School of Informatics, University of Edinburgh

Institute for Adaptive and Neural Computation

Temporal Hidden Hopfield Models

by

Felix Agakov, David Barber

Informatics Research Report EDI-INF-RR-0156

School of Informatics **November 2002**
http://www.informatics.ed.ac.uk/

Temporal Hidden Hopfield Models

Felix Agakov, David Barber

Informatics Research Report EDI-INF-RR-0156

SCHOOL *of* INFORMATICS
Institute for Adaptive and Neural Computation
November 2002

Abstract :
Many popular probabilistic models for temporal sequences assume simple hidden dynamics or low dimensionality of discrete variables. For higher dimensional discrete hidden variables, recourse is often made to approximate mean field theories, which to date have been applied to models with only simple hidden unit dynamics. We consider a class of models in which the discrete hidden space is defined by parallel dynamics of densely connected high-dimensional stochastic Hopfield networks. For these Hidden Hopfield Models (HHMs), mean field methods are derived for learning discrete and continuous temporal sequences. We discuss applications of HHMs to classification and reconstruction of non-stationary time series. We also demonstrate a few problems (learning of incomplete binary sequences and reconstruction of 3D occupancy graphs) where distributed discrete hidden space representation may be useful. We show that while these problems cannot be easily solved by other dynamic belief networks, they are efficiently addressed by HHMs.

Keywords : Dynamic belief networks, Hidden Markov Models, Hopfield networks, variational learning, missing data, temporal sequences.

Temporal Hidden Hopfield Models

Felix V. Agakov and **David Barber**
Division of Informatics, University of Edinburgh, Edinburgh EH1 2QL, UK
felixa@anc.ed.ac.uk, *dbarber@anc.ed.ac.uk*
http://anc.ed.ac.uk

November 4, 2002

Abstract

Many popular probabilistic models for temporal sequences assume simple hidden dynamics or low-dimensionality of discrete variables. For higher dimensional discrete hidden variables, recourse is often made to approximate mean field theories, which to date have been applied to models with only simple hidden unit dynamics. We consider a class of models in which the discrete hidden space is defined by parallel dynamics of densely connected high-dimensional stochastic Hopfield networks. For these Hidden Hopfield Models (HHMs), mean field methods are derived for learning discrete and continuous temporal sequences. We discuss applications of HHMs to classification and reconstruction of non-stationary time series. We also demonstrate a few problems (learning of incomplete binary sequences and reconstruction of 3D occupancy graphs) where distributed discrete hidden space representation may be useful. We show that while these problems cannot be easily solved by other dynamic belief networks, they are efficiently addressed by HHMs.

1 Markovian Dynamics for Temporal Sequences

Dynamic Bayesian networks are popular tools for modeling temporally correlated patterns. Included in this class of models are Hidden Markov Models (HMMs), auto-regressive HMMs (see e.g. Rabiner (1989)), and Factorial HMMs (Ghahramani and Jordan, 1995). These models are special cases of a generalized Markov chain

$$p(\{\mathsf{h}\},\{\mathsf{v}\}) = p(\mathsf{h}^{(0)})p(\mathsf{v}^{(0)})\prod_{t=0}^{T-1} p(\mathsf{h}^{(t+1)}|\mathsf{h}^{(t)},\mathsf{v}^{(t)})p(\mathsf{v}^{(t+1)}|\mathsf{h}^{(t)},\mathsf{v}^{(t)}), \tag{1}$$

where $\{\mathsf{h}\} = \{\mathsf{h}^{(0)},\dots,\mathsf{h}^{(T)}\}$ and $\{\mathsf{v}\} = \{\mathsf{v}^{(0)},\dots,\mathsf{v}^{(T)}\}$ are hidden and visible variables [see Figure 1 (a)–(c)].

A general procedure for learning the model parameters $\mathbf{\Theta}$ by maximum likelihood training is the EM algorithm, which optimizes a lower bound on the likelihood

$$\Phi(\{\mathsf{v}\};q,\mathbf{\Theta}) = \langle \log p(\{\mathsf{h}\},\{\mathsf{v}\}) + \log q(\{\mathsf{h}\}|\{\mathsf{v}\})\rangle_{q(\{\mathsf{h}\}|\{\mathsf{v}\})} \tag{2}$$

with respect to the parameters [the M-step] and an auxiliary distribution $q(\{\mathsf{h}\}|\{\mathsf{v}\})$ [the E-step]. The bound on the likelihood $\mathcal{L}$ is exact if and only if $q(\{\mathsf{h}\}|\{\mathsf{v}\})$ is identical to the true posterior $p(\{\mathsf{h}\}|\{\mathsf{v}\})$. However, in general, the problem of evaluating the averages over the discrete $p(\{\mathsf{h}\}|\{\mathsf{v}\})$ is exponential in the dimension of $\{\mathsf{h}\}$.

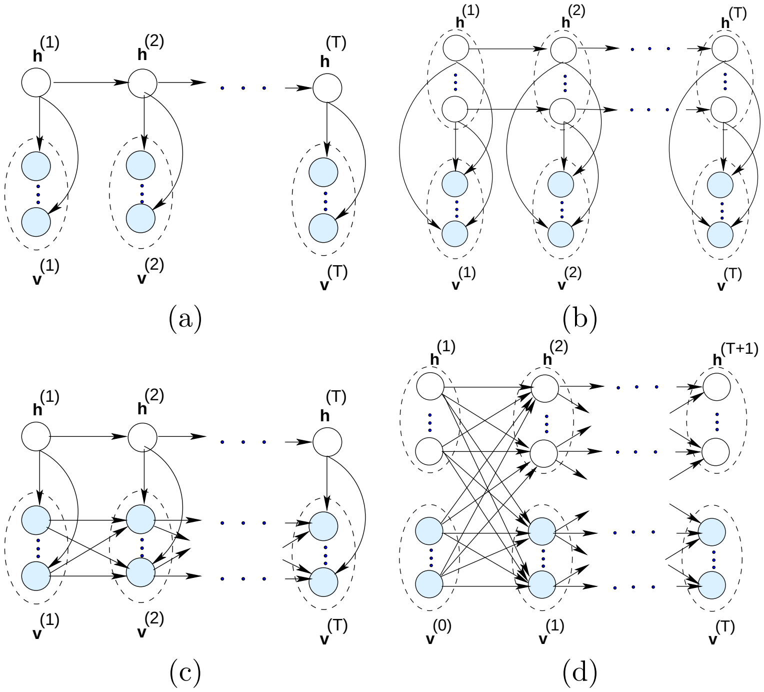


Figure 1: Graphical models for temporal sequences: (a) Hidden Markov Model; (b) Factorial HMM; (c) Auto-regressive HMM; (d) generalized Markov chain with temporally shifted observations.

This computational intractability of learning is one of the fundamental problems of probabilistic graphical modeling. Many popular models for temporal sequences therefore assume that the hidden variables are either very low-dimensional, in which case $\mathcal{L}$ can be optimized exactly (e.g. HMMs), or have very simple temporal dependencies, so that $p(\{\mathsf{h}\}|\{\mathsf{v}\})$ is approximately factorized.

Our work here is motivated by the observation that mean field theories succeed in the contrasting limits of extremely sparse connectivity (models are then by construction approximately factorized), and extremely dense connectivity (for distributions with probability tables dependent on a linear combination of parental states). This latter observation raises the possibility of using mean field methods for approximate learning in dynamic networks with *high dimensional, densely connected* discrete hidden spaces.

The resulting model with a large discrete hidden dimension can be used for learning highly non-stationary data of coupled dynamical systems. Moreover, as we show in section 5, it yields a fully probabilistic way of addressing some problems of image processing (half-toning and binary super-resolution of video sequences) and scanning (3D shape reconstruction). We also demonstrate that the model can be naturally applied to reconstruction of incomplete discrete temporal sequences.

2 Hidden Hopfield Models

To fully specify the model (1) we need to define the transition probabilities $p(\mathsf{h}^{(t+1)}|\mathsf{x}^{(t)})$ and $p(\mathsf{v}^{(t+1)}|\mathsf{x}^{(t)})$, where $\mathsf{x} = [\mathsf{h}^T\ \mathsf{v}^T]^T$. For large models and discrete hidden variables the conditionals $p(h_i^{(t+1)}|\mathsf{x}^{(t)})$ cannot be defined by probability tables, and some parameterization needs to be considered. It should be specified in such a form that computationally tractable approximations of $p(\mathsf{h}^{(t+1)}|\mathsf{x}^{(t)})$ are sufficiently accurate. We consider $h_i^{(t+1)} \in \{-1, +1\}$ and

$$p(h_i^{(t+1)}|\mathsf{x}^{(t)}; \mathsf{w}_i, b_i) = \sigma\left(h_i^{(t+1)}(\mathsf{w}_i^T \mathsf{x}^{(t)} + b_i)\right), \tag{3}$$

where w_i is a weight vector connecting node i with all of the nodes, b_i is the node's bias, and $\sigma(a) = 1/(1 + e^{-a})$.

The model has a graphical structure, temporal dynamics, and parametrization of the conditionals $p(h_i|\mathsf{x})$ similar to a synchronous Hopfield network (e.g. Hertz et al. (1991)) amended with hidden variables and a full generally *non-symmetric* weight matrix. This motivates us to refer to generalized Markov chains (1) with parameterization (3) as a Hidden Hopfield Model (HHM).

