Asynchronous Bayesian Learning over a Network

Kinjal Bhar He Bai Jemin George Carl Busart

***Abstract*— We present a practical asynchronous data fusion model for networked agents to perform distributed Bayesian learning without sharing raw data. Our algorithm employs unadjusted Langevin dynamics with a gossip-based protocol for sampling, coupled with an event-triggered mechanism to further reduce communication between gossiping agents. These mechanisms drastically reduce communication overhead and help avoid bottlenecks commonly experienced with distributed algorithms. In addition, the algorithm is expected to increase resilience to occasional link failure. We establish mathematical guarantees for our algorithm and demonstrate its effectiveness via a numerical experiment.**

***Index Terms*— Distributed Bayesian learning, Unadjusted Langevin algorithm, Asynchronous Gossip protocol, Event-triggered mechanism, Multi-agent systems**

I. Introduction

Distributed learning in machine learning applications has gained much attention recently due to ubiquitous applications in sensor networks and multi-agent systems where the data is distributed at multiple computing nodes, yet a common model needs to be trained. Such situations arise when constrained by memory, inefficient data sharing means, or confidentiality requirements for sensitive data. Overfitting may occur when isolated agents train on their local data. Comprehensive parameter updates across isolated models also introduces inefficiencies for reaching threshold accuracy levels when compared to unconstrained information sharing. Distributed learning aims to leverage the full distributed data by a coordinated training among all the agents where the agents are allowed to share partial information (usually the learned model parameters or their gradients) without sharing any raw data. The information shared is significantly lower compared to sharing the raw data and does not compromise confidentiality.

In this paper, we focus on Bayesian inference techniques since they have been established as a reliable method for training machine learning models involving large datasets and a large number of trainable parameters. Additionally, since they are based on *sampling* from posterior distributions, they provide a built-in mechanism to quantify uncertainty. However, computing exact posteriors in most practical scenarios is analytically or computationally impossible. In this paper we employ Markov Chain Monte Carlo (MCMC) with an unadjusted Langevin algorithm (ULA) as the sampling method. Convergence of centralized Langevin method has been established for strongly log-concave posterior [1]–[4] and for non log-concave posterior [5]–[12]. Distributed [13]–[16] and federated [17], [18] formulations of various Bayesian based algorithms have been developed as well. However, most literature on distributed Bayesian learning deals with synchronized updates by all agents at any given time [13]–[16], which is not practical. Synchronized updates have immense communication overhead at every time instant and may be stymied due to lagging agents.

We seek to develop an algorithm to circumvent the aforementioned shortcomings. Motivated by optimization literature [19], [20], we introduce the concept of asynchronous gossip updates to the ULA. The gossip algorithm allows asynchronous updates where at any time only *two* agents make updates and share information. In addition to reduced communication overhead, it is more robust to occasional link failures since at most a single link is active at any time.

Furthermore, we incorporate an event-triggered information sharing scheme where information between the two active agents does not need to be exchanged unless some event is triggered, further mitigating the communication overhead issue. We present rigorous convergence proofs for the proposed algorithm. The results obtained in this paper are of practical relevance as they model the information exchange over a graph much more pragmatically. To make the updates truly asynchronous, we propose using a constant step size which does result in a bias in the convergence. However, bias exists even for centralized implementations [21]. Detailed discussion on how to minimize the bias in the convergence and one illustrative example supporting our results are provided.

The rest of the paper is organized as follows. We start with an introduction of the Bayesian learning framework and the ULA algorithm in Section II. In Section III, we introduce the key aspects of the gossip protocol and the event-triggering scheme followed by mathematical guarantees in Section IV. Section V provides further insight of our results, while we conclude with a numerical example in Section VI.

Notation: An $n \times n$ identity matrix is denoted as I_n. $\mathbf{1}_n$ denotes a n-dimensional vector of all ones and $\mathbf{e}_i$ is a n-dimensional vector with all 0s except the i-th element being 1. The L2-norm of a vector $\mathbf{x}$ is denoted as $\|\mathbf{x}\|_2$. Given matrices A and B, $A \otimes B$ denotes their Kronecker product. For a graph $\mathcal{G}(\mathcal{V}, \mathcal{E})$ of order n, $\mathcal{V} \triangleq \{v_1, \ldots, v_n\}$ represents the agents or nodes and the communication links between

K. Bhar and H. Bai are with Oklahoma State University, Stillwater, OK, 74078, USA. kbhar@okstate.edu, he.bai@okstate.edu

J. George and C. Busart are with the U.S. Army Research Laboratory, Adelphi, MD, 20783, USA. jemin.george.civ@army.mil, carl.e.busart.civ@army.mil

This work was partly supported by DEVCOM Army Research Laboratory (ARL) under Cooperative Agreement W911NF2120219. Views and conclusions contained in this document are those of the authors and should not be interpreted as representing the official policies, either expressed or implied, of ARL or the U.S. Government.

the agents are represented as $\mathcal{E} \triangleq \{\varepsilon_1, \ldots, \varepsilon_\ell\} \subseteq \mathcal{V} \times \mathcal{V}$. A Gaussian distribution with a mean $\mu \in \mathbb{R}^m$ and a covariance $\Sigma \in \mathbb{R}^{m\times m}_{\geq 0}$ is denoted by $\mathcal{N}(\mu, \Sigma)$.

II. Preliminaries

A. Bayesian inference framework

Consider a network of n agents characterized by an undirected communication graph $\mathcal{G}(\mathcal{V}, \mathcal{E})$ of order n. The entire data $\boldsymbol{X} = \{\boldsymbol{X}_i\}_{i=1}^n$ is distributed among n agents with the i-th agent having access only to its local dataset $\boldsymbol{X}_i = \{x_i^j\}_{j=1}^{M_i}$, where $x_i^j \in \mathbb{R}^d$.

Bayesian learning provides a framework for leaning unknown parameters by sampling from a posterior distribution. The probability of the unknown parameter $\boldsymbol{w}$ given the data $\boldsymbol{X}$, denoted by $p(\boldsymbol{w}|\boldsymbol{X})$, is the posterior distribution of interest. Assuming that the individual datasets of the agents are conditionally independent, the target posterior distribution $p^*(\boldsymbol{w}) \triangleq p(\boldsymbol{w}|\boldsymbol{X})$ is given by

$$p(\boldsymbol{w}|\boldsymbol{X}) \propto p(\boldsymbol{w}) \prod_{i=1}^n p(\boldsymbol{X}_i|\boldsymbol{w}) = \prod_{i=1}^n p(\boldsymbol{X}_i|\boldsymbol{w}) p(\boldsymbol{w})^{\frac{1}{n}}. \quad (1)$$

Thus, the objective of the inference problem is to determine p^*. As analytical solutions to p^* are often intractable, MCMC algorithms aim at sampling from p^*.

