Learning Causally Linked Markov Random Fields

G. E. Hinton, S. Osindero and K. Bao
Department of Computer Science
University of Toronto
Toronto, Canada M5S 3G4

Abstract

We describe a learning procedure for a generative model that contains a hidden Markov Random Field (MRF) which has directed connections to the observable variables. The learning procedure uses a variational approximation for the posterior distribution over the hidden variables. Despite the intractable partition function of the MRF, the weights on the directed connections and the variational approximation itself can be learned by maximizing a lower bound on the log probability of the observed data. The parameters of the MRF are learned by using the mean field version of contrastive divergence [1]. We show that this hybrid model simultaneously learns parts of objects and their inter-relationships from intensity images. We discuss the extension to multiple MRF's linked into in a chain graph by directed connections.

1 Introduction

Generative models are widely used within machine learning. However, in many applications the graphical models involve exclusively causal, or exclusively undirected edges. In this paper we consider models that contain *both* types of edge, and suggest approximate learning methods for such models. The main contribution of this paper is the proposal of combining variational inference with the contrastive divergence algorithm to facilitate learning in systems involving causally linked Markov Random Fields (MRF's). We support our proposal with examples of learning in several domains.

2 Learning Causal Models

One way to make generative models with stochastic hidden variables is to use a directed acyclic graph as shown in Figure 1 (a). The difficulty in learning such "causal" models is that the posterior distribution over the hidden variables is intractable (except in certain special cases such as factor analysis, mixture models, square ICA or graphs that are very sparsely connected). Despite the intractability of the posterior, it is possible to optimize a bound on the log probability of the data by using a simple factorial distribution, $Q(\mathbf{h}|\mathbf{x})$, as an approximation to the true posterior, $P(\mathbf{h}|\mathbf{x})$ over hidden configurations, $\mathbf{h}$, given a data-vector, $\mathbf{x}$. If the hidden variables are binary, a factorial distribution can be represented by assigning a probability, q_j to each hidden variable, j:

$$Q(\mathbf{h}|\mathbf{x}) = \prod_j q_j^{h_j}(1-q_j)^{1-h_j} \tag{1}$$

where h_j is the binary state of hidden unit j in hidden configuration $\mathbf{h}$. Neal and Hinton [2] show that:

$$-\log P(\mathbf{x}) = \mathcal{F}(\mathbf{x}) - \mathrm{KL}(Q(\mathbf{h}|\mathbf{x})||P(\mathbf{h}|\mathbf{x})) \tag{2}$$

where the $\mathcal{F}$ denotes the 'variational free-energy' of the data and is given by

$$\mathcal{F}(\mathbf{x}) = \sum_{\mathbf{h}} Q(\mathbf{h}|\mathbf{x}) \log Q(\mathbf{h}|\mathbf{x}) - \sum_{\mathbf{h}} Q(\mathbf{h}|\mathbf{x}) \log P(\mathbf{h}, \mathbf{x}) \tag{3}$$

where $\mathbf{x}$ is a data-vector and $P(\mathbf{h}, \mathbf{x})$ is the joint probability of first generating $\mathbf{h}$ from the model, and then generating $\mathbf{x}$ from $\mathbf{h}$.

Since the intractable KL divergence term in equation 2 is non-negative, the variational free-energy, $\mathcal{F}$, gives a tractable upper bound on the negative log probability of the data. Minimizing this bound also has the useful property that it tends to adjust the parameters to make the true posterior distribution as factorial as possible which makes factorial approximate inference work well in the learned model.

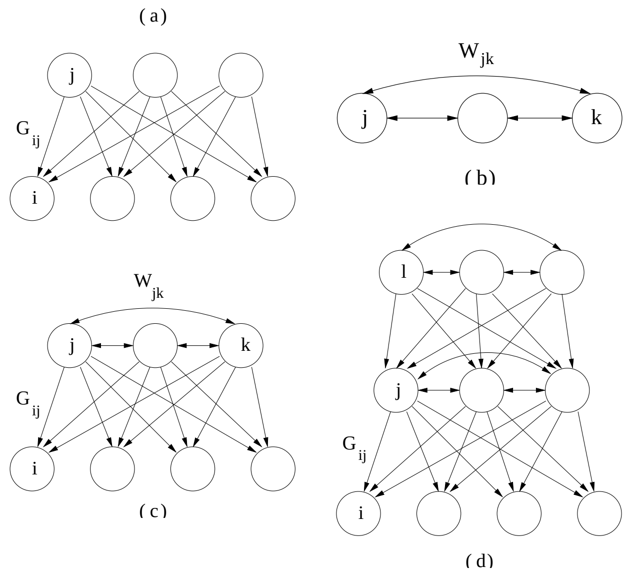

Figure 1: (a) A "causal" generative model. (b) A Markov random field (MRF) with pairwise interactions between the variables. (c) A hybrid model in which the hidden variables of a causal generative model form a Markov random field. (d) A causal hierarchy of MRF's.

