Stochastic gradient method with accelerated stochastic dynamics

2016 J. Phys.: Conf. Ser. 699 012019

(http://iopscience.iop.org/1742-6596/699/1/012019)

Journal of Physics: Conference Series **699** (2016) 012019 doi:10.1088/1742-6596/699/1/012019

Stochastic gradient method with accelerated stochastic dynamics

Masayuki Ohzeki

Department of Systems Science, Graduate School of Informatics, Kyoto University, 36-1 Yoshida Hon-machi, Sakyo-ku, Kyoto, 606-8501, Japan

E-mail: mohzeki@i.kyoto-u.ac.jp

Abstract. We implement the simple method to accelerate the convergence speed to the steady state and enhance the mixing rate to the stochastic gradient Langevin method. The ordinary stochastic gradient method is based on mini-batch learning for reducing the computational cost when the amount of data is extraordinary large. The stochasticity of the gradient can be mitigated by the injection of Gaussian noise, which yields the stochastic Langevin gradient method; this method can be used for Bayesian posterior sampling. However, the performance of the stochastic Langevin gradient method depends on the mixing rate of the stochastic dynamics. In this study, we propose violating the detailed balance condition to enhance the mixing rate. Recent studies have revealed that violating the detailed balance condition accelerates the convergence to a stationary state and reduces the correlation time between the samplings. We implement this violation of the detailed balance condition in the stochastic gradient Langevin method and test our method for a simple model to demonstrate its performance.

1. Introduction

Since massive amounts of data can be acquired from various sources, the importance of the so-called big-data analysis is rapidly increasing. In particular, extracting relevant parameters that facilitate describing the acquired data is critical for analyzing high-dimensional data. For high-dimensional data analysis, Boltzmann machine learning can be used to extract parameters that describe the structure of given data [1]. Boltzmann machine learning has proven to be effective and has stimulated increasing interest in deep learning [2, 3, 4, 5]. We can here implement the L_1-norm regularization to determine the sparse parameters for characterizing given data [5] and and a type of the relaxation method to the L_0 norm [6, 7]. Although Boltzmann machine learning has been successfully implemented to extract features from high-dimensional data, the computational cost is usually expensive for high-precision estimations. The crucial reason for the high computational cost is the expectation value in the Gibbs-Boltzmann distribution. In order to mitigate the computational cost, some approximations such as the mean-field analysis, belief propagation, and their improvements can be adopted [8, 9, 10, 11, 12, 13, 14]. In addition, increase of the amount of the data also increases the computational cost for evaluating the update rules in the learning algorithms.

In this study, we develop a technique to mitigate the computational cost of learning from large amount of high-dimensional data. In the middle of the 20th century, Robbins and Monro proposed a stochastic gradient method, which utilized a mini-batch dataset for gradient computation to update the tentative estimations at each iteration in the learning process [15].

The stochasticity of the gradient will be averaged out because the entire dataset is used by the end of the learning process. Then, the learning rate is gradually decreased to ensure convergence to a local maximum of the likelihood function. In order to mitigate the stochasticity of the gradient, Welling and Teh proposed an improved version of the stochastic gradient method by introducing Gaussian noise at each iteration such as by using the Langevin dynamics [16]. At the first stage of the method, the stochastic gradient noise dominates the injected noise in the learning process. On the other hand, at the last stage, the injected noise equilibrates the system to converge to a stationary state.

For enhancing the precision of the learning, it is beneficial to accelerate the convergence to a stationary state in the stochastic gradient Langevin method. Considering the recent advancement in stochastic dynamics in the field of nonequilibrium statistical physics, a novel technique to accelerate the convergence to the desired stationary distribution has been proposed, which performs the violation of the detailed balance condition (vDBC). The detailed balance condition (DBC) is typically satisfied in stochastic dynamics simulated by the Langevin equation, the master equation, and the Fokker-Planck equation. One such representative method is the Markov-chain Monte-Carlo method. The condition assures convergence to a stationary distribution and facilitates the construction of the stochastic rule as in the transition matrix. However, the DBC is just a sufficient condition for the convergence and thus is not necessarily to be satisfied. Hence, we may violate the DBC to attain faster convergence to the stationary distribution. One of the techniques that do not hold the DBC is the Suwa-Todo method [17]. In this method, the elements of the transfer matrix are optimized in the master equation simulated in the Markov-Chain Monte-Carlo method, while ignoring the DBC. Another method is the skewed DBC, which utilizes double stochastic rules for acceleration while violating the DBC [18]. The vDBC is characterized by the inherent asymmetry of the transfer matrix at the master equation or the Fokker-Planck equation level. This asymmetry leads to an eigenvalue shift as compared to the case when the DBC is satisfied and results in the acceleration of the convergence to a stationary state [19]. From a different point of view, the vDBC can be captured as rare-event sampling, which produces an alternative pathway to the stationary state [20]. A simple implementation of the vDBC has been performed in the Langevin dynamics especially for a system with continuous-valued degrees of freedom [21, 22]. The vDBC can be recast as the resultant feature of the optimization of stochastic dynamics from a point of view of the variational principle [23]. Although several innovative studies to enhance the sampling efficacy and accelerate the convergence to a predetermined distribution have been extensively reported [24, 25, 26, 27, 28, 29, 30, 31], the concept of vDBC is much easier to implement. In this study, we aim to implement the concept of the vDBC in the stochastic gradient Langevin method and verify its performance.

The remainder of the paper is organized as follows: First, we briefly review the stochastic gradient Langevin method. In the next section, we introduce the scheme of the accelerated stochastic dynamics through the vDBC. In the fourth section, we formulate the application of the accelerated stochastic dynamics to the stochastic gradient Langevin method. We test our scheme, in the following section, for the simplest learning model, which is a mixture of the Gaussian distribution, to confirm the efficacy of our method. The last section is devoted for discussion on the future direction of our study.

2. Stochastic gradient Langevin method

In this section, we briefly review the stochastic gradient Langevin method. The stochastic gradient Langevin method has been invented for Bayesian learning, which captures the uncertainty in the learned parameters and avoids overfitting. This method combines the Robbins-Monro method, which optimizes a likelihood function in a stochastic manner, with overdamped Langevin dynamics, which injects noise into the parameter updates. Each

stochasticity has a different property. The former stochasticity is a resultant property to reduce the computational cost for large-scale data. In the latter, the noise makes the trajectory of the parameters converge to full posterior distribution rather than just the maximum a posteriori mode.

