Iterative Project Quasi-Newton Algorithm for Training RBM

Shuai Mi
Tianjin University
Tianjin, China
watermelon0519@163.com

Xiaozhao Zhao
Tianjin University
Tianjin, China
0.25eye@gmail.com

Yuexian Hou
Tianjin University
Tianjin, China
yxhou@tju.edu.cn

Peng Zhang
Tianjin University
Tianjin, China
pzhang@tju.edu.cn

Wenjie Li
The Hong Kong Polytechnic University
Hong Kong
cswjli@comp.polyu.edu.hk

Dawei Song
Tianjin University
Tianjin, China
dwsong@tju.edu.cn

Abstract

The restricted Boltzmann machine (RBM) has been used as building blocks for many successful deep learning models, e.g., deep belief networks (DBN) and deep Boltzmann machine (DBM) etc. The training of RBM can be extremely slow in pathological regions. The second order optimization methods, such as quasi-Newton methods, were proposed to deal with this problem. However, the non-convexity results in many obstructions for training RBM, including the infeasibility of applying second order optimization methods. In order to overcome this obstruction, we introduce an em-like iterative project quasi-Newton (IPQN) algorithm. Specifically, we iteratively perform the sampling procedure where it is not necessary to update parameters, and the sub-training procedure that is convex. In sub-training procedures, we apply quasi-Newton methods to deal with the pathological problem. We further show that Newton's method turns out to be a good approximation of the natural gradient (NG) method in RBM training. We evaluate IPQN in a series of density estimation experiments on the artificial dataset and the MNIST digit dataset. Experimental results indicate that IPQN achieves an improved convergent performance over the traditional CD method.

Introduction

A RBM is a non-convex graphical model which contains a layer of visible units and a layer of hidden units. The hidden units give rise to the unidentifiability which in turn directly results in the non-convexity of RBM.

Typically, RBM is trained using Contrastive Divergence (CD) learning, which can be considered as a stochastic gradient descent (SGD) method. This optimization method is concise. However, it is a well known fact in the optimization community that gradient descent is unsuitable for optimizing objectives that exhibit pathological curvature (Martens 2010).

Martens (2010) proposed a second order Hessian-Free optimization method, which can effectively deal with pathological curvature, to train deep auto-encoders and neural networks. However, the non-convexity of RBM could result in many obstructions, including the infeasibility of applying 2-nd order optimization methods.

We are committed to deal with this obstruction. We introduce a novel iterative project quasi-Newton (IPQN) method for training restricted Boltzmann machine. With the iterative project method (Zhao et al. 2013), our algorithm transforms the overall training procedure into an iterative procedure, i.e., iteratively perform the sampling procedure where it is not necessary to update parameters, and the sub-training procedure that is convex. We apply the curvature information based second order quasi-Newton methods to the sub-training procedures to deal with the pathological problem. In addition, our algorithm can deal with another training obstruction caused by the non-convexity, i.e., the saddle points whose neighborhood is similar to optima.

Furthermore, we investigate the nature of the Hessian in the RBM training, and based on this nature, two main advantages of our algorithm can be realized. Firstly, in IPQN, the Hessian of the objective function is relatively well conditioned. Secondly, Newtons method turns out to be a good approximation of the natural gradient method (Amari 1998) in our algorithm. Besides, there are still more advantages of IPQN, i.e., Newtons method can achieve multi-times quadratically convergent stage by applying IP method, and Newtons method is independent of affine changes. We show some experiment results in this paper.

Experimental Study

We experimentally investigate the IPQN algorithm in density estimation tasks for restricted Boltzmann machines. In our experiments, we used the artificial data and the MNIST digit data. We compare iterative project quasi-Newton method (IPQN) with Contrastive Divergence (CD-1), iterative project (IP) and Hessian-Free (HF), and show that IPQN achieves better convergence in our experiments.

Experiments on the Artificial Data

The artificial binary dataset: we first randomly select the target distribution $q(x)$, which is randomly chosen from the open probability simplex over the n random variables using the Dirichlet prior (with parameters alpha = 0.5). Then, the dataset with N samples are generated from $q(x)$. In order

to accurately and effectively evaluate the contrastive divergence, the artificial dataset is set to be 10-dimensional. For experiments on the artificial data, the baseline is CD-1, and other three methods are compared:

- IPQN: we change the overall training procedure into an iterative procedure using the iterative project method and apply quasi-Newton methods in sub-training procedures;
- IP: we change the overall training procedure into an iterative procedure using the iterative project method and apply CD-1 in sub-training procedures;
- HF: we apply Hessian-Free method without changing the overall training procedure into an iterative procedure.

