Distributed Event-Triggered Unadjusted Langevin Algorithm for Bayesian Learning

K. Bhar [a], H. Bai [a], J. George [b], C. Busart [b]

[a] *Oklahoma State University, Stillwater, OK 74078, USA.*

[b] *U.S. Army Research Laboratory, Adelphi, MD 20783, USA.*

Abstract

This paper presents a distributed event-triggered unadjusted Langevin algorithm (DETULA) to address the Bayesian learning problem. We consider a set of networked learning agents who have access to their own independently distributed data sets. The objective of each agent is to reconstruct the global posterior of the unknown model parameters through local learning along with local interaction with neighboring agents. We propose an event-triggered communication mechanism for a distributed Langevin algorithm to limit the inter-agent interactions and thus reduce the communication overhead. We provide conditions on the algorithm step sizes and the triggering threshold to ensure mean-square consensus of the agents' parameter estimates and convergence of the estimates to the global posterior as if the data sets were aggregated at a central location. A major improvement of our result over previous studies is the establishment of said consensus without imposing any bounded restriction of the gradient of the objective function. Additionally, we establish probabilistic guarantees to prevent consecutive triggering by any agent while maintaining the same rate of convergence as in the case without event-triggering. We demonstrate the DETULA using distributed supervised-learning problems. Our results indicate that the agents successfully recover the global posterior by periodically sharing their samples with neighbors.

Key words: Bayesian learning, Langevin dynamics, distributed multi-agent networks, event-triggered communication

1 Introduction

Bayesian learning is a principled approach to estimating model parameters in machine learning problems. Motivations for pursuing Bayesian learning are to avoid overfitting and to provide an uncertainty measure of the model parameter estimates [1]. Compared with Maximum A Posteriori and Maximum Likelihood estimation, Bayesian approach captures the epistemic uncertainties and estimates the posterior distribution of model parameters given a data set and a prior distribution.

An analytical solution to the exact posterior is often intractable. Therefore, one needs to resort to numerical methods such as Markov Chain Monte Carlo (MCMC) methods. MCMC methods approximate the posterior distribution using samples from a Markov chain with the posterior being the equilibrium distribution [1]. A class of gradient based MCMC schemes, known as Hamiltonian Monte Carlo (HMC), is derived from Hamiltonian dynamics [37]. Examples of HMC methods include Langevin algorithms [4, 6, 8, 9, 13–15] and higher-order HMC algorithms [29, 30, 33, 34, 36]. Stochastic gradient-based Langevin algorithm [47] and its variations have been widely used for large scale data applications.

The aforementioned MCMC approaches are applicable when all available training data is accessible to a single centralized entity. However in distributed situations where dispersed networked agents cannot share training data,

Email addresses: kbhar@okstate.edu (K. Bhar), he.bai@okstate.edu (H. Bai), jemin.george.civ@mail.mil (J. George), carl.e.busart.civ@mail.mil (C. Busart).

distributed Bayesian learning algorithms must be used to fuse information from the distributed data sets. Parallel MCMC methods [18,20,21,49] have been proposed to tackle the distributed data problem in a master-slave communication architecture, where a central node aggregates samples from Markov chains on individual computing nodes. Alternatively, distributed Bayesian learning schemes [19,25,26,39] generalize the communication topology considered in distributed and parallel MCMC schemes to any connected undirected graph.

The existing distributed Bayesian learning algorithms require constant communication between the agents. In this paper, we demonstrate that such constant communication can be reduced by introducing an *event-triggering* mechanism for inter-agent communication. The idea of event-triggered communication has been widely used in distributed control applications [10, 17, 20] and in distributed optimization [11, 24, 51]. These papers focus on continuous-time, deterministic optimization algorithms for convex problems. More recently, event-driven stochastic gradient descent algorithms for non-convex problems have been considered in [16, 43]. To reduce the communication overhead associated with the existing distributed Bayesian learning schemes [19, 25, 39], we propose a distributed event-triggered unadjusted Langevin algorithm (DETULA). The proposed algorithm does not require inter-agent communication at every iteration. Instead, it enables each agent to determine whether communication is necessary at each iteration. The key idea behind the triggering mechanism is to communicate only if the difference between the previously communicated estimate and the current estimate is above a triggering threshold. The DETULA employs diminishing sequences for the triggering threshold and for the step sizes of its gradient and consensus terms. Assuming that the target posterior distribution satisfies a log-Sobolev inequality (LSI) [18], we provide conditions on the triggering threshold and the step sizes that guarantee consensus in mean-square and convergence to the target posterior. Empirical results from a Gaussian mixture example and a logistic regression example show that the DETULA performance is similar to centrally trained models with significantly reduced inter-agent communication compared to existing distributed Bayesian learning algorithms.

Related literature: There has been active research on analyzing convergence properties of the unadjusted Langevin algorithm (ULA). The unadjusted Langevin algorithm (ULA) is obtained by discretizing the continuous time Langenvin dynamics [28] and ignoring any rejection-acceptance criteria. Continuous time Langevin dynamics is shown to be the steepest descent flow of the KL-divergence with respect to the Wasserstein metric [22, 46, 48] and converge exponentially to the target posterior under a LSI assumption on the posterior distribution [38, 46, 48]. For the ULA, [48] shows that a bias exists for any arbitrarily small (fixed) step size. For strongly log-concave posterior distributions, convergence properties of the ULA are discussed in [4, 8, 9, 13, 14, 48]. Analysis of ULA without the log-concave target distribution often requires additional assumption on the negative log of the posterior distribution, including dissipative property [3, 35, 41, 50, 52], relaxed dissipativity conditions [3, 52], contractivity condition [32], or limiting the non-convexity to a local region [5, 31]. In [31, 44, 50], computational efficiency of sampling algorithm compared to optimization methods is reported in the nonconvex setting. The convergence is shown to be polynomial in dimension and error tolerance [5, 32, 35].

Recently, distributed ULA (DULA) and Bayesian learning algorithms have been proposed in [19, 25, 26, 39]. [26] introduces a distributed learning algorithm that projects the local posterior onto an allowed family of posteriors and then performs consensus based on the projected posterior. In [25], convergence of distributed Lagenvin dynamics with strongly log-concave posteriors is investigated. A DULA for non-log-concave posterior distributions is analyzed in [39]. In [19], distributed stochastic gradient Langevin dynamics and HMC methods for strongly log-concave posterior distributions are studied. These algorithms all require inter-agent communication at each iteration.

Contribution: The major contribution of this paper is to introduce the first-ever distributed Bayesian learning algorithm with an event-triggering mechanism for significantly reduced communication. Theoretical convergence properties are established and linked to triggering-threshold tuning parameters applied within and across dispersed agents under the assumption that the target distribution satisfies a log-*Sobolev inequality*. Compared with the prior work in [39], our analysis successfully removes the bounded gradient assumption. The event-triggering mechanism and the removal of the bounded gradient assumption introduce nontrivial complexity in the analysis of both consensus and convergence to the target posterior. To bound the error terms induced by the triggering mechanism, we design the triggering threshold as a diminishing sequence and establish a lower bound for its decay rate. Without the bounded gradient assumption, the consensus is established by analyzing the coupled consensus error and average dynamics, whereas in [39], the consensus is established independently of the average dynamics. In addition, our numerical results demonstrate how the event-triggering mechanism reduces inter-agent communication (by more than 50%) while maintaining inference performance. Indeed, our result shows that the triggering threshold can be selected to prevent consecutive triggering events for the same agent on expectation.

Notation: Let $\mathbb{R}^{n\times m}$ denote the set of $n \times m$ real matrices. For a vector $\boldsymbol{\phi}$, ϕ_i is the i-th entry of $\boldsymbol{\phi}$. An $n \times n$ identity matrix is denoted as I_n and $\mathbf{1}_n$ denotes an n-dimensional vector of all ones. For $p \in [1, \infty]$, the p-norm of a vector $\mathbf{x}$ is denoted as $\|\mathbf{x}\|_p$. For matrices $A \in \mathbb{R}^{m\times n}$ and $B \in \mathbb{R}^{p\times q}$, $A \otimes B \in \mathbb{R}^{mp\times nq}$ 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 and the inter-agent communication links are represented as $\mathcal{E} \triangleq \{e_1, \ldots, e_\ell\} \subseteq \mathcal{V} \times \mathcal{V}$. Let $\mathcal{A} = [a_{i,j}] \in \mathbb{R}^{n\times n}$ be the *adjacency matrix* with entries of $a_{i,j} = 1$ if $(v_i, v_j) \in \mathcal{E}$ and zero otherwise. Let $\mathcal{N}_i$ be the set of neighbors of agent i in $\mathcal{G}(\mathcal{V}, \mathcal{E})$, i.e., $\mathcal{N}_i = \{j \,|\, a_{i,j} = 1, j = 1, \cdots, n\}$. Define $\Delta = \text{diag}(\mathcal{A}\mathbf{1}_n)$ as the n-degree matrix and $\mathcal{L} = \Delta - \mathcal{A}$ as the graph *Laplacian*. Denote by $\mathcal{N}(\boldsymbol{\mu}, M)$ the normal distribution with a mean $\boldsymbol{\mu}$ and a covariance matrix M.

2 Problem formulation

We consider a connected network of n agents, each with a randomly distributed set of m_i data items, $\boldsymbol{X}_i = \left\{\boldsymbol{x}_i^j\right\}_{j=1}^{j=m_i}$, $\forall\, i = 1, \ldots, n$, where $\boldsymbol{x}_i^j \in \mathbb{R}^{d_x}$ is the j-th data element in a set of m_i data items available to the i-th agent. Denote by $\boldsymbol{X}$ the entire data set $\{\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n\}$. Let $\boldsymbol{w} \in \mathbb{R}^{d_w}$ be the parameter vector associated with the model and $p(\boldsymbol{w})$ the prior associated with the model parameters. The *global posterior* distribution of $\boldsymbol{w}$ given the n independent data sets distributed among the agents is expressed as

$$p(\boldsymbol{w}|\boldsymbol{X}) \propto p(\boldsymbol{w}) \prod_{i=1}^{n} p(\boldsymbol{X}_i|\boldsymbol{w}) = \prod_{i=1}^{n} \underbrace{p(\boldsymbol{X}_i|\boldsymbol{w}) p(\boldsymbol{w})^{\frac{1}{n}}}_{\text{pseudo local posterior}} . \tag{1}$$

For the ease of notation, we define $p(\boldsymbol{w}|\boldsymbol{X}_i)$ as the *pseudo local posterior* given by

$$p(\boldsymbol{w}|\boldsymbol{X}_i) = p(\boldsymbol{X}_i|\boldsymbol{w}) p(\boldsymbol{w})^{\frac{1}{n}} . \tag{2}$$

Then the global posterior $p(\boldsymbol{w}|\boldsymbol{X})$ can be written as the product of pseudo local posteriors as

$$p(\boldsymbol{w}|\boldsymbol{X}) \propto \prod_{i=1}^{n} p(\boldsymbol{w}|\boldsymbol{X}_i). \tag{3}$$

This paper is aimed at developing a communication efficient method for collaborative Bayesian learning from large scale data sets distributed among a networked set of agents as a solution to the numerous issues associated with the point estimation schemes. In particular, we present an event-triggered, distributed version of the unadjusted Langevin algorithm to distributively obtain samples from the global posterior $p(\boldsymbol{w}|\boldsymbol{X})$. Compared to the existing distributed Bayesian learning schemes [19,25,26,39], the event-triggered communication scheme significantly reduces the inter-agent communication. We also show that by tuning the triggering threshold, each agent will not triggered at any two consecutive time steps with high probability.

3 Centralized unadjusted Langevin algorithm

In this section, we briefly review a centralized unadjusted Langevin algorithm (CULA) for addressing the learning problem in Section 2. The CULA assumes that all the data sets are collected at a central server. The Langevin algorithm is a well-known class of Monte Carlo sampling algorithms based on the gradient of the negative log-likelihood. However, like any Bayesian technique, a major issue is that the normalizing factor becomes computationally intractable. To that end, an energy function E is defined [12,27,37] as follows

$$p(\boldsymbol{w}|\boldsymbol{X}) \propto \exp(-E(\boldsymbol{w}, \boldsymbol{X})). \tag{4}$$

It then follows from Bayes rule that

$$E(\boldsymbol{w}, \boldsymbol{X}) = -\log\left(p(\boldsymbol{X}|\boldsymbol{w})\right) - \log\left(p(\boldsymbol{w})\right). \tag{5}$$

Therefore, computing $E(\boldsymbol{w}, \boldsymbol{X})$ does not require the normalization constant in the posterior $p(\boldsymbol{w}|\boldsymbol{X})$. The global posterior distribution, i.e., the target distribution is denoted as p^*. It then follows from (4) that

$$p^*(\cdot) = \exp\big(-E(\cdot, \boldsymbol{X}) + C\big), \tag{6}$$

where C is the normalizing constant.

Thereafter, samples in CULA are drawn at each time step as

$$\boldsymbol{w}(k+1) = \boldsymbol{w}(k) - \alpha_k \nabla E(\boldsymbol{w}(k), \boldsymbol{X}) + \sqrt{2\alpha_k}\boldsymbol{v}(k) \tag{7}$$

where $\boldsymbol{w}(k)$ is the sample at the k-th time step, α_k is a time-varying step-size of the algorithm, and $\boldsymbol{v}(k) \sim \mathcal{N}(\boldsymbol{0}, I_{d_w})$, $\forall k \geq 0$ is the additive noise. In essence, (7) performs a gradient ascent on $p(\boldsymbol{w}|\boldsymbol{X})$ with the additive noise $\boldsymbol{v}$ to effectively search the sample space. Substituting (5) in (7) yields

$$\boldsymbol{w}(k+1) = \boldsymbol{w}(k) - \alpha_k \nabla E(\boldsymbol{w}(k), \boldsymbol{X}) + \sqrt{2\alpha_k}\boldsymbol{v}(k). \tag{8}$$

The algorithm (8) is a discrete time approximation of the continuous time Langevin dynamics given by the following Stochastic Differential Equation (SDE) known as the Langevin equation [28]:

$$d\bar{\boldsymbol{w}}^*(t) = -\nabla E(\bar{\boldsymbol{w}}^*(t), \boldsymbol{X})dt + \sqrt{2}d\boldsymbol{B}(t), \tag{9}$$

where $\boldsymbol{B}(t)$ is a d_w-dimensional Brownian motion.

Continuous-time Langevin dynamics (9) have been shown to exponentially converge to $p^*(\cdot)$ for various classes of distributions [22, 42, 48]. Convergence properties of the discrete algorithm (8) have also been widely studied for log-concave target distributions [4,6,9,13–15] and without the strong log-concavity assumption on target distributions [3, 5, 31, 32, 35, 41, 50, 52].

4 The proposed DETULA

In practice, due to network limitations, computational constraints, data privacy restrictions, and other logistical constraints, the learning data set $\boldsymbol{X}$, in its entirety, is often unavailable to a single central server for processing. Such scenarios necessitate the implementation of distributed learning algorithms. To that effect recent studies such as [19, 25, 39] have extended the Langevin algorithm to a distributed setting. Particularly, the distributed ULA (DULA) [39] takes the following form:

$$\boldsymbol{w}_i(k+1) = \boldsymbol{w}_i(k) - \beta_k \sum_{j=1}^{n} a_{i,j}\left(\boldsymbol{w}_i(k) - \boldsymbol{w}_j(k)\right) - \alpha_k n \nabla E_i(\boldsymbol{w}_i(k), \boldsymbol{X}_i) + \sqrt{2\alpha_k}\boldsymbol{v}_i(k), \tag{10}$$

where $a_{i,j}$ denotes the entries of the adjacency matrix corresponding to the communication network $\mathcal{G}(\mathcal{V}, \mathcal{E})$, β_k is the consensus step-size, the additive noise $\boldsymbol{v}_i(k)$ satisfies $\boldsymbol{v}_i(k) \sim \mathcal{N}(\boldsymbol{0}, nI_{d_w})$, $\forall i \in \{1, \ldots, n\}$ and $E_i(\cdot, \boldsymbol{X}_i) = -\log\big(p(\boldsymbol{X}_i|\cdot)\big) - \log\big(p(\cdot)\big)$. In (10) the learning occurs over a distributed network of agents, each having a local sample $\boldsymbol{w}_i(k)$ at the k-th time step and with access to a local data set $\boldsymbol{X}_i$. Restricted information sharing via term $\sum_{j=1}^{n} a_{i,j}(\boldsymbol{w}_i(k) - \boldsymbol{w}_j(k))$ is performed to ensure consensus among agents. However, this leads to a large number of communication at every time step. To reduce the communication overhead for consensus among agents, while still guaranteeing convergence to the target posterior, we propose the following distributed ULA algorithm with an event-triggering mechanism:

$$\boldsymbol{w}_i(k+1) = \boldsymbol{w}_i(k) - \beta_k \sum_{j=1}^{n} a_{i,j}\left(\hat{\boldsymbol{w}}_i(k) - \hat{\boldsymbol{w}}_j(k)\right) - \alpha_k n \nabla E_i(\boldsymbol{w}_i(k), \boldsymbol{X}_i) + \sqrt{2\alpha_k}\boldsymbol{v}_i(k), \tag{11}$$

In (11), for $i = 1, \ldots, n$, the piece-wise constant signal $\hat{\boldsymbol{w}}_i(k)$, defined as

$$\hat{\boldsymbol{w}}_i(k) = \boldsymbol{w}_i(t_q^i),\ \forall\, k \in \left\{t_q^i,\, t_q^i + 1,\, \ldots,\, t_{q+1}^i - 1\right\},$$

denotes agent i's last broadcast sample. The set $\left\{t_q^i : q = 0, 1, \ldots\right\}$ with $t_0^i = 0$ denotes triggering instants, i.e., the time instants when agent i broadcasts $\boldsymbol{w}_i$ to its neighbors.

