Neural Networks 79 (2016) 78–87

Contents lists available at ScienceDirect

Neural Networks

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

Dynamical analysis of contrastive divergence learning: Restricted Boltzmann machines with Gaussian visible units

Ryo Karakida [a,*], Masato Okada [a,b], Shun-ichi Amari [b]

[a] Department of Complexity Science and Engineering, The University of Tokyo, 5-1-5 Kashiwanoha, Kashiwa-shi, Chiba 277-8561, Japan

[b] RIKEN Brain Science Institute, 2-1 Hirosawa, Wako-shi, Saitama 351-0198, Japan

ARTICLE INFO

Article history:
Received 2 October 2015
Received in revised form 23 March 2016
Accepted 31 March 2016
Available online 12 April 2016

Keywords:
Deep learning
Restricted Boltzmann machine
Contrastive divergence
Component analysis
Stability of learning algorithms

ABSTRACT

The restricted Boltzmann machine (RBM) is an essential constituent of deep learning, but it is hard to train by using maximum likelihood (ML) learning, which minimizes the Kullback–Leibler (KL) divergence. Instead, contrastive divergence (CD) learning has been developed as an approximation of ML learning and widely used in practice. To clarify the performance of CD learning, in this paper, we analytically derive the fixed points where ML and CD_n learning rules converge in two types of RBMs: one with Gaussian visible and Gaussian hidden units and the other with Gaussian visible and Bernoulli hidden units. In addition, we analyze the stability of the fixed points. As a result, we find that the stable points of CD_n learning rule coincide with those of ML learning rule in a Gaussian–Gaussian RBM. We also reveal that larger principal components of the input data are extracted at the stable points. Moreover, in a Gaussian–Bernoulli RBM, we find that both ML and CD_n learning can extract independent components at one of stable points. Our analysis demonstrates that the same feature components as those extracted by ML learning are extracted simply by performing CD_1 learning. Expanding this study should elucidate the specific solutions obtained by CD learning in other types of RBMs or in deep networks.

1. Introduction

The restricted Boltzmann machine (RBM) is a bipartite graphical model widely used as an essential constituent of deep neural networks. The visible and hidden units of the RBM are conditionally independent of each other (Hinton, 2002; Smolensky, 1986). Contrastive divergence (CD) learning, an approximate algorithm of maximum likelihood (ML) learning, efficiently uses this conditional independence (Hinton, 2002). If ML learning is used to train an RBM, it requires many iterations of Gibbs sampling at each update step and takes too much computational time. In contrast, CD learning requires only a few iterations of Gibbs sampling, iterated transitions between visible units and hidden units. CD_n learning uses n steps of Gibbs sampling. In particular, CD_1 learning is widely used and can complete the training in a short time. As evidenced empirically, an RBM trained by CD_1 learning achieves solutions close enough to those of an RBM trained by ML learning (Carreira-Perpinan & Hinton, 2005). Stacked RBMs pre-trained by CD_1 learning have performed well in practical applications such as visual image classification (Hinton & Salakhutdinov, 2006) and acoustic modeling (Dahl, Mohamed, & Hinton, 2010).

CD learning performs well enough to achieve success in practice, but there is little theoretical evidence that shows that it performs well. The previous theoretical studies demonstrated that the properties of CD learning are quite different from those of ML learning. For instance, there are certain cases where CD learning does not converge because its gradient does not obey any objective function (Sutskever & Tieleman, 2010). In a simple case of an RBM with continuous 1-hidden and 1-visible units, Williams and Agakov (2002) gained theoretical insights into how the gradients of CD learning are biased in comparison with those of ML learning. In general, the gradient of CD learning is interpreted as a truncated expansion of the log-likelihood gradient (Bengio & Delalleau, 2009). Even if the learning procedure converges to equilibrium solutions, these solutions do not necessarily maximize the likelihood function (Carreira-Perpinan & Hinton, 2005).