For each data-vector in the training set, a locally optimal factorial approximation to the true posterior can be found by following the gradient of the bound w.r.t. Q. Alternatively, the same gradient can be used to train a feedforward "recognition" network to map each training case to a good Q. Once it has been learned, the feedforward network can be viewed as a way of caching the results of iterative settling whilst also acting as a regularizer that encourages similar data-vectors to use similar Q distributions.

3 Learning Markov Random Fields

Hidden latent causes are a good way to model some types of correlation, but they are not good at modeling *constraints* between variables[1]. Consider, for example, a spherical, zero-mean, 20-dimensional Gaussian that has been projected onto the plane in which the sum of the coordinates is 1. To capture this constrained distribution, factor analysis requires 19 hidden factors because it must use a very tight noise model on all 20 variables and then use hidden factors to increase the variance in the 19 allowable directions of variation. Hidden ancestral variables cannot be used to decrease variance[2].

A better way to model constraints is to use an "energy-based" model that associates high energies with data-vectors that violate constraints. The probability of a data-vector is then defined in terms of its energy using the Boltzmann distribution:

$$P(\mathbf{x}) = \frac{e^{-E(\mathbf{x})}}{Z}, \qquad Z = \sum_{\mathbf{u}} e^{-E(\mathbf{u})} \tag{4}$$

where $\mathbf{x}$ is a data-vector, $E(\mathbf{x})$ is its energy, and $\mathbf{u}$ is an index over all possible data-vectors.

The main difficulty in learning energy-based models comes from the normalizing term, Z, (called the partition function) in Eq 4. This is an intractable sum or integral over all possible data-vectors. If a Markov chain is used to sample vectors, $\mathbf{u}$ from the distribution defined by the model, it is possible to get an unbiased estimate of the gradient of the log probability of the data:

$$\frac{\partial \log P(\mathbf{x})}{\partial \theta} = -\frac{\partial E(\mathbf{x})}{\partial \theta} + \sum_{\mathbf{u}} P(\mathbf{u}) \frac{\partial E(\mathbf{u})}{\partial \theta} \tag{5}$$

However, the estimate of the gradient will be very noisy and it is typically hard to know how long to run the Markov chain before it is sampling from the model's distribution. In practice, it is common to assume that if the learning works, the Markov chain must have been close to its equilibrium distribution — a dubious inference.

In some energy-based models, such as a Boltzmann machine with interconnected hidden variables, it is necessary to sum over all possible configurations of the hidden variables to compute the numerator in Eq 4. In other energy-based models, such as "fully visible" Boltzmann machines that just have lateral connections between the visible units it is easy to compute the energy of a data-vector[3] but it is still hard to get the exact derivatives of the partition function. For models of this type, Hinton [3] has shown that learning can still work very well if a Markov chain is started at the data and then run for just a few steps instead of being run all the way to equilibrium.

The use of a brief Markov chain can be combined with the mean field approximation in which the distribution over binary configurations is represented by a factorial distribution Q [1]. For fully visible Boltzmann machines, this leads to a learning algorithm in which the network starts at a data-vector and then updates the q values of all the units in parallel using the rule:

$$q_j^{t+1} = \lambda q_j^t + \frac{1-\lambda}{1+\exp(-b_j - \sum_k q_k^t w_{jk})} \tag{6}$$

[1]In a directed graph, this requires *observed* descendants.

[2]Assuming the factor loadings do not use imaginary components to create negative variance.

[3]Other models that fall within this class include "restricted Boltzmann machines" in which there are no interconnections between hidden units and also models in which the global energy is a function of the activities of multiple layers of deterministic, non-linear hidden units.

where b_j is the bias of unit j, w_{jk} is a symmetric connection between unit j and unit k, and λ is a damping coefficient between 0 and 1 that is used to prevent oscillations. Using the parallel updates in Eq. 6, the learning rule in Eq. 5 becomes:

$$\Delta w_{jk} \propto \sum_{cases} q_j^+ q_k^+ - q_j^- q_k^- \tag{7}$$

where the q^+ values are the the components of a training vector and the q^- values are produced by allowing the mean field net to run for a few iterations of equation 6. The q^+ values would normally be binary, but the learning procedure can still be applied if each training case is a factorial distribution over binary vectors.

4 Causally Linked Markov Random Fields

Both purely causal models and MRF's are used extensively within machine learning, but there are noticeably fewer models in the literature that employ both causal and undirected connections [4]. Causal hierarchies of MRF's (chain-graphs) have some very attractive properties as generative models (see below) but the problem of *learning* them efficiently when there is dense connectivity has not been adequately addressed.

