MAP 5932 Stochastic Optimization

FSU Mathematics Spring 2023 | Course Final Review Presentation.

Presented by Rafiq Islam

2023-10-05

An article from the Journal of Machine Learning Research

Article:

Decentralized Stochastic Gradient Langevin Dynamics and Hamiltonian Monte Carlo

- **Mert Gürbüzbalaban**
  Department of Management Science and Information Systems
  Rutgers Business School, Piscataway, NJ 008854, USA
- **Xuefeng Gao**
  Department of Systems Engineering and Engineering Management, The Chinese University of Hong Kong
- **Yuanhan Hu**
  Department of Management Science and Information Systems
  Rutgers Business School, Piscataway, NJ 008854, USA
- **Lingjiong Zhu**
  Department of Mathematics
  Florida State University, Tallahassee, FL 32306, USA

Introduction

- Recent decades have witnessed the era of big data, and there has been an exponential growth in the amount of data collected and stored with ever-increasing rates

Introduction

- Recent decades have witnessed the era of big data, and there has been an exponential growth in the amount of data collected and stored with ever-increasing rates
- Since the rate at which data is generated is often outpacing our ability to analyze it in terms of computational resources at hand, there has been a lot of recent interests for developing scaleable machine learning algorithms which are efficient on large datasets.

Introduction

- Recent decades have witnessed the era of big data, and there has been an exponential growth in the amount of data collected and stored with ever-increasing rates
- Since the rate at which data is generated is often outpacing our ability to analyze it in terms of computational resources at hand, there has been a lot of recent interests for developing scaleable machine learning algorithms which are efficient on large datasets.
- In the modern world, digital devices such as smart phones, tablets, wearables, sensors or video cameras are major sources of data generation.

Introduction (cont.)

- Often these devices are connected over a communication network (such as a wireless network or a sensor network) that has a high latency or a limited bandwidth.

Introduction (cont.)

- Often these devices are connected over a communication network (such as a wireless network or a sensor network) that has a high latency or a limited bandwidth.
- Because of communication constraints and privacy constraints, gathering all these data for centralized processing is often impractical or infeasible.

Introduction (cont.)

- Often these devices are connected over a communication network (such as a wireless network or a sensor network) that has a high latency or a limited bandwidth.
- Because of communication constraints and privacy constraints, gathering all these data for centralized processing is often impractical or infeasible.
- Decentralized machine learning algorithms have received a lot of attention for such applications where agents can collaboratively learn a predictive model without sharing their own data but sharing only their local models with their immediate neighbors at some frequency to generate a global model.

Introduction (cont.)

- In this presentation, we consider both distributed optimization and decentralized sampling problems

Introduction (cont.)

- In this presentation, we consider both distributed optimization and decentralized sampling problems
- In particular, for decentralized sampling algorithms, we propose and study decentralized stochastic gradient Langevin dynamics (DE-SGLD) and decentralized stochastic gradient Hamiltonian Monte Carlo (DE-SGHMC)

Introduction (cont.)

- In this presentation, we consider both distributed optimization and decentralized sampling problems
- In particular, for decentralized sampling algorithms, we propose and study decentralized stochastic gradient Langevin dynamics (DE-SGLD) and decentralized stochastic gradient Hamiltonian Monte Carlo (DE-SGHMC)
- For distributed optimization, we will consider distributed stochastic gradient (D-SG) and distributed accelerated stochastic gradient (D-ASG)

Decentralized Bayesian Inference

- Before introducing the DE-SGLD algorithm, we consider the problem of decentralized Bayesian inference.

Decentralized Bayesian Inference

- Before introducing the DE-SGLD algorithm, we consider the problem of decentralized Bayesian inference.
- We have N agents connected over a network $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ where $\mathcal{V} = \{1, 2, 3, \cdots, N\}$ 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 $(i, j) \in \mathcal{E}$ where the network is undirected, i.e., $(i, j) \in \mathcal{E}$ then $(j, i) \in \mathcal{E}$

Decentralized Bayesian Inference

- Before introducing the DE-SGLD algorithm, we consider the problem of decentralized Bayesian inference.
- We have N agents connected over a network $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ where $\mathcal{V} = \{1, 2, 3, \cdots, N\}$ 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 $(i, j) \in \mathcal{E}$ where the network is undirected, i.e., $(i, j) \in \mathcal{E}$ then $(j, i) \in \mathcal{E}$
- Let $A = [a_1, a_2, \cdots, a_n]$ be a dataset consisting of n independent and identically distributed (i.i.d) data vectors sampled from a parameterized distribution $p(A|x)$ where the parameter $x \in \mathbb{R}^d$ has a common prior distribution $p(x)$

Decentralized Bayesian Inference (cont.)

- Due to the decentralization in the data collection, each agent i possesses a subset A_i of the data where $A_i = \{a_1^i, a_2^i, \cdots .a_{ni}^i\}$ and n_i is the number of samples of agent i

Decentralized Bayesian Inference (cont.)

- Due to the decentralization in the data collection, each agent i possesses a subset A_i of the data where $A_i = \{a_1^i, a_2^i, \cdots .a_{ni}^i\}$ and n_i is the number of samples of agent i
- The data is held disjointly over agents, i.e., $A = \cup_i A_i$ with $A_i \cap A_j = \emptyset$ for $i \neq j$
- The goal is to sample from the posterior distribution

$$p(x|A) \propto p(A|x)p(x)$$

Decentralized Bayesian Inference (cont.)

- Due to the decentralization in the data collection, each agent i possesses a subset A_i of the data where $A_i = \{a_1^i, a_2^i, \cdots .a_{ni}^i\}$ and n_i is the number of samples of agent i
- The data is held disjointly over agents, i.e., $A = \cup_i A_i$ with $A_i \cap A_j = \emptyset$ for $i \neq j$
- The goal is to sample from the posterior distribution

$$p(x|A) \propto p(A|x)p(x)$$

- Since the data points are independent, the log-likelihood function will be additive;

$$\log p(A|x) = \sum_{i=1}^{N} \sum_{j=1}^{n_i} \log p(a_j^i|x)$$

Decentralized Bayesian Inference (cont.)

- Thus if we set

$$f(x) = \sum_{i=1}^{N} f_i(x), \quad f_i(x) = -\sum_{j=1}^{n_i} \log p(a_j^i|x) - \frac{1}{N} \log p(x) \quad (1)$$

the aim is to sample from the posterior distribution with density $\pi(x) = p(x|A) \propto e^{-f(x)}$

Decentralized Bayesian Inference (cont.)

- Thus if we set

$$f(x) = \sum_{i=1}^{N} f_i(x), \quad f_i(x) = -\sum_{j=1}^{n_i} \log p(a_j^i|x) - \frac{1}{N} \log p(x) \quad (1)$$

  the aim is to sample from the posterior distribution with density $\pi(x) = p(x|A) \propto e^{-f(x)}$
- The functions $f_i(x)$ are called "component functions" where $f_i(x)$ is associated to the local data of agent i and is only accessible by the agent i.

Decentralized Stochastic Gradient Langevin Dynamics (SGLD)

- Let $x_i^{(k)}$ denote the local variable of node i at iteration k

Decentralized Stochastic Gradient Langevin Dynamics (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 a stochastic gradient step over the node's component function $f_i(x)$, i.e.,

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

Decentralized Stochastic Gradient Langevin Dynamics (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 a stochastic gradient step over the node's component function $f_i(x)$, i.e.,

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

- $\eta > 0$ is the step size

Decentralized Stochastic Gradient Langevin Dynamics (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 a stochastic gradient step over the node's component function $f_i(x)$, i.e.,

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

- $\eta > 0$ is the step size
- W_{ij} are the entries of a doubly stochastic wight matrix W with $W_{ij} > 0$ only if i is connected to j

Decentralized Stochastic Gradient Langevin Dynamics (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 a stochastic gradient step over the node's component function $f_i(x)$, i.e.,

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

- $\eta > 0$ is the step size
- W_{ij} are the entries of a doubly stochastic wight matrix W with $W_{ij} > 0$ only if i is connected to j
- For example, we can take $W = I - \delta L$ where $\delta > 0$ and L is the graph Laplacian

Decentralized SGLD (cont.)

