NOTE **Communicated by Yoshua Bengio**

Bounding the Bias of Contrastive Divergence Learning

Asja Fischer
asja.fischer@ini.rub.de
Institut für Neuroinformatik, Ruhr-Universität Bochum, 44780 Bochum, Germany

Christian Igel
igel@diku.dk
Department of Computer Science, University of Copenhagen,
2100 Copenhagen Ø, Denmark

Optimization based on k-step contrastive divergence (CD) has become a common way to train restricted Boltzmann machines (RBMs). The k-step CD is a biased estimator of the log-likelihood gradient relying on Gibbs sampling. We derive a new upper bound for this bias. Its magnitude depends on k, the number of variables in the RBM, and the maximum change in energy that can be produced by changing a single variable. The last reflects the dependence on the absolute values of the RBM parameters. The magnitude of the bias is also affected by the distance in variation between the modeled distribution and the starting distribution of the Gibbs chain.

1 Training RBMs Using Contrastive Divergence

Restricted Boltzmann machines (RBMs) are undirected graphical models (Smolensky, 1986; Hinton, 2002). The RBM structure is a bipartite graph consisting of one layer of observable variables $\boldsymbol{V} = (V_1, \dots, V_m)$ and one layer of hidden (latent) variables $\boldsymbol{H} = (H_1, \dots, H_n)$. The modeled distribution is given by $p(\boldsymbol{v}, \boldsymbol{h}) = e^{-\mathcal{E}(\boldsymbol{v},\boldsymbol{h})}/Z$, where $Z = \sum_{\boldsymbol{v},\boldsymbol{h}} e^{-\mathcal{E}(\boldsymbol{v},\boldsymbol{h})}$, and the energy $\mathcal{E}$ is given by

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

with real-valued parameters w_{ij}, b_j, and c_i ($i \in \{1, \dots, n\}$, $j \in \{1, \dots, m\}$) jointly denoted as $\boldsymbol{\theta}$. In the following, we restrict our considerations to RBMs with binary units for which $E_{p(h_i|\boldsymbol{v})}[h_i] = \text{sigmoid}(c_i + \sum_{j=1}^{m} w_{ij} v_j)$ with $\text{sigmoid}(x) = (1 + \exp(-x))^{-1}$.

Neural Computation **23, 664–673** (2011) © 2011 Massachusetts Institute of Technology

Differentiating the log likelihood $\ell(\boldsymbol{\theta} \mid \boldsymbol{v}_l)$ of the model parameters $\boldsymbol{\theta}$ given one training example $\boldsymbol{v}_l$ with respect to $\boldsymbol{\theta}$ yields

$$\nabla_{\boldsymbol{\theta}}\ell(\boldsymbol{\theta} \mid \boldsymbol{v}_l) = -\sum_{\boldsymbol{h}} p(\boldsymbol{h} \mid \boldsymbol{v}_l)\nabla_{\boldsymbol{\theta}}\mathcal{E}(\boldsymbol{v}_l, \boldsymbol{h}) + \sum_{\boldsymbol{v}} p(\boldsymbol{v})\sum_{\boldsymbol{h}} p(\boldsymbol{h} \mid \boldsymbol{v})\nabla_{\boldsymbol{\theta}}\mathcal{E}(\boldsymbol{v}, \boldsymbol{h}). \tag{1.1}$$

Computing the first term on the right side of the equation is straightforward because it factorizes nicely. The computation of the second term is intractable for regular-sized RBMs because its complexity is exponential in the size of the smallest layer.

Therefore, k-step contrastive divergence (CD-k) learning (Hinton, 2002) approximates the second term by a sample obtained by k-steps of Gibbs sampling. Starting from an example $\boldsymbol{v}^{(0)}$ of the training set, the Gibbs chain is run for only k steps, yielding the sample $\boldsymbol{v}^{(k)}$. Each step t consists of sampling $\boldsymbol{h}^{(t)}$ from $p(\boldsymbol{h} \mid \boldsymbol{v}^{(t)})$ and sampling $\boldsymbol{v}^{(t+1)}$ from $p(\boldsymbol{v} \mid \boldsymbol{h}^{(t)})$ subsequently. The gradient (1.1) with respect to $\boldsymbol{\theta}$ of the log likelihood for one training pattern $\boldsymbol{v}^{(0)}$ is then approximated by

$$\mathrm{CD}_k\big(\boldsymbol{\theta}, \boldsymbol{v}^{(0)}\big) = -\sum_{\boldsymbol{h}} p\big(\boldsymbol{h} \mid \boldsymbol{v}^{(0)}\big)\nabla_{\boldsymbol{\theta}}\mathcal{E}\big(\boldsymbol{v}^{(0)}, \boldsymbol{h}\big) +\sum_{\boldsymbol{h}} p\big(\boldsymbol{h} \mid \boldsymbol{v}^{(k)}\big)\nabla_{\boldsymbol{\theta}}\mathcal{E}\big(\boldsymbol{v}^{(k)}, \boldsymbol{h}\big). \tag{1.2}$$

