Survey on Stochastic Gradient Langevin Dynamics

Yash Travadi
Peshal Agarwal
Asim Unmesh
Javed Akhtar

Mentor: Prof. Piyush Rai

April 2018

Motivation

- Monte-Carlo approaches have been extensively used for Bayesian inference, and proven to be superior over many methods. However, its computationally expensive setup makes it impractical for large scale datasets.

Motivation

- Monte-Carlo approaches have been extensively used for Bayesian inference, and proven to be superior over many methods. However, its computationally expensive setup makes it impractical for large scale datasets.
- Multiple approaches have been tried for bigdata which allow updates to be performed in an online fashion.

Motivation

- Monte-Carlo approaches have been extensively used for Bayesian inference, and proven to be superior over many methods. However, its computationally expensive setup makes it impractical for large scale datasets.
- Multiple approaches have been tried for bigdata which allow updates to be performed in an online fashion.
- We explore an old concept of Physics called the Langevin Dynamics and Reimannian geometry which have been recently applied scale machine learning problems.

Problem Statement

- We proposed to delve deeper into some of the recent advances in the field of highly scalable MCMC, methods:-
  1. Stochastic Gradient Langevin Dynamics (SGLD)
  2. Stochastic Gradient Reimannian Langevin Dynamics (SGRLD)

Problem Statement

- We proposed to delve deeper into some of the recent advances in the field of highly scalable MCMC, methods:-
  1. Stochastic Gradient Langevin Dynamics (SGLD)
  2. Stochastic Gradient Reimannian Langevin Dynamics (SGRLD)
- Surveyed the applications of SGLD and SGRLD on Logistic Regression and LDA respectively

Problem Statement

- We proposed to delve deeper into some of the recent advances in the field of highly scalable MCMC, methods:-
  1. Stochastic Gradient Langevin Dynamics (SGLD)
  2. Stochastic Gradient Reimannian Langevin Dynamics (SGRLD)
- Surveyed the applications of SGLD and SGRLD on Logistic Regression and LDA respectively
- Theoretically developed the LDA model including the metadata via SGRLD

Problem Statement

- We proposed to delve deeper into some of the recent advances in the field of highly scalable MCMC, methods:-
  1. Stochastic Gradient Langevin Dynamics (SGLD)
  2. Stochastic Gradient Reimannian Langevin Dynamics (SGRLD)
- Surveyed the applications of SGLD and SGRLD on Logistic Regression and LDA respectively
- Theoretically developed the LDA model including the metadata via SGRLD
- Implemented the theoretically developed model on a real dataset

Background and Related Work

- A new class of MCMC algorithms called Hamiltonian MCMC algorithm was introduced recently by Neal[3] (2010).
  - In SGLD[5] an additional noise term added to the stochastic gradient update.
  - It made a huge impact on reducing learning time for non-Bayesian algorithms and SGLD extends its use for Bayesian learning as well.
  - SGLD gives a very general framework to perform Bayesian Inference, demonstrating its use on some simple models like mixture of Gaussians, logistic regression, etc.

Background and Related Work

- A new class of MCMC algorithms called Hamiltonian MCMC algorithm was introduced recently by Neal[3] (2010).
  - In SGLD[5] an additional noise term added to the stochastic gradient update.
  - It made a huge impact on reducing learning time for non-Bayesian algorithms and SGLD extends its use for Bayesian learning as well.
  - SGLD gives a very general framework to perform Bayesian Inference, demonstrating its use on some simple models like mixture of Gaussians, logistic regression, etc.
- More recently, Patterson and Teh [4](2013), proposed a new method called Stochastic gradient Riemannian Langevin dynamic (SGRLD) and implemented on LDA in an online mini-batch setting.