- $w_i^{(k)}$ are independent and identically distributed (i.i.d.) Gaussian random variables with zero mean and identity covariance matrix for every i and k.

Decentralized SGLD (cont.)

- $w_i^{(k)}$ are independent and identically distributed (i.i.d.) Gaussian random variables with zero mean and identity covariance matrix for every i and k.
- $\tilde{\nabla} f_i(x_i^{(k)})$ is an unbiased stochastic estimate of the deterministic gradient $\nabla f_i(x_i^{(k)})$ with a bounded variance.

Decentralized SGLD (cont.)

- $w_i^{(k)}$ are independent and identically distributed (i.i.d.) Gaussian random variables with zero mean and identity covariance matrix for every i and k.
- $\tilde{\nabla} f_i(x_i^{(k)})$ is an unbiased stochastic estimate of the deterministic gradient $\nabla f_i(x_i^{(k)})$ with a bounded variance.
- When the number of data points n_i is large, stochastic estimates $\tilde{\nabla} f_i(x_i^{(k)})$ are cheaper to compute compared to actual gradients $\nabla f_i(x_i^{(k)})$ and can for instance be estimated from a minibatch of data, i.e. from randomly selected smaller subsets of data. This allows the DE-SGLD method to be scaleable to big data settings when n_i can be large.

Decentralized SGLD (cont.)

- Our objective is to sample from a target distribution with density $\pi(x) \propto e^{-f(x)}$ on $\mathbb{R}^d$ where

$$f(x) := \sum_{i=1}^{N} f_i(x) \tag{3}$$

Decentralized SGLD (cont.)

- Our objective is to sample from a target distribution with density $\pi(x) \propto e^{-f(x)}$ on $\mathbb{R}^d$ where

$$f(x) := \sum_{i=1}^{N} f_i(x) \tag{3}$$

- We assume for every $i = 1, 2, \cdots, N$, f_i is $\mu-$strongly convex and $L-$smooth, that is for every $x, y \in \mathbb{R}^d$

$$\frac{L}{2}||x-y||^2 \geq f_i(x) - f_i(y) - \nabla f_i(y)^T(x-y) \geq \frac{\mu}{2}||x-y||^2 \tag{4}$$

Decentralized SGLD (cont.)

Assumption 1

We assume that the gradient noise defined as

$$\xi_i^{(k+1)} := \tilde{\nabla} f_i\left(x_i^{(k)}\right) - \nabla f_i\left(x_i^{(k)}\right) \tag{5}$$

is unbiased with a finite second moment, i.e.,

$$\mathbb{E}\left[\xi_i^{(k+1)} \middle| \mathcal{F}_k\right] = 0, \quad \mathbb{E}\left\|\xi_i^{(k+1)}\right\|^2 \leq \sigma^2 \tag{6}$$

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

Decentralized SGLD (cont.)

- Based on (5), we can rewrite the DE-SGLD iterations in (2) in terms of the gradient noise $\xi_i^{(k+1)}$ as

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

Decentralized SGLD (cont.)

- Based on (5), we can rewrite the DE-SGLD iterations in (2) in terms of the gradient noise $\xi_i^{(k+1)}$ as

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

- By defining the column vector

$$x^{(k)} := \left[\left(x_1^{(k)}\right)^T, \left(x_2^{(k)}\right)^T, \cdots, \left(x_N^{(k)}\right)^T\right]^T \in \mathbb{R}^{Nd}$$

concetenates the local decision variables into a single vector, we can express the DE-SGLD iterations further as

$$x^{(k+1)} = \mathcal{W} x^k - \eta \nabla F\left(x^{(k)}\right) - \eta \xi^{(k+1)} + \sqrt{2\eta} w^{(k+1)} \quad (7)$$

with $\mathcal{W} = W \otimes I_d$, and
$F(x) := F(x_1, x_2, \cdots, x_N) = \sum_{i=1}^N f_i(x_i)$

Decentralized SGLD (cont.)

- In equation (7),

$$w^{(k+1)} := \left[\left(w_1^{(k)} \right)^T, \left(w_2^{(k)} \right)^T, \cdots, \left(w_N^{(k)} \right)^T \right]^T$$

are i.i.d. Gaussian noise with mean 0 and with a covariance matrix given by the identity matrix.

Decentralized SGLD (cont.)

- In equation (7),

$$w^{(k+1)} := \left[\left(w_1^{(k)}\right)^T, \left(w_2^{(k)}\right)^T, \cdots, \left(w_N^{(k)}\right)^T\right]^T$$

are i.i.d. Gaussian noise with mean 0 and with a covariance matrix given by the identity matrix.

- In equation (7),

$$\xi^{(k+1)} := \left[\left(\xi_1^{(k)}\right)^T, \left(\xi_2^{(k)}\right)^T, \cdots, \left(\xi_N^{(k)}\right)^T\right]^T$$

are the gradient nose so that

$$\mathbb{E}\left[\xi^{(k+1)}\middle|\mathcal{F}_k\right] = 0, \quad \mathbb{E}\|\xi^{(k+1)}\|^2 \leq \sigma^2 N \tag{8}$$

Decentralized SGLD (cont.)

- Let us define the average at k-th iteration

$$\bar{x}^{(k)} := \frac{1}{N} \sum_{i=1}^{N} x_i^{(k)}$$

Decentralized SGLD (cont.)

- Let us define the average at k-th iteration

$$\bar{x}^{(k)} := \frac{1}{N}\sum_{i=1}^{N} x_i^{(k)}$$

- Since $\mathcal{W}$ is doubly stochastic, we get

$$\bar{x}^{(k+1)} = \bar{x}^{(k)} - \eta\frac{1}{N}\sum_{i=1}^{N}\nabla f_i\left(x_i^{(k)}\right) - \eta\bar{\xi}^{(k+1)} + \sqrt{2\eta}\bar{w}^{(k+1)} \quad (9)$$

Decentralized SGLD (cont.)

- Let us define the average at k-th iteration

$$\bar{x}^{(k)} := \frac{1}{N}\sum_{i=1}^{N} x_i^{(k)}$$

- Since $\mathcal{W}$ is doubly stochastic, we get

$$\bar{x}^{(k+1)} = \bar{x}^{(k)} - \eta\frac{1}{N}\sum_{i=1}^{N}\nabla f_i\left(x_i^{(k)}\right) - \eta\bar{\xi}^{(k+1)} + \sqrt{2\eta}\bar{w}^{(k+1)} \quad (9)$$

$$\bar{w}^{(k+1)} := \frac{1}{N}\sum_{i=1}^{N} w_i^{(k+1)} \sim \frac{1}{\sqrt{N}}\mathcal{N}(0, I_d), \quad \bar{\xi}^{(k+1)} := \frac{1}{N}\sum_{i=1}^{N}\xi_i^{(k+1)} \quad (10)$$

that satisfies

$$\mathbb{E}\left[\bar{\xi}^{(k+1)}\middle|\mathcal{F}_k\right] = 0, \quad \mathbb{E}\|\bar{\xi}^{(k+1)}\|^2 \leq \frac{\sigma^2}{N} \quad (11)$$

Decentralized SGLD (cont.)

- We now state the main result of DE-SGLD, which bounds the average of $\mathcal{W}_2$ distance between the distribution of $x_i^{(k)}$ and the target distribution π (that has a density proportional to $exp(-f(x))$) over $1 \leq i \leq N$.

Decentralized SGLD (cont.)

- We now state the main result of DE-SGLD, which bounds the average of $\mathcal{W}_2$ distance between the distribution of $x_i^{(k)}$ and the target distribution π (that has a density proportional to $exp(-f(x))$) over $1 \leq i \leq N$.
- This result provides also a bound on the $\mathcal{W}_2$ distance of the node averages $\bar{x}^{(k)}$ and the target distribution π

Decentralized SGLD (cont.)