Define $\mathbf{w}(k) \triangleq \left[\boldsymbol{w}_1(k)^\top, \ldots, \boldsymbol{w}_n(k)^\top\right]^\top \in \mathbb{R}^{nd_w}$, $\hat{\mathbf{w}}(k) \triangleq \left[\hat{\boldsymbol{w}}_1(k)^\top, \ldots, \hat{\boldsymbol{w}}_n(k)^\top\right]^\top \in \mathbb{R}^{nd_w}$, $\mathbf{v}(k) \triangleq \left[\boldsymbol{v}_1(k)^\top, \ldots, \boldsymbol{v}_n(k)^\top\right]^\top \in \mathbb{R}^{nd_w}$ and $\widehat{\nabla E}(\mathbf{w}(k), \boldsymbol{X}) \triangleq \left[\nabla E_1(\boldsymbol{w}_1(k), \boldsymbol{X}_1)(k)^\top, \ldots, \nabla E_n(\boldsymbol{w}_n(k), \boldsymbol{X}_n)(k)^\top\right]^\top \in \mathbb{R}^{nd_w}$. Now (11) can be written in a vectorized form as

$$\mathbf{w}(k+1) = \mathbf{w}(k) - \beta_k \left(\mathcal{L} \otimes I_{d_w}\right) \hat{\mathbf{w}}(k) - \alpha_k n \widehat{\nabla E}(\mathbf{w}(k), \boldsymbol{X}) + \sqrt{2\alpha_k}\mathbf{v}(k), \tag{12}$$

where $\mathcal{L}$ is the network Laplacian. Ignoring the additive noise $\mathbf{v}_k$, (12) can be considered a distributed optimization algorithm which seeks to minimize $E(\boldsymbol{w}, \boldsymbol{X})$ defined as

$$E(\boldsymbol{w}, \boldsymbol{X}) \triangleq \sum_{i=1}^{n} E_i(\boldsymbol{w}, \boldsymbol{X}_i). \tag{13}$$

Define the network weight-matrix $\mathcal{W}_k = (I_n - \beta_k \mathcal{L})$ and $\mathbf{e}(k) = \mathbf{w}(k) - \hat{\mathbf{w}}(k)$. We obtain from (12)

$$\mathbf{w}(k+1) = \left(\mathcal{W}_k \otimes I_{d_w}\right) \mathbf{w}(k) - \alpha_k n \widehat{\nabla E}(\mathbf{w}(k), \boldsymbol{X}) + \sqrt{2\alpha_k}\mathbf{v}(k) + \beta_k \left(\mathcal{L} \otimes I_{d_w}\right) \mathbf{e}(k), \tag{14}$$

Let $\boldsymbol{e}_i(k) = \boldsymbol{w}_i(k) - \hat{\boldsymbol{w}}_i(k)$, and $e^m(k)$ be a tuning parameter. The triggering event instants for any i-th agent are designed as

$$t_{q+1}^i = \inf\left\{k > t_q^i \;\middle|\; \|\boldsymbol{e}_i(k)\|_2^2 > e^m(k)\right\}. \tag{15}$$

where the selection scheme for $e^m(k)$ is discussed in detail in Section 5. The index k in the paranthesis of $e^m(k)$ denotes that it could be a function of time step to allow for more flexibility. As we shall see later in Condition 2, the time index dependence of the threshold $e^m(k)$ shall be required to guarantee consensus.

For implementation, agent i stores its most recent broadcast sample in $\hat{\boldsymbol{w}}_i^{(i)}(k)$, and the most recent sample received from neighbor j in $\hat{\boldsymbol{w}}_j^{(i)}(k)$. Note that $\hat{\boldsymbol{w}}_j^{(i)}(k)$ is equivalent to $\hat{\boldsymbol{w}}_j(k)$ in (10) and that $\hat{\boldsymbol{w}}_i^{(j)}(k) = \hat{\boldsymbol{w}}_i^{(l)}(k)$, $\forall j, l \in \mathcal{N}_i$. Then the triggering mechanism for agent i $(i = 1, \ldots, n)$ at time step k is as follows:

If $\|\boldsymbol{e}_i(k)\|_2^2 > e^m(k)$, then agent i sets $\hat{\boldsymbol{w}}_i^{(i)}(k) = \boldsymbol{w}_i(k)$ and broadcasts $\boldsymbol{w}_i(k)$ to its neighboring agents. Otherwise, agent i does not communicate and $\hat{\boldsymbol{w}}_i^{(i)}(k) = \hat{\boldsymbol{w}}_i^{(i)}(k-1)$.

To compute $\boldsymbol{e}_i(k)$, agent i only makes use of its current sample $\boldsymbol{w}_i(k)$ and the past broadcast sample $\hat{\boldsymbol{w}}_i^{(i)}(k)$, no communication with other agents is needed. The pseudocode of the proposed DETULA is given in Algorithm 1. Additional details on the design parameters of the algorithm are discussed in Section 5. In particular, $e^m(k)$ must be chosen carefully. Increasing $e^m(k)$ results in less communication, but consensus and convergence could be compromised if $e^m(k)$ is too large. As $e^m(k) \to 0$, the DETULA approaches the DULA that communicates at every iteration. The exact form of $e^m(k)$ and associated bounds are established in Condition 2 in Section 5. Mathematical guarantees on the triggering frequency are established in Theorem 3.

5 Main results

We address the two major objectives of our algorithm in this section, namely, consensus and convergence to target distribution p^*. For the distributed approach to be reliable, we first ensure consensus among the agents, and thereafter convergence to the global posterior distribution. Finally, we include some analysis on the frequency of the event-triggering and conditions to ensure meaningful reduction in communication.

Algorithm 1 Distributed Event-Triggered ULA (DETULA)

Input : (i) a according to Condition 1(i)
(ii) b according to Condition 1(ii),
(iii) μ_e satisfying (125),
(iv) δ_1 and δ_2 according to Condition 1(iii)
(v) δ_3 according to Condition 2, and optionally Condition 3 (or simply following (36)).

Initialization : $\mathbf{w}(0) = \left[\boldsymbol{w}_1^\top(0) \ \dots \ \boldsymbol{w}_n^\top(0)\right]^\top$

for $i = 1$ to n **do**
  Broadcast $\boldsymbol{w}_i(0)$ & *set* $\hat{\boldsymbol{w}}_i^{(i)}(0) = \boldsymbol{w}_i(0)$
  Receive $\boldsymbol{w}_j(0)$ & *set* $\hat{\boldsymbol{w}}_j^{(i)}(0) = \boldsymbol{w}_j(0), \forall j \in \mathcal{N}_i$.
  Compute $\nabla E_i(\boldsymbol{w}_i(0), \boldsymbol{\xi}_i(0))$
  Sample $\boldsymbol{v}_i(0) \sim \mathcal{N}(\mathbf{0}, nI_{d_w})$
  Update $\boldsymbol{w}_i(1) = \boldsymbol{w}_i(0) - \alpha_0 \nabla E_i(\boldsymbol{w}_i(0)_{,i}) - \beta_0 \sum_{j=1}^{n} a_{ij}\left(\hat{\boldsymbol{w}}_i^{(i)}(0) - \hat{\boldsymbol{w}}_j^{(i)}(0)\right) + \sqrt{2\alpha_0}\boldsymbol{v}_i(0)$
end for
for $k \geq 0$ **do**
  for $i = 1$ to n **do**
    Sample $\boldsymbol{v}_i(k) \sim \mathcal{N}(\mathbf{0}, nI_{d_w})$
    Compute $\nabla E_i\left(\boldsymbol{w}_i(k), \boldsymbol{X}_i\right)$
    Compute $\boldsymbol{e}_i(k) = \boldsymbol{w}_i(k) - \hat{\boldsymbol{w}}_i^{(i)}(k)$
    if $\|\boldsymbol{e}_i(k)\|_2^2 > e^m(k)$ **then**
      Broadcast $\boldsymbol{w}_i(k)$ & *set* $\hat{\boldsymbol{w}}_i^{(i)}(k) = \boldsymbol{w}_i(k)$
    else
      Set $\hat{\boldsymbol{w}}_i^{(i)}(k) = \hat{\boldsymbol{w}}_i^{(i)}(k-1)$
    end if
    if any $\boldsymbol{w}_j(k), j \in \mathcal{N}_i$ received **then**
      Set $\hat{\boldsymbol{w}}_j^{(i)}(k) = \boldsymbol{w}_j(k)$
    else
      Set $\hat{\boldsymbol{w}}_j^{(i)}(k) = \hat{\boldsymbol{w}}_j^{(i)}(k-1)$
    end if
    Update $\boldsymbol{w}_i(k+1) = \boldsymbol{w}_i(k) - \alpha_k \nabla E_i\left(\boldsymbol{w}_i(k), \boldsymbol{X}_i\right) - \beta_k \sum_{j=1}^{n} a_{ij}\left(\hat{\boldsymbol{w}}_i^{(i)}(k) - \hat{\boldsymbol{w}}_j^{(i)}(k)\right) + \sqrt{2\alpha_k}\boldsymbol{v}_i(k)$
  end for
end for

5.1 Consensus in mean-square expectation

Define the average-consensus error as

$$\tilde{\mathbf{w}}(k) = \left(I_{nd_w} - \frac{1}{n}\mathbf{1}_n\mathbf{1}_n^\top \otimes I_{d_w}\right)\mathbf{w}(k). \tag{16}$$

Our first result shows that the mean-square expectation of $\tilde{\mathbf{w}}(k)$ in (16) is decreasing at the rate $\mathcal{O}\left(\frac{1}{(k+1)^{\delta_2-2\delta_1}}\right)$ when operated within certain limits of event-triggering. Our consensus result relies on the following commonly made assumptions.

Assumption 1 *The interaction topology of n networked agents is given as a connected undirected graph $\mathcal{G}(\mathcal{V}, \mathcal{E})$.*

For the connected undirected graph $\mathcal{G}(\mathcal{V}, \mathcal{E})$, the graph Laplacian $\mathcal{L}$ is a positive semi-definite matrix with one eigenvalue at 0 corresponding to the eigenvector $\mathbf{1}_n$. Furthermore, it follows from Lemma 3 in [21] that for all $\mathbf{x} \in \mathbb{R}^n$, such that $\mathbf{1}_n^T\mathbf{x} = 0$, we have $\mathbf{x}^T\mathcal{L}\left(\mathcal{L}\right)^+\mathbf{x} = \mathbf{x}^T\mathbf{x}$, where $(\cdot)^+$ denoted the pseudo-inverse. Let $\mathcal{F}_k$ denotes a filtration generated by the sequence $\{\mathbf{w}_0, \dots, \mathbf{w}_k\}$, i.e., $\mathbb{E}[\,\mathbf{v}_k\,|\mathcal{F}_k] = \mathbf{0}$.

Assumption 2 *For all $i = \{1, 2, \ldots, n\}$, the individual gradients $\nabla E_i : \mathbb{R}^{d_w} \mapsto \mathbb{R}^{d_w}$ are Lipschitz continuous with Lipschitz constant $L_i > 0$, i.e., $\forall\ \boldsymbol{w}_a, \boldsymbol{w}_b \in \mathbb{R}^{d_w}$,*

$$\|\nabla E_i(\boldsymbol{w}_a, \boldsymbol{X}_i) - \nabla E_i(\boldsymbol{w}_b, \boldsymbol{X}_i)\|_2 \leq L_i \|\boldsymbol{w}_a - \boldsymbol{w}_b\|_2. \tag{17}$$

As a result of (17) we have $\nabla E : \mathbb{R}^{d_w} \mapsto \mathbb{R}^{d_w}$ and $\widehat{\nabla E} : \mathbb{R}^{nd_w} \mapsto \mathbb{R}^{nd_w}$ are both Lipschitz continuous with Lipschitz constant $L > 0$ and $\bar{L} > 0$ respectively. Hence, $\forall\ \boldsymbol{w}_a, \boldsymbol{w}_b \in \mathbb{R}^{d_w}$ and $\forall\ \mathbf{w}_a, \mathbf{w}_b \in \mathbb{R}^{nd_w}$ we respectively have

$$\begin{aligned} \|\nabla E(\boldsymbol{w}_a, \boldsymbol{X}) - \nabla E(\boldsymbol{w}_b, \boldsymbol{X})\|_2 &\leq \bar{L} \|\boldsymbol{w}_a - \boldsymbol{w}_b\|_2, \\ \|\widehat{\nabla E}(\mathbf{w}_a, \boldsymbol{X}) - \widehat{\nabla E}(\mathbf{w}_b, \boldsymbol{X})\|_2 &\leq L \|\mathbf{w}_a - \mathbf{w}_b\|_2, \end{aligned} \tag{18}$$

where $\bar{L} \leq \max\limits_{i=\{1,2,\ldots,n\}} nL_i$ and $L \leq \max\limits_{i=\{1,2,\ldots,n\}} L_i$.

Next, we list the essential conditions that need to be fulfilled for successful operation of the proposed DETULA algorithm.

Condition 1 *Sequences $\{\alpha_k\}$ and $\{\beta_k\}$ are selected as*

$$\alpha_k \triangleq \frac{a}{(k+1)^{\delta_2}} \quad \text{and} \quad \beta_k \triangleq \frac{b}{(k+1)^{\delta_1}}, \tag{19}$$

where the positive constants a, b, δ_1 and δ_2 satisfy:

(i) $a < \frac{1}{nL}$ *and* $\phi(a) < 0$ *where* $\phi(a)$ *is given in* (78),
(ii) $\frac{anL}{\lambda_2(\mathcal{L})} < b < \frac{1}{\lambda_2(\mathcal{L})}$ *and* $\frac{b}{1 - b\lambda_2(\mathcal{L})} < \frac{\lambda_2(\mathcal{L})}{d_m^2}$,
(iii) $\frac{1}{2} + \delta_1 < \delta_2 < 1$.

Condition 1(*iii*) puts restrictions on the decay rates of α_k and β_k, where the consensus gain β_k decreases at a slower rate than the gradient step-size α_k. For sequences $\{\alpha_k\}$ and $\{\beta_k\}$ that satisfy Condition 1, α_k, β_k and β_k^2 are not summable sequences while α_k is square-summable. Additionally, we have proved in Section 8.1 that there always exists $\bar{a} \in (0, 1)$ such that for all $a < \bar{a}$ we have $\phi(a) < 0$, i.e., for small enough values of a, $\phi(a) < 0$ can always be ensured.

Condition 2 *The triggering threshold $e^m(k)$ in* (15) *is chosen as*

$$e^m(k) = \frac{\mu_e}{(k+1)^{\delta_3}}, \tag{20}$$

where $\mu_e = e^m(0)$ is a tuning parameter of the algorithm and δ_3 satisfies $0 < 1 - \delta_2 < \delta_3$.

Condition 2 gives a suitable way to choose the error threshold before event is triggered. $e^m(k)$ signifies the limit up to which the error (in the estimation of the samples of its neighbors) an agent is allowed to incur before triggering (i.e., sharing information with its neighbors) without compromising consensus. Furthermore, we design a diminishing threshold for asymptotic consensus.

Theorem 1 *Given Conditions 1 and 2 along with Assumptions 1 - 2, the average-consensus error $\tilde{\mathbf{w}}$ for the DETULA given in Algorithm 1 satisfies*

$$\mathbb{E}\left[\|\tilde{\mathbf{w}}_{k+1}\|_2^2\right] \leq \frac{W_5}{\exp\left(W_1(k+1)^{1-\delta_1}\right)} + \frac{W_2\mu_e}{(k+1)^{\delta_3}} + \frac{W_3 C_{\bar{w}}}{(k+1)^{2\delta_2 - 2\delta_1}} + \frac{W_4}{(k+1)^{\delta_2 - 2\delta_1}}, \tag{21}$$

where W_2, W_3, W_4 and W_5 are positive constants defined in (62)– (65), *respectively.*

Detailed proof of Theorem 1 can be found in Section 8.1. The important conclusion from Theorem 1 is the guarantee of consensus on mean-square expectation with increasing time steps. Since the first term in (21) decays exponentially,

we expect $\mathbb{E}\left[\|\tilde{\mathbf{w}}(k+1)\|_2^2\right] \propto (k+1)^{-\delta_p}$ for large k, where $\delta_p = \min\{\delta_3, \delta_2 - 2\delta_1\}$. Moreover, if $\delta_3 \geq \delta_2 - 2\delta_1$ (which is possible to attain within the restrictions imposed on δ_3), then $\mathbb{E}\left[\|\tilde{\mathbf{w}}(k+1)\|_2^2\right] \propto (k+1)^{-(\delta_2 - 2\delta_1)}$ which is the same as without event-triggering. Additionally, we observe from (21) that a larger μ_e, although allows for higher accumulation of the error and thus less communications, results in a larger upper bound of the consensus error.

5.2 Average Langevin dynamics

With the average consensus result in Theorem 1, we proceed to analyze the average dynamics from the DETULA in (14). Let $\bar{\boldsymbol{w}}(k) = \frac{1}{n}\sum_{i=1}^n \boldsymbol{w}_i(k)$ and $\bar{\boldsymbol{v}}(k) = \frac{1}{n}\sum_{i=1}^n \boldsymbol{v}_i(k)$. It follows that

$$\bar{\boldsymbol{w}}(k+1) = \bar{\boldsymbol{w}}(k) - \alpha_k \sum_{i=1}^n \nabla E_i\left(\boldsymbol{w}_i(k), \boldsymbol{X}_i\right) + \sqrt{2\alpha_k}\bar{\boldsymbol{v}}(k). \tag{22}$$

Because $\boldsymbol{v}_i(k)$'s are independent and satisfy $\boldsymbol{v}_i(k) \sim \mathcal{N}(\mathbf{0}, nI_{d_w})$, $\bar{\boldsymbol{v}}(k)$ satisfies $\bar{\boldsymbol{v}}(k) \sim \mathcal{N}(\mathbf{0}, I_{d_w})$. Define

$$\zeta(\bar{\boldsymbol{w}}(k), \tilde{\mathbf{w}}(k)) \triangleq \sum_{i=1}^n \Big(\nabla E_i\left(\bar{\boldsymbol{w}}(k) + \tilde{\boldsymbol{w}}_i(k), \boldsymbol{X}_i\right) - \nabla E_i\left(\bar{\boldsymbol{w}}(k), \boldsymbol{X}_i\right)\Big), \tag{23}$$

where $\tilde{\boldsymbol{w}}_i(k) \triangleq \boldsymbol{w}_i(k) - \bar{\boldsymbol{w}}(k)$. Also, from (13) that $\sum_{i=1}^n \nabla E_i\left(\bar{\boldsymbol{w}}(k), \boldsymbol{X}_i\right) = \nabla E\left(\bar{\boldsymbol{w}}(k), \boldsymbol{X}\right)$. Therefore, we can rewrite (22) as

$$\bar{\boldsymbol{w}}(k+1) = \bar{\boldsymbol{w}}(k) - \alpha_k \nabla E\left(\bar{\boldsymbol{w}}(k), \boldsymbol{X}\right) - \alpha_k \zeta(\bar{\boldsymbol{w}}(k), \tilde{\mathbf{w}}_k) + \sqrt{2\alpha_k}\bar{\boldsymbol{v}}(k). \tag{24}$$

Missing from the CULA (7), the ζ term in (24) represents the effect due to the consensus error $\tilde{\mathbf{w}}(k)$. If $\tilde{\mathbf{w}}(k) = 0$, the ζ term becomes zero.

