High-Order and Bias-Free Distributed Langevin Dynamics for Scalable Bayesian Inference

Mohammad Rafiqul Islam

Florida State University

Jan 30, 2026

Table of Contents

Background and Preliminaries

The overdamped Langevin diffusion is defined by the stochastic differential equation (SDE)[1] [2]:

$$dX_t = -\nabla f(X_t)dt + \sqrt{2}dW_t, \qquad t \geq 0 \tag{1}$$

where $\{X_t : t \in \mathbb{R}_+\}$ is a continuous-time diffusion process, $W_t : t \geq 0$ is a d-dimensional Brownian motion

- Under some mild conditions, the SDE in (1) has a strong solution that admits Gibbs distribution $\pi(x) \propto e^{-f(x)}$ as the unique invariant distribution.

[1] Pavliotis, G.A. (2014). Stochastic Processes and Applications: Diffusion Processes, The Fokker-Planck and Langevin Equations. Springer.

[2] Roberts, G.O. and Tweedie, R.L. (1996). Exponential convergence of Langevin distributions and their discrete approximations. Bernoulli. 2(4):341-363.

Background and Preliminaries

- The (first-order) overdamped Langevin Monte Carlo (LMC) algorithm is a well-known sampling method that is the Euler-Maruyama discretization of the SDE (1):

$$x_{k+1} = x_k - \eta \nabla f(x_k) + \sqrt{2\eta}\xi_{k+1} \tag{2}$$

- A variant of (1) is the (second-order) underdamped LMC algorithm (uLMC), that is based on the SDE:

$$\begin{aligned} dV_t &= -\gamma V_t dt - \nabla f(X_t)dt + \sqrt{2\gamma}dW_t, \qquad t \geq 0, \\ dX_t &= V_t dt, \end{aligned} \tag{3}$$

where $(V_t, X_t) \in \mathbb{R}^d \times \mathbb{R}^d$, $W_t : t \geq 0$ is a d-dimensional Brownian motion, $\gamma > 0$ is the friction coefficient

Background and Preliminaries

- Many novel Langevin algorithms have been proposed based on the optimization literature.
- Decentralized Stochastic Gradient Langevin Dynamics[3]

$$x_i^{(k+1)} = \sum_{j \in \Omega_i} W_{ij} x_j^{(k)} - \eta \tilde{\nabla} f_i \left(x_i^{(k)} \right) + \sqrt{2\eta} \xi_i^{(k+1)}. \tag{4}$$

  where W_{ij} are the entries of a doubly stochastic weight matrix W with $W_{ij} > 0$ only if i is connected to j.
- Projected Langevin Monte Carlo[4]

$$dX_t = \mathcal{P}_C \left(x_k - \eta \nabla f \left(x_k \right) + \sqrt{2\eta} \xi_{k+1} \right) \tag{5}$$

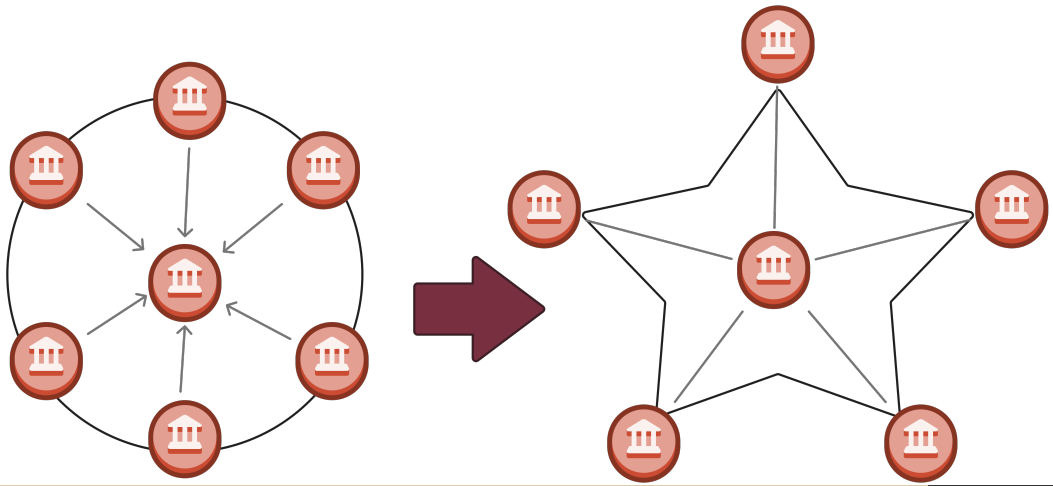
  where $\mathcal{P}_C(\cdot)$ is the projection onto the closed-convex set C.

[3] Gurbuzbalaban, M., Gao, X., Hu, Y. and L. Zhu (2021). Decentralized stochastic gradient Langevin dynamics and Hamiltonian Monte Carlo. Journal of Machine Learning Research. 22, 1-69.