- We now state the main result of DE-SGLD, which bounds the average of $\mathcal{W}_2$ distance between the distribution of $x_i^{(k)}$ and the target distribution π (that has a density proportional to $exp(-f(x))$) over $1 \leq i \leq N$.
- This result provides also a bound on the $\mathcal{W}_2$ distance of the node averages $\bar{x}^{(k)}$ and the target distribution π
- To facilitate the presentation, we define the second largest magnitude of the eigenvalues of W as

$$\bar{\gamma} := \max\left\{\left|\lambda_2^W\right|, \left|\lambda_N^W\right|\right\} \tag{12}$$

which is related to the connectivity of the graph $\mathcal{G}$

Theorem 1

Assume $\mathbb{E}||x^{(0)}||^2 < \infty$ and $\eta \in (0, \bar{\eta})$ where $\bar{\eta} = \min\left(\frac{1+\lambda_N^W}{L}, \frac{1}{L+\mu}\right)$. Then, for every k, DE-SGLD iterates $x_i^{(k)}$ given by (2) and their average $\bar{x}^{(k)}$ satisfy

$$\mathcal{W}_2\left(\mathcal{L}\left(\bar{x}^{(k)}\right), \pi\right)$$

$$\leq (1-\mu\eta)^k \left(\sqrt{\mathbb{E}\|\bar{x}^{(0)} - x_*\|^2} + \sqrt{2\mu^{-1}dN^{-1}}\right)$$

$$+ \left(\bar{\gamma}^2 \frac{1 - \eta\mu\left(1-\frac{\eta L}{2}\right)^k - \bar{\gamma}^{2k}}{1 - \eta\mu\left(1-\frac{\eta L}{2}\right) - \bar{\gamma}^2}\right)^{1/2} \frac{2L}{\sqrt{N}} \left(\mathbb{E}||x^{(0)}||^2\right)^{1/2} + \sqrt{\eta} E_1$$

where $E_1 := \frac{1.65L}{\mu}\sqrt{dN^{-1}} + \frac{\sigma}{\sqrt{\mu(1-\frac{\eta L}{2})N}}$

$$+ \left(\frac{\eta}{\mu(1-\frac{\eta L}{2})} + \frac{}{\mu^2(1-\frac{\eta L}{2})^2}\right)^{1/2} \cdot \left(\frac{4L^2D^2\eta}{N(1-\bar{\gamma}^2)} + \frac{4L^2D^2\eta}{(1-\bar{\gamma}^2)} + \frac{8L^2d}{(1-\bar{\gamma}^2)}\right)$$

Theorem 1 (cont.)

Furthermore,

$$\frac{1}{N}\sum_{i=1}^{N}\mathcal{W}_2\left(\mathcal{L}\left(\bar{x}^{(k)}\right),\pi\right)$$
$$\leq (1-\mu\eta)^k\left(\sqrt{\mathbb{E}\|\bar{x}^{(0)}-x_*\|^2}+\sqrt{2\mu^{-1}dN^{-1}}\right)+\frac{2\bar{\gamma}^k}{\sqrt{N}}\left(\mathbb{E}||x^{(0)}||^2\right)^{1/2}$$
$$+\left(\bar{\gamma}^2\frac{1-\eta\mu\left(1-\frac{\eta L}{2}\right)^k-\bar{\gamma}^{2k}}{1-\eta\mu\left(1-\frac{\eta L}{2}\right)-\bar{\gamma}^2}\right)^{1/2}\frac{2L}{\sqrt{N}}\left(\mathbb{E}||x^{(0)}||^2\right)^{1/2}+\sqrt{\eta}E_2+\eta E_3 \tag{13}$$

with $E_2 = E_1 + \frac{2\sqrt{2d}}{\sqrt{1-\bar{\gamma}^2}}$ and $E_3 = \frac{2D}{\sqrt{N}(1-\bar{\gamma})} + \frac{2\sigma}{\sqrt{1-\bar{\gamma}^2}}$, where x_* is the minimizer of $f, \bar{x}^{(0)} = \frac{1}{N}\sum_{i=1}^{N} x_i^{(0)}$, and D is defined in (16)

Discussions

- We observe that

$$\lim_{k\to\infty} \sup \mathcal{W}_2 \left(\mathcal{L}\left(\bar{x}^{(k)}\right), \pi \right) = \mathcal{O}(\sqrt{\eta})$$

where $\mathcal{O}(.)$ hides other constants (d, μ, L, σ, N) and $\bar{\gamma}$

Discussions

- We observe that

$$\lim_{k\to\infty} \sup \mathcal{W}_2 \left(\mathcal{L}\left(\bar{x}^{(k)}\right), \pi \right) = \mathcal{O}(\sqrt{\eta})$$

  where $\mathcal{O}(.)$ hides other constants (d, μ, L, σ, N) and $\bar{\gamma}$
- With the iteration budget K, we can choose $\eta = \frac{c \log \sqrt{K}}{\mu K}$ for a constant $c > 1$

Discussions

- We observe that

$$\lim_{k\to\infty} \sup \mathcal{W}_2 \left(\mathcal{L}\left(\bar{x}^{(k)}\right), \pi \right) = \mathcal{O}(\sqrt{\eta})$$

  where $\mathcal{O}(.)$ hides other constants (d, μ, L, σ, N) and $\bar{\gamma}$
- With the iteration budget K, we can choose $\eta = \frac{c \log \sqrt{K}}{\mu K}$ for a constant $c > 1$
- Consequently,

$$\mathcal{W}_2(\mathcal{L}(\bar{x}^{(2k)}), \pi) = \mathcal{O}\left(\frac{1}{\left(\sqrt{K}\right)^c} + \frac{\sqrt{c \log K}}{\sqrt{K}} \right) = \mathcal{O}\left(\frac{\sqrt{\log K}}{\sqrt{K}} \right)$$

  where the last $\mathcal{O}(.)$ term hides constants that depends on $x^{(0)}, d, \mu, L, \sigma, \bar{\gamma}, N$ and c

Outline of the Proof

- To facilitate the analysis, let us define xk from the iterates:

$$x_{k+1} = x_k - \eta \frac{1}{N} \nabla f(x_k) + \sqrt{2\eta} \bar{w}^{(k+1)} \tag{14}$$

where $x_0 = \bar{x}_0 = \frac{1}{N} \sum_{i=1}^{N} x_i^{(0)}$

Outline of the Proof

- To facilitate the analysis, let us define xk from the iterates:

$$x_{k+1} = x_k - \eta \frac{1}{N} \nabla f(x_k) + \sqrt{2\eta} \bar{w}^{(k+1)} \tag{14}$$

where $x_0 = \bar{x}_0 = \frac{1}{N} \sum_{i=1}^{N} x_i^{(0)}$

- This is an Euler-Maruyama discretization (with stepsize η) of the continuous-time overdamped Langevin diffusion:

$$dX_t = -\frac{1}{N} \nabla f(X_t) dt + \sqrt{2N^{-1}} dW_t \tag{15}$$

wher W_t is a standard d-dimensional Brownian motion

Outline of the Proof (cont.)

- To bound the average of $\mathcal{W}_2$ distance between $\mathcal{L}(x_i^{(k)})$ and π over $1 \leq i \leq N$, main idea of our proof technique is to bound the following three terms:

Outline of the Proof (cont.)

- To bound the average of $\mathcal{W}_2$ distance between $\mathcal{L}(x_i^{(k)})$ and π over $1 \leq i \leq N$, main idea of our proof technique is to bound the following three terms:

1. The L^2 distance between $\bar{x}^{(k)}$ and their average (mean)

$$\bar{x}^{(k)} = \frac{\sum_{i=1}^{N}}{N}$$

for $1 \leq i \leq N$

Outline of the Proof (cont.)

