Künstl Intell (2015) 29:441–444
DOI 10.1007/s13218-015-0371-2

DOCTORAL AND POSTDOCTORAL DISSERTATIONS

Training Restricted Boltzmann Machines

Asja Fischer[1]

Published online: 12 May 2015
© Springer-Verlag Berlin Heidelberg 2015

Abstract Restricted Boltzmann Machines (RBMs), two-layered probabilistic graphical models that can also be interpreted as feed forward neural networks, enjoy much popularity for pattern analysis and generation. Training RBMs however is challenging. It is based on likelihood maximization, but the likelihood and its gradient are computationally intractable. Therefore, training algorithms such as Contrastive Divergence (CD) and learning based on Parallel Tempering (PT) rely on Markov chain Monte Carlo methods to approximate the gradient. The presented thesis contributes to understanding RBM training methods by presenting an empirical and theoretical analysis of the bias of the CD approximation and a bound on the mixing rate of PT. Furthermore, the thesis improves RBM training by proposing a new transition operator leading to faster mixing Markov chains, by investigating a different parameterization of the RBM model class referred to as centered RBMs, and by exploring estimation techniques from statistical physics to approximate the likelihood. Finally, an analysis of the representational power of deep belief networks with real-valued visible variables is given.

Keywords Restricted Boltzmann machines · Contrastive divergence · Parallel tempering · Mixing rate · Likelihood estimation · Deep belief networks

✉ Asja Fischer
asja.fischer@gmail.com

[1] Department of Computer Science and Operations Research, University of Montreal, Montreal, QC, Canada

1 Introduction

In the last years, models extending or borrowing concepts from Restricted Boltzmann Machines (RBMs, [19]) have been applied with great success as generative models, classifiers, feature extractors or as building blocks of deep architectures such as Deep Belief Networks (DBNs, [12]). Restricted Boltzmann machines are undirected graphical models that can also be interpreted as two-layered stochastic neural networks, see Fig. 1. As an undirected graphical model, an RBM represents a probability distribution that can be used in an unsupervised learning problem to model some distribution over some input space. Given a set of samples as training data, learning corresponds to adjusting the model parameters of the RBM such that the represented probability distribution fits the training data as well as possible. That is, the model parameters are adjusted such that the likelihood of the parameters given the training data is maximized. Maximum likelihood learning is in general challenging for undirected graphical models because maximum likelihood parameters cannot be found analytically and the log-likelihood gradient needed for gradient-based optimization is not tractable, since it involves averages over a number of terms exponential in the size of the model. Therefore, common training algorithms conduct Markov Chain Monte Carlo (MCMC) methods to approximate the gradient. An overview about RBMs and their most common training techniques can be found in our recent review article [10].

The presented thesis investigates existing training algorithms for RBMs empirically and theoretically, contributes to improving training, and analyses the representational power of DBNs.