To analyze the convergence of the distribution of $\bar{\boldsymbol{w}}(k)$, we convert (24) to a continuous time system (denoted by t) that yields the same distribution of $\bar{\boldsymbol{w}}$ at the discrete time step k, $k \geq 0$. Towards this end, we define a continuous time $t_k = \sum_{j=0}^{k-1} \alpha_j$, which corresponds to the k-th time step in the discrete time system in (24). We now rewrite (24) as

$$\bar{\boldsymbol{w}}(t_{k+1}) = \bar{\boldsymbol{w}}(t_k) - \alpha_k \nabla E\left(\bar{\boldsymbol{w}}(t_k), \boldsymbol{X}\right) - \alpha_k \zeta(\bar{\boldsymbol{w}}(t_k), \tilde{\mathbf{w}}_{t_k}) + \sqrt{2}\Big(\boldsymbol{B}(t_{k+1}) - \boldsymbol{B}(t_k)\Big), \tag{25}$$

which is further represented as a SDE in continuous time for $t \in [t_k, t_{k+1})$, given by

$$d\bar{\boldsymbol{w}}(t) = -\Big(\nabla E\left(\bar{\boldsymbol{w}}(t_k), \boldsymbol{X}\right) - \zeta(\bar{\boldsymbol{w}}(t_k), \tilde{\omega}(t_k))\Big)dt + \sqrt{2}d\boldsymbol{B}(t), \tag{26}$$

where we set $\tilde{\omega}(t_k) = \tilde{\mathbf{w}}(k),\ \forall t \in [t_k,\ t_{k+1})$ for any $k \geq 0$. Since the gradient terms in (26) are constant for the entire interval $t \in [t_k, t_{k+1})$, (26) can be integrated within $[t_k, t_{k+1})$ to give exactly (25). With the same initial distribution of $\bar{\boldsymbol{w}}$, the distributions of $\bar{\boldsymbol{w}}$ in (25) and (26) are the same at the discrete time instants t_k, $\forall k \geq 0$.

5.3 Convergence to $p^*(\cdot)$

Let $\bar{\boldsymbol{w}}(t)$ in (26) admit a probability distribution $p_t(\bar{\boldsymbol{w}})$ for $t_k \leq t < t_{k+1}$. Our objective is to prove $p_{t_k}(\bar{\boldsymbol{w}}) \to p^*$ as $k \to \infty$. Motivated by [5, 31, 46, 48], we analyze (26) using the KL-divergence of $p_t(\bar{\boldsymbol{w}})$ to the target distribution p^* as a Lyapunov functional. Denote such KL divergence by $F(p_t(\bar{\boldsymbol{w}}))$, which is given by

$$F(p_t(\bar{\boldsymbol{w}})) = \int p_t(\bar{\boldsymbol{w}}) \log\left(\frac{p_t(\bar{\boldsymbol{w}})}{p^*(\bar{\boldsymbol{w}})}\right) d\bar{\boldsymbol{w}}. \tag{27}$$

Note that $F(p_t(\bar{\boldsymbol{w}})) \geq 0$ and $F(p_t(\bar{\boldsymbol{w}})) = 0$ if and only if $p_t = p^*$. Here, we introduce $G : \mathbb{R}^{nd_w} \times \mathbb{R}^{\sum_i m_i d_x} \mapsto \mathbb{R}$, an aggregate potential function of local variables $\boldsymbol{w}_i(k)$ and local data $\boldsymbol{X}_i$

$$G(\mathbf{w}_k, \boldsymbol{X}) \triangleq \sum_{i=1}^{n} E_i\left(\boldsymbol{w}_i(k), \boldsymbol{X}_i\right), \tag{28}$$

where $\mathbf{w} \triangleq [\boldsymbol{w}_1^\top, \ldots, \boldsymbol{w}_n^\top]^\top$. From (28), $G(\mathbf{w}, \boldsymbol{X})$ is continuously differentiable and its gradient $\nabla G(\mathbf{w}, \boldsymbol{X}) \in \mathbb{R}^{nd_w}$ is given as $\nabla G(\mathbf{w}, \boldsymbol{X}) = \widehat{\nabla E}(\mathbf{w}, \boldsymbol{X})$ (which is defined prior to (12)). It follows from (18) that $\nabla G : \mathbb{R}^{nd_w} \mapsto \mathbb{R}^{nd_w}$ is Lipschitz continuous as well with a Lipschitz constant L, i.e., $\forall\, \mathbf{w}_a,\, \mathbf{w}_b \in \mathbb{R}^{nd_w}$

$$\|\nabla G(\,\mathbf{w}_a, \boldsymbol{X}\,) - \nabla G(\,\mathbf{w}_b, \boldsymbol{X}\,)\|_2 \leq L\|\mathbf{w}_a - \mathbf{w}_b\|_2. \tag{29}$$

Assumption 3 *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*

$$\int g(\bar{\boldsymbol{w}}) \log g(\bar{\boldsymbol{w}}) p^*(\bar{\boldsymbol{w}})\, d\bar{\boldsymbol{w}} \leq \frac{1}{2\rho_U} \int \frac{\|\nabla g(\bar{\boldsymbol{w}})\|^2}{g(\bar{\boldsymbol{w}})} p^*(\bar{\boldsymbol{w}})\, d\bar{\boldsymbol{w}}, \tag{30}$$

where ρ_U is the log-Sobolev constant.

If $g(\bar{\boldsymbol{w}}) = \dfrac{p_t(\bar{\boldsymbol{w}})}{p^*(\bar{\boldsymbol{w}})}$, the inequality (30) becomes

$$F(p_t(\bar{\boldsymbol{w}})) \;\triangleq\; \mathbb{E}_{p_t(\bar{\boldsymbol{w}})}\left[\log\left(\frac{p_t(\bar{\boldsymbol{w}})}{p^*(\bar{\boldsymbol{w}})}\right)\right] \leq \frac{1}{2\rho_U}\mathbb{E}_{p_t(\bar{\boldsymbol{w}})}\left[\left\|\nabla \log\left(\frac{p_t(\bar{\boldsymbol{w}})}{p^*(\bar{\boldsymbol{w}})}\right)\right\|_2^2\right]. \tag{31}$$

The LSI assumption on the *global* posterior distribution is common in literature for analyzing Langevin dynamics and algorithms, see e.g., [31,46]. LSI is satisfied by strongly log-concave $p^*(\bar{\boldsymbol{w}})$; namely, when $-\log p^*(\bar{\boldsymbol{w}})$ is strongly convex. However, LSI applies to a much broader class of probability measures, including many examples of non-log-concave distributions. For example, in [31], it was shown that a posterior distribution that is strongly convex outside of a bounded region, but nonconvex inside of it satisfies an LSI.

Theorem 2 below shows that $F(p_t(\bar{\boldsymbol{w}}))$ converges to zero, indicating that $p_t(\bar{\boldsymbol{w}})$ converges to the target distribution p^*.

Theorem 2 *Consider the DETULA given in Algorithm 1 under Assumptions 1 - 3 along with Conditions 1 and 2. Suppose that the target distribution p^* satisfies the LSI* (30) *with a constant $\rho_U > 0$, and has a bounded second moment, i.e., $\int \|\bar{\boldsymbol{w}}\|_2^2 p^*(\bar{\boldsymbol{w}})\, d\bar{\boldsymbol{w}} \leq c_1$ for some bounded positive constant c_1. Then for all initial distributions $p_{t_0}(\bar{\boldsymbol{w}})$ satisfying $F(p_{t_0}(\bar{\boldsymbol{w}})) \leq c_2$, we have*

$$\begin{aligned} F(p_{t_{k+1}}(\bar{\boldsymbol{w}})) \leq \bar{C}_{_{F_1}} \exp\left(-\rho_U \sum_{\ell=0}^{k} \alpha_\ell\right) + \mu_e\left(\frac{\bar{C}_{_{F_2}}}{(k+1)^{\delta_2-2\delta_1}} + \frac{\bar{C}_{_{F_3}}}{(k+1)^{\delta_3}}\right) + \frac{\bar{C}_{_{F_4}}}{(k+1)^{\delta_2-2\delta_1}} \\ + \bar{C}_{_{F_5}} \exp\left(-\frac{\rho_U a}{1-\delta_2}(k+1)^{1-\delta_2}\right), \end{aligned} \tag{32}$$

where the positive constants $\bar{C}_{_{F_1}}$, $\bar{C}_{_{F_2}}$, $\bar{C}_{_{F_3}}$, $\bar{C}_{_{F_4}}$ and $\bar{C}_{_{F_5}}$ other associated parameters are defined in (100)-(103).

The constants c_1 and c_2 are assumed for establishing the result in Theorem 2 which appear in the constants $\bar{C}_{_{F_4}}$ and $\bar{C}_{_{F_1}}$ respectively. Existence of c_2 has been proven for certain random initializations [31, 45] in the centralized case. For our purpose, c_2 can be derived for $\boldsymbol{w}_i(0) \sim \mathcal{N}\left(\boldsymbol{w}_i^*, \frac{1}{L_i} I_{d_w}\right)$ where $\boldsymbol{w}_i^*$ is a stationary point of $E_i(\cdot)$. Hence, one may choose $\boldsymbol{w}_i(0) \sim \mathcal{N}\left(\boldsymbol{w}_i^*, \sigma_i^2 I_{d_w}\right)$, where $\boldsymbol{w}_i^*$ can be individually estimated by a few steps of local gradient

descent on the local data by each agent and $\sigma_i^2 > 0$ is hueristically chosen. For all the experiments in this paper, we use $\boldsymbol{w}_i(0) \sim \mathcal{N}(\mathbf{0}, \sigma^2)$ which also produce satisfactory results.

The proof of Theorem 2 is given in Section 8.2. Since $\boldsymbol{w}_i = \bar{\boldsymbol{w}} + \tilde{\boldsymbol{w}}_i$, the convergence results in Theorem 1 and 2 ensure that the distribution of $\boldsymbol{w}_i$ converges to the target distribution p^*. The exponential decay rate of $F(p_{t_0}(\bar{\boldsymbol{w}}))$ is similar to the results available in the literature, e.g., [4,31,46]. The second term in (32), $\mu_e \left(\frac{\bar{C}_{F_2}}{(k+1)^{\delta_2 - 2\delta_1}} + \frac{\bar{C}_{F_3}}{(k+1)^{\delta_3}} \right)$, is the contribution of the event-triggering scheme in the convergence of the DETULA algorithm. When $\mu_e = 0$, we recover the convergence of a DULA where the agents communicate at every iteration. The strength of the DETULA lies in the guaranteed convergence as shown in Theorem 2 despite the reduced inter-agent communication.

Corollary 1 *For the DETULA given in Algorithm 1 under Assumptions 1 - 2 and Conditions 1 and 2, it follows from Theorem 2 that for any $\epsilon > 0$, we have $F(p_{t_k}(\bar{\boldsymbol{w}})) \leq \epsilon$, $\forall k \geq k^*$, if μ_e is selected such that:*

$$\mu_e \leq \left(\frac{(k^*+1)^{\delta_2 - 2\delta_1}}{3\bar{C}_{F_3}} \right) \epsilon, \tag{33}$$

where

$$k^* = \max \left\{ \left(\frac{1-\delta_2}{a\rho_U} \log \left(\frac{3Q_1}{\epsilon} \right) \right)^{\frac{1}{1-\delta_2}} - 1, \left(\frac{3Q_2}{\epsilon} \right)^{\frac{1}{\delta_2 - 2\delta_1}} - 1 \right\}, \tag{34}$$

and the constants Q_1 and Q_2 are given in (106) *and* (107)*, respectively.*

Given a certain threshold ϵ for the KL divergence, Corollary 1 gives an estimate of the number of iterations k^* and the upper bound on μ_e to guarantee that the KL divergence is within the threshold ϵ after k^* iterations.

5.4 Frequency of triggering

Since a key focus of this paper is the reduction in the communication of the DETULA algorithm via event triggering, it is of interest to analyze the frequency of triggering for any particular agent. Ideally, we would like to avoid instances of consecutive triggering by any agent. We next show that μ_e, δ_1, δ_2 and δ_3 can be adjusted to prevent such occurences on expectation. Specifically, we analyze the accumulation of the error $\boldsymbol{e}_i$ right after a triggering event by the i-th agent and establish the conditions on μ_e such that $\boldsymbol{e}_i$ is sufficiently small on expectation to prevent a consecutive triggering. Towards this end, we introduce additional conditions on δ_1, δ_2 and δ_3.

Condition 3 *δ_1, δ_2 and δ_3 satisfy the following conditions:*

$$\delta_3 \leq \delta_m, \tag{35}$$

where $\delta_m = \min\{2\delta_1, \delta_1 + \delta_3, \delta_2 - \delta_1\}$

Note that the lower bound on δ_3 given in Condition 2 is dictated by the consensus requirement of the algorithm as too low δ_3 will lead to higher accumulation of error due to estimation and if unchecked may lead to failure of consensus. On the other hand, the upper bound on δ_3 given in (35) is obtained from the frequency of event-triggering analysis, as too high δ_3 will likely cause consecutive event-triggering, marring the benefits of it in the first place. Thus, we suggest choosing δ_3 as

$$\max\{1 - \delta_1, \delta_2 - 2\delta_1\} < \delta_3 \leq \delta_m. \tag{36}$$

Theorem 3 *Consider the DETULA given in Algorithm 1 under Assumptions 1 - 3 with Conditions 1 - 3. Suppose*

that k_q is some arbitrary time step at which i-th agent is triggered, i.e., $\boldsymbol{e}_i(k_q) = 0$, then we have

$$\mathbb{E}[\|\boldsymbol{e}_i(k_q+1)\|_2^2] \leq \frac{\xi_q}{(k_q+1)^{\delta_m}}, \tag{37}$$

where ξ_q is given as in (120)*. Under the assumption that μ_e satisfies* (125)*, the probability of triggering at k_q+1 for agent i satisfies*

$$p\left(\|\boldsymbol{e}_i(k_q+1)\|_2^2 \geq \frac{\mu_e}{(k_q+2)^{\delta_3}}\right) \leq \frac{(k_q+2)^{\delta_3}}{(k_q+1)^{\delta_m}}\left(\frac{\bar{c}_1}{\mu_e}+\bar{c}_2\right), \tag{38}$$

where $\bar{c}_1$ and $\bar{c}_2$ are constants given in (121) *and* (122) *respectively.*

In effect, (38) in Theorem 3 gives an upper bound on the probability of any i-th agent triggering at consecutive time instants. Thus, keeping this upper bound low enough by appropriately choosing the design parameters, we can prevent consecutive triggering by any agent on expectation. This results in at least a 50% expected reduction in the total communication as compared to the non-triggered DULA algorithm.

The coefficient $\frac{(k_q+2)^{\delta_3}}{(k_q+1)^{\delta_m}}$ on the right hand side of (38) drastically decays to $\frac{\delta_3}{\delta_m} \leq 1$ with increasing k_q. In fact, the larger is the difference in the powers, i.e., the greater is $\delta_m - \delta_3$, the faster is the convergence of $\frac{(k_q+2)^{\delta_3}}{(k_q+1)^{\delta_m}}$ and the smaller is the limiting value $\frac{\delta_3}{\delta_m}$. Our next objective is to adjust the term $\left(\frac{\bar{c}_1}{\mu_e}+\bar{c}_2\right)$ to be as low as possible to reduce triggering. From (122) we note that $\bar{c}_2 = \frac{(1-b\lambda_2(\mathcal{L}))^{-1}bd_m^2}{\lambda_2(\mathcal{L})}$, where d_m is the maximal number of neighbors across all the agents. Thus, $\bar{c}_2 \propto d_m^2$ which means more number of individual connections of agents (i.e., more neighbors) will increase chances of frequent triggering. This is intuitively expected as more neighbors would mean higher accumulation of error from approximating the samples of all those neighbors. Also, note that $\bar{c}_2 \propto b$, hence, we can adjust the value of b to keep $\bar{c}_2$ sufficiently low. As for the term $\frac{\bar{c}_1}{\mu_e}$, it can be appropriately tuned to a suitably low value by increasing μ_e with the minimum value of μ_e given by (125).

6 Numerical experiments

6.1 1D Gaussian toy problem

To demonstrate the proposed algorithm, we first use a 1D Gaussian toy problem, for which we can analytically compute the posterior. We then make a comparison of the posterior approximated by our algorithm with the true analytical posterior. Let

$$\theta \sim \mathcal{N}(0, \sigma_\theta^2), \tag{39}$$

$$x_i|\theta \sim \mathcal{N}(\theta, \sigma_x^2), \qquad i = 1, 2, \ldots, N, \tag{40}$$

where we use $\sigma_\theta = 10$, $\theta_x = 0.1$ and $N = 100$. The analytical expression for the posterior is given as

$$\pi = \mathcal{N}\left(\mu_p, \sigma_p^2\right) = \mathcal{N}\left(\frac{\sum_{i=1}^N x_i}{\frac{\sigma_x^2}{\sigma_\theta^2}+N}, \left(\frac{1}{\sigma_\theta^2}+\frac{N}{\sigma_x^2}\right)^{-1}\right).$$

The entire data was distributed equally among 5 agents. The communication topology is a ring graph where each agent communicates with two neighbors. The true value of θ is 12.22, while $\mu_p = 12.23$ and $\sigma_p^2 = 0.01$. The estimation error of the mean and the variance by DULA ($\mu_e = 0$) and DETULA ($\mu_e = 0.001, \delta_3 = 0.16$) are plotted in Figure 1, which shows that the estimates converge to the true values. The values of other hyperparameters used are: $a = 10^{-5}$, $b = 0.1443$, $\delta_1 = 0.08$ and $\delta_2 = 0.84$.

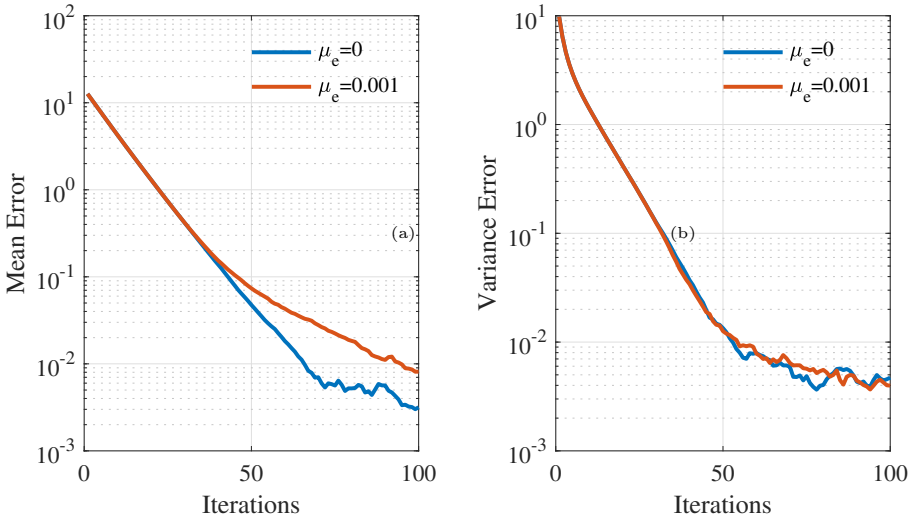

Fig. 1. Comparison of the absolute error between the (a) mean and (b) variance estimation by the agents and the analytical solution. The plots show the average result for all the agents. For the DETULA, we obtain over 50% reduction in communication compared to without event-triggering.

