Journal of Machine Learning **ISSN**: 2790-2048(e), 2790-203X(p)

Batch Normalization Preconditioning for Stochastic Gradient Langevin Dynamics

Susanna Lange [*] [1], Wei Deng [†] [2], Qiang Ye [‡] [3], and Guang Lin [§] [4]

[1]Data Science Institute, University of Chicago, Chicago, IL 60615, USA.
[2]Machine Learning Research, Morgan Stanley, New York City, NY 10036, USA.
[3]Department of Mathematics, University of Kentucky, Lexington, KY 40506, USA.
[4]Departments of Mathematics, & School of Mechanical Engineering, Purdue University, West Lafayette, IN 47907, USA.

Abstract. Stochastic gradient Langevin dynamics (SGLD) is a standard sampling technique for uncertainty estimation in Bayesian neural networks. Past methods have shown improved convergence by including a preconditioning of SGLD based on RMSprop. This preconditioning serves to adapt to the local geometry of the parameter space and improve the performance of deep neural networks. In this paper, we develop another preconditioning technique to accelerate training and improve convergence by incorporating a recently developed batch normalization preconditioning (BNP), into our methods. BNP uses mini-batch statistics to improve the conditioning of the Hessian of the loss function in traditional neural networks and thus improve convergence. We will show that applying BNP to SGLD will improve the conditioning of the Fisher information matrix, which improves the convergence. We present the results of this method on three experiments including a simulation example, a contextual bandit example, and a residual network which show the improved initial convergence provided by BNP, in addition to an improved condition number from this method.

Keywords:
Bayesian neural networks,
Preconditioning,
Batch normalization,
Stochastic gradient Langevin dynamics.

Article Info.:
Volume: 2
Number: 1
Pages: 65- 82
Date: March/2023
doi.org/10.4208/jml.220726a

Article History:
Received: 26/07/2022
Accepted: 10/03/2023

Handler:
Jiequn Han

1 Introduction

Markov chain Monte Carlo (MCMC) provides a principled framework for simulating the distribution of interest. During the simulation, the entire dataset is often used to compute the energy or the gradient, which, however, is not scalable enough in big data problems. To tackle this issue, stochastic gradient Langevin dynamics [23] proposes to inject additional Gaussian noise to stochastic gradient descent and smoothly transitions into an MCMC sampler as the step size goes to zero. The explorative feature of the sampler not only captures uncertainty for reliable decision-making but also facilitates non-convex optimization to alleviate over-fitting [19,25]. Since then, many interesting stochastic gradient Markov chain Monte Carlo (SG-MCMC) methods are proposed to accelerate the convergence [2,4,6,14]. However, these sampling algorithms still suffer from a slow convergence given morbid curvature information. To handle this issue, Girolami *et al.* [7] and Patterson

[*] `susannalange@uchicago.edu`
[†] `weideng056@gmail.com`
[‡] `qiang.ye@uky.edu`.
[§] Corresponding author. `guanglin@purdue.edu`.

and Teh [18] propose to adjust the Langevin algorithm on the Riemann manifold. Despite the correctness of the simulations, it is challenging to conduct the transformation in high-dimensional problems. Motivated by the adaptive preconditioner as in root mean squared propagation (RMSprop), the preconditioned SGLD algorithm (pSGLD) proposes to accelerate SGLD through a diagonal approximation of the Fisher information to resolve the scalability issue [13]. This uses gradient information to construct a preconditioner that can be interpreted to have an adaptive step size, with a smaller step size for curved directions and a larger step size for flat directions. This combats the slow training related to saddle points in neural networks. Other preconditioning methods have been investigated, including dense approximations of the inverse Hessian, as in [1,21]. There have also been approaches to use non-linear averaging methods to accelerate network convergence. He *et al.* uses a truncated generalized conjugate residual method that uses symmetry of the Hessian to improve convergence [9], and combines gradient descent ascent with Anderson mixing in generative adversarial networks, which was shown to improve adversarial training [10].

Another approach to accelerate convergence is to incorporate batch normalization (BN) layers into the network architecture [11]. BN uses mini-batch statistics to normalize hidden variables of a network and has been shown to decrease training times and improve network regularization. BN and its connection to Bayesian neural networks have been studied in [22], in particular, a network with BN can be interpreted as an approximate Bayesian model. Batch normalization has also been successfully applied to Bayesian models as studied in [16] which shows that including BN layers does not affect the probabilistic inference of variational methods. Batch normalization preconditioning is a technique that also uses mini-batch statistics but does so by transforming a network's trainable parameters using a preconditioner [12]. This is done by applying a preconditioning transformation on the parameter gradients during training. This transformation has been shown to improve the conditioning of the Hessian of the loss function which corresponds to a major advantage of the BNP transformation, that is, improvement in the convergence of the method. More importantly, BNP is a general framework that is applicable to different neural network architectures and in different settings, such as Bayesian models.

In this paper, we develop BNP for Bayesian neural networks to be used as a sampling method and examine its effects. We show that we can develop a similar preconditioning technique for SGLD that further improves initial convergence by improving the condition number of the Fisher information matrix. Additionally, we provide experimental results on three different methods, each showing improvement in convergence over our comparative baselines. We also compute the condition number of the approximate empirical Fisher information, which demonstrates the improvement in the condition number in our method.

The paper is organized as follows. In Section 2 we provide background information on stochastic gradient Langevin dynamics as well as a preconditioned version of SGLD. Section 2.2 introduces the basics of a batch normalization architecture. In Section 3 we expand upon a preconditioning method, batch normalization preconditioning, to a Bayesian setting that improves the conditioning of the Fisher information matrix and Section 4 showcases the benefits of BNP applied to SLGD in three different experiments.