Bengio and Delalleau (2009) show that CD-k is an approximation of the true log-likelihood gradient by finding an expansion of the gradient that considers the kth sample in the Gibbs chain and showing that CD-k is equal to a truncation of this expansion. Furthermore, they prove that the left-out term converges to zero as k goes to infinity:

Theorem 1 (Bengio & Delalleau, 2009, p. 1608). *For a converging Gibbs chain*

$$\boldsymbol{v}^{(0)} \Rightarrow \boldsymbol{h}^{(0)} \Rightarrow \boldsymbol{v}^{(1)} \Rightarrow \boldsymbol{h}^{(1)} \ldots$$

starting at data point $\boldsymbol{v}^{(0)}$, the log-likelihood gradient can be written as

$$\begin{aligned}\nabla_{\boldsymbol{\theta}} \log p(\boldsymbol{v}^{(0)}) = &-\sum_{\boldsymbol{h}} p(\boldsymbol{h} \mid \boldsymbol{v}^{(0)})\nabla_{\boldsymbol{\theta}}\mathcal{E}(\boldsymbol{v}^{(0)}, \boldsymbol{h})\\ &+ E_{p(\boldsymbol{v}^{(k)}|\boldsymbol{v}^{(0)})}\left[\sum_{\boldsymbol{h}} p(\boldsymbol{h} \mid \boldsymbol{v}^{(k)})\nabla_{\boldsymbol{\theta}}\mathcal{E}(\boldsymbol{v}^{(k)}, \boldsymbol{h})\right]\\ &+ E_{p(\boldsymbol{v}^{(k)}|\boldsymbol{v}^{(0)})}\left[\nabla_{\boldsymbol{\theta}} log p(\boldsymbol{v}^{(k)})\right]\end{aligned}$$

and the final term converges to zero as k goes to infinity.

In addition, Bengio and Dellalleau deduce a bound for the final term, which we refer to as the bias of *CD-k*, relating it to the magnitude of the RBM parameters:

$$\left| E_{p(\boldsymbol{v}^{(k)}|\boldsymbol{v}^{(0)})}\left[\frac{\partial \log p(\boldsymbol{v}^{(k)})}{\partial \theta_a}\right]\right| \leq 2^m (1 - 2^m 2^n \mathrm{sigmoid}(-\alpha)^m \mathrm{sigmoid}(-\beta)^n)^k .$$

Here θ_a denotes a single parameter of the RBM, $\alpha = \max_j(\sum_i |w_{ij}| + |b_j|)$, $\beta = \max_i(\sum_j |w_{ij}| + |c_i|)$. But the bound gets loose very fast if the norm of the parameters increases. Note that the absolute value of

$$E_{p(\boldsymbol{v}^{(k)}|\boldsymbol{v}^{(0)})}\left[\frac{\partial \log p(\boldsymbol{v}^{(k)})}{\partial \theta_a}\right] = \sum_{\boldsymbol{v}} p(\boldsymbol{v}^{(k)} = \boldsymbol{v} \mid \boldsymbol{v}^{(0)}) \frac{\partial \log p(\boldsymbol{v})}{\partial \theta_a}$$

is never larger than one for binary RBMs (this follows from $|\partial \log p(\boldsymbol{v})/\partial \theta_a| \leq 1$; e.g., see Bengio & Delalleau, 2009, p. 1611 above equation 4.3), while the bound given above grows quickly with α and β and approaches 2^m, the number of configurations of the visible units.

