A New Learning Algorithm for Mean Field Boltzmann Machines

Max Welling and Geoffrey E. Hinton

Depart. of Computer Science, Univ. of Toronto
10 King's College Road, Toronto, M5S 3G5 Canada.
{welling,hinton}@cs.toronto.edu

Abstract. We present a new learning algorithm for Mean Field Boltzmann Machines based on the *contrastive divergence* optimization criterion. In addition to *minimizing* the divergence between the data distribution and the equilibrium distribution, we *maximize* the divergence between one-step reconstructions of the data and the equilibrium distribution. This eliminates the need to estimate equilibrium statistics, so we do not need to approximate the multimodal probability distribution of the free network with the unimodal mean field distribution. We test the learning algorithm on the classification of digits.

1 Introduction

A network of symmetrically-coupled binary (0/1) threshold units has a simple quadratic energy function that governs its dynamic behavior [4].

$$E(\mathbf{v},\mathbf{h}) = -(\frac{1}{2}\mathbf{v}^T\mathbf{V}\mathbf{v} + \frac{1}{2}\mathbf{h}^T\mathbf{W}\mathbf{h} + \mathbf{v}^T\mathbf{J}\mathbf{h}) \tag{1}$$

where $\mathbf{v}$ represent visible units whose states are fixed by the data $\{\mathbf{d}_{1:N}\}$, $\mathbf{h}$ represent hidden units, and where we have added one unit with value always 1, whose weights to all other units represent the biases. The energy function can be viewed as an indirect way of defining a probability distribution over all the binary configurations of the network [2] and if the right stochastic updating rule is used, the dynamics eventually produces samples from this Boltzmann distribution, $P(\mathbf{v},\mathbf{h}) = e^{-E(\mathbf{v},\mathbf{h})}/Z$ where Z denotes the normalization constant or partition function. This "Boltzmann machine" (BM) has a simple learning rule [2] which minimizes the Kullback-Leibler divergence between the data distribution $P_0(\mathbf{v},\mathbf{h}) = P(\mathbf{h}|\mathbf{v})\tilde{P}_0(\mathbf{v})$ (where $\tilde{P}_0(\mathbf{v})$ is the empirical data distribution) and the equilibrium distribution $P_{\text{EQ}}(\mathbf{v},\mathbf{h})$.

$$\delta\mathbf{W} \propto \langle\mathbf{h}\mathbf{h}^T\rangle_0 - \langle\mathbf{h}\mathbf{h}^T\rangle_{\text{EQ}} \qquad \delta\mathbf{V} \propto \langle\mathbf{v}\mathbf{v}^T\rangle_0 - \langle\mathbf{v}\mathbf{v}^T\rangle_{\text{EQ}} \qquad \delta\mathbf{J} \propto \langle\mathbf{v}\mathbf{h}^T\rangle_0 - \langle\mathbf{v}\mathbf{h}^T\rangle_{\text{EQ}} \tag{2}$$

This learning rule is both simple and local, but the settling time required to get samples from the right distribution and the high noise in the estimates of the correlations make learning slow and unreliable.

To improve the efficiency of the BM learning algorithm Peterson and Anderson [6] introduced the mean field (MF) approximation which replaces the averages in eqn. 2 with averages over factorized distributions.

$$\langle \mathbf{h}\mathbf{h}^T\rangle_0 \rightarrow \frac{1}{N}\sum_{n=1}^{N}\mathbf{q}_{0,n}\mathbf{q}_{0,n}^T \qquad \langle \mathbf{v}\mathbf{v}^T\rangle_{\mathrm{EQ}} \rightarrow \mathbf{r}_{\mathrm{EQ}}\mathbf{r}_{\mathrm{EQ}}^T \qquad \langle \mathbf{h}\mathbf{h}^T\rangle_{\mathrm{EQ}} \rightarrow \mathbf{q}_{\mathrm{EQ}}\mathbf{q}_{\mathrm{EQ}}^T \quad (3)$$

where the parameters $\mathbf{q}_{0,n}$, $\mathbf{r}_{\mathrm{EQ}}$ and $\mathbf{q}_{\mathrm{EQ}}$ are computed through the mean field equations,