We consider the learning of the parameters $\boldsymbol{\theta}$ with a prior distribution $p(\boldsymbol{\theta})$ and the probability of the datum $\mathbf{x}$ given the parameter $\boldsymbol{\theta}$, $P(\mathbf{x}|\boldsymbol{\theta})$; namely the posterior distribution $P(\mathbf{x}|\boldsymbol{\theta})p(\boldsymbol{\theta})$. The basic task is to find the maximum a posteriori (MAP) parameters $\boldsymbol{\theta}$ by maximizing the following logarithm of the posterior distribution:

$$\boldsymbol{\theta}^* = \arg\max_{\boldsymbol{\theta}} \left\{ \sum_{k=1}^{D} \log P(\mathbf{x}^{(k)}|\boldsymbol{\theta}) + \log p(\boldsymbol{\theta}) \right\}. \tag{1}$$

In the context of the optimization, the prior distribution regularizes the parameters while the likelihood terms constitute the cost function to be optimized. In the basic framework of the maximum a posteriori, by taking the derivative of the parameters $\boldsymbol{\theta}$, we obtain the following iteration for seeking the maximum of the posterior distribution.

$$\boldsymbol{\theta}' = \boldsymbol{\theta} + dt \frac{\partial \mathcal{L}(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}}, \tag{2}$$

where

$$\frac{\partial \mathcal{L}(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} = \frac{\partial}{\partial \boldsymbol{\theta}} \log p(\boldsymbol{\theta}) + \sum_{k=1}^{D} \frac{\partial}{\partial \boldsymbol{\theta}} \log P(\mathbf{x}^{(k)}|\boldsymbol{\theta}). \tag{3}$$

The logarithm of the posterior distribution is defined as $\mathcal{L}(\boldsymbol{\theta})$. Increasing the amount of given data results in the computational cost to sum over the gradient of the log-likelihood in the second term. To mitigate the problem, a stochastic gradient method has been proposed. In this method, the gradient of the log-likelihood function for a randomly chosen subset of the given data is used:

$$\frac{\partial \mathcal{L}(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} = \frac{\partial}{\partial \boldsymbol{\theta}} \log p(\boldsymbol{\theta}) + \frac{D}{d} \sum_{k=1}^{d} \frac{\partial}{\partial \boldsymbol{\theta}} \log P(\mathbf{x}^{(k)}|\boldsymbol{\theta}). \tag{4}$$

where $d < D$. The subset of the given data is randomly chosen. Over a number of iterations, all the data is used and the noise in the gradient caused by using subsets is averaged out. To ensure convergence to a local maximum, the step size is gradually decreased. The step size dt depends on the step as dt_i and satisfies $\sum_i dt_i = \infty$ and $\sum_t dt_i^2 < \infty$ [15]. Typically, step size is set as $dt_i = a(b+i)^{-\gamma}$ where $\gamma \in (0.5, 1]$. This is the stochastic gradient method and just an approximation of the MAP estimation (or also applicable to the maximum likelihood estimation).

To avoid the problem that it does not capture parameter uncertainty and can potentially overfit data in the MAP estimation, stochastic noise as shown below can be introduced.

$$\boldsymbol{\theta}' = \boldsymbol{\theta} + dt \frac{\partial \mathcal{L}(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} + \sqrt{2T} dW, \tag{5}$$

where T is the temperature representing the strength of the injected noise (usually set as unity) and $d\mathbf{W}$ denotes the Wiener process, whose order is $O(\sqrt{dt})$, i. e., the Gaussian noise with a vanishing mean and a variance of dt. This simply represents the overdamped Langevin equation in the context of statistical physics. The overdamped Langevin equation has a corresponding

Fokker-Planck equation, which describes the time-evolution of the distribution function. It can be confirmed that the equilibrium distribution can be the posterior distribution. Welling and Teh have proposed the combination of Langevin dynamics with the stochastic gradient method, i.e., the stochastic gradient Langevin method, to generate the posterior distribution for learning from large-scale data [16]. The step size is then decreased similar to the stochastic gradient method introduced above. By decreasing the step size gradually, the injected noise will become dominant and the effective dynamics will converge to the Langevin equation with the exact gradient. A number of the iterations will thus generate the posterior distribution. In other words, the performance of the stochastic gradient Langevin method strongly depends on the convergence rate of the Langevin equation to the equilibrium distribution, i.e., the posterior distribution. The increase in the convergence speed improves the performance and reduces the computational cost of the method. This fact motivates the study of the stochastic gradient Langevin method from a point of view of nonequilibrium statistical physics.

3. Accelerated stochastic dynamics

The stochastic gradient Langevin method aims at the sampling following the posterior distribution while mitigating the computational cost for evaluation of the gradient. Hereafter let us consider how to perform sampling from the desired distribution function. The distribution function can be characterized by the Gibbs-Boltzmann distribution associated with many degrees of freedom. Therefore it is difficult to straightforwardly perform sampling following the distribution function. We then assume that a fictitious system consisting of several degrees of freedom is governed by the Langevin equation. After sufficient long equilibration, the degrees of freedom follows the Gibbs-Boltzmann distribution. However the relaxation to the Gibbs-Boltzmann distribution takes extremely long time. Here, we employ the recent advancement in stochastic dynamics, i.e., the vDBC. For the case of the standard Langevin equation, the DBC holds to ensure the convergence to a stationary state. However, this is just a sufficient condition and can be violated. We modify the standard form of the Langevin equation to accelerate the convergence speed towards a stationary state. This is referred as the accelerated stochastic dynamics.