- To bound the average of $\mathcal{W}_2$ distance between $\mathcal{L}(x_i^{(k)})$ and π over $1 \leq i \leq N$, main idea of our proof technique is to bound the following three terms:

1. The L^2 distance between $\bar{x}^{(k)}$ and their average (mean)

$$\bar{x}^{(k)} = \frac{\sum_{i=1}^{N}}{N}$$

for $1 \leq i \leq N$

2. The L^2 distance between the average iterate $\bar{x}^{(k)}$ and iterates x_k obtained from Euler-Maruyama discretization of overdamped Langevin SDE; and

Outline of the Proof (cont.)

- To bound the average of $\mathcal{W}_2$ distance between $\mathcal{L}(x_i^{(k)})$ and π over $1 \leq i \leq N$, main idea of our proof technique is to bound the following three terms:

1. The L^2 distance between $\bar{x}^{(k)}$ and their average (mean)

$$\bar{x}^{(k)} = \frac{\sum_{i=1}^{N}}{N}$$

   for $1 \leq i \leq N$
2. The L^2 distance between the average iterate $\bar{x}^{(k)}$ and iterates x_k obtained from Euler-Maruyama discretization of overdamped Langevin SDE; and
3. The $\mathcal{W}_2$ distance between $\mathcal{L}(x_k)$ and π, i.e., the convergence of Euler-Maruyama discretization of the overdamped Langevin SDE.

Uniform L^2 bounds between $x_i^{(k)}$ and their average

Let $x_* \in \mathbb{R}^d$ denote the unique minimizer of $f(x)$, and $x^* = [x_*^T, x_*^T, x_*^T, \cdots, x_*^T]^T$ is an Nd-dimensional vector.

Lemma 2

Under the assumptions of Theorem 1, we have, $\mathbb{E}\|\nabla F(x^{(k)})\|^2 \leq D^2$ for an k, where

$$D^2 = 4L^2\mathbb{E}\|x^{(0)} - x^*\|^2 + 8L^2\frac{C_1^2\eta^2 N}{(1-\bar{\gamma})^2} + \frac{2L^2(\eta\sigma^2 N + 2dN)}{\mu(1+\lambda_N^W - \eta L)} + 4\|\nabla F(x^*)\|^2 \tag{16}$$

where

$$C_1 = \bar{C}_1\left(1 + \frac{2(L+\mu)}{\mu}\right), \text{ and } \bar{C}_1 = \sqrt{2L\sum_{i=1}^{N}(f_i(0) - f_i^*)}, f_i^* = \min_{x\in\mathbb{R}^d} f_i(x) \tag{17}$$

Outline of the Proof (cont.)

- It is clear from the DE-SGLD iterations that the deviations between the iterates $x_i^{(k)}$ and their means $\bar{x}^{(k)}$ depend on the magnitude of the gradients $\nabla F(x^{(k)})$, the stepsize as well as the magnitude of the injected Gaussian noise.

Outline of the Proof (cont.)

- It is clear from the DE-SGLD iterations that the deviations between the iterates $x_i^{(k)}$ and their means $\bar{x}^{(k)}$ depend on the magnitude of the gradients $\nabla F(x^{(k)})$, the stepsize as well as the magnitude of the injected Gaussian noise.
- Building on Lemma 2 which gives us a control over the second moment of the gradients, we provide uniform L_2 bounds between the iterates $x_i^{(k)}$ and their means.

Outline of the Proof (cont.)

- It is clear from the DE-SGLD iterations that the deviations between the iterates $x_i^{(k)}$ and their means $\bar{x}^{(k)}$ depend on the magnitude of the gradients $\nabla F(x^{(k)})$, the stepsize as well as the magnitude of the injected Gaussian noise.
- Building on Lemma 2 which gives us a control over the second moment of the gradients, we provide uniform L_2 bounds between the iterates $x_i^{(k)}$ and their means.

Lemma 3

Under the assumptions of Theorem 1, for any k, we have

$$\sum_{i=1}^{N} \mathbb{E}\|x_i^{(k)} - \bar{x}^{(k)}\|^2 \leq 4\bar{\gamma}^{2k}\mathbb{E}\|x^{(0)}\|^2 + \frac{4D^2\eta^2}{(1-\bar{\gamma})^2} + \frac{4\sigma^2 N\eta^2}{(1-\bar{\gamma}^2)} + \frac{8dN\eta}{(1-\bar{\gamma}^2)}$$

where D is defined in (16) and $\bar{\gamma}$ is given in (12)

Outline of the Proof (cont.)

Note that we can deduce from (9) that

$$\bar{x}^{(k+1)} = \bar{x}^{(k)} - \eta \frac{1}{N} \nabla f(\bar{x}^{(k)}) + \eta \mathcal{E}_{k+1} - \eta \bar{\xi}^{(k+1)} + \sqrt{2\eta} \bar{w}^{(k+1)} \quad (18)$$

Outline of the Proof (cont.)

Note that we can deduce from (9) that

$$\bar{x}^{(k+1)} = \bar{x}^{(k)} - \eta \frac{1}{N} \nabla f(\bar{x}^{(k)}) + \eta \mathcal{E}_{k+1} - \eta \bar{\xi}^{(k+1)} + \sqrt{2\eta} \bar{w}^{(k+1)} \quad (18)$$

where

$$\mathcal{E}_{k+1} = \frac{1}{N} \sum_{i=1}^{N} \left[\nabla f_i(\bar{x}^{(k)}) - \nabla f_i(x_i^{(k)}) \right] \quad (19)$$

Lemma 4

Under the assumptions of Theorem 1, for any k, we have

$$\mathbb{E}\|\mathcal{E}_{k+1}\|^2 \leq \frac{4\bar{\gamma}^{2k}}{N} \mathbb{E}\|x^{(0)}\|^2 + \frac{4L^2 D^2 \eta^2}{N(1-\bar{\gamma})^2} + \frac{4L^2 \sigma^2 \eta^2}{(1-\bar{\gamma}^2)} + \frac{8d\eta}{(1-\bar{\gamma}^2)}$$

where $\mathcal{E}_{k+1}$ is defined in (19)

L^2 distance between the mean and the discretized overdamped SDE

Lemma 5

Under the assumptions of Theorem 1, for every k,

$$\mathbb{E}\|\bar{x}^{(k)} - x_k\|^2$$

$$\leq \eta \left(\frac{\eta}{\mu(1-\frac{\eta L}{2})} + \frac{(1+\eta L)^2}{\mu^2(1-\frac{\eta L}{2})^2} \right) \left(\frac{4L^2 D^2 \eta}{N(1-\bar{\gamma})^2} + \frac{4L^2\sigma^2\eta}{(1-\bar{\gamma}^2)} + + \frac{8L^2 d}{(1-\bar{\gamma}^2)} \right)$$

$$+ \frac{\eta\sigma^2}{\mu(1-\frac{\eta L}{2})N} + \frac{\bar{\gamma}^{2k} - \left(1-\eta\mu(1-\frac{\eta L}{2})\right)^k}{\bar{\gamma}^2 - 1 + \eta\mu(1-\frac{\eta L}{2})} \cdot \frac{4L^2\bar{\gamma}^2}{N} \mathbb{E}\|x^{(0)}\|^2$$

$\mathcal{W}_2$ Distance between the iterates and the Gibbs distribution

Bounds on the $\mathcal{W}_2$ distance between the Euler-Maruyama discretization x_k of the overdamped Langevin diffiusion and Gibbs distribution π has been established in the literature[1]

[1] Dalalyan, A.S. and Karagulyan, A.G. (2019). User-friendly guarantees for the Langevin Monte Carlo with inaccurate gradient. Stochastic Processes and their Applications. 129(12), 5278-5311

$\mathcal{W}_2$ Distance between the iterates and the Gibbs distribution

Bounds on the $\mathcal{W}_2$ distance between the Euler-Maruyama discretization x_k of the overdamped Langevin diffiusion and Gibbs distribution π has been established in the literature[1]

Lemma 6

For any $\eta \in \left(0, \frac{2N}{L+\mu}\right]$, we have

$$\mathcal{W}_2(\mathcal{L}(x_k), \pi) \leq (1 - \mu n)^k \mathcal{W}_2(\mathcal{L}(x_0), \pi) + \frac{1.65L}{\mu}\sqrt{\eta d N^{-1}}$$

[1] Dalalyan, A.S. and Karagulyan, A.G. (2019). User-friendly guarantees for the Langevin Monte Carlo with inaccurate gradient. Stochastic Processes and their Applications. 129(12), 5278-5311

Decentralized SGHLMC

- We introduce the following algorithm which we call decentralized stochastic gradient Hamiltonian Monte Carlo (DE-SGHMC): For each agent $i = 1, 2, \cdots, N$,

$$v_i^{(k+1)} = v_i^{(k)} - \eta \left[\gamma v_i^{(k)} + \tilde{\nabla} f_i \left(x_i^{(k)} \right) \right] + \sqrt{2\gamma\eta} w_i^{(k+1)} \quad (20)$$

$$x_i^{(k+1)} = \sum_{j \in \Omega_i} W_{ij} x_j^{(k)} + \eta v_i^{(k+1)} \quad (21)$$

where $w_i^{(k+1)}$ is the Gaussian noise and $\tilde{\nabla} f_i$ is the noisy gradient introduced just as before for DE-SGLD.

