1–24, 2025

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 JEMIN.GEORGE.CIV@MAIL.MIL
DEVCOM Army Research Laboratory, Adelphi, MD 20783

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

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 we use in this paper is the 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. 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 decentralized 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 gradients across agents.

1.2. Contribution

The major contributions of this paper are as follows. This is the first study on the effect of gradient tracking on decentralized ULA to the best of our knowledge. We do not make any convexity assumption 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.

Stochastic gradients are known to be unstable in some cases without bounded gradient assumptions Toulis et al. (2016); Mai and Johansson (2021); Asi and Duchi (2019a,b). We derive such bounds from Lipschitz continuity without assuming bounded gradients. Hence, stochastic gradients can be used for GT-DULA.

Furthermore, unlike most literature with stochastic gradients Li and Orabona (2019); Pu and Nedić (2021); Xin et al. (2019b); Koloskova et al. (2021); Zhang and You (2019); Liu et al. (2024); Lu and De Sa (2021); Dandi et al. (2022); Sun et al. (2023); Lin et al. (2021); Le Bars et al. (2023), we do not explicitly assume bounded variance for the stochastic gradient noise either, but rather derive this bound as part of our analysis. We address the issue of data heterogeneity without assuming any bounds on the variance of the gradients across agents beforehand. Typically, a bound on the variance is assumed and the dependence 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 $\mathcal{O}(\alpha)$. Thus, we derive appropriate bounds for the gradients along the topology of the samples generated, and show that α is a controlling factor to handle any degree of heterogeneity. 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 $\boldsymbol{w}$ 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. Let $p^* = p(\boldsymbol{w}|\boldsymbol{X})$.

The DULA presented in Parayil et al. (2020) provides an efficient framework for this problem by exploiting the Langevin dynamics. The posterior posterior distribution is rewritten in terms of a 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 $\boldsymbol{v}_{i,k} \sim \mathcal{N}(\mathbf{0}_{d_w}, I_{d_w})$ is the injected standard Gaussian noise in the system. 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 size) are presented in Section 4.4 to compare with the results for GT-DULA.

3. Methodology

In the traditional formulation of the DULA in (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. A common approach is to initialize $\boldsymbol{d}_{i,0} = \hat{\boldsymbol{g}}_{i,0}$ for any $i \in \mathcal{V}$ and use

$$\triangle \boldsymbol{d}_{i,k} \triangleq \hat{\boldsymbol{g}}_{i,k+1} - \hat{\boldsymbol{g}}_{i,k}, \tag{6}$$

where $\hat{\boldsymbol{g}}_{i,k} = -\frac{1}{f_i}\nabla \log p(\hat{\boldsymbol{X}}_i|\boldsymbol{w}_i) - \frac{1}{n}\nabla \log p(\boldsymbol{w}_i)$ is the unbiased estimator of the i-th agent's local gradient using the stochastic mini-batch $\hat{X}_i \subset \boldsymbol{X}_i$, wherein $|\hat{\boldsymbol{X}}_i| = m_i$ and $f_i = \frac{m_i}{M_i}$, and γ is a constant consensus step size corresponding to the mixing of the gradient estimates. Since values of $\boldsymbol{w}_{i,k}$ and $\boldsymbol{d}_{i,k}$ are needed for mixing within the graph, the information shared by each agent with its neighbors entails $[\boldsymbol{w}_{i,k}^\top, \boldsymbol{d}_{i,k}^\top]^\top$. The global gradient estimate at the beginning of the learning is simply each agent's local gradient. However, it is improved by (5) gradually. Intuitively, it is expected that $\{\boldsymbol{d}_{i,k}\}$ shall achieve consensus and converge towards the global gradient. Thus, with each step k, the estimate $\boldsymbol{d}_{i,k}$ is expected to give a better update with respect to the global data $\boldsymbol{X}$ than the local gradient $\nabla E_i(\boldsymbol{w}_{i,k})$.

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. Define $\mathbf{w}_k \triangleq [\boldsymbol{w}_{1,k}^\top, \ldots, \boldsymbol{w}_{n,k}^\top]^\top$, $\mathbf{d}_k \triangleq [\boldsymbol{d}_{1,k}^\top, \ldots, \boldsymbol{d}_{n,k}^\top]^\top$, $\hat{\mathbf{g}}_k \triangleq [\hat{\boldsymbol{g}}_{1,k}^\top, \ldots, \hat{\boldsymbol{g}}_{n,k}^\top]^\top$, $\mathbf{v}_k \triangleq [\boldsymbol{v}_{1,k}^\top, \ldots, \boldsymbol{v}_{n,k}^\top]^\top$, $\mathcal{W}_\beta \triangleq I_n - \beta\mathcal{L}$ and $\mathcal{W}_\gamma \triangleq I_n - \gamma\mathcal{L}$. The GT-DULA proposed in (4) and (5) combined with the update rule in (6) can be concisely written for all $i \in \mathcal{V}$ as

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

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

4. Theoretical Results

In this section, we lay out the relevant results for GT-DULA. Some notations used throughout the analysis are provided. (i) L_{ij}, L'_i: Lipschitz constant for the i-th agent corresponding to the likelihood for the j-th data point and the prior, respectively, (ii) λ_2, λ_n : second smallest and the largest eigenvalue of the Laplacian $\mathcal{L}$ of the network, respectively, (iii) $L_i = \sum_{i\in\mathcal{V}} L_{ij}$, (iv) $\sigma_{L,i}^2 = \mathbb{E}_{\mathbf{B}}\left[\left(\frac{M_i}{m_i}\sum_{j\in\hat{\boldsymbol{X}}_i} L_{ij} - \sum_{j\in\boldsymbol{X}_i} L_{ij}\right)^2\right]$, where $\mathbf{B}$ represents the set of all possible combinations of mini-batch, (v) $L^2 = \max_{i\in\mathcal{V}}\left\{\left(L_i + \frac{L'_i}{n}\right)^2\right\}$, (vi) $\sigma_L^2 = \max_{i\in\mathcal{V}}\{\sigma_{L,i}^2\}$, (vii) $\bar{L}^2 = \sum_{i\in\mathcal{V}}\left(L_i + \frac{L'_i}{n}\right)^2$, $\bar{\sigma}_L^2 = \sum_{i\in\mathcal{V}} \sigma_{L,i}^2$, (vii) $L_1^2 = L^2 + \sigma_L^2$, (viii) $L_2^2 = \bar{L}^2 + \bar{\sigma}_L^2$, (ix) $\mathbb{E}[\|\bar{\boldsymbol{w}}_k\|^2] \leq C_{\bar{\boldsymbol{w}}}$, $\forall k \geq 0$, (x) $\mathbb{E}[\|\hat{\boldsymbol{w}}^*\|^2] \leq C_{\hat{\boldsymbol{w}}^*}$, where $\hat{\boldsymbol{w}}^*$ is a local extremum of $\sum_{i\in\mathcal{V}} \widehat{\nabla E}_i(\cdot)$ and $\widehat{\nabla E}_i(\cdot)$ is the stochastic gradient of the i-th agent.

Next, we state the assumptions used to establish the results.

Assumption 1 *All gradients of likelihood and prior for all agents are* Lipschitz *continuous. For agent $i \in \mathcal{V}$ and $x_i^j \in \boldsymbol{X}_i$, let the gradient of the likelihood corresponding to x_i^j and of the prior be $\nabla \log p(x_i^j|\cdot)$ and $\nabla \log p(\cdot)$, respectively. Then, for any $\boldsymbol{w}_a, \boldsymbol{w}_b \in \mathbb{R}^{d_w}$ we have*

$$\|\nabla \log p(x_i^j|\boldsymbol{w}_a) - \nabla \log p(x_i^j|\boldsymbol{w}_b)\| \leq L_{ij}\|\boldsymbol{w}_a - \boldsymbol{w}_b\|, \tag{9}$$

$$\|\nabla \log p(\boldsymbol{w}_a) - \nabla \log p(\boldsymbol{w}_b)\| \leq L'_i\|\boldsymbol{w}_a - \boldsymbol{w}_b\|. \tag{10}$$

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.*

$$\mathbb{E}_{p^*}\left[g(\bar{\boldsymbol{w}})\log g(\bar{\boldsymbol{w}})\right] \leq \frac{1}{2\rho_U}\mathbb{E}_{p^*}\left[\frac{\|\nabla g(\bar{\boldsymbol{w}})\|^2}{g(\bar{\boldsymbol{w}})}\right]. \tag{11}$$

Assumption 1 implies smoothness of the log of the posterior function, Assumption 2 implies sufficient connectivity of the graph to ensure consensus while Assumption 3 is needed for the convergence of the distribution.

Below are the conditions that the hyperparameters need to satisfy in order for our results to hold.