3.1. Langevin equation and its corresponding Fokker-Planck equation

We begin with the N-dimensional overdamped Langevin dynamics defined as

$$d\boldsymbol{\theta} = \mathbf{A}(\boldsymbol{\theta})dt + \sqrt{2T}d\mathbf{W}, \tag{6}$$

where $\boldsymbol{\theta}$ is the vector representing the N-dimensional degrees of freedom (parameters to be estimated for machine learning) and $d\boldsymbol{\theta}$ denotes the infinitesimal change in $\boldsymbol{\theta}$ during dt. Here we assume that the step size dt is homogeneous for the sake of simplicity. In addition, $\mathbf{A}(\boldsymbol{\theta})$ is a time-independent force vector, T is the temperature, and $d\mathbf{W}$ denotes the Wiener process for the N-dimensional degrees of freedom as stated earlier. The force usually takes the gradient of the energy as $-\partial E(\boldsymbol{\theta})/\partial\boldsymbol{\theta}$ in the context of the statistical physics. or our problem, the force is denoted as $\partial \log p(\boldsymbol{\theta})/\partial\boldsymbol{\theta} + (1/D)\sum_{k=1}^{D} \partial \log P(\mathbf{x}^{(k)}|\boldsymbol{\theta})/\partial\boldsymbol{\theta}$. In other words, the energy function is defined as

$$E(\boldsymbol{\theta}) = -\log p(\boldsymbol{\theta}) - \sum_{k=1}^{D} \log P(\mathbf{x}^{(k)}|\boldsymbol{\theta}) \tag{7}$$

For simplicity, we use the notion of the energy function hereafter. In accelerated stochastic dynamics, we do not stick with the standard case. In order to accelerate the convergence to a stationary state, we modify the force from its usual form. The probability distribution for the

degrees of freedom evolved in the Langevin dynamics can be formulated as the Fokker-Planck equation as follows:

$$\frac{\partial P(\boldsymbol{\theta}, t)}{\partial t} = -\sum_{k=1}^{N} \frac{\partial}{\partial \theta_k} J_k(\boldsymbol{\theta}, t), \tag{8}$$

where $\mathbf{J}(\boldsymbol{\theta}, t) = (J_1(\boldsymbol{\theta}, t), J_2(\boldsymbol{\theta}, t), \cdots, J_N(\boldsymbol{\theta}, t))$ denotes the probabilistic flow defined as

$$\mathbf{J}(\boldsymbol{\theta}, t) = \left(\mathbf{A}(\boldsymbol{\theta}) - T \frac{\partial}{\partial \boldsymbol{\theta}} \right) P(\boldsymbol{\theta}, t). \tag{9}$$

When the system is in any stationary state, the divergence of the probabilistic flow vanishes. Therefore, the following equality should be satisfied in the steady state:

$$0 = -\sum_{k=1}^{N} \frac{\partial}{\partial \theta_k} J_k^{(\mathrm{ss})}(\boldsymbol{\theta}), \tag{10}$$

where

$$\mathbf{J}^{(\mathrm{ss})}(\boldsymbol{\theta}) = \left(\mathbf{A}(\boldsymbol{\theta}) - T \frac{\partial}{\partial \boldsymbol{\theta}} \right) P^{(\mathrm{ss})}(\boldsymbol{\theta}). \tag{11}$$

We refer to the above equality as the divergence-free condition. This is a balanced condition in the context of the master equation. The stationary state should be set for the accomplishment of our aim as follows:

$$P^{(\mathrm{ss})}(\boldsymbol{\theta}) = \prod_{k=1}^{D} P(\mathbf{x}^{(k)}|\boldsymbol{\theta}) p(\boldsymbol{\theta}), \tag{12}$$

where we set $T = 1$ (in machine learning, the inverse of temperature is absorbed into the energy definition or the likelihood function and prior distribution).

We can immediately obtain a trivial solution of the divergence-free condition. When the probabilistic flow in the steady state does not exist, the divergence-free condition is satisfied.

$$\mathbf{J}^{(\mathrm{eq})}(\boldsymbol{\theta}) = \mathbf{0}, \tag{13}$$

where the superscript has been changed from "ss" to "eq". This stationary state is in particular called the equilibrium state because current does not exist here. Then, the force can be defined as

$$\mathbf{A}(\boldsymbol{\theta}) = -\frac{\partial}{\partial \boldsymbol{\theta}} E(\boldsymbol{\theta}). \tag{14}$$

If we simulate the Langevin dynamics with the trivial force as shown in Equation (14), the long-time relaxation yields the equilibrium distribution.

3.2. Nontrivial solution

We impose the probabilistic flow in a steady state to obtain the following form:

$$\mathbf{J}^{(\mathrm{ss})}(\boldsymbol{\theta}) = \gamma \mathbf{B}(\boldsymbol{\theta}) P^{(\mathrm{ss})}(\boldsymbol{\theta}), \tag{15}$$

where γ is a degree of vDBC. Following the divergence-free condition, the vector field $\mathbf{B}(\boldsymbol{\theta}) = (B_1(\boldsymbol{\theta}), B_2(\boldsymbol{\theta}), \cdots, B_N(\boldsymbol{\theta}))$ must satisfy

$$0 = \gamma \sum_{k=1}^{N} \left(\frac{\partial B_k(\boldsymbol{\theta})}{\partial \theta_k} - B_k \frac{\partial E(\boldsymbol{\theta})}{\partial \theta_k} \right) P^{(\mathrm{ss})}(\boldsymbol{\theta}), \tag{16}$$

where we have used the fact that the distribution function is set as $P^{(\mathrm{ss})}(\boldsymbol{\theta}) \propto \exp(-E(\boldsymbol{\theta}))$ and the subscript for the bracket denotes the element of the vector. A trivial solution can be given by

$$B_k(\boldsymbol{\theta}) \propto \exp(E(\boldsymbol{\theta})). \tag{17}$$

However, we have to be careful about the instability when using the exponential term, which often leads to instability in the integration of the Langevin equality

Let us find an alternative to the force obtained above to remove the instability of the exponential term. We set the vector field as

$$B_k(\boldsymbol{\theta}) = \left(\frac{\partial E(\boldsymbol{\theta})}{\partial \theta_{k-1}} - \frac{\partial E(\boldsymbol{\theta})}{\partial \theta_{k+1}} \right), \tag{18}$$

where $\theta_{N+1} = \theta_1$ and $\theta_0 = \theta_N$. Then the divergence-free condition holds. The resulting force is given as

$$A_k(\boldsymbol{\theta}) = -\frac{\partial E(\boldsymbol{\theta})}{\partial \theta_k} + \gamma \left(\frac{\partial E(\boldsymbol{\theta})}{\partial \theta_{k-1}} - \frac{\partial E(\boldsymbol{\theta})}{\partial \theta_{k+1}} \right). \tag{19}$$

This force drives the system while satisfying the divergence-free condition but not the DBC. The violation of the DBC can be confirmed through the formulation of the path probability during infinitesimal time. For a two-dimensional case, the above nontrivial force can be implemented as

$$A_1(\boldsymbol{\theta}) = -\frac{\partial E(\boldsymbol{\theta})}{\partial \theta_1} - \gamma \frac{\partial E(\boldsymbol{\theta})}{\partial \theta_2} \tag{20}$$

$$A_2(\boldsymbol{\theta}) = -\frac{\partial E(\boldsymbol{\theta})}{\partial \theta_2} + \gamma \frac{\partial E(\boldsymbol{\theta})}{\partial \theta_1}. \tag{21}$$

Notice that the nontrivial force can be simply denoted as the combination of the gradient of the energy.

