Pseudo Temperature of Observed Data

Muneki YASUDA*

Graduate School of Information Sciences, Tohoku University, Sendai 980-8579, Japan

Received May 17, 2012; final version accepted August 28, 2012

A Boltzmann machine (BM) is a basic learning model forming a Markov random field, and many approximate learning algorithms for it so far. In the present paper, a new strategy for approximate BM learnings is proposed by introducing a temperature of observed data which controls a smoothness of empirical distribution. By controlling the temperature, one can obtain better solutions to BM learning with an existing approximate learning algorithm.

1. Introduction

A Boltzmann machine (BM) is a basic learning model forming a Markov random field [1]. So far, many approximate learning algorithms for it have been proposed because an exact learning algorithm for a BM, i.e., an exact minimization of a Kullback–Leibler divergence (KLD) between a BM and an empirical distribution which is a histogram of given observed data set [3, 4], is computationally expensive one which grows exponentially with the increase of size of system. Because BMs are mathematically equivalent to Ising models known in statistical physics, BM learnings have been investigated not only in information sciences but also in physics.

In a BM learning, one must treat an intractable quantity termed a free energy of given system. Therefore, it often be the central key of BM learnings that how approximate the free energy. Mean-field methods, which are physical approaches, are effective techniques for approximating free energies, and most approximate learning algorithms are constructed by approximating free energies by using those methods [2]. In a conventional way, since an empirical distribution is uniquely determined by a given data set, once one fixes an approximate method, a solution to a BM learning is determined. Hence, if one wants to obtain a better solution, one has to employ a higher order approximation. Researching more effective approximations for free energies is important for BM learnings, but we approach BM learnings from a different way in the present paper.

In this paper, a new strategy for approximate BM learnings is introduced by adding a controllable parameter, which is regarded as the inverse temperature in statistical physics, to given empirical distributions. The controllable parameter, which is referred as *the pseudo-temperature* of observed data in the present paper, can control a smoothness of empirical distribution and can help a fixed approximate BM learning.

2. Boltzmann Machine Learning

Let us consider an undirected graph $G(V, E)$ consisting of n nodes and some links; each node is labeled as i . Let $V = \{1, 2, \dots, n\}$ be the set of all nodes. A link between nodes i and j is labeled as (i, j) and let E be the set of all links. Because we consider undirected links, (i, j) and (j, i) denote the same link. We assign ± 1 binary random variable $x_i \in \{+1, -1\}$ to node i . On the graph, we define a BM as

$$P_B(\mathbf{x} \mid \boldsymbol{\theta}, \mathbf{w}) := \frac{1}{Z_B(\boldsymbol{\theta}, \mathbf{w})} \exp\left(\sum_{i \in V} \theta_i x_i + \sum_{(i,j) \in E} w_{ij} x_i x_j\right), \quad (1)$$

where $Z_B(\boldsymbol{\theta}, \mathbf{w})$ is the partition function

$$Z_B(\boldsymbol{\theta}, \mathbf{w}) := \sum_{\mathbf{x}} \exp\left(\sum_{i \in V} \theta_i x_i + \sum_{(i,j) \in E} w_{ij} x_i x_j\right).$$

The parameters $\boldsymbol{\theta} := \{\theta_i \mid i \in V\}$ and $\mathbf{w} := \{w_{ij} \mid (i, j) \in E\}$ are the vector of biases and the matrix of symmetric connections, respectively. The ij elements of $\mathbf{w}$, w_{ij} , are zero if $(i, j) \notin E$ or $i = j$.