Condition 1 $\beta \leq \min\left\{1, \frac{n}{\lambda_n}\right\}$, $\alpha \leq \min\left\{\frac{\beta\lambda_2}{2nL_1^2}, \frac{\rho_U}{4n\bar{L}L_2}\right\}$, $\beta\lambda_2 - \frac{4\alpha^2 n^2 L_1^2}{\beta\lambda_2} < 1$, $\gamma \leq \min\left\{\beta, \frac{1}{\lambda_2}\right\}$.

There always exist α, β such that the first three constraints in Condition 1 are met. Once the value of β is finalized, a feasible value of γ can be chosen from the last constraint of Condition 1.

We now introduce 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

$$\tilde{\mathbf{w}}_k \triangleq \mathbf{w}_k - (\mathbf{1}_n \otimes \bar{\boldsymbol{w}}_k) = \left(\left(I_n - \frac{1}{n}\mathbf{1}_n\mathbf{1}_n \right) \otimes I_{d_w} \right) \mathbf{w}_k, \tag{12}$$

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*

$$\mathbb{E}[\|\Psi_{k+1}\|^2] \leq \sigma^{k+1}\mathbb{E}[\|\Psi_0\|^2] + B_{GT}, \tag{13}$$

in which

$$\sigma \triangleq 1 - \beta\lambda_2 + \frac{4\alpha^2 n^2 L_1^2}{\beta\lambda_2} < 1, \tag{14}$$

$$B_{GT} \triangleq \frac{2\alpha n^2 d_w (1 + 2\alpha^2 L_1^2)}{\beta^2\lambda_2^2 - 4\alpha^2 n^2 L_1^2}. \tag{15}$$

Theorem 1 proves that consensus is achieved at an exponential rate up to a bias B_{GT}. Smaller network size n and low stochastic gradient variance lead to faster consensus as well as lower B_{GT} while lower problem dimensionality d_w lowers B_{GT} only. Among the hyperparameters, 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 given as

$$F(p(\bar{\boldsymbol{w}})) \triangleq \int p(\bar{\boldsymbol{w}}) \log\left(\frac{p(\bar{\boldsymbol{w}})}{p^*(\bar{\boldsymbol{w}})}\right) d\bar{\boldsymbol{w}}. \tag{16}$$

The corresponding convergence result is presented in Theorem 2.

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

$$F\Big(p(\bar{\boldsymbol{w}}_{k+1})\Big) \leq \exp(-\alpha\rho_U(k+1))F\Big(p(\bar{\boldsymbol{w}}_0)\Big) + C_F\sigma^k + O_{GT}, \tag{17}$$

where

$$C_F \triangleq \frac{(2n\alpha^2\bar{L}^2 + 1)nL_1^2}{\rho_U}\mathbb{E}[\|\Psi_0\|^2], \tag{18}$$

$$O_{GT} \triangleq \frac{2n\alpha\bar{L}^2 d_w + 4n^2\alpha^2 L_2^2\bar{L}^2(C_{\bar{\boldsymbol{w}}} + C_{\hat{\boldsymbol{w}}^*})}{\rho_U} + \frac{2\alpha n^3 d_w L_1^2(2n\alpha^2\bar{L}^2+1)(1+2\alpha^2 L_1^2)}{\rho_U(\beta^2\lambda_2^2 - 4\alpha^2 n^2 L_1^2)}. \tag{19}$$

Theorem 2 shows exponential convergence of the KL divergence up to a bias O_{GT}. A smaller network size n, lower stochastic gradient variance, 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.

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

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*

$$\alpha \leq \min\left\{ \left(\frac{(1-\sigma)\beta\lambda_2}{20d_w L_1^2(2L_1^2\bar{L}^2 + \bar{L}^2 + L_1^2)}\frac{\epsilon\rho_U}{n^4}\right)^{\frac{1}{3}}, \left(\frac{1}{20L_2^2\bar{L}^2(C_{\bar{\boldsymbol{w}}} + C_{\hat{\boldsymbol{w}}^*})}\frac{\epsilon\rho_U}{n^3}\right)^{\frac{1}{2}}, \right.$$
$$\left. \left(\frac{(1-\sigma)\beta\lambda_2}{10d_w(L_1^2 + (\beta^2\lambda_2^2 - 4\alpha^2n^2L_1^2)\bar{L}^2)}\frac{\epsilon\rho_U}{n^3}\right) \right\} \sim \mathcal{O}\left(\frac{\epsilon\rho_U}{n^3}\right), \tag{20}$$

$$k \geq \max\left\{\frac{\log(5F(p(\bar{\boldsymbol{w}}_0))/\epsilon)}{\alpha\rho_U} + 1, \frac{\log(5C_F/\epsilon)}{\log(1/\sigma)}\right\} \sim \mathcal{O}\left(\frac{n^3\log(1/\epsilon)}{\epsilon}\right). \tag{21}$$

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.

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*

$$\mathbb{E}[\|\tilde{\mathbf{d}}_{k+1}\|^2] = (1-\gamma\lambda_2)^{k+1}\mathbb{E}[\|\tilde{\mathbf{d}}_0\|^2] + \frac{8nL_1^2(n+\alpha^2L_2^2)}{\gamma^2\lambda_2^2}\mathbb{E}[\|\Psi_0\|^2]\sigma^k + E_{GT}, \tag{22}$$

where

$$E_{GT} = \frac{1}{\gamma^2\lambda_2^2}\Big(16n\alpha^2c_4^4(C_{\bar{\boldsymbol{w}}} + C_{\hat{\boldsymbol{w}}^*}) + 8n\alpha d_w(nL_1^2 + L_2^2)\Big) + \frac{16\alpha n^3L_1^2(n+\alpha^2L_2^2)(1+2\alpha^2L_1^2)d_w}{(\beta^2\lambda_2^2 - 4\alpha^2n^2L_1^2)\gamma^2\lambda_2^2}. \tag{23}$$

Theorem 4 shows an exponential reduction of the gradient error up to a bias E_{GT}. Smaller n, lower stochastic gradient variance 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 $\boldsymbol{d}_{i,k}$ for all $i \in \mathcal{V}$ to the true global gradient implies GT-DULA's robustness towards data heterogeneity in the training data. Since the corresponding bias E_{GT} is proportional to α, effects of data heterogeneity can be mitigated by appropriately choosing 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. A key point in the table is that the B for the GT-DULA scales with $\mathcal{O}(n^2)$ while that of the DULA is $\mathcal{O}(n^3)$. Similarly the O of the GT-DULA ($\mathcal{O}(n^4)$) is an order of magnitude lower than that of the DULA ($\mathcal{O}(n^5)$). Thus, the asymptotic biases of the GT-DULA scale better with the graph size than the DULA. The E for the GT-DULA scales with $\mathcal{O}(\alpha)$. Thus, by choosing a smaller α, the E in the GT-DULA can be made arbitrarily small, making it robust for data heterogeneity. However, the E in the DULA scales with $\mathcal{O}(1)$ (w.r.t. α), which implies the existence of a fixed bias that is uncontrollable by the choice of α and depends solely on the specific problem parameters. Thus, E cannot be made arbitrarily small for the DULA. Also note that $L_1^2 = L^2 + \sigma_L^2$ and $L_2^2 = \bar{L}^2 + \bar{\sigma}_L^2$, where σ_L^2 and $\bar{\sigma}_L^2$ are the contributions of the variance of the stochastic gradients (both of which become zero for full-batch gradients, thus reducing the asymptotic biases in the results). 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.

