ESANN 2016 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 27-29 April 2016, i6doc.com publ., ISBN 978-287587027-8.
Available from http://www.i6doc.com/en/.

Stochastic Gradient Estimate Variance in Contrastive Divergence and Persistent Contrastive Divergence

Mathias Berglund

Aalto University - Department of Information and Computer Science
Espoo - Finland

Abstract. Contrastive Divergence (CD) and Persistent Contrastive Divergence (PCD) are popular methods for training Restricted Boltzmann Machines. However, both methods use an approximate method for sampling from the model distribution. As a side effect, these approximations yield significantly different biases and variances for stochastic gradient estimates of individual data points. It is well known that CD yields a biased gradient estimate. In this paper we however show empirically that CD has a lower stochastic gradient estimate variance than unbiased sampling, while the mean of subsequent PCD estimates has a higher variance than independent sampling. The results give one explanation to the finding that CD can be used with smaller minibatches or higher learning rates than PCD.

1 Introduction

Popular methods to train Restricted Boltzmann Machines [1] include Contrastive Divergence [2, 3] and Persistent Contrastive Divergence[1] [4, 5]. Although some theoretical research has focused on the properties of these two methods [6, 7, 5], both methods are still used in similar situations, where the choice is often based on intuition or heuristics.

One known feature of Contrastive Divergence (CD) learning is that it yields a biased estimate of the gradient [6, 7]. On the other hand, it is known to be fast for reaching good results [7, 5]. In addition to the computationally light sampling procedure in CD, it is claimed to benefit from a low variance of the gradient estimates [2, 7]. However, the current authors are not aware of any rigorous research on whether this claim holds true, and what the magnitude of the effect is[2].

On the other hand, Persistent Contrastive Divergence (PCD) has empirically been shown to require a lower learning rate and longer training than CD[3] [5]. In that work, the authors propose that the low learning rate is required since the model weights are updated while the Markov chain runs, which means that in order to sample from a distribution close to the stationary distribution the weight cannot change too rapidly. However, for similar reasons that CD updates are assumed to have low variance, subsequent PCD updates are likely to be correlated leading to a possibly undesirable "momentum" in the updates. This behavior would effectively increase the variance of the mean of subsequent updates, requiring either larger minibatches or smaller learning rates.

In this paper we explore the variances of CD, PCD and unbiased and independent stochastic gradient estimates. We hope to shed light on the observed fast speed of CD learning, and on the required low learning rate for PCD compared to CD.

[1] Persistent Contrastive Divergence is also known as Stochastic Maximum Likelihood

[2] The topic has been covered in e.g. [8], although for a Boltzmann machine with only one visible and hidden neuron.

[3] There are however tricks to be able to increase the learning rate of PCD, see e.g. [9]

ESANN 2016 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 27-29 April 2016, i6doc.com publ., ISBN 978-287587027-8. Available from http://www.i6doc.com/en/.

2 Contrastive Divergence and Persistent Contrastive Divergence

A restricted Boltzmann machine (RBM) is a Boltzmann machine where each visible neuron x_i is connected to all hidden neurons h_j and each hidden neuron to all visible neurons, but there are no edges between hidden and hidden or between visible and visible neurons. An RBM defines an energy of each state $(\mathbf{x}, \mathbf{h})$ by

$$-E(\mathbf{x}, \mathbf{h} \mid \boldsymbol{\theta}) = \mathbf{b}^\top \mathbf{x} + \mathbf{c}^\top \mathbf{h} + \mathbf{x}^\top \mathbf{W} \mathbf{h}$$

and assigns the following probability to the state via the Boltzmann distribution:

$$p(\mathbf{x}, \mathbf{h} \mid \boldsymbol{\theta}) = \frac{1}{Z(\boldsymbol{\theta})} \exp\left\{-E\left(\mathbf{x}, \mathbf{h} \mid \boldsymbol{\theta}\right)\right\}$$

where $\boldsymbol{\theta} = \{\mathbf{b}, \mathbf{c}, \mathbf{W}\}$ is a set of parameters and $Z(\boldsymbol{\theta})$ normalizes the probabilities to sum up to one. The log likelihood of one training data point is hence

$$\phi = \log P(\mathbf{x}) = \log\left(\sum_{\mathbf{h}} \exp\left\{-E\left(\mathbf{x}, \mathbf{h} \mid \boldsymbol{\theta}\right)\right\}\right) - \log Z(\boldsymbol{\theta}) = \phi^+ - \phi^-$$