[4] Bubeck et al. (2015). Finite-time analysis of Projected Langevin Monte Carlo. Advances in Neural Information Processing Systems.

Decentralized Learning System

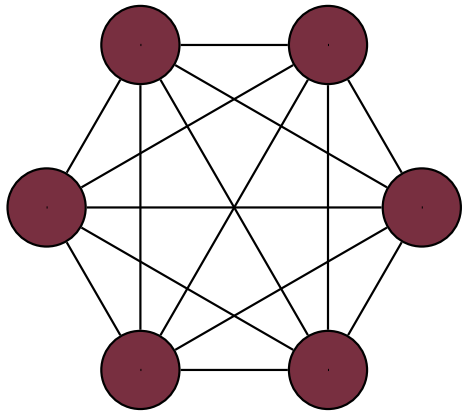
- The present era is the era of Machine Learning (ML), Artificial Intelligence (AI), and Big Data Analytics.
- The rate at which the data is generated is often outpaced by our ability to analyze in terms of computational resources. This causes the need of scalable machine learning algorithms.
- The data are collected through many digital devices, and those devices are connected through a communication network



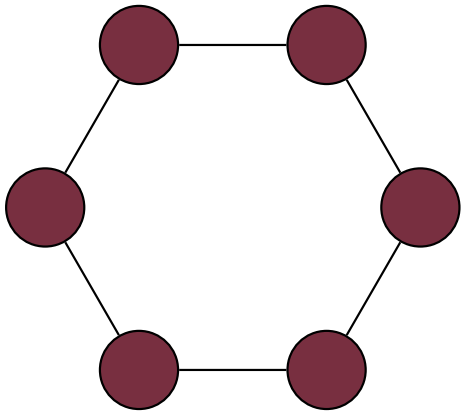
Decentralized Learning System

- We have N computational agents or nodes connected over a network $\mathcal{G}$ where $\mathcal{V} = \{1, 2, \cdots, N\}$ represents the agents and $\mathcal{E} \subseteq \mathcal{V} \times \mathcal{V}$ is the set of edges; i.e., i and j are connected if $(i, j) \in \mathcal{E}$.
- Let $A = [a_1, a_2, \cdots, a_n]$ be the input-output data pair with n i.i.d. observations sampled from a parametersized distribution $p(A \mid x)$ where the parameter $x \in \mathbb{R}^d$ has a common prior distribution $p(x)$
- In the decentralized setting, agent i possesses a subset A_i of the data where $A_i = \{a_1^i, a_2^i, \cdots, a_{n_i}^i\}$, and n_i is the number of samples of the agent i.
- The data is held disjointly; i.e., $A = \bigcup_i A_i$ with $A_i \cap A_j = \emptyset$ for $j \neq i$.
- The goal is to sample from the posterior distribution $p(x \mid A) \propto p(A \mid x)p(x)$

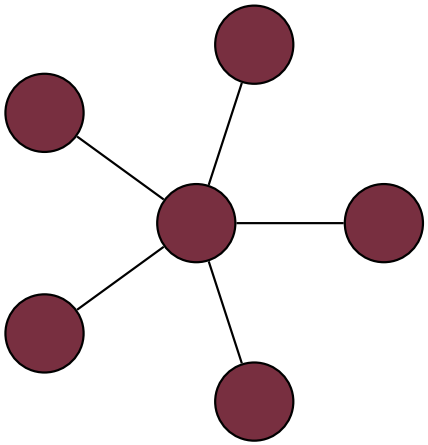
Network Topologies



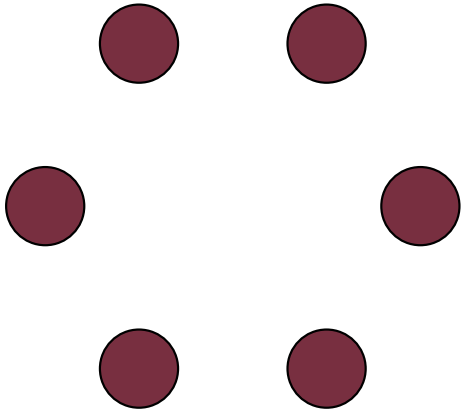
(a) Fully connected



(b) Circular



(c) Star



(d) Disconnected

Figure 2: Different types of network structures.

Decentralized Learning System

- Since the data are i.i.d., the log-likelihood is additive, i.e.,
$$\log p(A \mid x) = \sum_{i=1}^{N} \sum_{j=1}^{n_i} \log p(a_j^i \mid x)$$
- Thus, if we set the potential
$$f(x) := \sum_{i=1}^{N} f_i(x), \quad f_i(x) := -\sum_{j=1}^{n_i} \log p\left(a_j^i \mid x\right) - \frac{1}{N} \log p(x) \tag{6}$$
then the goal is to sample from the posterior distribution with density $\pi(x) := p(x \mid A) \propto e^{-f(x)}$
- The functions $f_i(x)$ are called "component functions" where $f_i(x)$ is associated tot he local data of agent i and is only accessible by the agent i.