Given M observed data, $\mathcal{D} := \{\mathbf{d}^{(\mu)} \in \{+1, -1\}^n \mid \mu = 1, 2, \dots, M\}$, we define a Kullback–Leibler divergence (KLD) presented as

* Corresponding author. E-mail: muneki@smapi.is.tohoku.ac.jp

$$D^{(M)}(\boldsymbol{\theta}, \mathbf{w}) := \sum_{\mathbf{x}} Q(\mathbf{x}) \ln \frac{Q(\mathbf{x})}{P_B(\mathbf{x} | \boldsymbol{\theta}, \mathbf{w})}. \quad (2)$$

The distribution $Q(\mathbf{x})$ is defined by

$$Q(\mathbf{x}) := \frac{1}{M} \sum_{\mu=1}^M \delta(\mathbf{x}, \mathbf{d}^{(\mu)})$$

which is the empirical distribution constructed by the given data, where the function $\delta(\mathbf{x}, \mathbf{d})$ is the Kronecker delta function. The BM learning is achieved by minimizing the KLD with respect to $\boldsymbol{\theta}$ and $\mathbf{w}$. However, the direct minimization of the KLD is computationally hard due to the multiple summations and it requires an approximation.

The KLD is rewritten as

$$D^{(M)}(\boldsymbol{\theta}, \mathbf{w}) = - \sum_{i \in V} \theta_i \langle x_i \rangle_{\mathcal{D}} - \sum_{(i,j) \in E} w_{ij} \langle x_i x_j \rangle_{\mathcal{D}} - F_B(\boldsymbol{\theta}, \mathbf{w}) + c, \quad (3)$$

where the notation $\langle \cdots \rangle_{\mathcal{D}}$ denotes the average over the empirical distribution and the expression c is a constant. The expression $F_B(\boldsymbol{\theta}, \mathbf{w})$ is the free energy of the BM defined by $F_B(\boldsymbol{\theta}, \mathbf{w}) := -\ln Z_B(\boldsymbol{\theta}, \mathbf{w})$. Most approximate learning algorithm for the BM based on mean-field approaches, called *mean-field learning algorithms*, are constructed by approximating the free energy by mean-field approximations: the belief propagation for instance. Hence, the quality of the BM learning directly depends on the quality of employed approximation.

3. Pseudo-Temperature of Observed Data

Let us extend the empirical distribution to add a certain positive constant β as

$$\widehat{Q}(\mathbf{x} | \beta) \propto Q(\mathbf{x})^\beta. \quad (4)$$

As $\beta \rightarrow 0$ the distribution $\widehat{Q}(\mathbf{x} | \beta)$ approaches a uniform distribution, while as $\beta \rightarrow \infty$ it approaches a delta function. Hence, we see that the constant β controls smoothness of $\widehat{Q}(\mathbf{x} | \beta)$.

To study a role of β in detail, let us consider an ideal situation in which we obtain a sufficient number of observed data, i.e., $M \rightarrow \infty$, which are generated from a BM, $P_B(\mathbf{x} | \boldsymbol{\theta}^*, \mathbf{w}^*)$. In this situation, the empirical distribution can be well approximated by the generative BM, i.e., $Q(\mathbf{x}) = P_B(\mathbf{x} | \boldsymbol{\theta}^*, \mathbf{w}^*)$ and the extended empirical distribution $\widehat{Q}(\mathbf{x} | \beta)$ is equivalent to $P_B(\mathbf{x} | \beta\boldsymbol{\theta}^*, \beta\mathbf{w}^*)$. Thus, the constant β can control the magnitudes of $\boldsymbol{\theta}^*$ and $\mathbf{w}^*$. Such parameter is known as the inverse temperature in statistical physics. Smaller β represents higher temperature. Since it is known that most mean-field learning algorithms succeed when $\boldsymbol{\theta}^*$ and $\mathbf{w}^*$ are small, we expect that a mean-field learning algorithm gives results with higher qualities for smaller β .