The message of the above calculation is as follows. Even if we use the nontrivial force, we can obtain the same distribution function in the steady state as that of the equilibrium case. We may use the nontrivial force to evolve the system to reach the steady state. The question is which is faster. The answer is the case with the nontrivial force. The nontrivial force violates the DBC and the resultant effect found even in the Fokker-Planck equation. The Fokker-Planck equation can be reformulated by the operator form with the partial derivative. The ordinary Fokker-Planck equation only has the Hermitian operators. On the other hand, the nontrivial force yields the anti-Hermitian operator as in Reference [22]. The part of the Hermitian operators does not change even after induction of the nontrivial force. This is a remarkable property of the nontrivial force. In the same sense as in the case of the master equation [19], the anti-Hermitian (asymmetric) part of the operators shift the eigenvalues characterizing the convergence speed to the steady state. The previous study has proved that the asymmetric part of the transition matrix in the master equation (finite-dimensional case) affects the eigenvalues so that the convergence speed can be faster than that in the case without asymmetric part. In a strict sense, the case for the infinite-dimensional system as in the Fokker-Planck equation has not been yet established but the same discussion can be applied to this case. Therefore the nontrivial force accelerates the convergence to the steady state, which is characterized by the desired distribution function.

3.3. Replication of the system

In addition to the above nontrivial force, we further find a nontrivial solution satisfying the divergence-free condition by considering the replication of the system. Note that this is slightly

different from the replica exchange Monte-Carlo simulation [25]. The replication stands for that we use the replicas of the system governed by the Langevin dynamics. The energy function and the temperature are common. We then consider the evolution of the system following the Langevin equation. In order to accelerate the stochastic dynamics, we introduce additional force in the replicate system. As in the previous subsection, we find the nontrivial force to hold the same distribution function in the steady state as that of the equilibrium case. Let us look at the replicate Fokker-Planck equation

$$\frac{\partial P(\Theta,t)}{\partial t} = -\sum_{r=1}^{R}\sum_{k=1}^{N}\frac{\partial}{\partial\theta_{rk}}J_{rk}(\Theta,t), \tag{22}$$

where Θ is a matrix $\Theta = (\boldsymbol{\theta}_1, \boldsymbol{\theta}_2, \cdots, \boldsymbol{\theta}_R)$ containing the N-dimensional vector for each system denoted by $r = 1, 2, \cdots, R$, and $\mathbf{J}_r(\Theta,t) = (J_{r1}(\Theta,t), J_{r2}(\Theta,t), \cdots, J_{rN}(\Theta,t))$ is the probabilistic flow defined as

$$\mathbf{J}_r(\Theta,t) = \left(\mathbf{A}_r(\Theta) - T\frac{\partial}{\partial\boldsymbol{\theta}_r}\right)P(\Theta,t). \tag{23}$$

The gradient indexed by r is considered for each system. Let us impose the divergence-free condition. The divergence-free condition for the replicate system can be recast as

$$0 = -\sum_{r=1}^{R}\sum_{k=1}^{N}\frac{\partial}{\partial\theta_{rk}}J_{rk}^{(\mathrm{ss})}(\Theta), \tag{24}$$

where

$$\mathbf{J}_r^{(\mathrm{ss})}(\Theta) = \left(\mathbf{A}_r(\Theta) - T\frac{\partial}{\partial\boldsymbol{\theta}_r}\right)\prod_{r=1}^{R}P^{(\mathrm{ss})}(\boldsymbol{\theta}_r). \tag{25}$$

We impose the stationary state of the replicate system as the the product of the identical independent distribution. We can then determine a nontrivial solution by considering a combination of the forces on the replicate system as

$$\mathbf{A}_r(\Theta) = -\frac{\partial}{\partial\boldsymbol{\theta}_r}E(\boldsymbol{\theta}_r) + \gamma\left(\frac{\partial}{\partial\boldsymbol{\theta}_{r+1}}E(\boldsymbol{\theta}_{r+1}) - \frac{\partial}{\partial\boldsymbol{\theta}_{r-1}}E(\boldsymbol{\theta}_{r-1})\right), \tag{26}$$

where the boundary index is taken periodically. Then the probabilistic flow in the steady state is given as

$$\mathbf{J}_r^{(\mathrm{ss})}(\Theta) = \gamma\left(\frac{\partial}{\partial\boldsymbol{\theta}_{r+1}}E(\boldsymbol{\theta}_{r+1}) - \frac{\partial}{\partial\boldsymbol{\theta}_{r-1}}E(\boldsymbol{\theta}_{r-1})\right)\prod_{r=1}^{R}P(\boldsymbol{\theta}_r). \tag{27}$$

We can immediately confirm that the divergence of the above probabilistic flow will vanish because the result of the summation becomes zero. However, the probabilistic flow exists even in the stationary state. This is quite different from an equilibrium system, where there is no divergence of flow. We can confirm that the DBC does not hold in the formulation of the path probability [21, 22]. In other words, the remainder of the probabilistic flow is obtained from the vDBC. This is the Ohzeki-Ichiki method. In the original formulation in the literature, a duplicate system was considered, where $R = 2$. Here, we generalize the original formulation into the replicate system for it to be suitable in the application of the stochastic gradient Langevin method.

A series of the previous studies reveal that faster convergence can be assured by a mathematical argument [19] and is understood as the rare-event sampling [20]. In addition, the Ohzeki-Ichiki method may avoid the critical slowing down that occurs during the phase transition [21]. The method is powerful in this sense and very simple to implement because only the force modification results in remarkable performance. The previous study confirms the reduction in the correlation time as well as faster convergence to the stationary state [22].

Recall the replica exchange Monte Carlo simulation, in which multiple systems with "different" temperatures are driven simultaneously, while each realization is exchanged several times. The technique exhibits outstanding performance with fast convergence to the equilibrium state. However, in the above formulation, the replicate system has a common temperature. We can find a nontrivial solution with the inhomogeneous temperature in the replicate system. In this study, we restrict ourselves to the case with homogeneous temperature, by considering inhomogeneous temperature to be a part of future work.