2 Bounding the CD Approximation Error

In this section, we derive a tighter bound for the bias in CD-k based on general results for the convergence rate of the Gibbs sampler (see Brémaud, 1999). The convergence rate depends on the distance between the distribution of the initial states μ (the starting distribution of the chain) and the stationary distribution. A measure of distance between two distributions α and β on a countable set Ω is the total variation distance defined as

$$d_V(\alpha, \beta) = \frac{1}{2}\|\alpha - \beta\|_1 = \frac{1}{2}\sum_{i \in \Omega} |\alpha(i) - \beta(i)|.$$

The total variation distance between two distributions is bounded by one. We make use of the following theorem:

Theorem 2. *A Markov random field* $\boldsymbol{X} = (X_1, \ldots, X_n)$ *with random variables taking values in a finite set* Ω *and a Markov chain* $(\boldsymbol{X}^{(k)})_{k \geq 0}$ *produced by periodic Gibbs sampling is given. Let* $\mathbf{T}$ *be the transition matrix,* μ *the starting distribution, and* p *the stationary distribution (i.e., the Gibbs distribution) of the Gibbs chain. It holds*

$$\|\mu \mathbf{T}^k - p\|_1 \leq \frac{1}{2}\|\mu - p\|_1 (1 - e^{-N\Delta})^k, \tag{2.1}$$

where

$$\triangle = \sup_{l\in\{1,\dots,n\}} \{|\mathcal{E}(\boldsymbol{x}) - \mathcal{E}(\boldsymbol{y})|; \boldsymbol{x}, \boldsymbol{y} \in \Omega^n \textit{ and } \forall i \in \{1, \dots, n\} \setminus \{l\} : x_i = y_i\},$$

and $\mathcal{E}$ denotes the energy function of the Gibbs distribution.

A proof is given by Brémaud (1999).

In the case of an RBM with hidden variables $\boldsymbol{H}$ and the visible variables $\boldsymbol{V}$ fixed to a pattern $\boldsymbol{v}^{(0)}$, the joint starting distribution is given by

$$\mu(\boldsymbol{v}, \boldsymbol{h}) = \begin{cases} p(\boldsymbol{h} \mid \boldsymbol{v}^{(0)}) & \text{if } \boldsymbol{v} = \boldsymbol{v}^{(0)} \\ 0 & \text{otherwise} \end{cases}. \tag{2.2}$$

Now we can state our main result:

Theorem 3. *Given is an RBM $(V_1, \dots, V_m, H_1, \dots, H_n)$ and a Markov chain produced by periodic Gibbs sampling starting from $\boldsymbol{v}^{(0)}$ ($\boldsymbol{v}^{(0)} \Rightarrow \boldsymbol{h}^{(0)} \Rightarrow \boldsymbol{v}^{(1)} \Rightarrow \boldsymbol{h}^{(1)} \dots$). Let the initial states $(\boldsymbol{v}^{(0)}, \boldsymbol{h}^{(0)})$ be distributed according to μ as defined in equation 2.2, and let p be the joint probability distribution of $\boldsymbol{V}$ and $\boldsymbol{H}$ of the RBM (i.e., the stationary distribution of the Markov chain). Then we can bound the error of the CD-k approximation of the log-likelihood derivative with regard to some RBM parameter θ_a (i.e., $\partial\ell(\boldsymbol{\theta} \mid \boldsymbol{v}^{(0)})/\partial\theta_a$) by*

$$\begin{aligned} \left| E_{p(\boldsymbol{v}^{(k)} \mid \boldsymbol{v}^{(0)})} \left[\frac{\partial log p(\boldsymbol{v}^{(k)})}{\partial \theta_a} \right] \right| &\leq \frac{1}{2} \|\mu - p\|_1 \left(1 - e^{-(m+n)\triangle}\right)^k \\ &\leq \left(1 - e^{-(m+n)\triangle}\right)^k \end{aligned}$$

with

$$\Delta = max \left\{ \max_{l\in\{1,\dots,m\}} \vartheta_l, \max_{l\in\{1,\dots,n\}} \xi_l \right\},$$

where

$$\vartheta_l = max \left\{ \left| \sum_{i=1}^{n} I_{\{w_{il}>0\}} w_{il} + b_l \right|, \left| \sum_{i=1}^{n} I_{\{w_{il}<0\}} w_{il} + b_l \right| \right\}$$

and

$$\xi_l = max \left\{ \left| \sum_{j=1}^{m} I_{\{w_{lj}>0\}} w_{lj} + c_l \right|, \left| \sum_{j=1}^{m} I_{\{w_{lj}<0\}} w_{lj} + c_l \right| \right\}.$$

