar{\boldsymbol{w}})\|^2}{g(\bar{\boldsymbol{w}})}\right]. \tag{10}$$

Assumption 1 implies smoothness of the log of the posterior function, Assumption 2 implies sufficient connectivity of the graph to ensure consensus and Assumption 3 is needed for the convergence of the distribution. As noted in Ma et al. (2021), when $g = \frac{p}{p^*}$, the LSI condition is similar to the Polyak-Łojasiewicz condition used in optimization theory.

Denote by λ_2 and λ_n the second smallest and the largest eigenvalue of the Laplacian $\mathcal{L}$ of the network, respectively. Below are the conditions that the hyperparameters need to satisfy in order for our results to hold.

***Condition* 1** *(i)* $\beta < \min\left\{1, \frac{n}{\lambda_n}\right\}$, *(ii)* $\alpha < \min\left\{\frac{\beta\lambda_2}{2n\bar{L}}, \frac{\rho_U}{4n\bar{L}^2}\right\}$, *(iii)* $\gamma < \min\left\{\beta, \frac{1}{\lambda_2}\right\}$.

Condition 1*(i)* gives an upper bound on β. Once an appropriate value of β that satisfies Condition 1*(i)* is chosen, permissible values of α and γ can be selected to satisfy Condition 1*(ii)* and Condition 1*(iii)* respectively.

We next present the key results of the proposed algorithm (7)–(8), which have three main components, namely, consensus, convergence, and gradient error, as outlined in Sections 4.1, 4.2 and 4.3, respectively.

4.1. Consensus of the samples

Since the aim is to collectively learn a common model, it is vital to establish consensus in the value of $\{\boldsymbol{w}_i\}$ sampled by each agent. To that end, we define the consensus error as

$$\tilde{\mathbf{w}}_k \triangleq \mathbf{w}_k - (\mathbf{1}_n \otimes \bar{\boldsymbol{w}}_k) = \left(\left(I_n - \frac{1}{n}\mathbf{1}_n\mathbf{1}_n\right)\otimes I_{d_w}\right)\mathbf{w}_k, \tag{11}$$

where $\bar{\boldsymbol{w}}_k \triangleq \frac{1}{n}\sum_{i\in\mathcal{V}}\boldsymbol{w}_{i,k}$. Defining $\Psi_k \triangleq [\tilde{\mathbf{w}}_k^\top, \alpha\mathbf{d}_k^\top]^\top \in \mathbb{R}^{2nd_w}$, we present the consensus result in Theorem 1. The proofs for all the theorems are given in the Appendix available online[1].

Theorem 1 *Suppose that Assumptions 1–2 hold and Condition 1 is satisfied. The consensus error $\tilde{\mathbf{w}}_{k+1}$ satisfies $\mathbb{E}[\|\tilde{\mathbf{w}}_{k+1}\|^2] \leq \mathbb{E}[\|\Psi_{k+1}\|^2]$ where*

$$\mathbb{E}[\|\Psi_{k+1}\|^2] \leq \sigma^{k+1}\mathbb{E}[\|\Psi_0\|^2] + B_{GT}, \tag{12}$$

in which

$$\sigma \triangleq 1 - \beta\lambda_2 + \frac{4\alpha^2n^2L^2}{\beta\lambda_2} < 1, \tag{13}$$

$$B_{GT} \triangleq \frac{2\alpha n^2 d_w(1+2\alpha^2L^2)}{\beta^2\lambda_2^2 - 4\alpha^2n^2L^2}. \tag{14}$$

Theorem 1 proves that consensus is achieved at an exponential rate up to a bias B_{GT}. Smaller network size n leads to faster consensus as well as lower B_{GT}, while lower problem dimensionality d_w lowers B_{GT} only. Lower α and higher β increase the rate of consensus and reduce B_{GT}.

4.2. Convergence of the posterior

The *convergence in distribution* of the GT-DULA is analyzed via the KL divergence of the distribution of the average sample $\bar{\boldsymbol{w}}_k$ (denoted by $p(\bar{\boldsymbol{w}}_k)$) from the target posterior p^*. The KL divergence, denoted by $F(p(\bar{\boldsymbol{w}}))$, is defined as

$$F(p(\bar{\boldsymbol{w}})) \triangleq \int p(\bar{\boldsymbol{w}}) \log\left(\frac{p(\bar{\boldsymbol{w}})}{p^*(\bar{\boldsymbol{w}})}\right) d\bar{\boldsymbol{w}}. \tag{15}$$

Theorem 2 *Suppose that Assumptions 1-3 hold. Under Condition 1,* $F\Big(p(\bar{\boldsymbol{w}}_{k+1})\Big)$ *satisfies*

$$F\Big(p(\bar{\boldsymbol{w}}_{k+1})\Big) \leq \Big(F\Big(p(\bar{\boldsymbol{w}}_0)\Big) + C_{_F}\Big)\exp(-\alpha\rho_U(k+1)) + C_{_F}\sigma^{k+1} + O_{_{GT}} \tag{16}$$

where

$$C_{_F} \triangleq \frac{(2n^2\alpha^2L^2\bar{L}^2 + nL^2)\alpha}{|\sigma - \exp(-\alpha\rho_U)|}\mathbb{E}\|\Psi_0\|^2, \tag{17}$$

$$O_{_{GT}} \triangleq \frac{2n\alpha\bar{L}^2d_w + 4n^2\alpha^2\bar{L}^4(C_{_{\bar{\boldsymbol{w}}}} + C_{_{\boldsymbol{w}^*}})}{\rho_U} + \frac{2\alpha n^3d_wL^2(2n\alpha^2\bar{L}^2+1)(1+2\alpha^2L^2)}{\rho_U(\beta^2\lambda_2^2 - 4\alpha^2n^2L^2)} \tag{18}$$

in which $C_{_{\bar{\boldsymbol{w}}}}$ *is an upper bound of* $\mathbb{E}[\|\bar{\boldsymbol{w}}_k\|^2]$, $\forall k \geq 0$ *and* $C_{_{\boldsymbol{w}^*}}$ *is an upper bound of* $\mathbb{E}[\|\boldsymbol{w}^*\|^2]$, *where* $\boldsymbol{w}^*$ *is some local extremum of* $\sum_{i\in\mathcal{V}}\nabla E_i(\cdot)$.

Theorem 2 shows exponential convergence of the KL divergence up to a bias $O_{_{GT}}$. A smaller network size n, smaller Lipschitz constants and higher ρ_U help with faster convergence rates and reduce $O_{_{GT}}$, while smaller dimensionality d_w reduces $O_{_{GT}}$ only. Higher α and β are likely to increase convergence rate while for reducing $O_{_{GT}}$ it helps to have a smaller α and larger β. Please refer to Lemma 5 in the Appendix[1] for the proof of the existence of $C_{_{\bar{\boldsymbol{w}}}}$.

Corollary 3 *Suppose that Assumptions 1-3 hold. Under Condition 1, for any* $\epsilon \in (0,1)$ *the requirements to ensure* $F\Big(p(\bar{\boldsymbol{w}}_{k+1})\Big) \leq \epsilon$ *are given by*

$$\alpha \leq \min\left\{\left(\frac{\beta^2\lambda_2^2\rho_U}{40(L^4\bar{L}^2d_w + L^4d_w)}\frac{\epsilon}{n^4}\right)^{\frac{1}{3}}, \left(\frac{\beta^2\lambda_2^2\rho_U}{20\bar{L}^4(C_{_{\bar{\boldsymbol{w}}}} + C_{_{\boldsymbol{w}^*}}) + 20L^2\bar{L}^2d_w + 12\rho_UL^2}\frac{\epsilon}{n^4}\right)^{\frac{1}{2}}, \left(\frac{\beta^2\lambda^2\rho_U}{10(\bar{L}^2d_w + L^2d_w)}\frac{\epsilon}{n^3}\right)\right\} \sim \mathcal{O}\left(\frac{\epsilon}{n^3}\right), \tag{19}$$

$$k \geq \max\left\{\frac{\log\Big(5(F(p(\bar{\boldsymbol{w}}_0)) + C_{_F})/\epsilon\Big)}{\alpha\rho_U} - 1, \frac{\log(5C_{_F}/\epsilon)}{\log(1/\sigma)} - 1\right\} \sim \mathcal{O}\left(\frac{n^3\log(1/\epsilon)}{\epsilon}\right). \tag{20}$$

The convergence to the ϵ-neighborhood of GT-DULA is $\mathcal{O}\left(\frac{\log(1/\epsilon)}{\epsilon}\right)$ which is the same as the centralized ULA with constant step size in Vempala and Wibisono (2019) under similar assumptions.

1. https://coral-osu.github.io/assets/pdf/GT_DULA_L4DC_2025_Appendix.pdf

4.3. Convergence of gradient estimates

Define the true global gradient as $\bar{\boldsymbol{u}}_k \triangleq \frac{1}{n}\sum_{i\in\mathcal{V}} \nabla E_i(\bar{\boldsymbol{w}}_k)$ and the gradient error as $\tilde{\mathbf{d}}_k \triangleq \mathbf{d}_k - \mathbf{1}_n \otimes \bar{\boldsymbol{u}}_k$. Our final result for the gradient error is presented in Theorem 4.