To generate data from such a model[6], we first run the top-level MRF to equilibrium and pick a configuration from the distribution defined by its energy function. This configuration then provides top-down input to the MRF at the next level down via the causal connections. The top-down input modifies the energy function of the second level MRF by changing the effective biases of its units[5]. We then run the second level MRF using its modified energy function and pick a configuration from its distribution. This can be repeated for as many levels as desired, with the bottom level being the "visible" units which may or may not be connected together in an MRF.

This generative model has a major advantage over a purely causal hierarchy: At each level of the hierarchy, learned constraints can be used to "clean-up" the representations generated from the level above. Consider, for example, a generative model in which the top level represents the pose parameters of a face and the next level down represents the pose parameters of each of the two eyes. The height of an eye within the face is somewhat variable, but the two eyes are constrained to have the same height. This creates a problem for a purely causal hierarchy in which the poses of the left and right eye are conditionally independent given the representation at the level above. The height of both eyes must be chosen at the top level and then the height of each eye must be communicated very accurately to the level below. But if an MRF can be used for clean-up at the level below, the height of each eye can be loosely determined by the top-down input, and the MRF can then enforce the constraint on the two heights. So the top-down input to each level can be used to select between (and distort) highly structured and finely balanced alternatives rather than having to specify a pattern in full detail. The causal connections are adept at suggesting which 'parts' to instantiate and roughly where to put them, whilst the undirected connections within the MRF are ideal for enforcing consistency relationships between these parts.

As we shall see, combining multiple MRF's into causal hierarchies also has a major advantage over combining them into one big MRF by using undirected connections: The causal connections between layers act as insulators that prevent the partition functions of the individual MRF's from combining together into one large partition function.

5 A simple version of the model

We begin by presenting the simplest architecture from the framework we have just described: a single, hidden MRF layer with causal connections to a layer of observed variables as illustrated by the network shown in Figure 1 (c).

For concreteness, we will work with a particular simple form for the model's interactions, although more elaborate cases can be treated in essentially the same way. The hidden MRF layer will consist of a Boltzmann machine which has binary nodes with pairwise interaction energies of the form $E(h_i, h_j) = h_i h_j w_{ij}$, and single node energies of the form $E(h_i) = b_i h_i$ where h_k is the binary state of node k and $\{w_{ij}, b_i\}$ are free parameters to be learned. Conditioned upon these hidden variables, the directed connections in our model specify a Gaussian distribution on the observables with $P(\mathbf{x}|\mathbf{h}) = \mathcal{N}(\mathbf{G}\mathbf{h} + \mathbf{m}; \sigma\mathbf{I})$ where σ is a pre-specified noise variance[6].

We use a single-layer sigmoid recognition network to specify the q's of the posterior approximation in equation 1 and the probabilities are given by

$$q_i = \left(1 + e^{-\sum_j R_{ij} x_j + c_i}\right)^{-1} \tag{8}$$

where $\{R_{ij}, c_i\}$ are parameters to be learned [7].

[4] Such models are formally referred to as chain-graphs; see for example [4, 5].

[5] It could also modify pairwise interactions between units in the lower-level MRF.

[6] We fix σ for simplicity, but it could also be learned.

[7] The derivatives that are used to train this recognition

Our formalism leads to the following expression for the variational free energy,

$$
\begin{aligned}
\mathcal{F} &= \mathcal{F}_{\text{MRF}} + \mathcal{F}_{\text{Gauss}} && (9) \\
\mathcal{F}_{\text{MRF}} &= \sum_i \left[q_i \log q_i + (1 - q_i)\log(1 - q_i)\right] \\
&\quad -\frac{1}{2}\mathbf{q}^T\mathbf{W}\mathbf{q} - \mathbf{b}^T\mathbf{q} + \log Z && (10) \\
\mathcal{F}_{\text{Gauss}} &= \frac{1}{2\sigma^2}\left(\mathbf{q}^T\mathbf{G}^T\mathbf{G}\mathbf{q} - 2\mathbf{x}^T\mathbf{G}\mathbf{q}\right) \\
&\quad +\mathbf{q}^T\mathbf{K}(1 - \mathbf{q}) + c && (11)
\end{aligned}
$$

where $K_{ij} = \delta_{ij}(\mathbf{G}^T\mathbf{G})_{ij}$, and c denotes constants that do not affect the derivatives of $\mathcal{F}$ *w.r.t.* the parameters. Minimising $\mathcal{F}$ is equivalent to maximising a lower bound on the data log-likelihood.

A crucial property of this model is that the intractable $\log Z$ term only depends on the biases and lateral connections of the hidden units. It does not enter into the derivatives of either the q^+ values or the weights on the causal connections. So the recognition weights $(\mathbf{R}, \mathbf{c})$ that determine the q values, and also the causal generative parameters $(\mathbf{G}, \mathbf{m})$, can be learned by using the exact gradient of the cost function To learn the hidden biases and the lateral weights $(\mathbf{b}, \mathbf{W})$ between hidden units, we allow the hidden units to run for a few mean field iterations from their initial q values and then use the contrastive divergence learning rule [3], as given in equation 7.

