Approximate Learning in Temporal Hidden Hopfield Models

Felix V. Agakov[1] and David Barber[1]

University of Edinburgh, Division of Informatics, Edinburgh EH1 2QL, UK,
felixa@anc.ed.ac.uk, http://anc.ed.ac.uk

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 (e.g. learning of incomplete binary sequences and reconstruction of 3D occupancy graphs) where distributed discrete hidden space representation may be useful.

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

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

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

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 $\{\mathsf{h}\}$.

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.

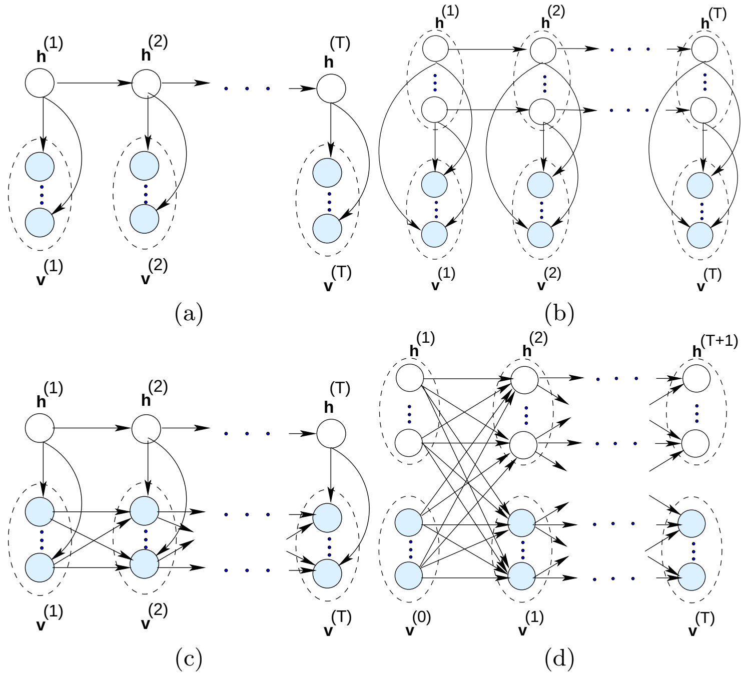


Fig. 1. Graphical models for temporal sequences: (a) Hidden Markov Model; (b) Factorial HMM; (c) Auto-regressive HMM; (d) generalized Markov chain with temporally shifted observations.

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 Sec. 5, it yields a fully probabilistic way of addressing some problems of scanning (3D shape reconstruction). We also demonstrate that the model can be applied to classification and 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

$$p(h_i^{(t+1)}|\mathsf{x}^{(t)}; \mathsf{w}_i, b_i) = \sigma\left(h_i^{(t+1)}(\mathsf{w}_i^T \mathsf{x}^{(t)} + b_i)\right), \quad \text{where} \quad \sigma(a) = 1/(1+e^{-a}), \tag{3}$$

w_i is a weight vector connecting node i with all of the nodes, and b_i is the node's bias.

The model has a graphical structure, temporal dynamics, and parameterization 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 approximate Gaussianity of the nodes' fields for numeric evaluation of the gradients, yielding a fast rule for learning incomplete discrete sequences. HHM learning of continuous time series results in a related, but different rule (Sec. 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 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 data. Derivations of these results and the simpler learning rule for discrete-data HHMs are given in Agakov and Barber, 2002.

3.1 Variational 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 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 its expected posterior firing rate and must be learned from data.