Theorem 4 *Under Assumptions 1-2 and Condition 1, the gradient error $\tilde{\mathbf{d}}_k$ satisfies*

$$\begin{aligned}\mathbb{E}[\|\tilde{\mathbf{d}}_{k+1}\|^2] \leq \left[\mathbb{E}[\|\tilde{\mathbf{d}}_0\|^2] + \frac{8nL^2(n+\alpha^2\bar{L}^2)}{\gamma\lambda_2|\sigma+\gamma\lambda_2-1|}\mathbb{E}\|\Psi_0\|^2\right](1-\gamma\lambda_2)^{k+1} \\ + \frac{8nL^2(n+\alpha^2\bar{L}^2)\mathbb{E}\|\Psi_0\|^2}{\gamma\lambda_2|\sigma+\gamma\lambda_2-1|}\sigma^{k+1} + E_{GT},\end{aligned} \tag{21}$$

where

$$E_{GT} = \frac{\left(16n\alpha^2c_4^4(C_{\bar{\boldsymbol{w}}} + C_{\boldsymbol{w}^*}) + 8n\alpha d_w(nL^2+\bar{L}^2)\right)}{\gamma^2\lambda_2^2} + \frac{16\alpha n^3L^2(n+\alpha^2\bar{L}^2)(1+2\alpha^2L^2)d_w}{(\beta^2\lambda_2^2 - 4\alpha^2n^2L^2)\gamma^2\lambda_2^2}. \tag{22}$$

Theorem 4 shows an exponential reduction of the gradient error up to a bias E_{GT}. Smaller n and smaller Lipschitz constants are likely to increase the rate and reduce E_{GT}. Lower d_w reduces E_{GT} without affecting the rate. Reducing α and increasing β, γ improve the rate as well as reduce E_{GT}. The true global gradient $\bar{\boldsymbol{u}}_k$ is the ideal gradient required for convergence of the mean dynamics $\bar{\boldsymbol{w}}$, but inaccessible by any agent. Thus, convergence of the gradient estimates $\boldsymbol{d}_{i,k}$ to $\bar{\boldsymbol{u}}_k$, $\forall i \in \mathcal{V}$, implies GT-DULA's robustness towards data heterogeneity in the training data. Since E_{GT} scales with α, the effects of data heterogeneity can be mitigated by appropriately tuning α in GT-DULA.

4.4. Comparison of GT-DULA with DULA and Discussion

The table below compares the asymptotic biases in the consensus error (B), KL divergence (O) and gradient error (E) between GT-DULA and DULA. Since $\alpha \sim \mathcal{O}\left(\frac{1}{n}\right)$ from Condition 1*(i)*, B for both GT-DULA and DULA scales with $\mathcal{O}(n)$. However, DULA has an additional term that scales with $\mathcal{O}(n)$ when compared to GT-DULA. Thus, the $\mathcal{O}(n)$ terms in B of DULA are larger than that in GT-DULA. In the case of O, GT-DULA scales with $\mathcal{O}(n^2)$ which is an order of magnitude improvement over DULA which scales with $\mathcal{O}(n^3)$. Thus, the asymptotic biases of GT-DULA scale better with the graph size than DULA. By choosing a small α, the E in GT-DULA can be made arbitrarily small, making it robust for data heterogeneity. However, the E in DULA contains a term independent of α, which implies the existence of a fixed bias that depends solely on the specific problem parameters. Also note that the degree of heterogeneity plays a role through L^2 and $\bar{L}^2$ which are expected to be larger for higher data heterogeneity and vice-versa.

The consensus result for GT-DULA has a bias ($B_{GT} > 0$) which is different from some literature on gradient tracking based distributed optimization, e.g., Pu and Nedić (2021). The B_{GT} stems directly from the injected Gaussian noise which is the characteristics of the Langevin algorithm but absent in optimization. This noise does not vanish for a constant step size and manifests as the bias B_{GT}. The bias O_{GT} results from the Gaussian noise, the consensus error bias, and the discretization error from the SDE in (38) (via the bounds obtained for the gradients). Likewise, E_{GT} is a result of the Gaussian noise, the consensus error bias, and the bounds of the gradients. Convergence to the (unique) minima as the convergence metric in strongly convex optimization helps circumvent

the biases that result from the bounds of the gradients. However, for GT-DULA, we employ the KL divergence to measure the convergence of the posterior rather than convergence of the samples to a local minima.

	GT-DULA	DULA
B	$\frac{2\alpha n^2 d_w(1+2\alpha^2 L^2)}{\beta^2\lambda_2^2-4\alpha^2 n^2 L^2}$	$\frac{2\alpha(4\alpha n^3 L^2 C_{\bar{\boldsymbol{w}}}+4\alpha n^2 L^2 C_{\hat{\mathbf{W}}^*}+2n^2 d_w)}{\beta^2\lambda_2^2-4\alpha^2 n^2 L^2}$
O	$\frac{1}{\rho_U}\left(2n\alpha\bar{L}^2 d_w + 4n^2\alpha^2\bar{L}^4(C_{\bar{\boldsymbol{w}}}+C_{\boldsymbol{w}^*})\right)$ $+\frac{2\alpha n^2 d_w(2n^2\alpha^2 L^2\bar{L}^2+nL^2)(1+2\alpha^2 L^2)}{\rho_U(\beta^2\lambda_2^2-4\alpha^2 n^2 L^2)}$	$\frac{1}{\rho_U}\left(2n\alpha\bar{L}^2 d_w + 4n^2\alpha^2\bar{L}^4(C_{\bar{\boldsymbol{w}}}+C_{\boldsymbol{w}^*})\right)$ $+\frac{(2n^2\alpha^2 L^2\bar{L}^2+nL^2)(2\alpha n^3 L^2 C_{\bar{\boldsymbol{w}}}+2\alpha n^2 L^2 C_{\hat{\mathbf{W}}^*}+n^2 d_w)}{\rho_U(\beta^2\lambda_2^2-4\alpha^2 n^2 L^2)}$
E	$\frac{16\alpha n^3 L^2(n+\alpha^2\bar{L}^2)(1+2\alpha^2 L^2)d_w}{(\beta^2\lambda_2^2-4\alpha^2 n^2 L^2)\gamma^2\lambda_2^2}$ $+\frac{1}{\gamma^2\lambda_2^2}\left(16n\alpha^2 c_4^4(C_{\bar{\boldsymbol{w}}}+C_{\boldsymbol{w}^*})+8n\alpha d_w(nL^2+\bar{L}^2)\right)$	$\frac{4\alpha L^2(4\alpha n^3 L^2 C_{\bar{\boldsymbol{w}}}+4\alpha n^2 L^2 C_{\hat{\mathbf{W}}^*}+2n^2 d_w)}{\beta^2\lambda_2^2-4\alpha^2 n^2 L^2}$ $+4L^2(nC_{\bar{\boldsymbol{w}}}+C_{\hat{\mathbf{W}}^*})$

5. Numerical Simulations

In this section we present simulation results performed on synthetic and real world data. All the results of GT-DULA are benchmarked against the corresponding results from DULA.

5.1. 1D Gaussian Problem

Consider the simple 1D Gaussian problem presented in Teh et al. (2016). Let $\theta \sim \mathcal{N}(0, \sigma_\theta^2)$ and the data for each agent is generated as $x_i|\theta \sim \mathcal{N}(\theta, \sigma_i^2)$ for $i = \{1, 2, \ldots, 5\}$; where $\sigma_i = \{10, 5, 16, 2, 18\}$ with 80 data points for each σ_i value (200 in total). The analytical expression of the true posterior from all the data is given by $\pi = \mathcal{N}\left(\frac{\sum_{i=1}^N x_i}{\frac{\sigma_x^2}{\sigma_\theta^2}+N}, \left(\frac{1}{\sigma_\theta^2}+\frac{N}{\sigma_x^2}\right)^{-1}\right)$. The data is distributed across 5 agents with the i-th agent receiving $\{x_i\}$ to simulate data heterogeneity. The graph is a ring graph. Figure 1(a) presents a comparison of the KL divergence between GT-DULA and DULA from MCMC simulations and analytical results. It shows a good match between the analytical and simulation results and clearly proves that the GT-DULA results in a lower bias than DULA. Next, we ran experiments with $n = 5$ and starting with $\alpha = 1e-2$ reduced α by $1/10$-th every 500 iterations for a total 3000 iterations to show the effect of reducing α on the asymptotic KL divergence and the gradient error which are shown in Figure 1(b) and (c) respectively. Figure 1(b) compares the effect of α on the asymptotic KL divergence between DULA and GT-DULA. It shows that while the asymptotic KL divergence of GT-DULA is lower than the DULA, they both decrease with every decrement of α. However, beyond 1500 iterations, the decrements in Figure 1(b) are less noticeable due to the slow convergence of the KL divergence at the low values of α. Figure 1(c) compares the corresponding effect on the gradient errors between DULA and GT-DULA. The gradient error of DULA saturates at around 20 and is unaffected by reducing α, while that in GT-DULA keeps reducing with every reduction in α.

Figure 2 shows how the asymptotic KL divergence scales with network size n. We split the 400 data points randomly into $n = \{5, 10, 20, 25, 40, 50, 80, 100, 200\}$ groups to simulate distributed data. For each n, 1000 *random* splits were performed and the asymptotic KL divergence value for each random split was computed analytically. Figure 2(a) compares the expected asymptotic KL divergence for every network size, which was obtained by taking the mean of all the 1000 cases for

each n. It clearly shows that although the expected asymptotic bias in the KL divergence increases with n, it is lower for GT-DULA than DULA in each corresponding case, and the gap increases with n. Figure 2(b) shows the distribution of the asymptotic KL divergence for different n. The lower variance in the asymptotic KL divergence of GT-DULA compared to DULA in Figure 2(b) for each n is indicative of GT-DULA's better handling of data heterogeneity, since each of the cases 1000 encompasses varying degrees of heterogeneity due to the random split of data. This benefit is also more pronounced with increasing n.