Background and Related Work

- A new class of MCMC algorithms called Hamiltonian MCMC algorithm was introduced recently by Neal[3] (2010).
  - In SGLD[5] an additional noise term added to the stochastic gradient update.
  - It made a huge impact on reducing learning time for non-Bayesian algorithms and SGLD extends its use for Bayesian learning as well.
  - SGLD gives a very general framework to perform Bayesian Inference, demonstrating its use on some simple models like mixture of Gaussians, logistic regression, etc.
- More recently, Patterson and Teh [4](2013), proposed a new method called Stochastic gradient Riemannian Langevin dynamic (SGRLD) and implemented on LDA in an online mini-batch setting.
- The basic idea of SGRLD is to apply online MCMC in a constrained parameter setting and obtain full posterior of each of the parameters.

Background and Related Work

- A new class of MCMC algorithms called Hamiltonian MCMC algorithm was introduced recently by Neal[3] (2010).
  - In SGLD[5] an additional noise term added to the stochastic gradient update.
  - It made a huge impact on reducing learning time for non-Bayesian algorithms and SGLD extends its use for Bayesian learning as well.
  - SGLD gives a very general framework to perform Bayesian Inference, demonstrating its use on some simple models like mixture of Gaussians, logistic regression, etc.
- More recently, Patterson and Teh [4](2013), proposed a new method called Stochastic gradient Riemannian Langevin dynamic (SGRLD) and implemented on LDA in an online mini-batch setting.
- The basic idea of SGRLD is to apply online MCMC in a constrained parameter setting and obtain full posterior of each of the parameters.
- We use this idea to solve the LDA model with metadata

Stochastic Gradient Langevin Dynamics

- **Basic Idea:** Combining stochastic optimization with Langevin Dynamics, which injects noise into the parameter updates in such a way that the trajectory of the parameters converge to the full posterior distribution rather than just converging to MAP estimate

Stochastic Gradient Langevin Dynamics

- **Basic Idea:** Combining stochastic optimization with Langevin Dynamics, which injects noise into the parameter updates in such a way that the trajectory of the parameters converge to the full posterior distribution rather than just converging to MAP estimate
- Let θ denote parameter vector, $p(\theta)$ a prior distribution and $p(x|\theta)$, likelihood. Let $X = \{x_i\}_{i=1}^N$ be data. Then posterior is given by

$$p(\theta|X) \propto p(\theta) \prod_{i=1}^{N} p(x_i|\theta)$$

Stochastic Gradient Langevin Dynamics

- **Stochastic Optimization:** At each iteration t, a subset of n data items $X_t = \{xt1, ..., x_{tn}\}$ is given, and the parameters are updated as follows:

$$\Delta\theta_t = \frac{\epsilon_t}{2}\left(\nabla \log p(\theta_t) + \frac{N}{n}\sum_{i=1}^{n} \nabla \log p(x_{ti}|\theta)\right) \tag{1}$$

, where step size satisfies,

$$\sum_{t=1}^{\infty} \epsilon_t = \infty \quad \text{and} \quad \sum_{t=1}^{\infty} \epsilon_t^2 < \infty \tag{2}$$

Stochastic Gradient Langevin Dynamics

- **Langevin Dynamics:** As before, these take gradient steps, but also injects Gaussian noise into the parameter updates so that they do not collapse to just the MAP solution:

$$\Delta\theta_t = \frac{\epsilon_t}{2}\left(\nabla \log p(\theta_t) + \sum_{i=1}^{n} \nabla \log p(x_t|\theta)\right) + \eta_t \tag{3}$$

, where $\eta_t \sim \mathcal{N}(0, \epsilon_t)$

Stochastic Gradient Langevin Dynamics

- **Langevin Dynamics:** As before, these take gradient steps, but also injects Gaussian noise into the parameter updates so that they do not collapse to just the MAP solution:

$$\Delta\theta_t = \frac{\epsilon_t}{2}\left(\nabla \log p(\theta_t) + \sum_{i=1}^{n} \nabla \log p(x_t|\theta)\right) + \eta_t \tag{3}$$