Proof. Bengio and Delalleau (2009) show that

$$E_{p(\boldsymbol{v}^{(k)}|\boldsymbol{v}^{(0)})}\left[\frac{\partial \log p(\boldsymbol{v}^{(k)})}{\partial \theta_a}\right] = \sum_{\boldsymbol{v}} \Big(p(\boldsymbol{v}^{(k)} = \boldsymbol{v} \mid \boldsymbol{v}^{(0)}) - p(\boldsymbol{v})\Big) \frac{\partial \log p(\boldsymbol{v})}{\partial \theta_a}$$

and use the inequality

$$\left|E_{p(\boldsymbol{v}^{(k)}|\boldsymbol{v}^{(0)})}\left[\frac{\partial \log p(\boldsymbol{v}^{(k)})}{\partial \theta_a}\right]\right| \leq \sum_{\boldsymbol{v}} \left|\Big(p(\boldsymbol{v}^{(k)}{=}\boldsymbol{v} \mid \boldsymbol{v}^{(0)}){-}p(\boldsymbol{v})\Big)\right| \left|\frac{\partial \log p(\boldsymbol{v})}{\partial \theta_a}\right|.$$

Instead of upper-bounding the right-hand side of this equation by

$$\max_{\boldsymbol{v}} \left|\Big(p(\boldsymbol{v}^{(k)} = \boldsymbol{v} \mid \boldsymbol{v}^{(0)}) - p(\boldsymbol{v})\Big)\right| \left[2^m \max_{\boldsymbol{v}} \frac{\partial \log p(\boldsymbol{v}^{(k)})}{\partial \theta_a}\right],$$

as in the proof by Bengio and Delalleau (2009, equation 3.5), we bound it by

$$\begin{aligned}
&\sum_{\boldsymbol{v}} \left|\big(p(\boldsymbol{v}^{(k)} = \boldsymbol{v} \mid \boldsymbol{v}^{(0)}) - p(\boldsymbol{v})\big)\right| \left|\frac{\partial \log p(\boldsymbol{v})}{\partial \theta_a}\right| \\
&= \sum_{\boldsymbol{v}} \left|\sum_{\boldsymbol{h}} \big(p^{(k)}(\boldsymbol{v}, \boldsymbol{h}) - p(\boldsymbol{v}, \boldsymbol{h})\big)\right| \left|\frac{\partial \log p(\boldsymbol{v})}{\partial \theta_a}\right| \\
&\leq \sum_{\boldsymbol{v}} \sum_{\boldsymbol{h}} \left|\big(p^{(k)}(\boldsymbol{v}, \boldsymbol{h}) - p(\boldsymbol{v}, \boldsymbol{h})\big)\right| \left|\frac{\partial \log p(\boldsymbol{v})}{\partial \theta_a}\right| \\
&\leq \sum_{\boldsymbol{v}} \sum_{\boldsymbol{h}} \left|\big(p^{(k)}(\boldsymbol{v}, \boldsymbol{h}) - p(\boldsymbol{v}, \boldsymbol{h})\big)\right|.
\end{aligned}$$

Here we use the notation $p^{(k)}(\boldsymbol{v}, \boldsymbol{h}) = p(\boldsymbol{v}^{(k)} = \boldsymbol{v}, \boldsymbol{h}^{(k)} = \boldsymbol{h} \mid \boldsymbol{v}^{(0)})$ and the fact that in binary RBMs, we have $|\frac{\partial \log P(x)}{\partial \theta_a}| \leq 1$. The right-hand side is twice the total variation distance between the distribution of the variables of the RBM after k steps of Gibbs sampling and the stationary distribution of the chain.

Now we can apply theorem 2 and get

$$\begin{aligned}
\left|E_{p(\boldsymbol{v}^{(k)}|\boldsymbol{v}^{(0)})}\left[\frac{\partial \log p(\boldsymbol{v}^{(k)})}{\partial \theta_a}\right]\right| &\leq \sum_{\boldsymbol{v}} \sum_{\boldsymbol{h}} \left|p^{(k)}(\boldsymbol{v}, \boldsymbol{h}) - p(\boldsymbol{v}, \boldsymbol{h})\right| \\
&\leq \frac{1}{2} \sum_{\boldsymbol{v}} \sum_{\boldsymbol{h}} |(\mu(\boldsymbol{v}, \boldsymbol{h}){-}p(\boldsymbol{v}, \boldsymbol{h}))| \big(1{-}e^{-(m+n)\Delta}\big)^k \\
&= \frac{1}{2} \|\mu - p\|_1 \big(1 - e^{-(m+n)\Delta}\big)^k.
\end{aligned}$$