6 A toy example

To illustrate the model we used 50,000 24x24 images of the digit seven that were generated by small rotations, translations and scalings of 1000 normalized 16x16 images from the Cedar CD-Rom. The distortions reduced the long-range correlations introduced by the normalization. We trained a network with 64 fully inter-connected hidden units for 500 sweeps through the training set updating the weights after every 250 examples. There was very little change in the weights after 80 sweeps. We used a momentum of 0.9 with learning rates of 10^{-4} for the causal generative connections and visible biases and for the recognition connections and biases, and 10^{-5} for the lateral connections and hidden generative biases. We also implemented L1 weight-decay corresponding to a Laplacian prior on the lateral connections. This aids interpretability by making most lateral connections small or zero, whilst also allowing large values for a few weights.

Figure 2 (a) and (b) show the generative weights of all 64 hidden units, along with examples of lateral interaction patterns for 4 representative units. The figure caption highlights some salient aspects of the learned lateral connections.

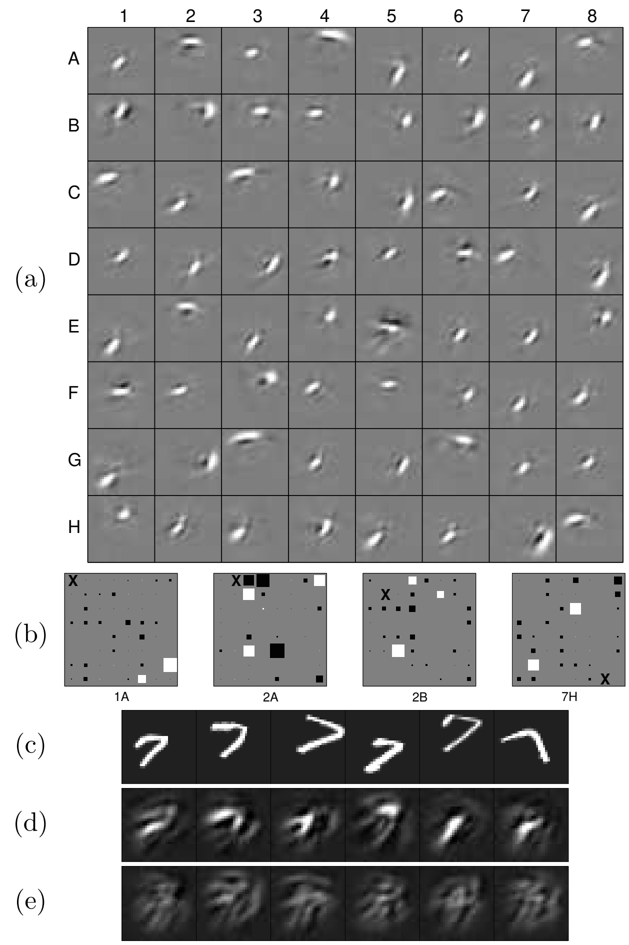

Figure 2: (a) The generative weights of all 64 hidden units in a model of handwritten 7's. (b) The lateral connection patterns for units 1A, 1C, 3F and 7H. The X marks the location of the unit itself. Note the positive interactions between units with collinear generative fields (e.g. 7H and 2G) and also the sizeable negative weights between mutually exclusive alternatives (e.g. 2A and 4A). Unit 2B appears to be a corner detector, and its interactions with 4A and 6B match this intuition. (c) Examples of the training data used. (d) Samples from the distribution learned by the model (obtained using prolonged Gibbs sampling.) (e) Samples from a model with the same generative parameters as in (a,d) but with the lateral connections set to zero, and the biases re-learned to compensate. Notice that there is much less consistency between the strokes in the samples generated from the model without lateral connections.

7 Learning to model natural objects as inter-related parts

It is hard to model real-valued images using binary hidden units so we use binomial units that are equiv-

---

network could be used to train a far more powerful recognition network that contained hidden layers.