B. Sampling method

We use the unadjusted Langevin algorithm (ULA) which is a first order gradient method for sampling from p^*. Define an energy function $E(\boldsymbol{w}) = -\log(p(\boldsymbol{w}|\boldsymbol{X}))$. It follows from (1) that for some constant C,

$$E(\boldsymbol{w}, \boldsymbol{X}) = \sum_{i=1}^n E_i(\boldsymbol{w}, \boldsymbol{X}_i) + C, \quad (2)$$

where $E_i(\boldsymbol{w}) = -\log p(\boldsymbol{X}_i|\boldsymbol{w}) - \frac{1}{n}\log p(\boldsymbol{w})$. In the centralized sampling scenario, the ULA is given as

$$\boldsymbol{w}(k+1) = \boldsymbol{w}(k) - \alpha \nabla E(\boldsymbol{w}(k), \boldsymbol{X}) + \sqrt{2\alpha}\boldsymbol{v}(k), \quad (3)$$

where $\alpha > 0$ is the gradient step size, the gradient is given as $\nabla E = -\nabla \log p\left(\boldsymbol{X}|\boldsymbol{w}\right) - \nabla \log p(\boldsymbol{w})$, and $\boldsymbol{v}(k) \sim \mathcal{N}(\boldsymbol{0}_{d_w}, I_{d_w})$ is an injected Gaussian noise. A distributed version of (3) was introduced in [13] which is given by

$$\begin{aligned} \boldsymbol{w}_i(k+1) = \boldsymbol{w}_i(k) - \beta_k \sum_{j\in\mathcal{N}_i} (\boldsymbol{w}_i(k) - \boldsymbol{w}_j(k)) \\ - \alpha_k n \nabla E_i(\boldsymbol{w}_i(k), \boldsymbol{X}_i) + \sqrt{2\alpha_k}\boldsymbol{v}_i(k), \end{aligned} \quad (4)$$

where $\boldsymbol{w}_i(k)$ is the sample of the i-th agent, $\mathcal{N}_i$ denotes the set of neighbors of the i-th agent, α_k is the time-dependent gradient step size, β_k is a time-dependent fusion weight, the individual agent's gradients are given as $\nabla E_i = -\nabla \log p\left(\boldsymbol{X}_i|\boldsymbol{w}_i\right) - \frac{1}{n}\nabla \log p(\boldsymbol{w}_i)$, and $\boldsymbol{v}_i(k) \sim \mathcal{N}(\boldsymbol{0}_{d_w}, nI_{d_w})$.

III. Asynchronous Gossip with Event-Triggering

A. Gossip protocol

One of the major drawbacks of the algorithm in (4) is the communication overhead presented by the fusion term $\sum_{j\in\mathcal{N}_i}(\boldsymbol{w}_i(k) - \boldsymbol{w}_j(k))$. This necessitates communication between all the neighbors at all time instants in a synchronized fashion. We propose the asynchronous gossip protocol [20] which circumvents this issue by needing only two agents to update their samples at any given time instant.

Consider that each agent has local clock that ticks at a Poisson rate of 1 at the tick of which, it randomly chooses one of its neighbors and together they make updates. We assume that no two ticks of the local clocks of the agents coincide. For analysis, we consider a universal clock which ticks at a rate of n and is indexed by k. Suppose that the k-th tick of the universal clock coincides with the i_k-th agent's local clock, then agent i_k chooses agent j_k from $\mathcal{N}_{i_k}$ uniformly at random. The probability of agent i, $\forall\, i \in \{1, \cdots, n\}$, being active at the k-th tick of the universal clock is given by $p_i = \frac{1}{n}\left(1 + \sum_{j\in\mathcal{N}_i}\frac{1}{|\mathcal{N}_j|}\right)$. Note that p_i, $\forall i$, is time-invariant and depends on the graph only. Thus, it can be computed and stored by each agent a priori and subsequently used when needed.

Let $\mathcal{A}_k = \{i_k, j_k\}$ be the set of two agents activated at the k-th tick of the universal clock. Denote by $\tau_i(k)$ the number of times agent i has been active until the k-th tick of the universal clock. The update algorithm for the active agents, i.e., $i \in \mathcal{A}_k$, is given by

$$\begin{aligned} \boldsymbol{w}_i(\tau_i(k)+1) = \boldsymbol{w}_i(\tau_i(k)) - \beta \sum_{j\in\mathcal{A}_k}\Big(\boldsymbol{w}_i(\tau_i(k)) - \boldsymbol{w}_j(\tau_j(k))\Big) \\ - \frac{n\alpha}{2p_i}\nabla E_i(\boldsymbol{w}_i(\tau_i(k)), \boldsymbol{X}_i) + \sqrt{2\alpha}\boldsymbol{v}_i(\tau_i(k)), \end{aligned} \quad (5)$$

where α and β are constant gradient step size and fusion weight, respectively, $\nabla E_i = -\nabla \log p\left(\boldsymbol{X}_i|\boldsymbol{w}_i\right) - \frac{1}{n}\nabla \log p(\boldsymbol{w}_i)$, and $\boldsymbol{v}_i$ is the injected noise given by $\boldsymbol{v}_i \sim \mathcal{N}\left(\boldsymbol{0}_{d_w}, \frac{n^2}{2}I_{d_w}\right)$. Define $\delta_i(k)$ as the indicator function such that $\delta_i(k) = 1$ if $i \in \mathcal{A}_k$ and otherwise $\delta_i(k) = 0$. Thus, for agent i, $\forall\, i \in \{1, \cdots, n\}$, the gossip-based sampling protocol (5) can be represented in the universal clock index k as

$$\begin{aligned} \boldsymbol{w}_i(k+1) = \boldsymbol{w}_i(k) - \delta_i(k)\beta \sum_{j\in\mathcal{A}_k}(\boldsymbol{w}_i(k) - \boldsymbol{w}_j(k)) \\ - \delta_i(k)\frac{n\alpha}{2p_i}\nabla E_i(\boldsymbol{w}_i(k), \boldsymbol{X}_i) + \delta_i(k)\sqrt{2\alpha}\boldsymbol{v}_i(k). \end{aligned} \quad (6)$$

For any agent i, $\boldsymbol{w}_i(\tau_i(k)) = \boldsymbol{w}_i(k)$. For all the ticks of the universal clock between the $\tau_i(k)^{\text{th}}$ and the $(\tau_i(k)+1)^{\text{th}}$ ticks of the i-th agent's local clock, $\boldsymbol{w}_i(k)$ remains unchanged.