Here Δ denotes the maximum change in energy that can be produced by changing a single variable. We distinguish the two cases whether the maximum change is produced by a hidden or visible variable and define $\Delta = \max\{\Delta_v, \Delta_h\}$ using $\Delta_v = \max_{l\in\{1,\dots,m\}} \vartheta_l$ and $\Delta_h = \max_{l\in\{1,\dots,n\}} \xi_l$. For the visible units, we have

$$\vartheta_l = \max\left\{|\mathcal{E}(\boldsymbol{v}, \boldsymbol{h}) - \mathcal{E}(\boldsymbol{v}', \boldsymbol{h})|\right\},$$

where we maximize over $\boldsymbol{v}'$, $\boldsymbol{v}$, and $\boldsymbol{h}$ under the constraint that $\forall j \in \{1,\dots,m\}$, $j \neq l : v_j = v'_j$ (i.e., that only one unit changes its state). Thus,

$$\begin{aligned}\vartheta_l =& \max\Bigg\{\Bigg| -\sum_{i=1}^{n}\sum_{j=1}^{m} h_i w_{ij} v_j - \sum_{j=1}^{m} v_j b_j - \sum_{i=1}^{n} h_i c_i \\ &- \left(-\sum_{i=1}^{n}\sum_{j=1}^{m} h_i w_{ij} v'_j - \sum_{j=1}^{m} v'_j b_j - \sum_{i=1}^{n} h_i c_i\right)\Bigg|\Bigg\} \\ =& \max\left\{\left|\sum_{i=1}^{n} h_i w_{il}(v'_l - v_l) + b_l(v'_l - v_l)\right|\right\} \\ =& \max\left\{\left|\sum_{i=1}^{n} h_i w_{il} + b_l\right|\right\} \\ =& \max\left\{\left|\sum_{i=1}^{n} I_{\{w_{il}>0\}} w_{il} + b_l\right|, \left|\sum_{i=1}^{n} I_{\{w_{il}<0\}} w_{il} + b_l\right|\right\},\end{aligned}$$

where the indicator function I is one if its argument is true and zero otherwise. The third equality holds because $(v'_l - v_l)$ is either -1 or 1 and can be pulled out as a common factor. The absolute value of the resulting sum is maximized if the h_i exclusively "select" either all positive or all negative w_{il}, which leads to the final expression. Analogously, we compute ξ_s.

The result for a single initial observed pattern $\boldsymbol{v}^{(0)}$ is appropriate for online learning. It is straightforward to extend the theorem to batch learning in which the gradient and the CD-k approximation are averages over a set of observed patterns defining an empirical distribution:

Corollary 1. *Let p denote the marginal distribution of the visible units of an RBM, and let p_e be the empirical distribution defined by a set of samples $\boldsymbol{v}_1, \dots, \boldsymbol{v}_\ell$. Then* *an upper bound on the expectation of the error of the CD-k approximation of the log-likelihood derivative with regard to some RBM parameter θ_a is given by*

$$\left| E_{p_e(\boldsymbol{v}^{(0)})}\left[E_{p(\boldsymbol{v}^{(k)}|\boldsymbol{v}^{(0)})}\left[\frac{\partial \log p(\boldsymbol{v}^{(k)})}{\partial \theta_a} \right]\right]\right| \leq \frac{1}{2} \| p_e - p \|_1 \left(1 - e^{-(m+n)\Delta}\right)^k$$

with Δ as defined in theorem 3.

We can use p_e instead of the joint starting distribution $\mu(\boldsymbol{v}, \boldsymbol{h}) = p(\boldsymbol{h} \mid \boldsymbol{v}) p_e(\boldsymbol{v})$ on the right-hand side of the equation because

$$\sum_{\boldsymbol{v}} \sum_{\boldsymbol{h}} |(\mu(\boldsymbol{v}, \boldsymbol{h}) - p(\boldsymbol{v}, \boldsymbol{h}))| = \sum_{\boldsymbol{v}} \sum_{\boldsymbol{h}} p(\boldsymbol{h} \mid \boldsymbol{v}) |(p_e(\boldsymbol{v}) - p(\boldsymbol{v}))|$$
$$= \sum_{\boldsymbol{v}} |(p_e(\boldsymbol{v}) - p(\boldsymbol{v}))|.$$

Our bounds shows that the bias is determined by two antagonistic terms. The dependence on $\| p_e - p \|_1$ is an important difference between our results and those derived by Bengio and Delalleau (2009). Since p_e is the target distribution for the RBM learning process, the variation distance between p_e and p should decrease during successful RBM learning. At the same time, the magnitudes of the parameters—if not controlled by weight decay—tend to increase in practice (see, e.g., Bengio & Delalleau, 2009; Fischer & Igel, 2009; Desjardins, Courville, Bengio, Vincent, & Dellaleau, 2010). Thus Δ increases and $(1 - e^{-(m+n)\Delta})^k$ approaches one.