Our model is motivated by the observation that, according to the Central Limit Theorem, for large densely connected models without strongly dependent weights, the posteriors (3) are approximately uni-modal. Therefore, the mean field approximation

$$q(\{\mathsf{h}\}|\{\mathsf{v}\}; \boldsymbol{\lambda}) = \prod_k \lambda_k^{(1+h_k)/2}(1 - \lambda_k)^{(1-h_k)/2}, \quad \lambda_k \overset{\text{def}}{=} q(h_k = 1|\{\mathsf{v}\}) \tag{4}$$

is expected to be reasonably accurate. During learning we optimize the bound (2) with respect to this factorized approximation q and the model parameters $\boldsymbol{\Theta} = \{\mathsf{W}, \mathsf{b}, p(\mathsf{h}^{(0)})\}$ for two types of visible variables v. In the first case $\mathsf{v} \in \{-1, +1\}^n$ and the conditionals $p(v_i|\mathsf{x})$ are defined similarly to expression (3). Essentially, this specific case of discrete visible variables is equivalent to sigmoid belief networks (Neal, 1992) with hidden and visible variables in each layer. In the second considered case the observations $\mathsf{v} \in \mathbb{R}^n$ with $p(v_i|\mathsf{x}) \sim \mathcal{N}(\mathsf{w}_i^T \mathsf{x}, s^2)$, where s^2 is the variance of isotropic Gaussian noise. Note that in both cases sparser variants of the generalized chains can be obtained by fixing certain HHM weights at zeros.

Previously, Saul et al. (1996) used a similar approximation for learning in sigmoid belief networks. Their approach suggests to optimize a variational lower bound on Φ, which is itself a lower bound on $\mathcal{L}$. For HHM learning of discrete time series we adopt a slightly different strategy and exploit Gaussianity of the nodes' fields for numeric evaluation of the gradients. An outline of the learning algorithm is given in Appendix A.1. This results in a fast learning rule, which smooths differences between discrete $\{\mathsf{h}\}$ and $\{\mathsf{v}\}$ and makes it easy to learn discrete sequences of irregularly observed data. HHM learning of continuous time series results in a related, but different rule (section 3.1).

Note that although both HMMs and Hidden Hopfield models can be used for learning of non-stationary time series with long temporal dependencies, they fundamentally differ in representations of the hidden spaces. HMMs capture non-stationarities by expanding the number of states of a single multinomial variable. As opposed to HMMs, Hidden Hopfield models have a more efficient, distributed hidden space representation. Moreover, the model allows intra-layer connections between the hidden variables, which yields a much richer hidden state structure compared with Factorial HMMs.

3 Learning in Hidden Hopfield Models

Here we outline the variational EM algorithm for HHMs with continuous observations. Derivations of these results, along with the learning rule for discrete-data HHMs are described in Appendix A.

3.1 EM algorithm

Let $H^{(t)}$, $V^{(t)}$ denote sets of variables hidden or visible at time t, and $x_i^{(t)}$ be the i^{th} variable at time t. For each such variable we introduce an auxiliary parameter $\lambda_i^{(t)}$, such that

$$\lambda_i^{(t)} \overset{\text{def}}{=} \begin{cases} q(x_i^{(t)} = 1|\mathsf{v}^{(t)}) \in [0,1] & \text{if } i \in H^{(t)}; \\ (x_i^{(t)} + 1)/2 \in \mathbb{R} & \text{if } i \in V^{(t)}. \end{cases} \tag{5}$$

Note that in the case when $x_i^{(t)}$ is hidden, $\lambda_i^{(t)}$ is effectively the mean-field parameter of the variational distribution $q(\{\mathsf{h}\}|\{\mathsf{v}\})$ and must be learned from data.

M-step:

Let w_{ij} be connecting $x_i^{(t+1)}$ and $x_j^{(t)}$. Then, as derived in Appendix A,

$$\frac{\partial \Phi}{\partial w_{ij}} = \sum_{t=0}^{T-1} f_i^{(t+1)} \frac{\partial \Phi^v(t)}{\partial w_{ij}} + (1 - f_i^{(t+1)}) \frac{\partial \Phi^h(t)}{\partial w_{ij}}, \tag{6}$$

where

$$\begin{aligned} \frac{\partial \Phi^h(t)}{\partial w_{ij}} \approx \lambda_i^{(t+1)}(2\lambda_j^{(t)} - 1) \quad & - \quad (1 - f_j^{(t)}) \left[\lambda_j^{(t)} \langle \sigma(c_{ij}^t) \rangle_{\mathcal{N}_{ij}^c(t)} + (\lambda_j^{(t)} - 1) \langle \sigma(d_{ij}^t) \rangle_{\mathcal{N}_{ij}^d(t)} \right] \\ & - \quad f_j^{(t)} (2\lambda_j^{(t)} - 1) \langle \sigma(e_i) \rangle_{\mathcal{N}_i^e(t)}, \end{aligned} \tag{7}$$

$$\frac{\partial \Phi^v(t)}{\partial w_{ij}} \approx \frac{1}{s^2} \left((2\lambda_j^{(t)} - 1) \left[v_i^{(t+1)} - \mathsf{w}_i^T (2\boldsymbol{\lambda}^{(t)} - \mathbf{1}) \right] + 4(1 - f_j^{(t)}) w_{ij} (\lambda_j^{(t)} - 1) \lambda_j^{(t)} \right) \tag{8}$$

and $f_j^{(t)} \in \{0, 1\}$ is an indicator variable equal to 1 if and only if x_j is visible at time instance t [i.e. $j \in V^{(t)}$]. The fields c_{ij}^t, d_{ij}^t, and e_i^t are approximately normally distributed according to

$$c_{ij}^t \sim \mathcal{N}_{ij}^d(t) \equiv \mathcal{N} \left(\mathsf{w}_i^T (2\boldsymbol{\lambda}^{(t)} - \mathbf{1}) - 2w_{ij}(\lambda_j^{(t)} - 1) + b_i, 4 \sum_{k \neq j}^{|\mathsf{x}^{(t)}|} \lambda_k^{(t)} (1 - \lambda_k^{(t)}) w_{ik}^2 \right) \tag{9}$$

$$d_{ij}^t \sim \mathcal{N}_{ij}^d(t) \equiv \mathcal{N} \left(\mathsf{w}_i^T (2\boldsymbol{\lambda}^{(t)} - \mathbf{1}) - 2w_{ij}\lambda_j^{(t)} + b_i, 4 \sum_{k \neq j}^{|\mathsf{x}^{(t)}|} \lambda_k^{(t)} (1 - \lambda_k^{(t)}) w_{ik}^2 \right) \tag{10}$$

$$e_i^t \sim \mathcal{N}_i^e(t) \equiv \mathcal{N} \left(\mathsf{w}_i^T (2\boldsymbol{\lambda}^{(t)} - \mathbf{1}) + b_i, 4 \sum_{k=1}^{|\mathsf{x}^{(t)}|} \lambda_k^{(t)} (1 - \lambda_k^{(t)}) w_{ik}^2 \right). \tag{11}$$

Analogously, the derivative w.r.t. the biases $\partial \Phi / \partial b_i$ is given by

$$\frac{\partial \Phi}{\partial b_i} \approx \sum_{t=0}^{T-1} \left[\lambda_i^{(t+1)} - \langle \sigma(e_i^t) \rangle_{\mathcal{N}_i^e(t)} \right]. \tag{12}$$

The resulting averages may be efficiently evaluated by using numerical Gaussian integration, and even crude approximation at the means often leads to good results (see section 5).

E-step:

Optimizing the bound (2) w.r.t. the mean field parameters $\lambda_i^{(t)}$ of non-starting and non-ending *hidden* nodes, we get the fixed point equations of the form $\lambda_k^{(t)} = \sigma(l_k^{(t)})$, where

$$l_k^{(t)} = \tilde{l_k}^{(t)} - \frac{2}{s^2} \sum_{i \in V^{(t+1)}} w_{ik} \left[\mathsf{w}_i^T (2\boldsymbol{\lambda}^{(t)} - \mathbf{1}) - w_{ik}(2\lambda_k^{(t)} - 1) - v_i^{(t+1)} \right], \tag{13}$$

and

$$\begin{aligned} \tilde{l_k}^{(t)} = \mathsf{w}_k^T (2\boldsymbol{\lambda}_k^{(t-1)} - 1) + b_k + \sum_{m \in H^{(t+1)}} \Bigg[& \left\langle \log \left\{ \sigma(c_{mk}^t)^{\lambda_i^{(t+1)}} \sigma(-c_{mk}^t)^{1-\lambda_m^{(t+1)}} \right\} \right\rangle_{\mathcal{N}_{ij}^c(t)} \\ & - \left\langle \log \left\{ \sigma(d_{mk}^t)^{\lambda_m^{(t+1)}} \sigma(-d_{mk}^t)^{1-\lambda_m^{(t+1)}} \right\} \right\rangle_{\mathcal{N}_{ij}^d(t)} \Bigg]. \end{aligned} \tag{14}$$

Here c_{mk}^t and d_{mk}^t are distributed according to the Gaussians (9), (10).