For the extended empirical distribution, we extend the BM by the same manner,

$$\widehat{P}_B(\mathbf{x} | \boldsymbol{\theta}, \mathbf{w}, \beta) \propto P_B(\mathbf{x} | \boldsymbol{\theta}, \mathbf{w})^\beta \propto P_B(\mathbf{x} | \beta\boldsymbol{\theta}, \beta\mathbf{w}), \quad (5)$$

and, for given observed data and β , we introduce a KLD between $\widehat{Q}(\mathbf{x} | \beta)$ and $P_B(\mathbf{x} | \beta\boldsymbol{\theta}, \beta\mathbf{w})$ as

$$\begin{aligned} D_\beta^{(M)}(\boldsymbol{\theta}, \mathbf{w}) &:= \sum_{\mathbf{x}} \widehat{Q}(\mathbf{x} | \beta) \ln \frac{\widehat{Q}(\mathbf{x} | \beta)}{P_B(\mathbf{x} | \beta\boldsymbol{\theta}, \beta\mathbf{w})} \\ &= - \sum_{i \in V} \theta_i \langle x_i \rangle_\beta - \sum_{(i,j) \in E} w_{ij} \langle x_i x_j \rangle_\beta - F_B(\beta\boldsymbol{\theta}, \beta\mathbf{w}) + \widehat{c}(\beta), \end{aligned} \quad (6)$$

where notation $\langle \cdots \rangle_\beta$ denotes the average over $\widehat{Q}(\mathbf{x} | \beta)$ and the expression $\widehat{c}(\beta)$ is a function unrelated to $\boldsymbol{\theta}$ and $\mathbf{w}$. We minimize the new KLD instead of the original one. Note that the new KLD coincides to the original one when $\beta = 1$.

By applying the approximate learning algorithm using the belief propagation to minimizing the new KLD, $D_\beta^{(M)}(\boldsymbol{\theta}, \mathbf{w})$, solutions are obtained as [2, 5]

$$\begin{aligned} \theta_i &= \frac{(1 - |\partial(i)|)}{\beta} \tanh^{-1} \langle x_i \rangle_\beta \\ &\quad + \frac{1}{2\beta} \sum_{j \in \partial(i)} \left(\tanh^{-1} \frac{\langle x_i \rangle_\beta + \langle x_i x_j \rangle_\beta}{1 + \langle x_j \rangle_\beta} + \tanh^{-1} \frac{\langle x_i \rangle_\beta - \langle x_i x_j \rangle_\beta}{1 - \langle x_j \rangle_\beta} \right), \end{aligned} \quad (7)$$

$$w_{ij} = \frac{1}{2\beta} \left(\tanh^{-1} \frac{\langle x_i \rangle_\beta + \langle x_i x_j \rangle_\beta}{1 + \langle x_j \rangle_\beta} - \tanh^{-1} \frac{\langle x_i \rangle_\beta - \langle x_i x_j \rangle_\beta}{1 - \langle x_j \rangle_\beta} \right) \quad (8)$$

for $\forall i \in V$ and $(i, j) \in E$, where the notation $\partial(i)$ is the set of nodes directly connecting to node i . Note that Eqs. (7) and (8) are reduced to the BM learning using the belief propagation when $\beta = 1$.

4. Numerical Experiments

Let us suppose that the generative model which generate observed data is a BM, $P_B(\mathbf{x} | \boldsymbol{\theta}^*, \mathbf{w}^*)$, defined on a fully-connected graph with $n = 10$. The parameters of the generative model are randomly generated as follows:

$$\theta_i^* \underset{\text{i.i.d.}}{\sim} \mathcal{N}(0, 0.2^2), \quad w_{ij}^* \underset{\text{i.i.d.}}{\sim} \mathcal{N}(0, 0.8^2/\sqrt{n}),$$

where the expression $\mathcal{N}(a, b^2)$ is a Gaussian with the mean a and the variance b^2 . We generate M data from the generative model, and we run the BM learning with the generated data by using Eqs. (7) and (8) with a certain β . We measure the quality of learning by using the log-likelihood function

$$\mathcal{L}(\boldsymbol{\theta}, \mathbf{w} | \mathcal{D}) := \sum_{\mathbf{x}} Q(\mathbf{x}) \ln P_B(\mathbf{x} | \boldsymbol{\theta}, \mathbf{w}), \quad (9)$$

where $\boldsymbol{\theta}$ and $\mathbf{w}$ are solution to the learning. Higher value of the log-likelihood function means better solution, because the maximum point of the log-likelihood function with respect to the parameters corresponds to the minimum point of the KLD in Eq. (2).