$$\mathbf{q}_{0,n} = \sigma\left(\mathbf{W}\mathbf{q}_{0,n} + \mathbf{J}^T\mathbf{d}_n\right) \quad \mathbf{r}_{\mathrm{EQ}} = \sigma\left(\mathbf{V}\mathbf{r}_{\mathrm{EQ}} + \mathbf{J}\mathbf{q}_{\mathrm{EQ}}\right) \quad \mathbf{q}_{\mathrm{EQ}} = \sigma\left(\mathbf{W}\mathbf{q}_{\mathrm{EQ}} + \mathbf{J}^T\mathbf{r}_{\mathrm{EQ}}\right) \quad (4)$$

and σ denotes the sigmoid function. These learning rules perform gradient descent on the cost funtion

$$F^{\mathrm{MF}} = F_0^{\mathrm{MF}} - F_{\mathrm{EQ}}^{\mathrm{MF}} \qquad \text{with} \qquad F_Q^{\mathrm{MF}} = \langle E\rangle_Q - H(Q) \quad (5)$$

where $Q = \prod_i q_i^{s_i}(1-q_i)^{1-s_i}$ is a factorized MF distribution, E is the energy in eqn. 1 and H denotes the entropy of Q.

The main drawback of training BMs using MF distributions is that we are approximating distributions which are potentially highly multimodal with a unimodal factorized distribution. This is especially dangerous in the negative phase where no units are clamped to data and the equilibrium distribution is expected to have many modes.

2 Contrastive Divergence Learning

Contrastive Divergence (CD) learning was introduced in [1], to train"Products of Experts" models from data. We start by recalling that the KL-divergence between the data distribution and the model distribution can be written as a difference between two free energies,

$$\mathrm{KL}[P_0(\mathbf{v})||P_{\mathrm{EQ}}(\mathbf{v})] = \mathrm{KL}[P_0(\mathbf{v},\mathbf{h})||P_{\mathrm{EQ}}(\mathbf{v},\mathbf{h})] = F_0 - F_{\mathrm{EQ}} \geq 0 \quad (6)$$

To get samples from the equilibrium distribution we imagine running a Markov chain, first sampling the hidden units with the data clamped to the visible units, then fixing the hidden units and sampling the visible units and so on until we eventually reach equilibrium. It is not hard to show that at every step of Gibbs sampling the free energy decreases on average, $F_0 \geq F_i \geq F_{\mathrm{EQ}}$. It must therefore be true that if the free energy hasn't changed after i steps of Gibbs sampling (for any i), either $P_0 = P_{\mathrm{EQ}}$ or the Markov chain does not mix (which must therefore be avoided). The above suggests that we could use the following contrastive free energy (setting $i = 1$),

$$\mathrm{CD} = F_0 - F_1 = \mathrm{KL}\left[P_0(\mathbf{v},\mathbf{h})||P_{\mathrm{EQ}}(\mathbf{v},\mathbf{h})\right] - \mathrm{KL}\left[P_1(\mathbf{v},\mathbf{h})||P_{\mathrm{EQ}}(\mathbf{v},\mathbf{h})\right] \geq 0 \quad (7)$$

as an objective to minimize. The big advantage is that we do not have to wait for the chain to reach equilibrium. Learning proceeds by taking derivatives with respect to the parameters and performing gradient descent on CD. The derivative is given by,

$$\frac{\partial \mathrm{CD}}{\partial \theta} = \left\langle \frac{\partial E}{\partial \theta} \right\rangle_0 - \left\langle \frac{\partial E}{\partial \theta} \right\rangle_1 - \frac{\partial F_1}{\partial P_1} \frac{\partial P_1}{\partial \theta} \tag{8}$$

with $\theta = \{\mathbf{V}, \mathbf{W}, \mathbf{J}\}$. The last term is hard to evaluate, but small compared with the other two. Hinton [1] shows that this awkward term can be safely ignored. For the BM, this results in the following learning rules,

$$\delta \mathbf{W} \propto \langle \mathbf{h}\mathbf{h}^T \rangle_0 - \langle \mathbf{h}\mathbf{h}^T \rangle_1 \quad \delta \mathbf{V} \propto \langle \mathbf{v}\mathbf{v}^T \rangle_0 - \langle \mathbf{v}\mathbf{v}^T \rangle_1 \quad \delta \mathbf{J} \propto \langle \mathbf{v}\mathbf{h}^T \rangle_0 - \langle \mathbf{v}\mathbf{h}^T \rangle_1 \tag{9}$$

Intuitively, these update rules decrease any systematic tendency of the one-step reconstructions to move away from the data-vectors.

Although some progress has been been made, this algorithm still needs equilibrium samples from the conditional distribution $P(\mathbf{h}|\mathbf{v})$ [1]. Unfortunately, this implies that in the presence of lateral connections among hidden units further approximations remain desirable.