B. Event-triggering mechanism

We next introduce an event-triggering mechanism that further reduces the need to exchange samples at all the time instants between the active agents. Unless an agent is triggered, it does not communicate its sample to its gossiping neighbor and the neighbor proceeds with the last communicated sample of that agent. Denote by $\hat{\boldsymbol{w}}_i(k)$ the last communicated sample of i-th agent until the the kth tick of the universal clock. Agent i is triggered again to communicate $\boldsymbol{w}_i(k)$ if and only if $\delta_i(k) = 1$ and

$$\|\boldsymbol{e}_i(k)\|_2^2 = \|\boldsymbol{w}_i(k) - \hat{\boldsymbol{w}}_i(k)\|_2^2 > \epsilon_i(k). \quad (7)$$

Incorporating the event-triggering mechanism (7) into (6), we propose the following sampling algorithm for agent i, $\forall i$

$$\begin{aligned}\boldsymbol{w}_i(k+1) &= \boldsymbol{w}_i(k) - \delta_i(k)\beta \sum_{j\in\mathcal{A}_k} (\hat{\boldsymbol{w}}_i(k) - \hat{\boldsymbol{w}}_j(k)) \\ &- \delta_i(k)\frac{n\alpha}{2p_i}\nabla E_i(\boldsymbol{w}_i(k), \boldsymbol{X}_i) + \delta_i(k)\sqrt{2\alpha}\boldsymbol{v}_i(k).\end{aligned} \tag{8}$$

We choose the triggering threshold $\epsilon_i(k)$ as

$$\epsilon_i(k) = \frac{\mu_i^e}{(\tau_i(k)+1)^{\delta_i^e}} \leq \frac{\mu_e}{(k+1)^{\delta_e}}, \tag{9}$$

where $\delta_i^e, \mu_i^e > 0$ are agent-specific parameters independently chosen to control the event-triggering rate, while $\mu_e = \frac{1}{(2n)^{\delta_e}}\max_i\{\mu_i^e\} > 0$ and $\delta_e = \min_i\{\delta_i^e\} > 0$. The last inequality in (9) holds for sufficiently large k with probability 1 (see [22, Lemma 3]).

IV. Results

We present the key results of our analysis in this section.

A. Consensus and average dynamics

We define the following notation. $\mathbf{w}(k) = \left[\boldsymbol{w}_1(k)^\top, \ldots, \boldsymbol{w}_n(k)^\top\right]^\top$, $\mathbf{v}(k) = \left[\boldsymbol{v}_1(k)^\top, \ldots, \boldsymbol{v}_n(k)^\top\right]^\top$, $\mathbf{e}(k) = \left[\boldsymbol{e}_1(k)^\top, \ldots, \boldsymbol{e}_n(k)^\top\right]^\top$ and $\nabla\mathbf{E}(k) = \left[\nabla E_1(\boldsymbol{w}_1(k), \boldsymbol{X}_1)^\top, \ldots, \nabla E_n(\boldsymbol{w}_n(k), \boldsymbol{X}_n)^\top\right]^\top$.

We rewrite (8) in the vector form as

$$\begin{aligned}\mathbf{w}(k+1) =& \mathcal{W}_k\mathbf{w}(k) - \alpha n S_k\nabla\mathbf{E}(k) + \sqrt{2\alpha}S_k'\mathbf{v}(k) \\ &+ \beta(\mathcal{L}_k \otimes I_{d_w})\mathbf{e}(k),\end{aligned} \tag{10}$$

where $\mathcal{L}_k = (\mathbf{e}_{i_k} - \mathbf{e}_{j_k})(\mathbf{e}_{i_k} - \mathbf{e}_{j_k})^\top$, $\mathcal{W}_k = \left(I_n - \beta\mathcal{L}_k\right) \otimes I_{d_w}$, $S_k = \left(\frac{1}{2p_{i_k}}\mathbf{e}_{i_k}\mathbf{e}_{i_k}^\top + \frac{1}{2p_{j_k}}\mathbf{e}_{j_k}\mathbf{e}_{j_k}^\top\right) \otimes I_{d_w}$ and $S_k' = (\mathbf{e}_{i_k}\mathbf{e}_{i_k}^\top + \mathbf{e}_{j_k}\mathbf{e}_{j_k}^\top) \otimes I_{d_w}$. Let $\bar{\boldsymbol{w}}(k) = \frac{1}{n}\sum_{i=1}^n \boldsymbol{w}_i(k)$ and $\tilde{\boldsymbol{w}}_i(k) = \boldsymbol{w}_i(k) - \bar{\boldsymbol{w}}(k)$. Define the consensus error $\tilde{\mathbf{w}}(k) = \left[\tilde{\boldsymbol{w}}_1(k)^\top, \ldots, \tilde{\boldsymbol{w}}_n(k)^\top\right]^\top$ and note that $\tilde{\mathbf{w}}(k) = (M \otimes I_{d_w})\mathbf{w}(k)$ where $M = I_n - \frac{1}{n}\mathbf{1}_n\mathbf{1}_n^\top$. Pre-multiplying (10) with $(M \otimes I_{d_w})$ yields the evolution of the consensus dynamics:

$$\tilde{\mathbf{w}}(k+1) = \mathcal{W}_k\tilde{\mathbf{w}}(k) + (M \otimes I_{d_w})\mathbf{g}(k), \tag{11}$$

where $\mathbf{g}(k) = -\alpha n S_k\nabla\mathbf{E}(k) + \sqrt{2\alpha}S_k'\mathbf{v}(k) + \beta(\mathcal{L}_k \otimes I_{d_w})\mathbf{e}(k)$ and $(M \otimes I_{d_w})\mathcal{W}_k = \mathcal{W}_k(M \otimes I_{d_w})$.

Next, we derive the dynamics of the averaged sample $\bar{\boldsymbol{w}}(k)$ generated at each tick of the universal clock as

$$\bar{\boldsymbol{w}}(k+1) = \bar{\boldsymbol{w}}(k) - \alpha\widehat{\nabla E}(k) + \sqrt{2\alpha}\bar{\boldsymbol{v}}(k), \tag{12}$$

where $\widehat{\nabla E}(k) = \sum_{i\in\mathcal{A}_k}\nabla E_i(\boldsymbol{w}_i(k), \boldsymbol{X}_i)$ and $\bar{\boldsymbol{v}}(k) = \frac{1}{n}\sum_{i\in\mathcal{A}_k}\boldsymbol{v}_i(k) \sim \mathcal{N}(\mathbf{0}_{d_w}, I_{d_w})$. The $\widehat{\nabla E}(k)$ can be considered a stochastic gradient and is related to the full gradient $\nabla E(\bar{\boldsymbol{w}}(k)) = \sum_{i=1}^n \nabla E_i(\bar{\boldsymbol{w}}(k), \boldsymbol{X})$ by

$$\widehat{\nabla E}(k) = \nabla E(\bar{\boldsymbol{w}}(k)) - \xi(\bar{\boldsymbol{w}}(k), \mathcal{A}_k) + \zeta(\bar{\boldsymbol{w}}(k), \tilde{\mathbf{w}}(k), \mathcal{A}_k), \tag{13}$$

where

$$\xi(\bar{\boldsymbol{w}}(k), \mathcal{A}_k) = \nabla E(\bar{\boldsymbol{w}}(k)) - \sum_{i\in\mathcal{A}_k}\frac{1}{2p_i}\nabla E_i(\bar{\boldsymbol{w}}(k), \boldsymbol{X}_i), \tag{14}$$

$$\begin{aligned}\zeta(\bar{\boldsymbol{w}}(k), \tilde{\mathbf{w}}(k), \mathcal{A}_k) = \sum_{i\in\mathcal{A}_k}\frac{1}{2p_i}\Big(&\nabla E_i(\boldsymbol{w}_i(k), \boldsymbol{X}_i) \\ &- \nabla E_i(\bar{\boldsymbol{w}}(k), \boldsymbol{X}_i)\Big).\end{aligned} \tag{15}$$

The $\xi(k)$ represents the stochasticity from the gossip protocol while $\zeta(k)$ denotes the gradient noise due to consensus error. It follows that $\mathbb{E}_{p_t(\mathcal{A}_k)}[\xi(k)] = 0$.