It can be easily seen that the mean field parameter $\pi_k \overset{\text{def}}{=} \lambda_k^{(0)}$ of the starting hidden node can be obtained by replacing the contribution of the previous states $b_k + \mathsf{w}_k^T(2\boldsymbol{\lambda}^{(t-1)} - 1)$ in the r.h.s. of (14) by $\log \left\{ \lambda_k^{(0)} / (1 - \lambda_k^{(0)}) \right\}$. Finally, since $\mathsf{h}^{(T)}$ is unrepresentative of the data (see figure 1), the mean field parameters $\lambda_i^{(T-1)}$ of the ending nodes are obtained from (13) by setting $\tilde{l_k}^{(t)} = \mathsf{w}_k^T(2\boldsymbol{\lambda}_k^{(t-1)} - 1) + b_k$.

3.2 Multiple sequences

To learn multiple sequences we need to estimate separate mean field parameters $\{\lambda_{ks}^{(t)}\}$ for each node k of time series s at $t > 0$. This does not change the fixed point equations of the E-step of the algorithm. From expression (2) it is clear that the gradients $\partial\Phi/\partial w_{ij}$ and $\partial\Phi/\partial b_i$ in the M-step will be expressed according to (7), (8), (12) [continuous case] and (31), (12) [discrete case] and with an additional summation over the training sequences.

3.3 Annealing

In some cases it may be useful to bias posterior probabilities of the hidden variables $\lambda_i^{(t)}$ toward deterministic values. This may be particularly important if it is known that the true values of the hidden variables, giving rise to the observations, are intrinsically deterministic (see example in section 5.5). Moreover, if several hidden state configurations result in approximately the same visible patterns, we may be interested in learning one of such representations instead of their smooth average, and adapt the parameters accordingly. One way to promote *approximate determinism* of the hidden variables is by following an annealing scheme, so that state fluctuations become less likely as training or inference continues (see e.g. Hertz et al. (1991)). Alternatively, a related effect can be achieved by introducing an inertia factor, so that for each hidden variable $h_i^{(t)}$ with the old estimate of the posterior $\tilde{\lambda}_i^{(t)} \overset{\text{def}}{=} q(\tilde{h}_i^{(t)}|\mathsf{v})$ the transition is defined as

$$p(h_i^{(t)}|\mathsf{h}^{(t-1)}, \tilde{h}_i^{(t)}) \propto p(h_i^{(t)}|\mathsf{h}^{(t-1)})(\tilde{\lambda}_i^{(t)})^{h_i}(1 - \tilde{\lambda}_i^{(t)})^{1-h_i}. \tag{15}$$

This is analogous to introducing a time-variant prior on activation of each hidden variable, which is defined by previous estimates of the mean field parameters and which encourages consistency of hidden unit activations.

From (15) it is easy to see that the resulting forms of the E-step expressions (13) and (28) for the fields $l_k^{(t)}$ are incremented by

$$\gamma_i^{(t)}(\tilde{\lambda}_i^{(t)}) \stackrel{\text{def}}{=} \log \frac{\tilde{\lambda}_i^{(t)}}{(1-\tilde{\lambda}_i^{(t)})}. \tag{16}$$

Note that $\lim_{\tilde{\lambda}_i^{(t)} \to 1} \gamma_i^{(t)} \to \infty$, i.e. units $h_i^{(t)}$ which were likely to be on at the previous iteration of learning give rise to a large positive field contribution $\gamma_i^{(t)}$. Clearly, each such $h_i^{(t)}$ is more likely to stay on at the current iteration, unless there are particularly strong indications (e.g. from the emissions) that it should change under the new values of the parameters $\boldsymbol{\Theta}$ obtained in the M step. Analogously, $\lim_{\tilde{\lambda}_i^{(t)} \to 0} \gamma_i^{(t)} \to -\infty$, i.e. units which were likely to be off will more likely remain off. Finally, note that the contribution (16) to the field of $h_i^{(t)}$ is significant only in deterministic limits of $\tilde{\lambda}_i^{(t)}$ (in practice, it is negligible for $\tilde{\lambda}_i^{(t)} \in [0.05, 0.95]$, with $\gamma_i^{(t)}(1/2) = 0$).

3.4 Constrained parametrization

It is clear that if the model has n binary hidden variables it is capable of representing 2^n states for each time slice. Note, however, that the full transition matrix comprises n^2 weights, which may lead to prohibitively large amounts of training data and high computational complexity of learning. In section 5 we demonstrate ways of imposing neighborhood sparsity constraints on the weight transition and emission matrices so that the number of adaptive parameters is significantly decreased. We also show that while the exact learning and inference in general remain computationally intractable, the Gaussian field approximation remains accurate and leads to good classification and reconstruction results.

4 Inference

A simple way to perform inference (estimation of the posterior probability $p(\{\mathsf{h}\}|\{\mathsf{v}\})$) is by clamping the observed sequence $\{\mathsf{v}\}$ on the visible variables, fixing the model parameters $\boldsymbol{\Theta}$ and performing the E-step of the variational EM algorithm described in section 3. This results in a set of mean-field parameters $\{\lambda_k^{(t)}\}$, which can be used for obtaining a hidden space representation of the sequence.

Alternatively, we can draw samples from $p(\{\mathsf{h}\}|\{\mathsf{v}\})$ by using Gibbs sampling. We can make it more efficient by utilizing the *red-black* scheme, where we first condition on the odd layers of a high-dimensional chain and sample nodes in the even layers in parallel, and then flip the conditioning (all the visible variables are assumed to stay fixed). Sampling from $p(\mathsf{x}^{(t)}|\mathsf{x}^{(t-1)}, \mathsf{x}^{(t+1)})$ cannot be performed directly, since $p(\mathsf{x}^{(t)}|\mathsf{x}^{(t-1)}, \mathsf{x}^{(t+1)}) \propto p(\mathsf{x}^{(t)}|\mathsf{x}^{(t-1)})p(\mathsf{x}^{(t+1)}|\mathsf{x}^{(t)}))$ cannot be easily normalized for large-scale models. In general we may need to use another Gibbs sampler for hidden components of $\mathsf{x}^{(t)}$, which results in

$$\begin{aligned} x_i^{(t)} \quad \leftarrow \quad & p(x_i^{(t)} = 1|\mathsf{x}^{(t-1)}, \mathsf{x}^{(t+1)}, \mathsf{x}^{(t)} \backslash x_i^{(t)}) = \\ & \sigma \left\{ b_i + \mathsf{w}_i^T \mathsf{x}^{(t-1)} + \sum_{j=1}^{|\mathsf{h}^{(t+1)}|} \log \frac{\sigma\left(x_j^{(t+1)}\left[\mathsf{w}_i^T \mathsf{x} - w_{ji} x_i + w_{ji}\right]\right)}{\sigma\left(x_j^{(t+1)}\left[\mathsf{w}_i^T \mathsf{x} - w_{ji} x_i - w_{ji}\right]\right)} \right. \\ & \left. + \frac{2}{\sigma^2} \sum_{k=1}^{|\mathsf{v}^{(t+1)}|} w_{ki}\left(v_k^{(t+1)} - \mathsf{w}_k^T \mathsf{x}^{(t)} + w_{ki} h_i^{(t)}\right) \right\} \end{aligned} \tag{17}$$

for continuous HHMs (see Appendix A.2).

Expression (17) is exactly equivalent to the E-step update (13), (14) where the current estimates of the mean field parameters $\{\boldsymbol{\lambda}\}$ are used instead of the current values of the hidden variables $\{\mathbf{h}\}$, and the intractable Gaussian averages of (14) are approximated at the mean values of the arguments.

5 Experimental results

Here we briefly describe applications of HHMs to reconstruction and classification of incomplete non-stationary discrete temporal sequences. We also present a toy problem of learning a binary video sequence (transformation of a digit) from incomplete noisy data and inference of its missing fragments. Finally, we apply constrained continuous-data HHMs to two toy problems of video halftoning (constrained compression) and 3D shape reconstruction, which cannot be easily addressed by other known probabilistic graphical models.

5.1 Reconstruction of discrete sequences

One way to validate correctness of the HHM learning rule is by performing deterministic reconstruction of learned temporal sequences from noiseless initializations at the starting patterns. For discrete sequences we expect such reconstructions to be good if there are sufficiently many hidden variables to capture long temporal dependencies, i.e. if the total number of nodes is of the same order as $s \times T$ (the number of training sequences and their length respectively).

Figures 2 (a), (b) illustrate reconstruction of a 7-d discrete time series of length 15, performed by a network with 7 visible and 3 hidden units[1]. The initial training pattern $\mathsf{v}^{(0)}$ was set at uniform random, and each subsequent observation vector $\mathsf{v}^{(t+1)}$ was generated from $\mathsf{v}^{(t)}$ by flipping each bit with probability 0.2 [Figure 2 (a)]. The model parameters $\boldsymbol{\Theta}$ were learned by the EM algorithm (section 3). The reconstructed sequence was generated from the initial state $\mathsf{x}^{(0)}$, sampled from the learned prior $p(\mathsf{x}^{(0)})$, by deterministically setting subsequent variables $x_i^{(t+1)}$ according to $sgn(\sigma(x_i^{(t+1)}(\mathsf{w}_i^T\mathsf{x}+b_i))-1/2)$ [Figure 2 (b)]. Note that without hidden variables deterministic reconstruction of the visible training sequence would be impossible. Indeed, patterns $\mathsf{v}^{(7)}$ and $\mathsf{v}^{(8)}$ are identical, and it is the learned activation of the hidden variables $\mathsf{h}^{(7)}$ and $\mathsf{h}^{(8)}$ which distinguishes mapping $\mathsf{x}^{(7)} \to \mathsf{x}^{(8)}$ from $\mathsf{x}^{(8)} \to \mathsf{x}^{(9)}$.