3 Contrastive Divergence Mean Field Learning

In this section we formulate the deterministic mean field variant of the contrastive divergence learning objective. First, let's assume that the MF equations minimize the MF free energy $F_Q^{\mathrm{MF}} = \langle E \rangle_Q - H(Q)$. Imagine N independent systems where data-vectors $\mathbf{d}_n$ are clamped to the visible units and MF equations are run to solve for the means of the hidden units $\mathbf{q}_{0,n}$. The sum of the resultant MF free energies is denoted with $F_0^{\mathrm{MF}} = \sum_n F_{0,n}^{\mathrm{MF}}$. Next, we fix the means of the hidden units, initialize the means of the visible units at the data and take a few steps downhill on the MF free energy. For convenience we will assume that a few iterations of the MF equations achieves this[2] but alternative descent methods are certainly allowed. Finally, we fix these *reconstructions* of the data $\mathbf{r}_{1,n}$, initialize the means of the hidden units at $\mathbf{q}_{0,n}$ and run the MF equations to compute $\mathbf{q}_{1,n}$. Call the sum of the resultant free energies $F_1^{\mathrm{MF}} = \sum_n F_{1,n}^{\mathrm{MF}}$. Summarizing the above with equations we have

$$\mathbf{q}_{0,n} = \sigma(\mathbf{W}\mathbf{q}_{0,n} + \mathbf{J}^T\mathbf{d}_n) \rightarrow \mathbf{r}_{1,n} = \sigma(\mathbf{V}\mathbf{r}_{1,n} + \mathbf{J}\mathbf{q}_{0,n}) \rightarrow \mathbf{q}_{1,n} = \sigma(\mathbf{W}\mathbf{q}_{1,n} + \mathbf{J}^T\mathbf{r}_{1,n}) \tag{10}$$

The last argument in the sigmoid is fixed and acts as a bias term. By the assumption that the MF equations minimize the MF free energy, we may interpret the above procedure as *coordinate descent* on the MF free energy in the variables $\{\mathbf{q}, \mathbf{r}\}$. When this coordinate descent procedure is performed until convergence, each chain, initialized at a particular data-vector, ends up in some local minimum. The sum of the resultant free energies will be called $F_\infty^{\mathrm{MF}} = \sum_n F_{\infty,n}^{\mathrm{MF}}$. The

[1] However, when initialized at the data, brief sampling from $P(\mathbf{v}|\mathbf{h})$ is sufficient.

[2] By running the MF equations sequentially, or by damping them sufficiently this can easily be achieved.

global minimum is denoted as $F^{\text{MF}}_{\text{EQ}}$. It is now easy to verify that the following inequalities must hold.

$$F_0^{\text{MF}} \geq F_1^{\text{MF}} \geq F_\infty^{\text{MF}} \geq N\ F_{\text{EQ}} \tag{11}$$

By analogy with the stochastic contrastive divergence objective we now propose the following 1-step MF contrastive divergence (CD^{MF}) objective,

$$\text{CD}^{\text{MF}} = F_0^{\text{MF}} - F_1^{\text{MF}} = \text{KL}[Q_0(\mathbf{v},\mathbf{h})||P_{\text{EQ}}(\mathbf{v},\mathbf{h})] - \text{KL}[Q_1(\mathbf{v},\mathbf{h})||P_{\text{EQ}}(\mathbf{v},\mathbf{h})] \geq 0 \tag{12}$$

where $Q_0(\mathbf{v},\mathbf{h}) = \tilde{P}_0(\mathbf{v})Q_0(\mathbf{h}|\mathbf{v})$ and Q_1 is the MF distribution after one step of coordinate descent in the variables $\{\mathbf{q},\mathbf{r}\}$. Due to the inequalities in eqn. 11 this objective is always positive. Notice that the only difference with the usual MF objective (eqn. 5) is the fact that we have replaced Q_{EQ} with Q_1. The above cost-function is minimized when the distribution of reconstructions $Q_1 \sim \{\mathbf{r}_{1,n},\mathbf{q}_{1,n}\}$ after one step of MF coordinate descent does not show any average tendency to drift away from the data distribution $Q_0 \sim \{\mathbf{d}_n,\mathbf{q}_{0,n}\}$. One could envision balls initialized at the data which roll down towards their respective local minima in the MF free energy surface over a distance $\mathbf{F}\delta t$. When the shape of the surface is such that the outer products of all *forces* $\mathbf{F}$ (instead of *distances* to the minima) cancel, learning stops.