Decentralized SGHLMC (cont.)

- Let us define the average at k-th iteration as:

$$\bar{x}^{(k)} = \frac{1}{N}\sum_{i=1}^{N} x_i^{(k)}, \quad \bar{v}^{(k)} = \frac{1}{N}\sum_{i=1}^{N} v_i^{(k)} \tag{22}$$

Decentralized SGHLMC (cont.)

- Let us define the average at k-th iteration as:

$$\bar{x}^{(k)} = \frac{1}{N}\sum_{i=1}^{N} x_i^{(k)}, \quad \bar{v}^{(k)} = \frac{1}{N}\sum_{i=1}^{N} v_i^{(k)} \tag{22}$$

- Since $\mathcal{W}$ is doubly stochastic, we get

$$\bar{v}^{(k+1)} = \bar{v}^{k} - \eta\gamma\bar{v}^{(k)} - \eta\frac{1}{N}\sum_{i=1}^{N}\nabla f_i\left(x_i^{(k)}\right) - \eta\bar{\xi}^{(k+1)} + \sqrt{2\gamma\eta}\,\bar{w}^{(k+1)}$$

$$\bar{x}^{(k+1)} = \bar{x}^{(k)} + \eta\bar{v}^{(k+1)}$$

- And, $\bar{\xi}^{(k+1)} = \frac{1}{N}\sum_{i=1}^{N}\xi_i^{(k+1)}$, $\bar{w}^{(k+1)} = \frac{1}{N}\sum_{i=1}^{N} w_i^{(k+1)}$

Decentralized SGHLMC (cont.)

Theorem 7

Assume $\mathbb{E}\|x^{(0)}\|^2$ and $\mathbb{E}\|v^{(0)}\|^2$ are finite. Let η be given satisfying

$$\eta^2 \in \left(0, \frac{1+\lambda_N^W}{2(L+\mu)}\right) \tag{23}$$

Then, we can choose $\gamma \in (0, \frac{1}{\eta}]$ such that $\beta = 1-\gamma\eta \in [0,1)$ and satisfies. the inequality

$$\beta \leq \bar{\beta} = \min\left(\frac{1+\lambda_N^W - 4\eta^2\mu}{4}, \eta^3\sqrt{c_1\mu^3\frac{(1+\lambda_N^W)}{64}}\right) \tag{24}$$

where

$$c_1 = \frac{1}{2}\frac{\eta^2\mu}{(1+\beta)+(1-\beta)\left(\frac{\eta^2\mu}{1-\lambda_N^W+\eta^2 L}\right)}$$

Decentralized SGHLMC (cont.)

Theorem 7 (cont.)

For every k, DE-SGHMC iterates $x_i^{(k)}$ given by (21) and their average $\bar{x}^{(k)}$ satisfy

$$\begin{aligned}
&\mathcal{W}_2\left(\mathcal{L}\left(\bar{x}^{(k)}\right), \pi\right) \\
&\leq (1-\mu\eta^2)^k \left(\left(\mathbb{E}\|\bar{x}^{(0)} - x_*\|^2\right)^{1/2} + \sqrt{2\mu^{-1}} dN^{-1}\right) \\
&+ \left(\bar{\gamma}^2 \frac{\left(1-\eta^2\mu\left(1-\frac{\eta^2 L}{2}\right)\right)^k - \bar{\gamma}^{2k}}{\left(1-\eta^2\mu\left(1-\frac{\eta^2 L}{2}\right)\right) - \bar{\gamma}^2}\right)^{1/2} \frac{2L}{\sqrt{N}} \left(\mathbb{E}\|x^{(0)}\|^2\right)^{1/2} + \eta E_4
\end{aligned} \tag{25}$$

with $E_4 = \mathcal{O}(1)$

Decentralized SGHLMC (cont.)

Theorem 7 (cont.)

Furthermore,

$$\frac{1}{N}\sum_{1}^{N}\mathcal{W}_2\left(\mathcal{L}\left(x^{(k)}\right),\pi\right)$$

$$\leq (1-\mu\eta^2)^k\left(\left(\mathbb{E}\|\bar{x}^{(0)}-x_*\|^2\right)^{1/2}+\sqrt{2\mu^{-1}dN^{-1}}\right)+\frac{\sqrt{2}\bar{\gamma}^k}{\sqrt{N}}\left(\mathbb{E}\|x^{(0)}\|^2\right)$$

$$+\left(\frac{\left(1-\eta^2\mu\left(1-\frac{\eta^2 L}{2}\right)\right)^k-\bar{\gamma}^{2k}}{\left(1-\eta^2\mu\left(1-\frac{\eta^2 L}{2}\right)\right)-\bar{\gamma}^2}\right)^{1/2}\frac{2L\bar{\gamma}}{\sqrt{N}}\left(\mathbb{E}\|x^{(0)}\|^2\right)^{1/2}+\eta E_5 \tag{26}$$

with $E_5=\mathcal{O}(1)$, and $\beta=\mathcal{O}(\eta^4)$ where $\mathcal{O}(.)$ hides the constants that depend on d,μ,L,σ, and $\bar{\gamma}$ and N

Discussions

- We observe that

$$\lim_{k\to\infty} \sup \mathcal{W}_2 \left(\mathcal{L}\left(\bar{x}^{(k)}\right), \pi \right) = \mathcal{O}(n)$$

where $\mathcal{O}(.)$ hides other constants ($d, \mu, L, \sigma, N,$ and $\bar{\gamma}$)

Discussions

- We observe that

$$\lim_{k\to\infty} \sup \mathcal{W}_2 \left(\mathcal{L}\left(\bar{x}^{(k)}\right), \pi \right) = \mathcal{O}(n)$$

  where $\mathcal{O}(.)$ hides other constants (d, μ, L, σ, N, and $\bar{\gamma}$)
- With the iteration K, we can choose $\eta = \sqrt{\frac{c \log \sqrt{K}}{\mu K}}$ for a constant $c > 1$

Discussions

- We observe that

$$\lim_{k\to\infty} \sup \mathcal{W}_2\left(\mathcal{L}\left(\bar{x}^{(k)}\right), \pi\right) = \mathcal{O}(n)$$

  where $\mathcal{O}(.)$ hides other constants (d, μ, L, σ, N, and $\bar{\gamma}$)
- With the iteration K, we can choose $\eta = \sqrt{\frac{c \log \sqrt{K}}{\mu K}}$ for a constant $c > 1$
- Consequently,

$$\mathcal{W}_2\left(\mathcal{L}\left(\bar{x}^{(2k)}\right), \pi\right) = \mathcal{O}\left(\sqrt{\frac{\sqrt{\log K}}{\sqrt{K}}}\right)$$

  where the last $\mathcal{O}(\cdot)$ term hides constants that depends on $x^{(0)}, v^{(0)}, d, \mu, L, \sigma, \bar{\gamma}, N$ and c

Numerical Experiments

- We present our numerical results to validate our theory and investigate the performance of DE-SGLD and DE-SGHMC