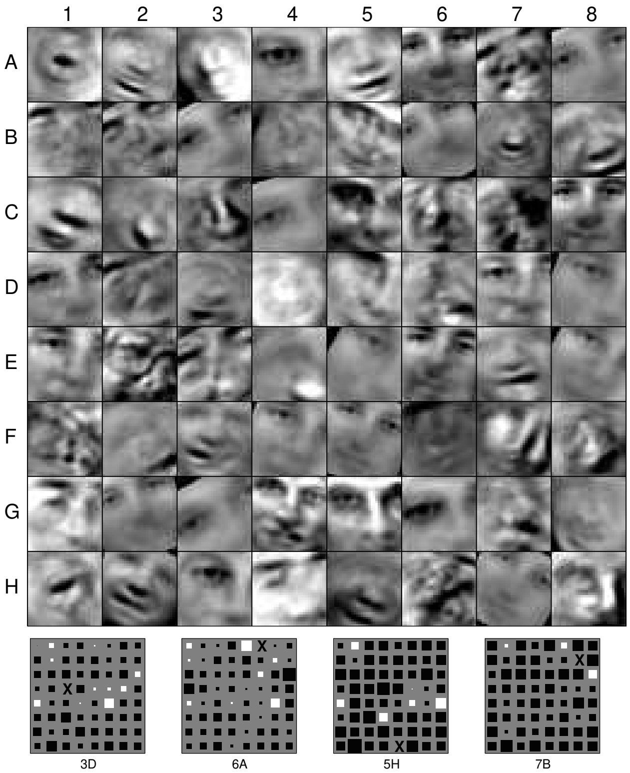

Figure 3: The results of applying the learning algorithm to images of faces. The generative weights of the hidden units are shown at the top and the lateral connections of some of the hidden units are shown beneath. The $8,400$ 31x31 training images were created by rotating ($\pm 30^{\circ}$), scaling (1.0 to 1.5), cropping and subsampling the 400 face images of 40 different people in the Olivetti face dataset. Each cropped image was then centred (zero pixel mean) and PCA was used to whiten the data and reduce the dimensionality from 961 to 144 by maintaining the normalised projections on the leading 144 eigenvectors.

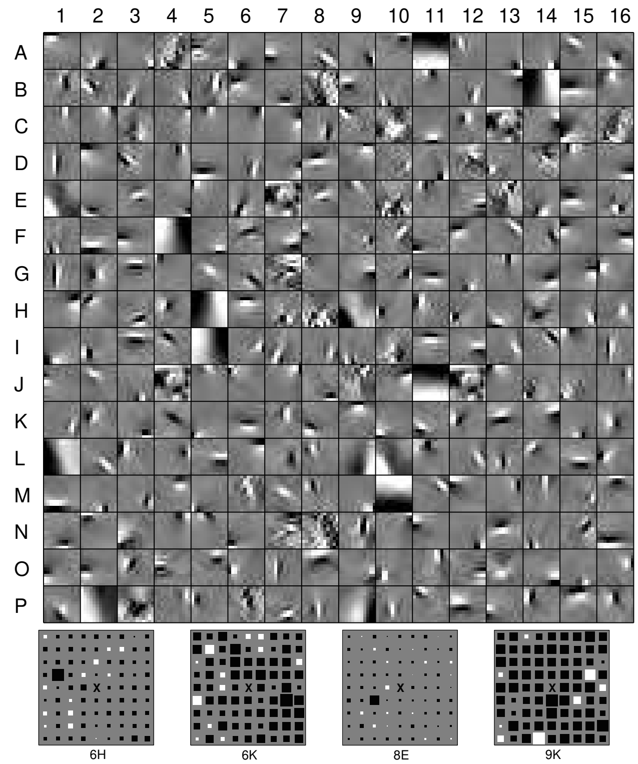

Figure 4: The generative weights of 256 hidden units trained on $150,000$ 12x12 patches of natural images extracted from Hyvarinen's natural image data. The images were whitened and reduced to 100 dimensions using centering and PCA. The lateral interactions were restricted to a 9x9 neighborhood with wraparound. There are strong negative interactions between anti-phase pairs {6H, 3G} & {8E, 6F} and also between highly non-collinear pairs {6H, 5H}, {6K, 5H}, {6K, 5I}, {6K, 9L}. The interactions between approximately collinear pairs with consistent phase are usually positive: {6H, 2D}, {6H, 5G}, {6K, 2L}.

alent to replicating each hidden unit (together with all its weights) $N = 100$ times [7]. We also make an additional modification that is motivated by a desire to produce more neurally plausible representation schemes. The variance contributed by a binomial pool of N binary units each of which has a probability of q of turning on is $Nq(1-q)$. (This appears through the term $\mathbf{q}^T\mathbf{K}(1-\mathbf{q})$ in equation 11.) If we omit the $(1-q)$ term, binomial units cannot use values of q near 1 to achieve low variance and so they learn to use small values of q and behave like Poisson units whose variance is linear in their "firing rate".

Figure 3 shows the weights learned by a network with 64 hidden Poisson units when it was trained on images of faces.

After learning, the hidden activities are sparse with a small subset of the units having activities significantly above their baseline for each image. The ability to learn parts and their relationships simultaneously should make it easier to achieve the goal of finding natural parts of objects in sets of unlabelled images [8], but we have not yet had time to explore this issue in detail. Unlike non-negative matrix factorization [9] our model learns parts without requiring any restrictions on the weights, but it is possible that it would be even better at extracting parts if we restricted the weights on the causal connections to be positive.