6.2 Gaussian mixture

We consider parameter inference of a Gaussian mixture with tied means [47]. The Gaussian mixture is given by

$$\theta_1 \sim \mathcal{N}(0, \sigma_1^2) \quad ; \quad \theta_2 \sim \mathcal{N}(0, \sigma_2^2) \tag{41}$$

$$x_i \sim \frac{1}{2}\mathcal{N}(\theta_1, \sigma_x^2) + \frac{1}{2}\mathcal{N}(\theta_1 + \theta_2, \sigma_x^2), \tag{42}$$

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 $\theta_1 = 0$ and $\theta_2 = 1$. The 100 data samples were equally distributed among n agents, each receiving a set of $\frac{100}{n}$ data samples. The communication topology between the agents is a ring graph.

Monte Carlo simulations with 100 random trials were conducted for 100000 iterations. The samples from the DETULA were compared with the non-triggered DULA ($\mu_e = 0$) and an approximated true posterior distribution in Figure 2(a). To compare the accuracy of our results, we used [2] to compute Wasserstein distances as a metric. The Wasserstein distances were averaged over the 100 trials and the n agents involved. Note that the presented values of Wasserstein distances (in Figures 3, 4; Tables 1, 2) are approximations because the target posterior itself is approximated and because the algorithm [2] only produces rough estimates of the Wasserstein distances. The presented values of Wasserstein distances are approximations since the target posterior itself is approximated. However, it serves a good metric for comparison between the DULA and the DETULA performances.

We set $\alpha_k = \frac{a}{(\gamma k+1)^{\delta_2}}$ and $\beta_k = \frac{b}{(\gamma k+1)^{\delta_1}}$, where $a = \frac{0.2}{230^{0.55}}$, $b = \frac{1.01}{\sigma_{max}(\mathcal{L})}$, $\delta_1 = \frac{1}{6}$, $\delta_2 = \frac{2}{3}$, $\gamma = \frac{1}{230}$ for both DULA and DETULA, and $\sigma_{max}(\cdot)$ denotes the largest singular value. Here we use 'γk', instead of 'k' in the step sizes for fine-tuning the step sizes to ensure numerical stability of the algorithm. Using a scale factor γ does not affect the theoretical results provided in the main paper. The error bound in event-triggering of DETULA was chosen as $\frac{\mu_e}{(k+1)^{\delta_3}}$, with $\delta_3 = \frac{1}{3}$ in all cases. Also, each agent used the full gradient from the batch of its corresponding data for all the experiments in this section.

A visual comparison between the approximate target posterior and the estimated posteriors from the DETULA and the DULA for one of the 5 agents is shown in Figure 2. To illustrate consensus, we randomly choose an agent as the reference agent and define the consensus metric C_{err} as the mean of the Wasserstein distances of the posteriors of the remaining 4 agents from that of the reference agent averaged over all 100 random trials. Figure 3 demonstrates the improvement of the consensus with increasing iterations. Further, a comparison of the performances of DULA

and DETULA for 5 agents at different iterations (but keeping all other parameters same) is presented in Figure 4. An overall improvement of the performance with increasing iterations is observed in all cases.

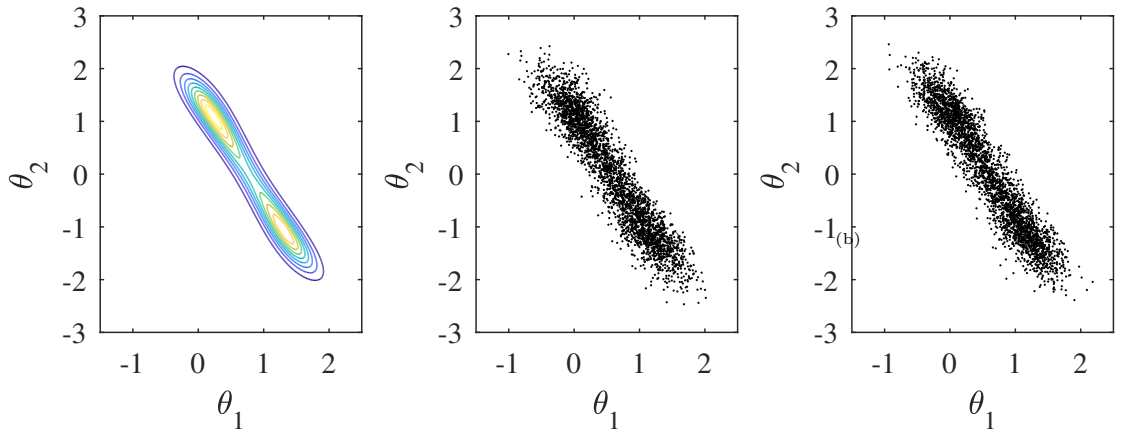

Fig. 2. Comparison of the estimate of the posterior distribution constructed by agent 4 using the DETULA and the DULA with 5 total agents. (a) (Approximate) true posterior distribution (b) DETULA ($\mu_e = 0.3, b = 0.14$) (c) DULA ($\mu_e = 0, b = 0.14$).

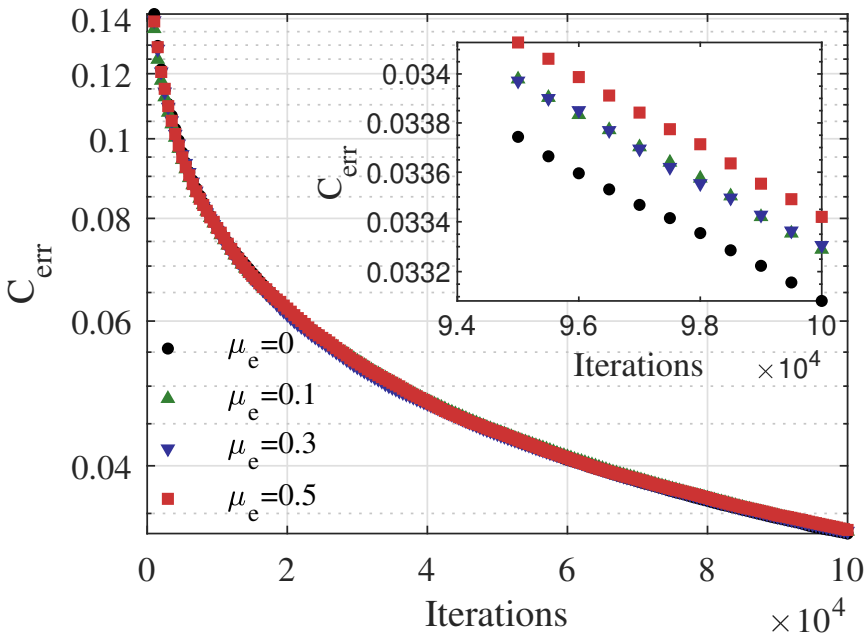

Fig. 3. Comparison of the consensus metric C_{err} of DULA ($\mu_e = 0$) with DETULA ($\mu_e \in \{0.1, 0.3, 0.5\}$) with 5 agents at different iterations from 1000 to 100000 iterations, computed at every 500 iterations for $b = 0.14$. The inset shows the magnified plot of C_{err} for iterations from 95000 to 100000 computed at every 500 iterations.

Table 1 demonstrates the effect of μ_e on the triggering frequency and Wasserstein distance of the final samples. We fixed $n = 5$ and performed the simulations for $\mu_e \in \{0, 0.1, 0.3, 0.5\}$ (where $\mu_e = 0$ corresponds to non-triggered DULA) in Table 1. We observe that the DETULA's performance is on par with the DULA for a wide range of μ_e. A clear benefit of the DETULA is its significant reduction in inter-agent communication. For example, inter-agent communication can be reduced by 75% while minimal loss in estimation quality for $\mu_e = 0.5$.

Next, we explore the effect of varying b in β_k on triggering frequency in Table 2 for fixed $\mu_e = 0.3$. A clear trend of diminishing number of average communications between the agents is noted with decreasing value of b, thus suggesting a reduction in triggering frequency with lower b values. This validates our results in (38) from Theorem 3.

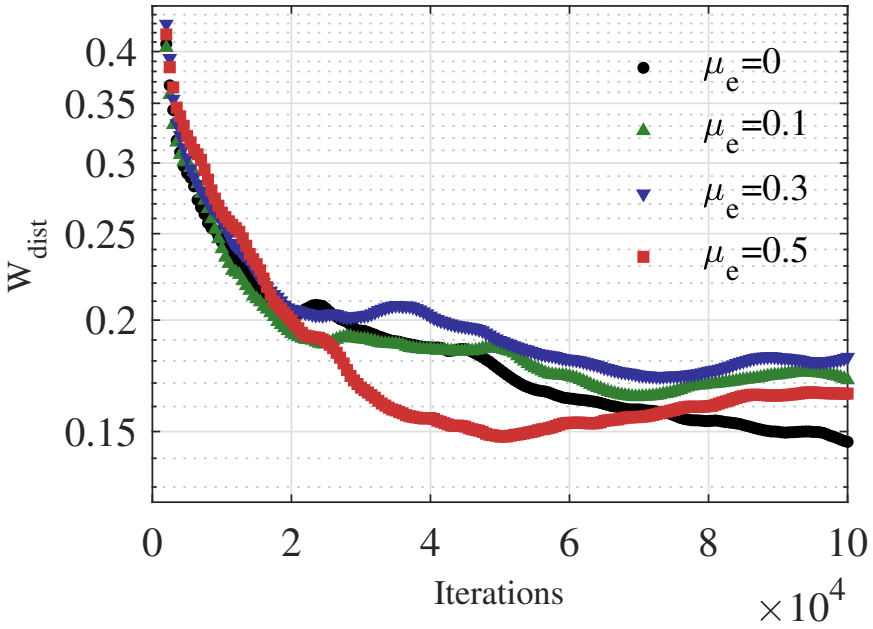

Fig. 4. Comparison of the performance of DULA ($\mu_e = 0$) with DETULA ($\mu_e \in \{0.1, 0.3, 0.5\}$) with 5 agents at different iterations from 2000 to 100000 iterations, computed at every 500 iterations. The plotted W_{dist} values denote average over all the 5 agents and all 100 random trials for $b = 0.14$.

$(n = 5, b = 0.14)$	W_{dist}	N_{comm}	PR
$\mu_e = 0$	0.1771	100000	0%
$\mu_e = 0.1$	0.1906	66669	33%
$\mu_e = 0.3$	0.2190	37072	63%
$\mu_e = 0.5$	0.2027	24735	75%

Table 1
Comparison of results from the DULA ($\mu_e = 0$) and the DETULA (varying μ_e) with 5 agents. (N_{comm}: Average number of communications, PR: Percentage reduction in communication, W_{dist}: Average Wasserstein distances.)

$(n = 5, \mu_e = 0.3)$	W_{dist}	N_{comm}	PR
$b = 0.01$	0.2256	37301	75%
$b = 0.14$	0.2190	37072	63%
$b = 0.30$	0.1906	37146	63%
$b = 0.80$	0.1765	41783	58%

Table 2
Comparison of results from the DETULA ($\mu_e = 0.3$) with 5 agents, but varying b.

6.3 Logistic regression

We use the *a9a* dataset available at the UCI machine learning repository to evaluate the performance of the DETULA for Bayesian logistic regression models. There are 32561 observations and 123 hidden parameters in the dataset. A Laplace prior with a scale of 1 is placed on the parameters. We implemented a "mini-batch" stochastic gradient version of Algorithm 1, where at each time instant, each agent uses a batch of 10 data points randomly drawn from

$\boldsymbol{X}_i$ to compute an approximated $\nabla \log p(\boldsymbol{X}_i|\boldsymbol{w}_i(k))$ in (10).

We examined the performance of the DETULA with respect to the number of agents n for $n = \{5, 10, 25\}$. For the DETULA and the DULA, we considered the ring communication topology for each n. During each run, 80% of data were randomly chosen for training and the remaining 20% for testing as in [47]. The training data were divided into random sets of equal sizes and distributed to each agent. The same 20% testing data during each run were used to compare the performance of the CULA, the DULA and the DETULA. For the decentralized algorithms, 1000 iterations were simulated while for the CULA, 10000 iterations were simulated. A Monte-Carlo simulation of 50 runs was conducted.

In Table 3, we compare the mean accuracy on the testing data set and the average number of inter-agent communications for the DULA ($\mu_e = 0$) and the DETULA ($\mu_e = 0.5$) with $n = 5$, 10, and 25 agents. We emphasize on the percentage reduction in communication that the DETULA achieves over DULA. The mean accuracy produced by the CULA is 84.03%. From the results, the DETULA clearly performs on par with both the CULA and the DULA with only half of the communication needed. Figure 5 shows the mean and standard deviation of the accuracy for the DETULA and the DULA. The shaded regions in the figure indicates one standard deviation. From Figure 5, we see that similar to the DULA, the DETULA produces convergence results with lower standard deviations as n increases.

n	5	10	25
PA ($\mu_e = 0$)	84.17%	84.41%	84.18%
PA ($\mu_e = 0.5$)	84.11%	84.29%	84.26%
N_{comm} ($\mu_e = 0$)	2086	2090	2102
N_{comm} ($\mu_e = 0.5$)	1130	902	864
PR ($\mu_e = 0.5$)	46%	57%	59%

Table 3
Comparison of results from the DULA ($\mu_e = 0$) and the DETULA ($\mu = 0.5$) with varying number of agents n after about 1000 iterations for each case. PA: Percentage accuracy, N_{comm}: Average number of communications, PR: Percentage reduction in communication.

7 Conclusion

In this paper, we investigate a distributed Bayesian learning problem, where a group of agents collaboratively approximate a posterior distribution from locally available data sets and inter-agent communications. We introduce the first ever distributed event-triggered unadjusted Langevin algorithm and establish conditions on its time-varying step sizes and on the triggering threshold such that the agents' samples converge asymptotically to the target distribution satisfying the LSI assumption. The event-triggering mechanism for communication allows the agents to communicate periodically, thereby relaxing the constant communication requirement in existing literature. Moreover, we establish guidelines for preventing consecutive triggering on expectation while still maintaining the same rate of convergence as without event-triggering. The presented numerical experiments demonstrate that the proposed algorithm significantly reduced inter-agent communications while generating posterior samples of the same quality as the centralized algorithm.

8 Appendix

This section provides the detailed analysis of the results discussed in Section 5.

8.1 Proof of Theorem 1

From (14), we have

$$\mathbf{w}(k+1) = (\mathcal{W}(k) \otimes I_{d_w})\,\mathbf{w}(k) - \alpha_k n \widehat{\nabla E}(\mathbf{w}(k), \mathbf{X}) + \sqrt{2\alpha_k}\mathbf{v}(k) + \beta_k(\mathcal{L} \otimes I_{d_w})\mathbf{e}(k). \tag{43}$$

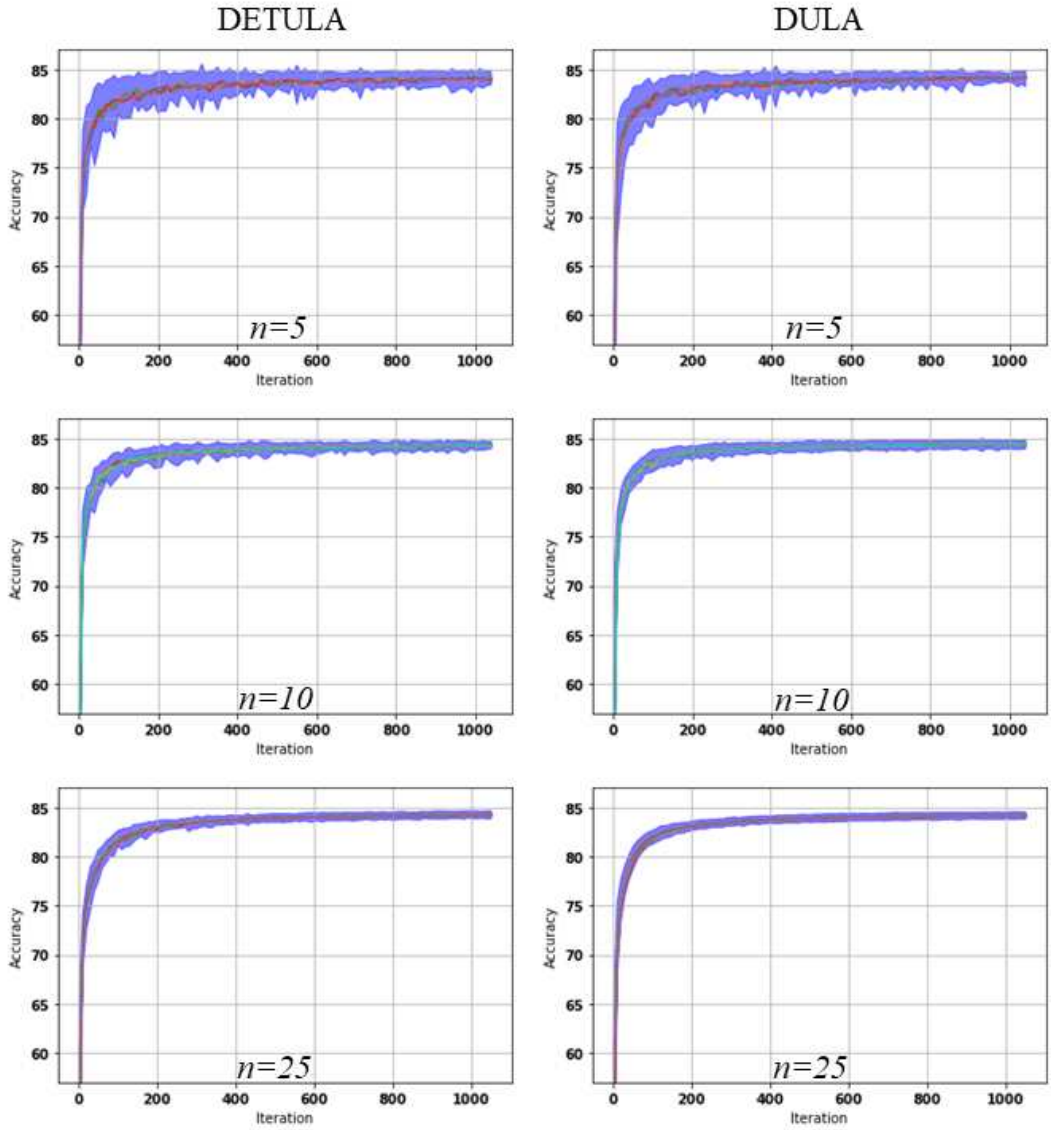

Fig. 5. Comparison between the DETULA with $\mu_e = 0.5$ and the DULA for 5, 10, and 25 agents. *Left column*: DETULA for 5, 10, and 25 agents, respectively. *Right column*: DULA for 5, 10, and 25 agents, respectively.