Sampling the *positive phase*, i.e. the data-drive phase of the gradient of the log likelihood $\frac{\partial \phi^+}{\partial \mathbf{W}}$ is easy, but sampling the *negative phase*, i.e. the model-driven phase $\frac{\partial \phi^-}{\partial \mathbf{W}}$ is intractable.

A popular method to solve sampling of the negative phase is Contrastive Divergence (CD). In CD, the negative particle is sampled only approximately by running a Markov Chain only k steps (often only one step) from the positive particle [2]. Another method, called Persistent Contrastive Divergence (PCD) solves the sampling with a related method, only that the negative particle is not sampled from the positive particle, but rather from the negative particle from the last data point [5].

3 Experiments

In order to examine the variance of CD and PCD gradient estimates, we use an empirical approach. We train an RBM and evaluate the variance of gradient estimates from different sampling strategies at different stages of the training process. The sampling strategies are CD-k with k ranging from 1 to 10, PCD, and CD-1000 that is assumed to correspond to an almost independent and unbiased stochastic gradient. In addition, we test CD-k with independent samples (I-CD), where the negative particle is sampled from a random training example. The variance of I-CD separates the effect of the negative particle being close to the data distribution in general, and the effect of the negative particle being close to the positive particle in question.

We use three different data sets. The first is a reduced size MNIST [10] set with 14×14 pixel images of the first 1 000 training set data points of each digit, totaling 10 000 data points. The second data set are the center 14×14 pixels of the first 10 000 CIFAR [11] images converted into gray scale. The third are the Caltech 101 Silhouettes [12], with 8 641 16×16 pixel black and white images. We binarize the grayscale images by sampling the visible units with activation probabilities equal to the pixel intensity.

ESANN 2016 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 27-29 April 2016, i6doc.com publ., ISBN 978-287587027-8.
Available from http://www.i6doc.com/en/.

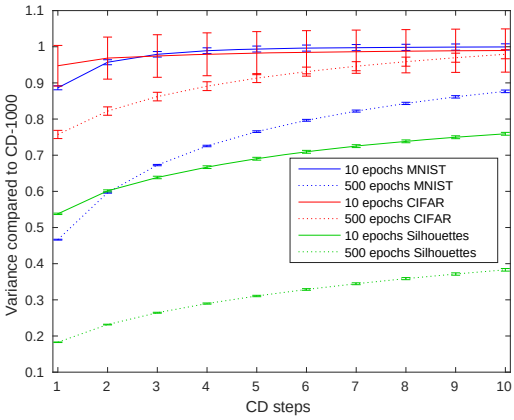

Fig. 1: The CD-k vs. CD-1000 (assumed to be unbiased and independent) ratio of the variance of the gradient estimate for different values of k after 10 and 500 epochs of training. Error bars indicate standard deviation between iterations.

We set the number of hidden neurons equal to the number of visible neurons. The biases are initialized to zero, while the weights are initially sampled from a zero-mean normal distribution with standard deviation $1/\sqrt{n_v + n_h}$ where n_v and n_h are the number of visible and hidden neurons, respectively. We train the model with CD-1, and evaluate the variance of the gradient estimates after 10, and 500 epochs. We use Adaptive learning rate [13] with an initial learning rate of 0.01. We do not use weight decay.

In all of the gradient estimates, the final sampling step for the probabilities of the hidden unit activations is omitted. The gradient estimate is therefore based on sampled binary visible unit activations, but continuous hidden unit activation probabilities conditional on these visible unit activations. This process is called Rao-Blackwellisation [9], and is often used in practice. The variance is calculated on individual gradient estimates based on only one positive and negative particle each. In practice, the gradient is usually estimated by averaging over a mini-batch of N independent samples, which diminishes the variance N-fold. We ignore the bias gradient estimates.