, where $\eta_t \sim \mathcal{N}(0, \epsilon_t)$

- **Stochastic Gradient Langevin Dynamics:** Combining the stochastic optimization and Langevin dynamics, the SGLD update is proposed as

$$\Delta\theta_t = \frac{\epsilon_t}{2}\left(\nabla \log p(\theta_t) + \frac{N}{n}\sum_{i=1}^{n} \nabla \log p(x_{ti}|\theta_t)\right) + \eta_t \tag{4}$$

, where $\eta_t \sim \mathcal{N}(0, \epsilon_t)$

Stochastic Gradient Riemannian Langevin Dynamics

- **Why needed ?** SGLD can't be applied when parameters are constrained.
- Patterson and Teh[4](2013), proposed solution to such a specific constrain i.e parameter that lie on a probability simplex, given by,

$$\Delta_K = \{(\pi_1, \pi_2, ...\pi_K) : \pi_k \geq 0, \sum_k \pi_k = 1\} \subseteq \mathbb{R}^K \tag{5}$$

Stochastic Gradient Riemannian Langevin Dynamics

- **Why needed ?** SGLD can't be applied when parameters are constrained.
- Patterson and Teh[4](2013), proposed solution to such a specific constrain i.e parameter that lie on a probability simplex, given by,

$$\Delta_K = \{(\pi_1, \pi_2, ...\pi_K) : \pi_k \geq 0, \sum_k \pi_k = 1\} \subseteq \mathbb{R}^K \tag{5}$$

- **Difficulties with above Constraint:**
  - Probability simplex (5) is compact.
  - Dirichlet priors over the probability simplex place most of its mass close to the boundaries and corners of the simplex.

Stochastic Gradient Riemannian Langevin Dynamics

- **Solution to constraints:**
  - Parameterization of probability simplex
  - Choice of a good parameterization is not obvious
  - The choice is guided by Riemannian geometry of the simplex [1]
  - Un-normalized parameterization, using a mirroring trick to remove boundaries, coupled with a natural gradient update is used

Stochastic Gradient Riemannian Langevin Dynamics

- **Solution to constraints:**
  - Parameterization of probability simplex
  - Choice of a good parameterization is not obvious
  - The choice is guided by Riemannian geometry of the simplex [1]
  - Un-normalized parameterization, using a mirroring trick to remove boundaries, coupled with a natural gradient update is used
- **Different kinds of Parameterization**

Parameterisation	Reduced-Mean	Reduced-Natural	Expanded-Mean	Expanded-Natural
θ	$\theta_k = \pi_k$	$\theta_k = \log \frac{\pi_k}{1-\sum_k^{K-1} \pi_k}$	$\pi_k = \frac{\lvert\theta_k\rvert}{\sum_{k=1} \lvert\theta_k\rvert}$	$\pi_k = \frac{e^{\theta_k}}{\sum_{k=1} e^{\theta_k}}$
$\nabla_\theta \log p(\theta \vert \mathrm{x})$	$\frac{n+\alpha}{\theta} - 1\frac{n_K+\alpha-1}{\pi_K}$	$n + \alpha - (n. + K\alpha)\pi$	$\frac{n+\alpha-1}{\theta} - \frac{n.}{\theta.} - 1$	$n + \alpha - n.\pi - e^\theta$
$G(\theta)$	$n. \left(\mathrm{diag}(\theta)^{-1} + \frac{1}{1-\sum_k \theta_k} 11^T\right)$	$\frac{1}{n.}\left(\mathrm{diag}(\pi) - \pi\pi^T\right)$	$\mathrm{diag}\left(\theta\right)^{-1}$	$\mathrm{diag}\left(e^\theta\right)$
$G^{-1}(\theta)$	$\frac{1}{n.}\left(\mathrm{diag}(\theta) - \theta\theta^T\right)$	$n. \left(\mathrm{diag}(\pi)^{-1} + \frac{1}{1-\sum_k \pi_k} 11^T\right)$	$\mathrm{diag}\left(\theta\right)$	$\mathrm{diag}\left(e^{-\theta}\right)$
$\sum_{k=1}^D \left(G^{-1} \frac{\partial G}{\partial \theta_k} G^{-1}\right)_{jk}$	$K\theta_j - 1$	$\frac{1}{\pi_j^2} - \frac{K-1}{(1-\sum_k \pi_k)^2}$	-1	$e^{-\theta_j}$
$\sum_{k=1}^D \left(G^{-1}(\theta)\right)_{jk} \mathrm{Tr}\left(G^{-1}(\theta)\frac{\partial G}{\partial \theta_k}\right)$	$K\theta_j - 1$	$\frac{1}{\pi_j^2} - \frac{K-1}{(1-\sum_k \pi_k)^2}$	-1	$e^{-\theta_j}$