Define the average-consensus error as: $\tilde{\mathbf{w}}(k) = (M \otimes I_{d_w})\,\mathbf{w}(k)$, where $M = I_n - \frac{1}{n}\mathbf{1}_n\mathbf{1}_n^\top$. We then have

$$\tilde{\mathbf{w}}(k+1) = (\mathcal{W}(k) \otimes I_{d_w})\,\tilde{\mathbf{w}}(k) + \beta_k(\mathcal{L} \otimes I_{d_w})\tilde{\mathbf{e}}(k) - \alpha_k n(M \otimes I_{d_w})\widehat{\nabla E}(\mathbf{w}(k), \mathbf{X}) + \sqrt{2\alpha_k}\tilde{\mathbf{v}}(k). \tag{44}$$

where $\tilde{\mathbf{v}}(k) = (M \otimes I_{d_w})\,\mathbf{v}(k), \tilde{\mathbf{e}}(k) = (M \otimes I_{d_w})\,\mathbf{e}(k)$ and we used the following identities:

$$(A \otimes I_{d_w})(B \otimes I_{d_w}) = AB \otimes I_{d_w}, \;\; \forall A, B \in \mathbb{R}^{n\times n},$$
$$\mathcal{L}\mathbf{1}\mathbf{1}^\top = \mathbf{1}\mathbf{1}^\top\mathcal{L}.$$

Taking the norm on both sides of (44) yields

$$\|\tilde{\mathbf{w}}(k+1)\|_2 \leq \|\,((I_n - \beta_k\mathcal{L}) \otimes I_{d_w})\,\tilde{\mathbf{w}}(k)\|_2 + \beta_k\|(\mathcal{L} \otimes I_{d_w})\tilde{\mathbf{e}}(k)\|_2 + \sqrt{2\alpha_k}\|\tilde{\mathbf{v}}(k)\|_2 + \alpha_k n\|(M \otimes I_{d_w})\widehat{\nabla E}(\mathbf{w}(k), \mathbf{X})\|_2. \tag{45}$$

The individual terms on the right hand side of (45) are then analysed below. Since $\mathbf{1}_{nd_w}^\top \tilde{\mathbf{w}}_k = 0$, it follows from Lemma 4.4 in [23] that for sufficiently large k

$$\| \left((I_n - \beta_k \mathcal{L}) \otimes I_{d_w} \right) \tilde{\mathbf{w}}(k) \|_2 \;\leq\; (1 - \beta_k \lambda_2(\mathcal{L})) \, \|\tilde{\mathbf{w}}(k)\|_2, \tag{46}$$

since $0 < 1 - \beta_k \lambda_2(\mathcal{L}) < 1$ from Condition $1(ii)$.
Also, we have

$$\beta_k \|(\mathcal{L} \otimes I_{d_w}) \tilde{\mathbf{e}}_k\|_2 \;\leq\; \beta_k \lambda_n(\mathcal{L}) \, \|\tilde{\mathbf{e}}_k\|_2, \tag{47}$$

and using $\|M \otimes I_{d_w}\|_2 = 1$ yields

$$\begin{aligned}
\|(M \otimes I_{d_w}) \widehat{\nabla E}(\mathbf{w}(k), \boldsymbol{X})\|_2 &\leq \|\widehat{\nabla E}(\mathbf{w}(k), \boldsymbol{X})\|_2 \\
&\leq \|\widehat{\nabla E}(\mathbf{w}(k), \boldsymbol{X}) - \widehat{\nabla E}(\mathbf{1}_n \otimes \bar{\boldsymbol{w}}(k), \boldsymbol{X}) + \widehat{\nabla E}(\mathbf{1}_n \otimes \bar{\boldsymbol{w}}(k), \boldsymbol{X}) - \widehat{\nabla E}(\mathbf{w}^*, \boldsymbol{X})\|_2, \\
&\leq \|\widehat{\nabla E}(\mathbf{w}(k), \boldsymbol{X}) - \widehat{\nabla E}(\mathbf{1}_n \otimes \bar{\boldsymbol{w}}(k), \boldsymbol{X})\|_2 + \|\widehat{\nabla E}(\mathbf{1}_n \otimes \bar{\boldsymbol{w}}(k), \boldsymbol{X}) - \widehat{\nabla E}(\mathbf{w}^*, \boldsymbol{X})\|_2, \\
&\leq L\|\tilde{\mathbf{w}}(k)\|_2 + L\|\mathbf{1}_n \otimes \bar{\boldsymbol{w}}(k)\|_2 + L\|\mathbf{w}^*\|_2.
\end{aligned} \tag{48}$$

where $\lambda_2(\cdot)$ and $\lambda_n(\cdot)$ denote the second smallest and the largest eigenvalue respectively, and $\mathbf{w}^*$ is a local minima of the function $E(\cdot)$, hence the gradient $\widehat{\nabla E}(\mathbf{w}^*) = 0$. Thereon, combining the results from (46), (47) and (48) into (45) yields

$$\begin{aligned}
\|\tilde{\mathbf{w}}(k+1)\|_2 \leq (1 - \beta_k \lambda_2(\mathcal{L}) + \alpha_k n L)\|\tilde{\mathbf{w}}(k)\|_2 + \alpha_k n L \|\mathbf{w}^*\|_2 + \alpha_k n L \|\mathbf{1}_n \otimes \bar{\boldsymbol{w}}(k)\|_2 + \sqrt{2\alpha_k}\|\mathbf{v}(k)\|_2 \\
+ \beta_k \lambda_n(\mathcal{L}) \, \|\mathbf{e}(k)\|_2,
\end{aligned} \tag{49}$$

Define $\sigma_k = \frac{b\lambda_2(\mathcal{L}) - anL}{(k+1)^{\delta_1}} < \beta_k \lambda_2(\mathcal{L}) - \alpha_k n L$, which then results in

$$\|\tilde{\mathbf{w}}(k+1)\|_2 \leq (1 - \sigma_k)\|\tilde{\mathbf{w}}(k)\|_2 + \alpha_k n L \|\mathbf{w}^*\|_2 + \alpha_k n L \|\mathbf{1}_n \otimes \bar{\boldsymbol{w}}(k)\|_2 + \sqrt{2\alpha_k}\|\mathbf{v}(k)\|_2 + \beta_k \lambda_n(\mathcal{L}) \, \|\mathbf{e}(k)\|_2. \tag{50}$$

For the stability of the algorithm, we need $\sigma_k \in (0, 1)$, $\forall\, k \geq 0$ which leads to $b\lambda_2(\mathcal{L}) - anL < 1$ and $b\lambda_2(\mathcal{L}) > anL$. We next introduce the inequality

$$(x + y)^2 \;\leq\; (1 + \theta)\, x^2 + \left(1 + \frac{1}{\theta}\right) y^2, \tag{51}$$

$\forall\, x, y \in \mathbb{R}$ and $\theta > 0$. Making use of (51) multiple times with $\theta = (1 - \sigma_k)^{-\frac{1}{3}} - 1 \;> 0$ and in conjunction with the relations

$$\begin{aligned}
&(1 - \sigma_k)^{-p} \leq (1 - \sigma_0)^{-p}, \;\; \forall p \geq 0, \\
&\frac{1}{1 - (1 - \sigma_k)^{\frac{1}{2}}} \leq \frac{3}{\sigma_k}.
\end{aligned}$$

gives

$$\begin{aligned}
\|\tilde{\mathbf{w}}(k+1)\|_2^2 \leq (1 - \sigma_k)\|\tilde{\mathbf{w}}(k)\|_2^2 + \frac{3(1 - \sigma_0)^{-\frac{2}{3}} \beta_k^2 \lambda_n^2(\mathcal{L})}{\sigma_k} \|\mathbf{e}(k)\|_2^2 + \frac{9\alpha_k^2 n^2 L^2}{\sigma_k^2} \|\mathbf{w}^*\|_2^2 \\
+ \frac{3\alpha_k^2 n^2 L^2 (1 - \sigma_0)^{-\frac{1}{3}}}{\sigma_k} \|\mathbf{1}_n \otimes \bar{\boldsymbol{w}}(k)\|_2^2 + \frac{6\alpha_k (1 - \sigma_0)^{-\frac{1}{3}}}{\sigma_k} \|\mathbf{v}(k)\|_2^2.
\end{aligned} \tag{52}$$

Now, taking the conditional expectation $\mathbb{E}\left[\,\cdot\,|\mathcal{F}_k\right]$ and thereafter total expectation $\mathbb{E}[\,\cdot\,]$ yields

$$\begin{aligned}\mathbb{E}\left[\,\|\tilde{\mathbf{w}}(k+1)\|_2^2\right] \leq (1-\sigma_k)\mathbb{E}\left[\|\tilde{\mathbf{w}}(k)\|_2^2\right] + \frac{3n\mu_e(1-\sigma_0)^{-\frac{2}{3}}b^2\lambda_n^2(\mathcal{L})}{\sigma_0(k+1)^{\delta_1+\delta_3}} + \frac{9a^2n^2L^2C^*}{\sigma_0^2(k+1)^{2\delta_2-2\delta_1}} \\ + \frac{3a^2n^2L^2(1-\sigma_0)^{-\frac{1}{3}}}{\sigma_0(k+1)^{2\delta_2-\delta_1}}\mathbb{E}\left[\|\mathbf{1}_n\otimes\bar{\boldsymbol{w}}(k)\|_2^2\right] + \frac{6an^2d_w(1-\sigma_0)^{-\frac{1}{3}}}{\sigma_0(k+1)^{\delta_2-\delta_1}},\end{aligned} \tag{53}$$

where we used the following relations:

$$\mathbb{E}\left[\|\mathbf{v}(k)\|_2^2\right] = n^2d_w, \tag{54}$$

$$\mathbb{E}\left[\|\mathbf{w}^*\|_2^2\right] = C^* \leq \infty. \tag{55}$$

and the relation from (20) in Condition 2. From (53), we see the need to have a bound on $\mathbb{E}\left[\|\boldsymbol{w}(k)\|_2^2\right]$, i.e., the expectation of the average of the samples $\boldsymbol{w}_i(k)$ be bounded for all $i = \{1, 2, \ldots, n\}$. To establish that bound, we resort to an induction approach, Assuming that $\mathbb{E}\left[\|\bar{\boldsymbol{w}}(\ell)\|_2^2\right] \leq C_{\bar{w}}\,,\ \forall\, 0 \leq \ell \leq k$, we seek to show that $\mathbb{E}\left[\|\bar{\boldsymbol{w}}(k+1)\|_2^2\right] \leq C_{\bar{w}}$ for any $k \geq 0$, thus establishing $C_{\bar{w}}$ as the desired bound. Towards this end, substituting $\mathbb{E}\left[\|\bar{\boldsymbol{w}}(k)\|_2^2\right] \leq C_{\bar{w}}$ in (53) gives

$$\mathbb{E}\left[\|\tilde{\mathbf{w}}(k+1)\|_2^2\right] \leq (1-\sigma_k)\mathbb{E}\left[\|\tilde{\mathbf{w}}_k\|_2^2\right] + \frac{d\mu_e}{(k+1)^{\delta_1+\delta_3}} + \frac{d'C_{\bar{w}}}{(k+1)^{2\delta_2-\delta_1}} + \frac{d''}{(k+1)^{\delta_2-\delta_1}}, \tag{56}$$

where

$$d = \frac{3n(1-\sigma_0)^{-\frac{2}{3}}b^2\lambda_n^2(\mathcal{L})}{\sigma_0}, \tag{57}$$

$$d' = \frac{3a^2n^3L^2(1-\sigma_0)^{-\frac{1}{3}}}{\sigma_0}, \tag{58}$$

$$d'' = \frac{6(1-\sigma_0)^{-\frac{1}{3}}an^2d_w}{\sigma_0} + \frac{9a^2n^2L^2C^*}{\sigma_0^2}. \tag{59}$$

From an extension of Lemma S4 in [39], we have

$$\mathbb{E}[\|\tilde{\mathbf{w}}(k+1)\|_2^2] \leq \frac{W_5}{\exp(W_1(k+1)^{1-\delta_1})} + \frac{W_2\mu_e}{(k+1)^{\delta_3}} + \frac{W_3C_{\bar{w}}}{(k+1)^{2\delta_2-2\delta_1}} + \frac{W_4}{(k+1)^{\delta_2-2\delta_1}}, \tag{60}$$

where

$$W_1 = \frac{\sigma_0}{1-\delta_1}, \tag{61}$$

$$W_2 = \frac{d(\delta_1+\delta_3)}{\sigma_0\delta_1}\exp(W_1 2^{1-\delta_1}), \tag{62}$$

$$W_3 = \frac{d'(2\delta_2-\delta_1)}{\sigma_0\delta_1}\exp(W_1 2^{1-\delta_1}), \tag{63}$$

$$W_4 = \frac{d''(\delta_2-\delta_1)}{\sigma_0\delta_1}\exp(W_1 2^{1-\delta_1}), \tag{64}$$

$$\begin{aligned}W_5 = \exp(W_1)\Bigg[\mathbb{E}[\|\tilde{\mathbf{w}}_0\|_2^2] + \sum_{\ell=0}^{\bar{k}_1}\left(\frac{1}{(1-\sigma_0)^\ell}\frac{d}{(\ell+1)^{\delta_1+\delta_3}}\right) + \sum_{\ell=0}^{\bar{k}_2}\left(\frac{1}{(1-\sigma_0)^\ell}\frac{d'C_{\bar{w}}}{(\ell+1)^{2\delta_2-\delta_1}}\right) \\ + \sum_{\ell=0}^{\bar{k}_3}\left(\frac{1}{(1-\sigma_0)^\ell}\frac{d''}{(\ell+1)^{\delta_2-\delta_1}}\right)\Bigg],\end{aligned} \tag{65}$$

wherein $\bar{k}_1 = \left\lceil \left(\frac{\delta_1+\delta_3}{\sigma_0}\right)^{\frac{1}{1-\delta_1}} \right\rceil$, $\bar{k}_2 = \left\lceil \left(\frac{2\delta_2-\delta_1}{\sigma_0}\right)^{\frac{1}{1-\delta_1}} \right\rceil$ and $\bar{k}_3 = \left\lceil \left(\frac{\delta_2-\delta_1}{\sigma_0}\right)^{\frac{1}{1-\delta_1}} \right\rceil$. The second term on the right hand side of (60) sheds light on the rationale of choosing $e^m(k)$ as in (20). Next, following the convergence analysis using Fokker Plank equation in (S153) of [39], we have

$$\begin{aligned} F(p_{t_{k+1}}(\bar{\boldsymbol{w}})) &\leq \exp(-\rho_U \alpha_k) F(p_{t_k}(\bar{\boldsymbol{w}})) + 2\alpha_k^2 \bar{L}^2 d_w + 2\alpha_k^3 \bar{L}^4 \mathbb{E}_{p(\bar{\boldsymbol{w}}(t_k))} \|\bar{\boldsymbol{w}}(t_k)\|_2^2 + (2\alpha_k^3 \bar{L} L^2 + L^2 \alpha_k) \mathbb{E}_{p(\tilde{\omega}(t_k))} \|\tilde{\omega}(t_k)\|_2^2 \\ &\leq \exp(-\rho_U \alpha_k) F(p_{t_k}(\bar{\boldsymbol{w}})) + 2\alpha_k^3 \bar{L}^4 C_{\bar{w}} + g_k, \\ &\leq \exp\left(-\rho_U \sum_{i=0}^{k} \alpha_i\right) F(p_{t_0}(\bar{\boldsymbol{w}})) + 2\bar{L}^4 C_{\bar{w}} \sum_{i=0}^{k} \alpha_i^3 + \sum_{i=0}^{k} g_i, \end{aligned} \tag{66}$$

where

$$g_k = 2\alpha_k^2 \bar{L}^2 d_w + (2\alpha_k^3 \bar{L}^2 L^2 + L^2 \alpha_k) \mathbb{E}_{p(\tilde{\omega}(t_k))} \|\tilde{\omega}(t_k)\|_2^2.$$

Since $\|\tilde{\omega}(t_k)\|_2$ is the continuous-time counterpart of $\|\tilde{\mathbf{w}}(k)\|_2$, they both have the same decay rate at the discrete time steps. Thus, substituting $\mathbb{E}\left[\|\tilde{\mathbf{w}}(k)\|_2^2\right]$ from (60), we get

$$\begin{aligned} g_k &\leq (2a^3 \bar{L}^2 L^2 + aL^2) \left[\frac{W_5}{\exp(W_1 (k+1)^{1-\delta_1})(k+1)^{\delta_2}} + \frac{W_2 \mu_e}{(k+1)^{\delta_2+\delta_3}} + \frac{W_4}{(k+1)^{2\delta_2-2\delta_1}} \right] + \frac{2a^2 \bar{L}^2 d_w}{(k+1)^{2\delta_2}} \\ &\quad + \frac{W_3'}{(k+1)^{3\delta_2-2\delta_1}} C_{\bar{w}}, \end{aligned} \tag{67}$$

$$\leq \bar{g}_k + \frac{W_3'}{(k+1)^{3\delta_2-2\delta_1}} C_{\bar{w}}, \tag{68}$$

where $\{\bar{g}_k\} \sim \mathcal{O}\left(\frac{1}{k^{\delta_2+\delta_3}}\right) + \mathcal{O}\left(\frac{1}{k^{2\delta_2-2\delta_1}}\right)$ and $W_3' = (2a^3 \bar{L}^2 L^2 + aL^2) W_3$. Let $\delta_2 + \delta_3 > 1$ and $2\delta_2 - 2\delta_1 > 1$, then $\{\bar{g}_k\}$ is a summable sequence, i.e., there exists an $0 < s < \infty$ such that $\sum_{i=0}^{\infty} \bar{g}_k = s$. Thus,

$$\begin{aligned} \sum_{i=0}^{k} g_i &\leq \sum_{i=0}^{k} \bar{g}_i + \left(\sum_{i=0}^{k} \frac{W_3'}{(k+1)^{3\delta_2-2\delta_1}} \right) C_{\bar{w}}, \leq \sum_{i=0}^{\infty} \bar{g}_i + \left(1 + \int_0^{\infty} \frac{dt}{(t+1)^{3\delta_2-2\delta_1}} \right) W_3' C_{\bar{w}}, \\ &\leq s + \left(\frac{3\delta_2 - 2\delta_1}{3\delta_2 - 2\delta_1 - 1} \right) W_3' C_{\bar{w}}. \end{aligned} \tag{69}$$

Also,

$$\sum_{i=0}^{k} \alpha_i^3 \leq a^3 + \int_0^{\infty} \frac{a^3}{(t+1)^{3\delta_2}} dt = \frac{3\delta_2 a^3}{3\delta_2 - 1}. \tag{70}$$