3 Experimental Results

We empirically studied the development of bias and bound during learning of the diag- and the 1DBall data set described by Bengio and Delalleau (2009). Small RBMs with six visible and six hidden neurons were trained by batch learning based on the expected value of the CD-1-update. Their parameters were initialized with weights drawn uniformly from $[-0.5, 0.5]$ and bias parameters set to $c_i = b_j = 0$ for all i and j. Each experiment was repeated 25 times with different initializations. The learning rates were set to 0.1, and no weight decay was used. The results are shown in Figure 1. The bias value plotted is the maximum $\max_{\theta_a} |E_{p(\boldsymbol{v}^{(1)}|\boldsymbol{v}^{(0)})}[\partial \log p(\boldsymbol{v}^{(1)})/\partial \theta_a]|$ over all parameters.

The results show the tightness of the new bound. Only in the initial phase of learning, when $\| p_e - p \|_1$ is large, was the bound rather loose (but always nontrivial, i.e., below one; this is not shown in the left plot). After 50,000 iterations, the differences between bound and bias value as defined above are ≈ 0.00138 and ≈ 0.02628 for Diag and 1DBall, respectively. In the beginning, the bias is small because the models with weights close to zero

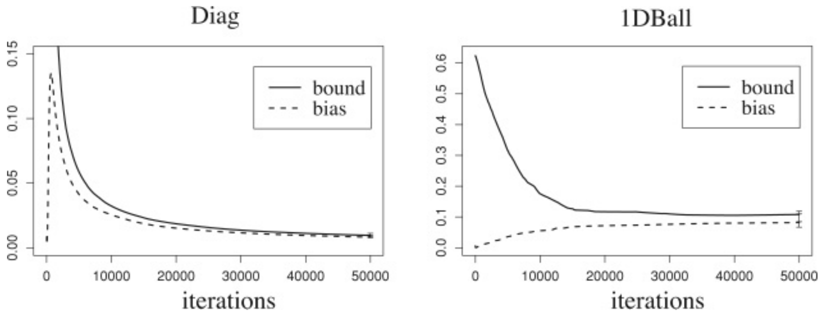

Figure 1: The development of bias and bound during learning of the Diag- (left) and the 1DBall-data set (right) described by Bengio and Delalleau (2009). The bias value plotted is the maximum $\max_{\theta_a} |E_{p(\boldsymbol{v}^{(1)}|\boldsymbol{v}^{(0)})}[\partial \log p(\boldsymbol{v}^{(1)})/\partial\theta_a]|$ over all parameters. Shown are the medians over 25 trials; error bars indicate quartiles.

mix fast (if the weights were all zero, the RBM would model a uniform distribution, which is sampled after a single Gibbs sampling step). We refer to Bengio and Delalleau (2009) for a detailed empirical analysis of CD-k learning of RBMs applied to the Diag and 1DBall benchmark (e.g., showing the dependence on k and the dimensionality of the problem).

4 Discussion and Conclusion

We derived a new upper bound for the bias when estimating the log-likelihood gradient by k-step contrastive divergence (CD-k) for training RBMs. It is considerably tighter than a recently published result. The main reason for that is that it incorporates the decrease of the bias for decreasing distance between the modeled distribution and the starting distribution of the Gibbs chain.

Learning based on CD-k has been successfully applied to RBMs (e.g., Hinton, 2002, 2007; Hinton, Osindero, & Teh, 2006; Hinton & Salakhutdinov, 2006; Bengio, Lamblin, Popovici, Larochelle, & Montreal, 2007; Bengio & Delalleau, 2009). However, it need not converge to a maximum likelihood (ML) estimate of the model parameters (conditions for convergence with probability one are given by Yuille, 2005). Analytical counterexamples are presented by MacKay (2001). Carreira-Perpiñán and Hinton (2005) show that in general, CD learning does not lead to the ML solution. In their experiments, it reaches solutions that are close. However, empirical evidence for misled RBM learning using approximations of the true log-likelihood gradient is, for example, given by Fischer and Igel (2009, 2010), as well as Desjardins et al. (2010). Intuitively, the smaller the bias of the log-likelihood gradient estimation, the higher the chances of converging to an ML solution quickly. Still, even small deviations of a few gradient components can deteriorate the learning process.