To compute the update rules we take the derivatives of the CD^{MF} objective with respect to the weights,

$$\frac{\partial \text{CD}^{\text{MF}}}{\partial \theta} = \left\langle \frac{\partial E}{\partial \theta} \right\rangle_{Q_0} - \left\langle \frac{\partial E}{\partial \theta} \right\rangle_{Q_1} - \frac{\partial F^{\text{MF}}}{\partial Q_1}\frac{\partial Q_1}{\partial \theta} \tag{13}$$

where $\theta = \{\mathbf{V},\mathbf{W},\mathbf{J}\}$. The last term represents the effect that the parameters $\{\mathbf{q}_{1,n},\mathbf{r}_{1,n}\}$ will have different values when we change the shape of the surface on which we perform coordinate descent to compute them. This term vanishes for the usual MF objective since in that case we have $\partial F^{\text{MF}}/\partial Q_\infty = 0$. Although this term is awkward to compute, it turns out to be much smaller than the other two in eqn. 13 and can be safely ignored. In section 4 we show experimental evidence to support this claim. Thus, the following update rules can be derived to minimize the CD^{MF} objective eqn. 12,

$$\delta\mathbf{W} \propto \sum_n \left(\mathbf{q}_{0,n}\mathbf{q}_{0,n}^T - \mathbf{q}_{1,n}\mathbf{q}_{1,n}^T\right) \qquad \delta\mathbf{V} \propto \sum_n \left(\mathbf{d}_n\mathbf{d}_n^T - \mathbf{r}_{1,n}\mathbf{r}_{1,n}^T\right)$$

$$\delta\mathbf{J} \propto \sum_n \left(\mathbf{d}_n\mathbf{q}_{0,n}^T - \mathbf{r}_{1,n}\mathbf{q}_{1,n}^T\right) \tag{14}$$

The main advantage of the above learning algorithm is that it only runs MF equations (until convergence) over the hidden units *conditioned* on data-vectors or one-step reconstructions of these data-vectors[3]. Most importantly, MF equations on the highly multimodal energy surface of the free network are entirely avoided.

[3] In fact, for the $\mathbf{q}_{1,n}$ a few steps downhill on the MF free energy is sufficient.

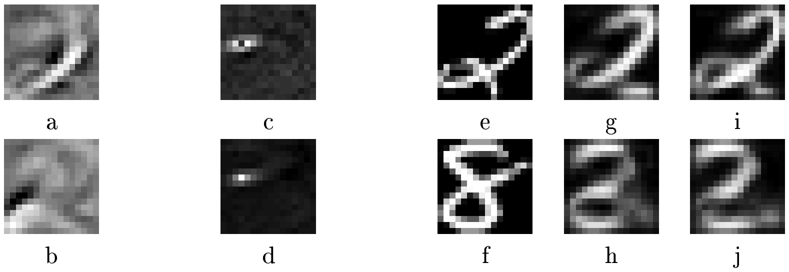


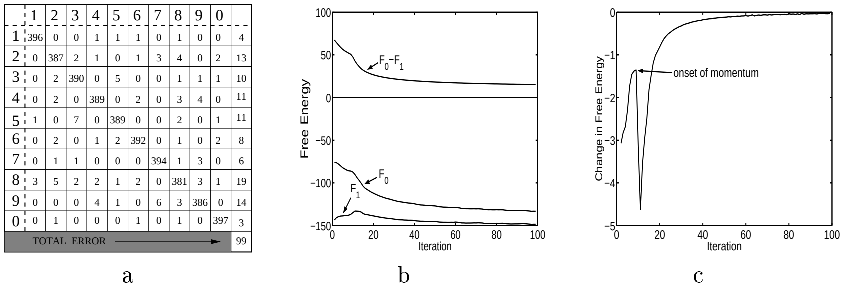
Fig. 1. (a,b)-Two examples of the weights from one hidden unit to all visible units (features) which can be interpreted as thinning (a) and shifting (b) operators. (c)-Visible-to-visible weights for one unit. (d)-ZCA-whitening filter for the same unit as in (c), providing evidence that the visible weights decorrelate the data. (e,f)-Two data vectors. (g,h)-One-step reconstructions of (e,f) by the two-model. (i,j)-Local minima of the two-model corresponding to (e,f). Note that the "8" is being reconstructed as a "2".