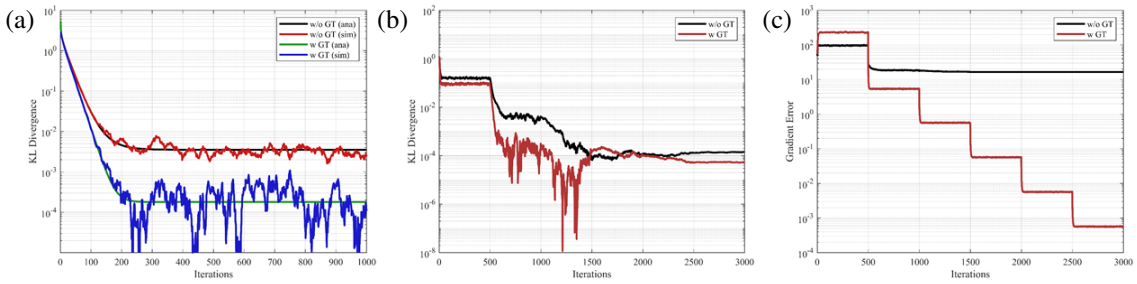

Figure 1: (a) Comparison of the KL divergence between GT-DULA with DULA via analytical and simulation results. (b) asymptotic KL divergence and (c) the average gradient error.

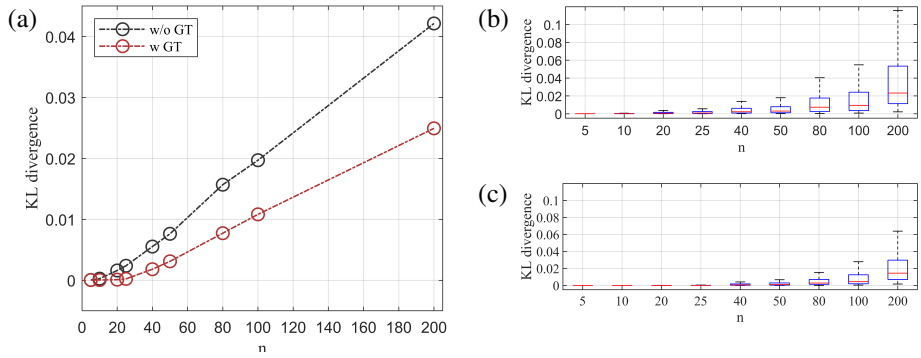

Figure 2: (a) Expected asymptotic KL divergence values for the posteriors generated by DULA vs GT-DULA for different network sizes. Distribution of the asymptotic KL divergence values for the posteriors generated by (b) DULA and (c) GT-DULA for different network sizes.

5.2. Multi-class Classification

In this section, we show experimental results for digit identification on the MNIST dataset. The training data was heterogeneously distributed among 10 agents (with a ring graph) where the i-th agent received 4500 training samples of the digit $i \in \{0, 1, \ldots, 9\}$ and 100 samples of each of other digits. For testing the standard test dataset for MNIST was used. Each agent uses LeNet-5 LeCun et al. (1998) initialized randomly with Kaiming uniform prior on the parameters of the network

and uses a stochastic mini-batch of size of about 5% of the full batch for computing the gradients. We empirically noticed that GT-DULA performs optimally at a higher step size (10^{-4}) than DULA (10^{-6}), the latter becoming unstable at 10^{-4}. We kept $\beta = 0.6$ the same for both algorithms while $\gamma = \beta$ was used in GT-DULA. We ran another experiment where both GT-DULA and DULA started with their corresponding optimal step sizes, which were reduced by a factor of $1/10$ after certain iterations (with the same β and γ values as earlier). We noticed that this strategy slightly improved the performance of GT-DULA while deteriorated the performance achieved by DULA in comparison to keeping a constant step size throughout. The results for the aforementioned experiments are presented in Figure 3 where the solid lines represent the mean values of all agents with the shaded region representing 1 standard deviation (SD) of the inter-agent variations. Figure 3(a) shows that although the steady-state accuracy is similar, it converges significantly faster for GT-DULA than DULA owing to its larger step size. Additionally, the consensus (marked by the shaded region) is better for GT-DULA than DULA. Furthermore, for GT-DULA the consensus improves significantly by reducing the step size although the steady-state accuracy values are similar. Figure 3(b) shows the faster convergence of GT-DULA's training loss compared to DULA. The strategy of reducing step size appears to slightly under-perform for GT-DULA while it drastically under-performs for DULA, both in the consensus and accuracy. Thus, we conclude from Figure 3 that for the MNIST problem, a constant step size can be found for both GT-DULA and DULA for optimal performance. However, due to the larger optimal step size of GT-DULA, it performs significantly faster with better consensus than DULA. The achieved asymptotic accuracy on the test data is slightly less than that in (Parayil et al., 2020, Table 1) since they did not use heterogeneous training data.

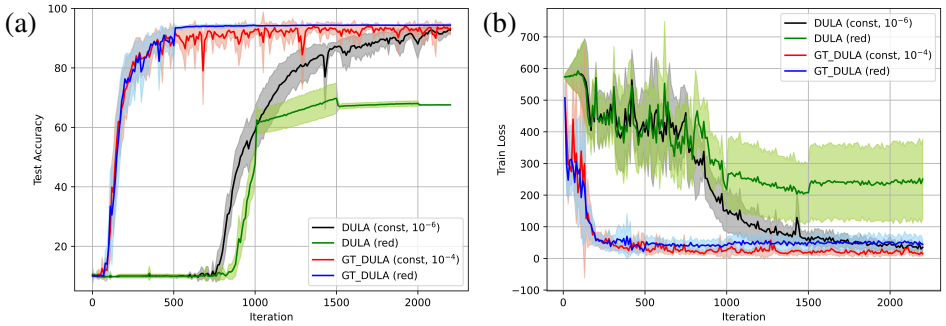

Figure 3: (a) Mean test data accuracy (b) Mean training loss for heterogeneously distributed MNIST dataset across 10 agents. Here, 'const' refers to constant step size throughout the simulation, while 'red' refers to the reduced step size case.

6. Conclusions

We introduce GT-DULA for decentralized Bayesian learning with constant step sizes. Different from gradient tracking based optimization, GT-DULA exhibits convergence biases due to the injected Gaussian noise for Bayesian learning and bounding the gradients. When compared with DULA, GT-DULA scales better with the network size and is more robust to data heterogeneity. The theoretical results are established rigorously under minimal conditions (Lipschitz continuity and LSI) and supported by simulation results.

Acknowledgments

This work was partially supported by the U.S. DEVCOM Army Research Laboratory (ARL) under Cooperative Agreement W911NF2120219 and National Science Foundation (NSF) award # 2241585. J. George's contributions are based upon work while serving at the NSF. Any opinion, findings, and conclusions or recommendations expressed in this material are those of the authors and should not be interpreted as representing the official policies, either expressed or implied, of ARL, NSF or the U.S. Government.

References

Alekh Agarwal, Martin J Wainwright, and John C Duchi. Distributed dual averaging in networks. *Advances in Neural Information Processing Systems*, 23, 2010.

Kinjal Bhar, He Bai, Jemin George, and Carl Busart. Asynchronous Bayesian learning over a network. In *2022 IEEE 61st Conference on Decision and Control (CDC)*, pages 2393–2398. IEEE, 2022.

David M Blei, Alp Kucukelbir, and Jon D McAuliffe. Variational inference: A review for statisticians. *Journal of the American statistical Association*, 112(518):859–877, 2017.

Xiang Cheng and Peter Bartlett. Convergence of Langevin MCMC in KL-divergence. In *Algorithmic Learning Theory*, pages 186–211. PMLR, 2018.

Xiang Cheng, Niladri S Chatterji, Peter L Bartlett, and Michael I Jordan. Underdamped Langevin MCMC: A non-asymptotic analysis. In *Conference on learning theory*, pages 300–323. PMLR, 2018.

Siddhartha Chib. Markov chain monte carlo methods: computation and inference. *Handbook of econometrics*, 5:3569–3649, 2001.

Igor Colin, Aurélien Bellet, Joseph Salmon, and Stéphan Clémençon. Gossip dual averaging for decentralized optimization of pairwise functions. In *International conference on machine learning*, pages 1388–1396. PMLR, 2016.

Arnak Dalalyan. Further and stronger analogy between sampling and optimization: Langevin Monte Carlo and gradient descent. In *Conference on Learning Theory*, pages 678–689. PMLR, 2017a.

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, 2017b.

Yatin Dandi, Anastasia Koloskova, Martin Jaggi, and Sebastian U Stich. Data-heterogeneity-aware mixing for decentralized learning. *arXiv preprint arXiv:2204.06477*, 2022.

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.

John C Duchi, Alekh Agarwal, and Martin J Wainwright. Dual averaging for distributed optimization: Convergence analysis and network scaling. *IEEE Transactions on Automatic control*, 57(3): 592–606, 2011.

Alain Durmus and Éric Moulines. Sampling from a strongly log-concave distribution with the unadjusted Langevin algorithm. *arXiv preprint arXiv:1605.01559*, 2016.

Alain Durmus and Eric Moulines. Nonasymptotic convergence analysis for the unadjusted Langevin algorithm. 2017.

Alain Durmus and Eric Moulines. High-dimensional Bayesian inference via the unadjusted Langevin algorithm. 2019.

Charles W Fox and Stephen J Roberts. A tutorial on variational Bayesian inference. *Artificial intelligence review*, 38:85–95, 2012.