Our bound for the bias increases with the maximum possible change in energy that can be produced by changing a single variable. This indicates the relevance of controlling the absolute values of the RBM parameters, for example, by using weight-decay (see the discussion by Fischer & Igel, 2010). Further, the bound increases with the number of RBM variables and decreases with increasing k. The latter underpins that larger values of k stabilize CD learning and that increasing k dynamically when the weights increase may be a good learning strategy (Bengio & Delalleau, 2009).

Acknowledgments

We acknowledge support from the German Federal Ministry of Education and Research within the National Network Computational Neuroscience under grant number 01GQ0951 (Bernstein Fokus "Learning Behavioral Models: From Human Experiment to Technical Assistance").

References

Bengio, Y., & Delalleau, O. (2009). Justifying and generalizing contrastive divergence. *Neural Computation*, *21*(6), 1601–1621.

Bengio, Y., Lamblin, P., Popovici, D., Larochelle, H., & Montreal, U. (2007). Greedy layer-wise training of deep networks. In B. Schölkopf, J. Platt, & T. Hoffman (Eds.), *Advances in neural information processing, 19* (pp. 153–160). Cambridge, MA: MIT Press.

Brémaud, P. (1999). *Markov chains: Gibbs fields, Monte Carlo simulation, and queues*. New York: Springer.

Carreira-Perpiñán, M. Á., & Hinton, G. E. (2005). On contrastive divergence learning. In R. Cowell & Z. Ghahramani (Eds.), *10th International Workshop on Artificial Intelligence and Statistics (AISTATS)* (pp. 59–66). N.p.: Society for Artificial Intelligence and Statistics.

Desjardins, G., Courville, A., Bengio, Y., Vincent, P., & Dellaleau, O. (2010). Parallel tempering for training of restricted Boltzmann machines. *Journal of Machine Learning Research Workshop and Conference Proceedings*, *9*, 145–152.

Fischer, A., & Igel, C. (2009). Contrastive divergence learning may diverge when training restricted Boltzmann machines. *Frontiers in Computational Neuroscience. Bernstein Conference on Computational Neuroscience (BCCN)*. 10.3389/conf.neuro.10.2009.14.121.

Fischer, A., & Igel, C. (2010). Empirical analysis of the divergence of Gibbs sampling based learning algorithms for restricted Boltzmann machines. In K. Diamantaras, W. Duch, & L. S. Iliadis (Eds.), *International Conference on Artificial Neural Networks* (*LNCS: 6354*) (pp. 208–217). New York: Springer.

Hinton, G. E. (2002). Training products of experts by minimizing contrastive divergence. *Neural Computation*, *14*, 1771–1800.

Hinton, G. E. (2007). Learning multiple layers of representation. *Trends in Cognitive Sciences*, *11*(10), 428–434.

Hinton, G. E., Osindero, S., & Teh, Y.-W. (2006). A fast learning algorithm for deep belief nets. *Neural Computation*, *18*(7), 1527–1554.

Hinton, G. E., & Salakhutdinov, R. R. (2006). Reducing the dimensionality of data with neural networks. *Science*, *313*(5786), 504–507.

MacKay, D.J.C. (2001). *Failures of the one-step learning algorithm.* Cambridge, UK: Cavendish Laboratory. Available online at http://www.cs.toronto.edu/∼mackay/gbm.pdf.

Smolensky, P. (1986). Information processing in dynamical systems: Foundations of harmony theory. In D. E. Rumelhart & J. L. McClelland (Eds.), *Parallel distributed processing: Explorations in the microstructure of cognition, Vol. 1: Foundations* (pp. 194–281). Cambridge, MA: MIT Press.

Yuille, A. L. (2005). The convergence of contrastive divergence. In L. Saul, Y. Weiss, & L. Bottou (Eds.), *Advances in neural processing systems, 17* (pp. 1593–1600). Cambridge, MA: MIT Press.

Received May 18, 2010; accepted August 4, 2010.