4 Experiments

In the experiments described below we have used 16 × 16 real valued digits from the "br" set on the CEDAR cdrom # 1. There are 11000 digits available equally divided into 10 classes. The first 7000 were used for training, while we cycled through the last 4000, using 3000 as a validation set and testing on the remaining 1000 digits. The final test-error was averaged over the 4 test-runs. All digit-images were separately scaled (linearly) between 0 and 1, before presentation to the algorithm. Separate models were trained for each digit, using 700 training examples. Each model was a fully connected MF-BM with 50 hidden units. A total of 2000 updates were performed on mini-batches of 100 data-vectors using a small weight-decay term and a momentum term. When training was completed, we computed the free energy F_0^{MF} for all data on all models (including validation and test data). Since it is very hard to compute the term $F_{\mathrm{EQ}}^{\mathrm{MF}}$, we fit a multinomial logistic regression model to the training data *plus* the validation data, using the 10 free energies F_0^{MF} for each model as "features". The prediction of this logistic regression model on the test data is finally compared with ground truth, from which a confusion matrix is calculated (figure 2-a). The total averaged classification error is 2.5% on this data set, which is a significant improvement over simple classifiers such as a 1-nearest-neighbor (5.5%) and multinomial logistic regression (6.4%). By comparison, a (stochastic) RBM with 50 and 100 hidden units, trained and tested using the same procedure, score 3.1% and 2.4% misclassification respectively. Figures 1 and 2 show some further results for this experiment (see figure captions for explanation).

5 Discussion

In this paper we have shown that efficient *contrastive divergence learning* can be used for BMs with lateral connections by replacing expensive Gibbs sampling with MF equations. During learning the negative phase is replaced with a "one-step-reconstruction" phase, for which the unimodal mean field approximation is



	1	2	3	4	5	6	7	8	9	0	
1	396	0	0	1	1	1	0	1	0	0	4
2	0	387	2	1	0	1	3	4	0	2	13
3	0	2	390	0	5	0	0	1	1	1	10
4	0	2	0	389	0	2	0	3	4	0	11
5	1	0	7	0	389	0	0	2	0	1	11
6	0	2	0	1	2	392	0	1	0	2	8
7	0	1	1	0	0	0	394	1	3	0	6
8	3	5	2	2	1	2	0	381	3	1	19
9	0	0	0	4	1	0	6	3	386	0	14
0	0	1	0	0	0	1	0	1	0	397	3
TOTAL ERROR →											99

Fig. 2. (a)-Confusion matrix for the digit classification task. (b)-Contrastive MF free energy (computed every 20 iterations). (c)-Change in contrastive MF free energy. Note that this change is always negative supporting our claim that the ignored term in eqn. 13 is much smaller than the other two.

expected to be appropriate. Recently (see [7] in this volume) this algorithm has been succesfully applied to the study of associative mental arithmetic.

The approach presented in this paper is straightforwardly extended to supervised learning (see [5] for related work) but seems less successful on the digit recognition task.

CD-learning is a very general method for training undirected graphical models from data. The ideas presented in this paper are easily modified to more sophisticated deterministic approximations of the free energy like the TAP and Bethe approximations. Also, both the stochastic and deterministic versions are easily extended to discrete models with an arbitrary number of states per unit. We have recently also applied CD-learning to models with continuous states, where Hybrid Monte Carlo sampling was used to compute the one-step reconstructions of the data [3].

References

1. G.E. Hinton. Training products of experts by minimizing contrastive divergence. Technical Report GCNU TR 2000-004, Gatsby Computational Neuroscience Unit, University College London, 2000.
2. G.E. Hinton and T.J. Sejnowski. *Learning and relearning in Boltzmann machines*, volume Volume 1: Foundations. MIT Press, 1986.
3. G.E. Hinton, M. Welling, Y.W. Teh, and K. Osindero. A new view of ICA. In *Int. Conf. on Independent Component Analysis and Blind Source Separation*, 2001.
4. J.J. Hopfield. Neural networks and physical systems with emergent collective computational abilities. In *Proceedings of the National Academy of Sciences*, volume 79, pages 2554–2558, 1982.
5. J.R. Movellan. Contrastive hebbian learning in the continuous hopfield model. In *Connectionist Models, Proceedings of the 1990 Summer School*, pages 10–17, 1991.
6. C. Peterson and J. Anderson. A mean field theory learning algorithm for neural networks. *Complex Systems*, 1:995–1019, 1987.
7. I. Stoianov, M. Zorzi, S. Becker, and C. Umilta. Associative arithmetic with Boltzmann machines: the role of number representations. In *International Conference on Artificial Neural Networks*, 2002.