B. Assumptions

***Assumption* 1.** *The gradients ∇E_i are Lipschitz continuous with Lipschitz constant $L_i > 0$ for all $i \in \{1, \ldots, n\}$, i.e., $\forall\, \boldsymbol{w}_a, \boldsymbol{w}_b \in \mathbb{R}^{d_w}$, we have*

$$\|\nabla E_i(\boldsymbol{w}_a, \boldsymbol{X}_i) - \nabla E_i(\boldsymbol{w}_b, \boldsymbol{X}_i)\|_2 \leq L_i\|\boldsymbol{w}_a - \boldsymbol{w}_b\|_2. \tag{16}$$

From (16) it follows that for $E(\boldsymbol{w}, \boldsymbol{X})$ in (2), there exists some $\bar{L} > 0$ such that $\forall\, \boldsymbol{w}_a, \boldsymbol{w}_b \in \mathbb{R}^{d_w}$ we have

$$\|\nabla E(\boldsymbol{w}_a, \boldsymbol{X}) - \nabla E(\boldsymbol{w}_b, \boldsymbol{X})\|_2 \leq \bar{L}\|\boldsymbol{w}_a - \boldsymbol{w}_b\|_2. \tag{17}$$

For the function $G(\mathbf{w}, \boldsymbol{X})$ defined as

$$G(\mathbf{w}, \boldsymbol{X}) = \sum_{i=1}^n \nabla E_i(\boldsymbol{w}_i, \boldsymbol{X}_{\boldsymbol{i}}), \tag{18}$$

where $\mathbf{w} = \left[\boldsymbol{w}_1^\top, \ldots, \boldsymbol{w}_n^\top\right]^\top$, we also conclude from (16) that there exists $L = \max_i\{L_i\} > 0$ such that $\forall\, \mathbf{w}_a, \mathbf{w}_b \in \mathbb{R}^{nd_w}$ we have

$$\|G(\mathbf{w}_a, \boldsymbol{X}) - G(\mathbf{w}_b, \boldsymbol{X})\|_2 \leq L\|\mathbf{w}_a - \mathbf{w}_b\|_2. \tag{19}$$

***Assumption* 2.** *The overall interaction topology of the n networked agents is given as a* connected, undirected *graph denoted by $\mathcal{G}(\mathcal{V}, \mathcal{E})$.*

For a connected undirected graph $\mathcal{G}(\mathcal{V}, \mathcal{E})$, the expected graph Laplacian, denoted by $\bar{\mathcal{L}} = \mathbb{E}[\mathcal{L}_k]$, is a positive semi-definite matrix with exactly one eigenvalue at 0 corresponding to the eigenvector $\mathbf{1}_n$.

***Assumption* 3.** *There exists some $0 < \mu_g < \infty$ such that for any $\boldsymbol{w} \in \mathbb{R}^{d_w}$, we have*

$$\sup_{i\in\{1,\ldots,n\}} \mathbb{E}[\|\nabla E_i(\boldsymbol{w}, \boldsymbol{X})\|_2] \leq \sqrt{\mu_g}. \tag{20}$$

Note that (20) can be equivalently represented as

$$\mathbb{E}[\|\nabla\mathbf{E}(\mathbf{w}, \boldsymbol{X})\|_2^2] \leq n\mu_g. \tag{21}$$

Assumption 3 has been used in many non-convex optimization references.

***Assumption* 4.** *The target distribution p^* satisfies a* log-Sobolev inequality *(LSI) defined as follows. For any smooth function g satisfying $\int g(\bar{\boldsymbol{w}})p^*(\bar{\boldsymbol{w}})\,d\bar{\boldsymbol{w}} = 1$, a constant $\rho_U > 0$ exists such that*

$$\int g(\bar{\boldsymbol{w}})\log g(\bar{\boldsymbol{w}})p^*(\bar{\boldsymbol{w}})\,d\bar{\boldsymbol{w}} \leq \frac{1}{2\rho_U}\int \frac{\|\nabla g(\bar{\boldsymbol{w}})\|^2}{g(\bar{\boldsymbol{w}})}p^*(\bar{\boldsymbol{w}})\,d\bar{\boldsymbol{w}}, \tag{22}$$

where ρ_U is the log-Sobolev constant.

***Assumption* 5.** *The second moment of the stochastic noise due to gossip in the average gradient ξ is bounded, i.e., for*

all $k \geq 0$ there exists some $0 < C_\xi < \infty$ such that

$$\mathbb{E}_{p_t(\mathcal{A}_k)}[\|\xi(\bar{\boldsymbol{w}}(k), \mathcal{A}_k)\|_2^2] \leq C_\xi. \tag{23}$$

***Condition* 1.** *The step size α is chosen to satisfy*

$$\frac{8\alpha^3 \bar{L}^4}{(1 - \exp(-\alpha\rho_U))} < \rho_U. \tag{24}$$

***Condition* 2.** *The fusion weight β is chosen to satisfy*

$$\beta(1-\beta) < \frac{1}{2\lambda_{n-1}(\bar{\mathcal{L}})}, \tag{25}$$

where $\lambda_{n-1}(\bar{\mathcal{L}})$ is the second smallest eigenvalue of $\bar{\mathcal{L}}$.

Note that the left hand side of (24) decreases monotonically with a decreasing α and approaches 0 as α approaches 0. Thus, given a ρ_U, there always exists an $\alpha^* > 0$ such that for any $\alpha \in (0, \alpha^*]$, (24) holds. Similarly, for given $\frac{1}{2\lambda_{n-1}(\bar{\mathcal{L}})}$ (constant for a graph) a $\beta^* > 0$ such that (25) holds for any $\beta \in (0, \beta^*]$.

C. Consensus analysis

Theorem 1 below shows that consensus is achieved at the rate of $\mathcal{O}\left(\frac{1}{k^{\delta_e}}\right)$ with an offset Y_3 given after (26).