Ruoming Geng. Ergodic foundations of Langevin-based MCMC. *International Journal of Applied Science*, 7(2):p8–p8, 2024.

Robert M Gower, Mark Schmidt, Francis Bach, and Peter Richtárik. Variance-reduced methods for machine learning. *Proceedings of the IEEE*, 108(11):1968–1983, 2020.

Justin Grimmer. An introduction to Bayesian inference via variational approximations. *Political Analysis*, 19(1):32–47, 2011.

Saghar Hosseini, Airlie Chapman, and Mehran Mesbahi. Online distributed optimization via dual averaging. In *52nd IEEE Conference on Decision and Control*, pages 1484–1489. IEEE, 2013.

Yan Huang, Ying Sun, Zehan Zhu, Changzhi Yan, and Jinming Xu. Tackling data heterogeneity: A new unified framework for decentralized SGD with sample-induced topology. *arXiv preprint arXiv:2207.03730*, 2022.

Xia Jiang, Xianlin Zeng, Jian Sun, and Jie Chen. Distributed stochastic gradient tracking algorithm with variance reduction for non-convex optimization. *IEEE Transactions on Neural Networks and Learning Systems*, 34(9):5310–5321, 2022.

Anastasiia Koloskova, Tao Lin, and Sebastian U Stich. An improved analysis of gradient tracking for decentralized machine learning. *Advances in Neural Information Processing Systems*, 34: 11422–11435, 2021.

Vyacheslav Kungurtsev, Adam Cobb, Tara Javidi, and Brian Jalaian. Decentralized Bayesian learning with Metropolis-adjusted Hamiltonian Monte Carlo. *Machine Learning*, 112(8):2791–2819, 2023.

Batiste Le Bars, Aurélien Bellet, Marc Tommasi, Erick Lavoie, and Anne-Marie Kermarrec. Refined convergence and topology learning for decentralized SGD with heterogeneous data. In *International Conference on Artificial Intelligence and Statistics*, pages 1672–1702. PMLR, 2023.

Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998.

Qing Li, Runze Gan, and Simon Godsill. Decentralised variational inference frameworks for multi-object tracking on sensor network. *arXiv preprint arXiv:2408.13689*, 2024.

Qunwei Li, Bhavya Kailkhura, Ryan Goldhahn, Priyadip Ray, and Pramod K Varshney. Robust decentralized learning using ADMM with unreliable agents. *IEEE Transactions on Signal Processing*, 70:2743–2757, 2022.

Tao Lin, Sai Praneeth Karimireddy, Sebastian U Stich, and Martin Jaggi. Quasi-global momentum: Accelerating decentralized deep learning on heterogeneous data. *arXiv preprint arXiv:2102.04761*, 2021.

Qing Ling, Yaohua Liu, Wei Shi, and Zhi Tian. Weighted ADMM for fast decentralized network optimization. *IEEE Transactions on Signal Processing*, 64(22):5930–5942, 2016.

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

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

Yucheng Lu and Christopher De Sa. Optimal complexity in decentralized training. In *International conference on machine learning*, pages 7111–7123. PMLR, 2021.

Adolphus Lye, Alice Cicirello, and Edoardo Patelli. A review of stochastic sampling methods for Bayesian inference problems. In *29th European Safety and Reliability Conference, ESREL 2019*, pages 1866–1873, 2019.

Yi-An Ma, Yuansi Chen, Chi Jin, Nicolas Flammarion, and Michael I Jordan. Sampling can be faster than optimization. *Proceedings of the National Academy of Sciences*, 116(42):20881–20885, 2019.

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

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

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

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.

Anjaly Parayil, He Bai, Jemin George, and Prudhvi Gurram. Decentralized Langevin dynamics for Bayesian learning. *Advances in Neural Information Processing Systems*, 33:15978–15989, 2020.

Shi Pu and Angelia Nedić. Distributed stochastic gradient tracking methods. *Mathematical Programming*, 187(1):409–457, 2021.

Song S Qian, Craig A Stow, and Mark E Borsuk. On monte carlo methods for Bayesian inference. *Ecological modelling*, 159(2-3):269–277, 2003.

Kevin Scaman, Francis Bach, Sébastien Bubeck, Yin Tat Lee, and Laurent Massoulié. Optimal algorithms for smooth and strongly convex distributed optimization in networks. In *international conference on machine learning*, pages 3027–3036. PMLR, 2017.

Matthias W Seeger and David P Wipf. Variational Bayesian inference techniques. *IEEE Signal Processing Magazine*, 27(6):81–91, 2010.

Wei Shi, Qing Ling, Kun Yuan, Gang Wu, and Wotao Yin. On the linear convergence of the ADMM in decentralized consensus optimization. *IEEE Transactions on Signal Processing*, 62(7):1750–1761, 2014.

Haoran Sun, Songtao Lu, and Mingyi Hong. Improving the sample and communication complexity for decentralized non-convex optimization: Joint gradient estimation and tracking. In *International conference on machine learning*, pages 9217–9228. PMLR, 2020.

Yuchang Sun, Jiawei Shao, Yuyi Mao, Jessie Hui Wang, and Jun Zhang. Semi-decentralized federated edge learning with data and device heterogeneity. *IEEE Transactions on Network and Service Management*, 20(2):1487–1501, 2023.

Yuki Takezawa, Han Bao, Kenta Niwa, Ryoma Sato, and Makoto Yamada. Momentum tracking: Momentum acceleration for decentralized deep learning on heterogeneous data. *arXiv preprint arXiv:2209.15505*, 2022.

Hanlin Tang, Xiangru Lian, Ming Yan, Ce Zhang, and Ji Liu. D^2: Decentralized training over decentralized data. In *International Conference on Machine Learning*, pages 4848–4856. PMLR, 2018.

Y Teh, Alexandre Thiéry, and Sebastian Vollmer. Consistency and fluctuations for stochastic gradient Langevin dynamics. *Journal of Machine Learning Research*, 17, 2016.

Luke Tierney and Antonietta Mira. Some adaptive monte carlo methods for Bayesian inference. *Statistics in medicine*, 18(17-18):2507–2515, 1999.

Konstantinos I Tsianos, Sean Lawlor, and Michael G Rabbat. Push-sum distributed dual averaging for convex optimization. In *2012 ieee 51st ieee conference on decision and control (cdc)*, pages 5453–5458. IEEE, 2012.

John N Tsitsiklis. *Problems in decentralized decision making and computation*. PhD thesis, Massachusetts Institute of Technology, 1984.

Dimitris G Tzikas, Aristidis C Likas, and Nikolaos P Galatsanos. The variational approximation for Bayesian inference. *IEEE Signal Processing Magazine*, 25(6):131–146, 2008.

Santosh Vempala and Andre Wibisono. Rapid convergence of the unadjusted Langevin algorithm: Isoperimetry suffices. *Advances in neural information processing systems*, 32, 2019.

Ermin Wei and Asuman Ozdaglar. Distributed alternating direction method of multipliers. In *2012 IEEE 51st IEEE Conference on Decision and Control (CDC)*, pages 5445–5450. IEEE, 2012.

Tongle Wu and Ying Sun. The effectiveness of local updates for decentralized learning under data heterogeneity. *arXiv preprint arXiv:2403.15654*, 2024.

Ran Xin, Soummya Kar, and Usman A Khan. An introduction to decentralized stochastic optimization with gradient tracking. *arXiv preprint arXiv:1907.09648*, 2019a.

Ran Xin, Anit Kumar Sahu, Usman A Khan, and Soummya Kar. Distributed stochastic optimization with gradient tracking over strongly-connected networks. In *2019 IEEE 58th Conference on Decision and Control (CDC)*, pages 8353–8358. IEEE, 2019b.

Yonggui Yan, Jie Chen, Pin-Yu Chen, Xiaodong Cui, Songtao Lu, and Yangyang Xu. Compressed decentralized proximal stochastic gradient method for nonconvex composite problems with heterogeneous data. In *International Conference on Machine Learning*, pages 39035–39061. PMLR, 2023.

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.

Scalability Enhancement and Data-Heterogeneity Awareness in Gradient Tracking based Decentralized Bayesian Learning (Appendix)