4. Application to the stochastic gradient method

We are now in a position to apply the Ohzeki-Ichiki method to the stochastic gradient Langevin method. We propose two ways to apply this method to the stochastic gradient Langevin method. Whichever we choose, the Ohzeki-Ichiki method is very simple. In order to perform sampling following the posterior distribution, we employ the stochastic gradient Langevin method. We only evaluate the numerical integration over the Langevin equation, in which the energy function is given by the logarithm of the posterior distribution. In addition to the ordinary force, namely the gradient of the energy function, we employ the nontrivial force as described in the previous section. The mathematical proof ensures the faster convergence to the steady state [19]. Therefore the nontrivial force always accelerate the stochastic dynamics to make the system converge to the steady state. The convergence speed can be increased depending on the strength of the nontrivial force, namely the value of γ. However the upper limit of the value of γ depends on the numerical recipe to integrate the stochastic differential equation. In the following calculation, we employ the Heun scheme not Euler-Maruyama scheme to mitigate the errors in the numerical computation errors. The remaining arbitrariness is choice of the nontrivial force. One is the nontrivial force between the different components in the single system. The other is the nontrivial force in the replicate system. The latter one is advantage for the parallel computation.

4.1. Ohzeki-Ichiki method

In the stochastic gradient Langevin method, we compute the gradient of the subset of the given data ($k = 1, 2, \cdots, d$). The indices of the data are randomly chosen. In the stochastic gradient method, we approximate the force term, i.e., gradient, as follows:

$$-\frac{\partial}{\partial \boldsymbol{\theta}} E(\boldsymbol{\theta}) \approx \frac{\partial}{\partial \boldsymbol{\theta}} \log p(\boldsymbol{\theta}) + \frac{D}{d} \sum_{k=1}^{d} \frac{\partial}{\partial \boldsymbol{\theta}} \log P(\mathbf{x}^{(k)}|\boldsymbol{\theta}). \tag{28}$$

A simple way to apply the Ohzeki-Ichiki method is the direct insertion of this approximate gradient into Equation (19). In other words, we only change the force term for the kth element in the Langevin equation as

$$-\frac{\partial}{\partial \theta_k} E(\boldsymbol{\theta}) \to -\frac{\partial}{\partial \theta_k} E(\boldsymbol{\theta}) + \gamma \left(\frac{\partial}{\partial \theta_{k-1}} E(\boldsymbol{\theta}) - \frac{\partial}{\partial \theta_{k+1}} E(\boldsymbol{\theta}) \right). \tag{29}$$

Therefore no more computational cost is needed for implementation of the Ohzeki-Ichiki method. This is a remarkable property of the Ohzeki-Ichiki method. To enhance precision and improve

the computation speed, additional efforts for its performance are usually needed. However this is not the case. In addition, we propose an alternative application of the Ohzeki-Ichiki method by considering the replicate system as shown below.

4.2. Ohzeki-Ichiki method for replicate system

In order to obtain more sampling points following the posterior distribution, we may perform parallel computation of the stochastic gradient Langevin method with a smaller batch size. This motivation is consistent with the concept of the Ohzeki-Ichiki method, in which we may consider that the replication of the system will generate a nontrivial force. We then distribute the approximate gradient into the replicated system. In other words, we apply the gradient determined by a different dataset, i.e., the further divided subset of a given data where the number of subsets is set to be d_r such that $d = \sum_{r=1} d_r$. For simplicity, we use $d_r = d/R$ in the numerical test.

We first divide the given subset of d into the replicate system as d_r for each r system. Then we evaluate the gradient for each replicate system. The computational cost for evaluation of the gradient can be reduced to a factor d_r/d. We have the value of the gradient for each replicate system as follows:

$$-\frac{\partial}{\partial \boldsymbol{\theta}_r} E(\boldsymbol{\theta}_r) \approx \frac{\partial}{\partial \boldsymbol{\theta}} \log p(\boldsymbol{\theta}) + \frac{D}{d_r} \sum_{k_r=1}^{d_r} \frac{\partial}{\partial \boldsymbol{\theta}} \log P(\mathbf{x}^{(k_r)}|\boldsymbol{\theta}). \tag{30}$$

Then we make the combination of the gradient conforming the nontrivial force as in Equation (27). In other words, we only change the force term in the Langevin equation as

$$-\frac{\partial}{\partial \boldsymbol{\theta}_r} E(\boldsymbol{\theta}_r) = -\frac{\partial}{\partial \boldsymbol{\theta}_r} E(\boldsymbol{\theta}_r) + \gamma \left(\frac{\partial}{\partial \boldsymbol{\theta}_{r+1}} E(\boldsymbol{\theta}_{r+1}) - \frac{\partial}{\partial \boldsymbol{\theta}_{r-1}} E(\boldsymbol{\theta}_{r-1}) \right) \tag{31}$$

In this sense, the nontrivial force is a collection of the gradient computed in different replica. The different replica has the individual data as the subset for evaluation of the gradient. To make the nontrivial force, all we need to know is the value of the gradient for the different data set distributed to each replica. This is done by communication between different replicas. The communication cost to send the data flow between different replicas is demanded but not significant. The gain by the replicate system overcomes the communication cost. The computational cost can be basically reduced to a factor d_r/d due to the parallel computation. In addition, because we replicate the system, we generate more sampling points from the given datasets. In this sense, the proposed scheme is very important for sampling following the posterior distribution.

The convergence speed compared to the ordinary evolution by the Langevin dynamics is definitely faster since the Ohzeki-Ichiki method holds mathematical assurance as in the previous study [19]. We should emphasize that its implementation is extremely simple. Additional computation does not need to perform the Ohzeki-Ichiki method. We only reuse the value of the gradient to make the nontrivial force. In the present paper, we propose the application of the Ohzeki-Ichiki method to the stochastic gradient Langevin method. In the next section, we test our method in the simple model although we push up the quantitative evaluation of performance in the future study.