The previous studies left a question unanswered: what specific solutions are commonly or differently found by ML and CD learning? Although there are general conditions under which CD learning gives the ML solutions (Akaho & Takabatake, 2008; Yuille, 2005), these conditions are loose, and CD solutions are hard to

* Corresponding author. Tel.: +81 471364085.
E-mail addresses: karakida@mns.k.u-tokyo.ac.jp (R. Karakida), okada@k.u-tokyo.ac.jp (M. Okada), amari@brain.riken.jp (S.-i. Amari).

http://dx.doi.org/10.1016/j.neunet.2016.03.013
0893-6080/

identify. For using CD learning in practice, it is important to identify the specific solutions obtained by CD learning and clarify what features are extracted from input data. A way to identify the solutions obtained by a learning rule is dynamical analysis of equilibrium and its stability (Amari, 1977). By obtaining fixed points of the learning rule and checking their stability by using the perturbation method, the dynamical analysis has revealed what weight matrix can be extracted as a stable fixed point. For instance, it has clarified principal or minor components extracted in linear neural networks (Baldi & Hornik, 1989; Chen & Amari, 2001; Oja, 1989), principal components extracted by ML learning in the probabilistic PCA model (Tipping & Bishop, 1999), and independent components extracted by ICA algorithms (Amari, Chen, & Cichocki, 1997; Hyvärinen, Karhunen, & Oja, 2001). If dynamical analysis can be carried out on CD learning, we can understand the features extracted by CD learning.

In this paper, we used the dynamical analysis to identify the fixed points of ML and CD_n learning rules in two types of RBMs. First, we derived an exact analytical form of the fixed points in a Gaussian–Gaussian RBM whose visible and hidden units are continuous real values (Hinton, 2010; Williams & Agakov, 2002). The ML and CD_n learning rules were explicitly formulated with model parameters. The analytical form demonstrated that ML learning extracts principal components whose eigenvalues are larger than a certain value. In addition, we analyzed the stability of the fixed points by using the perturbation method and revealed that a set of the largest principal components is extracted at stable fixed points. Next, we derived the analytical form for fixed points of CD_n learning rule and found that it coincides with that of ML learning. In addition, their stability also coincides with that of ML learning. We thus concluded that CD_n learning maximizes the likelihood function and extracts the same principal components as ML learning. Moreover, we also apply the same dynamical analysis to a Gaussian–Bernoulli RBM whose hidden units are binary (Hinton & Salakhutdinov, 2006; Lee, Ekanadham, & Ng, 2008). Under certain conditions, we revealed that both ML and CD_n learning in the Gaussian–Bernoulli RBM have one common stable fixed point, where the Gaussian–Bernoulli RBM decomposes mixed input signals to independent source signals.

This paper is a complete version of our unpublished results (Karakida, Okada, & Amari, Unpublished results). Unlike in the previous results, we generalize the analyses for the stability of the fixed points in Gaussian–Gaussian RBM to the case where there is no constraint on the number of hidden units. In addition, we demonstrate the previously omitted proofs of theories on the stable fixed point in Gaussian–Bernoulli RBM. We also discuss the relationship between our theoretical results and the previous studies such as experiments on natural images (Lee et al., 2008; Wang, Melchior, & Wiskott, 2014) and nonlinear PCA (Oja, 1997). Moreover, in both RBMs, we added the learning rules with bias parameters in the Appendix.

The results for CD_n learning are independent of n. Because CD_1 learning can extract the same features as ML learning, CD_1 learning seems to be efficient to train RBMs. Expanding our analysis would help to elucidate features that can be extracted in RBMs with binary visible units or stacked RBMs.

2. Model

2.1. Gaussian–Gaussian RBM

The probability distribution of a Gaussian–Gaussian RBM is defined as follows (Hinton, 2010; Williams & Agakov, 2002):

$$p(\mathbf{h},\mathbf{v}) = \exp\left\{-\sum_{i=1}^{M}\frac{(h_i-c_i)^2}{2s_i^2}-\sum_{j=1}^{N}\frac{(v_j-b_j)^2}{2\sigma_j^2} + \sum_{i,j}W_{ij}\frac{h_i}{s_i}\frac{v_j}{\sigma_j}-\psi\right\}, \tag{1}$$

where $\mathbf{v}$ and $\mathbf{h}$ are random variables representing the states of the visible and hidden units, respectively. Both hidden h_i and visible v_j take continuous values and obey Gaussian distributions characterized by variances s_i^2 ($i = 1,\dots,M$) and σ_j^2 ($j = 1,\dots,N$). Let us denote an $M \times N$ weight matrix by W, biases by c_i and b_j, and a normalization factor by ψ. The joint probability $p(\mathbf{h},\mathbf{v})$ yields the following conditional probability:

$$p(\mathbf{h}|\mathbf{v}) = \mathcal{N}\left(\mathbf{h};\, SW\Sigma^{-1}\mathbf{v}+\mathbf{c},\, S^2\right), \tag{2}$$