When analyzing subsequent PCD gradient estimates, the negative particles of the first estimate are sampled 1 000 steps from a random training example. Subsequent m estimates are then averaged, where the positive particle is randomly sampled from the data for each step while the negative particle is sampled from the previous negative particle. No learning occurs between the subsequent estimates. We can therefore disentangle the effects of weight updates from the effect of correlation between subsequent estimates.

We iterate all results for 10 different random initializations of the weights, and evaluate the variance by sampling gradient estimates of individual training examples 10 times for each training example in the data set. The variance is calculated for each weight matrix element separately, and the variances of the individual weights are then

ESANN 2016 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 27-29 April 2016, i6doc.com publ., ISBN 978-287587027-8.
Available from http://www.i6doc.com/en/.

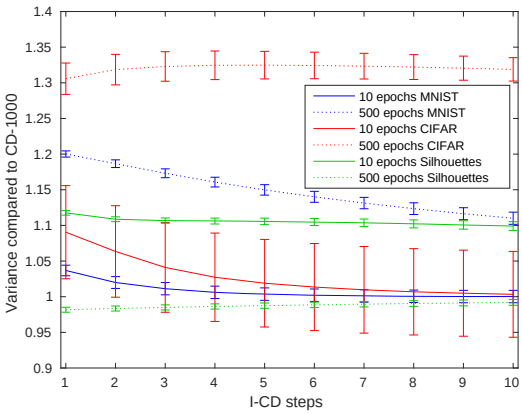

Fig. 2: The I-CD-k vs. CD-1000 (assumed to be unbiased and independent) ratio of the variance of the gradient estimate for different values of k as a ratio compared to after 10 and 500 epochs of training. Error bars indicate standard deviation between iterations.

averaged.

4 Results

As we can see from Figure 1, the variance of Contrastive Divergence is indeed smaller than for unbiased sampling of the negative particle. We also see that the variance of CD estimates quickly increases with the number of CD steps. However, this effect is significant only in later stages of training. This phenomenon is expected, as the model is expected not to mix as well in later stages of training compared to when the weights are close to the small initial random weights.

If we sample the negative particle from a different training example than the positive particle (I-CD), in Figure 2 we see that the variance is similar or even larger compared to the variance with unbiased sampling. Although it is trivial that the variance of the I-CD estimates is higher than for CD, the interesting result is that I-CD loses all of the variance advantage against unbiased sampling. The result supports the hypothesis that the low variance of CD precisely stems from the fact that the negative particle is sampled from the positive particle, and not from that the negative particle is sampled only a limited number of steps from a random training example.

For subsequent PCD updates, we see in Figure 3 that the variance indeed is considerably higher than for independent sampling. E.g. assuming a small enough learning rate, the variance of the mean of six subsequent gradient estimates is roughly three times as high as if the negative particles were sampled independently for the Silhouettes dataset after 500 epochs of training. Again, as expected this effect is stronger the later during training the evaluation is done.

When looking at the magnitude of the variance difference, we see that for CD-1,

ESANN 2016 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 27-29 April 2016, i6doc.com publ., ISBN 978-287587027-8.
Available from http://www.i6doc.com/en/.

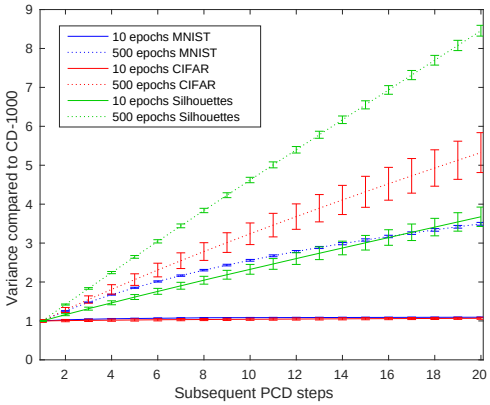

Fig. 3: The PCD vs CD-1000 (assumed to be unbiased and independent) ratio of the variance for the mean of m subsequent estimates after 10 and 500 epochs of training. Error bars indicate standard deviation between iterations.

the mean of 10 subsequent updates have a multiple times smaller variance than PCD. In effect, this means that ignoring any other effects and the effect of weight updates, PCD would need considerably smaller learning rates or larger minibatches to reach the same variance per minibatch. This magnitude is substantial, and might explain the empirical finding that PCD performs best with smaller learning rates than CD.