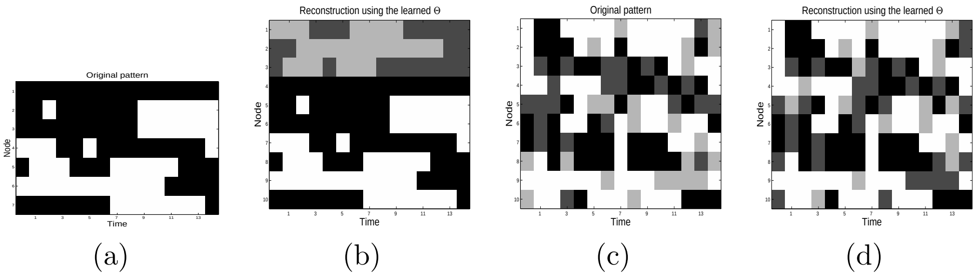
Figures 2 (c), (d) show a variant of the previous experiment for a discrete 10-d time series with irregularly missing data. It is pleasing that the model perfectly reproduces the visible patterns, although nothing in the framework explicitly suggests perfect reconstruction of the hidden variables from noisy initialization at the starting visible state.

Note that in order to deterministically reconstruct a binary sequence in HHMs it is sufficient to ensure that each data point $v_i^{(t)}$ maps into $v_i^{(t+1)}$ lying on the correct side of the hyperplane $(\mathsf{w}_i; b_i)$. Reconstruction of continuous temporally correlated patterns is in general more complex.

5.2 Classification of discrete sequences

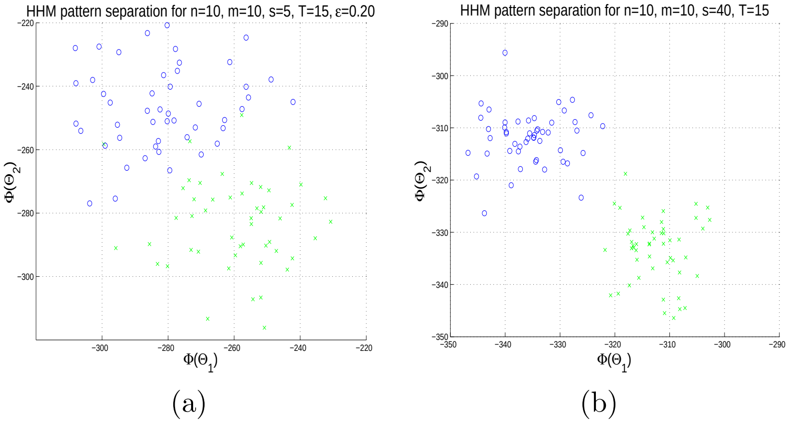
In large scale HHMs computation of the likelihood of a sequence is intractable. One possible discriminative criterion for classification of a given new sequence $\mathsf{v}^\star$ is the lower bound on the likelihood $\Phi(\mathsf{v}^\star; q^\star, \boldsymbol{\Theta})$ given by expression (2). Once HHM parameters $\boldsymbol{\Theta}$ are optimized for the training set $\{\mathsf{v}\}$, the distribution $q^\star$ may be evaluated by fixing $\boldsymbol{\Theta}$, clamping $\mathsf{v}^\star$ on the visible variables, and performing just the E-step of the algorithm.

[1] From now on we imply multiplication of the network size by the sequence length T.



(a) (b) (c) (d)

Figure 2: Training and reconstructed sequences with regular (a), (b) and irregular (c), (d) observations. Black and white squares correspond to -1 and +1 for the visible variables; dark gray and light gray – to -1 and +1 for the hidden variables.



(a) (b)

Figure 3: Temporal sequence classification in HHMs. The circles and crosses show the true class labels of the testing sequences; the axis define approximations of the bounds on their likelihoods for $\mathcal{M}_1$ and $\mathcal{M}_2$. *Discrete data*: (a): testing data is constructed by flipping each bit of each training sequence with probability 0.2; $n = 10, m = 10, T = 15, s = 5$. (b): testing data is drawn from the data-generating processes; $n = 10, m = 10, T = 15, s = 40$.

To demonstrate HHM classification we generated discrete sequences from two noisy non-stationary 15-d Markov processes $\mathcal{C}_1, \mathcal{C}_2$ with the conditionals (3) parametrized by the weights $\mathsf{W} \in \mathbb{R}$. At each time instance the weights were modified according to $\mathsf{W}^{(t+1)} = \mathsf{W}^{(t)} + \beta(r_1^{(t)}\mathsf{A} + r_2^{(t)}\mathsf{B}^{(t)})$, where r_1, r_2 are small random terms, $\mathsf{B}^{(t)}, \mathsf{A}$ are matrices of random elements and β is a small scaling factor. Moreover, at a certain time instance the weights were transformed by a rigid rotation factor. During training we fitted two models $\mathcal{M}_1$, $\mathcal{M}_2$ to 10-dimensional noisy subsets of the data, generated by $\mathcal{C}_1$ and $\mathcal{C}_2$. Each of the models had $n = 10$ visible and $m = 10$ hidden units and was trained on temporal sequences of length $T = 15$.

Figure 3 (a) shows typical approximations of the lower bounds on the likelihoods of 100 testing sequences, generated from the training data by perturbing each bit with probability 0.2. The training set consisted of $s = 5$ sequences for each of the classes. Figure 3 (b) demonstrates a similar plot for the case when $s = 40$, and the testing data was generated by the processes $\mathcal{C}_1$ and $\mathcal{C}_2$. We see that the true labels of the testing data form two reasonably well-separated clusters in the space of approximate likelihoods for models $\mathcal{M}_1$ and $\mathcal{M}_2$. This supports the idea that HHMs can be used for classification of non-stationary temporal sequences and indicates

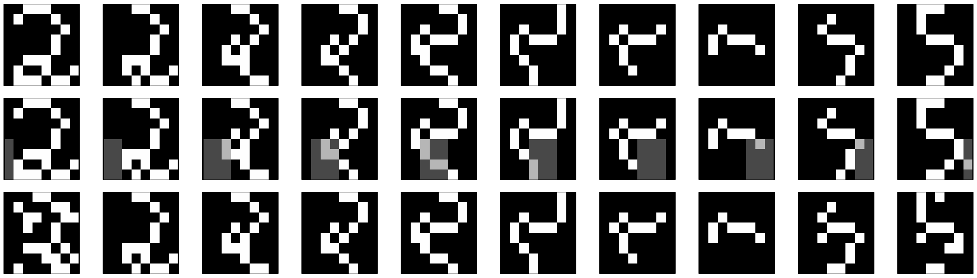


Figure 4: Learning and inference of incomplete temporally correlated patterns. *Top:* the true underlying sequence; *Middle:* the sequence clamped on the visible variables (black/white) with the inferred values of the missing variables (dark/light gray); *Bottom:* the sequence reconstructed from a complete noisy initialization by forward sampling.

that classification is robust both to external random perturbations of the training data and to noise in the generating processes. The experiment of Figure 3 (b) was performed for models with different dimensions m of the hidden space and resulted in approximate probabilities of misclassification $\delta \approx 0.11$ for $m = 1$, $\delta \approx 0.06$ for $m = 5$, and $\delta \approx 0.04$ for $m = 10$ (all results were averaged over a large number of independent runs for the same generating processes). For significantly larger hidden spaces classification was generally worse, what could be explained by overfitting.

We have also performed similar experiments for continuous temporal sequences and found qualitatively similar influence of m on performance of HHM classifiers. These results suggest that HHMs with higher-dimensional hidden spaces are able to capture regimes of non-stationarity and long temporal dependencies, and confirm that the exploited approximations may indeed be sufficiently accurate.

5.3 Learning incomplete discrete sequences

In the previous experiments we have shown that by expanding the HHM hidden space we can increase the length of sequences which can be successfully reconstructed from noisy initializations (by increasing the number of patterns which can be disambiguated), or improve time series classification. In all those cases there was no apparent interpretation of the hidden space or motivation for the choice of the parameterization (3), and in principle other dynamic belief networks could yield comparable performance. In this and the following sections we look at some problems for which a distributed binary hidden state representation arises naturally, and for which HHMs can be models of choice.

In many practical tasks it may often be the case that the data is incomplete, and some variables observed at one time instance are missing at the following time step (e.g. such data could arise as a result of temporally unavailable medical examinations, partial occlusion of images, etc.). Here we demonstrate that HHMs can be easily applied to learning incomplete temporally correlated patterns $\mathsf{x}^{(t)} = [\mathsf{h}^{(t)}\mathsf{v}^{(t)}]$, where the hidden variables $\mathsf{h}^{(t)}$ may be interpreted as a vector of missing observations at time t.

Figure (4) shows an example of applying an HHM to reconstruction of a temporal sequence from its incomplete noisy subsamples. The underlying data contained 10 8×8 binary images with an average Hamming distance of 7 bits between the subsequent patterns (see Figure 4 *top*). The model was trained on 4 sequences generated from the complete series by randomly omitting approximately 15% and permuting approximately 5% of each subsequent pattern. At the reconstruction stage the visible part of the sequence with different, systematically missing observations, was clamped on the HHM's visible units, and the missing observations were inferred variationally. As we see from Figure 4 *middle*, the resulting reconstruction is reasonably accurate.