Clearly, it would be better to perform some extraction of low level features before attempting to extract inter-related parts of complex objects. Figure 4 shows the results of applying exactly the same algorithm to patches of natural images.

8 Learning with multiple hidden layers

Ideally, a whole hierarchy of features at different levels should be learned cooperatively in order to encourage low-level features to be useful for extracting high-level parts that have consistent inter-relations. Our model is proposed with multiple hidden layers in mind, however we have only just started to investigate this empirically.

We now present the free energy, $\mathcal{F}_2$, for a model with two hidden MRF layers, with the 'top' layer having a directed influence on the layer below (as shown in Figure 1 (d)). If we are able to adequately tackle the extra complexity involved in learning such a model then the generalisation to hierarchies of arbitrary depth involves relatively little extra effort. We will now use h^m and h^t to denote the binary states of hidden units in the middle and top MRF layers respectively. As before, $Q(\mathbf{h}^m|\mathbf{x})$ will denote the a factorial approximation to the posterior probabilities for the MRF units connected to the observables, and we will use $R(\mathbf{h}^t|\mathbf{x})$ to denote the factorial approximation for the MRF units in the top layer.

$$
\begin{aligned}
\mathcal{F}_2 \;=\;& \sum_{\mathbf{h}^t} R(\mathbf{h}^t|\mathbf{x}) \log R(\mathbf{h}^t|\mathbf{x}) \\
&+ \sum_{\mathbf{h}^m} Q(\mathbf{h}^m|\mathbf{x}) \log Q(\mathbf{h}^m|\mathbf{x}) \\
&- \sum_{\mathbf{h}^t} R(\mathbf{h}^t|\mathbf{x}) \log P(\mathbf{h}^t) \\
&- \sum_{\mathbf{h}^t,\mathbf{h}^m} R(\mathbf{h}^t|\mathbf{x}) Q(\mathbf{h}^m|\mathbf{x}) \log P(\mathbf{h}^m|\mathbf{h}^t) \\
&- \sum_{\mathbf{h}^m} Q(\mathbf{h}^m|\mathbf{x}) \log P(\mathbf{x}|\mathbf{h}^m) \qquad (12)
\end{aligned}
$$

The main difference between this free energy and the one which we have already dealt with is due to the term $\sum_{\mathbf{h}^t,\mathbf{h}^m} R(\mathbf{h}^t|\mathbf{x})Q(\mathbf{h}^m|\mathbf{x}) \log P(\mathbf{h}^m|\mathbf{h}^t)$. The partition function of the middle layer MRF now depends on the states in the top layer MRF. Consequently we are required to deal with an *expectation* over partition functions as one of the terms within our free energy. Again for concreteness we first present the mathematical form of the free energy for a simple case before discussing an initial approximation for overcoming this difficulty. Our model now involves two Boltzmann machine layers, as illustrated by Figure 1 (d), and conditioning on the states of the top layer provides an additional bias term to the energy function of the layer below. The factorial approximation to the posterior on the middle layer units remains unchanged, and a similar approximation is used for the top level units, specifically $R(\mathbf{h}^t|\mathbf{x}) = \prod_j r_j^{h_j^t}(1-r_j)^{1-h_j^t}$. As before, the observables are given by a Gaussian distribution conditioned on the states of the middle layer units. The free energy is given by,

$$
\begin{aligned}
\mathcal{F}_2 \;=\;& \sum_j [r_j \log r_j + (1-r_j)\log(1-r_j)] \\
&+ \sum_i [q_i \log q_i + (1-q_i)\log(1-q_i)] \\
&- \frac{1}{2}\mathbf{r}^T\mathbf{H}\mathbf{r} - \mathbf{c}^T\mathbf{r} + \log Z_{\text{TOP}} \\
&- \frac{1}{2}\mathbf{q}^T\mathbf{W}\mathbf{q} - (\mathbf{b}+\mathbf{r})^T\mathbf{q} \\
&+ \quad \left\langle \log Z_{\text{MID}}(\mathbf{h}^t) \right\rangle_{\mathbf{h}^t \sim R(\mathbf{h}^t|\mathbf{x})} \\
&+ \mathcal{F}_{\text{Gauss}} \qquad (13)
\end{aligned}
$$

One strategy is to replace the expectation over partition functions with the partition function evaluated at the expected value of $\mathbf{h}^t$, i.e. at $\mathbf{h}^t = \mathbf{r}$. This can be viewed as a first order Taylor series approximation to $\log Z_{\text{MID}}(\mathbf{h}^t)$ about the mean of $R(\mathbf{h}^t|\mathbf{x})$ (higher order expansions might also be feasible, however the terms are much more complicated.) Such an approximation means that the free energy is no longer a bound on the true log likelihood, however we are at present unaware of any other tractable approximation that would allow us to maintain such a bound.