5. Numerical tests

Let us test the stochastic gradient Langevin method with the Ohzeki-Ichiki method. Similarly to the first study on the stochastic gradient Langevin method [16], we test our algorithm for the

Gaussian mixture distribution with tied mean, in which the probability distribution of the data is defined as

$$P(x|\boldsymbol{\theta}) = \frac{1}{2}\sqrt{\frac{b}{2\pi}}\exp\left(-\frac{1}{2}b\left(x-\theta_1\right)^2\right) + \frac{1}{2}\sqrt{\frac{b}{2\pi}}\exp\left(-\frac{1}{2}b\left(x-\theta_1-\theta_2\right)^2\right) \tag{32}$$

where b denotes the precision (the inverse of the variance) and the parameters $\boldsymbol{\theta} = (\theta_1, \theta_2)$ are related to the mean of the Gaussian distribution. The prior distribution is assumed to be a Gaussian distribution as

$$P(\boldsymbol{\theta}) = \left(\frac{a_1 a_2}{2\pi}\right)\exp\left(-\frac{1}{2}a_1\theta_1^2 - \frac{1}{2}a_2\theta_2^2\right) \tag{33}$$

where a_1 and a_2 denote the precisions for the parameters. This model consists of a bimodal distribution. Thus the MAP estimation cannot capture the uncertainty of the parameters. The posterior distribution can be computed exactly as

$$P(\boldsymbol{\theta}|x) = \frac{1}{Z(x)}P(x|\boldsymbol{\theta})P(\boldsymbol{\theta}), \tag{34}$$

where

$$Z(x) = \sqrt{\frac{\alpha_1}{2\pi}}\exp\left(-\frac{\alpha_1}{2}x^2\right) + \sqrt{\frac{\alpha_2}{2\pi}}\exp\left(-\frac{\alpha_2}{2}x^2\right) \tag{35}$$

and

$$\alpha_1 = \frac{a_1 b}{a_1 + b} \tag{36}$$

$$\alpha_2 = \frac{a_1 a_2 b}{a_1 a_2 + (a_1 + a_2)b}. \tag{37}$$

We generate $D = 100$ samples of $x^{(k)}$ from one of the Gaussian distribution at $(\theta_1, \theta_2) = (2, -2)$. The model has another-mode solution as $(\theta_1, \theta_2) = (0, 2)$, which has strong negative correlation between the parameters.

Let us perform the stochastic gradient Langevin method with the Ohzeki-Ichiki method. The gradient of the log-likelihood function is evaluated as

$$\begin{aligned}
\frac{\partial}{\partial\theta_1}\log P(x|\boldsymbol{\theta}) &= \frac{b(x-\theta_1)\exp(-b(x-\theta_1)^2/2) + b(x-\theta_1-\theta_2)\exp(-b(x-\theta_1-\theta_2)^2/2)}{\exp(-b(x-\theta_1)^2/2) + \exp(-b(x-\theta_1-\theta_2)^2/2)} \\
\frac{\partial}{\partial\theta_2}\log P(x|\boldsymbol{\theta}) &= \frac{b(x-\theta_1-\theta_2)\exp(-b(x-\theta_1-\theta_2)^2/2)}{\exp(-b(x-\theta_1)^2/2) + \exp(-b(x-\theta_1-\theta_2)^2/2)}.
\end{aligned} \tag{38}$$

We set $a_1 = a_2 = 0.1$ and $b = 10$, which yield a large barrier between the two modes at $(\theta_1, \theta_2) = (0, 2)$ and $(\theta_1, \theta_2) = (2, -2)$. We iterate the update process for 10^6 steps. The step size decreases as $\beta(t+\delta) - \epsilon$, where $\epsilon = 0.55$ and δ and β are determined such that step sizes changes from $dt = 0.01$ to $dt = 0.0001$. First let us compare the results with the batch size for the stochastic gradient $d = 1$. Under this setting, the original stochastic gradient Langevin method usually fails to generate the sampling points following the posterior distribution in short time. Here, we demonstrate a more difficult case for the generation of the sampling points following the posterior distribution than the previous study [16]. We show the obtained sampling points through a density plot. We compare the stochastic gradient Langevin method with and without the application of the Ohzeki-Ichiki method with $\gamma = 2$ and 5 using the nontrivial force in Equations (20) and (21) as shown in Fig. 1. We can confirm that a wide range of the sampling

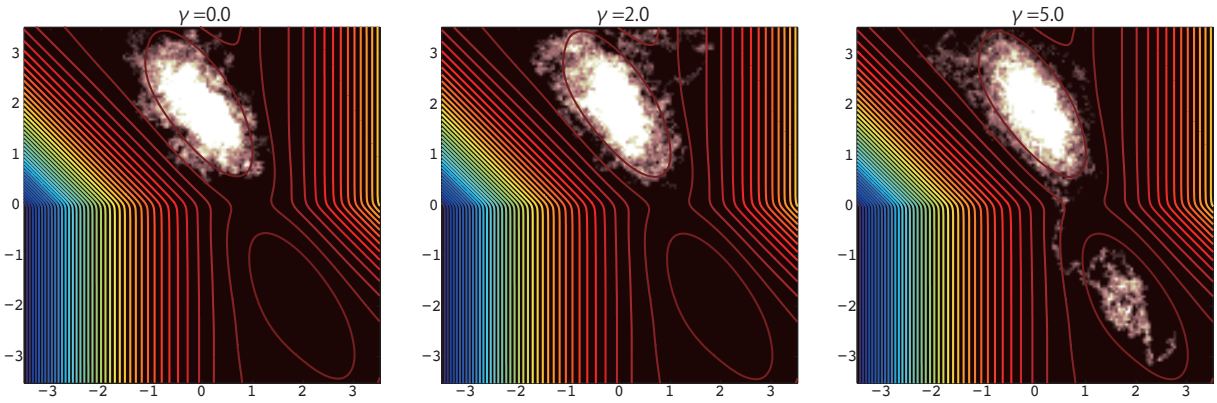

Figure 1. (Colour online) Density plot and the contour plot of the exact posterior distribution of the batch size $d = 1$. The number of iterations is $\Delta t = 10^6$. The left panel shows the stochastic gradient Langevin method, the centre one describes that with the Ohzeki-Ichiki method with $\gamma = 2$, and the right one represents that with $\gamma = 5$. The posterior distribution is plotted by the colour-gradation curves from red (higher values) to blue (lower values).