The M-Step. Let w_{ij} be connecting $x_i^{(t+1)}$ and $x_j^{(t)}$. By maximizing the bound on the likelihood (2) we get

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

where $f_i^{(t+1)} \in \{0, 1\}$ is an indicator equal to 1 if and only if x_i is visible at time $t+1$ [i.e. $i \in V^{(t+1)}$]. The gradient contributions depend on the units' observability and are given by

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

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

with the *fields*

$$c_{ij}^t = \mathsf{w}_i^T \mathsf{x}^{(t)} + b_i | h_j^{(t)} = 1, \quad d_{ij}^t = \mathsf{w}_i^T \mathsf{x}^{(t)} + b_i | h_j^{(t)} = -1, \quad e_i^t = \mathsf{w}_i^T \mathsf{x}^{(t)} + b_i. \tag{9}$$

Since c_{ij}^t, d_{ij}^t, and e_i^t are given by linear combinations of a large number of random variables, the Central Limit Theorem implies approximate Gaussianity of $p(c_{ij}^t)$, $p(d_{ij}^t)$, and $p(e_i^t)$ (Barber and Sollich, 2000) with the means expressed as

$$\mu_{ij}^d(t) = \mathsf{w}_i^T(2\boldsymbol{\lambda}^{(t)} - \mathbf{1}) - 2w_{ij}\lambda_j^{(t)} + b_i, \;\; \mu_{ij}^c(t) = \mu_{ij}^d(t) + 2w_{ij}, \;\; \mu_i^e(t) = \mathsf{w}_i^T(2\boldsymbol{\lambda}^{(t)} - \mathbf{1}) + b_i \tag{10}$$

and the variances

$$s_{ij}^d(t) = s_{ij}^c(t) = 4\sum_{k \neq j}^{|\mathsf{x}^{(t)}|} \lambda_k^{(t)}(1-\lambda_k^{(t)})w_{ik}^2, \quad s_i^e(t) = s_{ij}^d(t) + 4\lambda_j^{(t)}(1-\lambda_j^{(t)})w_{ij}^2. \tag{11}$$

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 $\langle x_j^2 \rangle = 1$. 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 1-D averages may be efficiently evaluated by using numerical Gaussian integration, and even crude approximation at the means often leads to good results (see Sec. 5).

The E-Step. Optimizing the bound (2) w.r.t. the expected firing rates $\lambda_i^{(t)}$ 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 approximately normally distributed with the moments (10), (11).

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 Fig. 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] with an additional summation over the training sequences.

3.3 Constrained Parameterization

It is clear that if the model has n binary hidden variables it is capable of representing 2^n states for each time slice. However, the full transition matrix comprises n^2 weights, which may require prohibitively large amounts of training data and high computational complexity of learning. In Sec. 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 results in reasonable performance.

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 Sec. 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 remain 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

$$\begin{aligned} x_i^{(t)} \leftarrow p(x_i^{(t)} = 1|\mathsf{x}^{(t-1)},\mathsf{x}^{(t+1)},\mathsf{x}^{(t)}\backslash x_i^{(t)}) = \\ \sigma\left\{ b_i + \mathsf{w}_i^T\mathsf{x}^{(t-1)} + \sum_{j=1}^{|\mathsf{h}^{(t+1)}|} \log \frac{\sigma\left(x_j^{(t+1)}\left[\mathsf{w}_i^T\mathsf{x} - w_{ji}x_i + w_{ji}\right]\right)}{\sigma\left(x_j^{(t+1)}\left[\mathsf{w}_i^T\mathsf{x} - w_{ji}x_i - w_{ji}\right]\right)} \right. \\ \left. + \frac{2}{s^2}\sum_{k=1}^{|\mathsf{v}^{(t+1)}|} w_{ki}\left(v_k^{(t+1)} - \mathsf{w}_k^T\mathsf{x}^{(t)} + w_{ki}h_i^{(t)}\right)\right\} \end{aligned} \tag{15}$$

for continuous-data HHMs (Agakov and Barber, 2002).