We have also tried to retrieve the underlying sequence by deterministic forward sampling (Figure 4 *bottom*) as described in section (5.1). The model was initialized at the completely visible starting pattern $\mathsf{x}^{(0)}$ perturbed with 10% noise. Each subsequent pattern $x_i^{(t+1)}$ was set according to $sgn(\sigma(x_i^{(t+1)}(\mathsf{w}_i^T\mathsf{x}^{(t)} + b_i)) - 1/2)$ and perturbed with additional 10% noise. We see that the underlying sequence can still be retrieved relatively accurately, though further experiments show that this reconstruction proves to be sensitive to the noise of the training sequences.

5.4 Constrained HHMs for sequence halftoning

As we noted in section 3.4, learning all $(m+n)\times(m+n)$ weights could often be prohibitive. One way to circumvent this problem is to impose sparsity constraints on the weight matrix, so that the transitions and emissions are defined by a small subset of the full weight matrix. In addition to decreasing computational effort and reducing the required amount of training data, carefully imposed constraints may yield a clear topological interpretation of the hidden variables.

Consider, for example, a special case of a generalized Markov chain with the joint distribution given by

$$p(\{\mathsf{h}\},\{\mathsf{v}\}) = p(\mathsf{h}^{(0)})\prod_{t=0}^{T-1} p(\mathsf{h}^{(t+1)}|\mathsf{h}^{(t)})p(\mathsf{v}^{(t+1)}|\mathsf{h}^{(t)}). \tag{18}$$

Graphically, the model corresponds to an HMM with a high-dimensional distributed hidden space representation. For some types of data (e.g. video sequences) it may be natural to assume that points which are spatially close to each other belong to the same object, and their colors are marginally dependent (assuming the objects are reasonably smoothly colored). On the other hand, colors of spatially distant points are likely to be marginally independent. From a single time slice of the model (18) it is clear that visible variables are marginally dependent if they share a common ancestor. By arranging parents sharing common children to be spatially close to each other in the hidden space, we can model *smoothness of images* by imposing local neighborhood constraints on the emission weights (so that each hidden parent is connected only to a small spacial neighborhood in the visible space, and each visible node is a direct offspring of a spacial neighborhood in the hidden space). Moreover, from the graphical structure it is clear that smooth transitions in the hidden space imply *smoothness of dynamical changes* in observations and vice versa, yielding local neighborhood constraints on the transition weights. In the extreme case of factorial transitions, marginal dependencies of the visible variables are time-invariant.

In the following experiments we consider the problem of *sequence halftoning*, performed by an HHM with local neighborhood constraints on the transition and emission weights. The problem is well known in industrial image processing and involves reduction of a stream of color images to a recognizable monochrome representation. A Hidden Hopfield model with 15×15 hidden states was trained on a sequence of 8 10×10 frames (Figure 5 *top*). The transition weights were constrained in such a way that each hidden variable $h_i^{(t)}$ at time $t > 1$ was linked to $h_i^{(t-1)}$ and 8 of its closest neighbors [i.e. to all the nodes within a 3×3 spacial neighborhood]. Analogously, each hidden variable was connected to a 3×3 spacial neighborhood in the visible space. Each weight w_{ij} was initialized as a "Mexican hat"-type radial basis function of the

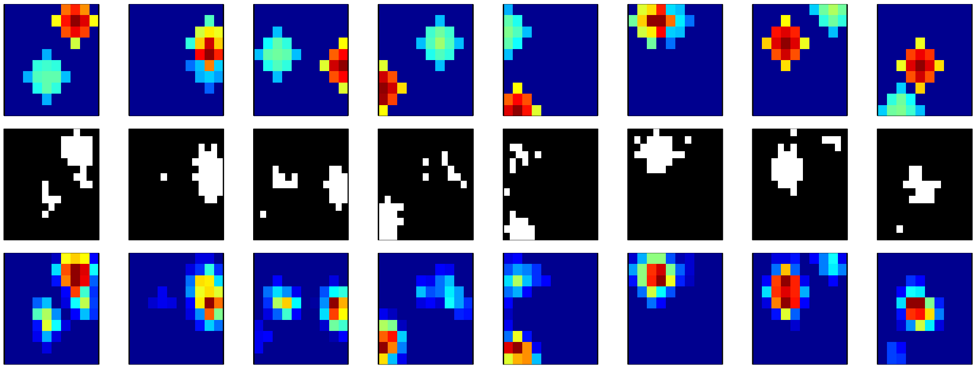


Figure 5: Halftoning of a continuous sequence. *Top:* true sequence; *Middle:* hidden space representation; *Bottom:* reconstructed visible sequence. Learning and inference were performed with the inertia factor.

topological distance between the linked nodes, defined within the range (-0.05, 0.1). The biases were initialized to -1, and both the biases and non-zero weights were fine-tuned by training. The variance of the Gaussian noise was set as $\sigma^2 = 1$.

Figure 5 *middle* shows a sample from the posterior distribution inferred variationally by clamping the continuous data on the visible variables and performing the E-step. Notice that color balls in the visible space correspond to clusters of activation in the hidden space, and density of each cluster roughly corresponds to intensity of the balls. The visible sequence emitted by the hidden variables is shown on Figure 5 *bottom*.

It is worth mentioning that if the visible sequence is fully observed (like in the considered example), temporal contribution to the posterior (17) is likely to be overweighted by the contribution from the continuous observations. However, if a continuous pattern $\mathsf{v}^{(t)}$ is incomplete then temporal information is important, and knowledge of the previous and future states $\mathsf{h}^{(t-1)}$, $\mathsf{h}^{(t+1)}$ may be essential for accurate inference of $\mathsf{h}^{(t)}$.

We believe that the experiment demonstrates potential applicability of constrained HHMs to inferring and learning temporal topographic mappings. Note that unlike constrained HMMs (Roweis, 2000) or temporal GTMs (Bishop et al., 1997), HHMs benefit from the distributed hidden space representation, high dimensionality of the hidden space, and simple definition of the transition probabilities. At this stage it remains unclear which conditions must be satisfied for HHM visualization to be good in general. We are currently investigating this and related issues.

5.5 Constrained HHMs for shape reconstruction

In the last experiment we demonstrate application of a constrained HHM to reconstruction of a 3D occupancy graph from a sequence of weighted 1D projections.

Imagine an object moving with uniform speed orthogonally to the scanning plane spanned by two mutually perpendicular linear scanners. The task is to infer the original shape of the object from a temporal sequence of scanner measurements. In the simplest case a real-life object may be described by its occupancy graph defined by a number of filled or empty discrete cells, and the scanner measurements can be given by the number of the filled cells along each line slice. The depth of each cell is given by the speed per unit time, divided by the frequency of the scans.

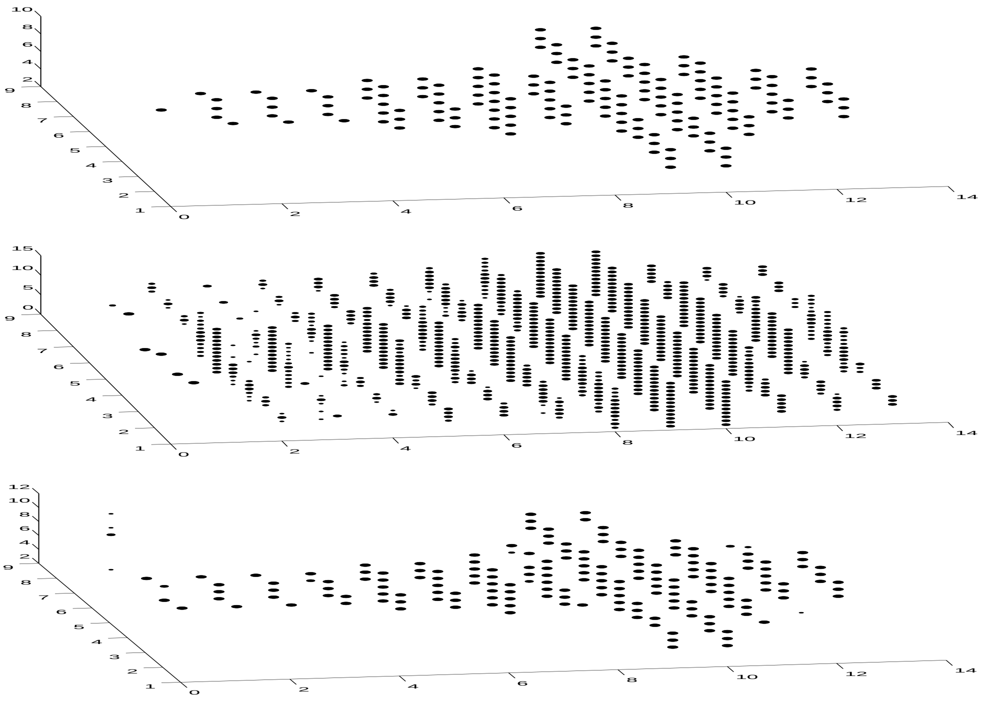


Figure 6: 3D shape reconstruction. The smaller axes define the scanning plane. The disk radii are set proportional to square roots of the posterior probabilities. *Top:* the true shape; *Middle:* the reconstructed shape (no inertia factor); *Bottom:* the reconstructed shape with the inertia factor.