Substituting (69) and (70) in (66) results in

$$F(p_{t_{k+1}}(\bar{\boldsymbol{w}})) \leq s' + \left(\frac{6n^4 a^3 L^4 \delta_2}{3\delta_2 - 1} + \frac{3\delta_2 - 2\delta_1}{3\delta_2 - 2\delta_1 - 1} W_3' \right) C_{\bar{w}}, \tag{71}$$

where $s' = s + F(p_{t_0}(\bar{\boldsymbol{w}}))$.
Next, we establish the relation between the expected value $\mathbb{E}\left[\|\bar{\boldsymbol{w}}(t_{k+1})\|_2^2\right]$ and the KL-divergence $F(p_{t_k+1}(\bar{\boldsymbol{w}}))$ of the average of the samples $\bar{\boldsymbol{w}}$. To this end, following the proof of Lemma 6 from [31], we couple $\bar{\boldsymbol{w}}^*$ optimally with

$\bar{\boldsymbol{w}}(t) \sim p_t(\bar{\boldsymbol{w}})$, i.e., $(\bar{\boldsymbol{w}}(t), \bar{\boldsymbol{w}}^*) \sim \gamma \in \Gamma_{\text{opt}}(p_t(\bar{\boldsymbol{w}}), p^*)$. This gives us

$$\begin{aligned}\mathbb{E}_{\bar{\boldsymbol{w}}(t_{k+1})\sim p_{t_{k+1}}}\left[\|\bar{\boldsymbol{w}}(t_{k+1})\|_2^2\right] &= \mathbb{E}_{(\bar{\boldsymbol{w}}(t_{k+1}),\bar{\boldsymbol{w}}^*)\sim\gamma}\left[\|\bar{\boldsymbol{w}}^* + \bar{\boldsymbol{w}}(t_{k+1}) - \bar{\boldsymbol{w}}^*\|_2^2\right], \\ &\leq 2\mathbb{E}_{\bar{\boldsymbol{w}}^*\sim p^*}\|\bar{\boldsymbol{w}}^*\|_2^2 + 2\mathbb{E}_{(\bar{\boldsymbol{w}}(t_{k+1}),\bar{\boldsymbol{w}}^*)\sim\gamma}\|\bar{\boldsymbol{w}}(t_{k+1}) - \bar{\boldsymbol{w}}^*\|_2^2, \\ &\leq 2c_1 + 2\mathcal{W}_2^2(p_{t_{k+1}}(\bar{\boldsymbol{w}}), p^*), \\ &\leq 2c_1 + \frac{4}{\rho_U}F(p_{t_{k+1}}(\bar{\boldsymbol{w}})),\end{aligned} \tag{72}$$

where $\mathcal{W}_2(\,\cdot\,,\,\cdot\,)$ denotes the Wasserstein metric between two distributions and and the relation in the last inequality comes from [38, Theorem 1]. Using (71) in (72), we obtain

$$\mathbb{E}\left[\|\bar{\boldsymbol{w}}(t_{k+1})\|_2^2\right] \leq 2c_1 + \frac{4}{\rho_U}\left[s' + \left(\frac{6n^4a^3L^4\delta_2}{3\delta_2 - 1} + \frac{3\delta_2 - 2\delta_1}{3\delta_2 - 2\delta_1 - 1}W_3'\right)C_{\bar{w}}\right]. \tag{73}$$

To establish a uniform bound on $\mathbb{E}\left[\|\bar{\boldsymbol{w}}(t_k)\|_2^2\right]$ for all $k \geq 0$ via induction, we need $\mathbb{E}\left[\|\bar{\boldsymbol{w}}(t_{k+1})\|_2^2\right] \leq C_{\bar{w}}$, i.e.,

$$2c_1 + \frac{4}{\rho_U}\left[s' + \left(\frac{6n^4a^3L^4\delta_2}{3\delta_2 - 1} + \frac{3\delta_2 - 2\delta_1}{3\delta_2 - 2\delta_1 - 1}W_3'\right)C_{\bar{w}}\right] \leq C_{\bar{w}}, \tag{74}$$

which results in

$$C_{\bar{w}} \geq \frac{2c_1 + \frac{4}{\rho_U}s'}{1 - \frac{4}{\rho_U}\left(\frac{6n^4a^3L^4\delta_2}{3\delta_2 - 1} + \frac{3\delta_2 - 2\delta_1}{3\delta_2 - 2\delta_1 - 1}W_3'\right)}. \tag{75}$$

For $0 < C_{\bar{w}} < \infty$ to exist as given by (75), we further have to ensure that

$$1 - \frac{4}{\rho_U}\left(\frac{6n^4a^3L^4\delta_2}{3\delta_2 - 1} + \frac{3\delta_2 - 2\delta_1}{3\delta_2 - 2\delta_1 - 1}W_3'\right) > 0,$$

i.e.,

$$1 - \frac{4}{\rho_U}\left(\frac{6n^4a^3L^4\delta_2}{3\delta_2 - 1} + \frac{3\delta_2 - 2\delta_1}{3\delta_2 - 2\delta_1 - 1}\times\ (2a^3\bar{L}^2L^2 + aL^2)\frac{d'(2\delta_2 - \delta_1)}{\sigma_0\delta_1}\exp(W_1 2^{1-\delta_1})\right) > 0, \tag{76}$$

where $W_1 = \frac{\sigma_0}{1-\delta_1} < \frac{b\lambda_2(\mathcal{L})}{1-\delta_1}$. Thus, (76) is guaranteed when

$$1 - \frac{4}{\rho_U}\left(\frac{6n^4a^3L^4\delta_2}{3\delta_2 - 1} + \frac{3\delta_2 - 2\delta_1}{3\delta_2 - 2\delta_1 - 1}(2a^3\bar{L}^2L^2 + aL^2)\frac{d'(2\delta_2 - \delta_1)}{\sigma_0\delta_1}\exp\left(\frac{b\lambda_2(\mathcal{L})}{1-\delta_1}2^{1-\delta_1}\right)\right) > 0. \tag{77}$$

Further, substituting d' from (58) in (76) results in the following polynomial condition in a:

$$\phi(a) = \theta_5a^5 + \theta_4a^4 + \theta_3a^3 + \theta_2a^2 + \theta_1a - \frac{\rho_U b^2\lambda_2^2(\mathcal{L})}{4} < 0, \tag{78}$$

where

$$\theta_5 = \frac{6n^6L^6\delta_2}{3\delta_2-1} + 6n^3\bar{L}^2L^4(1-b\lambda_2(\mathcal{L}))^{-\frac{1}{3}}\frac{(3\delta_2-2\delta_1)(2\delta_2-\delta_1)}{\delta_1(3\delta_2-2\delta_1-1)}\exp(W_1 2^{1-\delta_1}) > 0,$$
$$\theta_4 = -\frac{12bn^5L^5\lambda_2(\mathcal{L})\delta_2}{3\delta_2-1} < 0,$$
$$\theta_3 = \frac{6n^4L^4\delta_2 b^2\lambda_2^2(\mathcal{L})}{3\delta_2-1} + 3n^3L^2(1-b\lambda_2(\mathcal{L}))^{-\frac{1}{3}}\frac{(3\delta_2-2\delta_1)(2\delta_2-\delta_1)}{\delta_1(3\delta_2-2\delta_1-1)}\exp(W_1 2^{1-\delta_1}) > 0,$$
$$\theta_2 = -\frac{\rho_U n^2L^2}{4} < 0,$$
$$\theta_1 = \frac{\rho_U nbL\lambda_2(\mathcal{L})}{2} > 0.$$

Since, $\phi(0) = -\frac{\rho_U b^2\lambda_2^2(\mathcal{L})}{4} < 0$ and $\phi(a)$ is a continuous function of a, there exists $\bar{a} \in (0,1]$ such that $\forall\, a \in (0,\bar{a})$ we have $\phi(a) < 0$. Hence, we shall always be able to find values of a such that $\phi(a) < 0$ is satisfied and, thus, existence of $C_{\bar{w}}$ is guaranteed. This concludes

$$\mathbb{E}\left[\|\bar{\boldsymbol{w}}(k)\|_2^2\right] \leq C_{\bar{w}}, \quad \forall k \geq 0, \tag{79}$$

where the lower bound on $C_{\bar{w}}$ is given by (75). Finally, combining (79) with (53) yields

$$\begin{aligned}\mathbb{E}\left[\,\|\tilde{\mathbf{w}}(k+1)\|_2^2\right] \leq (1-\sigma_k)\mathbb{E}\left[\|\tilde{\mathbf{w}}(k)\|_2^2\right] + \frac{3n\mu_e(1-\sigma_0)^{-\frac{2}{3}}b^2\lambda_n^2(\mathcal{L})}{\sigma_0(k+1)^{\delta_1+\delta_3}} + \frac{9a^2n^2L^2C^*}{\sigma_0^2(k+1)^{2\delta_2-2\delta_1}} \\ + \frac{3a^2n^3L^2C_{\bar{w}}(1-\sigma_0)^{-\frac{1}{3}}}{\sigma_0(k+1)^{2\delta_2-\delta_1}} + \frac{6an^2d_w(1-\sigma_0)^{-\frac{1}{3}}}{\sigma_0(k+1)^{\delta_2-\delta_1}},\end{aligned} \tag{80}$$

which from the extension of Lemma 3 leads to $\forall\, k \geq 0$

$$\mathbb{E}[\|\tilde{\mathbf{w}}(k+1)\|_2^2] \leq \frac{W_5}{\exp(W_1(k+1)^{1-\delta_1})} + \frac{W_2\mu_e}{(k+1)^{\delta_3}} + \frac{W_3C_{\bar{w}}}{(k+1)^{2\delta_2-2\delta_1}} + \frac{W_4}{(k+1)^{\delta_2-2\delta_1}}, \tag{81}$$

where W_1, W_2, W_3, W_4 and W_5 are given in (61) – (65), respectively. Furthermore, it can be easily seen from (81) that for large k, $\mathbb{E}\left[\|\tilde{\mathbf{w}}(k+1)\|_2^2\right] \sim \mathcal{O}\left(\frac{1}{(k+1)^{\delta_p}}\right)$, where $\delta_p = \min\{\delta_3, \delta_2 - 2\delta_1\}$. This concludes the proof of Theorem 1. ■

8.2 Proof of Theorem 2

In order to establish convergence we start with the analysis of the continuous time evolution of the KL divergence of the distribution of the mean of the samples from all the agents. Thus, from (27) we note

$$\dot{F}(p_t(\bar{\boldsymbol{w}})) = \int \left(\log\left(\frac{p_t(\bar{\boldsymbol{w}})}{p^*(\boldsymbol{w})}\right)\right)\frac{\partial p_t(\bar{\boldsymbol{w}})}{\partial t}d\bar{\boldsymbol{w}}. \tag{82}$$

An expression of $\frac{\partial p_t(\bar{\boldsymbol{w}})}{\partial t}$ can be obtained from the Fokker-Plank (FP) equation (see 4.1 in [40]). Additionally, making use of the LSI assumption from Assumption 3, the following result is obtained.

$$\dot{F}(p_t(\bar{\boldsymbol{w}})) \leq -\rho_U F(p_t(\bar{\boldsymbol{w}})) + 2\alpha_k^2\bar{L}^4\mathbb{E}_{p(\bar{\boldsymbol{w}}(t_k))}\|\bar{\boldsymbol{w}}(t_k)\|_2^2 + (2\alpha_k^2\bar{L}^2L^2 + L^2)\mathbb{E}_{p(\tilde{\omega}(t_k))}\|\tilde{\omega}(t_k)\|_2^2 + 2\alpha_k\bar{L}^2d_w. \tag{83}$$

The detailed derivation of (83) can be found in Section S4 of [39]. Thereafter, integrating (83) from t_k to t_{k+1}, where k indicates the discrete time steps, and using $\frac{1-\exp(-\rho_U(t_{k+1}-t_k))}{\rho_U} \leq t_{k+1} - t_k = \alpha_k$ yields

$$\begin{aligned}F(p_{t_{t+1}}(\bar{\boldsymbol{w}})) \leq \exp(-\rho_U\alpha_k)F(p_{t_k}(\bar{\boldsymbol{w}})) + 2\alpha_k^2\bar{L}^2d_w + 2\alpha_k^3\bar{L}^4\mathbb{E}_{p(\bar{\boldsymbol{w}}(t_k))}\|\bar{\boldsymbol{w}}(t_k)\|_2^2 \\ + (2\alpha_k^3\bar{L}L^2 + L^2\alpha_k)\mathbb{E}_{p(\omega(\tilde{t}_k))}\|\tilde{\omega}(t_k)\|_2^2.\end{aligned} \tag{84}$$

Let

$$Z_k = \frac{W_5(2\alpha_k^3\bar{L}^2L^2 + L^2\alpha_k)}{\exp(W_1k^{1-\delta_1})}, \tag{85}$$

$$\xi_k = 2\alpha_k^3\bar{L}^4C_{\bar{w}} + (2\alpha_k^3\bar{L}^2L^2 + L^2\alpha_k)\left(\frac{W_2\mu_e}{k^{\delta_3}} + \frac{W_3C_{\bar{w}}}{k^{2\delta_2-2\delta_1}} + \frac{W_4}{k^{\delta_2-2\delta_1}}\right) + 2\alpha_k^2\bar{L}^2d_w + Z_k, \tag{86}$$

$$\theta_k = \frac{W_5(2a^3\bar{L}^2L^2 + L^2a)}{\exp(W_1k^{1-\delta_1})} = \frac{\bar{W}_5}{\exp(W_1k^{1-\delta_1})}, \tag{87}$$

where $\bar{W}_5 = (2a^3\bar{L}^2L^2 + L^2a)W_5$ and $Z_k \leq \theta_k$. Thus, (84) becomes

$$\begin{aligned} F(p_{t_{k+1}}(\bar{\boldsymbol{w}})) &\leq \exp(-\rho_U\alpha_k)F(p_{t_k}(\bar{\boldsymbol{w}})) + \xi_k, \\ &\leq \exp\left(-\rho_U\sum_{\ell=0}^{k}\alpha_k\right)F(p_{t_0}(\bar{\boldsymbol{w}})) \\ &\quad + \sum_{\ell=0}^{k}\xi_k\exp\left(-\rho_U\sum_{i=\ell+1}^{k}\alpha_i\right), \end{aligned} \tag{88}$$

Now,

$$\begin{aligned} \xi_k &\leq \frac{2a^3\bar{L}^4C_{\bar{w}}}{k^{3\delta_2}} + \left(\frac{2a^3\bar{L}^2L^2}{k^{3\delta_2}} + \frac{L^2a}{k^{\delta_2}}\right)\left(\frac{W_2\mu_e}{k^{\delta_3}} + \frac{W_3C_{\bar{w}}}{k^{2\delta_2-2\delta_1}} + \frac{W_4}{k^{\delta_2-2\delta_1}}\right) + \frac{2a^2\bar{L}^2d_w}{k^{2\delta_2}} + \theta_k, \\ &\leq \frac{2a^3\bar{L}^4C_{\bar{w}}}{k^{3\delta_2}} + \frac{2a^3\bar{L}^2L^2W_2\mu_e}{k^{3\delta_2+\delta_3}} + \frac{L^2aW_2\mu_e}{k^{\delta_2+\delta_3}} + \frac{2a^3\bar{L}^2L^2W_3C_{\bar{w}}}{k^{5\delta_2-2\delta_1}} + \frac{L^2aW_3C_{\bar{w}}}{k^{3\delta_2-2\delta_1}} + \frac{2a^3\bar{L}^2L^2W_4}{k^{4\delta_2-2\delta_1}} \\ &\quad + \frac{L^2aW_4}{k^{2\delta_2-2\delta_1}} + \frac{2a^2\bar{L}^2d_w}{k^{2\delta_2}} + \theta_k, \\ &\leq \frac{C_\xi}{k^{2\delta_2-2\delta_1}} + \frac{L^2aW_2\mu_e}{k^{\delta_2+\delta_3}} + \theta_k, \end{aligned} \tag{89}$$

where $C_\xi = 2a^3\bar{L}^4C_{\bar{w}} + 2a^3\bar{L}^2L^2W_2\mu_e + 2a^3\bar{L}^2L^2W_3C_{\bar{w}} + L^2aW_3C_{\bar{w}} + 2a^3\bar{L}^2L^2W_4 + L^2aW_4 + 2a^2\bar{L}^2d_w$.

Therefore,

$$\begin{aligned} \sum_{\ell=0}^{k}\xi_\ell\exp\left(-\rho_U\sum_{i=\ell+1}^{k}\alpha_i\right) &\leq \underbrace{\sum_{\ell=0}^{k}\frac{C_\xi}{\ell^{2\delta_2-2\delta_1}}\exp\left(-\rho_U\sum_{i=\ell+1}^{k}\alpha_i\right)}_{T_1} + \underbrace{\sum_{\ell=0}^{k}\frac{L^2aW_2\mu_e}{\ell^{\delta_2+\delta_3}}\exp\left(-\rho_U\sum_{i=\ell+1}^{k}\alpha_i\right)}_{T_2} \\ &\quad + \underbrace{\sum_{\ell=0}^{k}\theta_\ell\exp\left(-\rho_U\sum_{i=\ell+1}^{k}\alpha_i\right)}_{T_3} \end{aligned} \tag{90}$$

We will next analyze the terms T_1, T_2 and T_3 one at a time. For T_1, we have for some $\bar{k}_1 \in (0,k)$,

$$\begin{aligned}
T_1 &= \sum_{\ell=0}^{k} \frac{C_\xi}{\ell^{2\delta_2-2\delta_1}} \exp\left(-\rho_U \sum_{i=\ell+1}^{k} \alpha_i \right) \\
&= \sum_{\ell=0}^{\bar{k}_1} \frac{C_\xi \exp\left(\rho_U \sum_{i=0}^{\ell} \alpha_i\right)}{\ell^{2\delta_2-2\delta_1}} \exp\left(-\rho_U \sum_{i=0}^{k} \alpha_i \right) + \sum_{\ell=\bar{k}_1+1}^{k} \frac{C_\xi}{\ell^{2\delta_2-2\delta_1}} \exp\left(-\rho_U \sum_{i=\ell+1}^{k} \alpha_i \right), \\
&\leq \sum_{\ell=0}^{\bar{k}_1} \frac{C_\xi \exp(a\rho_U) \exp\left(\frac{a\rho_U}{1-\delta_2}\ell^{1-\delta_2}\right)}{\ell^{2\delta_2-2\delta_1}} \exp\left(-\rho_U \sum_{i=0}^{k} \alpha_i \right) + C_\xi \exp\left(-\frac{\rho_U a}{1-\delta_2}(k+1)^{1-\delta_2} \right) \times \\
&\quad \sum_{\ell=\bar{k}_1+1}^{k} \frac{\exp\left(\frac{a\rho_U}{1-\delta_2}(\ell+2)^{1-\delta_2}\right)}{\ell^{2\delta_2-2\delta_1}}, \qquad (91) \\
&\leq \sum_{\ell=0}^{\bar{k}_1} \frac{C_\xi \exp(a\rho_U) \exp\left(\frac{a\rho_U}{1-\delta_2}\ell^{1-\delta_2}\right)}{\ell^{2\delta_2-2\delta_1}} \exp\left(-\rho_U \sum_{i=0}^{k} \alpha_i \right) + C_\xi \exp\left(-\frac{\rho_U a}{1-\delta_2}(k+1)^{1-\delta_2} + \frac{a\rho_U}{1-\delta_2}2^{1-\delta_2} \right) \times \\
&\quad \sum_{\ell=\bar{k}_1+1}^{k} \frac{\exp\left(\frac{a\rho_U}{1-\delta_2}\ell^{1-\delta_2}\right)}{\ell^{2\delta_2-2\delta_1}}. \qquad (92)
\end{aligned}$$