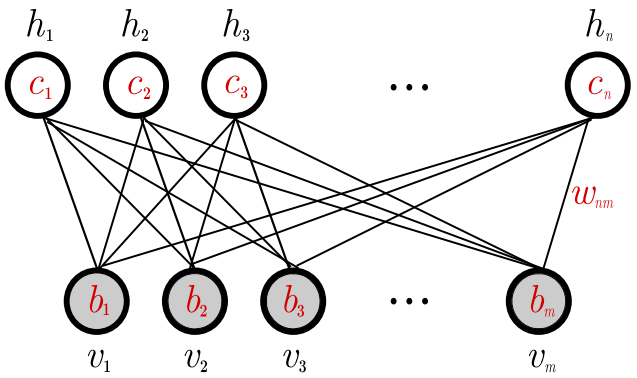

Fig. 1 The undirected, bipartite graph of an RBM with m visible variables $\mathbf{v} = (v_1, \ldots, v_m)$ to model observable data and n hidden variables $\mathbf{h} = (h_1, \ldots, h_n)$ to capture dependencies between the visible variables. The set of parameters of an RBM is given by the bias values b_j and c_i associated to the jth visible and ith hidden unit, respectively, and a weight parameter w_{ij} associated to the connection of both ($i = 1, \ldots, n; j = 1, \ldots, m$). The modeled probability distribution is given by $p(\mathbf{v}, \mathbf{h}) = \frac{1}{Z} \mathrm{e}^{-E(\mathbf{v},\mathbf{h})}$ with normalization constant $Z = \sum_{\mathbf{v},\mathbf{h}} \mathrm{e}^{-E(\mathbf{v},\mathbf{h})}$ and energy $E(\mathbf{v}, \mathbf{h}) = -\sum_{i,j} v_j w_{ij} h_i - \sum_j v_j b_j - \sum_i h_i c_i$

2 An Analysis of RBM Training Algorithms

Compared to 1986, when RBMs were introduced [19], RBMs can now be applied to more interesting problems. This is due to the increase in computational power and the development of new learning strategies which started around 2002 [11]. Obtaining unbiased estimates of the log-likelihood gradient by MCMC methods typically requires many sampling steps and thus is computationally too demanding. Hinton [11] however showed that the biased estimates obtained by initializing a Gibbs chain with a training example and running it for just a few steps can be sufficient for RBM training. The resulting approximation of the log-likelihood gradient is referred to as Contrastive Divergence (CD).

2.1 An Analysis of the Bias of CD Learning

The bias of CD depends not only on the number of sampling steps, but also on the mixing rate of the Gibbs chain (i.e. the speed of the convergence of the Markov chain to the model distribution) and mixing slows down with increasing magnitude of model parameters [1, 11]. The magnitude of the RBM parameters increases during training, and so does the CD-bias. We show in a detailed empirical analysis [6] that this can lead to a distortion of the learning process: after some learning iterations, the likelihood can start to diverge in the sense that the model systematically gets worse. This divergence depends on the number of Gibbs sampling steps used to gain a sample, the number of hidden variables of the RBM, and on the usage of weight-decay or an adaptive learning rate.

To deepen the theoretical understanding of the CD-bias we derive a new upper bound that reflects the dependencies of the bias on the number of sampling steps and the magnitude of the RBM parameters [9]. It is further affected by the distance in variation between the modeled distribution and the starting distribution of the Gibbs chain.

It was previously reported [1] that despite of the bias, the signs of most components of the CD update are equal to the corresponding signs of the log-likelihood gradient. Therefore, we investigate [7] training based on resilient backpropagation [17] as an optimization technique depending only on the signs. However, in our experiments this did not prevent the divergence caused by the approximation bias.

2.2 An Analysis of the Mixing Rate of PT Sampling

One of the most promising sampling techniques used for RBM training so far is Parallel Tempering (PT), which maintains several Gibbs chains in parallel and is designed to produce a faster mixing Markov chain. This can prevent the likelihood from diverging [5]. We analyze the convergence rate of PT for sampling from RBMs by deriving a lower bound on the spectral gap, which shows an exponential dependency on the size of the smallest layer and the sum of the absolute values of the RBM parameters ([8], Chapter 6).

3 Improvements for RBM Training

We contribute in different ways to improving RBM training as it is described in the following.

3.1 A New Transition Operator for Sampling in RBMs

Since the bias of the gradient approximation and the performance of RBM learning algorithms heavily depend on the mixing rate of the Markov chain employed for drawing samples, it is of high interest to use sampling techniques with a fast convergence rate. Likewise, when using RBMs as generative models, one is interested in sampling techniques leading to a fast convergence of the Markov chain to the stationary distribution, since this is the distribution one wishes to draw samples from. Contrastive divergence learning as well as PT rely on Gibbs sampling, which is a Metropolis-type transition operator. We propose [3] to replace Gibbs sampling by another transition operator from this family. This operator—we refer to as flip-the-state transition operator—maximizes the probability of state changes and can replace Gibbs sampling in RBM learning

algorithms without producing computational overhead. It is shown analytically that the operator induces an irreducible, aperiodic, and hence properly converging Markov chain, also for the typically used periodic update schemes. Furthermore, we demonstrate empirically that using the flip-the-state operator can lead to faster mixing and in turn to more accurate learning.