Proof of Theorem 1

Pre-multiplying (7) with $(M \otimes I_{d_w})$ where $M = I_n - \frac{1}{n}\mathbf{1}_n\mathbf{1}_n^\top$ the combined dynamics of $\Psi_k \triangleq [\tilde{\mathbf{w}}_k^\top, \alpha\mathbf{d}_k^\top]^\top \in \mathbb{R}^{2nd_w}$ from (7) and (8) can be written as

$$\Psi_{k+1} = \mathcal{W}\Psi_k + \mathbf{e}_k \tag{23}$$

where $\mathcal{W} \triangleq \begin{bmatrix} (\mathcal{W}_\beta \otimes I_{d_w}) & -n(M \otimes I_{d_w}) \\ \mathbf{0}_{nd_w} & (\mathcal{W}_\gamma \otimes I_{d_w}) \end{bmatrix}$ and $\mathbf{e}_k \triangleq \begin{bmatrix} \sqrt{2\alpha n}(M \otimes I_{d_w})\mathbf{v}_k \\ \alpha(\mathbf{g}_{k+1} - \mathbf{g}_k) \end{bmatrix}$. Taking the norm of (23) yields

$$\|\Psi_{k+1}\| \leq (1 - \beta\lambda_2)\|\Psi_k\| + \|\mathbf{e}_k\|, \tag{24}$$

where we use the results $\|\mathcal{W}\| \leq 1 - \beta\lambda_2$. Also, we have

$$\|\mathbf{e}_k\|^2 = 2\alpha n\|\mathbf{v}_k\|^2 + \alpha^2\|\mathbf{g}_{k+1} - \mathbf{g}_k\|^2 \tag{25}$$

where we used $\|M \otimes I_{d_w}\| \leq 1$. Squaring (24) and applying the identity $(x+y)^2 \leq (\theta+1)x^2 + \left(\frac{\theta+1}{\theta}\right)y^2$ for any $\theta > 0$ (with $\theta = (1-\beta\lambda_2)^{-1} - 1 > 0$) yields

$$\|\Psi_{k+1}\|^2 \leq (1 - \beta\lambda_2)\|\Psi_k\|^2 + \frac{1}{\beta\lambda_2}\|\mathbf{e}_k\|^2, \tag{26}$$

$$\leq (1 - \beta\lambda_2)\|\Psi_k\|^2 + \frac{1}{\beta\lambda_2}\Big(2\alpha n\|\mathbf{v}_k\|^2 + \alpha^2\|\mathbf{g}_{k+1} - \mathbf{g}_k\|^2\Big), \tag{27}$$

where we substituted (25) in (27). Next, we need to establish a bound for $\|\mathbf{g}_{k+1} - \mathbf{g}_k\|^2$. We have $\|\mathbf{g}_{k+1} - \mathbf{g}_k\|^2 = \sum_{i\in\mathcal{V}} \|\nabla E_i(\boldsymbol{w}_{i,k+1}) - \nabla E_i(\boldsymbol{w}_{i,k})\|^2$ which, based on Assumption 1, satisfies

$$\|\mathbf{g}_{k+1} - \mathbf{g}_k\|^2 \leq \sum_{i\in\mathcal{V}} \Big[L_i^2\|\boldsymbol{w}_{i,k+1} - \boldsymbol{w}_{i,k}\|^2\Big] \leq L^2\|\mathbf{w}_{k+1} - \mathbf{w}_k\|^2, \tag{28}$$

where we have used the definition $\max_{i\in\mathcal{V}}\{L_i^2\} = L^2$. From (7) we obtain

$$\mathbf{w}_{k+1} - \mathbf{w}_k = \beta(\mathcal{L} \otimes I_{d_w})\mathbf{w}_k - \alpha n\mathbf{d}_k + \sqrt{2\alpha n}\mathbf{v}_k$$

$$= \beta(\mathcal{L} \otimes I_{d_w})(\mathbf{w}_k - \mathbf{1}_n \otimes \bar{\boldsymbol{w}}_k) - \alpha n\mathbf{d}_k + \sqrt{2\alpha n}\mathbf{v}_k \tag{29}$$

$$= \beta(\mathcal{L} \otimes I_{d_w})\tilde{\mathbf{w}}_k - \alpha n\mathbf{d}_k + \sqrt{2\alpha n}\mathbf{v}_k. \tag{30}$$

In (29) we use the fact that $(\mathcal{L} \otimes I_{d_w})(\mathbf{1}_n \otimes \bar{\boldsymbol{w}}_k) = \mathbf{0}_{nd_w}$. Taking the norm of (30) yields

$$\|\mathbf{w}_{k+1} - \mathbf{w}_k\| \leq \beta\lambda_n\|\tilde{\mathbf{w}}_k\| + n\|\alpha\mathbf{d}_k\| + \sqrt{2\alpha n}\|\mathbf{v}_k\| \leq n\|\tilde{\mathbf{w}}_k\| + n\|\alpha\mathbf{d}_k\| + \sqrt{2\alpha n}\|\mathbf{v}_k\|, \tag{31}$$

where we use first inequality in Condition 1. Squaring (31) and noting that $\|\Psi_k\|^2 = \|\tilde{\mathbf{w}}_k\|^2 + \|\alpha\mathbf{d}_k\|^2$ gives

$$\|\mathbf{w}_{k+1} - \mathbf{w}_k\|^2 \leq 4n^2\|\Psi_k\|^2 + 4\alpha n\|\mathbf{v}_k\|^2. \tag{32}$$

Thereby, substituting (32) in (28) results in

$$\|\mathbf{g}_{k+1} - \mathbf{g}_k\|^2 \leq 4n^2L^2\|\Psi_k\|^2 + 4\alpha nL^2\|\mathbf{v}_k\|^2. \tag{33}$$

Substituting (33) in (27) gives

$$\|\Psi_{k+1}\|^2 \leq \left(1 - \beta\lambda_2 + \frac{4\alpha^2n^2L^2}{\beta\lambda_2}\right)\|\Psi_k\|^2 + \frac{2\alpha n(1+2\alpha^2L^2)}{\beta\lambda_2}\|\mathbf{v}_k\|^2. \tag{34}$$

Finally, taking the total expectation of (34) yields

$$\mathbb{E}[\|\Psi_{k+1}\|^2] \leq \left(1 - \beta\lambda_2 + \frac{4\alpha^2n^2L^2}{\beta\lambda_2}\right)\mathbb{E}[\|\Psi_k\|^2] + \frac{2\alpha n^2d_w(1+2\alpha^2L^2)}{\beta\lambda_2}, \tag{35}$$

where $\mathbb{E}[\|\mathbf{v}_k\|^2] \leq nd_w$. Note, that $\left(1 - \beta\lambda_2 + \frac{4\alpha^2n^2L^2}{\beta\lambda_2}\right) \in (0,1)$ from condition 1, which assures the convergence of $\mathbb{E}[\|\Psi_{k+1}\|^2]$. By iteratively using (35) we establish the rate of consensus for the proposed GT-DULA which is presented in Theorem 1.

Proof of Theorem 2

We start with the average dynamics generated by the GT-DULA. From (5) and noting that $\boldsymbol{d}_{i,0} = \nabla E_i(\boldsymbol{w}_{i,0})$, we have $\sum_{i\in\mathcal{V}}\boldsymbol{d}_{i,k} = \sum_{i\in\mathcal{V}}\nabla E_i(\boldsymbol{w}_{i,0})$. From (4), we establish the following average dynamics:

$$\bar{\boldsymbol{w}}_{k+1} = \bar{\boldsymbol{w}}_k - \alpha\boldsymbol{G}_k + \sqrt{2\alpha}\bar{\boldsymbol{v}}_k, \tag{36}$$

where $\boldsymbol{G}_k \triangleq \sum_{i\in\mathcal{V}}\nabla E_i(\boldsymbol{w}_{i,0})$ and $\bar{\boldsymbol{v}} \sim \mathcal{N}(\mathbf{0}_{d_w}, I_{d_w})$. Next, we split the gradient term $\boldsymbol{G}_k$ as

$$\boldsymbol{G}_k = \overline{\nabla E}_k + \xi_k, \tag{37}$$

where $\overline{\nabla E}_k \triangleq \sum_{i\in\mathcal{V}}\nabla E_i(\bar{\boldsymbol{w}}_k)$ and $\xi_k \triangleq \sum_{i\in\mathcal{V}}\Big(\nabla E_i(\boldsymbol{w}_{i,k}) - \nabla E_i(\bar{\boldsymbol{w}}_k)\Big)$. In essence, $\overline{\nabla E}_k$ represents the gradient computed at the average sample and ξ_k encompasses the deviation of due to local gradients and is a consequence of the distributed learning setup.

With (37) in mind, (36) can be written as a stochastic differential equation in continuous-time as below.

$$d\bar{\boldsymbol{w}}(t) = -\boldsymbol{G}_k dt + \sqrt{2}d\boldsymbol{B}(t) = -\Big(\overline{\nabla E}_k + \xi_k\Big)dt + \sqrt{2}d\boldsymbol{B}(t), \tag{38}$$

where $t \in [t_k, t_{k+1})$ such that continuous time $t_k = \alpha k$ corresponds to discrete-time instant k for any $k \geq 0$ and $\boldsymbol{B}(t)$ is a d_w-dimensional Brownian motion.

Next, defining $\boldsymbol{y}_{1,k} \triangleq \bar{\boldsymbol{w}}_k$, $\boldsymbol{y}_{2,k} \triangleq \tilde{\mathbf{w}}_k$ and $\boldsymbol{y}_k \triangleq [\boldsymbol{y}_{1,k}^\top, \boldsymbol{y}_{2,k}^\top]^\top$ and following a similar approach as in (33) of Bhar et al. (2022) we obtain the Fokker-Planck (FP) equation for (38) which gives the continuous-time evolution of the distribution of $\bar{\boldsymbol{w}}(t)$ as

$$\frac{\partial p(\bar{\boldsymbol{w}}(t))}{\partial t} = -\nabla\cdot\left[\int p(\bar{\boldsymbol{w}}(t)|\boldsymbol{y}_k)\Big(-\overline{\nabla E}_k - \xi_k\Big)p(\boldsymbol{y}_k)d\boldsymbol{y}_k\right] + \nabla^2 p(\bar{\boldsymbol{w}}(t)). \tag{39}$$

Proceeding with (39) in the same way as in (S101)-(S125) from Parayil et al. (2020) yields

$$\dot{F}\Big(p(\bar{\boldsymbol{w}}(t))\Big) = -\frac{1}{2}\mathbb{E}\left\|\nabla \log\left(\frac{p(\bar{\boldsymbol{w}}(t))}{p^*(\bar{\boldsymbol{w}}(t))}\right)\right\|^2 + \iint \left\|\overline{\nabla E}_t - \overline{\nabla E}_k\right\|^2 p(\bar{\boldsymbol{w}}(t))d\boldsymbol{y}_k + \iint \|\xi_k\|^2 p(\bar{\boldsymbol{w}}(t))d\boldsymbol{y}_k, \tag{40}$$

where $\overline{\nabla E}_t \triangleq \sum_{i\in\mathcal{V}} \nabla E_i(\bar{\boldsymbol{w}}(t))$.