Decentralized SGLD (DE-SGLD)

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

$$x_i^{(k+1)} = \sum_{j \in \Omega_i} W_{ij} x_j^{(k)} - \eta \tilde{\nabla} f_i\left(x_i^{(k)}\right) + \sqrt{2\eta} w_i^{(k+1)}. \qquad (7)$$

  - $\eta > 0$ is the stepsize.
  - W_{ij} are the entries of the doubly stochastic weight matrix W. For instance, $W = I - \delta L$ where $\delta > 0$ and $L = D - A$ is the graph Laplacian.

Decentralized SGLD (DE-SGLD)

- $w_i^{(k+1)}$ are the i.i.d. and $w_i^{(k+1)} \sim \mathcal{N}(0, \mathrm{I})$
- $\tilde{\nabla} f_i\left(x_i^{(k)}\right)$ is an unbiased stochastic estimate of the deterministic gradient $\nabla f_i\left(x_i^{(k)}\right)$ with bounded variance, i.e., $\xi_i^{(k+1)} := \tilde{\nabla} f_i\left(x_i^{(k)}\right) - \nabla f_i\left(x_i^{(k)}\right)$ and

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

where $\mathcal{F}_k$ is the natural filtration of the iterates $x_i^{(k)}$ up to (and including) k.

Why and What is EXTRA SGLD?

- Notice that when we use the full-batch, the iterations in (7) turn out to be

$$x_i^{(k+1)} = \sum_{j \in \Omega_i} W_{ij} x_j^{(k)} - \eta \nabla f_i \left(x_i^{(k)} \right) - \eta \xi_i^{(k+1)} + \sqrt{2\eta} w_i^{(k+1)}. \quad (8)$$

- The full-batch implementation introduces a bias term in the form of the Gaussian noise $\xi_i^{(k+1)}$
- As a result, the distribution of the iterates $x_i^{(k)}$ converges linearly to the neighborhood but not too close to the target distribution.
- Thus, we introduce EXact FirsT-OrdeR Algorithm (EXTRA) SGLD that allows to obtain non-asymptotic performance guarantees in terms of $\mathcal{W}_2$ distance.
- When each $f_i(x)$ and $f(x) = \sum f_i(x)$ are smooth and strongly convex, EXTRA iterates distribution converges to $\pi(x)$ linearly (but geometrically fast in k)

Wasserstein Distance

Define $\mathrm{P}(\mathbb{R}^d)$ as the space consisting of all the Borel probability measures μ on $\mathbb{R}^d$ with the finite second moment (based on the Euclidean norm).

Definition (2–Wasserstein distance)

For any two Borel probability measures μ, ν on $\mathbb{R}^d$ with finite second moments, the 2–Wasserstein distance[5] is defined as

$$\begin{aligned} \mathcal{W}_2(\mu,\nu) &:= \left(\inf \mathbb{E}\,\|x-y\|^2\right)^{\frac{1}{2}} \\ &= \left(\inf_{\gamma\in\Pi(\mu,\nu)} \int_{\mathbb{R}^d\times\mathbb{R}^d} \|x-y\|^2 \, d\gamma(x,y)\right)^{\frac{1}{2}}, \end{aligned}$$

where the infimum is taken over all joint distributions of the random variable x, y with marginal distributions μ, ν respectively.

[5] Villani, Cédric. Optimal transport: old and new. Vol. 338. Berlin: springer, 2008.

Assumptions

Assumption 1

We assume for every $i = 1, \cdots, N$:

- $f_i(x)$ is $\mu-$strongly convex and L-smooth, i.e., for every $x, y \in \mathbb{R}^d$,

$$\frac{L}{2}\|x-y\|^2 \geq f_i(x) - f_i(y) - \nabla f_i(y)^\top (x-y) \geq \frac{\mu}{2}\|x-y\|^2 \quad (9)$$

- $F : \mathbb{R}^{Nd} \mapsto \mathbb{R}$ with $F(x_1, \cdots, x_N) = \sum_{i=1}^N f_i(x_i)$ is also μ-strongly convex and L-smooth for any $x = (x_1, x_2, \cdots, x_N) \in \mathbb{R}^{Nd}$.