$$p(\mathbf{v}|\mathbf{h}) = \mathcal{N}\left(\mathbf{v};\, \Sigma W^T S^{-1}\mathbf{h}+\mathbf{b},\, \Sigma^2\right), \tag{3}$$

where we define a multivariate normal distribution with mean $\boldsymbol{\mu}$ and variance Σ^2 by $\mathcal{N}\left(\mathbf{v};\boldsymbol{\mu},\Sigma^2\right)$. Let us denote the covariance matrix of the hidden units by an $M \times M$ diagonal matrix $S = \mathrm{diag}(s_1, s_2, \dots, s_M)$, whose entries satisfy $S_{ii} = s_i$ and $S_{ij} = 0$ ($i \neq j$). In addition, we denote the covariance matrix of the visible units by an $N \times N$ diagonal matrix $\Sigma = \mathrm{diag}(\sigma_1, \sigma_2, \dots, \sigma_N)$.

When training examples of $\mathbf{v}$ are given from the outside, we need to estimate W, $\mathbf{b}$, and $\mathbf{c}$ such that the marginal distribution $p(\mathbf{v})$ of (1) is as close as $q(\mathbf{v})$, which is the distribution of $\mathbf{v}$ generating training examples. The model variances Σ^2 and S^2 are given and fixed. For mathematical simplicity, we set the mean of input data to $\boldsymbol{\mu} = \int d\mathbf{v} q(\mathbf{v})\mathbf{v} = \mathbf{0}$ and the bias parameters to $\mathbf{b} = \mathbf{c} = \mathbf{0}$ in the following learning rules. We can formulate a general case in the same way as explained in the Appendix. In Section 3, we will also assume that the variances of the visible and hidden units are homogeneous, i.e., $\Sigma = \sigma I_N$ and $S = sI_M$.

ML learning rule. The learning rule of the maximum likelihood (ML) estimate of W is derived by minimizing the Kullback–Leibler (KL) divergence between the input distribution and the model distribution (Hinton, 2002) and is given by:

$$\tau\frac{dW}{dt} = S^{-1}\left\{\langle\mathbf{h}\mathbf{v}^T\rangle_0 - \langle\mathbf{h}\mathbf{v}^T\rangle_\infty\right\}\Sigma^{-1}, \tag{4}$$

where τ is a learning constant. The first term is defined by $\langle\mathbf{h}\mathbf{v}^T\rangle_0 = \int d\mathbf{h}d\mathbf{v}p(\mathbf{h}|\mathbf{v})q(\mathbf{v})\mathbf{h}\mathbf{v}^T$, where $q(\mathbf{v})$ is the input data distribution. In contrast, the second term is defined by $\langle\mathbf{h}\mathbf{v}^T\rangle_\infty = \int d\mathbf{h}d\mathbf{v}p(\mathbf{h},\mathbf{v})\mathbf{h}\mathbf{v}^T$, which is the expectation with respect to the model distribution $p(\mathbf{h},\mathbf{v})$. In practical application of RBMs, the first term of ML learning is calculated by a finite number of training examples and the second term is calculated by samples of the model distribution obtained by Gibbs sampling. In this study, to analyze average behaviors of the learning rules, we neglect fluctuations caused by the training examples and the Gibbs sampler and try to analytically calculate each term by using its definition. In Gaussian–Gaussian RBM, the ML learning rule (4) becomes:

$$\tau\frac{dW}{dt} = W\Sigma^{-1}C\Sigma^{-1} - W(I_N - W^TW)^{-1}. \tag{5}$$

Let us denote the data covariance matrix by $C = \int d\mathbf{v}q(\mathbf{v})\mathbf{v}\mathbf{v}^T - \boldsymbol{\mu}\boldsymbol{\mu}^T$, where I_N is an $N \times N$ identity matrix.

CD_n **learning rule**. In practical application of RBMs, CD_n learning replaces the second term of ML learning with $\langle\mathbf{h}\mathbf{v}^T\rangle_n$, which is calculated by using samples obtained after n times iteration of alternating Gibbs sampling between the visible and hidden layers (Hinton, 2002):

$$\tau\frac{dW}{dt} = S^{-1}\left\{\langle\mathbf{h}\mathbf{v}^T\rangle_0 - \langle\mathbf{h}\mathbf{v}^T\rangle_n\right\}\Sigma^{-1}. \tag{6}$$