In this new approximation we use contrastive divergence both to estimate derivatives of the lateral connections and MRF biases, and also to compute a component of the derivative with respect to the top level activities, $\mathbf{r}$. (From the point of view of forming derivatives, the top level units simply act as case dependent biases.)

Preliminary experiments using models with two MRF layers causally linked into a hierarchy indicate that this approximation might be adequate for our gradient based learning. The 'middle' MRF layer typically develops features that are qualitatively similar to those in the single layer case. The 'top' level units tend to sensibly co-activate sets of units in the 'middle' layer, however it is hard to properly characterise the behaviour of units deeper within a densely connected network and their effects are not always apparent simply by studying the generative weights.

To illustrate the increased representational power achieved by adding an additional MRF layer, we present somewhat qualitative results from a simple experiment again using the Cedar digits. Our data consisted of 1100 16×16 images of each class type from 0 to 9 (that is 11000 training examples in total). Figure 5 (a) shows an example of the training data. Using this dataset, we trained two different model architectures: the first had a single hidden Boltzmann machine layer consisting of 256 fully interconnected units; the second had two hidden Boltzmann machine layers, again with 256 fully interconnected units within each layer, and with directed connections from the top layer providing additional biases to the middle layer. We trained both networks until the changes in parameters were very small (approximately 500 sweeps through the whole data set). Figures 5 (b) and (c) illustrate generative samples from models with one and two hidden MRF layers respectively. From this qualitative comparison it is immediately apparent that the model with a hierarchy of MRF layers has managed to capture more of the statistical structure within the dataset. The generated samples in Figure 5 (b) somewhat resemble single digits, but they are also rather contaminated by additional strokes — as if several digits classes were combined. This contamination is present to a much smaller degree in Figure 5 (c) in which we can see clearer examples of single digits being generated. We speculate that the additional hidden layer is beneficial by providing top down biases to shift the middle layer activities in favour of the strokes for particular digit classes, which might then make the task of ensuring 'stroke consistency' easier for the lateral connections within that layer.

9 Improving the accuracy of approximate inference

There are several reasons why one might wish to use models containing both directed and undirected connections. As discussed in Section 4 they are elegantly able to capture some kinds of statistical structure which would be difficult to capture using connections of just one type. In particular, hierarchies of MRF's have many appealing properties that make them suitable for learning parts-based representations.

Another quite different reason for choosing to combine elements of both kinds of model is to allow approximate inference techniques to work more effectively, and this benefit can be seen in the case of even just a single hidden MRF layer. Many approximate inference techniques assume some simplifying independence relationships, but such relationships generally do not, and cannot, hold in the true posterior. In particular, if the latent variables are assumed to be independent in the prior, an effect known as 'explaining-away' causes those variables to coupled in the posterior [10]. However, somewhat counter-intuitively, it is possible to reduce or eliminate this *posterior* dependence by using a model in which the variables are coupled in the opposite way in the *prior*. The required coupling depends on the parameters, but not on the data.

(a)

(b)

(c)

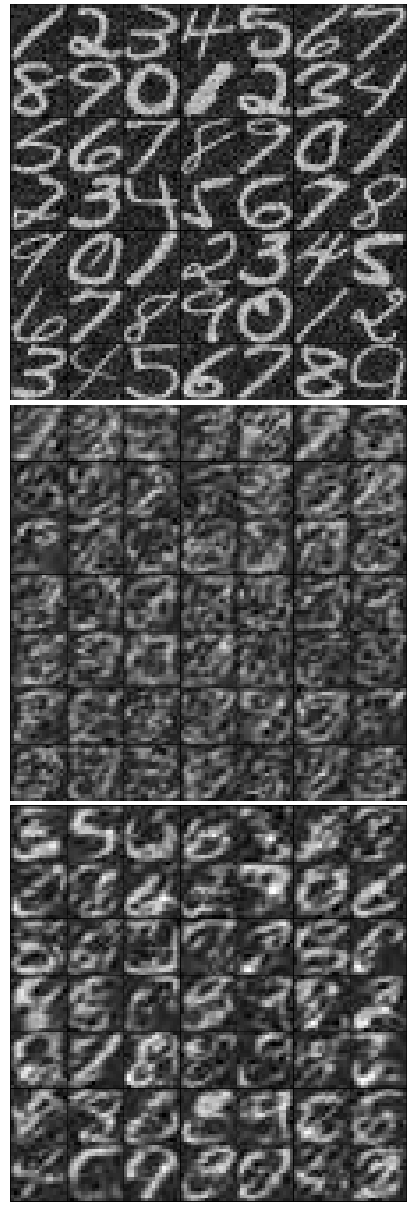

Figure 5: Illustrative results from learning with multiple hidden layers. (a) Examples of training data, corrupted with the same amount of Gaussian noise as assumed during learning. (b) Random selection of examples generated by Gibbs sampling from a model with a single hidden layer. (c) Random selection of examples generated by Gibbs sampling from a model with a hierarchy of two MRF layers. Each MRF layer had 256 fully interconnected hidden units, and there was full directed connectivity from the top MRF layer to the middle MRF layer, as well as full directed connectivity from the middle MRF layer to the observables.