	GT-DULA	DULA
B	$\frac{2\alpha n^2 d_w(1+2\alpha^2 L_1^2)}{\beta^2\lambda_2^2-4\alpha^2 n^2 L_1^2}$	$\frac{2\alpha(4\alpha n^3 L_1^2 C_{\bar{\boldsymbol{w}}}+4\alpha n^2 L_1^2 C_{\hat{\mathbf{w}}^*}+2n^2 d_w)}{\beta^2\lambda_2^2-4\alpha^2 n^2 L_1^2}$
O	$\frac{1}{\rho_U}\left(2n\alpha\bar{L}^2 d_w + 4n^2\alpha^2 L_2^2\bar{L}^2(C_{\bar{\boldsymbol{w}}} + C_{\hat{\boldsymbol{w}}^*})\right)$ $+\frac{2\alpha n^2 d_w(2n^2\alpha^2 L_1^2\bar{L}^2+nL_1^2)(1+2\alpha^2 L_1^2)}{\rho_U(\beta^2\lambda_2^2-4\alpha^2 n^2 L_1^2)}$	$\frac{1}{\rho_U}\left(2n\alpha\bar{L}^2 d_w + 4n^2\alpha^2 L_2^2\bar{L}^2(C_{\bar{\boldsymbol{w}}} + C_{\hat{\boldsymbol{w}}^*})\right)$ $+\frac{(2n^2\alpha^2 L_1^2\bar{L}^2+nL_1^2)(2\alpha n^3 L_1^2 C_{\bar{\boldsymbol{w}}}+2\alpha n^2 L_1^2 C_{\hat{\mathbf{w}}^*}+n^2 d_w)}{\rho_U(\beta^2\lambda_2^2-4\alpha^2 n^2 L_1^2)}$
E	$\frac{16\alpha n^3 L_1^2(n+\alpha^2 L_2^2)(1+2\alpha^2 L_1^2)d_w}{(\beta^2\lambda_2^2-4\alpha^2 n^2 L_1^2)\gamma^2\lambda_2^2}$ $+\frac{1}{\gamma^2\lambda_2^2}\left(16n\alpha^2 c_4^4(C_{\bar{\boldsymbol{w}}} + C_{\hat{\boldsymbol{w}}^*}) + 8n\alpha d_w(nL_1^2 + L_2^2)\right)$	$\frac{4\alpha L_1^2(4\alpha n^3 L_1^2 C_{\bar{\boldsymbol{w}}}+4\alpha n^2 L_1^2 C_{\hat{\mathbf{w}}^*}+2n^2 d_w)}{\beta^2\lambda_2^2-4\alpha^2 n^2 L_1^2}$ $+4L_1^2(nC_{\bar{\boldsymbol{w}}} + C_{\hat{\mathbf{w}}^*})$

5. Numerical Simulations

In this section we present the simulation results performed on synthetic and real world data. All the results of GT-DULA are benchmarked against 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) shows the comparison of the KL divergence between GT-DULA and DULA from MCMC and analytical results while Figure 1(b) shows the effect of different mini-batch sizes for $n = 5$. Figure 1(a) shows a good match between the analytical and simulation results and clearly proves that the GT-DULA results in a lower bias than DULA. Figure 1(b) shows that the bias increases with smaller gradient batch size due to the stochastic noise. Next, we ran experiments with $n = 5$ and starting with $\alpha = 1e-5$ reduced α by $1/10$-th every 100 iterations to show the effect of reducing α on the asymptotic KL divergence and the gradient error which are shown in Figure 1(c) and (d) respectively. In Figure 1(d), 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. To do that, we split the 400 data points randomly into n groups to simulate distributed data for $n = \{5, 10, 20, 25, 40, 50, 80, 100, 200\}$. For each value of 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 as well. 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 as each of the 1000 cases encompasse varying degrees of heterogeneity due to their random split. This benefit is also more pronounced with increasing n.

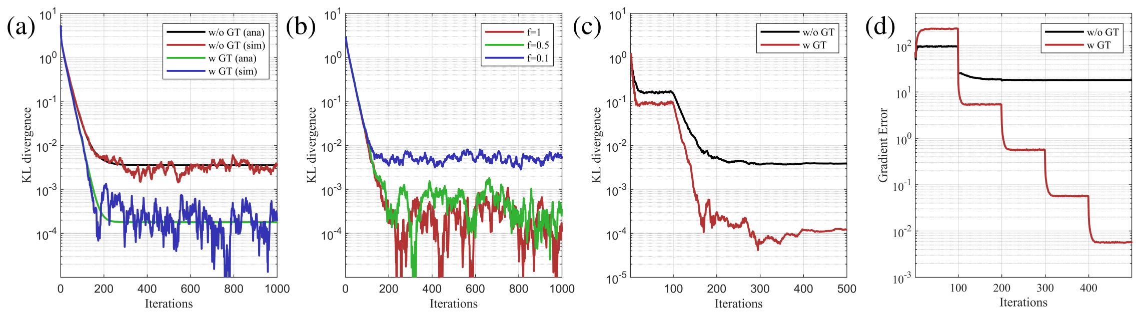

Figure 1: (a) Comparison of the KL divergence between GT-DULA with DULA via analytical and simulation results with full gradients. (b) Effect of different mini-batch sizes for stochastic gradient GT-DULA on the KL divergence; Comparison between GT-DULA and DULA of the effect of reducing step sizes on (c) asymptotic KL divergence and (d) the average gradient error.

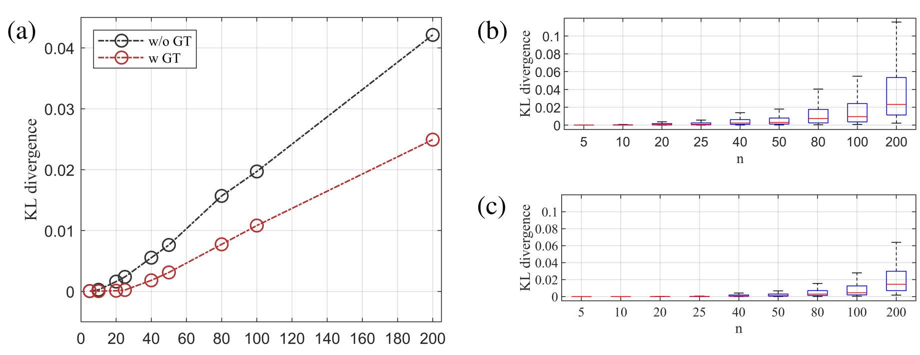

Figure 2: (a) Expected asymptotic KL divergence values for the posteriors generated by DULA vs GT-DULA for different network sizes. Distribution of the asymptotic KL divergence values for the posteriors generated by (b) DULA and (c) GT-DULA for different network sizes.

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.

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 as well compared to DULA. The strategy of reducing step size seems to slightly underperform for GT-DULA while it drastically underperforms 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, owing to the larger optimal step size of GT-DULA, it performs significantly faster with better consensus than DULA. Note that the asymptotic accuracy on the test data achieved here are slightly less than what was reported in Table 1 of Parayil et al. (2020) since they did not use heterogeneous training data.

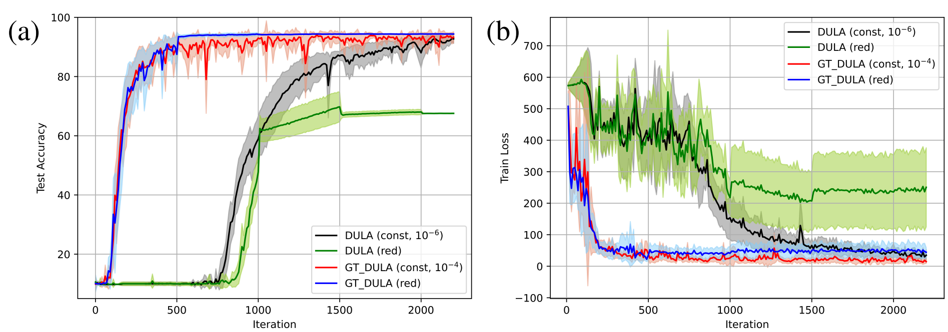

Figure 3: (a) Mean test data accuracy (b) Mean training loss for heterogeneously distributed MNIST dataset across 10 agents. (Here, 'const' refers to constant step size throughout case, while 'red' refers to the reduced step size case)

6. Conclusions

The GT-DULA presented in this paper provides a decentralized Bayesian learning algorithm with constant step sizes that 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. Furthermore, using only Lipschitz continuity, we prove stability of the GT-DULA with stochastic gradients without explicitly assuming bounded variance on the stochastic gradients.

References

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

Hilal Asi and John C Duchi. The importance of better models in stochastic optimization. *Proceedings of the National Academy of Sciences*, 116(46):22924–22930, 2019a.

Hilal Asi and John C Duchi. Stochastic (approximate) proximal point methods: Convergence, optimality, and adaptivity. *SIAM Journal on Optimization*, 29(3):2257–2290, 2019b.

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.

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.

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.

Xiaoyu Li and Francesco Orabona. On the convergence of stochastic gradient descent with adaptive stepsizes. In *The 22nd international conference on artificial intelligence and statistics*, pages 983–992. PMLR, 2019.

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 non-convex 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.

Vien V Mai and Mikael Johansson. Stability and convergence of stochastic gradient clipping: Beyond Lipschitz continuity and smoothness. In *International Conference on Machine Learning*, pages 7325–7335. PMLR, 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.

Panos Toulis, Dustin Tran, and Edo Airoldi. Towards stability and optimality in stochastic gradient descent. In *Artificial Intelligence and Statistics*, pages 1290–1298. PMLR, 2016.

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.

