Theoretical Computer Science 598 (2015) 102–117

Contents lists available at ScienceDirect

Theoretical Computer Science

www.elsevier.com/locate/tcs

A bound for the convergence rate of parallel tempering for sampling restricted Boltzmann machines

Asja Fischer[a,b,*], Christian Igel[b]

[a] Institut für Neuroinformatik, Ruhr-Universität Bochum, 44780 Bochum, Germany
[b] Department of Computer Science, University of Copenhagen, 2100 Copenhagen, Denmark

ARTICLE INFO

Article history:
Received 28 February 2014
Received in revised form 15 April 2015
Accepted 18 May 2015
Available online 27 May 2015
Communicated by A. Sperduti

Keywords:
Restricted Boltzmann machines
Parallel tempering
Mixing time

ABSTRACT

Sampling from restricted Boltzmann machines (RBMs) is done by Markov chain Monte Carlo (MCMC) methods. The faster the convergence of the Markov chain, the more efficiently can high quality samples be obtained. This is also important for robust training of RBMs, which usually relies on sampling. Parallel tempering (PT), an MCMC method that maintains several replicas of the original chain at higher temperatures, has been successfully applied for RBM training. We present the first analysis of the convergence rate of PT for sampling from binary RBMs. The resulting bound on the rate of convergence of the PT Markov chain shows an exponential dependency on the size of one layer and the absolute values of the RBM parameters. It is minimized by a uniform spacing of the inverse temperatures, which is often used in practice. Similarly as in the derivation of bounds on the approximation error for contrastive divergence learning, our bound on the mixing time implies an upper bound on the error of the gradient approximation when the method is used for RBM training.

1. Introduction

Restricted Boltzmann machines (RBMs) are probabilistic graphical models corresponding to stochastic neural networks [1,2] (see [3] for a recent review). They are applied in many machine learning tasks, notably they serve as building blocks of deep belief networks [4]. Markov chain Monte Carlo (MCMC) methods are used to sample from RBMs, and chains that quickly converge to their stationary distribution are desirable to efficiently get high quality samples. Adaptation of the RBM model parameters typically corresponds to gradient-based likelihood maximization given training data. As computing the exact gradient is usually computationally not tractable, sampling-based methods are employed to approximate the likelihood gradient. It has been shown that inaccurate approximations can deteriorate the learning process (e.g., [5]), and for the most popular learning scheme *contrastive divergence learning* (CD, [2]) the approximation quality has been analyzed [6,7]. The quality of the approximation depends, among other things, on how quickly the Markov chain approaches the stationary distribution, that is, on its mixing rate.

To improve RBM learning, *parallel tempering* (PT, [8]) has successfully been used as a sampling method in RBM training [9–11,3]. Parallel tempering introduces supplementary Gibbs chains that sample from smoothed replicas of the original distribution—with the goal of improving the mixing rate. However, so far there exist no published attempts to analyze the mixing rate of PT applied to RBMs. Based on the work by Woodard et al. [12], we provide the first such analysis. After

* Corresponding author.
E-mail addresses: asja.fischer@rub.de (A. Fischer), igel@diku.dk (C. Igel).

http://dx.doi.org/10.1016/j.tcs.2015.05.019
0304-3975/© 2015 Elsevier B.V. All rights reserved.

introducing the basic concepts, Section 3 states our main result. Section 4 summarizes general theorems required for our proof in Section 5, which is followed by a discussion and our conclusions.

2. Background

In the following, we will give a brief introduction to RBMs and the relation between the mixing rate and the spectral gap of a Markov chain. Afterwards we will describe the parallel tempering algorithm and its application to sampling from RBMs.

2.1. Restricted Boltzmann machines

Restricted Boltzmann machines (RBMs) are probabilistic undirected graphical models (Markov random fields). Their structure is a bipartite graph connecting a set of m visible random variables $\boldsymbol{V} = (V_1, V_2, \dots, V_m)$ modeling observations to n hidden (latent) random variables $\boldsymbol{H} = (H_1, H_2, \dots, H_n)$ capturing dependencies between the visible variables. In binary RBMs the state space of one single variable is given by $\Omega = \{0,1\}$ and accordingly $(\boldsymbol{V}, \boldsymbol{H}) \in \{0,1\}^{m+n}$. The joint probability distribution of $(\boldsymbol{V}, \boldsymbol{H})$ is given by the Gibbs distribution

$$\pi(\boldsymbol{v}, \boldsymbol{h}) = \frac{\exp(-E(\boldsymbol{v}, \boldsymbol{h}))}{Z} \ , \tag{1}$$

with energy

$$E(\boldsymbol{v}, \boldsymbol{h}) = -\sum_{i=1}^{n}\sum_{j=1}^{m} h_i w_{ij} v_j - \sum_{j=1}^{m} b_j v_j - \sum_{i=1}^{n} c_i h_i$$

and real-valued connection weights w_{ij} and bias parameters b_j and c_i for $i \in \{1, \dots, n\}$ and $j \in \{1, \dots, m\}$. The normalization constant Z, also called the partition function, is given by $Z = \sum_{\boldsymbol{v},\boldsymbol{h}} \exp(-E(\boldsymbol{v}, \boldsymbol{h}))$.