Assumptions

Assumption 2

The gradient noise $\xi_i^{(k+1)} := \tilde{\nabla} f_i\left(x_i^{(k)}\right) - \nabla f\left(x_i^{(k)}\right)$ at iteration k for any agent i is unbiased with a finite second moment, i.e.,

$$\mathbb{E}\left[\xi_i^{(k+1)} \mid \mathcal{F}_k\right] = 0, \qquad \mathbb{E}\left\|\xi_i^{(k+1)}\right\|^2 \leq \sigma^2, \tag{10}$$

where $\mathcal{F}_k$ is the natural filtration of the iterates $x_i^{(k)}$ up to (and including) time k.

Assumptions

Assumption 3

Consider a connected network $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ consisting of a set of agents $\mathcal{V} = \{1, 2, \cdots, N\}$ and a set of undirected edges $\mathcal{E}$. The doubly stochastic matrices $W = [W_{ij}] \in \mathbb{R}^{N\times N}$ and $= [\tilde{W}_{ij}] \in \mathbb{R}^{N\times N}$ satisfy

1. Null space property:

$$\text{null}\{W - \tilde{W}\} = \text{span}\{1_N\}, \qquad \text{null}\{I_N - \tilde{W}\} \supseteq \text{span}\{1_N\},$$

   where $\text{span}\{1_N\}$ is the span of the vector space supported by all-one vector $[1_N^\top, 1_N^\top, \cdots, 1_N^\top]$.

2. Spectral property:

$$\tilde{W} \succ 0, \qquad \frac{I_N + W}{2} \succeq \tilde{W} \succeq W.$$

Generalized EXTRA SGLD

- We assume that $\tilde{W} = h\, I_N + (1-h)W$, for $h \in (0, 1/2]$.
- Then the EXTRA stochastic gradient Langevin dynamics[6] as follows.

$$x_i^{(k+2)} = \sum_{j \in \Omega_i} W_{ij} x_j^{(k+1)} - \eta \nabla f_i \left(x_i^{(k+1)} \right) - \eta \xi_i^{(k+1)} + \sqrt{2\eta} w_i^{(k+2)}, \tag{11}$$

$$x_i^{(k+1)} = \sum_{j \in \Omega_i} \widetilde{W}_{ij} x_j^{(k)} - \eta \nabla f_i \left(x_i^{(k)} \right) - \eta \xi_i^{(k)} + \sqrt{2\eta} w_i^{(k+1)}. \tag{12}$$

[6] Gurbuzbalaban, Mert, Mohammad Rafiqul Islam, Xiaoyu Wang, and Lingjiong Zhu. "Generalized EXTRA stochastic gradient Langevin dynamics." arXiv preprint arXiv:2412.01993 (2024).

Generalized EXTRA SGLD

- These updates for N agents can also be expressed as

$$x^{(k+2)} = \mathcal{W}x^{(k+1)} - \eta\nabla F\left(x^{(k+1)}\right) - \eta\xi^{(k+1)} + \sqrt{2\eta}w^{(k+2)}, \tag{13}$$

$$x^{(k+1)} = \widetilde{\mathcal{W}}x^{(k)} - \eta\nabla F\left(x^{(k)}\right) - \eta\xi^{(k)} + \sqrt{2\eta}w^{(k+1)}, \tag{14}$$

- where $\mathcal{W} = W \otimes I_d$, $\widetilde{\mathcal{W}} = \widetilde{W} \otimes I_d$,
- $x^{(k)} = \left[\left(x_1^{(k)}\right)^T, \ldots, \left(x_N^{(k)}\right)^T\right]^T \in \mathbb{R}^{Nd}$, and
- $w^{(k)} = \left[\left(w_1^{(k)}\right)^T, \ldots, \left(w_N^{(k)}\right)^T\right]^T, \quad k = 0, 1, 2, \ldots,$

Generalized EXTRA SGLD

Let us define the average at the k-th iteration $\bar{x}^{(k)} := \frac{1}{N}\sum_{i=1}^{N} x_i^{(k)}$.

Theorem

Consider the generalized EXTRA Langevin dynamics with the network averaging matrix $\widetilde{W} = hI_N + (1-h)W$ where

$$0 < h \leq \frac{1-\bar{\gamma}_W}{4\bar{\gamma}^2_{I_N-W}} \wedge \frac{1}{2} \wedge \frac{1}{\gamma_1\gamma_2}, \tag{15}$$

and assume that the stepsize η is chosen satisfying

$$0 < \eta < \frac{1}{h\gamma_1\gamma_2} \wedge \frac{\gamma_{\tilde{W}}}{6(L+\mu)\vee 2A} \wedge 1 \wedge \frac{1}{L+\mu} \wedge \frac{\gamma_{\tilde{W}}}{6(L+\mu)}, \tag{16}$$

where $\gamma_1, \gamma_2, \gamma_{\tilde{W}}, \bar{\gamma}^2_{I_N-W}$ are constants are defined in our main paper ([Gurbuzbalaban et al.(2024)Gurbuzbalaban, Islam, Wang, and Zh