Jiaqi Zhang and Keyou You. Decentralized stochastic gradient tracking for non-convex empirical risk minimization. *arXiv preprint arXiv:1909.02712*, 2019.
iXtf l30

Gradient Tracking based Decentralized Bayesian Learning]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

$$\Psi_{k+1} = \mathcal{W}\Psi_k + \mathbf{e}_k \tag{24}$$

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 $\mathbf{e}_k \triangleq \begin{bmatrix} \sqrt{2\alpha n}(M \otimes I_{d_w})\mathbf{v}_k \\ \alpha(\hat{\mathbf{g}}_{k+1} - \hat{\mathbf{g}}_k) \end{bmatrix}$. Taking the norm of (24) yields

$$\|\Psi_{k+1}\| \leq (1 - \beta\lambda_2)\|\Psi_k\| + \|\mathbf{e}_k\|, \tag{25}$$

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

$$\|\mathbf{e}_k\|^2 = 2\alpha n\|\mathbf{v}_k\|^2 + \alpha^2\|\hat{\mathbf{g}}_{k+1} - \hat{\mathbf{g}}_k\|^2 \tag{26}$$

where we used $\|M \otimes I_{d_w}\| \leq 1$. Squaring (25) and applying the identity $(x + y)^2 \leq (\theta + 1)x^2 + \left(\frac{x+1}{x}\right)y^2$ for any $\theta > 0$ (with $\theta = (1 - \beta\lambda_2)^{-1} - 1 > 0$) yields

$$\|\Psi_{k+1}\|^2 \leq (1 - \beta\lambda_2)\|\Psi_k\|^2 + \frac{1}{\beta\lambda_2}\|\mathbf{e}_k\|^2, \tag{27}$$

$$\leq (1 - \beta\lambda_2)\|\Psi_k\|^2 + \frac{1}{\beta\lambda_2}\Big(2\alpha n\|\mathbf{v}_k\|^2 + \alpha^2\|\hat{\mathbf{g}}_{k+1} - \hat{\mathbf{g}}_k\|^2\Big), \tag{28}$$

where we substituted (26) in (28). Next, we need to establish a bound for $\|\hat{\mathbf{g}}_{k+1} - \hat{\mathbf{g}}_k\|^2$.

$$\|\mathbf{g}'_{k+1} - \mathbf{g}'_k\|^2 = \sum_{i\in\mathcal{V}}\left[\left(\frac{M_i}{m_i}\sum_{j=1}^{m_i} L_{i,j} + \frac{L'_i}{n}\right)^2 \|\boldsymbol{w}_{i,k+1} - \boldsymbol{w}_{i,k}\|^2\right] \tag{29}$$

Let $\mathcal{B}_k \in \mathbf{B}$ be the randomness generated by the stochastic gradients and define $Y_i \triangleq \left(\frac{M_i}{m_i}\sum_{j=1}^{m_1} L_{ij} + \frac{L'_i}{n}\right)$. Then, taking the expectation of (29) w.r.t. $\mathbf{B}$ gives

$$\mathbb{E}_{\mathbf{B}}\|\mathbf{g}'_{k+1} - \mathbf{g}'_k\|^2 = \sum_{i\in\mathcal{V}}\left[\mathbb{E}_{\mathbf{B}}[Y_i^2] \times \|\boldsymbol{w}_{i,k+1} - \boldsymbol{w}_{i,k}\|^2\right], \tag{30}$$

Applying the result from (82) in Lemma 6 to (30) yields

$$\mathbb{E}_{\mathbf{B}}\|\hat{\mathbf{g}}_{k+1} - \hat{\mathbf{g}}_k\|^2 \leq L_1^2\|\mathbf{w}_{k+1} - \mathbf{w}_k\|^2. \tag{31}$$

Again, from (7) we can write

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

$$= \beta(\mathcal{L} \otimes I_{d_w})(\mathbf{w}_k - \mathbf{1}_n \otimes \bar{\boldsymbol{w}}_k) - \alpha n\mathbf{d}_k + \sqrt{2\alpha n}\mathbf{v}_k, \tag{32}$$

$$= \beta(\mathcal{L} \otimes I_{d_w})\tilde{\mathbf{w}}_k - \alpha n\mathbf{d}_k + \sqrt{2\alpha n}\mathbf{v}_k. \tag{33}$$

In (32) we use the fact that $(\mathcal{L} \otimes I_{d_w})(\mathbf{1}_n \otimes \bar{\boldsymbol{w}}_k) = \mathbf{0}_{nd_w}$. Then, taking the norm of (33) yields

$$\|\mathbf{w}_{k+1} - \mathbf{w}_k\| \leq \beta\lambda_n\|\tilde{\mathbf{w}}_k\| + n\|\alpha\mathbf{d}_k\| + \sqrt{2\alpha n}\|\mathbf{v}_k\| \leq n\|\tilde{\mathbf{w}}_k\| + n\|\alpha\mathbf{d}_k\| + \sqrt{2\alpha n}\|\mathbf{v}_k\|, \quad (34)$$

where we use first inequality in Condition 1. Squaring (34) and noting that $\|\Psi_k\|^2 = \|\tilde{\mathbf{w}}_k\|^2 + \|\alpha\mathbf{d}_k\|^2$ gives

$$\|\mathbf{w}_{k+1} - \mathbf{w}_k\|^2 \leq 4n^2\|\Psi_k\|^2 + 4\alpha n\|\mathbf{v}_k\|^2. \quad (35)$$

Thereby, substituting (33) in(31) results in

$$\mathbb{E}_{\mathbf{B}}\|\hat{\mathbf{g}}_{k+1} - \hat{\mathbf{g}}_k\|^2 \leq 4n^2L_1^2\|\Psi_k\|^2 + 4\alpha nL_1^2\|\mathbf{v}_k\|^2. \quad (36)$$

Taking $\mathbb{E}_{\mathbf{B}}[\cdot]$ of (28) and substituting (36) in it gives

$$\mathbb{E}_{\mathbf{B}}[\|\Psi_{k+1}\|^2] \leq \left(1 - \beta\lambda_2 + \frac{4\alpha^2n^2L_1^2}{\beta\lambda_2}\right)\|\Psi_k\|^2 + \frac{2\alpha n(1+2\alpha^2L_1^2)}{\beta\lambda_2}\|\mathbf{v}_k\|^2, \quad (37)$$

Finally, taking the total expectation of (37) yields

$$\mathbb{E}[\|\Psi_{k+1}\|^2] \leq \left(1 - \beta\lambda_2 + \frac{4\alpha^2n^2L_1^2}{\beta\lambda_2}\right)\mathbb{E}[\|\Psi_k\|^2] + \frac{2\alpha n^2d_w(1+2\alpha^2L_1^2)}{\beta\lambda_2}, \quad (38)$$

where $\mathbb{E}[\|\mathbf{v}_k\|^2] \leq nd_w$. Note, that $\left(1 - \beta\lambda_2 + \frac{4\alpha^2n^2L_1^2}{\beta\lambda_2}\right) \in (0,1)$ from condition 1, which assures the convergence of $\mathbb{E}[\|\Psi_{k+1}\|^2]$. Further, iteratively using (38) we establish the rate of consensus for the proposed GT-DULA which is presented in Theorem 1.

Proof of Theorem 2

We start off with the average dynamics generated by the GT-DULA. From (5) and noting that $\boldsymbol{d}_{i,0} = \hat{\boldsymbol{g}}_{i,0}$, it is trivial that $\sum_{i\in\mathcal{V}}\boldsymbol{d}_{i,k} = \sum_{i\in\mathcal{V}}\hat{\boldsymbol{g}}_{i,k}$. Thereafter, from (4), the following average dynamics can be established.

$$\bar{\boldsymbol{w}}_{k+1} = \bar{\boldsymbol{w}}_k - \alpha\boldsymbol{G}_k + \sqrt{2\alpha}\bar{\boldsymbol{v}}_k, \quad (39)$$

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

$$\boldsymbol{G}_k = \overline{\nabla E}_k + \xi_k + \zeta_k, \quad (40)$$

where $\overline{\nabla E}_k \triangleq \sum_{i\in\mathcal{V}}\nabla E_i(\bar{\boldsymbol{w}}_k)$, $\xi_k \triangleq \sum_{i\in\mathcal{V}}\left(\widehat{\nabla E}_i(\boldsymbol{w}_{i,k}) - \widehat{\nabla E}_i(\bar{\boldsymbol{w}}_k)\right)$ and $\zeta_k \triangleq \sum_{i\in\mathcal{V}}\left(\widehat{\nabla E}_i(\bar{\boldsymbol{w}}_k) - \nabla E_i(\bar{\boldsymbol{w}}_k)\right)$. Hence, in essence, $\overline{\nabla E}_k$ represents the gradient computed at the average sample, ξ_k encompasses the deviation of due to local gradients and is a consequence of the distributed learning setup, and ζ_k is the error incurred by mini-batch gradients. Also, note that since the stochastic gradient $\widehat{\nabla E}_i(\cdot)$ is an unbiased estimator of the true gradient $\nabla E_i(\cdot)$, we have $\mathbb{E}_{\mathbf{B}}[\zeta_k] = \mathbf{0}_{d_w}$. From our succeeding analysis we can conclude that as long as the sources of additional deviation of the net gradient in (39) from $\overline{\nabla E}$ are bounded, the resulting algorithm asymptotically converges to some neighborhood of the target distribution.

