Physica A 491 (2018) 1014–1022

Contents lists available at ScienceDirect

Physica A

journal homepage: www.elsevier.com/locate/physa

Parameter inference in a probabilistic model from data: Regulation of transition rate in the Monte Carlo method

Hirohito Kiwata

Division of Natural Science, Osaka Kyoiku University, Kashiwara, Osaka 582-8582, Japan

HIGHLIGHTS

- A new method for inference in the inverse Ising problem.
- Our method is available by using the Metropolis–Hastings or heat-bath algorithm.
- The accuracy of the inference by our method is similar to that by pseudo-likelihood method.
- Our method is superior to minimum probability flow.
- The required computational task of the inference is less than that of pseudo-likelihood method.

ARTICLE INFO

Article history:
Received 12 November 2016
Received in revised form 27 September 2017
Available online 16 October 2017

Keywords:
Inverse Ising problem
Inference of parameters in a probabilistic model
Pseudo-likelihood method
Minimum probability flow

ABSTRACT

We consider the inference of parameters in a probabilistic model from a data set, which is generated by an unknown probabilistic model. The Monte Carlo method is a tool for obtaining a data set obeying a given probability distribution. A set of transition rates is required to satisfy three conditions (irreducible, aperiodic, and stationary) for a sampled data set to represent a probability distribution. We utilize the stationary condition of a probability distribution with respect to transition rates to infer parameters. A frequency distribution by a data set substitutes for an unknown probability distribution in the condition. Our method includes minimum probability flow as a special case and becomes superior to it as the number of samples increases.

1. Introduction

With the progress of information technology, we can obtain various types of data, and deriving useful information from big data has become necessary. In the fields of science and technology, large and complex data require new modeling methods and analytic techniques that surpass the traditional methods. We consider a modeling method that includes an inference method for parameters in a constructed model from a given data set. Particularly, a probabilistic model is investigated in this paper. The problem of obtaining various measurement quantities from a given probabilistic model corresponds to a forward problem, and the Monte Carlo method is a well-known standard method of obtaining data from a known probabilistic distribution. To infer parameters in probabilistic models by using a data set corresponds to an inverse problem; this has drawn significant attention recently and is becoming increasingly important [1].

Machine learning is a present-day subject of inverse problem. The system of a learning machine is designed to give desired outputs with sample inputs. The composition of the desired system from sample data corresponds to the inverse problem. The Boltzmann machine classified as a stochastic recurrent neural network is expected to be a learning machine for expressing data containing rich constitution [2]. To obtain desired outputs by the Boltzmann machine, we must regulate

E-mail address: kiwata@cc.osaka-kyoiku.ac.jp.

https://doi.org/10.1016/j.physa.2017.09.103
0378-4371/

the parameters of the model. The standard method of settling the problem of parameter inference in the Boltzmann machine is using the Kullback–Leibler divergence between the probabilistic model and frequency distribution derived from a data set, and the parameters are determined to minimize the Kullback–Leibler divergence. The learning of the Boltzmann machine corresponds to minimization of likelihood. However, the method used to tune the parameters in the Boltzmann machine requires huge arithmetic tasks to be performed as the number of units increases. To overcome this difficulty, various approximate methods have been developed over decades to enable the learning of the Boltzmann machine.

The pseudo-likelihood method is one of the approximate methods for inferring parameters from a data set and has a long history [3]. It has increasingly gained attention as the progress of computer's performance [4–7]. Since the Boltzmann machine resembles a spin-glass model in statistical physics, it has been gaining considerable attention not only of computer scientists but also physicists [8]. The problem of parameter inference has been addressed from the physics point of view. The Kullback–Leibler divergence is the same as the free energy of the system, which is analytically evaluated as an expanded form with respect to parameters [9–20]. The expanded form of the free energy [21,22] gives an accurate estimation of parameters as long as random variables are weakly coupled. This method has the advantage of requiring less evaluation effort; however, it has a limitation of no utilization in the range of strong coupling of random variables. As a recent useful method, contrastive divergence has been applied to multi-layer neural networks [23]. Multi-layer neural networks have gained significant attention in connection with deep learning [24,25].