Generalized EXTRA SGLD

Theorem (continues)

Then, for any $K \geq K_0$, the following bound holds:

$$\mathcal{W}_2\left(\mathcal{L}\left(\bar{x}^{(K)}\right), \pi\right) \leq \left(\frac{\bar{\gamma}_{\tilde{W}}^{2K} - \left(1 - \eta\mu\left(1 - \frac{\eta L}{2}\right)\right)^K}{\bar{\gamma}_{\tilde{W}}^2 - 1 + \eta\mu\left(1 - \frac{\eta L}{2}\right)}\right)^{1/2} \cdot \frac{2L\bar{\gamma}_{\tilde{W}}}{\sqrt{N}}\left(\mathbb{E}\left\|x^{(0)}\right\|^2\right)^{1/2} + (1 - \eta\mu)^K \mathcal{W}_2\left(\mathcal{L}(x_0), \pi\right) + \sqrt{\eta}\mathcal{E}_1$$

Generalized EXTRA SGLD

Theorem (continues)

where

$$\mathcal{E}_1 := \left(\frac{\eta}{\mu\left(1-\frac{\eta L}{2}\right)} + \frac{(1+\eta L)^2}{\mu^2\left(1-\frac{\eta L}{2}\right)^2} \right)^{1/2} \cdot \left(\frac{4L^2\left(R_h + R_h'\right)\eta}{N(1-\bar{\gamma}_{\tilde{W}})^2} + \frac{4L^2\sigma^2\eta}{1-\bar{\gamma}_{\tilde{W}}^2} + \frac{8L^2 d}{1-\bar{\gamma}_{\tilde{W}}^2} \right)^{1/2} + \frac{\sigma}{\sqrt{\mu\left(1-\frac{\eta L}{2}\right)N}} + \frac{1.65L}{\mu}\sqrt{dN^{-1}}.$$

Generalized EXTRA SGLD

Theorem (continues)

Moreover,

$$\frac{1}{N}\sum_{i=1}^{N}\mathcal{W}_2\left(\mathcal{L}\left(x_i^{(K)}\right),\pi\right)$$

$$\leq \eta\cdot\frac{D_1}{\sqrt{N}}+\sqrt{\eta}\cdot(D_2+\mathcal{E}_1)+\left(\frac{\bar{\gamma}_{\tilde{W}}^{2K}-\left(1-\eta\mu\left(1-\frac{\eta L}{2}\right)\right)^K}{\bar{\gamma}_{\tilde{W}}^{2}-1+\eta\mu\left(1-\frac{\eta L}{2}\right)}\right)^{1/2}\cdot$$

$$\frac{2L\bar{\gamma}_{\tilde{W}}}{\sqrt{N}}\left(\mathbb{E}\left\|x^{(0)}\right\|^2\right)^{1/2}+(1-\mu\eta)^K\mathcal{W}_2\left(\mathcal{L}\left(x_0\right),\pi\right)+$$

$$\frac{2\left(\bar{\gamma}_{\tilde{W}}\right)^K}{\sqrt{N}}\sqrt{\mathbb{E}\left\|x^{(0)}\right\|^2}$$

Why EXTRA SGLD is better then DE-SGLD

For DE-SGLD, under the assumptions in Theorem 1 in [7], as $\varepsilon \to 0$,

$$\frac{1}{N}\sum_{i=1}^{N} \mathcal{W}_2^{\text{de-sgld}}\left(\mathcal{L}\left(x_i^{(K)}\right), \pi\right) \leq O(\varepsilon), \tag{17}$$

provided that

$$K \geq K^{\text{de-sgld}} = \tilde{O}\left(\frac{L^4 d}{\varepsilon^2 \mu^3}\right) \tag{18}$$

where $\tilde{O}$ hides the logarithmic dependence on ε.

[7] Gürbüzbalaban, Mert, et al. "Decentralized stochastic gradient Langevin dynamics and Hamiltonian monte carlo." Journal of Machine Learning Research 22.239 (2021): 1-69.

Why EXTRA SGLD is better then DE-SGLD

For generalized EXTRA SGLD, under the assumptions in Theorem, as $\varepsilon \to 0$, by taking $h \geq \Omega(\eta\mu)$ and $h < \frac{1}{(L/\mu)^4(L+\|B\|^2)} \wedge \frac{1}{2\gamma_1\gamma_2}$, it holds that

$$\frac{1}{N}\sum_{i=1}^{N} \mathcal{W}_2^{\text{extra-sgld}}\left(\mathcal{L}\left(x_i^{(K)}\right), \pi\right) \leq \mathcal{O}(\varepsilon), \tag{19}$$

provided that

$$K \geq K^{\text{extra-sgld}} = \tilde{\mathcal{O}}\left(\frac{L^2 d}{\epsilon^2 \mu^3}\right), \tag{20}$$

where $\tilde{\mathcal{O}}$ hides the logarithmic dependence on ε and the constants γ_1, γ_2 are provided in the main paper