With (40) in mind, (39) can be written as a stochastic differential equation in continuous-time as below.

$$d\bar{\boldsymbol{w}}(t) = -\boldsymbol{G}_k dt + \sqrt{2}d\boldsymbol{B}(t) = -\Big(\overline{\nabla E}_k + \xi_k + \zeta_k\Big)dt + \sqrt{2}d\boldsymbol{B}(t), \tag{41}$$

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 $\boldsymbol{y}_{1,k} \triangleq \bar{\boldsymbol{w}}_k$, $\boldsymbol{y}_{2,k} \triangleq \tilde{\mathbf{w}}_k$, $\boldsymbol{y}_{3,k} \triangleq \mathcal{B}_k$ and $\boldsymbol{y}_k \triangleq [\boldsymbol{y}_{1,k}^\top, \boldsymbol{y}_{2,k}^\top, \boldsymbol{y}_{3,k}^\top]^\top$ and following a similar approach as in (33) of Bhar et al. (2022) we can write down the Fokker-Planck (FP) equation for (41) which gives the continuous-time evolution of the distribution of $\bar{\boldsymbol{w}}(t)$ as

$$\frac{\partial p(\bar{\boldsymbol{w}}(t))}{\partial t} = -\nabla \cdot \left[\int \sum_{\mathbf{B}} p(\bar{\boldsymbol{w}}(t)|\boldsymbol{y}_k)\Big(-\overline{\nabla E}_k - \xi_k\Big)p(\boldsymbol{y}_k)d\boldsymbol{y}_k\right] + \nabla^2 p(\bar{\boldsymbol{w}}(t)), \tag{42}$$

where we used the fact that $\sum_{\mathbb{B}} \zeta_k p(\boldsymbol{y}_{k,3}) = \mathbb{E}_{\mathbb{B}}[\zeta_k] = \mathbf{0}_{d_w}$. Thereafter, proceeding with (42) in the same way as in (S101)-(S125) from Parayil et al. (2020) yields

$$\begin{aligned}\dot{F}\Big(p(\bar{\boldsymbol{w}}(t))\Big) = -\frac{1}{2}\mathbb{E}\left\|\nabla \log\left(\frac{p(\bar{\boldsymbol{w}}(t))}{p^*(\bar{\boldsymbol{w}}(t))}\right)\right\|^2 + \iint \Big\|\overline{\nabla E}_t - \overline{\nabla E}_k\Big\|^2 p(\bar{\boldsymbol{w}}(t))d\boldsymbol{y}_k \\ + \iint \|\xi_k\|^2 p(\bar{\boldsymbol{w}}(t))d\boldsymbol{y}_k,\end{aligned} \tag{43}$$

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

$$\|\xi_k\|^2 = \left\|\sum_{i\in\mathcal{V}}\Big(\widehat{\nabla E}_i(\boldsymbol{w}_{i,k}) - \widehat{\nabla E}_i(\bar{\boldsymbol{w}}_k)\Big)\right\|^2 \leq n\sum_{i\in\mathcal{V}}\Big(Y_i^2\|\boldsymbol{w}_{i,k} - \bar{\boldsymbol{w}}_k\|^2\Big), \tag{44}$$

where Y_i is defined in Lemma 6. Taking the expectation of (44) w.r.t. $\mathbf{B}$ and thereby applying (82) from Lemma 6 we get

$$\mathbb{E}_{\mathbf{B}}[\|\xi_k\|^2] \leq nL_1^2\|\tilde{\mathbf{w}}_k\|^2, \tag{45}$$

which after marginalizing w.r.t. $\boldsymbol{y}_k$ yields

$$\mathbb{E}[\|\xi_k\|^2] \leq nL_1^2\mathbb{E}[\|\tilde{\mathbf{w}}_k\|^2], \tag{46}$$

Next, we analyze $\Big\|\overline{\nabla E}_t - \overline{\nabla E}_k\Big\|^2$.

$$\Big\|\overline{\nabla E}_t - \overline{\nabla E}_k\Big\|^2 \leq n\bar{L}^2\|\bar{\boldsymbol{w}}(t) - \bar{\boldsymbol{w}}(t_k)\|^2, \tag{47}$$

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 (41) from t_k to $t \in [t_k, t_{k+1})$ gives

$$\begin{aligned}\|\bar{\boldsymbol{w}}(t) - \bar{\boldsymbol{w}}(t_k)\|^2 &\leq \Big\| - \boldsymbol{G}_k(t - t_k) + \sqrt{2}\Big(\boldsymbol{B}(t) - \boldsymbol{B}(t_k)\Big)\Big\|^2, \\ &\leq 2\|\boldsymbol{B}(t) - \boldsymbol{B}(t_k)\|^2 + \|\boldsymbol{G}_k(t - t_k)\|^2 - 2\sqrt{2}\boldsymbol{S}_k, && (48) \\ &\leq 2\|\boldsymbol{B}(t) - \boldsymbol{B}(t_k)\|^2 + \alpha^2\|\boldsymbol{G}_k\|^2 - 2\sqrt{2}\boldsymbol{S}_k, && (49)\end{aligned}$$

where $\boldsymbol{S}_k \triangleq \Big(\boldsymbol{B}(t) - \boldsymbol{B}(t_k)\Big)^\top \Big(\boldsymbol{G}_k(t-t_k)\Big)$. In (49) we use $t - t_k < t_{k+1} - t_k = \alpha$ for any $t \in [t_k, t_{k+1})$. Substituting (49) in (47) results in

$$\left\|\overline{\nabla E}_t - \overline{\nabla E}_k\right\|^2 \le n\bar{L}^2\Big[2\|\boldsymbol{B}(t) - \boldsymbol{B}(t_k)\|^2 + \alpha^2\|\boldsymbol{G}_k\|^2 - 2\sqrt{2}\boldsymbol{S}_k\Big], \tag{50}$$

Now, the $\|\boldsymbol{G}_k\|^2$ can be bound as

$$\begin{aligned}
\|\boldsymbol{G}_k\|^2 &= \left\|\sum_{i\in\mathcal{V}} \widehat{\nabla E}_i(\boldsymbol{w}_{i,k})\right\|^2 = \left\|\sum_{i\in\mathcal{V}} \Big(\widehat{\nabla E}_i(\boldsymbol{w}_{i,k}) - \widehat{\nabla E}_i(\bar{\boldsymbol{w}}_k) + \widehat{\nabla E}_i(\bar{\boldsymbol{w}}_k) - \widehat{\nabla E}_i(\hat{\boldsymbol{w}}^*)\Big)\right\|^2, \\
&\le 2\,\|\xi_k\|^2 + 2\left\|\sum_{i\in\mathcal{V}} \Big(\widehat{\nabla E}_i(\bar{\boldsymbol{w}}_k) - \widehat{\nabla E}_i(\hat{\boldsymbol{w}}^*)\Big)\right\|^2 \le 2\|\xi_k\|^2 + 2n\left(\sum_{i\in\mathcal{V}} Y_i^2\right)\|\bar{\boldsymbol{w}}_k - \hat{\boldsymbol{w}}^*\|^2, \\
&\le 2n\sum_{i\in\mathcal{V}}\Big(Y_i^2\|\boldsymbol{w}_{i,k} - \bar{\boldsymbol{w}}_k\|^2\Big) + 4n\left(\sum_{i\in\mathcal{V}} Y_i^2\right)\Big(\|\bar{\boldsymbol{w}}_k\|^2 + \|\hat{\boldsymbol{w}}^*\|^2\Big),
\end{aligned} \tag{51}$$

where we used the bound from (44) and $\hat{\boldsymbol{w}}^*$ is some local extremum of $\sum_{i\in\mathcal{V}} \widehat{\nabla E}_i(\cdot)$, i.e., $\sum_{i\in\mathcal{V}} \widehat{\nabla E}_i(\hat{\boldsymbol{w}}^*) = \mathbf{0}_{d_w}$. Thereafter, taking the expectation w.r.t. $\mathbf{B}$ of (51) and using (82) and (83) in Lemma 6 yields

$$\mathbb{E}_{\mathbf{B}}[\|\boldsymbol{G}_k\|^2] \le 2nL_1^2\|\tilde{\mathbf{w}}_k\|^2 + 4nL_2^2\big(\|\bar{\boldsymbol{w}}_k\|^2 + \|\hat{\boldsymbol{w}}^*\|^2\big), \tag{52}$$