Figure 6 illustrates application of a constrained HHM (18) with 12×9 hidden and 12+9 visible variables to shape reconstruction from 13 consecutive frames. The hidden variables correspond to the binary occupancy cells, while the visible variables represent noisy scanner measurements ($\sigma^2 = 1$). As in section 5.4, the transition weights were set according to the local neighborhood constraints (justified by the presumed smoothness of scanned objects) and fixed at 0.2 (or at 0 outside the region). The emission weights connecting $v_i^{(t)}$ with $\mathsf{h}^{(t)}$ were set to 0.6 (or 0) to perform summations only along the i^{th} row ($i = 1 \dots 12$) or the i^{th} column ($i = 13 \dots 21$) of the discretized slice of the scanned space at time t. The biases of the transition probabilities were set to 0.

From Figure 6 *bottom* we see that impervious to the fact that the scanning data is noisy and the inference task is severely under-defined (more than 100 hidden variables with only 21 visible data points), the constrained HHM with the inertia factor can reasonably accurately reconstruct the underlying shape – a hunting knife (Figure 6 *top*). Note that since the true hidden variables giving rise to the observations are intrinsically deterministic, exclusion of the inertia factor (15) leads to much vaguer posteriors (Figure 6 *middle*). The performance could possibly be improved by fine-tuning the biases and non-zero weights, analogously to what is described in section 5.4. The results suggest that constrained versions of generalized Markov chains (e.g. HHMs with local neighborhood constraints on the weights, factorial Hidden Hopfield Models – HHMs with islands of transitional discontinuity, etc.), while still intractable, may be practical for learning or inferring inherently smooth or constrained data.

It is important to note that in spite of neighborhood or sparsity constraints the resulting

HHMs very quickly become intractable. For the 3×3 topological neighborhoods, the field contributions could be calculated exactly (though, as we see, the Gaussian field approximation also yields accurate predictions). However, the complexity of exact computations grows exponentially with the size of the neighborhood, and for the next closest 5×5 square region (2^{25} possible parental states) an approximation should be used for most practical purposes. This suggests possible combinations of exact and approximate methods in sparse HHMs.

6 Summary

Learning temporal sequences with discrete hidden units is typically achieved using only low dimensional hidden spaces due to the exponential increase in learning complexity with the hidden unit dimension. Motivated by the observation that mean field methods work well in the counter-intuitive limit of a large, densely connected graph with conditional probability tables dependent on a linear combination of parental states, we formulated the Hidden Hopfield Model for which the hidden unit dynamics is specified precisely by a form for which mean field theories may be accurate in large scale systems. For discrete or continuous observations, we derived fast EM-like algorithms exploiting mean and Gaussian field approximations, and demonstrated successful applications to classification and reconstruction of non-stationary and incomplete correlated temporal sequences. We have also discussed learning and inference applications of the constrained HHMs, which may be useful for learning smooth data. The models can be modified to allow other types of emission probabilities $p(v_i^{(t+1)}|\mathsf{v}^{(t)}, \mathsf{h}^{(t)})$, e.g. mixtures of Gaussians, or extended to handling mixed discrete and continuous observations.

A Appendix: Derivations

Here we briefly outline derivation of the variational EM algorithm and the block Gibbs sampling scheme discussed in section 3. The derivations are quite straight-forward; we include them here for completeness.

A.1 Variational EM algorithm

Let $x_i^{(t)}$ be the i^{th} variable of an HHM-induced chain at time t. For each such variable we introduce an auxiliary parameter $\lambda_i^{(t)}$, such that $\lambda_i^{(t)} \overset{\text{def}}{=} q(x_i^{(t)} = 1) \in [0, 1]$ if $x_i^{(t)}$ is hidden, and

$$\lambda_i^{(t)} \overset{\text{def}}{=} (x_i^{(t)} + 1)/2 \tag{19}$$

if $x_i^{(t)}$ is observable[2]. If x_i is hidden the parameter $\lambda_i^{(t)}$ must be learned from data, which is equivalent to the E-step of the variational EM algorithm. Otherwise, $\lambda_i^{(t)}$ is necessarily deterministic with $\lambda_i^{(t)} \in \{0, 1\}$ for discrete and $\lambda_i^{(t)} \in \mathbb{R}$ for continuous observations.

A.1.1 Discrete observations

It is intuitively clear that the parameters $\lambda_i^{(t)}$ corresponding to the visible variables need to remain fixed, which is equivalent to clamping training sequences on the visible nodes. There are no other principle differences between treating binary hidden and binary visible variables in the considered framework.

[2] To simplify the notation we let $\mathsf{x}^T \overset{\text{def}}{=} [\mathsf{h}^T \mathsf{v}^T]$ and $q(\{\mathsf{x}\}) \overset{\text{def}}{=} q(\{\mathsf{h}\}|\{\mathsf{v}\})$.

E-step:

The E-step involves optimization of the lower bound on the likelihood w.r.t. the parameters of the variational distribution $q(\{\mathsf{h}\})$.

From (1) and (2) it is clear that

$$\log p(\{\mathsf{v}\}|\mathsf{h}^{(0)}, \mathsf{v}^{(0)}) \geq \sum_{i=1}^{|\{\mathsf{h}\}|} H(q_i(h_i)) + \sum_{t=0}^{T-1} \langle \log p(\mathsf{v}^{(t+1)}, \mathsf{h}^{(t+1)}|\mathsf{v}^{(t)}, \mathsf{h}^{(t)}) \rangle_{q(\mathsf{h}^{(t)}, \mathsf{h}^{(t+1)})} \tag{20}$$

where $H(q_i(h_i))$ is the entropy of the mean field distribution of $q_i(h_i)$. Differentiation w.r.t. $q(h_k^{(t)})$ leads to

$$q(h_k^{(t)}) \propto \exp \left\{ \langle \log p(h_k^{(t)}|\mathsf{x}^{(t-1)}) \rangle_{q(\mathsf{h}^{(t-1)})} + \sum_{i \in V^{(t+1)}} \langle \log p(v_i^{(t+1)}|\mathsf{x}^{(t)}) \rangle_{q(\mathsf{h}^{(t)} \backslash h_k^{(t)})} \right.$$
$$\left. + \sum_{m \in H^{(t+1)}} \langle \log p(h_m^{(t+1)}|\mathsf{x}^{(t)}) \rangle_{q(h_m^{(t+1)}, \mathsf{h}^{(t)} \backslash h_k^{(t)})} \right\} \tag{21}$$

for each non-starting and non-ending hidden node $h_k^{(t)}$, where $V^{(t+1)}$ and $H^{(t+1)}$ are sets of nodes visible and hidden at $t+1$. From the parameterization given by (3) it is easily derived that $\lambda_k^{(t)} \overset{\text{def}}{=} q(h_k^{(t)} = 1) = \sigma(l_k)$, where

$$l_k^{(t)} = \mathsf{w}_k^T (2\boldsymbol{\lambda}^{(t-1)} - 1) + b_k + \sum_{i \in V^{(t+1)}} \left[\langle \log \sigma(v_i^{(t+1)} c_{ik}^t) \rangle_{p(c_{ik}^t)} - \langle \log \sigma(v_i^{(t+1)} d_{ik}^t) \rangle_{p(d_{ik}^t)} \right]$$
$$+ \sum_{m \in H^{(t+1)}} \left[\left\langle \log \left\{ \sigma(c_{mk}^t)^{\lambda_i^{(t+1)}} \sigma(-c_{mk}^t)^{1-\lambda_m^{(t+1)}} \right\} \right\rangle_{p(c_{mk}^t)} \right.$$
$$\left. - \left\langle \log \left\{ \sigma(d_{mk}^t)^{\lambda_m^{(t+1)}} \sigma(-d_{mk}^t)^{1-\lambda_m^{(t+1)}} \right\} \right\rangle_{p(d_{mk}^t)} \right] \tag{22}$$

and the *fields* c_{mk}^t and d_{mk}^t are given by

$$c_{mk}^t = \mathsf{w}_m^T \mathsf{x}^{(t)} + b_m | h_k^{(t)} = 1, \tag{23}$$
$$d_{mk}^t = \mathsf{w}_m^T \mathsf{x}^{(t)} + b_m | h_k^{(t)} = -1. \tag{24}$$

Since c_{mk}^t and d_{mk}^t are given by linear combinations of random variables, the Central Limit Theorem implies approximate Gaussianity of $p(c_{mk}^t)$ and $p(d_{mk}^t)$ (Barber and Sollich, 2000) with the means $\mu_{mk}^c(t)$, $\mu_{mk}^d(t)$ trivially given by

$$\mu_{mk}^d(t) = \mathsf{w}_m^T (2\boldsymbol{\lambda}^{(t)} - \mathbf{1}) - 2 w_{mk} \lambda_k^{(t)} + b_m, \tag{25}$$
$$\mu_{mk}^c(t) = \mu_{mk}^d(t) + 2 w_{mk} \tag{26}$$

and the variances

$$\begin{aligned} s_{mk}^d(t) = s_{mk}^c(t) &= \left\langle \left[\sum_{i \neq k} w_{mi} (x_i^{(t)} - \langle x_i^{(t)} \rangle) \right]^2 \right\rangle \\ &= \sum_{i,j \neq k} w_{mi} w_{mj} \left(\langle x_l^{(t)} x_j^{(t)} \rangle - \langle x_l^{(t)} \rangle \langle x_j^{(t)} \rangle \right) \\ &= 4 \sum_{i \neq k} w_{mi}^2 (1 - \lambda_i^{(t)}) \lambda_i^{(t)}. \end{aligned} \tag{27}$$

Here we have used the mean field approximation $\langle x_i x_j \rangle = \langle x_i \rangle \langle x_j \rangle$ and the fact that $\langle x_j^2 \rangle = 1$.