Experimentally: EXTRA-SGLD vs DE-SGLD

- Synthetic data that we generate by simulating the following model:

$$\delta_i \sim \mathcal{N}(0, \xi^2), \qquad X_i \sim \mathcal{N}(0, I_2), \qquad y_i = \beta^\top X_i + \delta_i, \tag{21}$$

  - white noise δ_i's are i.i.d. scalars with $\xi = 1, \beta \in \mathbb{R}^2$
  - I_2 is the 2×2 identity matrix
- The prior distribution of β follows $\mathcal{N}(0, \lambda I_2)$, and we set $\lambda = 10$ for this set of experiments.
- The posterior distribution can be derived from the following model

$$\pi(\beta) \sim \mathcal{N}(m, V) \quad m := \left(\Sigma^{-1} + \frac{X^\top X}{\xi^2}\right)^{-1} \left(\frac{X^\top y}{\xi^2}\right),$$

$$V := \left(\frac{X^\top X}{\xi^2} + \Sigma^{-1}\right)^{-1},$$

Experimentally: EXTRA-SGLD vs DE-SGLD

- where $\Sigma = \lambda I_2$ is the covariance matrix of the prior of β
- $X = \left[X_1^\top, X_2^\top, \cdots\right]^\top$ and $Y = [y_1, y_2, \cdots]^\top$ are the input and output matrices, respectively.
- For this experiment, we simulate 5000 data points using the model (21) and then we distribute these data points randomly among the $N = 20$ agents.
- All agents have an equal amount of data exclusively, and share only the parameter estimates.

Experimentally: EXTRA-SGLD vs DE-SGLD

The posterior distribution $\pi(\beta) \propto e^{-f(\beta)}$ where $f(\beta) = \sum_{i=1}^{N} f_i(\beta)$ with

$$f_i(\beta) := -\sum_{j=1}^{n_i} \log p\left(y_j^i | \beta, X_j^i\right) - \frac{1}{N} \log p(\beta) = \sum_{j=1}^{n_i} \left(y_j^i - \beta^\top X_j^i\right)^2 + \frac{1}{2\lambda N} \|\beta\|^2,$$

where

$$p\left(y_j^i | \beta, X_j^i\right) = \frac{1}{\sqrt{2\pi\xi^2}} e^{-\frac{1}{2\xi^2}(y_j^i - \beta^\top X_j^i)^2}, \quad p(\beta) \propto e^{-\frac{1}{2\lambda}\|\beta\|^2},$$

and each agent i has an equal number of $n_i = 50$ data points $\{(X_j^i, y_j^i)\}_{j=1}^{n_i}$.

Experimentally: EXTRA-SGLD vs DE-SGLD

- We restrict the experiments with a deterministic gradient, i.e. $\sigma = 0$ and a fixed stepsize $\eta = 0.009$.
- The doubly stochastic mixing matrix $\widetilde{W} = hI_N - (1-h)W$ is calculated for different values of the parameter h.
- We consider 5 linearly spaced h values with the minimum being 0.001 and the maximum being 0.5 and we tune up the parameter h to the network. For the fully connected network $h = 0.50$, circular network $h = 0.38$, star network $h = 0.13$, and for the disconnected network $h = 0.38$.
- For each agent i, the iterations $\beta_i^{(k)} \sim \mathcal{N}\left(m_i^{(k)}, \Sigma_i^{(k)}\right)$ and we compute the Wasserstein-2 distance based on the formula given by Givens[8]

[8] Givens, Clark R., and Rae Michael Shortt. "A class of Wasserstein metrics for probability distributions." Michigan Mathematical Journal 31.2 (1984): 231-240.