Our proposed learning method is able to take advantage of this fact, and to work within a space of models for which factorial inference is more accurate than it would be able to be if directed connections alone were used. This point is illustrated rather nicely by some of the lateral connections in Figure 4. The lateral interactions tend to cancel out the correlations in the posterior that would be introduced by explaining-away. Consider, for example, two hidden units such as 6H and 3G in Figure 4 that have highly anti-correlated weights on their causal connections. If both these units turn on together the image will be unchanged, so explaining-away would make their activities be strongly positively correlated in the posterior. By learning a strongly negative lateral interaction, the network manages to make them approximately independent in the posterior thus making the variational inference work well.

The idea of using a complicated prior distribution in order to achieve approximate independence in the posterior is a very different approach from Independent Components Analysis (ICA) [11, 12] which assumes independence in the prior and therefore gives rise to awkward posteriors when there are more hidden variables than observables.

10 Summary & Discussion

We have presented a learning procedure for training models that contain both directed and undirected connections; in particular we have focused on large densely connected MRF's that are linked to either observables or other MRF's via directed (causal) connections. Learning in such models is generally intractable, and so the learning task necessitates approximations. Our proposed method combines variational techniques with the contrastive divergence algorithm.

Whilst initial results are promising, there is clearly much more work to be done in developing more sophisticated approximation schemes and in exploring different model architectures for different types of problem. In addition to the approximation methods we have developed in this paper, there are other schemes that may be useful and indeed could be combined with our approach. One could, for instance, consider running our method until convergence and then using this solution as the starting point for a much slower, but potentially more accurate approach that uses Monte Carlo methods. Alternatively, the learned recognition model parameters could be used to initialise further learning using a version of the wake-sleep algorithm [13].

There are many domains in which hybrid models such as the ones we have presented here might be useful, and we hope that our suggested approximation techniques open up avenues for exploration.

Acknowledgements This research was funded by NSERC and CFI. We thank Peter Dayan, Zoubin Ghahramani, Javier Movellan, Sam Roweis, Terry Sejnowski, Yee Whye Teh, Max Welling and Rich Zemel for helpful discussions. GEH holds a Canada Research Chair and is a fellow of CIAR.

References

[1] M. Welling and G. E. Hinton. A new learning algorithm for mean field boltzmann machines. In *Proc. International Conference on Artificial Neural Networks*, pages 351–357. 2002.

[2] R. M. Neal and G. Hinton. A view of the EM algorithm that justifies incremental, sparse, and other variants. In M. I. Jordan, editor, *Learning in Graphical Models*, pages 355–368. Kluwer Academic Publishers, 1998.

[3] G. E. Hinton. Training products of experts by minimizing contrastive divergence. *Neural Computation*, 14:1771–1800, 2002.

[4] S. Lauritzen and N Wermuth. Graphical models for associations between variables, some of which are qualitative and some quantitative. *Annals of Statistics*, 17:31–57, 1989.

[5] W. L. Buntine. Chain graphs for learning. In *Uncertainty in Artificial Intelligence*, pages 46–54, 1995.

[6] S. L. Lauritzen and T. S. Richardson. Chain graphs and their causal interpretations. *Journal of the Royal Statistical Society. Series B. Statistical Methodology*, 64(3):321–361, 2002.

[7] Y. W. Teh and G. E. Hinton. Rate-coded restricted boltzmann machines for face recognition. In *Advances in Neural Information Processing Systems 13*. 2001.

[8] M. Weber, M. Welling, and P. Perona. Unsupervised learning of models for recognition. In *Proc. 6th European Conference on Computer Vision*, 2000.

[9] D.D. Lee and H. S. Seung. Algorithms for non-negative matrix factorization. In *Advances in Neural Information Processing Systems 13*. 2001.

[10] J. Pearl. *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference*. Morgan Kaufmann Publishers Inc., 1988.

[11] A. J. Bell and T. J. Sejnowski. An information maximisation approach to blind separation and blind devonvolution. *Neural Cmputation*, 7:1129–1159, 1995.

[12] J.F. Cardoso. Infomax and maximum likelihood for blind source separation. *IEEE Signal Processing Letters*, 4:112–114, 1997.

[13] G.E. Hinton, P. Dayan, B.J. Frey, and R.M. Neal. The "wake-sleep" algorithm for unsupervised neural networks. *Science*, 268:1158–1160, 1995.