We propose a new inference method, and the idea of our method originates from the generation of data from a given probability distribution by using the Monte Carlo method. The transition rate of the Monte Carlo method is a function with respect to parameters of the probabilistic model, and we utilize the transition rate for the inference of parameters inversely. In our previous paper, we proposed an inference method which makes use of the detail balance condition [26]. The accuracy of this inference method is similar to that of the pseudo-likelihood method. In this paper, we utilize another condition of the transition rate, i.e. the stationary condition of a probability distribution with respect to transition rates. Minimum probability flow is an inference method similar in concept to our method with regard to utilization of transition rate [27]. However, minimum probability flow is a special case of our method. We investigate the accuracy of inference and compare it with those of the pseudo-likelihood method and minimum probability flow. These methods are applied to two types of probabilistic models, i.e. dense and sparse connection of random variables of probabilistic models. We demonstrate that the accuracy of inference depends not only on the inference methods but also on probabilistic models.

2. Theory

Consider a set of N discrete numbers, $\boldsymbol{x} = \{x_1, x_2, \ldots, x_N\}$. A set $\boldsymbol{x}$ is generated by an unknown probability distribution. We want to estimate parameters in probabilistic models from given data $\boldsymbol{x}$. To explain our idea concretely, we prescribe x_i binary, i.e. $x_i \in \{-1, +1\}$. The prescription for x_i does not restrict the effectiveness of our method. Then, the set is given by $\boldsymbol{x} \in \{-1, +1\}^N$. We assume that data $\boldsymbol{x}$ is generated by the following probability distribution:

$$P(\boldsymbol{x}|\{J_{ij}\}, \{h_i\}) = \frac{1}{Z(\{J_{ij}\}, \{h_i\})} \exp\Big(\sum_{i=1}^{N}\sum_{j\in\partial i} J_{ij}x_i x_j + \sum_{i=1}^{N} h_i x_i\Big), \tag{1}$$

where J_{ij} is a coupling parameter, and h_i is a bias. The sign $j \in \partial i$ represents the summation of j connecting with i by J_{ij}. The denominator $Z(\{J_{ij}\}, \{h_i\})$ corresponds to a partition function, which is introduced to normalize the probability distribution. We consider a graph consisting of N vertices and some edges connecting each vertex. An ith element of $\boldsymbol{x}$, i.e. x_i, is located on the ith vertex. There is an edge between the ith and jth vertices in the case of $J_{ij} \neq 0$. For simplicity, we assume J_{ij} is symmetric, i.e. $J_{ij} = J_{ji}$, and set $J_{ii} = 0$. The above probability distribution corresponds to a graph. The problem is to infer a set of $\{J_{ij}\}$ and $\{h_i\}$ from M sets of $\boldsymbol{x}$, i.e. $\mathcal{D} = \{\boldsymbol{x}^{(1)}, \boldsymbol{x}^{(2)}, \ldots, \boldsymbol{x}^{(M)}\}$. For the inference of parameters in a probabilistic model, maximum likelihood estimation is useful. Given $\mathcal{D}$, we define a histogram corresponding to a frequency distribution function, such as

$$Q(\boldsymbol{x}) = \frac{1}{M}\sum_{\mu=1}^{M} \delta(\boldsymbol{x}, \boldsymbol{x}^{(\mu)}), \tag{2}$$

where $\delta(\boldsymbol{x}, \boldsymbol{x}')$ is the Kronecker delta, which is equal to unity when $\boldsymbol{x} = \boldsymbol{x}'$ and zero when $\boldsymbol{x} \neq \boldsymbol{x}'$. With the frequency distribution function, we define the log-likelihood as

$$\begin{aligned}\mathcal{L}(\{J_{ij}\}, \{h_i\}) &= \sum_{\boldsymbol{x}\in\{-1,+1\}^N} Q(\boldsymbol{x}) \ln P(\boldsymbol{x}|\{J_{ij}\}, \{h_i\}) = \frac{1}{M}\sum_{\mu=1}^{M} \ln P(\boldsymbol{x}^{(\mu)}|\{J_{ij}\}, \{h_i\}) \\ &= \frac{1}{M}\sum_{\mu=1}^{M}\Big(\sum_{i=1}^{N}\sum_{j\in\partial i} J_{ij}x_i^{(\mu)}x_j^{(\mu)} + \sum_{i=1}^{N} h_i x_i^{(\mu)}\Big) - \ln Z(\{J_{ij}\}, \{h_i\}).\end{aligned} \tag{3}$$

To find $\{\hat{J}_{ij}\}$ and $\{\hat{h}_i\}$, which maximize the log-likelihood Eq. (3), we differentiate this equation with respect to J_{ij} or h_i. However, the evaluation of the partition function requires tedious tasks and hinders the estimation of $\{\hat{J}_{ij}\}$ and $\{\hat{h}_i\}$. To resolve