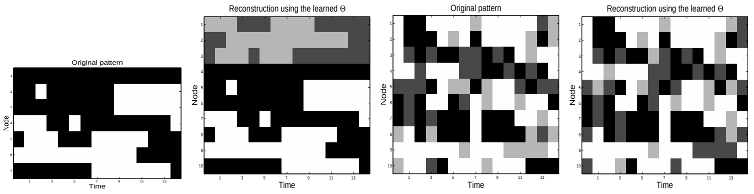
5 Experimental results

5.1 Sequence Reconstruction

One way to validate correctness of the HHM learning rule is by performing deterministic retrieval 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 retrieval 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 [Fig. 2 (a)]. The model parameters $\boldsymbol{\Theta}$ were learned by the EM algorithm (Sec. 3). The retrieved sequence was generated from the initial pattern $\mathsf{x}^{(0)}$ sampled from the learned prior $p(\mathsf{x}^{(0)})$, with each subsequent pattern $x_i^{(t+1)}$ set according to $sgn(\sigma(x_i^{(t+1)}(\mathsf{w}_i^T\mathsf{x} + b_i)) - 1/2)$ [Fig. 2 (b)]. Note that patterns $\mathsf{v}^{(7)}$ and $\mathsf{v}^{(8)}$ are identical, and it is the learned activation of the

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



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

Fig. 2. Training and retrieved 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.

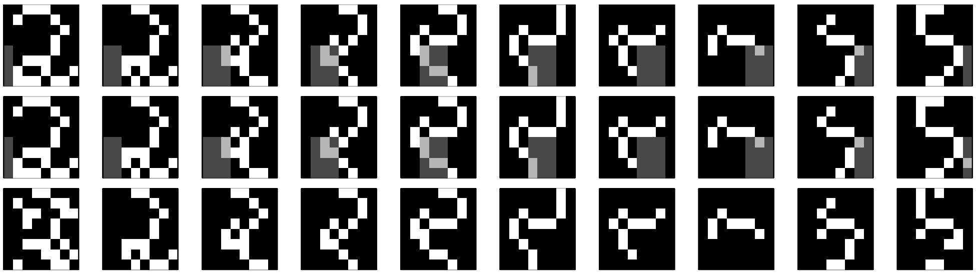


Fig. 3. Inference of incomplete discrete time series. *Top:* the true underlying sequence; *Middle:* the sequence clamped on the visible variables (black/white) with the inferred values of the missing variables (dark/light gray); *Bottom:* the sequence reconstructed from a complete noisy initialization by forward sampling.

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.

Figure 3 shows an example of applying an HHM to reconstruction of a temporal sequence from its incomplete noisy versions. The underlying data contained 10 8×8 binary images with an average Hamming distance of 7 bits between the subsequent patterns (see Fig. 3 *top*). The model was trained on 4 sequences generated from the complete series by randomly omitting about 15% and permuting 10% 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 Fig. 3 *middle*, the resulting reconstruction is reasonably accurate. We have also tried to retrieve the underlying sequence by deterministic forward sampling (Fig. 3 *bottom*) from noisy initialization, perturbing each subsequent pattern with an additional 10% noise. We see that the results are still accurate, though further experiments show that this reconstruction proves to be sensitive to the noise of the training sequences.

5.2 Sequence Classification

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), which can be evaluated variationally or by sampling.

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 smoothly changing in time. 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

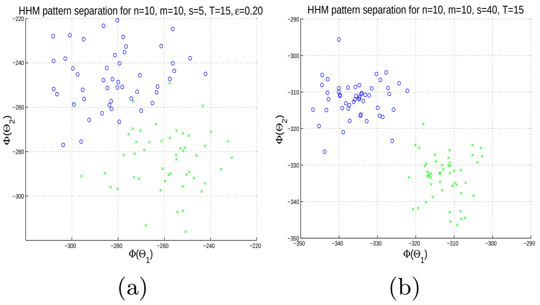


Fig. 4. Temporal sequence classification in HHMs. *Symbols:* true class labels of the testing sequences; *Axis:* approximate bounds on the likelihoods for $\mathcal{M}_1$ and $\mathcal{M}_2$. (a): testing data is constructed by flipping each bit of each training sequence with probability 0.2; $n = 10, m = 10, T = 15, s = 5$. (b): testing data is drawn from the data-generating processes; $n = 10, m = 10, T = 15, s = 40$.