We now derive the bounds for $\mathbb{E}\|\overline{\nabla E}_t - \overline{\nabla E}_k\|^2$, $\mathbb{E}\|\xi_k\|^2$ and $\mathbb{E}\|\zeta_k\|^2$ individually. First, from Assumption 1, we have

$$\|\xi_k\|^2 = \left\|\sum_{i\in\mathcal{V}}\Big(\nabla E_i(\boldsymbol{w}_{i,k}) - \nabla E_i(\bar{\boldsymbol{w}}_k)\Big)\right\|^2 \leq n\sum_{i\in\mathcal{V}}\Big[L_i^2\|\boldsymbol{w}_{i,k} - \bar{\boldsymbol{w}}_k\|^2\Big] \leq nL^2\|\tilde{\mathbf{w}}_k\|^2, \tag{41}$$

which after marginalizing with respect to (w.r.t.) $\boldsymbol{y}_k$ yields

$$\mathbb{E}[\|\xi_k\|^2] \leq nL^2\mathbb{E}[\|\tilde{\mathbf{w}}_k\|^2]. \tag{42}$$

Next, we note that $\left\|\overline{\nabla E}_t - \overline{\nabla E}_k\right\|^2$ can be bounded as

$$\left\|\overline{\nabla E}_t - \overline{\nabla E}_k\right\|^2 \leq n\bar{L}^2\|\bar{\boldsymbol{w}}(t) - \bar{\boldsymbol{w}}(t_k)\|^2, \tag{43}$$

where $\bar{L}^2 \triangleq \sum_{i\in\mathcal{V}}\left(\sum_{j=1}^{M_i} L_{ij} + \frac{L_i'}{n}\right)^2$. Integrating (38) from t_k to $t \in [t_k, t_{k+1})$ gives

$$\begin{aligned}
\|\bar{\boldsymbol{w}}(t) - \bar{\boldsymbol{w}}(t_k)\|^2 &\leq \Big\| - \boldsymbol{G}_k(t-t_k) + \sqrt{2}\Big(\boldsymbol{B}(t) - \boldsymbol{B}(t_k)\Big)\Big\|^2 \\
&\leq 2\|\boldsymbol{B}(t) - \boldsymbol{B}(t_k)\|^2 + \|\boldsymbol{G}_k(t-t_k)\|^2 - 2\sqrt{2}\boldsymbol{S}_k \qquad (44)\\
&\leq 2\|\boldsymbol{B}(t) - \boldsymbol{B}(t_k)\|^2 + \alpha^2\|\boldsymbol{G}_k\|^2 - 2\sqrt{2}\boldsymbol{S}_k \qquad (45)
\end{aligned}$$

where $\boldsymbol{S}_k \triangleq \Big(\boldsymbol{B}(t) - \boldsymbol{B}(t_k)\Big)^\top\Big(\boldsymbol{G}_k(t-t_k)\Big)$. In (45) we have used $t - t_k < t_{k+1} - t_k = \alpha$ for any $t \in [t_k, t_{k+1})$. Substituting (45) in (43) results in

$$\left\|\overline{\nabla E}_t - \overline{\nabla E}_k\right\|^2 \leq n\bar{L}^2\Big[2\|\boldsymbol{B}(t) - \boldsymbol{B}(t_k)\|^2 + \alpha^2\|\boldsymbol{G}_k\|^2 - 2\sqrt{2}\boldsymbol{S}_k\Big]. \tag{46}$$

We bound $\|\boldsymbol{G}_k\|^2$ as

$$\|\boldsymbol{G}_k\|^2 = \left\|\sum_{i\in\mathcal{V}}\nabla E_i(\boldsymbol{w}_{i,k})\right\|^2 = \left\|\sum_{i\in\mathcal{V}}\Big(\nabla E_i(\boldsymbol{w}_{i,k}) - \nabla E_i(\bar{\boldsymbol{w}}_k) + \nabla E_i(\bar{\boldsymbol{w}}_k) - \nabla E_i(\boldsymbol{w}^*)\Big)\right\|^2 \tag{47}$$

$$\begin{aligned}
&\leq 2\|\xi_k\|^2 + 2\left\|\sum_{i\in\mathcal{V}}\Big(\nabla E_i(\bar{\boldsymbol{w}}_k) - \nabla E_i(\boldsymbol{w}^*)\Big)\right\|^2 \leq 2\|\xi_k\|^2 + 2n\sum_{i\in\mathcal{V}}L_i^2\|\bar{\boldsymbol{w}}_k - \boldsymbol{w}^*\|^2 \\
&\leq 2\|\xi_k\|^2 + 4n\left(\sum_{i\in\mathcal{V}}L_i^2\right)\Big(\|\bar{\boldsymbol{w}}_k\|^2 + \|\boldsymbol{w}^*\|^2\Big) \leq 2nL^2\|\tilde{\mathbf{w}}_k\|^2 + 4n\bar{L}^2\big(\|\bar{\boldsymbol{w}}_k\|^2 + \|\boldsymbol{w}^*\|^2\big).
\end{aligned} \tag{48}$$

In (47) we have used the fact that $\boldsymbol{w}^*$ is some local extremum of $\sum_{i\in\mathcal{V}}\nabla E_i(\cdot)$, i.e., $\sum_{i\in\mathcal{V}}\nabla E_i(\boldsymbol{w}^*)=\mathbf{0}_{d_w}$ while in (48) we have used $\sum_{i\in\mathcal{V}}L_i^2=\bar{L}^2$ and the bound from (41). Substituting (48) in (46) yields

$$\begin{aligned}\left\|\overline{\nabla E}_t-\overline{\nabla E}_k\right\|^2 \leq 2n\bar{L}^2\|\boldsymbol{B}(t)-\boldsymbol{B}(t_k)\|^2+2n^2\alpha^2L^2\bar{L}^2\|\tilde{\mathbf{w}}_k\|^2+4n^2\alpha^2\bar{L}^4\|\bar{\boldsymbol{w}}_k\|^2\\ +4n^2\alpha^2\bar{L}^4\|\boldsymbol{w}^*\|^2-2\sqrt{2}n\bar{L}^2\boldsymbol{S}_k.\end{aligned} \tag{49}$$

Marginalizing (49) w.r.t. $\boldsymbol{y}_k$ gives

$$\mathbb{E}\left\|\overline{\nabla E}_t-\overline{\nabla E}_k\right\|^2 \leq 2n\alpha\bar{L}^2d_w+2n^2\alpha^2L^2\bar{L}^2\mathbb{E}[\|\tilde{\mathbf{w}}_k\|^2]+4n^2\alpha^2\bar{L}^4C_{\bar{\boldsymbol{w}}}+4n^2\alpha^2\bar{L}^4C_{\boldsymbol{w}^*}, \tag{50}$$

where $\mathbb{E}[\|\boldsymbol{w}^*\|^2]\leq C_{\boldsymbol{w}^*}$. For details on the derivation of (50), refer to (S135)-(S141) in Parayil et al. (2020). Finally, incorporating (42) and (50) in (40) yields

$$\begin{aligned}\dot{F}\Big(p(\bar{\boldsymbol{w}}(t))\Big) &\leq -\frac{1}{2}\mathbb{E}\left\|\nabla\log\left(\frac{p(\bar{\boldsymbol{w}}(t))}{p^*(\bar{\boldsymbol{w}}(t))}\right)\right\|^2+2n\alpha\bar{L}^2d_w+(2n^2\alpha^2L^2\bar{L}^2+nL^2)\mathbb{E}[\|\tilde{\mathbf{w}}_k\|^2]\\ &\quad+4n^2\alpha^2\bar{L}^4C_{\bar{\boldsymbol{w}}}+4n^2\alpha^2\bar{L}^4C_{\boldsymbol{w}^*}\\ &\leq -\frac{1}{2}\mathbb{E}\left\|\nabla\log\left(\frac{p(\bar{\boldsymbol{w}}(t))}{p^*(\bar{\boldsymbol{w}}(t))}\right)\right\|^2+(2n^2\alpha^2L^2\bar{L}^2+nL^2)\mathbb{E}[\|\tilde{\mathbf{w}}_k\|^2]+f,\end{aligned} \tag{51}$$

where $f\triangleq 2n\alpha\bar{L}^2d_w+4n^2\alpha^2\bar{L}^4C_{\bar{\boldsymbol{w}}}+4n^2\alpha^2\bar{L}^4C_{\boldsymbol{w}^*}$. Here we have utilized the LSI assumption in Assumption 3 with $g(\bar{\boldsymbol{w}})=\frac{p(\bar{\boldsymbol{w}}(t))}{p^*(\bar{\boldsymbol{w}}(t))}$, which results in

$$F\Big(p(\bar{\boldsymbol{w}}(t))\Big)\leq\frac{1}{2\rho_U}\mathbb{E}\left\|\nabla\log\left(\frac{p(\bar{\boldsymbol{w}}(t))}{p^*(\bar{\boldsymbol{w}}(t))}\right)\right\|^2. \tag{52}$$

Using (52) in (51) yields

$$\dot{F}\Big(p(\bar{\boldsymbol{w}}(t))\Big)\leq-\rho_UF\Big(p(\bar{\boldsymbol{w}}(t))\Big)+f+(2n^2\alpha^2L^2\bar{L}^2+nL^2)\mathbb{E}[\|\tilde{\mathbf{w}}_k\|^2]. \tag{53}$$