Figure: Table Courtesy Patterson and Teh[4]

DMR topic model with metadata via SGRLD

- Text data sometimes contain meta information such as authors, publication venues, and dates

DMR topic model with metadata via SGRLD

- Text data sometimes contain meta information such as authors, publication venues, and dates
- Extension of basic mixture models to account for metadata allows:
  1. Improve learning of topics
  2. Discover patterns such as author's profile based on topics, popularity of topic with time

DMR topic model with metadata via SGRLD

- Text data sometimes contain meta information such as authors, publication venues, and dates
- Extension of basic mixture models to account for metadata allows:
  1. Improve learning of topics
  2. Discover patterns such as author's profile based on topics, popularity of topic with time
- Models which generate both words as well as meta using topic variables have been studied

DMR topic model with metadata via SGRLD

- Text data sometimes contain meta information such as authors, publication venues, and dates
- Extension of basic mixture models to account for metadata allows:
  1. Improve learning of topics
  2. Discover patterns such as author's profile based on topics, popularity of topic with time
- Models which generate both words as well as meta using topic variables have been studied
- Mimno and McCallum[2] recently propose a new method for modeling the additional information called Dirichlet-multinomial Regression (DMR) topic models.

DMR topic model with metadata via SGRLD

- Text data sometimes contain meta information such as authors, publication venues, and dates
- Extension of basic mixture models to account for metadata allows:
  1. Improve learning of topics
  2. Discover patterns such as author's profile based on topics, popularity of topic with time
- Models which generate both words as well as meta using topic variables have been studied
- Mimno and McCallum[2] recently propose a new method for modeling the additional information called Dirichlet-multinomial Regression (DMR) topic models.
- They implemented stochastic EM sampling for inference, however, we formulate the DMR model via SGRLD

Generative Story of DMR

Notations

- x_d is the feature vector of the metadata
- N_d denotes the number of words in each document
- η_d is the topic proportion vector
- λ_k is the parameter corresponding to metadata

Data Generation

1. For each topic $k = 1, \ldots, K$
   - Draw $\lambda_k \sim \mathcal{N}(0, \sigma^2 I)$
   - Draw $\theta_{kw} \sim Gamma(\beta, 1)$
   - $\pi_{kw} = \frac{\theta_{kw}}{\sum_w \theta_{kw}}$

Generative Story of DMR

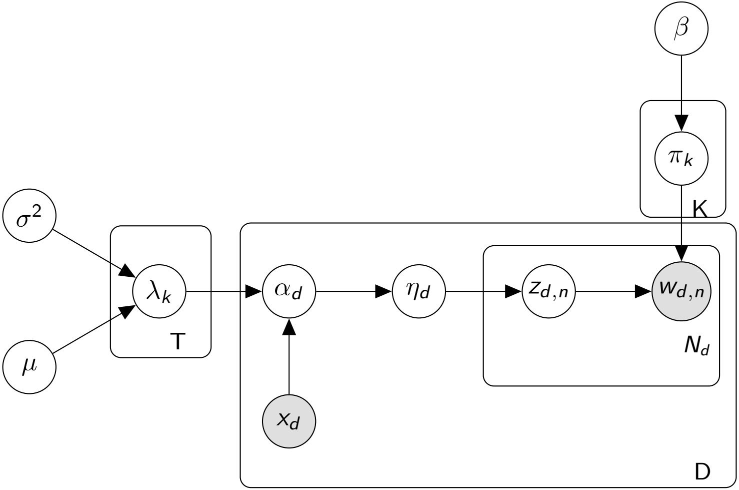
Notations

- x_d is the feature vector of the metadata
- N_d denotes the number of words in each document
- η_d is the topic proportion vector
- λ_k is the parameter corresponding to metadata