Next, taking $\mathbb{E}_{\mathbf{B}}[\cdot]$ of (50) and substituting (52) yields

$$\begin{aligned}
\mathbb{E}_{\mathbf{B}}\left\|\overline{\nabla E}_t - \overline{\nabla E}_k\right\|^2 \le 2n\bar{L}^2\|\boldsymbol{B}(t) - \boldsymbol{B}(t_k)\|^2 + 2n^2\alpha^2L_1^2\bar{L}^2\|\tilde{\mathbf{w}}_k\|^2 + 4n^2\alpha^2L_2^2\bar{L}^2\|\bar{\boldsymbol{w}}_k\|^2 \\
+ 4n^2\alpha^2L_2^2\bar{L}^2\|\hat{\boldsymbol{w}}^*\|^2 - 2\sqrt{2}n\bar{L}^2\boldsymbol{S}_k',
\end{aligned} \tag{53}$$

where $\boldsymbol{S}_k' \triangleq \Big(\boldsymbol{B}(t) - \boldsymbol{B}(t_k)\Big)^\top \Big(\mathbb{E}_{\mathbf{B}}[\boldsymbol{G}_k](t-t_k)\Big)$. Again, marginalizing (53) w.r.t. $\boldsymbol{y}_k$ gives

$$\mathbb{E}\left\|\overline{\nabla E}_t - \overline{\nabla E}_k\right\|^2 \le 2n\alpha\bar{L}^2d_w + 2n^2\alpha^2L_1^2\bar{L}^2\mathbb{E}[\|\tilde{\mathbf{w}}_k\|^2] + 4n^2\alpha^2L_2^2\bar{L}^2C_{\bar{\boldsymbol{w}}} + 4n^2\alpha^2L_2^2\bar{L}^2C_{\hat{\boldsymbol{w}}^*}, \tag{54}$$

where $\mathbb{E}[\|\hat{\boldsymbol{w}}^*\|^2] \le C_{\hat{\boldsymbol{w}}^*}$ for any choice of stochastic gradient. For details on the derivation of (54), refer to (S135)-(S141) in Parayil et al. (2020). Finally, incorporating (46) and (54) in (43) yields

$$\begin{aligned}
\dot{F}\Big(p(\bar{\boldsymbol{w}}(t))\Big) &\le -\frac{1}{2}\mathbb{E}\left\|\nabla\log\left(\frac{p(\bar{\boldsymbol{w}}(t))}{p^*(\bar{\boldsymbol{w}}(t))}\right)\right\|^2 + 2n\alpha\bar{L}^2d_w + (2n^2\alpha^2L_1^2\bar{L}^2 + nL_1^2)\mathbb{E}[\|\tilde{\mathbf{w}}_k\|^2] \\
&\qquad + 4n^2\alpha^2L_2^2\bar{L}^2C_{\bar{\boldsymbol{w}}} + 4n^2\alpha^2L_2^2\bar{L}^2C_{\hat{\boldsymbol{w}}^*}, \\
&\le -\frac{1}{2}\mathbb{E}\left\|\nabla\log\left(\frac{p(\bar{\boldsymbol{w}}(t))}{p^*(\bar{\boldsymbol{w}}(t))}\right)\right\|^2 + (2n^2\alpha^2L_1^2\bar{L}^2 + nL_1^2)\mathbb{E}[\|\tilde{\mathbf{w}}_k\|^2] + f,
\end{aligned} \tag{55}$$

where $f \triangleq 2n\alpha\bar{L}^2d_w + 4n^2\alpha^2L_2^2\bar{L}^2C_{\bar{\boldsymbol{w}}} + 4n^2\alpha^2L_2^2\bar{L}^2C_{\hat{\boldsymbol{w}}^*}$. Here we utilize the LSI assumption in Assumption 3 and putting $g(\bar{\boldsymbol{w}}) = \frac{p(\bar{\boldsymbol{w}}(t))}{p^*(\bar{\boldsymbol{w}}(t))}$ results in the following results

$$F\Big(p(\bar{\boldsymbol{w}}(t))\Big) \le \frac{1}{2\rho_U}\mathbb{E}\left\|\nabla\log\left(\frac{p(\bar{\boldsymbol{w}}(t))}{p^*(\bar{\boldsymbol{w}}(t))}\right)\right\|^2. \tag{56}$$

Using (56) in (55) yields

$$\dot{F}\Big(p(\bar{\boldsymbol{w}}(t))\Big) \leq -\rho_U F\Big(p(\bar{\boldsymbol{w}}(t))\Big) + f + (2n^2\alpha^2 L_1^2 \bar{L}^2 + nL_1^2)\mathbb{E}[\|\tilde{\mathbf{w}}_k\|^2]. \tag{57}$$

Next, integrating (57) w.r.t t within $t \in [t_k, t_{k+1}]$ and utilizing $t_{k+1} - t_k < \alpha$ gives us the evolution of the KL divergence of the posterior generated by GT-DULA samples as follows.

$$F\Big(p(\bar{\boldsymbol{w}}_{k+1})\Big) \leq \exp(-\alpha\rho_U) F\Big(p(\bar{\boldsymbol{w}}_k)\Big) + \frac{1-\exp(-\alpha\rho_U)}{\rho_U}\Big[f_k + (2n^2\alpha^2 L_1^2\bar{L}^2 + nL_1^2)\mathbb{E}[\|\tilde{\mathbf{w}}_k\|^2]\Big]. \tag{58}$$

Using (58) iteratively yields

$$\begin{aligned} F\Big(p(\bar{\boldsymbol{w}}_{k+1})\Big) &\leq \exp(-\alpha\rho_U(k+1))F\Big(p(\bar{\boldsymbol{w}}_0)\Big) + \frac{1-\exp(-\alpha\rho_U)}{\rho_U}\sum_{\ell=0}^{k}\Big[\exp(-\alpha\rho_U(k-\ell))\times \\ &\qquad \Big(f + (2n^2\alpha^2 L_1^2\bar{L}^2 + nL_1^2)\mathbb{E}[\|\tilde{\mathbf{w}}_\ell\|^2]\Big)\Big], \\ &\leq \exp(-\alpha\rho_U(k+1))F\Big(p(\bar{\boldsymbol{w}}_0)\Big) + \frac{f}{\rho_U} + \frac{2n^2\alpha^2 L_1^2\bar{L}^2 + nL_1^2}{\rho_U}\mathbb{E}[\|\tilde{\mathbf{w}}_k\|^2], \end{aligned} \tag{59}$$

where we used $\mathbb{E}[\|\tilde{\mathbf{w}}_\ell\|^2] \leq \mathbb{E}[\|\tilde{\mathbf{w}}_k\|^2] \leq \mathbb{E}[\|\Psi_k\|^2]$ for any $\ell \in [0,k]$ and $\sum_{\ell=0}^{k}\exp(-\alpha\rho_U(k-\ell)) \leq \sum_{\ell=0}^{\infty}\exp(-\alpha\rho_U\ell) = \frac{1}{1-\exp(-\alpha\rho_U)}$. Finally, substituting (13) in (59), we get the rate of convergence of the KL divergence of the generated posteriors which is presented in Theorem 2

Proof of Corollary 3

From (17), $F\Big(p(\bar{\boldsymbol{w}}_{k+1})\Big) \leq \epsilon$ can be satisfied if (i) $\exp(-\alpha\rho_U(k+1))F\Big(p(\bar{\boldsymbol{w}}_0)\Big) \leq \frac{\epsilon}{5}$, (ii) $C_d\sigma^k \leq \frac{\epsilon}{5}$ and (iii) $O_{GT} \leq \frac{3\epsilon}{5}$. (i) and (ii) respectively give the minimum k values in (21). Finally, (iii) can be satisfied if the we simultaneously satisfy the following conditions

$$\frac{8d_w L_1^4\bar{L}^2 n^4\alpha^5}{(1-\sigma)\rho_U\beta\lambda_2} + \frac{4d_w L_1^2\bar{L}^2 n^4\alpha^4}{(1-\sigma)\rho_U\beta\lambda_2} + \frac{4d_w L_1^4 n^3\alpha^3}{(1-\sigma)\rho_U\beta\lambda_2} \leq \frac{\epsilon}{5}, \tag{60}$$

$$\frac{4L_2^2\bar{L}^2(C_{\bar{\boldsymbol{w}}} + C_{\hat{\boldsymbol{w}}^*})n^2\alpha^2}{\rho_U} \leq \frac{\epsilon}{5}, \tag{61}$$

$$\frac{2d_w L_1^2 n^3\alpha}{(1-\sigma)\rho_U\beta\lambda_2} + \frac{2d_w\bar{L}^2 n\alpha}{\rho_U} \leq \frac{\epsilon}{5}, \tag{62}$$

each of which result in bounds in (20) respectively.