Training an RBM means adapting its parameters such that the distribution of $\boldsymbol{V}$ models a distribution underlying some observed data. In practice, this training corresponds to performing stochastic gradient ascent on the log-likelihood of the weight and bias parameters given sample (training) data. The gradient of the log-likelihood given a single training sample $\boldsymbol{v}_{\text{train}}$ is given by

$$\frac{\partial \ln \pi(\boldsymbol{v}_{\text{train}} \,|\, \boldsymbol{\theta})}{\partial \boldsymbol{\theta}} = -\sum_{\boldsymbol{h}} \pi(\boldsymbol{h} \,|\, \boldsymbol{v}_{\text{train}}) \frac{\partial E(\boldsymbol{v}_{\text{train}}, \boldsymbol{h})}{\partial \boldsymbol{\theta}} + \sum_{\boldsymbol{v},\boldsymbol{h}} \pi(\boldsymbol{v}, \boldsymbol{h}) \frac{\partial E(\boldsymbol{v}, \boldsymbol{h})}{\partial \boldsymbol{\theta}} \ , \tag{2}$$

where $\boldsymbol{\theta}$ is the vector collecting all parameters. Since the expectation under the model distribution in the second term on the right hand side can not be computed efficiently (it is exponential in $\min(n,m)$), it is approximated by MCMC methods in RBM training algorithms. Typically, the expected value under the model distribution is approximated by $\sum_{\boldsymbol{h}} \pi(\boldsymbol{h} \,|\, \boldsymbol{v}^{(k)}) \frac{\partial E(\boldsymbol{v}^{(k)}, \boldsymbol{h})}{\partial \boldsymbol{\theta}}$ given a sample $\boldsymbol{v}^{(k)}$ obtained by running a Markov chain for k steps. Alternatively, we can consider $\sum_{\boldsymbol{v}} \pi(\boldsymbol{v} \,|\, \boldsymbol{h}^{(k)}) \frac{\partial E(\boldsymbol{v}, \boldsymbol{h}^{(k)})}{\partial \boldsymbol{\theta}}$ given a sample $\boldsymbol{h}^{(k)}$ to save computation time if $m < n$.

2.2. Mixing rates and the spectral gap

A homogeneous Markov chain on a discrete state space Ω can be described by a transition matrix $P = (p_{\boldsymbol{x},\boldsymbol{y}})_{\boldsymbol{x},\boldsymbol{y} \in \Omega}$, where $p_{\boldsymbol{x},\boldsymbol{y}}$ is the probability to move from $\boldsymbol{x}$ to $\boldsymbol{y}$ in one step of the Markov chain. We also refer to this probability as $P(\boldsymbol{x}, \boldsymbol{y})$, and accordingly $P^k(\boldsymbol{x}, \boldsymbol{y})$ gives the probability to move from $\boldsymbol{x}$ to $\boldsymbol{y}$ in k steps of the chain.

Markov chain Monte Carlo methods make use of the fact that an ergodic Markov chain on Ω with transition matrix P and equilibrium distribution π satisfies $P^k(\boldsymbol{x}, \boldsymbol{y}) \to \pi(\boldsymbol{y})$ as $k \to \infty$ for all $\boldsymbol{x}, \boldsymbol{y} \in \Omega$. It is important to study the convergence rate of the Markov chain in order to find out how large k has to be to ensure that $P^k(\boldsymbol{x}, \boldsymbol{y})$ is suitably close to $\pi(\boldsymbol{y})$. One way to measure the closeness to stationarity is the total variation distance $\|P^k(\boldsymbol{x}, \cdot) - \pi\| = \frac{1}{2}\sum_{\boldsymbol{y}} |P^k(\boldsymbol{x}, \boldsymbol{y}) - \pi(\boldsymbol{y})|$ for an arbitrary starting state $\boldsymbol{x}$. Reversibility of the chain implies that the eigenvalues of P are real-valued, and we sort them by value $1 = \lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_r \geq -1$. The value $\text{Gap}(P) = 1 - \lambda_2$ is called the *spectral gap* of P, and the eigenvalue with the second largest absolute value $\lambda_{\text{SLEM}} = \max\{\lambda_2, |\lambda_r|\}$ is often referred to as *second largest eigenvalue modulus* (SLEM).

Many bounds on convergence rates are based on the SLEM. Diaconis and Saloff-Coste [13], for example, prove

$$\|P^k(\boldsymbol{x}, \cdot) - \pi\| \leq \frac{1}{2\sqrt{\pi(\boldsymbol{x})}} {\lambda_{\text{SLEM}}}^k \ . \tag{3}$$

If P is positive definite all eigenvalues are non-negative. In this case $\lambda_{\text{SLEM}} = \lambda_2$ and we can replace λ_{SLEM} in (3) (and similar bounds) by $1 - \text{Gap}(P)$. We can make an arbitrary transition matrix Q positive definite by skipping the move each