We used the result $\sum_{i=\ell+1}^{k} \alpha_i \geq \int_{\ell+1}^{k} \frac{a}{(t+1)^{\delta_2}} dt = \frac{a((k+1)^{1-\delta_2}-(\ell+2)^{1-\delta_2})}{1-\delta_2}$ in (91). Now, $\frac{\exp\left(\frac{a\rho_U}{1-\delta_2}\ell^{1-\delta_2}\right)}{\ell^{2\delta_2-2\delta_1}}$ in the last term of (92) has a local minima at $\left\lceil \left(\frac{2\delta_2-2\delta_1}{a\rho_U}\right)^{\frac{1}{1-\delta_2}} \right\rceil$. So, we choose $\bar{k}_1 = \left\lceil \left(\frac{2\delta_2-2\delta_1}{a\rho_U}\right)^{\frac{1}{1-\delta_2}} \right\rceil$. Thus, we have

$$\begin{aligned}
\sum_{\ell=\bar{k}_1+1}^{k} \frac{\exp\left(\frac{a\rho_U}{1-\delta_2}\ell^{1-\delta_2}\right)}{\ell^{2\delta_2-2\delta_1}} &\leq \int_{\bar{k}_1}^{k+1} \frac{\exp\left(\frac{a\rho_U}{1-\delta_2}t^{1-\delta_2}\right)}{t^{2\delta_2-2\delta_1}} dt \leq \frac{2\delta_2-2\delta_1}{a\rho_U\delta_2} \left(\frac{\exp\left(\frac{a\rho_U}{1-\delta_2}(k+1)^{1-\delta_2}\right)}{(k+1)^{\delta_2-2\delta_1}} - \frac{\exp\left(\frac{a\rho_U}{1-\delta_2}\bar{k}^{1-\delta_2}\right)}{\bar{k}_1^{\delta_2-2\delta_1}} \right) \\
&\leq \frac{2\delta_2-2\delta_1}{a\rho_U\delta_2} \frac{\exp\left(\frac{a\rho_U}{1-\delta_2}(k+1)^{1-\delta_2}\right)}{(k+1)^{\delta_2-2\delta_1}}. \qquad (93)
\end{aligned}$$

Combining (92) with (93) results in

$$T_1 \leq \sum_{\ell=0}^{\bar{k}_1} \frac{C_\xi \exp(a\rho_U) \exp\left(\frac{a\rho_U}{1-\delta_2}\ell^{1-\delta_2}\right)}{\ell^{2\delta_2-2\delta_1}} \exp\left(-\rho_U \sum_{i=0}^{k} \alpha_i \right) + \frac{2C_\xi(\delta_2-\delta_1)}{a\rho_U\delta_2} \frac{\exp\left(\frac{a\rho_U}{1-\delta_2}2^{1-\delta_2}\right)}{(k+1)^{\delta_2-2\delta_1}}. \qquad (94)$$

Following a similar analysis as we did for term T_1, we get the result below for term T_2.

$$\begin{aligned}
T_2 &= \sum_{\ell=0}^{k} \frac{L^2 a W_2 \mu_e}{\ell^{\delta_2+\delta_3}} \exp\left(-\rho_U \sum_{i=\ell+1}^{k} \alpha_i \right) \\
&\leq \sum_{\ell=0}^{\bar{k}_2} \frac{L^2 a W_2 \mu_e \exp(a\rho_U) \exp\left(\frac{a\rho_U}{1-\delta_2}\ell^{1-\delta_2}\right)}{\ell^{\delta_2+\delta_3}} \exp\left(-\rho_U \sum_{i=0}^{k} \alpha_i \right) + \frac{L^2 a W_2 \mu_e(\delta_2+\delta_3)}{a\rho_U\delta_2} \frac{\exp\left(\frac{a\rho_U}{1-\delta_2}2^{1-\delta_2}\right)}{(k+1)^{\delta_3}},
\end{aligned} \qquad (95)$$

where $\bar{k}_2 = \left\lceil \left(\frac{\delta_2+\delta_3}{a\rho_U} \right)^{\frac{1}{1-\delta_2}} \right\rceil$. Finally for T_3, we first note that

$$\theta_\ell \exp\left(-\rho_U \sum_{i=\ell+1}^{k} \alpha_i \right) \leq \theta_\ell \exp(a\rho_U) \exp\left(-\rho_U \sum_{i=\ell}^{k} \alpha_i \right), \tag{96}$$

which leads to

$$\begin{aligned} T_3 &= \sum_{\ell=0}^{k} \theta_\ell \exp\left(-\rho_U \sum_{i=\ell+1}^{k} \right) \\ &\leq \bar{W}_5 \exp(a\rho_U) \exp\left(-\frac{a\rho_U}{1-\delta_2}(k+1)^{1-\delta_2} \right) \exp\left(\frac{a\rho_U}{1-\delta_2} \right) \sum_{\ell=0}^{k} \exp\left(\ell^{1-\delta_1} \left(\frac{a\rho_U}{1-\delta_2} \ell^{\delta_1-\delta_2-W_1} \right) \right), \\ &\leq C_\theta \exp\left(-\frac{a\rho_U}{1-\delta_2}(k+1)^{1-\delta_2} \right), \end{aligned} \tag{97}$$

where

$$\begin{aligned} C_\theta &= \bar{W}_5 \exp(a\rho_U) \exp\left(\frac{a\rho_U}{1-\delta_2} \right) \left[\sum_{\ell=0}^{\bar{\ell}} \exp\left(\frac{a\rho_U}{1-\delta_2} \ell^{1-\delta_2} - W_1 \ell^{1-\delta_1} \right) + \frac{\kappa^{-\frac{1}{1-\delta_1}}}{1-\delta_1} \Gamma\left(\frac{1}{1-\delta_1} \right) \right], \\ \bar{\ell} &= \left\lceil \left(\frac{a\rho_U}{(1-\delta_2)W_1} \right)^{\frac{1}{\delta_2-\delta_1}} \right\rceil, \\ \kappa &= \left(W_1 - \frac{a\rho_U}{(1-\delta_2)\bar{\ell}^{\delta_2-\delta_1}} \right) > 0, \end{aligned}$$

and $\Gamma(\cdot)$ is the gamma function given as

$$\Gamma(z) = \int_0^\infty x^{z-1} \exp(-x) dx, \quad \forall\ z > 0.$$

Replacing the results from (94), (95), (97) in (90) and thereafter combining it with (88) yields

$$\begin{aligned} F(p_{t_{k+1}}(\bar{\boldsymbol{w}})) \leq & \left[F(p_{t_0}(\bar{\boldsymbol{w}})) + \sum_{\ell=0}^{\bar{k}_1} \frac{C_\xi \exp\left(\rho_U \sum_{i=0}^{\ell} \alpha_i \right)}{\ell^{2\delta_2-2\delta_1}} + \sum_{\ell=0}^{\bar{k}_2} \frac{L^2 a W_2 \mu_e \exp\left(\rho_U \sum_{i=0}^{\ell} \alpha_i \right)}{\ell^{\delta_2+\delta_3}} \right] \exp\left(-\rho_U \sum_{i=0}^{i} \alpha_k \right) \\ & + \frac{\exp\left(\frac{a\rho_U}{1-\delta_2} 2^{1-\delta_2} \right)}{a\rho_U \delta_2} \left[\frac{2C_\xi(\delta_2-\delta_1)}{(k+1)^{\delta_2-2\delta_1}} + \frac{L^2 a W_2 \mu_e (\delta_2+\delta_3)}{(k+1)^{\delta_3}} \right] + C_\theta \exp\left(-\frac{a\rho_U}{1-\delta_2}(k+1)^{1-\delta_2} \right), \end{aligned} \tag{98}$$

where

$$C_\xi = a^3\left[2\bar{L}^4 + L^2(2a^2\bar{L}^2+1)\bar{W}_3\right]C_{\bar{w}} + 2a^3\bar{L}^2L^2W_2\mu_e + a^2\left[L^2(2a^2\bar{L}^2+1)(\bar{W}_4 a + \bar{\bar{W}}_4)\right] + 2a^2\bar{L}^2 d_w,$$

$$\bar{W}_3 = \frac{3n^3L_V^2(1-\sigma_0)^{-\frac{1}{3}}(2\delta_2-\delta_1)}{\sigma_0^2\delta_1}\exp\left(W_1 2^{1-\delta_1}\right),$$

$$\bar{W}_4 = \frac{9n^2L_V^2C^*(\delta_2-\delta_1)}{\sigma_0^3\delta_1}\exp\left(W_1 2^{1-\delta_1}\right),$$

$$\bar{\bar{W}}_4 = \frac{6n^2d_w(1-\sigma_0)^{-\frac{1}{3}}}{\sigma_0^2\delta_1}\exp\left(W_1 2^{1-\delta_1}\right),$$

$$C_\theta = aL^2(2a^2\bar{L}^2+1)W_5\exp\left(a\rho_U\left(\frac{2-\delta_2}{1-\delta_2}\right)\right)\left[\sum_{\ell=0}^{\bar{\ell}}\exp\left(\frac{a\rho_U}{1-\delta_2}\ell^{1-\delta_2} - W_1\ell^{1-\delta_1}\right) + \frac{\kappa^{-\frac{1}{1-\delta_1}}}{1-\delta_1}\Gamma\left(\frac{1}{1-\delta_1}\right)\right],$$

$$\bar{\ell} = \left\lceil\left(\frac{a\rho_U}{(1-\delta_2)W_1}\right)^{\frac{1}{\delta_2-\delta_1}}\right\rceil,$$

$$\kappa = W_1 - \frac{a\rho_U}{(1-\delta_2)\bar{\ell}^{\delta_2-\delta_1}}.$$

Finally, (98) can be rewritten as

$$\begin{aligned}F(p_{t_{k+1}}(\bar{\boldsymbol{w}})) \leq \bar{C}_{_{F_1}}\exp\left(-\rho_U\sum_{i=0}^{k}\alpha_i\right) + \mu_e\left(\frac{\bar{C}_{_{F_2}}}{(k+1)^{\delta_2-2\delta_1}} + \frac{\bar{C}_{_{F_3}}}{(k+1)^{\delta_3}}\right) + \frac{\bar{C}_{_{F_4}}}{(k+1)^{\delta_2-2\delta_1}} \\ + \bar{C}_{_{F_5}}\exp\left(-\frac{\rho_U a}{1-\delta_2}(k+1)^{1-\delta_2}\right)\end{aligned} \tag{99}$$

where

$$\bar{C}_{_{F_1}} = F(p_{t_0}(\bar{\boldsymbol{w}})) + \sum_{\ell=0}^{\bar{k}_1}\frac{C_\xi\exp\left(\rho_U\sum_{i=0}^{\ell}\alpha_i\right)}{\ell^{2\delta_2-2\delta_1}} + \sum_{\ell=0}^{\bar{k}_2}\frac{L^2aW_2\mu_e\exp\left(\rho_U\sum_{i=0}^{\ell}\alpha_i\right)}{\ell^{\delta_2+\delta_3}}, \tag{100}$$

$$\bar{C}_{_{F_2}} = \frac{4a^2\bar{L}^2L^2W_2(\delta_2-\delta_1)}{\rho_U\delta_2}\exp\left(\frac{a\rho_U}{1-\delta_2}2^{1-\delta_2}\right), \tag{101}$$

$$\bar{C}_{_{F_3}} = \frac{L^2W_2(\delta_2+\delta_3)}{\rho_U\delta_2}\exp\left(\frac{a\rho_U}{1-\delta_2}2^{1-\delta_2}\right), \tag{102}$$

$$\begin{aligned}\bar{C}_{_{F_4}} = \frac{2a(\delta_2-\delta_1)}{\rho_U\delta_2}\exp\left(\frac{a\rho_U}{1-\delta_2}2^{1-\delta_2}\right)\Big[a\Big(2\bar{L}^4 + L^2(2a^2\bar{L}^2+1)\bar{W}_3\Big)C_{\bar{w}} + L^2(2a^2\bar{L}^2+1)(\bar{W}_4a + \bar{\bar{W}}_4) \\ + 2\bar{L}^2d_w\Big],\end{aligned} \tag{103}$$

$$\bar{C}_{_{F_5}} = C_\theta. \tag{104}$$

This concludes the proof of Theorem 2. ■

8.3 Proof of Corollary 1

Using (A12) from Appendix A2 in [7]

$$\sum_{l=0}^{k}\alpha_l \geq \int_0^k\frac{a}{(x+1)^{\delta_2}}dx = \frac{a(k+1)^{1-\delta_1}}{1-\delta_1} - \frac{a}{1-\delta_1}. \tag{105}$$

Substituting (105) in (99), we have

$$F(p_{t_{k+1}}(\bar{\boldsymbol{w}})) \leq Q_1 \exp\left(-\frac{a\rho_U}{1-\delta_2}(k+1)^{1-\delta_2}\right) + \frac{\bar{C}_{F_4}}{(k+1)^{\delta_2-2\delta_1}} + \frac{\mu_e Q_2}{(k+1)^{\delta_c}}$$

where

$$Q_1 = \bar{C}_{F_1} \exp\left(\frac{a\rho_U}{1-\delta_2}\right) + \bar{C}_{F_5}, \tag{106}$$

$$Q_2 = \bar{C}_{F_2} + \bar{C}_{F_3}, \tag{107}$$

$$\delta_c = \min\{\delta_3, \delta_2 - 2\delta_1\}. \tag{108}$$

Now, $F(p_{t_{k+1}}(\bar{\boldsymbol{w}})) \leq \epsilon$ is guaranteed if we satisfy the following criteria

$$Q_1 \exp\left(-\frac{a\rho_U}{1-\delta_2}(k+1)^{1-\delta_2}\right) \leq \frac{\epsilon}{3}, \tag{109}$$

$$\frac{\bar{C}_{F_4}}{(k+1)^{\delta_2-2\delta_1}} \leq \frac{\epsilon}{3}, \tag{110}$$

$$\frac{\mu_e Q_2}{(k+1)^{\delta_c}} \leq \frac{\epsilon}{3}. \tag{111}$$

From (109) and (110), we get the following conditions respectively.

$$k \geq \left(\frac{1-\delta_2}{a\rho_U}\log\left(\frac{3Q_1}{\epsilon}\right)\right)^{\frac{1}{1-\delta_2}} - 1,$$

$$k \geq \left(\frac{3\bar{C}_{F_4}}{\epsilon}\right)^{\frac{1}{\delta_2-2\delta_1}} - 1.$$

Therefore, to satisfy both (109) and (110) simultaneously, we choose k^* as:

$$k^* = \max\left\{\left(\frac{1-\delta_2}{a\rho_U}\log\left(\frac{3Q_1}{\epsilon}\right)\right)^{\frac{1}{1-\delta_2}} - 1, \left(\frac{3\bar{C}_{F_4}}{\epsilon}\right)^{\frac{1}{\delta_2-2\delta_1}} - 1\right\}. \tag{112}$$

In addition, to satisfy (111), we substitute k^* from (112) into (111) and find a bound on μ_e as follows.

$$\mu_e \leq \left(\frac{(k^*+1)^{\delta_c}}{3Q_2}\right)\epsilon. \tag{113}$$

This concludes the proof of Corollary 1. ■

8.4 Proof of Theorem 3

Let k_q be any time step of event triggering for one specific agent, say, the i-th agent. Then the error term of agent i for the next time step (k_q+1) is given by

$$\boldsymbol{e}_i(k_q+1) = \boldsymbol{w}_i(k_q+1) - \hat{\boldsymbol{w}}_i(k_q+1) = \boldsymbol{w}_i(k_q+1) - \boldsymbol{w}_i(k_q). \tag{114}$$

Substituting (10) in (114), we get

$$\begin{aligned}\boldsymbol{e}_i(k_q+1) = \beta_{k_q}\sum_{j=1}^{n} a_{ij}(\boldsymbol{e}_i(k_q)-\boldsymbol{e}_j(k_q)) - \beta_{k_q}\sum_{j=1}^{n}(\boldsymbol{w}_i(k_q)-\boldsymbol{w}_i(k_q)) + \alpha_{k_q} n\nabla E_i(\boldsymbol{w}_i(k_q),\boldsymbol{X}_i)\\ + \sqrt{2\alpha_{k_q}}\, d\boldsymbol{v}_i(k_q).\end{aligned} \tag{115}$$

Taking the norm of (115), we have

$$\begin{aligned}\|\boldsymbol{e}_i(k_q+1)\|_2 &\le \beta_{k_q}\left\|\sum_{j=1}^{n} a_{ij}(\boldsymbol{e}_i(k_q)-\boldsymbol{e}_j(k_q))\right\|_2 + \beta_{k_q}\left\|\sum_{j=1}^{n} a_{ij}(\boldsymbol{w}_i(k_q)-\boldsymbol{w}_j(k_q))\right\|_2 + \sqrt{2\alpha_{k_q}}\|\boldsymbol{v}_i(k_q)\|_2\\ &\quad + \alpha_{k_q} n\|\nabla E_i(\boldsymbol{w}_i(k_q),\boldsymbol{X}_i)\|_2,\\ &\le (2\beta_{k_q}d_i + \alpha_{k_q}nL)\|\boldsymbol{w}(k_q)\|_2^{\max} + \beta_{k_q}d_i\|\boldsymbol{e}_i(k_q)\|_2^{\max} + \sqrt{2\alpha_{k_q}}\|\boldsymbol{v}_i(k_q)\|_2 + \alpha_{k_q}nL\|\boldsymbol{w}^*\|_2,\end{aligned} \tag{116}$$

where $\|\boldsymbol{w}(k_q)\|_2^{\max} = \max\limits_{j=\{1,2,\ldots,n\}}\{\|\boldsymbol{w}_j(k_q)\|_2, \|\boldsymbol{w}^*\|_2\}$, $\|\boldsymbol{e}(k_q)\|_2^{\max} = \max\limits_{j=\{1,2,\ldots,n\}}\|\boldsymbol{e}_j(k_q)\|_2$ and d_i is the number of neighbors of the i-th agent, i.e., $d_i = |\mathcal{N}_i|\big($where$|\cdot|$ denotes the cardinality$\big)$. In (116), $\boldsymbol{e}_i(k_q)$ vanishes since i-th agent triggers at the k_q-th time step.