Theorem 1. *Suppose that Assumptions 1–5 hold and that α and β satisfy Conditions 1 and 2, respectively. Define $\lambda = 1 - 2\beta(1-\beta)\lambda_{n-1}(\bar{\mathcal{L}})$ where $\lambda_i(\cdot)$ denotes the i-th largest eigenvalue of the positive semi-definite matrix $\bar{\mathcal{L}} = \mathbb{E}[\mathcal{L}_k]$. Then the consensus error $\tilde{\mathbf{w}}(k+1)$ satisfies*

$$\mathbb{E}[\|\tilde{\mathbf{w}}(k+1)\|_2^2] \leq Y_1 \sqrt{\lambda}^{k+1} + \frac{Y_2}{(k+1)^{\delta_e}} + Y_3, \tag{26}$$

where $Y_1 = \mathbb{E}[\|\tilde{\mathbf{w}}(0)\|_2^2] + \frac{2\beta^2 n \mu_e}{1-\sqrt{\lambda}} \sum_{t=0}^{\bar{t}-1} \frac{\sqrt{\lambda}^{-(t+1)}}{(t+1)^{\delta_e}}$, $Y_2 = -\frac{2\beta^2 n \mu_e}{\sqrt{\lambda}(1-\sqrt{\lambda})} \left(\ln\sqrt{\lambda} + \frac{\delta_e}{\bar{t}+1}\right)^{-1}$, $Y_3 = \frac{2\alpha n^2(\alpha\mu_g/2p_m + 2d_w)}{(1-\sqrt{\lambda})^2}$, $\bar{t} = \max\left\{0, \left\lceil \frac{\delta_e}{|\ln\lambda|} - 1 \right\rceil\right\}$ *and* $p_m = \min_i\{p_i\}$.

Proof. To analyze the consensus error, we start with the consensus dynamics in (11) and take the norm on both sides, yielding

$$\|\tilde{\mathbf{w}}(k+1)\|_2 \leq \|\mathcal{W}_k \tilde{\mathbf{w}}(k)\|_2 + \|\mathbf{g}(k)\|_2, \tag{27}$$

where we used the result $\|(M \otimes I_{d_w})\mathbf{g}(k)\|_2 \leq \|M \otimes I_{d_w}\|_2 \|\mathbf{g}(k)\|_2 = \|\mathbf{g}(k)\|_2$ since $\|M \otimes I_{d_w}\|_2 = 1$. Denoting by $\mathcal{F}_k$ be the filtration generated by randomized sampling of $\{\mathbf{w}(\ell)\}_{\ell=0}^k$, it can be shown that the conditional expectation $\mathbb{E}[\|\mathcal{W}_k \tilde{\mathbf{w}}(k)\|_2^2 | \mathcal{F}_k]$ follows the relation below:

$$\mathbb{E}[\|\mathcal{W}_k \tilde{\mathbf{w}}(k)\|_2^2 | \mathcal{F}_k] = \tilde{\mathbf{w}}(k)^\top \mathbb{E}[\mathcal{W}_k^\top \mathcal{W}_k] \tilde{\mathbf{w}}(k) \leq \lambda \|\tilde{\mathbf{w}}(k)\|_2^2. \tag{28}$$

Note from (25) that $0 < \lambda$. It also follows from $2\beta(1-\beta)\lambda_{n-1}(\bar{\mathcal{L}}) > 0$ that $\lambda < 1$. The conditional expectation $\mathbb{E}[\|\mathbf{g}(k)\|_2^2 | \mathcal{F}_k]$ can be shown to satisfy

$$\mathbb{E}[\|\mathbf{g}(k)\|_2^2 | \mathcal{F}_k] \leq 2\left(\frac{\alpha^2 n^2 \mu_g}{2p_m} + 2\alpha n^2 d_w\right) + \frac{2\beta^2 n \mu_e}{(k+1)^{\delta_e}}. \tag{29}$$

Recall the identity $(x+y)^2 \leq (\theta+1)x^2 + \left(\frac{\theta+1}{\theta}\right) y^2$ for any $x, y, \theta \in \mathbb{R}$ and $\theta > 0$. We use this identity with $\theta = \sqrt{\lambda}^{-1} - 1 > 0$ on (27), subsequently take the conditional expectation $\mathbb{E}[\cdot|\mathcal{F}_k]$, and substitute (28) and (29). Further, taking the total expectation yields

$$\mathbb{E}[\|\tilde{\mathbf{w}}(k+1)\|_2^2] \leq \sqrt{\lambda}\mathbb{E}[\|\tilde{\mathbf{w}}(k)\|_2^2] + \frac{1}{1-\sqrt{\lambda}} \frac{2b^2 n \mu_e}{(k+1)^{\delta_e}} + \frac{2\alpha n^2(\alpha\mu_g/2p_m + 2d_w)}{1-\sqrt{\lambda}}. \tag{30}$$

Finally, using (30) iteratively with some additional algebra results in the consensus error bound (26). □

D. Convergence analysis

We denote by $p(\bar{\boldsymbol{w}}(k))$ the probability distribution of $\bar{\boldsymbol{w}}(k)$ admitted by the average dynamics (12) and analyze its evolution. To do so, we first reformulate (12) as a stochastic differential equation (SDE). For any $t \in [t_k, t_{k+1})$ where $t_k = k\alpha$, the SDE form of (12) is given by

$$\begin{aligned} d\bar{\boldsymbol{w}}(t) = -\Big(&\nabla E(\bar{\boldsymbol{w}}(t_k)) - \xi(\bar{\boldsymbol{w}}(t_k), \mathcal{A}_k) \\ &+ \zeta(\bar{\boldsymbol{w}}(t_k), \tilde{\mathbf{w}}(t_k), \mathcal{A}_k)\Big)dt + \sqrt{2} d\boldsymbol{B}(t), \end{aligned} \tag{31}$$

where $\boldsymbol{B}(t)$ represents a d_w dimensional Brownian motion, $\bar{\boldsymbol{w}}(t_k) = \bar{\boldsymbol{w}}(k)$, and $\tilde{\mathbf{w}}(t_k) = \tilde{\mathbf{w}}(k)$. Denote by $p_t(\bar{\boldsymbol{w}})$ the distribution of $\bar{\boldsymbol{w}}(t)$ from (31). Since the gradient terms in (31) remain constant within $t \in [t_k, t_{k+1})$, $p_{t_{k+1}}(\bar{\boldsymbol{w}})$ is the same as $p(\bar{\boldsymbol{w}}(k+1))$ from (12), $\forall k \geq 0$. Thus, we analyze the evolution of $p_t(\bar{\boldsymbol{w}})$ from (31). Let $y_{k,1} = \bar{\boldsymbol{w}}(t_k)$, $y_{k,2} = \tilde{\mathbf{w}}(t_k)$, $y_{k,3} = \mathcal{A}_k$, and $y_k = [y_{k,1}^\top, y_{k,2}^\top, y_{k,3}^\top]^\top$. Using the Fokker Planck (FP) equation for the SDE in (31) we have

$$\begin{aligned} \frac{\partial p_t(\bar{\boldsymbol{w}}|y_k)}{\partial t} = -\nabla \cdot \Big[p_t(\bar{\boldsymbol{w}}|y_k)\Big(&- \nabla E(y_{k,1}) + \xi(y_{k,1}, y_{k,3}) \\ &- \zeta(y_k)\Big)\Big] + \nabla^2 p_t(\bar{\boldsymbol{w}}|y_k). \end{aligned} \tag{32}$$