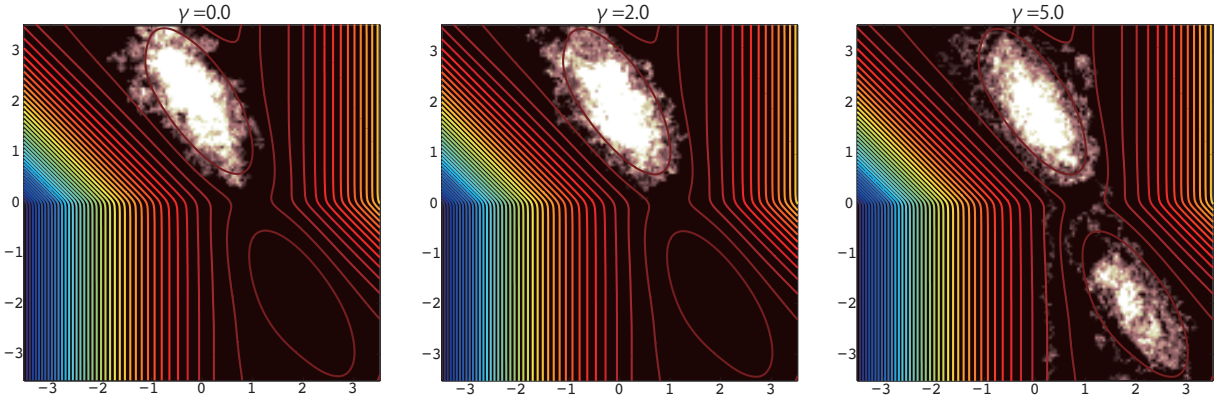

Figure 2. (Colour online) Density plot and the contour plot of the exact posterior distribution of the batch size $d = 10$. The number of iterations is $\Delta t = 10^6$. The same symbols are used in Figure 1.

points are obtained by the Ohzeki-Ichiki method. In particular, the case where $\gamma = 5$ exhibits a leap from one mode to the other mode. In other words, for the case with the application of the Ohzeki-Ichiki method with a larger value of the degree of vDBC estimates more accurate the posterior distribution.

Figure 2 shows the results of the stochastic gradient Langevin method and its improved version by the application of the Ohzeki-Ichiki method with the batch size $d = 10$. Similar to the previous case, for $d = 1$, we again confirm the efficient sampling by the application of the Ohzeki-Ichiki method.

Let us consider another application of the Ohzeki-Ichiki method dealing with the replicate system. We employ the Ohzeki-Ichiki method for the parallel update of the $R = d = 10$ replicate system. For each system, we use one datum randomly chosen for the gradient to update the tentative estimation value. We perform parallel computation of the update equation over $R = d = 10$ replicate system. Then we introduce the nontrivial force. We only need to

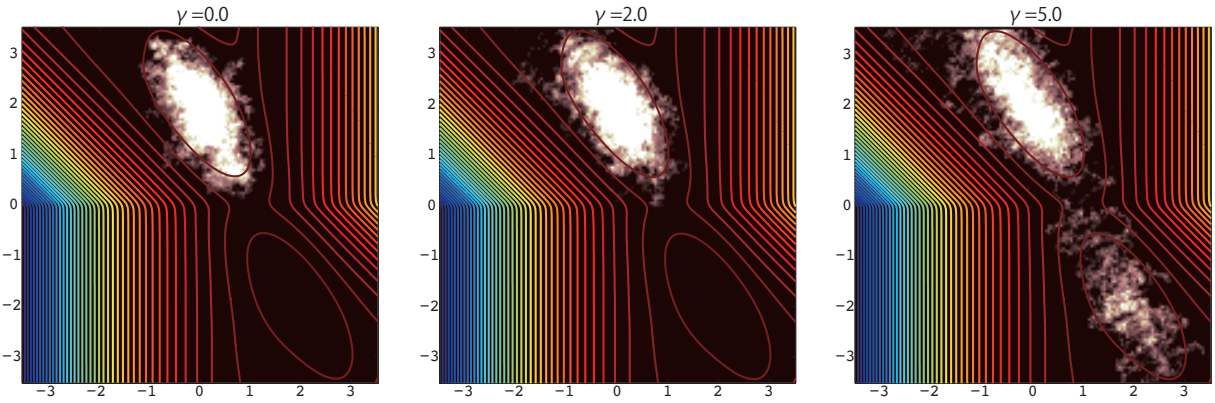

Figure 3. (Colour online) Density plot and the contour plot of the exact posterior distribution by the replicate system $R = 10$ and the batch size $d = 10$. We divide the mini batch d into d/R. The number of iterations is $\Delta t = 10^5$. The same symbols are used in Figure 1.

communicate the value of the gradient for the given subset $d_r = d/R = 1$ with different replicas. In other words, we "reuse" the data for efficient sampling as the nontrivial force. Figure 3 shows the results of the parallel computation of the stochastic gradient Langevin method and those of the case induced by the nontrivial force with $\gamma = 2$ and $\gamma = 5$. In this computation, we reduce the number of steps to $\Delta t = 10^5$. Efficient sampling can be performed by using the replicate system.

We have not yet compared the quantitative performance of the methods. However one can confirm their qualitative performance from the results shown in the figures. We find that a wide range of the sampling can be obtained by use of the larger value of γ. Similarly to the previous case, the sampling points without the nontrivial force ($\gamma = 0$) can not go over the ridge to the other mode of the posterior distribution. On the other hand, the nontrivial force of $\gamma = 5$ in the replicate system brings the sampling point in the other mode.

In addition, we plot the results of the long-time computation up to $\Delta t = 10^6$ by using the replicate system in Figure 4. Even for the case without the nontrivial force, the sampling points are found in a relatively wide range. The sampling from the bimodal distribution seems to be performed. However increase in the γ yields the more efficient sampling more correctly following the posterior distribution, as shown in Figure 4.

It may be assumed that the infinitely large nontrivial force causes the system to be in a stationary state immediately. This may or may not be true. Indeed increase in γ makes the system mixed rapidly and accelerates the convergence to a stationary state. However, depending on the value of γ, the numerical integration of the Langevin dynamics may become worse. To stabilize the computation of the Langevin dynamics, more sophisticated integration methods or the Metropolis-Hasting method may be utilized as in the original proposal of the stochastic gradient Langevin method. From mathematical and physical point of view, the Ohzeki-Ichiki method can ensure that a stationary state is reached faster than the case under the DBC. Thus, the increase in the degree of vDBC has some tradeoff between performance and instability. The optimal value of γ will be part of future work. The answer may be obtained from the variational principle discussed in a previous study [23].