Proof of Theorem 4

From (8) we have

$$\tilde{\mathbf{d}}_{k+1} = (\mathcal{W}_\gamma \otimes I_{d_w})\tilde{\mathbf{d}}_k + (\hat{\mathbf{g}}_{k+1} - \hat{\mathbf{g}}_k) - \mathbf{1}_n \otimes (\bar{\boldsymbol{u}}_{k+1} - \bar{\boldsymbol{u}}_k). \tag{63}$$

Taking the square of the norm of (63) yields

$$\|\tilde{\mathbf{d}}_{k+1}\|^2 \leq (1-\gamma\lambda_2)\|\tilde{\mathbf{d}}_k\|^2 + \frac{2}{\gamma\lambda_2}\|\hat{\mathbf{g}}_{k+1} - \hat{\mathbf{g}}_k\|^2 + \frac{2n}{\gamma\lambda_2}\|\bar{\boldsymbol{u}}_{k+1} - \bar{\boldsymbol{u}}_k\|, \tag{64}$$

and thereafter taking the total expectation of (64) gives

$$\mathbb{E}[\|\tilde{\mathbf{d}}_{k+1}\|^2] \leq (1-\gamma\lambda_2)\mathbb{E}[\|\tilde{\mathbf{d}}_k\|^2] + \frac{2}{\gamma\lambda_2}\mathbb{E}[\|\hat{\mathbf{g}}_{k+1} - \hat{\mathbf{g}}_k\|^2] + \frac{2n}{\gamma\lambda_2}\mathbb{E}[\|\bar{\boldsymbol{u}}_{k+1} - \bar{\boldsymbol{u}}_k\|], \tag{65}$$

Next, we derive bounds for the individual terms on the right hand side of (65). First, from (36) we have

$$\mathbb{E}[\|\hat{\mathbf{g}}_{k+1} - \hat{\mathbf{g}}_k\|^2] \leq 4n^2 L_1^2 \mathbb{E}[\|\Psi_k\|^2] + 4\alpha n^2 L_1^2 d_w. \tag{66}$$

Then,

$$\mathbb{E}[\|\bar{\boldsymbol{u}}_{k+1} - \bar{\boldsymbol{u}}_k\|^2] = \frac{L_2^2}{n}\mathbb{E}[\|\bar{\boldsymbol{w}}_{k+1} - \bar{\boldsymbol{w}}_k\|^2], \tag{67}$$

and from (39) we can write

$$\bar{\boldsymbol{w}}_{k+1} - \bar{\boldsymbol{w}}_k = -\alpha \boldsymbol{G}_k + \sqrt{2\alpha}\bar{\boldsymbol{v}}_k. \tag{68}$$

Thus, from (68)

$$\|\bar{\boldsymbol{w}}_{k+1} - \bar{\boldsymbol{w}}_k\|^2 \leq 2\alpha^2\|\boldsymbol{G}_k\|^2 + 4\alpha\|\bar{\boldsymbol{v}}_k\|^2. \tag{69}$$

Taking the total expectation of (69) and substituting (52) yields

$$\mathbb{E}[\|\bar{\boldsymbol{w}}_{k+1} - \bar{\boldsymbol{w}}_k\|^2] \leq 4n\alpha^2 L_1^2 \mathbb{E}[\|\tilde{\mathbf{w}}_k\|^2] + 8n\alpha^2 L_2^2 \big(C_{\bar{\boldsymbol{w}}} + C_{\hat{\boldsymbol{w}}^*}\big) + 4n\alpha d_w, \tag{70}$$

where we also use $\mathbb{E}[\|\bar{\boldsymbol{v}}_k\|^2] \leq n d_w$. Substituting (70) in (67) gives

$$\mathbb{E}[\|\bar{\boldsymbol{u}}_{k+1} - \bar{\boldsymbol{u}}_k\|^2] \leq 4\alpha^2 L_1^2 L_2^2 \mathbb{E}[\|\tilde{\mathbf{w}}_k\|^2] + 8\alpha^2 L_2^4 \big(C_{\bar{\boldsymbol{w}}} + C_{\hat{\boldsymbol{w}}^*}\big) + 4\alpha d_w L_2^2, \tag{71}$$

Finally, substituting (66) and (71) in (65) leads to

$$\begin{aligned}\mathbb{E}[\|\tilde{\mathbf{d}}_{k+1}\|^2] \leq (1-\gamma\lambda_2)\mathbb{E}[\|\tilde{\mathbf{d}}_k\|^2] &+ \frac{8nL_1^2(n+\alpha^2 L_2^2)}{\gamma\lambda_2}\mathbb{E}[\|\Psi_k\|^2] + \frac{16n\alpha^2 L_2^4(C_{\bar{\boldsymbol{w}}} + C_{\hat{\boldsymbol{w}}^*})}{\gamma\lambda_2} \\ &+ \frac{8n\alpha d_w(nL_1^2 + L_2^2)}{\gamma\lambda_2},\end{aligned} \tag{72}$$

where we used $\mathbb{E}[\|\tilde{\mathbf{w}}_k\|^2] \leq \mathbb{E}[\|\Psi_k\|^2]$. Next, iteratively using (72) yields

$$\begin{aligned}\mathbb{E}[\|\tilde{\mathbf{d}}_{k+1}\|^2] \leq (1-\gamma\lambda_2)^{k+1}\mathbb{E}[\|\tilde{\mathbf{d}}_0\|^2] &+ \frac{8nL_1^2(n+\alpha^2 L_2^2)}{\gamma^2\lambda_2^2}\mathbb{E}[\|\Psi_k\|^2] + \frac{16n\alpha^2 L_2^4(C_{\bar{\boldsymbol{w}}} + C_{\hat{\boldsymbol{w}}^*})}{\gamma^2\lambda_2^2} \\ &+ \frac{8n\alpha d_w(nL_1^2 + L_2^2)}{\gamma^2\lambda_2^2},\end{aligned} \tag{73}$$

where we use $\mathbb{E}[\|\Psi_\ell\|^2] \leq \mathbb{E}[\|\Psi_k\|^2]$ for all $\ell \in [0,k]$ and $\sum_{\ell=0}^{k}(1-\gamma\lambda_2)^\ell < \sum_{\ell=0}^{\infty}(1-\gamma\lambda_2)^\ell = \frac{1}{\gamma\lambda_2}$ has been used. Next, substituting (13) in (73) yields our final result for the gradient error presented in Theorem 4 below.

Useful Lemmas

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

$$\mathbb{E}[\|\bar{\boldsymbol{w}}(t_k)\|^2] \leq C_{\bar{\boldsymbol{w}}}, \tag{74}$$

where

$$C_{\bar{\boldsymbol{w}}} = \max\Bigg\{\mathbb{E}[\|\bar{\boldsymbol{w}}_0\|^2], \frac{1}{\rho_U^2 - 16n^2\alpha^2 L_2^2 \bar{L}^2}\Bigg(2\rho_U^2 c_1 + 4\rho_U F\Big(p(\bar{\boldsymbol{w}}_0)\Big) + 8\big(n\alpha\bar{L}^2 d_w + 2n^2\alpha^2 L_2^2 \bar{L}^2 C_{\hat{\boldsymbol{w}}^*}\big) + 4(2n^2\alpha^2 L_1^2\bar{L}^2 + nL_1^2)\big(\mathbb{E}[\|\Psi_0\|^2] + B_{GT}\big)\Bigg)\Bigg\}. \tag{75}$$

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

$$\mathbb{E}[\|\bar{\boldsymbol{w}}_{k+1}\|^2] \leq 2c_1 + \frac{4}{\rho_U} F\Big(p(\bar{\boldsymbol{w}}_{k+1})\Big), \tag{76}$$

where $\mathbb{E}_{p^*}[\|\bar{\boldsymbol{w}}\|^2] \leq c_1$. From (59), we can write

$$F\Big(p(\bar{\boldsymbol{w}}_{k+1})\Big) \leq F\Big(p(\bar{\boldsymbol{w}}_0)\Big) + \frac{1}{\rho_U}\Big(2n\alpha\bar{L}^2 d_w + 4n^2\alpha^2 L_2^2\bar{L}^2 C_{\bar{\boldsymbol{w}}} + 4n^2\alpha^2 L_2^2\bar{L}^2 C_{\hat{\boldsymbol{w}}^*} + (2n^2\alpha^2 L_1^2\bar{L}^2 + nL_1^2)\mathbb{E}[\|\tilde{\mathbf{w}}_k\|^2]\Big). \tag{77}$$