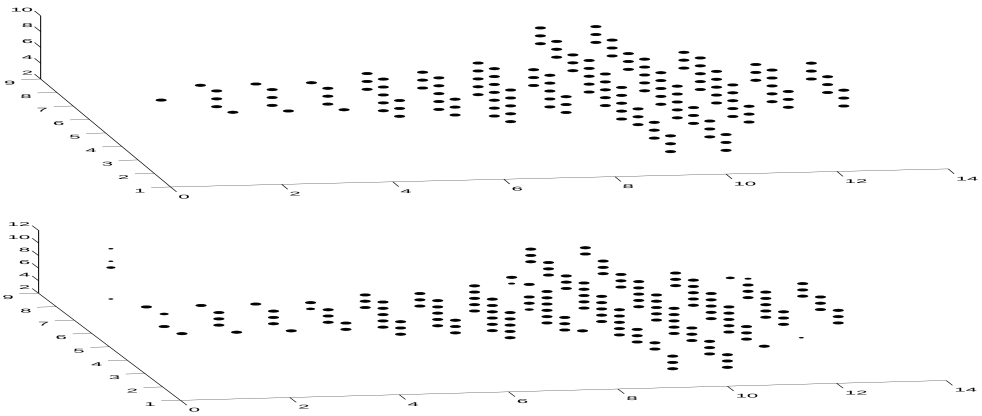


Fig. 5. 3D shape reconstruction. *Axis*: the scanning plane, time. The disk radii are proportional to square roots of the posterior probabilities. *Top:* the true shape; *Bottom:* the reconstructed shape.

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 4 (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 4 (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 likelihood bounds, and may be used as features for a linear classifier. As expected, up to a certain point increase in m generally leads to decrease in probability of misclassification (see Agakov and Barber, 2002 for details). Analogous experiments for continuous sequences lead to qualitatively similar class separation.

5.3 Constrained HHMs for Shape Reconstruction

To demonstrate a potential practical application of HHMs we consider the problem of reconstructing a 3D binary occupancy graph of an object from continuous sequences of noisy 1D scanner measurements. It is assumed that the object moves with uniform speed orthogonally to the plane scanned by two mutually perpendicular linear scanners, with the measurements given by the number of the filled cells along each planar slice.

Figure 5 shows application of a constrained HHM $p(\{\mathsf{h}\},\{\mathsf{v}\}|\mathsf{h}^{(0)}) = \prod_{t=0}^{T-1} p(\mathsf{h}^{(t+1)}|\mathsf{h}^{(t)})p(\mathsf{v}^{(t+1)}|\mathsf{h}^{(t)})$ with 12×9 hidden and $12 + 9$ visible variables to shape reconstruction for 13 time frames. In fact, it is an HMM with a high-dimensional distributed hidden space representation, where the hidden and visible variables correspond to the binary occupancy cells and noisy scanner measurements ($s^2 = 1$) respectively. 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 Fig. 5 *bottom* we see that impervious to the constraints and the facts that the scanner data is noisy and the inference task is severely under-determined, the model accurately reconstructs the underlying shape – a hunting knife (Fig. 5 *top*). 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.) may be practical for learning or inferring inherently smooth or constrained data. Moreover, they suggest possible combinations of exact and approximate learning methods in sparse HHMs. Finally, note that unlike constrained HMMs (Roweis, 1999) or temporal GTMs (Bishop et al., 1997), HHMs feature distributed discrete hidden spaces, suggesting potential applications to half-toning and spiking temporal topographic maps of incomplete discrete and continuous sequences under carefully imposed topological constraints (Agakov and Barber, 2002).

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 temporal sequences. We also discussed inference applications of the constrained Hidden Hopfield Models.

References

Agakov, F. V. and Barber, D. (2002). Temporal Hidden Hopfield Models. Technical Report EDI-INF-RR-0156, Division of Informatics, University of Edinburgh.

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. (1999). 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.