Marginalizing out y_k from (32), we get the evolution of $p_t(\bar{\boldsymbol{w}})$ for $t \in [t_k, t_{k+1})$ corresponding to any $k \geq 0$ as

$$\begin{aligned} \frac{\partial p_t(\bar{\boldsymbol{w}})}{\partial t} = \nabla \cdot \Bigg[\iint \sum_{y_{k,3} \in \mathbf{A}} p_t(\bar{\boldsymbol{w}}|y_k) \Big(&\nabla E(\bar{\boldsymbol{w}}(t_k)) + \zeta(y_k) \\ &- \xi(y_{k,1}, y_{k,3})\Big) p(y_k) dy_{k,1}\, dy_{k,2} \Bigg] + \nabla^2 p_t(\bar{\boldsymbol{w}}), \end{aligned} \tag{33}$$

where $\mathbf{A}$ is the finite set of all possible values of $y_{k,3} = \mathcal{A}_k$, i.e., the set of all possible gossiping partners at any time instant of the universal clock.

We next employ the KL divergence between the probability distribution $p_t(\bar{\boldsymbol{w}})$ and the target distribution $p^*(\bar{\boldsymbol{w}})$, denoted by $F(p_t(\bar{\boldsymbol{w}}))$, to prove convergence of the posterior of $\bar{\boldsymbol{w}}$ in (12). Specifically, $F(p_t(\bar{\boldsymbol{w}}))$ is defined as

$$F(p_t(\bar{\boldsymbol{w}})) = \int p_t(\bar{\boldsymbol{w}}) \log\left(\frac{p_t(\bar{\boldsymbol{w}})}{p^*(\bar{\boldsymbol{w}})}\right) d\bar{\boldsymbol{w}}. \tag{34}$$

Theorem 2 below establishes that $F(p_t(\bar{\boldsymbol{w}}))$ decreases at the rate of $\mathcal{O}\left(\frac{1}{k^{\delta_e}}\right)$ to a bias B given in (40). The proof makes use of (33) and the LSI (22) to obtain $\dot{F}(p_t(\bar{\boldsymbol{w}}))$ and subsequently bound $F(p_t(\bar{\boldsymbol{w}}))$.

Theorem 2. *Suppose that all the assumptions and conditions in Theorem 1 hold. Then*

1) If $\alpha\rho_U + \ln\sqrt{\lambda} < 0$, then

$$F(p_{t_{k+1}}(\bar{\boldsymbol{w}})) \leq \exp\big(-\alpha\rho_U(k+1)\big)\big)F(p_{t_0}(\bar{\boldsymbol{w}})) + \bar{Y}_1^{'}\exp(-\alpha\rho_U k) + \frac{\bar{Y}_2^{'}}{(k+1)^{\delta_e}} + B, \quad (35)$$

2) if $\alpha\rho_U + \ln\sqrt{\lambda} > 0$*, then*

$$F(p_{t_{k+1}}(\bar{\boldsymbol{w}})) \leq \exp\big(-\alpha\rho_U(k+1)\big)\big)F(p_{t_0}(\bar{\boldsymbol{w}})) + \bar{Y}_1^{''}\sqrt{\lambda}^{k+1} + \frac{\bar{Y}_2^{'}}{(k+1)^{\delta_e}} + B, \quad (36)$$

where $\bar{Y}_1^{'}$*,* $\bar{Y}_1^{''}$*,* $\bar{Y}_2^{'}$ *and* B *are positive constants given by*

$$\bar{Y}_1^{'} = \left(1 - \frac{1}{\alpha\rho_U + \ln\sqrt{\lambda}}\right)\left(\frac{\alpha^3\bar{L}^2L^2}{p_m} + \frac{\alpha L^2}{4p_m}\right)Y_1, \quad (37)$$

$$\bar{Y}_1^{''} = \frac{\sqrt{\lambda}^{-k+1}}{\alpha\rho_U + \ln\sqrt{\lambda}}\left(\frac{\alpha^3\bar{L}^2L^2}{p_m} + \frac{\alpha L^2}{4p_m}\right)Y_1, \quad (38)$$

$$\bar{Y}_2^{'} = \left(\alpha\rho_U - \frac{\delta_e}{\bar{k}_2}\right)^{-1}\left(\frac{\alpha^3\bar{L}^2L^2}{p_m} + \frac{\alpha L^2}{4p_m}\right)Y_2, \quad (39)$$

$$B = \frac{\nu}{1-\exp(-\alpha\rho_U)}, \quad (40)$$

in which $\nu = 2\alpha^2\bar{L}^2d_w + 2\alpha^3\bar{L}^4C_{\bar{\boldsymbol{w}}} + 4\alpha^3\bar{L}^2C_{\xi} + \left(\frac{\alpha^3\bar{L}^2L^2}{p_m} + \frac{\alpha L^2}{4p_m}\right)Y_3$ *and* Y_3 *is given after* (26)*.*

Proof. From (34) the evolution of $F(p_t(\bar{\boldsymbol{w}}))$ is related to $\frac{\partial p_t(\bar{\boldsymbol{w}})}{\partial t}$ by

$$\dot{F}(p_t(\bar{\boldsymbol{w}})) = \int\left(1 + \nabla\log\left(\frac{p_t(\bar{\boldsymbol{w}})}{p^*(\bar{\boldsymbol{w}})}\right)\right)\frac{\partial p_t(\bar{\boldsymbol{w}})}{\partial t}d\bar{\boldsymbol{w}}. \quad (41)$$

Substituting (33) into (41) and performing all the appropriate marginalization yield

$$\dot{F}(p_t(\bar{\boldsymbol{w}})) \leq -\frac{1}{2}\mathbb{E}_{p_t(\bar{\boldsymbol{w}})}\left\|\nabla\log\frac{p_t(\bar{\boldsymbol{w}})}{p^*(\bar{\boldsymbol{w}})}\right\|_2^2 + 2\alpha\bar{L}^2d_w + 2\alpha^2\times \bar{L}^2(\bar{L}^2C_{\bar{\boldsymbol{w}}} + 2C_{\xi}) + \left(\frac{\alpha^2\bar{L}^2L^2}{p_m} + \frac{L^2}{4p_m}\right)\mathbb{E}_{p_{t_k}(\tilde{\mathbf{w}})}\|\tilde{\mathbf{w}}(t_k)\|_2^2, \quad (42)$$

where $\mathbb{E}_{p_t(\xi)}[\|\xi(\bar{\boldsymbol{w}}(t_k)\tilde{\mathbf{w}}(t_k), \mathcal{A}_k)\|_2^2] \leq C_{\xi}$ and $\mathbb{E}_{p_t(\bar{\boldsymbol{w}})}[\|\bar{\boldsymbol{w}}(t_k)\|_2^2] \leq C_{\bar{\boldsymbol{w}}}$. Note that the existence of $C_{\bar{\boldsymbol{w}}}$ can be explicitly proven, which is skipped due to space constraints.. Thereafter, we employ the LSI (22) with $g(\bar{\boldsymbol{w}}) = \frac{p_t(\bar{\boldsymbol{w}})}{p^*(\bar{\boldsymbol{w}})}$ to obtain