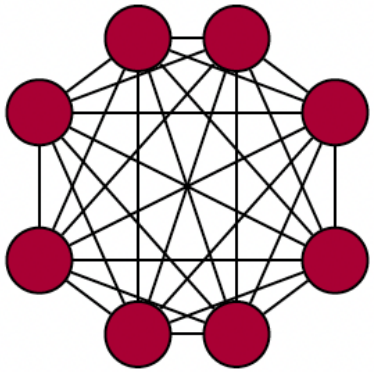
Numerical Experiments

- We present our numerical results to validate our theory and investigate the performance of DE-SGLD and DE-SGHMC
- We mainly focus on Bayesian linear regression and Bayesian logistic regression

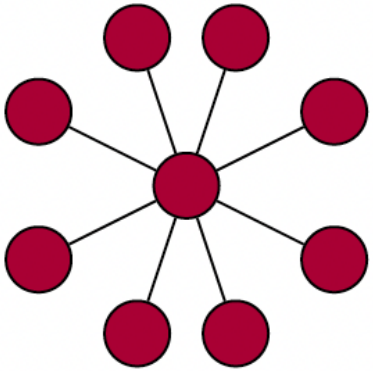
Numerical Experiments

- We present our numerical results to validate our theory and investigate the performance of DE-SGLD and DE-SGHMC
- We mainly focus on Bayesian linear regression and Bayesian logistic regression
- We consider mainly three network architectures
  - Fully-connected
  - Circular
  - A disconnected

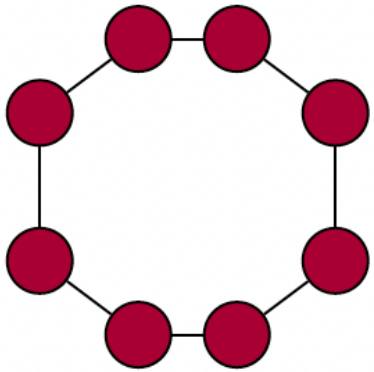
Network Architecture



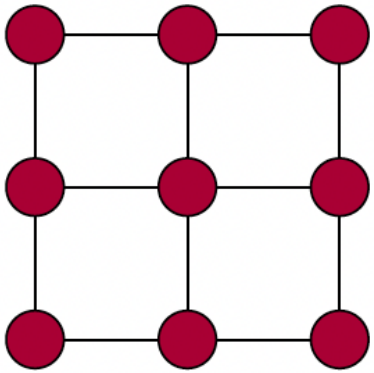
(a) Fully-connected



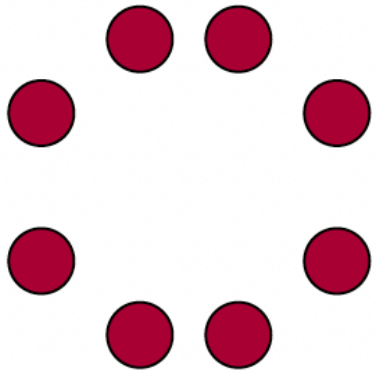
(b) Star



(c) Circular



(d) Grid



(e) Disconnected

Experiment Design

- Data

$$\delta_j \sim \mathcal{N}(0, \xi^2), \quad X_j \sim \mathcal{N}(0, I), \quad y_j = x^T X_j + \delta_j$$

where δ_j are i.i.d scalars with $\xi = 1, x \in \mathbb{R}^2$ and the prior distribution of $x \sim \mathcal{N}(0, \lambda I)$ with $\lambda = 10$

Experiment Design

- Data

$$\delta_j \sim \mathcal{N}(0, \xi^2), \quad X_j \sim \mathcal{N}(0, I), \quad y_j = x^T X_j + \delta_j$$

where δ_j are i.i.d scalars with $\xi = 1, x \in \mathbb{R}^2$ and the prior distribution of $x \sim \mathcal{N}(0, \lambda I)$ with $\lambda = 10$

- For Bayesian Linear regression we have the posterior distribution

$$\pi(x) \sim \mathcal{N}(m, V), \quad m = (\Sigma^{-1} + X^T X/\xi^2)^{-1}(X^T y/\xi^2)$$

,

$$V = (X^T X/\xi^2 + \Sigma^{-1})^{-1}$$

where $\Sigma = \lambda I$ is the covariance matrix of the prior distribution of x, $X = [X_1^T, X_2^T, X_3^T, \cdots]^T$ and $Y = [y_1, y_2, \cdots]^T$ are the matrices containing all the data points.

Experiment Design

- Data

$$\delta_j \sim \mathcal{N}(0, \xi^2), \quad X_j \sim \mathcal{N}(0, I), \quad y_j = x^T X_j + \delta_j$$

where δ_j are i.i.d scalars with $\xi = 1, x \in \mathbb{R}^2$ and the prior distribution of $x \sim \mathcal{N}(0, \lambda I)$ with $\lambda = 10$

- For Bayesian Linear regression we have the posterior distribution

$$\pi(x) \sim \mathcal{N}(m, V), \quad m = (\Sigma^{-1} + X^T X/\xi^2)^{-1}(X^T y/\xi^2)$$

,

$$V = (X^T X/\xi^2 + \Sigma^{-1})^{-1}$$

where $\Sigma = \lambda I$ is the covariance matrix of the prior distribution of x, $X = [X_1^T, X_2^T, X_3^T, \cdots]^T$ and $Y = [y_1, y_2, \cdots]^T$ are the matrices containing all the data points.

- We simulate 5000 data points and partition them randomly among $N = 100$ agents.

Experiment Design (cont.)

- Each agent has access to its own data but not to other agents' data.

Experiment Design (cont.)

- Each agent has access to its own data but not to other agents' data.
- The posterior distribution $\pi(x) \propto e^{-f(x)}$ is of the form $f(x) = \sum_{i=1}^{N} f_i(x)$ with

$$f_i(x) = -\sum_{j=1}^{n_i} \log p(y_j^i | x, X_j^i) - \frac{1}{N} \log p(x)$$
$$= \sum_{j=1}^{n_i} (y_j^i - x^T X_j^i) + \frac{1}{2\lambda N} \|x\|^2$$

where,

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

Results (DE-SGLD Method)

- Tune the step size $\eta = 0.009$ and consider the deterministic gradient (i.e., $\sigma = 0$)

Results (DE-SGLD Method)

- Tune the step size $\eta = 0.009$ and consider the deterministic gradient (i.e., $\sigma = 0$)
- It follows that $x_i^{(k)} \sim \mathcal{N}(m_i^k, \Sigma_i^{(k)})$ for some mean vector m_i^k and covariance matrix $\Sigma_i^{(k)}$

Results (DE-SGLD Method)

- Tune the step size $\eta = 0.009$ and consider the deterministic gradient (i.e., $\sigma = 0$)
- It follows that $x_i^{(k)} \sim \mathcal{N}(m_i^k, \Sigma_i^{(k)})$ for some mean vector m_i^k and covariance matrix $\Sigma_i^{(k)}$
- Based on 100 runs we estimate m_i^k and $\Sigma_i^{(k)}$ and compute $\mathcal{W}_2$ distance w.r.t $\pi(x) \sim \mathcal{N}(m, V)$

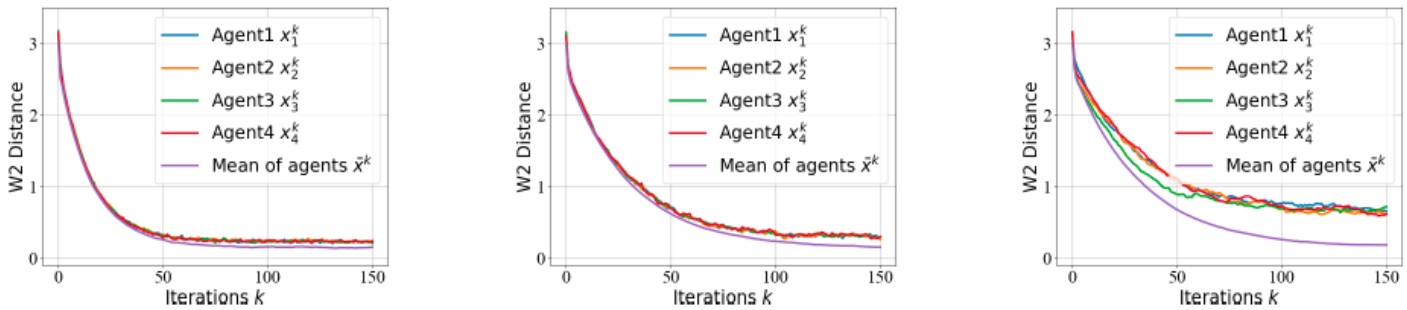
Results (DE-SGLD Method)