5 Conclusions

Contrastive Divergence or Persistent Contrastive Divergence are often used for training the weights of Restricted Boltzmann machines. Contrastive Divergence is claimed to benefit from low variance of the gradient estimates when using stochastic gradients. Persistent Contrastive Divergence could on the other hand suffer from high correlation between subsequent gradient estimates due to poor mixing of the Markov chain estimating the model distribution.

In this paper, we have empirically confirmed both of these findings. In experiments on three data sets, we find that the variance of CD-1 gradient estimates are considerably lower than when independently sampling with many steps from the model distribution. Conversely, the variance of the mean of subsequent gradient estimates using PCD is significantly higher than with independent sampling. This effect is mainly observable towards the end of training. In effect, this indicates that from a variance perspective, PCD would require considerably lower learning rates or larger minibatches than CD. As CD is known to be a biased estimator, it therefore seems that the choice between CD and PCD is a trade-off between bias and variance.

Although the results in this paper are practically significant, the approach in this paper is purely empirical. Further theoretical analysis of the variance of PCD and CD gradient estimates would therefore be warranted to confirm these findings.

ESANN 2016 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 27-29 April 2016, i6doc.com publ., ISBN 978-287587027-8.
Available from http://www.i6doc.com/en/.

References

[1] Paul Smolensky. Information processing in dynamical systems: Foundations of harmony theory. 1986.

[2] Geoffrey E Hinton. Training products of experts by minimizing contrastive divergence. *Neural computation*, 14(8):1771–1800, 2002.

[3] Geoffrey E Hinton and Ruslan R Salakhutdinov. Reducing the dimensionality of data with neural networks. *Science*, 313(5786):504–507, 2006.

[4] Laurent Younes. Parametric inference for imperfectly observed gibbsian fields. *Probability Theory and Related Fields*, 82(4):625–645, 1989.

[5] Tijmen Tieleman. Training restricted Boltzmann machines using approximations to the likelihood gradient. In *Proceedings of the 25th international conference on Machine learning*, pages 1064–1071. ACM, 2008.

[6] Yoshua Bengio and Olivier Delalleau. Justifying and generalizing contrastive divergence. *Neural Computation*, 21(6):1601–1621, 2009.

[7] Miguel A Carreira-Perpinan and Geoffrey E Hinton. On contrastive divergence learning. In *Artificial Intelligence and Statistics*, volume 2005, page 17, 2005.

[8] Christopher KI Williams and Felix V Agakov. An analysis of contrastive divergence learning in Gaussian Boltzmann machines. *Institute for Adaptive and Neural Computation*, 2002.

[9] Kevin Swersky, Bo Chen, Ben Marlin, and Nando de Freitas. A tutorial on stochastic approximation algorithms for training restricted boltzmann machines and deep belief nets. In *Information Theory and Applications Workshop (ITA), 2010*, pages 1–10. IEEE, 2010.

[10] Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998.

[11] Alex Krizhevsky and Geoffrey Hinton. Learning multiple layers of features from tiny images. *Master's thesis, Department of Computer Science, University of Toronto*, 2009.

[12] Benjamin M Marlin, Kevin Swersky, Bo Chen, and Nando D Freitas. Inductive principles for restricted Boltzmann machine learning. In *International Conference on Artificial Intelligence and Statistics*, pages 509–516, 2010.

[13] KyungHyun Cho, Tapani Raiko, and Alexander Ilin. Enhanced gradient and adaptive learning rate for training restricted Boltzmann machines. In *Proceedings of the 28th International Conference on Machine Learning (ICML-11)*, pages 105–112, 2011.