Integrating (53) w.r.t. t within $t\in[t_k,t_{k+1}]$ and utilizing $t_{k+1}-t_k<\alpha$ give the evolution of the KL divergence of the posterior generated by the average samples

$$F\Big(p(\bar{\boldsymbol{w}}_{k+1})\Big)\leq\exp(-\alpha\rho_U)F\Big(p(\bar{\boldsymbol{w}}_k)\Big)+\frac{1-\exp(-\alpha\rho_U)}{\rho_U}\Big[f+(2n^2\alpha^2L^2\bar{L}^2+nL^2)\mathbb{E}[\|\tilde{\mathbf{w}}_k\|^2]\Big]. \tag{54}$$

Using (54) iteratively yields

$$\begin{aligned}F\Big(p(\bar{\boldsymbol{w}}_{k+1})\Big) &\leq \exp(-\alpha\rho_U(k+1))F\Big(p(\bar{\boldsymbol{w}}_0)\Big)+\frac{1-\exp(-\alpha\rho_U)}{\rho_U}\sum_{\ell=0}^{k}\Big[\exp(-\alpha\rho_U(k-\ell))\times\\ &\qquad\Big(f+(2n^2\alpha^2L^2\bar{L}^2+nL^2)\mathbb{E}[\|\tilde{\mathbf{w}}_\ell\|^2]\Big)\Big]\\ &\leq \exp(-\alpha\rho_U(k+1))F\Big(p(\bar{\boldsymbol{w}}_0)\Big)+\frac{f+(2n^2\alpha^2L^2\bar{L}^2+nL^2)B_{GT}}{\rho_U}\\ &\quad+(2n^2\alpha^2L^2\bar{L}^2+nL^2)\alpha\mathbb{E}\|\Psi_0\|^2\sum_{\ell=0}^{k}\exp(-\alpha\rho_U(k-\ell))\sigma^\ell,\end{aligned} \tag{55}$$

where we have substituted (12) and used the relations $\sum_{\ell=0}^{k}\exp(-\alpha\rho_U(k-\ell)) \leq \sum_{\ell=0}^{\infty}\exp(-\alpha\rho_U\ell) = \frac{1}{1-\exp(-\alpha\rho_U)}$ and $1-\exp(-\alpha\rho_U) \leq \alpha\rho_U$. Note that

$$\begin{aligned}
&\sum_{\ell=0}^{k}\exp(-\alpha\rho_U(k-\ell))\sigma^{\ell} = \exp(-\alpha\rho_U k)\sum_{\ell=0}^{k}\Big(\exp(\alpha\rho_U))\sigma\Big)^{\ell} \\
&= \exp(-\alpha\rho_U k)\frac{|\sigma^{k+1}\exp(\alpha\rho_U(k+1))-1|}{|\sigma\exp(\alpha\rho_U)-1|} = \frac{|\sigma^{k+1}-\exp(-\alpha\rho_U(k+1))|}{|\sigma-\exp(-\alpha\rho_U)|} \\
&\leq \frac{\sigma^{k+1}+\exp(-\alpha\rho_U(k+1))}{|\sigma-\exp(-\alpha\rho_U)|}. \qquad (56)
\end{aligned}$$

Substituting (56) in (55), we get the rate of convergence of the KL divergence of the posterior of the average sample, which is presented in Theorem 2.

Proof of Corollary 3

From (16), $F\Big(p(\bar{\boldsymbol{w}}_{k+1})\Big) \leq \epsilon$ is satisfied if (i) $\Big(F\Big(p(\bar{\boldsymbol{w}}_0)\Big) + C_{_F}\Big)\exp(-\alpha\rho_U(k+1)) \leq \frac{\epsilon}{5}$, (ii) $C_{_F}\sigma^{k+1} \leq \frac{\epsilon}{5}$ and (iii) $O_{_{GT}} \leq \frac{3\epsilon}{5}$. (i) and (ii) respectively give the minimum k values in (20). Next, we have $O_{_{GT}} < \frac{\bar{O}_{GT}}{\rho_U(\beta^2\lambda_2^2-4\alpha^2n^2L^2)}$ where

$$\bar{O}_{_{GT}} = 2n\alpha\bar{L}^2d_w + 4n^2\alpha^2\bar{L}^4(C_{\bar{\boldsymbol{w}}} + C_{\boldsymbol{w}^*}) + 2\alpha n^3d_wL^2(2n\alpha^2\bar{L}^2+1)(1+2\alpha^2L^2). \qquad (57)$$

The condition in (iii) is satisfied if $\bar{O}_{_{GT}} \leq \frac{3\epsilon\rho_U(\beta^2\lambda_2^2-4\alpha^2n^2L^2)}{5}$ which can be further satisfied by $\bar{O}_{_{GT}} + \frac{12\rho_UL^2\alpha^2n^2}{5} \leq \frac{3\epsilon\rho_U\beta^2\lambda_2^2}{5}$ since $\epsilon < 1$. This final constraint is achieved by:

$$(8L^4\bar{L}^2d_w + 8L^4d_w)n^4\alpha^3 \leq \frac{\epsilon\rho_U\beta^2\lambda_2^2}{5}, \qquad (58)$$

$$\Big[4\bar{L}^4(C_{\bar{\boldsymbol{w}}} + C_{\boldsymbol{w}^*}) + \frac{12\rho_UL^2}{5} + 4L^2\bar{L}^2d_w\Big]n^4\alpha^2 \leq \frac{\epsilon\rho_U\beta^2\lambda_2^2}{5}, \qquad (59)$$

$$(2\bar{L}^2d_w + 2L^2d_w)n^3\alpha \leq \frac{\epsilon\rho_U\beta^2\lambda_2^2}{5}, \qquad (60)$$

each of which results in the bounds in (19), respectively.

Proof of Theorem 4

From (8) we have

$$\tilde{\mathbf{d}}_{k+1} = (\mathcal{W}_\gamma \otimes I_{d_w})\tilde{\mathbf{d}}_k + (\mathbf{g}_{k+1} - \mathbf{g}_k) - \mathbf{1}_n \otimes (\bar{\boldsymbol{u}}_{k+1} - \bar{\boldsymbol{u}}_k). \qquad (61)$$

Taking the square of the norm of (61) yields

$$\|\tilde{\mathbf{d}}_{k+1}\|^2 \leq (1-\gamma\lambda_2)\|\tilde{\mathbf{d}}_k\|^2 + \frac{2}{\gamma\lambda_2}\|\mathbf{g}_{k+1} - \mathbf{g}_k\|^2 + \frac{2n}{\gamma\lambda_2}\|\bar{\boldsymbol{u}}_{k+1} - \bar{\boldsymbol{u}}_k\|, \qquad (62)$$

whose total expectation is given by

$$\mathbb{E}[\|\tilde{\mathbf{d}}_{k+1}\|^2] \leq (1-\gamma\lambda_2)\mathbb{E}[\|\tilde{\mathbf{d}}_k\|^2] + \frac{2}{\gamma\lambda_2}\mathbb{E}[\|\mathbf{g}_{k+1} - \mathbf{g}_k\|^2] + \frac{2n}{\gamma\lambda_2}\mathbb{E}[\|\bar{\boldsymbol{u}}_{k+1} - \bar{\boldsymbol{u}}_k\|]. \qquad (63)$$

We derive bounds for the individual terms on the right hand side of (63). First, from (33) we have

$$\mathbb{E}[\|\mathbf{g}_{k+1}-\mathbf{g}_k\|^2] \leq 4n^2L^2\mathbb{E}[\|\Psi_k\|^2] + 4\alpha n^2L^2 d_w. \tag{64}$$

Then,

$$\mathbb{E}[\|\bar{\boldsymbol{u}}_{k+1}-\bar{\boldsymbol{u}}_k\|^2] = \frac{\bar{L}^2}{n}\mathbb{E}[\|\bar{\boldsymbol{w}}_{k+1}-\bar{\boldsymbol{w}}_k\|^2], \tag{65}$$

and from (36) we can write

$$\bar{\boldsymbol{w}}_{k+1}-\bar{\boldsymbol{w}}_k = -\alpha\boldsymbol{G}_k + \sqrt{2\alpha}\bar{\boldsymbol{v}}_k. \tag{66}$$

Thus, from (66)

$$\|\bar{\boldsymbol{w}}_{k+1}-\bar{\boldsymbol{w}}_k\|^2 \leq 2\alpha^2\|\boldsymbol{G}_k\|^2 + 4\alpha\|\bar{\boldsymbol{v}}_k\|^2. \tag{67}$$

Substituting (48) in (67) and taking the total expectation yield

$$\mathbb{E}[\|\bar{\boldsymbol{w}}_{k+1}-\bar{\boldsymbol{w}}_k\|^2] \leq 4n\alpha^2L^2\mathbb{E}[\|\tilde{\mathbf{w}}_k\|^2] + 8n\alpha^2\bar{L}^2\big(C_{\bar{\boldsymbol{w}}}+C_{\boldsymbol{w}^*}\big) + 4n\alpha d_w, \tag{68}$$

where we also use $\mathbb{E}[\|\bar{\boldsymbol{v}}_k\|^2] \leq nd_w$. Substituting (68) in (65) gives

$$\mathbb{E}[\|\bar{\boldsymbol{u}}_{k+1}-\bar{\boldsymbol{u}}_k\|^2] \leq 4\alpha^2L^2\bar{L}^2\mathbb{E}[\|\tilde{\mathbf{w}}_k\|^2] + 8\alpha^2\bar{L}^4\big(C_{\bar{\boldsymbol{w}}}+C_{\boldsymbol{w}^*}\big) + 4\alpha d_w\bar{L}^2. \tag{69}$$