Thus, the field update (22) may be re-written as

$$\begin{aligned} l_k^{(t)} \approx \mathsf{w}_k^T(2\boldsymbol{\lambda}^{(t-1)} - \mathbf{1}) + b_k + \sum_{i=1}^{|\mathsf{x}^{(t)}|} \left\langle \log\left\{ \sigma(c_{ik}^t)^{\lambda_i^{(t+1)}} \sigma(-c_{ik}^t)^{1-\lambda_i^{(t+1)}} \right\} \right\rangle_{\mathcal{N}_{ik}^c(t)} \\ - \sum_{i=1}^{|\mathsf{x}^{(t)}|} \left\langle \log\left\{ \sigma(d_{ik}^t)^{\lambda_i^{(t+1)}} \sigma(-d_{ik}^t)^{1-\lambda_i^{(t+1)}} \right\} \right\rangle_{\mathcal{N}_{ik}^d(t)}, \end{aligned} \tag{28}$$

reducing to

$$\lambda_k^{(t)} \approx \sigma \left\{ b_k + \mathsf{w}_k^T(2\boldsymbol{\lambda}^{(t-1)} - \mathbf{1}) + \sum_{i=1}^{|\mathsf{x}^{(t)}|} \left[\log \frac{\sigma(\mu_{ik}^c(t))}{\sigma(\mu_{ik}^d(t))} - 2w_{ik}(1 - \lambda_i^{(t+1)}) \right] \right\} \tag{29}$$

when the integrals are approximated at the means.

M-step:

The M-step optimizes the lower bound on the likelihood w.r.t. the parameters W and b of the conditional distributions (3).

Let w_{ij} be the weight connecting $x_i^{(t+1)}$ and $x_j^{(t)}$. From (20) and (3) we get

$$\begin{aligned} \frac{\partial \Phi}{\partial w_{ij}} &= \frac{\partial}{\partial w_{ij}} \sum_{t=0}^{T-1} \sum_{k=1}^{|\mathsf{x}^{(t+1)}|} \left\langle \log\left\{ \sigma\left(x_k^{(t+1)} [\mathsf{w}_k^T \mathsf{x}^{(t)} + b_k] \right) \right\} \right\rangle_{q(x_k^{(t+1)}, \mathsf{x}^{(t)})} \\ &= \sum_{t=0}^{T-1} \left\langle \left(1 - \sigma\left(x_i^{(t+1)} [\mathsf{w}_i^T \mathsf{x}^{(t)} + b_i] \right) \right) x_i^{(t+1)} x_j^{(t)} \right\rangle_{q(x_i^{(t+1)}, \mathsf{x}^{(t)})} \\ &= \sum_{t=0}^{T-1} \left[(2\lambda_j^{(t)} - 1)\lambda_i^{(t+1)} - \left\langle x_j^{(t)} \sigma\left(\mathsf{w}_i^T \mathsf{x}^{(t)} + b_i \right) \right\rangle_{q(\mathsf{x}^{(t)})} \right]. \end{aligned} \tag{30}$$

Once again applying the Central Limit Theorem, we obtain

$$\frac{\partial \Phi}{\partial w_{ij}} \approx \sum_{t=0}^{T-1} \left\{ \lambda_i^{(t+1)}(2\lambda_j^{(t)} - 1) - \left[\lambda_j^{(t)} \langle \sigma(c_{ij}^t) \rangle_{\mathcal{N}_{ij}^c(t)} + (\lambda_j^{(t)} - 1) \langle \sigma(d_{ij}^t) \rangle_{\mathcal{N}_{ij}^d(t)} \right] \right\} \tag{31}$$

with the moments of $\mathcal{N}_{ij}^c(t)$ and $\mathcal{N}_{ij}^d(t)$ given by (25) – (27). Analogously,

$$\begin{aligned} \frac{\partial \Phi}{\partial b_i} &= \sum_{t=0}^{T-1} \left\langle \left(1 - \sigma(x_i^{(t+1)} [\mathsf{w}_i^T \mathsf{x}^{(t)} + b_i]) \right) x_i^{(t+1)} \right\rangle_{q(x_i^{(t+1)}, \mathsf{x}^{(t)})} \\ &= \sum_{t=0}^{T-1} \left[\lambda_i^{(t+1)} - \langle \sigma(\mathsf{w}_i^T \mathsf{x}^{(t)} + b_i) \rangle_{q(\mathsf{x}^{(t)})} \right] \approx \sum_{t=0}^{T-1} \left[\lambda_i^{(t+1)} - \langle \sigma(e_i^t) \rangle_{\mathcal{N}_i^e(t)} \right] \end{aligned} \tag{32}$$

where $\mathcal{N}_i^e(t)$ is a Gaussian with the mean and the variance

$$\mu_i^e(t) = \mathsf{w}_i^T(2\boldsymbol{\lambda}^{(t)} - \mathbf{1}) + b_i, \tag{33}$$

$$s_i^e(t) = 4 \sum_{k=1}^{|\mathsf{x}^{(t)}|} \lambda_k^{(t)}(1 - \lambda_k^{(t)}) w_{ik}^2. \tag{34}$$

A.1.2 Continuous observations

The derivation is fully analogous to the one described in Appendix A.1.1, with the slight differences due to parameterization of the conditionals $p(v_i^{(t+1)}|\mathsf{x}^{(t)}) \sim \mathcal{N}(\mathsf{w}_i^T \mathsf{x}^{(t)}, s^2)$.

E-step:

From (21) we obtain $\lambda_k^{(t)} = \sigma(l_k)$ with

$$\begin{aligned} l_k^{(t)} = \tilde{l_k}^{(t)} - \frac{1}{2s^2} \sum_{i \in V^{(t+1)}} \left\langle (v_i^{(t+1)} - \mathsf{w}_i^T \mathsf{x}|h_k = 1)^2 - (v_i^{(t+1)} - \mathsf{w}_i^T \mathsf{x}|h_k = -1)^2 \right\rangle_{q(\mathsf{x}^{(t)})} \\ = \tilde{l_k}^{(t)} - \frac{2}{s^2} \sum_{i \in V^{(t+1)}} w_{ik} \left[\mathsf{w}_i^T (2\boldsymbol{\lambda}^{(t)} - \mathbf{1}) - w_{ik}(2\lambda_k^{(t)} - 1) - v_i^{(t+1)} \right], \qquad (35) \end{aligned}$$

where

$$\begin{aligned} \tilde{l_k}^{(t)} = \mathsf{w}_k^T (2\boldsymbol{\lambda}_k^{(t-1)} - 1) + b_k + \sum_{m \in H^{(t+1)}} \Bigg[\left\langle \log \left\{ \sigma(c_{mk}^t)^{\lambda_i^{(t+1)}} \sigma(-c_{mk}^t)^{1-\lambda_m^{(t+1)}} \right\} \right\rangle_{\mathcal{N}_{ij}^c(t)} \\ - \left\langle \log \left\{ \sigma(d_{mk}^t)^{\lambda_m^{(t+1)}} \sigma(-d_{mk}^t)^{1-\lambda_m^{(t+1)}} \right\} \right\rangle_{\mathcal{N}_{ij}^d(t)} \Bigg]. \qquad (36) \end{aligned}$$

As before, c_{mk}^t and d_{mk}^d are given by (23) and (24), the moments of $\mathcal{N}_{ij}^c(t)$ and $\mathcal{N}_{ij}^d(t)$ – by (25) – (27), and it is assumed that only those parameters λ_k which correspond to the hidden variables need to be adapted.

M-step:

By analogy with the previous case, optimization of the upper bound w.r.t. the model parameters leads to

$$\begin{aligned} \frac{\partial \Phi}{\partial w_{ij}} &= -\frac{1}{2s^2} \frac{\partial}{\partial w_{ij}} \sum_{t=0}^{T-1} \sum_{l \in V^{(t+1)}} \langle (\mathsf{w}_l^T \mathsf{x} - v_l^{(t+1)})^2 \rangle_{q(\mathsf{x}^{(t)})} \\ &+ \frac{\partial}{\partial w_{ij}} \sum_{k \in H^{(t+1)}} \left\langle \log \left\{ \sigma \left(x_k^{(t+1)} [\mathsf{w}_k^T \mathsf{x}^{(t)} + b_k] \right) \right\} \right\rangle_{q(x_k^{(t+1)}, \mathsf{x}^{(t)})}. \qquad (37) \end{aligned}$$