- Tune the step size $\eta = 0.009$ and consider the deterministic gradient (i.e., $\sigma = 0$)
- It follows that $x_i^{(k)} \sim \mathcal{N}(m_i^k, \Sigma_i^{(k)})$ for some mean vector m_i^k and covariance matrix $\Sigma_i^{(k)}$
- Based on 100 runs we estimate m_i^k and $\Sigma_i^{(k)}$ and compute $\mathcal{W}_2$ distance w.r.t $\pi(x) \sim \mathcal{N}(m, V)$
- We also compute the average $\bar{x}_i^{(k)}$ over the iterations and obtain the following

graph

Results (DE-SGLD Method)

- Tune the step size $\eta = 0.009$ and consider the deterministic gradient (i.e., $\sigma = 0$)
- It follows that $x_i^{(k)} \sim \mathcal{N}(m_i^k, \Sigma_i^{(k)})$ for some mean vector m_i^k and covariance matrix $\Sigma_i^{(k)}$
- Based on 100 runs we estimate m_i^k and $\Sigma_i^{(k)}$ and compute $\mathcal{W}_2$ distance w.r.t $\pi(x) \sim \mathcal{N}(m, V)$
- We also compute the average $\bar{x}_i^{(k)}$ over the iterations and obtain the following graph



(a) Fully-connected (b) Circular (c) Disconnected

Figure 2: Performance of DE-SGLD for Bayesian regression on different network structures with $N = 100$ agents. The results of the first 4 agents x_i^k and the node averages $\bar{x}^k = \sum_{i=1}^N x_i^{(k)}/N$ are reported.

Results (DE-SGHMC)

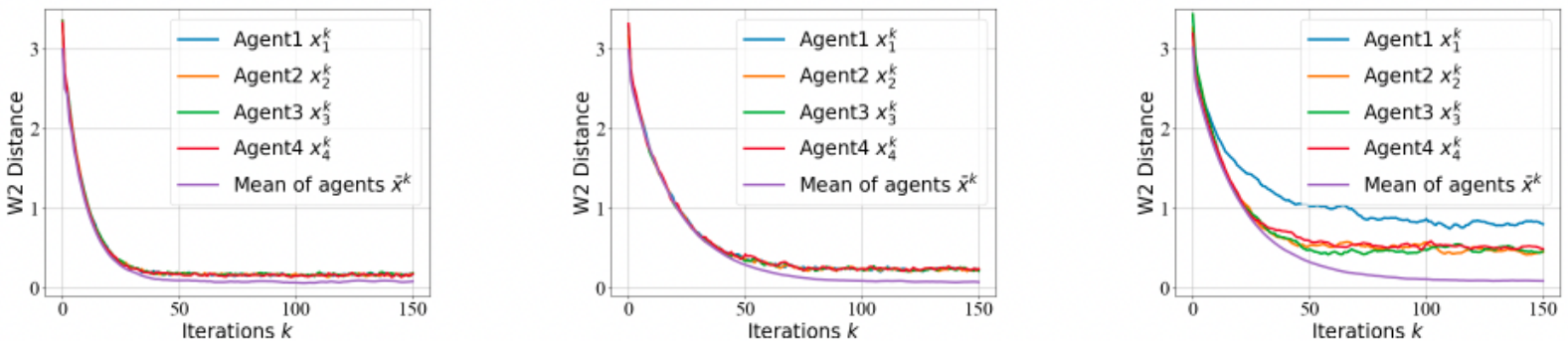
- We investigate the DE-SGHMC method on the same data set with the same three network structure

Results (DE-SGHMC)

- We investigate the DE-SGHMC method on the same data set with the same three network structure
- The stepsize and the friction coefficient are tuned to $\eta = 0.1$ and $\gamma = 7$, respectively. And we obtain the following graph

Results (DE-SGHMC)

- We investigate the DE-SGHMC method on the same data set with the same three network structure
- The stepsize and the friction coefficient are tuned to $\eta = 0.1$ and $\gamma = 7$, respectively. And we obtain the following graph



(a) Fully-connected (b) Circular (c) Disconnected

Figure 3: Performance of DE-SGHMC method for Bayesian regression on different network structures. The stepsize η and the friction coefficient γ are tuned to the dataset where we take $\eta = 0.1$ and $\gamma = 7$.

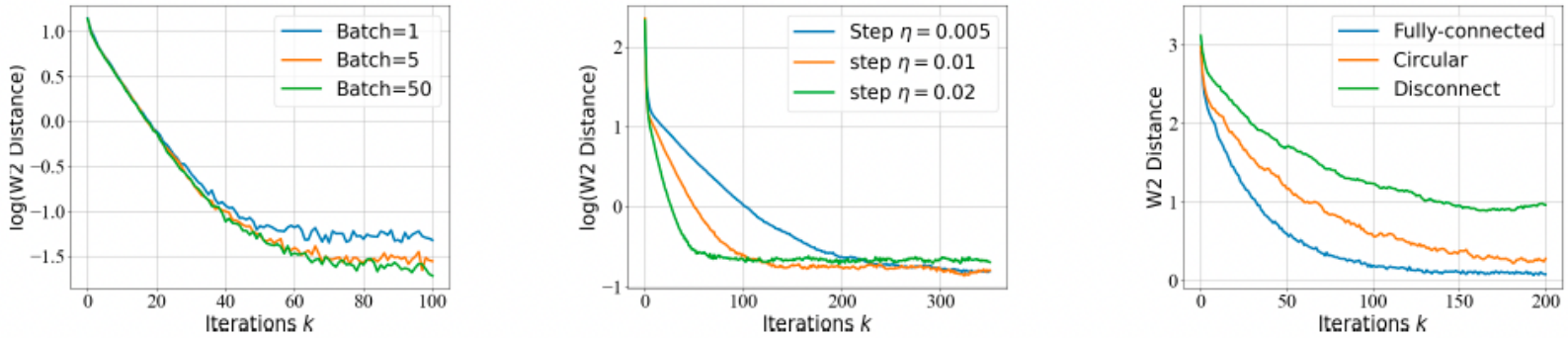
Results

- we investigate the effect of changing stepsize, batch size and the network structure on the speed of convergence where we stick to the DESGLD method for this set of experiments

Results

- we investigate the effect of changing stepsize, batch size and the network structure on the speed of convergence where we stick to the DESGLD method for this set of experiments
- We measure the 2-Wasserstein distance to the target π with a similar approach as before by fitting a Gaussian distribution $\mathcal{N}(m_i^{(k)}, \Sigma_i^{(k)})$ to the empirical distribution of $x_i^{(k)}$ over 100 independent runs.