Experimentally: EXTRA-SGLD vs DE-SGLD

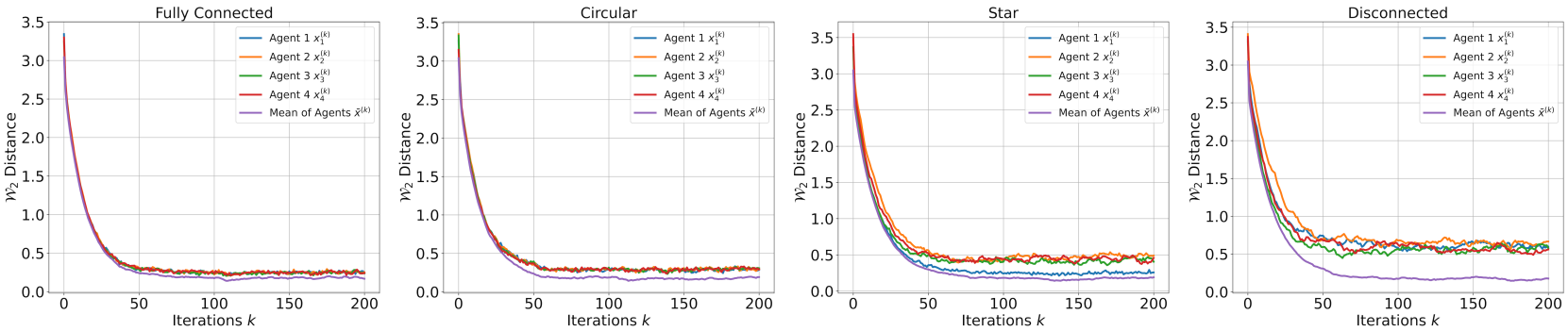


Figure 3: Performance of the EXTRA SGLD for Bayesian linear regression on four different network structures. Out of 20 agents, we report only the first 4 agents and the mean of the nodes $\bar{\beta}^{(k)} = \frac{1}{N}\sum_{i=1}^{N} \beta_i^{(k)}$.

Experimentally: EXTRA-SGLD vs DE-SGLD

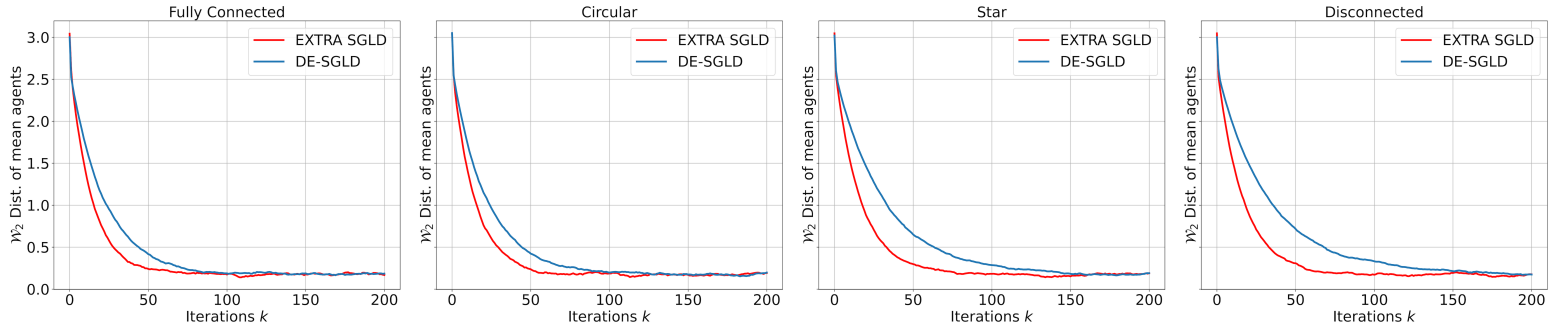


Figure 4: Comparative performance of the DE-SGLD and EXTRA SGLD for Bayesian linear regression on four different network structures in terms of the $\mathcal{W}_2$ distance of mean agents

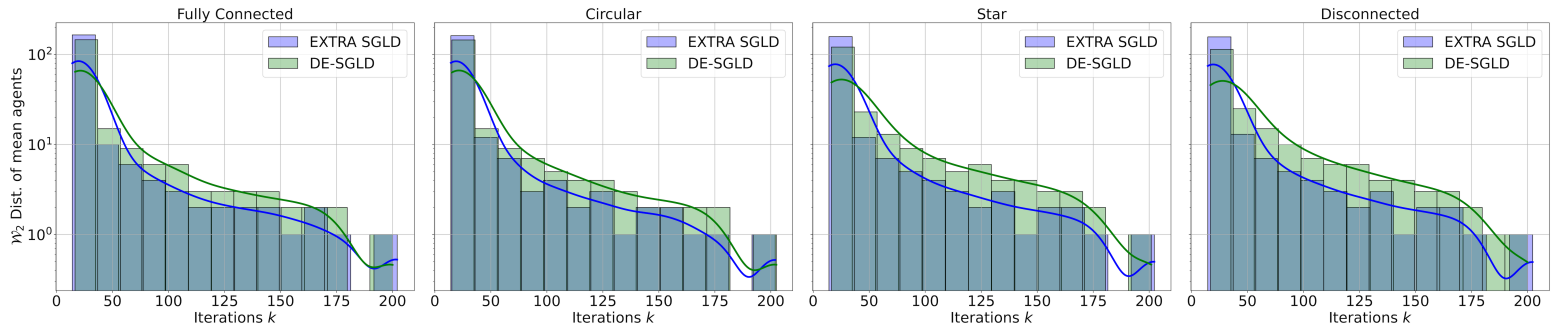


Figure 5: Histogram of the comparative performances of the DE-SGLD and EXTRA SGLD for Bayesian linear regression on four different network structures