Data Generation

1. For each topic $k = 1, \ldots, K$
    - Draw $\lambda_k \sim \mathcal{N}(0, \sigma^2 I)$
    - Draw $\theta_{kw} \sim Gamma(\beta, 1)$
    - $\pi_{kw} = \frac{\theta_{kw}}{\sum_w \theta_{kw}}$
2. For each document d
    - For each topic k, set $\alpha_{dk} = \exp(x_d^T \lambda_k)$
    - Draw $\eta_d \sim Dirichlet(\alpha_d)$
    - For each word $n = 1, \ldots, N_d$ in document d
        - Choose the topic for the word $z_{dn} \sim multinoulli(\eta_d)$
        - Generate word from the chosen topic $w_{dn} \sim multinoulli(\pi_{z_{dn}})$

Graphical topic model with metadata via SGRLD



SGRLD for DMR

- We assume model hyper-parameters μ, σ^2 and β to be fixed
- We integrate-out η from the posterior distribution
- α is deterministically derived from x and λ
- Since, π_k is constrained inside a probability simplex, we re-parameterize it with θ, such that $\theta_{kw} > 0$.

$$\pi_{kw} = \frac{\theta_{kw}}{\sum_w \theta_{kw}}$$

- Thus, we need to infer θ, Z and λ

SGRLD updates I

- Since $\theta_{kw} > 0$, we use mirror reflection

$$\theta^*_{kw} = \left|\theta_{kw} + \frac{\epsilon}{2}\left(\beta - \theta_{kw} + \frac{|D|}{|D_t|}\sum_{d \in D_t} \mathbb{E}_{z_d|w_d,\theta,\alpha}[n_{dkw} - \pi_{kw} n_{dk.}]\right) + (\theta_{kw})^{\frac{1}{2}}\zeta_{kw}\right| \quad (6)$$

, where $\zeta_{kw} \sim \mathcal{N}(0, \epsilon)$,
$n_{dkw} = \sum_{i=1}^{N_d} \delta(w_{di} = w, z_{di} = k)$.

SGRLD updates II

- We use Gibbs sampling to calculate the expectation required in the update of θ

$$p(z_{di} = k \mid w_d, \alpha_d, \theta) = \frac{(\alpha_{dk} + n_{dk.}^{\setminus i})\theta_{kw_{di}}}{\sum_k (\alpha_{dk} + n_{dk.}^{\setminus i})\theta_{kw_{di}}}$$

where $\setminus i$ represents a count excluding the topic assignment variable we are updating

SGRLD updates III

- We perform SGLD update on λ_k since it is unconstrained

$$\frac{\partial \mathcal{L}}{\partial \lambda_{kt}} = \sum_d x_{dt} \exp(x_d^T \lambda_k) \times$$
$$\Big(\Psi\big(\sum_k \exp(x_d^T \lambda_k)\big) - \Psi\big(\sum_k \exp(x_d^T \lambda_k) + n_d\big) +$$
$$\Psi\big(\exp(x_d^T \lambda_k) + n_{k|d}\big) - \Psi\big(\exp(x_d^T \lambda_k)\big)\Big) - \frac{\lambda_{kt}}{\sigma^2}$$