$$\begin{aligned} F(p_t(\bar{\boldsymbol{w}})) &= \mathbb{E}_{p_t(\bar{\boldsymbol{w}})}\left[\log\left(\frac{p_t(\bar{\boldsymbol{w}})}{p^*(\bar{\boldsymbol{w}})}\right)\right] \\ &\leq \frac{1}{2\rho_U}\mathbb{E}_{p_t(\bar{\boldsymbol{w}})}\left\|\nabla\log\left(\frac{p_t(\bar{\boldsymbol{w}})}{p^*(\bar{\boldsymbol{w}})}\right)\right\|_2^2, \end{aligned} \quad (43)$$

which when substituted in (42) gives a recursive relation in $F(p_t(\bar{\boldsymbol{w}}))$ for any $t \in [t_k, t_{k+1})$ as follows:

$$\dot{F}(p_t(\bar{\boldsymbol{w}})) \leq -\rho_U F(p_t(\bar{\boldsymbol{w}})) + 2\alpha\bar{L}^2d_w + 2\alpha^2\bar{L}^4C_{\bar{\boldsymbol{w}}} + 4\alpha^2\bar{L}^2C_{\xi} + \left(\frac{\alpha^2\bar{L}^2L^2}{p_m} + \frac{L^2}{4p_m}\right)\mathbb{E}_{p_{t_k}(\tilde{\mathbf{w}})}\|\tilde{\mathbf{w}}(t_k)\|_2^2. \quad (44)$$

Conducting further analysis on (44), we obtain the convergence rate for (12) in two cases depending on the sign of $\alpha\rho_U + \ln\sqrt{\lambda}$ (note that $\ln\sqrt{\lambda} < 0$ since $\lambda < 1$), which are shown in (35) and (36), respectively. □

V. Discussions

In this section, we highlight some key insights in our results. Firstly, from (26) we get the rate of consensus as $\mathcal{O}\left(\frac{1}{k^{\delta_e}}\right)$. with a constant offset Y_3 given after (26) in the asymptotic consensus error. This results from the usage of a constant gradient step size α. To keep Y_3 low, we may choose the step size α to be scaled as $\alpha \propto \frac{1}{n^2d_w}$. Also, increasing p_m reduces the Y_3 as higher p_m implies less randomness in the gossip.

Secondly, we conclude from (35) and (36) that in either case the rate of convergence is $\mathcal{O}\left(\frac{1}{k^{\delta_e}}\right)$ as well. It is tempting to conclude that a high value of δ_e is preferable since it fosters both consensus and convergence rate. However, a high δ_e value results in a quicker decay of the error threshold in (9), leading to increased communication overhead as k increases. Thus δ_e is an important hyperpaprameter trading off the rate of convergence against the communication overhead.

As observed from either (35) or (36) that, there is a constant bias B in the KL divergence bound. From (40), the most obvious dependence of B is on the step size α. For a sufficiently small α, $1-\exp(-\alpha\rho_U) \approx \alpha\rho_U$. Since the least power of α in any of the terms in ν is 2, we have $\nu = \alpha^2\bar{\nu}$. Hence, $B \approx \frac{\alpha\bar{\nu}}{\rho_U}$. Thus, lowering α is likely to reduce B, however, it may also compromise the rate of convergence. Furthermore, B linearly decreases with the reduction in C_{ξ} (variance of the stochasticity of gossip), $C_{\bar{\boldsymbol{w}}}$ (variance of the average of samples) and d_w (dimension of the samples). In addition, $B \propto \frac{n^2}{p_m}$, implying that reducing n (the number of agents) and increasing p_m (the least probability of any agent being active) reduces the bias. This is intuitive as reducing n or increasing p_m lowers the uncertainty in the random selection of gossiping agents which translates to a lower bias.

Finally, note that the last term of ν (given below (40)) contains the consensus error offset Y_3 while the other terms of ν are due to the variance from different sources (injected noise, gossip stochasticity, and average of samples).

VI. Numerical Experiments

A. Gaussian mixture

We consider parameter inference of a Gaussian mixture with tied means [23]. The Gaussian mixture is given by

$$\theta_1 \sim \mathcal{N}(0, \sigma_1^2) \quad ; \quad \theta_2 \sim \mathcal{N}(0, \sigma_2^2) \quad (45)$$

$$x_i \sim \frac{1}{2}\mathcal{N}(\theta_1, \sigma_x^2) + \frac{1}{2}\mathcal{N}(\theta_1 + \theta_2, \sigma_x^2), \quad (46)$$

where $\sigma_1^2 = 10$, $\sigma_2^2 = 1$, $\sigma_x^2 = 2$ and $\boldsymbol{w} \triangleq [\theta_1, \theta_2]^\top \in \mathbb{R}^2$. We draw 100 data samples x_i from the model with $[\theta_1, \theta_2] = [0, 1]$. These data points were equally randomly distributed among 5 agents. The communication topology between the agents is a ring graph.

Simulation results with 1 chain for $(100000 \times n)$ iterations is presented with: $\alpha = 1 \times 10^{-4}$, $\beta = 0.1$, $\mu_e = 8$ and $\delta_e = 0.51$. The samples from the gossip event-triggered algorithm (8) are compared with an approximated true posterior distribution in Figure 1. Wasserstein distances between the sampled posterior from the approximated posterior, calculated using [24], are presented as a metric of accuracy.

The average frequency of gossiping and event-triggering for each agent is listed in Table I which suggests that an average (over all agents) of 60% reduction in activity is achieved due to the gossiping protocol, while communication is reduced by more than 80% due to event-triggering. Note that the percentage reduction in communication due to event-triggering is computed based on the number of times each agent has been active.

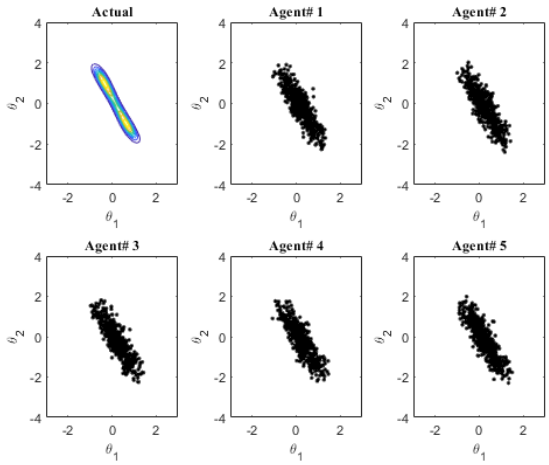

Fig. 1: Comparison of the posteriors constructed by the 5 agents with the actual approximate posterior. The Wassersterin distances between the agents' posteriors with the approximate true posterior are: 0.1089, 0.0942, 0.0964, 0.0963, and 0.1073, respectively.