Experimentally: EXTRA-SGLD vs DE-SGLD

- We have a dataset $Z = \{z_j\}_{j=1}^n$ where $z_j = (X_j, y_j), X_j \in \mathbb{R}^d$ are the features and $y_j \in \{0, 1\}$ are the labels with the assumption that X_j are independent.
- The probability distribution of y_j given X_j and regression coefficients $\beta \in \mathbb{R}^d$ is given by $\mathbb{P}(y_j = 1|X_j, \beta) = \frac{1}{1+e^{-\beta^\top X_j}}$.
- The prior distribution $p(\beta) \sim \mathcal{N}(0, \lambda I_d)$ for some $\lambda > 0$, where I_d is the $d \times d$ identity matrix.

Experimentally: EXTRA-SGLD vs DE-SGLD

- In a distributed network system, if each agent i contains a subset Z_i of data, then the goal of the Bayesian logistic regression is to sample from $\pi(\beta) \propto e^{-f(\beta)}$ with $f(\beta) = \sum_{i=1}^{N} f_i(\beta)$ where

$$\begin{aligned} f_i(\beta) &:= -\sum_{j=1}^{n_i} \log p\left(y_j^i = 1 | X_j^i, \beta\right) - \frac{1}{N} \log p(\beta) \\ &= \sum_{j=1}^{n_i} \log\left(1 + e^{-\beta^\top X_j}\right) + \frac{1}{2N\lambda}\|\beta\|^2 \end{aligned} \tag{22}$$

  is strongly convex and smooth.
- Data: Wisconsin Breast Cancer Data

Experimentally: EXTRA-SGLD vs DE-SGLD

- Data: Wisconsin Breast Cancer Data
- The data set contains 569 instances with 30 features which are computed from a digitized image of a fine needle aspirate (FNA) of a breast mass.
- For this experiment, we take:
  - $\eta = 0.005$
  - batch size 32
  - $N = 6$
- We take $h = 0.278, 0.389, 0.167$, and 0.278 for fully connected network, circular network, star network, and disconnected network, respectively.

Experimentally: EXTRA-SGLD vs DE-SGLD

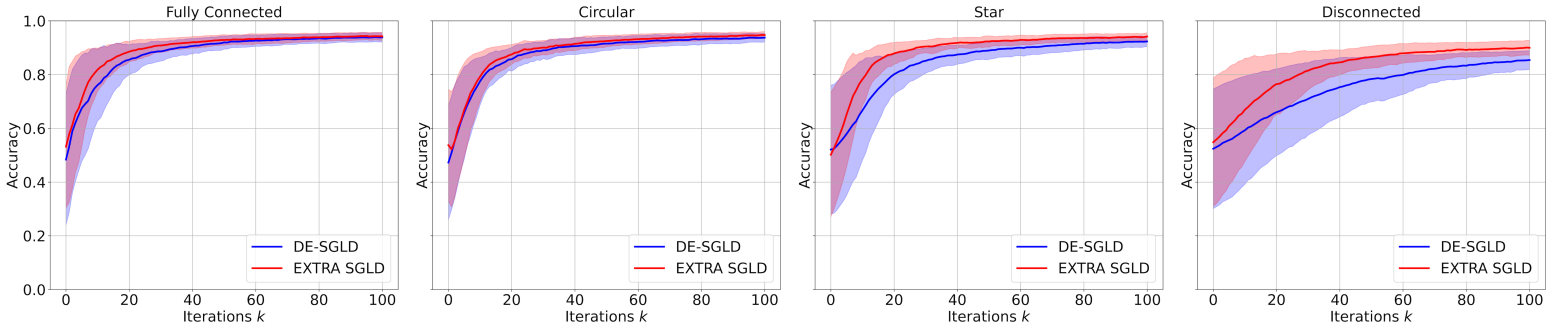


Figure 6: Comparative accuracy distribution of the DE-SGLD and EXTRA SGLD method across different network structures on Breast Cancer data set. The plots are from a randomly selected node.

Experimentally: EXTRA-SGLD vs DE-SGLD

	FCN	CN	SN	FDN (Cons.)	FDN (Ens.)
DE-SGLD	98.80%	98.79%	99.08%	11.79%	99.08%
EXTRA SGLD	98.91%	98.92%	98.96%	15.31%	99.18%

Table 1: MNIST classification accuracies across different network topologies

Bibliography


Mert Gurbuzbalaban, Mohammad Rafiqul Islam, Xiaoyu Wang, and Lingjiong Zhu.
Generalized extra stochastic gradient langevin dynamics.
arXiv preprint arXiv:2412.01993, 2024.