where $\mathcal{L}$ is the complete log-likelihood

Experimental Results I

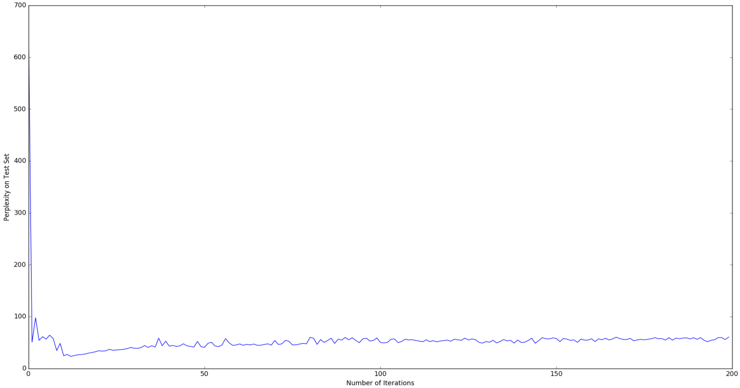
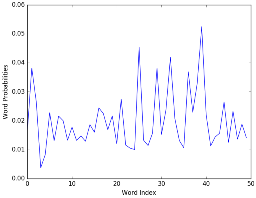
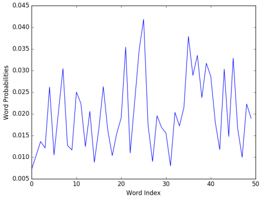


Figure: Perplexity vs number of iteration over dummy dataset

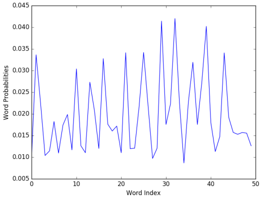
Experimental Results II



(a) Topic 0 at State 0



(b) Topic 0 at State 100



(c) Topic 0 at State 200

Experimental Results III

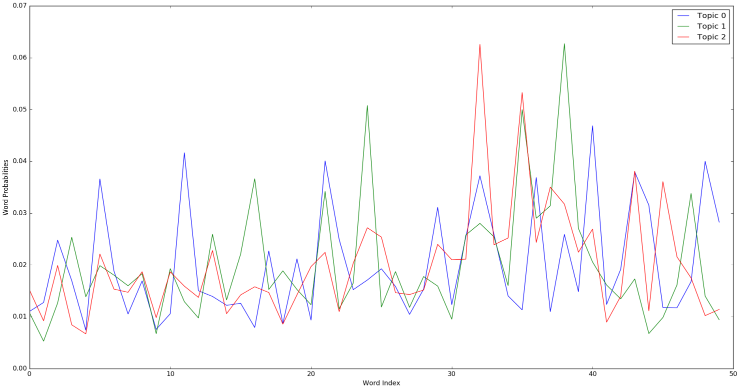


Figure: Word Proportion

Difficulties and work in progress

- While implementing on large dataset we are facing scalability issues
  - In the update of θ_{kw} we sample z using Gibbs sampling for each topic and word
  - The size of π_k becomes too large (size of vocabulary)
- Difficult to tune hyper-parameters since model is a bit sensitive to hyper-parameters
- We are trying to implement DMR topic model on a corpus of research papers drawn from the Rexa database

Summary

- Understood the concept of Langevin dynamics and its application in machine learning problems via SGLD
- Realized the shortcomings of SGLD and methods to overcome the same using SGRLD
- Studied a new topic model that incorporates meta information for each document, known as DMR (Dirichlet-Multinomial Regression)
- Derived update equations for inference over DMR model via SGRLD
- Implemented our model on a dummy dataset

References

[1] S.-I. Amari. Information geometry of the em and em algorithms for neural networks. *Neural networks*, 8(9):1379–1408, 1995.

[2] D. Mimno and A. McCallum. Topic models conditioned on arbitrary features with dirichlet-multinomial regression. *arXiv preprint arXiv:1206.3278*, 2012.

[3] R. M. Neal et al. Mcmc using hamiltonian dynamics. *Handbook of Markov Chain Monte Carlo*, 2(11), 2011.

[4] S. Patterson and Y. W. Teh. Stochastic gradient riemannian langevin dynamics on the probability simplex. In *Advances in Neural Information Processing Systems*, pages 3102–3110, 2013.

[5] M. Welling and Y. W. Teh. Bayesian learning via stochastic gradient langevin dynamics. In *Proceedings of the 28th International Conference on Machine Learning (ICML-11)*, pages 681–688, 2011.