6. Summary

In this study, we proposed the application of the Ohzeki-Ichiki method in the stochastic gradient Langevin method. The Ohzeki-Ichiki method violates the DBC, which is a sufficient condition

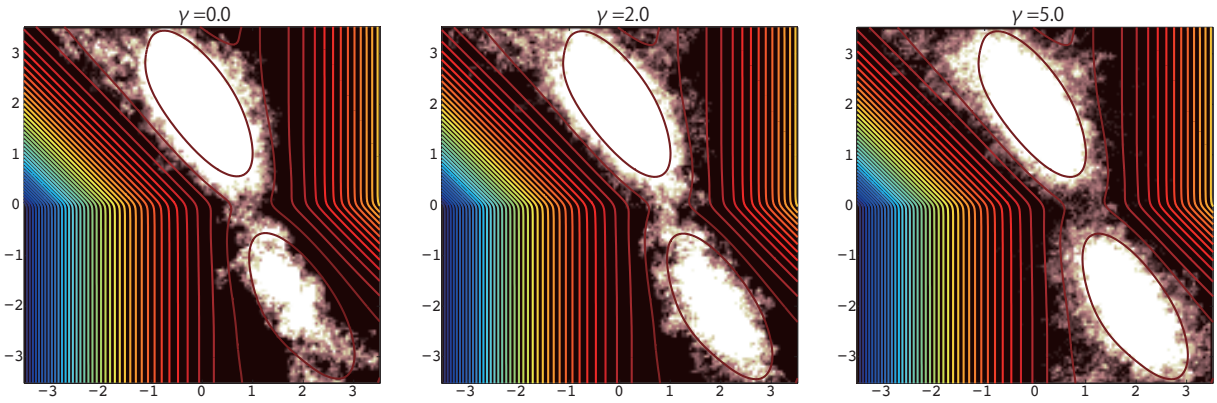

Figure 4. (Colour online) Density plot and the contour plot of the exact posterior distribution by the replicate system $R = 10$ and the batch size $d = 10$. We divide the mini batch d into d/R. The number of iterations is $\Delta t = 10^6$. The same symbols are used in Figure 1.

to ensure convergence to a stationary state, and shows remarkable performance for reaching a stationary state compared to the standard equilibrium case under the DBC. We applied this method to the stochastic gradient Langevin method, which approximates the gradient of the log-likelihood function in a stochastic manner and utilizes prior distribution to capture the uncertainty of the leaned parameters. The acceleration of the convergence to a stationary state improved the performance of the stochastic gradient Langevin method as expected. The range for the sampling points was wider as compared to the original method. In addition, we developed a new technique to implement the Ohzeki-Ichiki method based on the idea of the stochastic gradient method. We introduced the nontrivial force in the replicated system, which is the gradient determined by the distributed data from the given subset. This replication generated more sampling points and the nontrivial force made the replicate system converge to the stationary state faster than the ordinary approach. These are just the preliminary results of the Ohzeki-Ichiki method for machine learning problems. Future study will reveal its significance in this field and confirms the quantitative assessment of the Ohzeki-Ichiki method. We again emphasize that the implementation of the Ohzeki-Ichiki method is relatively simpler than other methods to accelerate the convergence to a stationary state. This method is simply the addition of the nontrivial force. Hence, the proposed method makes advancement in the machine learning process.

Acknowledgement

The author would like to thank M. Yasuda, and A. Ichiki for the fruitful discussions. This work is financially supported by the JST-CREST, JSPS KAKENHI Grants No. 25120008 and No. 15H03699, and the Kayamori Foundation of Informational Science Advancement.

References

[1] Ackley D H, Hinton G E and Sejnowski T J 1985 *Cognitive Science* **9** 147–169
[2] Hinton G E, Osindero S and Teh Y W 2006 *Neural Comput.* **18** 1527–1554
[3] Hinton G E and Salakhutdinov R R 2006 *Science* **313** 504–507
[4] Pankaj M and David J S 2014 **stat.ML/1410.3831**
[5] Ohzeki M 2015 *Journal of the Physical Society of Japan* **84** 054801
[6] Decelle A and Ricci-Tersenghi F 2014 *Phys. Rev. Lett.* **112**(7) 070603
[7] Yamanaka S, Ohzeki M and Decelle A 2015 *Journal of the Physical Society of Japan* **84** 024801
[8] Sessak V and Monasson R 2009 *Journal of Physics A: Mathematical and Theoretical* **42** 055001

[9] Cocco S and Monasson R 2011 *Phys. Rev. Lett.* **106**(9) 090601
[10] Cocco S and Monasson R 2012 *Journal of Statistical Physics* **147** 252–314
[11] Ricci-Tersenghi F 2012 *Journal of Statistical Mechanics: Theory and Experiment* **2012** P08015
[12] Yasuda M and Tanaka K 2013 *Phys. Rev.* E **87**(1) 012134
[13] Raymond J and Ricci-Tersenghi F 2013 *Phys. Rev.* E **87**(5) 052111
[14] Ohzeki M 2013 *Journal of Physics: Conference Series* **473** 012005
[15] Robbins H and Monro S 1951 *Ann. Math. Statist.* **22** 400–407
[16] Welling M and Teh Y W 2011 *Proceedings of the International Conference on Machine Learning*
[17] Suwa H and Todo S 2010 *Phys. Rev. Lett.* **105**(12) 120603
[18] Turitsyn K S, Chertkov M and Vucelja M 2011 *Physica D: Nonlinear Phenomena* **240** 410 – 414
[19] Ichiki A and Ohzeki M 2013 *Phys. Rev.* E **88**(2) 020101
[20] Ichiki A and Ohzeki M 2015 *Phys. Rev.* E **91**(6) 062105
[21] Ohzeki M and Ichiki A 2015 *Phys. Rev.* E **92**(1) 012105
[22] Ohzeki M and Ichiki A 2015 *Journal of Physics: Conference Series* **638** 012003
[23] Takahashi K and Ohzeki M 2016 *Phys. Rev.* E **93**(1) 012129
[24] Swendsen R H and Wang J S 1987 *Phys. Rev. Lett.* **58**(2) 86–88
[25] Hukushima K and Nemoto K 1996 *Journal of the Physical Society of Japan* **65** 1604–1608
[26] Neal R 2001 *Statistics and Computing* **11** 125–139
[27] Ohzeki M 2010 *Phys. Rev. Lett.* **105**(5) 050401
[28] Ohzeki M and Nishimori H 2010 *J. Phys. Soc. Jpn.* **79**(8) 084003
[29] Ohzeki M and Nishimori H 2011 *Journal of Physics: Conference Series* **302** 012047
[30] Ohzeki M and Nishimori H 2011 *Physica E: Low-dimensional Systems and Nanostructures* **43** 782–785
[31] Ohzeki M 2012 *Phys. Rev.* E **86**(6) 061110