Note, in the right hand side of (77), we have used the bound $C_{\bar{\boldsymbol{w}}}$ since it is assumed to hold up to $\ell \leq k$. Next, from Theorem 1, we can write $\mathbb{E}[\|\tilde{\mathbf{w}}_k\|^2] \leq \mathbb{E}[\|\Phi_0\|^2] + B_{GT}$ which when substituted in (77) yields

$$F\Big(p(\bar{\boldsymbol{w}}_{k+1})\Big) \leq F\Big(p(\bar{\boldsymbol{w}}_0)\Big) + \frac{1}{\rho_U}\Big(2n\alpha\bar{L}^2 d_w + 4n^2\alpha^2 L_2^2\bar{L}^2 C_{\bar{\boldsymbol{w}}} + 4n^2\alpha^2 L_2^2\bar{L}^2 C_{\hat{\boldsymbol{w}}^*} + (2n^2\alpha^2 L_1^2\bar{L}^2 + nL_1^2)\big(\mathbb{E}[\|\Phi_0\|^2] + B_{GT}\big)\Big). \tag{78}$$

Next, substituting (78) in (76) results in

$$\mathbb{E}[\|\bar{\boldsymbol{w}}_{k+1}\|^2] \leq 2c_1 + \frac{4}{\rho_U} F\Big(p(\bar{\boldsymbol{w}}_0)\Big) + \frac{4}{\rho_U^2}\Big(2n\alpha\bar{L}^2 d_w + 4n^2\alpha^2 L_2^2\bar{L}^2 C_{\bar{\boldsymbol{w}}} + 4n^2\alpha^2 L_2^2\bar{L}^2 C_{\hat{\boldsymbol{w}}^*} + (2n^2\alpha^2 L_1^2\bar{L}^2 + nL_1^2)\big(\mathbb{E}[\|\Phi_0\|^2] + B_{GT}\big)\Big). \tag{79}$$

For induction, we need to enforce $\mathbb{E}[\|\bar{\boldsymbol{w}}_{k+1}\|^2] \leq C_{\bar{\boldsymbol{w}}}$, thus, from (79) we have

$$2c_1 + \frac{4}{\rho_U} F\Big(p(\bar{\boldsymbol{w}}_0)\Big) + \frac{4}{\rho_U^2}\Big(2n\alpha\bar{L}^2 d_w + 4n^2\alpha^2 L_2^2\bar{L}^2 C_{\bar{\boldsymbol{w}}} + 4n^2\alpha^2 L_2^2\bar{L}^2 C_{\hat{\boldsymbol{w}}^*} + (2n^2\alpha^2 L_1^2\bar{L}^2 + nL_1^2)\big(\mathbb{E}[\|\Phi_0\|^2] + B_{GT}\big)\Big) \leq C_{\bar{\boldsymbol{w}}}. \tag{80}$$

i.e.,

$$\left(1 - \frac{16n^2\alpha^2 L_2^2 \bar{L}^2}{\rho_U^2}\right) C_{\bar{\boldsymbol{w}}} \geq 2c_1 + \frac{4}{\rho_U} F\Big(p(\bar{\boldsymbol{w}}_0)\Big) + \frac{4}{\rho_U^2}\Big(2n\alpha\bar{L}^2 d_w + 4n^2\alpha^2 L_2^2 \bar{L}^2 C_{\hat{\boldsymbol{w}}^*} + (2n^2\alpha^2 L_1^2 \bar{L}^2 + nL_1^2)\big(\mathbb{E}[\|\Phi_0\|^2] + B_{GT}\big)\Big). \tag{81}$$

For $C_{\bar{\boldsymbol{w}}}$ to exits we need $\alpha < \frac{\rho_U}{4nL_2\bar{L}}$, and the value in (75) follows from (81). ■

Lemma 6 *For any $i \in \mathcal{V}$, let $Y_i \triangleq \left(\frac{M_i}{m_i}\sum_{j=1}^{m_1} L_{ij} + \frac{L_i'}{n}\right)$, then we have the following bounds.*

$$\mathbb{E}_{\mathbf{B}}[Y_i^2] \leq L_1^2, \qquad \forall\, i \in \mathcal{V}, \tag{82}$$

and

$$\mathbb{E}_{\mathbf{B}}\left[\sum_{i\in\mathcal{V}} Y_i^2\right] \leq L_2^2, \tag{83}$$

where L_1^2 and L_2^2 are defined in (89) and (90) respectively.

Proof : We have

$$\mathbb{E}_{\mathbf{B}}[Y_i] = \mathbb{E}_{\mathbf{B}}\left[\frac{M_i}{m_i}\sum_{j=1}^{m_i} L_{i,j}\right] + \frac{L_i'}{n}, \tag{84}$$

Now, it is obvious that for full gradient $\mathbb{E}_{\mathbf{B}}\left[\frac{M_i}{m_i}\sum_{j=1}^{m_i} L_{i,j}\right] = \sum_{j=1}^{M_i} L_{ij}$. In the case of stochastic gradient, let the set of data in the mini-batch of the i-th agent at k-th time step be represented by $\mathcal{B}_{i,k} \in \mathbf{B}_i \subset \mathbf{B}$ where $\mathbf{B}_i$ is the set of all possible mini-batches for the i-th agent. Then the total number of mini-batches possible for any $i \in \mathcal{V}$ is $|\mathbf{B}_i| \triangleq b_i = \binom{M_i}{m_i}$. Now, the probability of choosing any one of these is equal, i.e., $p(\mathcal{B}_{i,k}) = b_i^{-1}$. Next, the number of mini-batches that would contain $x_i^j \in \boldsymbol{X}_i$ is $\binom{M_i-1}{m_i-1}$. Thus, the term $L_{i,j}$ for any $j \in \{1, 2, \ldots, M_i\}$ will show up in $\binom{M_i-1}{m_i-1}$ of the mini-batches in $\mathbf{B}_i$. Noting that $\mathbf{B} = \bigcup_{i\in\mathcal{V}} \mathbf{B}_i$ and $\mathbf{B}_i \cap \mathbf{B}_i = \emptyset$ for $i \neq j$, we can write

$$\mathbb{E}_{\mathbf{B}}\left[\frac{M_i}{m_i}\sum_{j=1}^{m_i} L_{i,j}\right] = \mathbb{E}_{\mathbf{B}_i}\left[\frac{M_i}{m_i}\sum_{j=1}^{m_i} L_{i,j}\right], = \frac{M_i}{m_i} b_i^{-1}\binom{M_i-1}{m_i-1}\sum_{j=1}^{M_i} L_{i,j} = \sum_{j=1}^{M_i} L_{i,j}. \tag{85}$$

Thus, substituting (85) in (84) gives

$$\mathbb{E}_{\mathbf{B}}[Y_i] = L_i + \frac{L_i'}{n}, \tag{86}$$

where $L_i \triangleq \sum_{j=1}^{M_i} L_{i,j}$. Next, we have

$$\mathrm{Var}(Y_i) = \mathbb{E}_{\mathbf{B}}\left[\left(\frac{M_i}{m_i}\sum_{j=1}^{m_i} L_{i,j} + \frac{L_i'}{n} - \sum_{j=1}^{M_i} L_{i,j} - \frac{L_i'}{n}\right)^2\right], \tag{87}$$

$$= \mathbb{E}_{\mathbf{B}}\left[\left(\frac{M_i}{m_i}\sum_{j=1}^{m_i} L_{i,j} - \sum_{j=1}^{M_i} L_{i,j}\right)^2\right] = \sigma_{L,i}^2. \tag{88}$$

Note $\sigma_{L,i}^2$ sure to exist for all $i \in \mathcal{V}$ since the number of data points are fixed and finite, hence, $\mathbb{E}_{\mathbf{B}}\left[\left(\frac{M_i}{m_i}\sum_{j=1}^{m_i} L_{i,j} - \sum_{j=1}^{M_i} L_{i,j}\right)^2\right]$ is finite. Thus, $\mathbb{E}_{\mathbf{B}}[Y_i^2] = \mathbb{E}_{\mathbf{B}}[Y_i]^2 + \mathrm{Var}[Y_i] \leq \left(L_i + \frac{L_i'}{n}\right)^2 + \sigma_{L,i}^2$. Defining

$$L_1^2 \triangleq L^2 + \sigma_L^2, \tag{89}$$

where $L \triangleq \max_{i\in\mathcal{V}}\left\{L_i + \frac{L_i'}{n}\right\}$ and $\sigma_L^2 \triangleq \max_{i\in\mathcal{V}}\{\sigma_{L,i}^2\}$, we derive (82). Finally, using $\mathbb{E}_{\mathbf{B}}\left[\sum_{i\in\mathcal{V}} Y_i^2\right] = \sum_{i\in\mathcal{V}} \mathbb{E}_{\mathbf{B}}[Y_i^2]$ and denoting

$$L_2^2 \triangleq \sum_{i\in\mathcal{V}}\left(L_i + \frac{L_i'}{n}\right)^2 + \sum_{i\in\mathcal{V}} \sigma_{L,i}^2. \tag{90}$$

the result in (83) can be derived. ■