$Agent\#$	1	2	3	4	5
gos	199481	200460	200817	199912	199328
$\%\, gos$	39.9%	40.1%	40.2%	40.0%	39.9%
ET	33735	33646	33695	33457	33477
$\%\, ET$	16.9%	16.8%	16.8%	16.7%	16.8%

TABLE I: Details about the frequency of gossiping and event-triggering averaged over all 5 trials for all agents. ($gos \equiv$ number of times the agents have been active among the total 500000 iterations, $\%\, gos \equiv$ gossip as a fraction of the total iterations; $ET \equiv$ number of times the agents have exchanged their samples, $\%\, ET \equiv$ fraction of triggers out of the total number of times each agent had been active).

VII. Conclusions and Future Work

In this paper, we propose an asynchronous, distributed ULA algorithm for Bayesian learning via an event-triggered gossip communication. We derive rigorous convergence guarantees for the proposed algorithm and illustrate its effectiveness using a numerical experiment. Though we obtain good empirical results, our mathematical analysis shows asymptotic bias in the convergence which stems from the use of a constant step size. Our future work involves the analysis of gossip algorithms with diminishing step sizes and other asynchronous algorithms for distributed Bayesian learning.

References

[1] A. Dalalyan, "Further and stronger analogy between sampling and optimization: Langevin monte carlo and gradient descent," in *Conference on Learning Theory*. PMLR, 2017, pp. 678–689.

[2] X. Cheng, N. S. Chatterji, P. L. Bartlett, and M. I. Jordan, "Underdamped langevin mcmc: A non-asymptotic analysis," in *Conference on learning theory*. PMLR, 2018, pp. 300–323.

[3] X. Cheng and P. Bartlett, "Convergence of langevin mcmc in kl-divergence," in *Algorithmic Learning Theory*. PMLR, 2018, pp. 186–211.

[4] A. Durmus and E. Moulines, "High-dimensional bayesian inference via the unadjusted langevin algorithm," *Bernoulli*, vol. 25, no. 4A, pp. 2854–2882, 2019.

[5] M. Raginsky, A. Rakhlin, and M. Telgarsky, "Non-convex learning via stochastic gradient langevin dynamics: a nonasymptotic analysis," in *Conference on Learning Theory*. PMLR, 2017, pp. 1674–1703.

[6] P. Xu, J. Chen, D. Zou, and Q. Gu, "Global convergence of langevin dynamics based algorithms for nonconvex optimization," *Advances in Neural Information Processing Systems*, vol. 31, 2018.

[7] Y. Zhang, Ö. D. Akyildiz, T. Damoulas, and S. Sabanis, "Nonasymptotic estimates for stochastic gradient langevin dynamics under local conditions in nonconvex optimization," *arXiv preprint arXiv:1910.02008*, 2019.

[8] N. H. Chau, É. Moulines, M. Rásonyi, S. Sabanis, and Y. Zhang, "On stochastic gradient langevin dynamics with dependent data streams: The fully nonconvex case," *SIAM Journal on Mathematics of Data Science*, vol. 3, no. 3, pp. 959–986, 2021.

[9] W. Mou, N. Flammarion, M. J. Wainwright, and P. L. Bartlett, "Improved bounds for discretization of langevin diffusions: Near-optimal rates without convexity," *arXiv preprint arXiv:1907.11331*, 2019.

[10] M. B. Majka, A. Mijatović, and Ł. Szpruch, "Nonasymptotic bounds for sampling algorithms without log-concavity," *The Annals of Applied Probability*, vol. 30, no. 4, pp. 1534–1581, 2020.

[11] Y.-A. Ma, Y. Chen, C. Jin, N. Flammarion, and M. I. Jordan, "Sampling can be faster than optimization," *Proceedings of the National Academy of Sciences*, vol. 116, no. 42, pp. 20 881–20 885, 2019.

[12] X. Cheng, N. S. Chatterji, Y. Abbasi-Yadkori, P. L. Bartlett, and M. I. Jordan, "Sharp convergence rates for langevin dynamics in the nonconvex setting," *arXiv preprint arXiv:1805.01648*, 2018.

[13] A. Parayil, H. Bai, J. George, and P. Gurram, "Decentralized langevin dynamics for bayesian learning," *Advances in Neural Information Processing Systems*, vol. 33, pp. 15 978–15 989, 2020.

[14] V. Kungurtsev, A. Cobb, T. Javidi, and B. Jalaian, "Decentralized bayesian learning with metropolis-adjusted hamiltonian monte carlo," *arXiv preprint arXiv:2107.07211*, 2021.

[15] A. Kolesov and V. Kungurtsev, "Decentralized langevin dynamics over a directed graph," *arXiv preprint arXiv:2103.05444*, 2021.

[16] M. Gürbüzbalaban, X. Gao, Y. Hu, and L. Zhu, "Decentralized stochastic gradient langevin dynamics and hamiltonian monte carlo," *Journal of Machine Learning Research*, vol. 22, no. 239, pp. 1–69, 2021.

[17] S. Lee, C. Park, S.-N. Hong, Y. C. Eldar, and N. Lee, "Bayesian federated learning over wireless networks," *arXiv preprint arXiv:2012.15486*, 2020.

[18] K. El Mekkaoui, D. Mesquita, P. Blomstedt, and S. Kaski, "Federated stochastic gradient langevin dynamics," in *Uncertainty in Artificial Intelligence*. PMLR, 2021, pp. 1703–1712.

[19] S. Boyd, A. Ghosh, B. Prabhakar, and D. Shah, "Gossip algorithms: Design, analysis and applications," in *Proceedings IEEE 24th Annual Joint Conference of the IEEE Computer and Communications Societies.*, vol. 3. IEEE, 2005, pp. 1653–1664.

[20] S. S. Ram, A. Nedić, and V. V. Veeravalli, "Asynchronous gossip algorithms for stochastic optimization," in *Proceedings of the 48h IEEE Conference on Decision and Control (CDC) held jointly with 2009 28th Chinese Control Conference*. IEEE, 2009, pp. 3581–3586.

[21] A. Wibisono, "Sampling as optimization in the space of measures: The langevin dynamics as a composite optimization problem," in *Conference on Learning Theory*. PMLR, 2018, pp. 2093–3027.

[22] A. Nedic, "Asynchronous broadcast-based convex optimization over a network," *IEEE Transactions on Automatic Control*, vol. 56, no. 6, pp. 1337–1351, 2010.

[23] M. Welling and Y. W. Teh, "Bayesian learning via stochastic gradient langevin dynamics," in *Proceedings of the 28th international conference on machine learning (ICML-11)*. Citeseer, 2011, pp. 681–688.

[24] Aslansefat, Koorosh, "Ecdf-based distance measure algorithms," 2020, available at: https://github.com/koo-ec/ECDF-based-Distance-Measure, Retrieved April 29, 2020.