In Fig. 1, we show the plot of the log-likelihood function, $\mathcal{L}(\boldsymbol{\theta}, \mathbf{w} | \mathcal{D})$, against β in the case of $M \rightarrow \infty$. In this case, the average values of $\hat{Q}(\mathbf{x} | \beta)$ are $\langle \cdots \rangle_\beta = \sum_{\mathbf{x}} (\cdots) P_B(\mathbf{x} | \beta \boldsymbol{\theta}^*, \beta \mathbf{w}^*)$, and they can be numerically obtained because we know the generative model and our system is not very large. The dot lines in the plot indicates the result when $\beta = 1$ that corresponds to the result obtained by belief propagation without consideration of the temperature. As we expected smaller β gives better result, because by using small β (smaller than one), we can translate the original problem into the easier problem in the perspective of mean-field approximations.

Next, we consider the case of a finite M where it is more important for practical applications. In Fig. 2, we show the plot of the log-likelihood function, $\mathcal{L}(\boldsymbol{\theta}, \mathbf{w} | \mathcal{D})$, against β in the case of $M = 5000$. In this plot, we see the different behavior from Fig. 1. Near $\beta = 0.4$, the log-likelihood functions are higher than one when $\beta = 1$ as we expected, but the results quite degrade for small β . The reason can be thought as follow. Typically, an empirical distribution with a finite data set does not keep details of structure of a generative model. Too small β makes an empirical distribution quite smooth and might break a structure inherited from a generative model.

Fig. 1. Plot of log-likelihood (9) against β in the case of $M \rightarrow \infty$. Each point is averaged over 5000 trials. The dot lines in this plot indicates the result when $\beta = 1$.

Fig. 2. Plot of log-likelihood (9) against β in the case of $M = 5000$. Each point is averaged over 2000 trials. The dot lines in this plot indicates the result when $\beta = 1$.

5. Conclusion

In this paper, by introducing the pseudo-temperature of observed data, the new strategy for mean-field learning algorithms was proposed. The pseudo-temperature can change the temperature of empirical distribution, and therefore, can translate a given learning problem into an easier problem. In particular, this was quite effective in the case where a sufficient number of data was obtained. Smaller β gives better result.

However, in the case where a number of given data was finite, too small β degraded solutions, and as we saw in Fig. 2, there was an optimal value of β . Investigating a method to find the optimal β is needed, and it should be addressed in the future.

Acknowledgments

This work was partly supported by Grants-In-Aid (Nos. 21700247 and 24700220) for Scientific Research from the Ministry of Education, Culture, Sports, Science and Technology, Japan.

REFERENCES

- [1] Ackley, D. H., Hinton, G. E., and Sejnowski, T. J., "A learning algorithm for Boltzmann machines," *Cognitive Science*, **9**: 147–169 (1985).
- [2] Yasuda, M., and Tanaka, K., "Approximate learning algorithm in Boltzmann machines," *Neural Computation*, **21**: 3130–3178 (2009).
- [3] MacKay, D. J., *Information Theory, Inference, and Learning Algorithm*, Cambridge University Press (2003).
- [4] Wainwright, M. J., and Jordan, M. I., "Graphical models, exponential families, and variational inference," *Foundations and Trends in Machine Learning*, **1**: 1–305 (2008).
- [5] Yasuda, M., and Horiguchi, T., "Triangular approximation for information Ising model and its application to Boltzmann machine," *Physica A*, **368**: 83–95 (2006).