3.2 An Analysis of Centered RBMs

An undesired property of training RBMs based on the log-likelihood gradient is that the learning procedure is not invariant to the data representation. For example training an RBM on the MNIST data set of handwritten digits (white digits on black background) leads to a better model than training it on the data set generated by flipping each bit (black digits on white background). So far, two ways of achieving invariance to such changes of the data representation have been described. First, the enhanced gradient was designed as an alternative update direction that can replace the gradient and leads to the desired invariance [4]. Second, subtracting the data mean from the visible variables was reported to lead to similar learning results on flipped and unflipped data sets [20]. Removing the data mean of all variables is generally known as the centering trick. It was recently applied to deep Boltzmann machines, where it leads to better conditioned optimization problems and improves some aspects of model performance [16]. We analyse centered binary RBMs (see [8], Chapter 8, and [15]), where we allow to subtract arbitrary offset values from visible and hidden variables. It is shown analytically that centering can be reformulated as a different update rule for training normal binary RBMs. The corresponding update direction becomes equivalent to the enhanced gradient for a certain choice of offsets and yields the desired invariance to the data representation for a broad set of offset values. Numerical simulations show that centering leads to better models in terms of the log-likelihood, and to an update direction closer to the natural gradient. Optimal model performance is achieved when subtracting mean values from both visible and hidden variables. It is further shown that the enhanced gradient suffers from divergence more often than other centering variants, which can be prevented by using an exponentially moving average for the offset estimation.

3.3 New Estimators for the Normalization Constant of RBMs

Assessing model performance is difficult since the likelihood of RBMs is not tractable due to a normalization constant which depends exponentially on the size of the RBM. It can be reliably estimated using Annealed Importance Sampling (AIS) [18], which however needs too much computation time to efficiently monitor the training process. Therefore, we explore ([8], Chapter 9) alternative techniques from statistical physics for estimating the normalization constant, including Bennetts Acceptance Ratio (BAR) [2]. A unifying framework for deriving these methods as well as AIS is given and an empirical analysis shows that BAR gives superior results and outperforms AIS. Moreover, BAR allows to reuse the samples generated for learning when employed to track the partition function during PT based training.

4 An Analysis of the Representational Power of DBNs with Real-valued Visible Variables

Deep belief networks are built by stacking RBMs, and known to be able to approximate any distribution over fixed-length binary vectors [14]. However, DBNs are often used for modeling distributions of real valued variables. Therefore, we analyze [13] the approximation properties of DBNs with two layers of binary hidden units, and visible units with conditional distributions from the exponential family. It is shown that they can, under mild assumptions, model any additive mixture of distributions from the exponential family with independent variables. An arbitrarily good approximation in terms of Kullback–Leibler divergence of an m dimensional mixture distribution with n components can be achieved by a DBN with a layer of m visible variables and n and $n + 1$ hidden variables in the first and second hidden layer, respectively. Furthermore, we show that relevant infinite mixtures can be approximated arbitrarily well by a DBN with a finite number of variables. This includes the important special case of infinite additive mixtures of Gaussian distributions, which in turn can model any strictly positive density over a compact domain with arbitrary high accuracy [21]. Therefore, DBNs with Gaussian visible and binary hidden variables can also model any strictly positive density over a compact domain arbitrarily well.

References

1. Bengio Y, Delalleau O (2009) Justifying and generalizing contrastive divergence. Neural Comput 21(6):1601–1621
2. Bennett CH (1976) Efficient estimation of free energy differences from Monte Carlo data. J Comput Phys 22(2):245–268
3. Brügge K, Fischer A, Igel C (2013) The flip-the-state transition operator for restricted Boltzmann machines. Mach Learn 13:53–69
4. Cho K, Raiko T, Ilin A (2011) Enhanced gradient and adaptive learning rate for training restricted Boltzmann machines. In: Getoor L, Scheffer T (eds) Proceedings of 28th international conference on machine learning (ICML). ACM, pp 105–112