There are four possible simplifications of this expression, depending on observability of nodes $x_i^{(t+1)}$ and $x_j^{(t)}$ connected by the weight w_{ij}, so that
$i \in V^{(t+1)}$, $j \in H^{(t)} \Rightarrow$

$$\begin{aligned} \frac{\partial \Phi}{\partial w_{ij}} &= -\frac{1}{s^2} \sum_{t=0}^{T-1} \langle (\mathsf{w}_i^T \mathsf{x} - v_i^{(t+1)}) x_j^{(t)} \rangle_{q(\mathsf{x}^{(t)})} \\ &= \frac{1}{s^2} \left[v_i^{(t+1)} (2\lambda_j - 1) - \sum_{k=1}^{|\mathsf{x}^{(t)}|} w_{ik} \langle x_k^{(t)} x_j^{(t)} \rangle_{q(\mathsf{x}^{(t)})} \right] \\ &= \frac{1}{s^2} \left((2\lambda_j^{(t)} - 1) \left[v_i^{(t+1)} - \mathsf{w}_i^T (2\boldsymbol{\lambda}^{(t)} - \mathbf{1}) \right] + 4 w_{ij} (\lambda_j^{(t)} - 1) \lambda_j^{(t)} \right), \qquad (38) \end{aligned}$$

$i \in V^{(t+1)}$, $j \in V^{(t)} \Rightarrow$

$$\frac{\partial \Phi}{\partial w_{ij}} = \frac{1}{s^2} \left(v_j^{(t)} \left[v_i^{(t+1)} - \mathsf{w}_i^T (2\boldsymbol{\lambda}^{(t)} - \mathbf{1}) \right] \right), \qquad (39)$$

$i \in H^{(t+1)}$, $j \in V^{(t)} \Rightarrow$

$$\frac{\partial \Phi}{\partial w_{ij}} = v_j^{(t)} \left[\lambda_i^{(t+1)} - \langle \sigma(\mathsf{w}_i^T \mathsf{x} + b_i) \rangle_{q(\mathsf{x})} \right] \approx v_j^{(t)} \left[\lambda_i^{(t+1)} - \langle \sigma(e_i^t) \rangle_{\mathcal{N}_i^e(t)} \right], \tag{40}$$

where the moments of $e_i^t \sim \mathcal{N}_i^e(t)$ are given by (33) – (34). Finally, in the last case when $i \in H^{(t+1)}$, $j \in H^{(t)}$ the gradients $\partial\Phi/\partial w_{ij}$ are given by (31). Note that all of the above gradients can be combined to give (6), and $\partial\Phi/\partial b_i$ is expressed as (32).

A.2 Gibbs sampling

Here we derive the conditional distributions used in the sampling scheme outlined in section 4.

Let $\hat{\mathsf{X}}$ and $\acute{\mathsf{X}}$ be the odd layers of a 3-layer fully-connected chain $\hat{\mathsf{X}} \to \mathsf{X} \to \acute{\mathsf{X}}$. We are interested in sampling the even layer X from $p(\mathsf{X}|\hat{\mathsf{X}}, \acute{\mathsf{X}})$. Note that $p(\hat{\mathsf{X}}, \mathsf{X}, \acute{\mathsf{X}}) = p(\hat{\mathsf{X}})p(\mathsf{X}|\hat{\mathsf{X}})p(\acute{\mathsf{X}}|\mathsf{X})$, i.e.

$$p(\mathsf{X}|\hat{\mathsf{X}}, \acute{\mathsf{X}}) \propto p(\mathsf{X}|\hat{\mathsf{X}})p(\acute{\mathsf{X}}|\mathsf{X}) = \prod_i p(X_i|\hat{\mathsf{X}})p(\acute{\mathsf{X}}|\mathsf{X}^{\backslash i}, X_i). \tag{41}$$

Here X_i is a hidden variable in layer X, and $\mathsf{X}^{\backslash i} \overset{\text{def}}{=} \mathsf{X}\backslash X_i$. Note that in general $p(\acute{\mathsf{X}}|\mathsf{X}^{\backslash i}, X_i)$ is not factorized over X_is. This complicates normalization of (41) and motivates usage of another sampler for X's components

$$\begin{aligned} X_i \leftarrow p(X_i|\mathsf{X}^{\backslash i}, \hat{\mathsf{X}}, \acute{\mathsf{X}}) &\propto p(X_i, \mathsf{X}^{\backslash i}|\hat{\mathsf{X}})p(\acute{\mathsf{X}}|\mathsf{X}^{\backslash i}, X_i) \\ &\propto p(X_i|\hat{\mathsf{X}}) \prod_j p(\acute{X}_j|\mathsf{X}^{\backslash i}, X_i), \end{aligned} \tag{42}$$

where we have used (41) and the fact that $p(X_i|\mathsf{X}^{\backslash i}, \hat{\mathsf{X}}) = p(X_i|\hat{\mathsf{X}})$.

For the HHMs, the conditional distributions of the hidden variables $p(h_i^{(t+1)}|\mathsf{x}^{(t)})$ are given by (3), and the conditional distributions of the visible variables $p(v_i^{(t+1)}|\mathsf{x}^{(t)}) \sim \mathcal{N}(\mathsf{w}_i^T\mathsf{x}^{(t)}, s^2)$. From (42) it is then clear that

$$\begin{aligned} p(h_i^{(t)} &= \pm 1|\mathsf{x}^{(t-1)}, \mathsf{x}^{(t+1)}, \mathsf{x}^{(t)}\backslash h_i^{(t)}) \propto \\ &\sigma(\pm[\mathsf{w}_i^T\mathsf{x}^{(t-1)} + b_i]) \prod_{j=1}^{|\mathsf{h}^{(t+1)}|} \sigma(h_j^{(t+1)}[\mathsf{w}_j^T\mathsf{x}^{(t)} - w_{ji}x_i^{(t)} \pm w_{ji}]) \\ &\prod_{k=1}^{|\mathsf{v}^{(t+1)}|} \exp\left\{ -\frac{1}{2s^2}(v_k^{(t+1)} - [\mathsf{w}_k^T\mathsf{x} - w_{ki}x_i^{(t)} \pm w_{ki}])^2 \right\}, \end{aligned} \tag{43}$$

which leads to

$$\begin{aligned} x_i^{(t)} \quad \leftarrow \quad & p(x_i^{(t)} = 1|\mathsf{x}^{(t-1)}, \mathsf{x}^{(t+1)}, \mathsf{x}^{(t)}\backslash x_i^{(t)}) = \\ & \sigma\left\{ b_i + \mathsf{w}_i^T\mathsf{x}^{(t-1)} + \sum_{j=1}^{|\mathsf{x}^{(t+1)}|} \log \frac{\sigma\left(x_j^{(t+1)}\left[\mathsf{w}_i^T\mathsf{x} - w_{ji}x_i + w_{ji})\right]\right)}{\sigma\left(x_j^{(t+1)}\left[\mathsf{w}_i^T\mathsf{x} - w_{ji}x_i - w_{ji})\right]\right)} \right\} \end{aligned} \tag{44}$$

for discrete and

$$\begin{aligned} x_i^{(t)} \quad \leftarrow \quad & p(x_i^{(t)} = 1 | \mathsf{x}^{(t-1)}, \mathsf{x}^{(t+1)}, \mathsf{x}^{(t)} \backslash x_i^{(t)}) = \\ & \sigma \left\{ b_i + \mathsf{w}_i^T \mathsf{x}^{(t-1)} + \sum_{j=1}^{|\mathsf{h}^{(t+1)}|} \log \frac{\sigma \left(x_j^{(t+1)} \left[\mathsf{w}_i^T \mathsf{x} - w_{ji} x_i + w_{ji}) \right] \right)}{\sigma \left(x_j^{(t+1)} \left[\mathsf{w}_i^T \mathsf{x} - w_{ji} x_i - w_{ji}) \right] \right)} \right. \\ & \left. + \frac{2}{\sigma^2} \sum_{k=1}^{|\mathsf{v}^{(t+1)}|} w_{ki} \left(v_k^{(t+1)} - \mathsf{w}_k^T \mathsf{x}^{(t)} + w_{ki} h_i^{(t)} \right) \right\} \end{aligned} \tag{45}$$

for continuous data models. Here we have used simple identities $a/(a+b) \equiv \sigma(\log(a/b))$ and $\log(\sigma(a)/\sigma(-a)) \equiv a$ which hold $\forall a, b > 0$.

References

Barber, D. and Sollich, P. (2000). Gaussian Fields for Approximate Inference. In Solla, S. A., Leen, T., and Müller, K.-R., editors, *Advances in Neural Information Processing Systems 12*, pages 393–399. MIT Press, Cambridge, MA.

Bishop, C. M., Hinton, G. E., and Strachan, I. G. D. (1997). GTM Through Time. NCRG/97/005, NCRG, Dept. of Computer Science and Applied Mathematics, Aston University.

Ghahramani, Z. and Jordan, M. (1995). Factorial hidden Markov models. In Touretzky, D. S., Mozer, M. C., and Hasselmo, M. E., editors, *Proc. Conf. Advances in Neural Information Processing Systems, NIPS*, volume 8, pages 472–478. MIT Press.

Hertz, J., Krogh, A., and Palmer, R. G. (1991). *Introduction to the Theory of Neural Computation.* MA: Addison-Wesley Publishing Company.

Neal, R. M. (1992). Connectionist learning of belief networks. *Artificial Intelligence*, (56):71 – 113.

Rabiner, L. (1989). A tutorial on hidden Markov models and selected applications in speech recognition. *Proc. of the IEEE*, 77(2).

Roweis, S. (2000). Constrained hidden Markov models. In *Proc. Conf. Advances in Neural Information Processing Systems, NIPS*, volume 12. MIT Press.

Saul, L., Jaakkola, T., and Jordan, M. (1996). Mean field theory for sigmoid belief networks. *Journal of Artificial Intelligence Research*, 4.