Next, squaring both sides of (116) and using (51) multiple times with $\theta = (1-\beta_{k_q}\lambda_2(\mathcal{L}))^{-1} - 1 > 0$, we obtain

$$\begin{aligned}\|\boldsymbol{e}_i(k_q+1)\|_2^2 \le (1-\beta_{k_q}\lambda_2(\mathcal{L}))^{-2}(2\beta_{k_q}d_i + \alpha_{k_q}nL)^2\big(\|\boldsymbol{w}(k_q)\|_2^{\max}\big)^2 + \frac{(1-\beta_{k_q}\lambda_2(\mathcal{L}))^{-1}\beta_{k_q}d_i^2}{\lambda_2(\mathcal{L})}\big(\|\boldsymbol{e}_i(k_q)\|_2^{\max}\big)^2\\ + \frac{(1-\beta_{k_q}\lambda_2(\mathcal{L}))^{-1}}{\beta_{k_q}\lambda_2(\mathcal{L})}2\alpha_{k_q}\|\boldsymbol{v}_i(k_q)\|_2^2 + \frac{\alpha_{k_q}^2 n^2L^2}{\beta_{k_q}^2\lambda_2^2(\mathcal{L})}\|\boldsymbol{w}^*\|_2^2.\end{aligned} \tag{117}$$

We then take the conditional expectation $\mathbb{E}[\cdot\,|\mathcal{F}_{k_q}]$ of (117) and then the total expectation given to yield

$$\begin{aligned}\mathbb{E}[\|\boldsymbol{e}_i(k_q+1)\|_2^2] &\le \frac{(1-b\lambda_2(\mathcal{L}))^{-2}(2bd_m+anL)^2C_w}{(k_q+1)^{2\delta_1}} + \frac{(1-b\lambda_2(\mathcal{L}))^{-1}bd_m^2\mu_e}{\lambda_2(\mathcal{L})(k+1)^{\delta_1+\delta_3}} + \frac{2and_w(1-b\lambda_2(\mathcal{L}))^{-1}}{b\lambda_2(\mathcal{L})(k_q+1)^{\delta_2-\delta_1}}\\ &\quad + \frac{a^2n^2L^2c^*}{b^2\lambda_2^2(\mathcal{L})(k_q+1)^{2\delta_2-2\delta_1}},\\ &\le \frac{\xi_q}{(k_q+1)^{\delta_m}},\end{aligned} \tag{118}$$

(119)

where $d_m = \max\limits_{i=\{1,2,\ldots,n\}} d_i$, $\delta_m = \min\{\delta_1+\delta_3, \delta_2-\delta_1, 2\delta_1\}$ and

$$\xi_q = \bar{c}_1 + \bar{c}_2\mu_e, \tag{120}$$

wherein

$$\bar{c}_1 = (1-b\lambda_2(\mathcal{L}))^{-2}(2bd_m+anL)^2C_w + \frac{2and_w(1-b\lambda_2(\mathcal{L}))^{-1}}{b\lambda_2(\mathcal{L})} + \frac{a^2n^2L^2c^*}{b^2\lambda_2^2(\mathcal{L})}, \tag{121}$$

$$\bar{c}_2 = \frac{(1-b\lambda_2(\mathcal{L}))^{-1}bd_m^2}{\lambda_2(\mathcal{L})}. \tag{122}$$

In (118) we have made use of the results $\mathbb{E}\Big[\big(\|\boldsymbol{w}(k_q)\|_2^{\max}\big)^2\Big] \le C_w$ (which follows from the conclusion in (79) and $\big(1-\beta_{k_q}\lambda_2(\mathcal{L})\big)^{-n} \le (1-b\lambda_2(\mathcal{L}))^{-n}, n\in\{1,2\}$.

To prevent triggering in consecutive time steps, on expectation, we need to ensure

$$\mathbb{E}[\|\boldsymbol{e}_i(k_q+1)\|_2^2] \leq \frac{\mu_e}{(k_q+2)^{\delta_3}},$$

which, from (119), is ensured if

$$\frac{\xi_q}{(k_q+1)^{\delta_m}} \leq \frac{\mu_e}{(k_q+2)^{\delta_3}}, \tag{123}$$

i.e.,

$$\xi_q \leq \frac{(k_q+1)^{\delta_m}}{(k_q+2)^{\delta_3}}\mu_e, \tag{124}$$

Furthermore, since $\delta_m \geq \delta_3$, then $\frac{(k_q+1)^{\delta_m}}{(k_q+2)^{\delta_3}}$ rapidly goes beyond 1 with increasing time steps, thus, we need only to ensure $\xi_q \leq \mu_e$ which from (120) gives

$$\mu_e > \frac{\bar{c}_1}{1-\bar{c}_2}. \tag{125}$$

The condition in (125) can be easily satisfied by adjusting the parameters (b, μ_e). Further, from Markov's inequality, we get

$$p\left(\|\boldsymbol{e}_i(k_q+1)\|_2^2 > \frac{\mu_e}{(k_q+2)^{\delta_3}}\right) \leq \frac{\xi_q}{\mu_e}\frac{(k_q+2)^{\delta_3}}{(k_q+1)^{\delta_m}}, \leq \left(\frac{\bar{c}_1}{\mu_e}+\bar{c}_2\right)\frac{(k_q+2)^{\delta_3}}{(k_q+1)^{\delta_m}}. \tag{126}$$

This concludes the proof of Theorem 3. ■

9 Appendix

In this section, we list all the useful lemmas that have been used for the analysis of our algorithm.

Lemma 1 *Given assumption 1, we have*

$$M \triangleq \left(I_n - \frac{1}{n}\mathbf{1}_n\mathbf{1}_n^\top\right) = \mathcal{L}\mathcal{L}^+, \tag{127}$$

where $(\cdot)$ *denotes the generalized inverse. Furthermore, for all* $\mathbf{x} \in \mathbb{R}^n$ *and* $\mathbf{x} \notin \mathbb{R}^n_{\mathbf{1}}$*, we have*

$$\tilde{\mathbf{x}}^\top\mathcal{L}\tilde{\mathbf{x}} = \mathbf{x}^\top\mathcal{L}\mathbf{x} > \lambda_2(\mathcal{L})\mathbf{x}^\top\mathbf{x}, \tag{128}$$

where $\tilde{\mathbf{x}} = M\mathbf{x}$ *and* $\lambda_2(\cdot)$ *denoted the second largest eigenvalue of* $\mathcal{L}$*.*

Refer to Lemma 3 in [21] for a detailed proof.

Lemma 2 *Let* $f(t)$ *be a non-negative and decreasing sequence for all* $k \leq t \leq K$*, then we have*

$$\int_k^K f(t)dt \leq \sum_{t=k}^{K} f(t) \leq \int_{k-1}^{K} f(t)dt. \tag{129}$$

Alternatively, for a non-negative and increasing sequence $f(t)$ for all $k \leq t \leq K$, we have

$$\int_{k-1}^{K} f(t)dt \leq \sum_{t=k}^{K} f(t) \leq \int_{k}^{K+1} f(t)dt. \tag{130}$$

Refer to Appendix A2 in [7] for the proof.

Lemma 3 *For a non-negative sequence $\{y_k\}$ satisfying:*

$$y_{k+1} \leq \left(1 - \frac{\mu_\beta}{(k+1)^{\delta_1}}\right) + \frac{\mu_\zeta}{(k+1)^{\delta_4}}, \tag{131}$$

for all $k \geq 0$ and where $0 < \mu_\beta \leq 1$, $\delta_1, \delta_4 \in (0,1)$ and $\delta_1 < \delta_4$, we have the following convergence rate.

$$y_{k+1} \leq \frac{Y_3}{\exp(Y_1(k+1)^{1-\delta_1})} + \frac{Y_2}{(k+1)^{\delta_4}}, \tag{132}$$

where

$$\begin{aligned}
Y_1 &= \frac{\mu_\beta}{1-\delta_1}, \\
Y_2 &= \frac{\mu_\zeta \delta_4}{\mu_\beta \delta_1} \exp(Y_1 2^{1-\delta_1}), \\
Y_3 &= \exp(Y_1)\left(y_0 + \sum_{t=0}^{\bar{k}} (1-\mu_\beta)^{-1} \frac{\mu_\zeta}{(k+1)_4^{\delta}}\right), \\
\bar{k} &= \left\lceil \left(\frac{\delta_4}{\mu_\beta}\right)^{\frac{1}{1-\delta_1}} \right\rceil.
\end{aligned}$$

Refer to Lemma S4 in [39] for a detailed proof.

References

[1] Christophe Andrieu, Nando De Freitas, Arnaud Doucet, and Michael I Jordan. An introduction to MCMC for machine learning. *Machine learning*, 50(1-2):5–43, 2003.

[2] Aslansefat, Koorosh. Ecdf-based distance measure algorithms, 2020. available at: `https://github.com/koo-ec/ECDF-based-Distance-Measure`, Retrieved April 29, 2021.

[3] Ngoc Huy Chau, Éric Moulines, Miklos Rásonyi, Sotirios Sabanis, and Ying Zhang. On stochastic gradient Langevin dynamics with dependent data streams: the fully non-convex case. *arXiv preprint*, arXiv:1905.13142, 2019.

[4] Xiang Cheng and Peter L Bartlett. Convergence of Langevin MCMC in KL-divergence. *Proceedings of Machine Learning Research*, (83):186–211, 2018.

[5] Xiang Cheng, Niladri S Chatterji, Yasin Abbasi-Yadkori, Peter L Bartlett, and Michael I Jordan. Sharp convergence rates for Langevin dynamics in the nonconvex setting. *arXiv e-print*, page arXiv:1805.01648, 2018.

[6] Xiang Cheng, Niladri S. Chatterji, Peter L. Bartlett, and Michael I. Jordan. Underdamped Langevin MCMC: A non-asymptotic analysis. In *Proceedings of the 31st Conference On Learning Theory*, volume 75, pages 300–323, 06–09 Jul 2018.

[7] T.H. Cormen, C.E. Leiserson, R.L. Rivest, and C. Stein. *Introduction to Algorithms*. Computer science. MIT Press, 2009.

[8] 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, 2017.

[9] Arnak S Dalalyan. Theoretical guarantees for approximate sampling from smooth and log-concave densities. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 79(3):651–676, 2017.

[10] D. V. Dimarogonas, E. Frazzoli, and K. H. Johansson. Distributed event-triggered control for multi-agent systems. *IEEE Transactions on Automatic Control*, 57(5):1291–1297, 2012.

[11] W. Du, X. Yi, J. George, K. H. Johansson, and T. Yang. Distributed optimization with dynamic event-triggered mechanisms. In *2018 IEEE Conference on Decision and Control (CDC)*, pages 969–974, 2018.

[12] Simon Duane, Anthony D Kennedy, Brian J Pendleton, and Duncan Roweth. Hybrid Monte Carlo. *Physics letters B*, 195(2):216–222, 1987.

[13] Alain Durmus and Eric Moulines. Sampling from strongly log-concave distributions with the unadjusted Langevin algorithm. *arXiv e-prints*, arXiv:1605.01559, 2016.

[14] Alain Durmus and Eric Moulines. Nonasymptotic convergence analysis for the unadjusted langevin algorithm. *The Annals of Applied Probability*, 27(3):1551–1587, 2017.

[15] Alain Durmus and Eric Moulines. High-dimensional bayesian inference via the unadjusted langevin algorithm. *Bernoulli*, 25(4A):2854–2882, 2019.

[16] Jemin George and Prudhvi Gurram. Distributed stochastic gradient descent with event-triggered communication. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 7169–7178, 2020.

[17] A. Girard. Dynamic triggering mechanisms for event-triggered control. *IEEE Transactions on Automatic Control*, 60(7):1992–1997, 2015.

[18] Leonard Gross. Logarithmic Sobolev inequalities. *American Journal of Mathematics*, 97(4):1061–1083, 1975.

[19] Mert Gürbüzbalaban, Xuefeng Gao, Yuanhan Hu, and Lingjiong Zhu. Decentralized Stochastic Gradient Langevin Dynamics and Hamiltonian Monte Carlo. *arXiv:2007.00590*, 2020.

[20] W. P. M. H. Heemels, K. H. Johansson, and P. Tabuada. An introduction to event-triggered and self-triggered control. In *IEEE 51st IEEE Conference on Decision and Control (CDC)*, pages 3270–3285, 2012.

[21] W. Xiao I. Gutman. Generalized inverse of the Laplacian matrix and some applications. *Bulletin, Classe des Sciences Mathématiques et Naturelles, Sciences mathématiques*, 129(29):15–23, 2004.

[22] Richard Jordan, David Kinderlehrer, and Felix Otto. The variational formulation of the Fokker–Planck equation. *SIAM Journal on Mathematical Analysis*, 29(1):1–17, 1998.

[23] S. Kar, J. Moura, and H. Poor. Distributed linear parameter estimation: Asymptotically efficient adaptive strategies. *SIAM Journal on Control and Optimization*, 51(3):2200–2229, 2013.

[24] Solmaz S. Kia, Jorge Cortés, and Sonia Martínez. Distributed convex optimization via continuous-time coordination algorithms with discrete-time communication. *Automatica*, 55:254 – 264, 2015.

[25] Vyacheslav Kungurtsev. Decentralized Langevin dynamics. *arXiv preprint arXiv:2001.00665*, 2020.

[26] Anusha Lalitha, Xinghan Wang, Osman Kilinc, Yongxi Lu, Tara Javidi, and Farinaz Koushanfar. Decentralized Bayesian learning over graphs. *arXiv preprint arXiv:1905.10466*, 2019.

[27] B. Leimkuhler, S. Reich, and Cambridge University Press. *Simulating Hamiltonian Dynamics*. Cambridge Monographs on Applied and Computational Mathematics. Cambridge University Press, 2004.

[28] Don S. Lemons and Anthony Gythiel. Paul Langevin's 1908 paper "On the Theory of Brownian Motion" ["Sur la ᴛhéorie du mouvement Brownien," C. R. Acad. Sci. (Paris) 146, 530–533 (1908)]. *American Journal of Physics*, 65(11):1079–1081, 1997.

[29] Yi-An Ma, Niladri Chatterji, Xiang Cheng, Nicolas Flammarion, Peter Bartlett, and Michael I. Jordan. Is There an Analog of Nesterov Acceleration for MCMC? *arXiv e-prints*, February 2019.

[30] Yi-An Ma, Tianqi Chen, and Emily Fox. A complete recipe for stochastic gradient MCMC. In *Advances in Neural Information Processing Systems 28*, pages 2917–2925. 2015.

[31] 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.

[32] Mateusz B Majka, Aleksandar Mijatović, and Lukasz Szpruch. Non-asymptotic bounds for sampling algorithms without log-concavity. *arXiv preprint*, arXiv:1808.07105, 2018.

[33] Oren Mangoubi and Aaron Smith. Mixing of Hamiltonian Monte Carlo on strongly log-concave distributions 2: Numerical integrators. In Kamalika Chaudhuri and Masashi Sugiyama, editors, *Proceedings of Machine Learning Research*, volume 89 of *Proceedings of Machine Learning Research*, pages 586–595. PMLR, 16-18 Apr 2019.

[34] Oren Mangoubi and Nisheeth Vishnoi. Dimensionally tight bounds for second-order Hamiltonian Monte Carlo. In *Advances in Neural Information Processing Systems*, pages 6027–6037. 2018.

[35] Wenlong Mou, Nicolas Flammarion, Martin J Wainwright, and Peter L Bartlett. Improved bounds for discretization of Langevin diffusions: near-optimal rates without convexity. *arXiv preprint*, arXiv:1907.11331, 2019.

[36] Wenlong Mou, Yi-An Ma, Martin J. Wainwright, Peter L. Bartlett, and Michael I. Jordan. High-Order Langevin Diffusion Yields an Accelerated MCMC Algorithm. *arXiv e-prints*, August 2019.

[37] Radford M. Neal. MCMC using Hamiltonian dynamics. *arXiv e-prints*, arXiv:1206.1901, June 2012.

[38] F. Otto and C. Villani. Generalization of an inequality by Talagrand and links with the logarithmic Sobolev inequality. *Journal of Functional Analysis*, 173(2):361 – 400, 2000.

[39] Anjaly Parayil, He Bai, Jemin George, and Prudhvi Gurram. Decentralized Langevin dynamics for Bayesian learning. *Advances in Neural Information Processing Systems*, 33, 2020.

[40] G.A. Pavliotis. *Stochastic Processes and Applications: Diffusion Processes, the Fokker-Planck and Langevin Equations*. Texts in Applied Mathematics. Springer New York, 2014.

[41] Maxim Raginsky, Alexander Rakhlin, and Matus Telgarsky. Non-convex learning via stochastic gradient Langevin dynamics: a nonasymptotic analysis. In *Proceedings of the 2017 Conference on Learning Theory*, volume 65, pages 1674–1703. PMLR, 07–10 Jul 2017.

[42] Gareth O. Roberts and Richard L. Tweedie. Exponential convergence of Langevin distributions and their discrete approximations. *Bernoulli*, 2(4):341–363, 12 1996.

[43] Navjot Singh, Deepesh Data, Jemin George, and Suhas Diggavi. SPARQ-SGD: Event-triggered and compressed communication in decentralized optimization. In *59th IEEE Conference on Decision and Control (CDC)*, pages 3449–3456. IEEE, 2020.

[44] Kunal Talwar. Computational separations between sampling and optimization. In *Advances in Neural Information Processing Systems*, pages 14997–15007, 2019.

[45] Santosh Vempala and Andre Wibisono. Rapid convergence of the unadjusted langevin algorithm: Isoperimetry suffices. In *Advances in Neural Information Processing Systems*, pages 8094–8106, 2019.

[46] Santosh Vempala and Andre Wibisono. Rapid convergence of the unadjusted Langevin algorithm: Isoperimetry suffices. In *Advances in Neural Information Processing Systems*, pages 8094–8106. 2019.

[47] Max Welling and Yee W Teh. Bayesian learning via stochastic gradient Langevin dynamics. In *Proceedings of the 28th international conference on machine learning (ICML-11)*, pages 681–688, 2011.

[48] Andre Wibisono. Sampling as optimization in the space of measures: The Langevin dynamics as a composite optimization problem. In *Proceedings of the 31st Conference On Learning Theory*, volume 75, pages 2093–3027, 06–09 Jul 2018.

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

[50] Pan Xu, Jinghui Chen, Difan Zou, and Quanquan Gu. Global convergence of Langevin dynamics based algorithms for nonconvex optimization. *Advances in Neural Information Processing Systems*, pages 3122–3133, 2018.

[51] X. Yi, L. Yao, T. Yang, J. George, and K. H. Johansson. Distributed optimization for second-order multi-agent systems with dynamic event-triggered communication. In *2018 IEEE Conference on Decision and Control (CDC)*, pages 3397–3402, 2018.

[52] Ying Zhang, Ömer Deniz Akyildiz, Theo Damoulas, and Sotirios Sabanis. Nonasymptotic estimates for stochastic gradient Langevin dynamics under local conditions in nonconvex optimization. *arXiv preprint*, arXiv:1910.02008, 2019.