K-L divergence is used to evaluate the goodness-of-fit of the RBM trained by these four algorithms. For all experiments, we run 20 randomly generated distributions and report the average K-L divergence. The sample size is from 50 to 500. Note that our experiments focus on the case that variable number is relatively small ($n = 10$) in order to analytically evaluate the contrastive divergence.

The average K-L divergences between RBMs and the underlying real distributions are shown in Fig. 1. Among these four algorithms, IPQN achieves better result. The performance of HF is even worse than CD-1 because of the non-convexity of RBM. Intuitively, along with the increasing of the sample size the estimated probability distribution should be closer to the underlying distribution. However, we can see that the average performance of CD-1 fluctuates wildly. This observation illustrated that CD-1 is more likely to achieve a fluctuant convergence under some special distributions even when the sample is relatively sufficient. On the other hand, the performances of two algorithms involved with IP method (i.e. IP and IPQN) improve reasonably along with the increasing of the sample size. Thus, we can experimentally find that the em based IP method is more likely to achieve a stable convergence. Two reasons of the stable convergence are introduced here: first, gradient based methods, such as CD-1, are proved to be an approximation of IP method (Zhao et al. 2013). Second, the monotonicity of IP method can be theoretically guaranteed.

The average performance of IPQN is better than IP especially under relatively small sample size (i.e. N = 50,100) (see Fig. 1). To explain this performance difference, we have theoretically explained at our homepage that quasi-Newton methods can avoid hang-up in pathological curvature regions and achieve a Fisher-efficient optimum.

Experiments on the MNIST Digit Data

In this subsection, we experimentally investigate the performance of IPQN on real-world datasets in the context of density estimation. We use log-likelihood to evaluate the goodness-of-fit of the training since it is not easy to compute the contrastive divergence due to the high dimensionality. We used a RBM that contains 10 hidden units to learn the distribution density over the MNIST digit data. In our experiments, the training set consists of 1000 cases and the test set consists of 1000 cases. Four algorithms are compared: CD-1, HF, IP and IPQN.

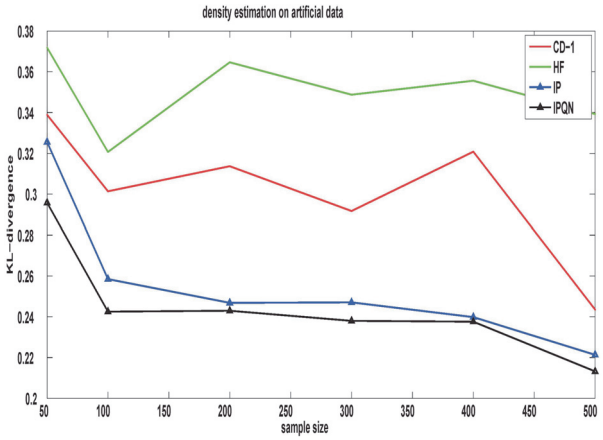

Figure 1: Density estimation for RBM

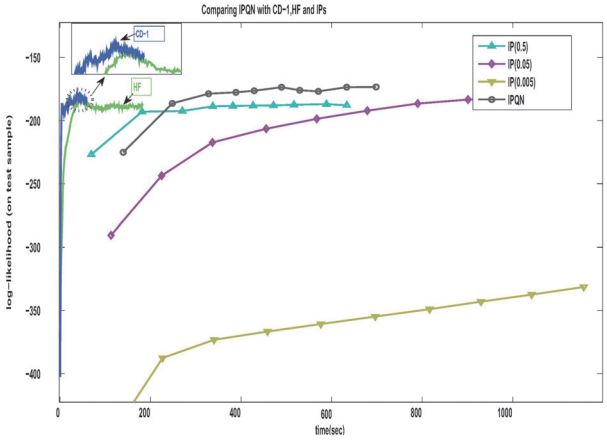

Figure 2: Trajectories of MNIST experiments

CD-1 and HF take less time to achieve convergence, but from Fig. 2 we can see that IPQN achieves better convergent value than CD-1 and HF. This observations agree with the experiments on the artificial data. We will use distributed system to speed up the IPQN training in future works.

Acknowledgments

This work is funded in part by the Chinese 863 Program (grant No. 2015AA015403).

References

Amari, S.-I. 1998. Natural gradient works efficiently in learning. *Neural computation* 10(2):251–276.

Martens, J. 2010. Deep learning via hessian-free optimization. In *Proceedings of the 27th International Conference on Machine Learning (ICML-10)*, 735–742.

Zhao, X.; Hou, Y.; Yu, Q.; Song, D.; and Li, W. 2013. Understanding boltzmann machine and deep learning via a confident information first principle. *arXiv preprint arXiv:1302.3931*.