Finally, substituting (64) and (69) in (63) leads to

$$\begin{aligned}\mathbb{E}[\|\tilde{\mathbf{d}}_{k+1}\|^2] \leq (1-\gamma\lambda_2)\mathbb{E}[\|\tilde{\mathbf{d}}_k\|^2] + \frac{8nL^2(n+\alpha^2\bar{L}^2)}{\gamma\lambda_2}\mathbb{E}[\|\Psi_k\|^2] + \frac{16n\alpha^2\bar{L}^4(C_{\bar{\boldsymbol{w}}}+C_{\boldsymbol{w}^*})}{\gamma\lambda_2} \\ + \frac{8n\alpha d_w(nL^2+\bar{L}^2)}{\gamma\lambda_2},\end{aligned} \tag{70}$$

where we used $\mathbb{E}[\|\tilde{\mathbf{w}}_k\|^2] \leq \mathbb{E}[\|\Psi_k\|^2]$. Iteratively using (70) yields

$$\begin{aligned}\mathbb{E}[\|\tilde{\mathbf{d}}_{k+1}\|^2] \leq (1-\gamma\lambda_2)^{k+1}\mathbb{E}[\|\tilde{\mathbf{d}}_0\|^2] + \frac{16n\alpha^2\bar{L}^4(C_{\bar{\boldsymbol{w}}}+C_{\boldsymbol{w}^*})}{\gamma^2\lambda_2^2} + \frac{8n\alpha d_w(nL^2+\bar{L}^2)}{\gamma^2\lambda_2^2} \\ \frac{8nL^2(n+\alpha^2\bar{L}^2)B_{GT}}{\gamma^2\lambda_2^2} + \frac{8nL^2(n+\alpha^2\bar{L}^2)\mathbb{E}\|\Psi_0\|^2}{\gamma\lambda_2}\sum_{\ell=0}^{k}(1-\gamma\lambda_2)^{k-\ell}\sigma^\ell,\end{aligned} \tag{71}$$

where we have substituted (12) and used $\sum_{\ell=0}^{k}(1-\gamma\lambda_2)^\ell < \sum_{\ell=0}^{\infty}(1-\gamma\lambda_2)^\ell = \frac{1}{\gamma\lambda_2}$. Note that

$$\begin{aligned}\sum_{\ell=0}^{k}(1-\gamma\lambda_2)^{k-\ell}\sigma^\ell &= (1-\gamma\lambda_2)^k\sum_{\ell=0}^{k}\left(\frac{\sigma}{1-\gamma\lambda_2}\right)^\ell = (1-\gamma\lambda_2)^k\frac{\left|\left(\frac{\sigma}{1-\gamma\lambda_2}\right)^{k+1}-1\right|}{\left|\frac{\sigma}{1-\gamma\lambda_2}-1\right|} \\ &= \frac{|\sigma^{k+1}-(1-\gamma\lambda_2)^{k+1}|}{|\sigma+\gamma\lambda_2-1|} \leq \frac{\sigma^{k+1}+(1-\gamma\lambda_2)^{k+1}}{|\sigma+\gamma\lambda_2-1|}.\end{aligned} \tag{72}$$

Substituting (72) in (71) yields our final result for the gradient error presented in Theorem 4.

Useful Lemmas

Lemma 5 *For any $k \geq 0$ the following bound holds.*

$$\mathbb{E}[\|\bar{\boldsymbol{w}}(t_k)\|^2] \leq C_{\bar{\boldsymbol{w}}}, \tag{73}$$

where

$$C_{\bar{\boldsymbol{w}}} = \max\Bigg\{\mathbb{E}[\|\bar{\boldsymbol{w}}_0\|^2], \frac{1}{\rho_U^2 - 16n^2\alpha^2\bar{L}_2^4}\Bigg(2\rho_U^2 c_1 + 4\rho_U F\Big(p(\bar{\boldsymbol{w}}_0)\Big) + 8\big(n\alpha\bar{L}^2 d_w \\ + 2n^2\alpha^2\bar{L}_2^4 C_{\boldsymbol{w}^*}\big) + 4(2n^2\alpha^2L^2\bar{L}^2 + nL^2)\big(\mathbb{E}[\|\Psi_0\|^2] + B_{GT}\big)\Bigg)\Bigg\}. \tag{74}$$

Proof: We follow a similar approach as used in Lemma S6 of Parayil et al. (2020) which uses induction to derive this bound. Assuming that $\mathbb{E}[\|\bar{\boldsymbol{w}}_\ell\|^2] \leq C_{\bar{\boldsymbol{w}}}$ for all $\ell \leq k$, we need to prove $\mathbb{E}[\|\bar{\boldsymbol{w}}_{k+1}\|^2] \leq C_{\bar{\boldsymbol{w}}}$. From (S252) in Parayil et al. (2020), we have

$$\mathbb{E}[\|\bar{\boldsymbol{w}}_{k+1}\|^2] \leq 2c_1 + \frac{4}{\rho_U}F\Big(p(\bar{\boldsymbol{w}}_{k+1})\Big), \tag{75}$$

where $\mathbb{E}_{p^*}[\|\bar{\boldsymbol{w}}^*\|^2] \leq c_1$. From (55), we can write

$$F\Big(p(\bar{\boldsymbol{w}}_{k+1})\Big) \leq F\Big(p(\bar{\boldsymbol{w}}_0)\Big) + \frac{1}{\rho_U}\Big(2n\alpha\bar{L}^2 d_w + 4n^2\alpha^2\bar{L}_2^4 C_{\bar{\boldsymbol{w}}} + 4n^2\alpha^2\bar{L}_2^4 C_{\boldsymbol{w}^*} \\ + (2n^2\alpha^2L^2\bar{L}^2 + nL^2)\mathbb{E}[\|\tilde{\mathbf{w}}_k\|^2]\Big). \tag{76}$$

Note that in the right hand side of (76), we have used the bound $C_{\bar{\boldsymbol{w}}}$ since it is assumed to hold up to $\ell \leq k$. From Theorem 1, we have $\mathbb{E}[\|\tilde{\mathbf{w}}_k\|^2] \leq \mathbb{E}[\|\Psi_0\|^2] + B_{GT}$ which when substituted in (76) yields

$$F\Big(p(\bar{\boldsymbol{w}}_{k+1})\Big) \leq F\Big(p(\bar{\boldsymbol{w}}_0)\Big) + \frac{1}{\rho_U}\Big(2n\alpha\bar{L}^2 d_w + 4n^2\alpha^2\bar{L}_2^4 C_{\bar{\boldsymbol{w}}} + 4n^2\alpha^2\bar{L}_2^4 C_{\boldsymbol{w}^*} \\ + (2n^2\alpha^2L^2\bar{L}^2 + nL^2)\big(\mathbb{E}[\|\Psi_0\|^2] + B_{GT}\big)\Big). \tag{77}$$

Substituting (77) in (75) results in

$$\mathbb{E}[\|\bar{\boldsymbol{w}}_{k+1}\|^2] \leq 2c_1 + \frac{4}{\rho_U}F\Big(p(\bar{\boldsymbol{w}}_0)\Big) + \frac{4}{\rho_U^2}\Big(2n\alpha\bar{L}^2 d_w + 4n^2\alpha^2\bar{L}_2^4 C_{\bar{\boldsymbol{w}}} + 4n^2\alpha^2\bar{L}_2^4 C_{\boldsymbol{w}^*} \\ + (2n^2\alpha^2L^2\bar{L}^2 + nL^2)\big(\mathbb{E}[\|\Psi_0\|^2] + B_{GT}\big)\Big). \tag{78}$$

For induction, we need to enforce $\mathbb{E}[\|\bar{\boldsymbol{w}}_{k+1}\|^2] \leq C_{\bar{\boldsymbol{w}}}$. Thus, from (78) we require

$$2c_1 + \frac{4}{\rho_U}F\Big(p(\bar{\boldsymbol{w}}_0)\Big) + \frac{4}{\rho_U^2}\Big(2n\alpha\bar{L}^2 d_w + 4n^2\alpha^2\bar{L}_2^4 C_{\bar{\boldsymbol{w}}} + 4n^2\alpha^2\bar{L}_2^4 C_{\boldsymbol{w}^*} \\ + (2n^2\alpha^2L^2\bar{L}^2 + nL^2)\big(\mathbb{E}[\|\Psi_0\|^2] + B_{GT}\big)\Big) \leq C_{\bar{\boldsymbol{w}}}. \tag{79}$$

i.e.,

$$\left(1-\frac{16n^2\alpha^2\bar{L}_2^4}{\rho_U^2}\right)C_{\bar{\boldsymbol{w}}} \geq 2c_1+\frac{4}{\rho_U}F\Big(p(\bar{\boldsymbol{w}}_0)\Big)+\frac{4}{\rho_U^2}\Big(2n\alpha\bar{L}^2d_w+4n^2\alpha^2\bar{L}_2^4C_{\boldsymbol{w}^*} + (2n^2\alpha^2L^2\bar{L}^2+nL^2)\big(\mathbb{E}[\|\Psi_0\|^2]+B_{GT}\big)\Big). \tag{80}$$

For $C_{\bar{\boldsymbol{w}}}$ to exist, we need $\alpha < \frac{\rho_U}{4n\bar{L}^2}$, and the value in (74) follows from (80). ■