5. Desjardins G, Courville A, Bengio Y, Vincent P, Dellaleau O (2010) Tempered Markov chain Monte Carlo for training of restricted Boltzmann machines. In: Teh YW, Titterington M (eds) Proceedings of the 13th international workshop on artificial intelligence and statistics (AISTATS), JMLR W&CP, vol 9, pp 145–152
6. Fischer A, Igel C (2010) Empirical analysis of the divergence of Gibbs sampling based learning algorithms for restricted Boltzmann machines. In: Diamantaras K, Duch W, Iliadis LS (eds) International conference on artificial neural networks (ICANN), LNCS, vol 6354. Springer, pp 208–217
7. Fischer A, Igel C (2011) Training RBMs based on the signs of the CD approximation of the log-likelihood derivatives. In: Verleysen M (ed) 19th European symposium on artificial neural networks (ESANN). d-side publications, Belgium, pp 495–500
8. Fischer A (2014) Training restricted Boltzmann machines. Ph.D. thesis, Univesity of Copenhagen, Denmark
9. Fischer A, Igel C (2011) Bounding the bias of contrastive divergence learning. Neural Comput 23:664–673
10. Fischer A, Igel C (2014) Training restricted Boltzmann machines: an introduction. Pattern Recogn 47:25–39
11. Hinton GE (2002) Training products of experts by minimizing contrastive divergence. Neural Comput 14:1771–1800
12. Hinton GE, Salakhutdinov RR (2006) Reducing the dimensionality of data with neural networks. Science 313(5786):504–507
13. Krause O, Fischer A, Glasmachers T, Igel C (2013) Approximation properties of DBNs with binary hidden units and real-valued visible units. In: Dasgupta S, McAllester D (eds) Proceedings of the 30th international conference on machine learning (ICML), JMLR W&CP, vol 28, pp 419–426
14. Le Roux N, Bengio Y (2010) Deep belief networks are compact universal approximators. Neural Comput 22(8):2192–2207
15. Melchior J, Fischer A, Wang N, Wiskott L (2013) How to center binary restricted Boltzmann machines. Techn Rep. arXiv preprint arXiv:1311.1354
16. Montavon G, Müller K (2012) Deep Boltzmann machines and the centering trick. Lect Notes Comput Sci 7700:621–637
17. Riedmiller M (1994) Advanced supervised learning in multi-layer perceptrons—from backpropagation to adaptive learning algorithms. Comput Stand Interfaces 16(5):265–278
18. Salakhutdinov R, Murray I (2008) On the quantitative analysis of deep belief networks. In: Cohen WW, McCallum A, Roweis ST (eds) Proceedings of the international conference on machine learning (ICML), vol 25. ACM
19. Smolensky P: Information processing in dynamical systems: foundations of harmony theory. In: Rumelhart DE, McClelland JL (eds) Parallel distributed processing: explorations in the microstructure of cognition, Foundations, vol. 1. MIT Press, pp 194–281 (1986)
20. Tang Y, Sutskever I (2011) Data normalization in the learning of restricted Boltzmann machines. In: Technical report, Department of computer science. University of Toronto, Toronto
21. Zeevi AJ, Meir R (1997) Density estimation through convex combinations of densities: approximation and estimation bounds. Neural Netw 10(1):99–109

Asja Fischer received her B.Sc. degree in Biology from the Ruhr-University Bochum, Germany, in 2005. After one year of postgraduate studies in Bioinformatics at the Universidade de Lisboa, Portugal, she studied Cognitive Science and Mathematics at the University of Osnabrück and the Ruhr-University Bochum, Germany, and received her M.Sc. degree in Cognitive Science in 2009. Between 2010 and 2015, Asja was employed both at the Institute for Neural Computation, Ruhr-University Bochum, and the Department of Computer Science, University of Copenhagen, working on her PhD in Computer Science, which she defended in Copenhagen in 2014. Asja is currently working as a post-doctoral researcher in the Machine Learning Laboratory at the Université de Montréal.