Results

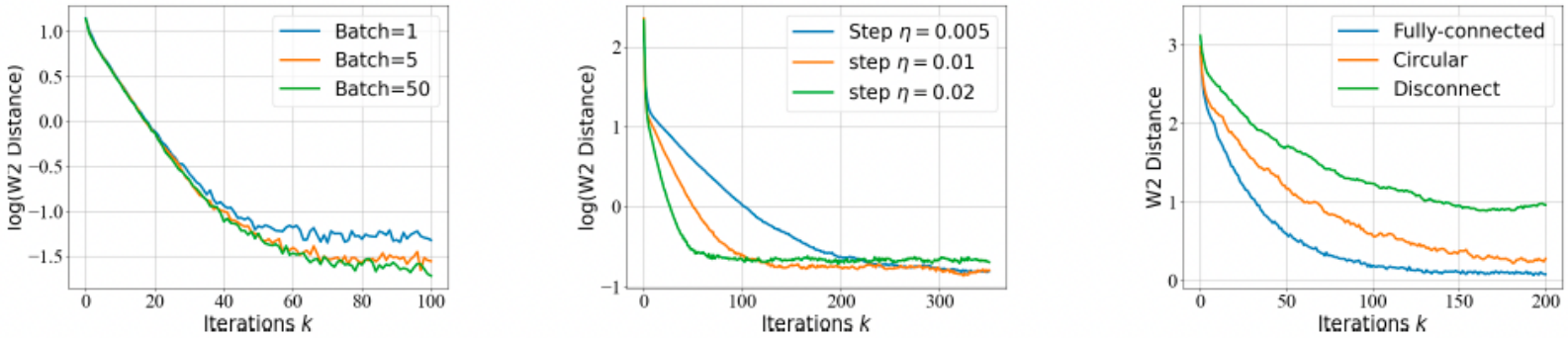


(a) Batch size (b) Stepsize (c) Network structure

Figure 4: Performance of DE-SGLD method for Bayesian regression under different settings. Figures are based on one randomly picked agent. The y-axis is presented in a logarithmic scale in (a) and (b) .

- Both Figure 4(a) and Figure 4(b) are based on the fully-connected network architecture. In Figure 4(a), we fixx the stepsize to $\eta = 0.009$ and vary the batch sizes (the number of data points sampled with replacement to estimate the gradient)

Results

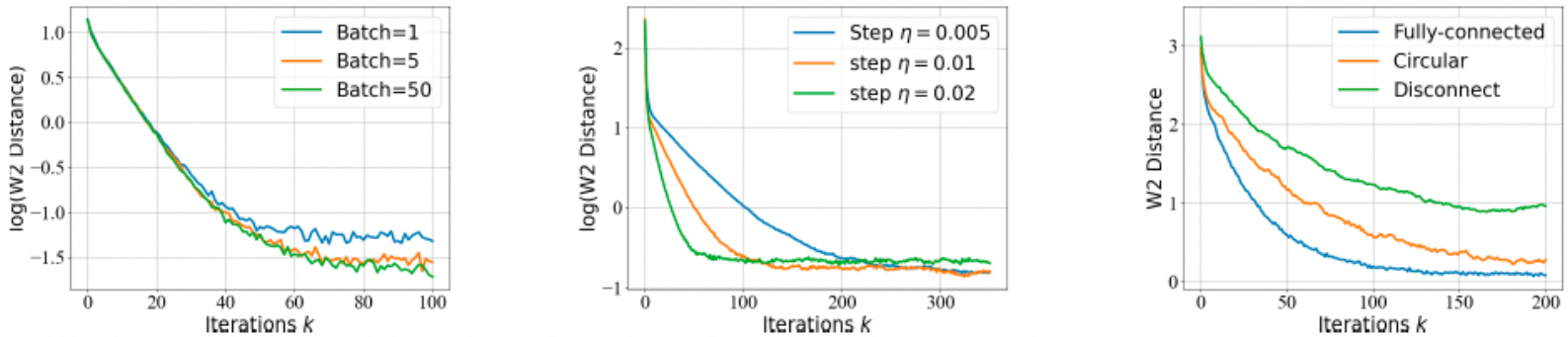


(a) Batch size (b) Stepsize (c) Network structure

Figure 4: Performance of DE-SGLD method for Bayesian regression under different settings. Figures are based on one randomly picked agent. The y-axis is presented in a logarithmic scale in (a) and (b) .

- Both Figure 4(a) and Figure 4(b) are based on the fully-connected network architecture. In Figure 4(a), we fixx the stepsize to $\eta = 0.009$ and vary the batch sizes (the number of data points sampled with replacement to estimate the gradient)
- We conclude that different batch sizes affect the asymptotic error the iterates have with respect to the 2-Wasserstein distance

Results

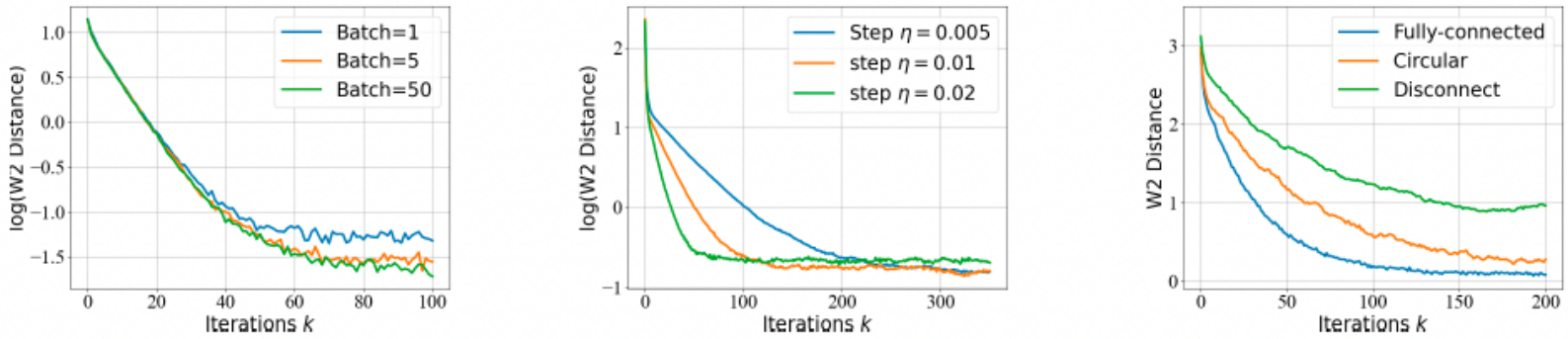


(a) Batch size (b) Stepsize (c) Network structure

Figure 4: Performance of DE-SGLD method for Bayesian regression under different settings. Figures are based on one randomly picked agent. The y-axis is presented in a logarithmic scale in (a) and (b) .

- In Figure 4(b), we used stochastic gradients with batch size $b = 25$ while we varied the stepsize
- The result clearly demonstrates the trade-off between the convergence rate and the asymptotic accuracy; for larger stepsize the algorithm converges faster to an asymptotic error region but the accuracy becomes worse as predicted by Theorem 1

Results



(a) Batch size (b) Stepsize (c) Network structure

Figure 4: Performance of DE-SGLD method for Bayesian regression under different settings. Figures are based on one randomly picked agent. The y-axis is presented in a logarithmic scale in (a) and (b) .

- In Figure 4(c) we report the effect of network structure with a constant stepsize $\eta = 0.008$ and batch size $b = 25$ where we report the performance of a randomly picked agent.
- The fastest convergence is observed for the fully-connected network.

Conclusion

- We studied DE-SGLD and DE-SGHMC methods which allow scalable Bayesian inference for decentralized learning settings

Conclusion

- We studied DE-SGLD and DE-SGHMC methods which allow scalable Bayesian inference for decentralized learning settings
- For both methods, we show that the distribution of the iterate $x^{(k)}$ of node i converges linearly (in k) to a neighborhood of the target distribution in the 2-Wasserstein metric when the target density $\pi(x) \propto e^{-f(x)}$ is strongly log-concave (i.e. f is strongly convex) and f is smooth

Conclusion

- We studied DE-SGLD and DE-SGHMC methods which allow scalable Bayesian inference for decentralized learning settings
- For both methods, we show that the distribution of the iterate $x^{(k)}$ of node i converges linearly (in k) to a neighborhood of the target distribution in the 2-Wasserstein metric when the target density $\pi(x) \propto e^{-f(x)}$ is strongly log-concave (i.e. f is strongly convex) and f is smooth
- Our results are non-asymptotic and provide performance bounds for any finite k.
- We also illustrated the efficiency of our methods on the Bayesian linear regression and Bayesian logistic regression problems