An Operator Interpretation of Message Passing

I. Rezek
Robotics Research Group
University of Oxford
Oxford, OX1 3PJ
irezek@robots.ox.ac.uk

S.J. Roberts
Robotics Research Group
University of Oxford
Oxford, OX1 3PJ
sjrob@robots.ox.ac.uk

Abstract

Message passing algorithms may be viewed from a purely probabilistic or statistical physics perspective. This work describes an alternative, linear algebraic view of message passing in trees. We demonstrate the construction of global belief operators on Markov Chains and Trees and compare these with classical results. By interpreting message passing as finding a global stable solution we wish to avoid application dependent development of message passing methods of more complex graph structures.

1 Introduction

Graphical Models depict graphically the relationships between parameters of a model. These relationships are probabilistic in nature and the aim is to infer the value of the model’s variables from some observed data. From a Bayesian probabilistic viewpoint, the goal is to find the posterior distributions of the parameters given the data.

If the graphical model is simple, then recursive algorithms can be found to propagate probabilities between nodes on the graph. Simplicity in this case means that either analytic solutions are possible and/or the topology of the graph is simple in that it contains no loops or cycles.

In the case of non-analytic solutions, the model can be estimated using sampling methodologies. In Markov Chain Monte Carlo methods this implies random sampling of potential values for each model variable. If this is done for a sufficiently long time (and assuming certain other conditions such as irreducibility) the hope is that one samples from the stationary distribution of the Markov Chain.

Another approach to finding the parameters’ probability distributions is that of variational inference. After some simplifying assumptions, whether mean field or structured approximations, the key is to iterate across all variables and compute their probability densities/distributions until a stationary point is found.

The key message here is that what we would like to infer is after is the stationary distribution of the model. This can also be seen from a Physics point of view. Consider, for instance, a discrete graphical model. It is not difficult to convert the probabilities relating the parameters to each other into an Ising-type Hamiltonian. Consider also a Kalman-filter type graphical model. This can be expressed as a diffusion equation with the appropriate
Hamiltonian [5]. In both cases one is interested in the stationary “wave-function” describing the state (probabilities) of the models’ parameters.

It is therefore not surprising to notice the close resemblance between belief propagation methods, for example in trees, to successive over-relaxation methods used in partial differential equations. In fact, the forward-backward iterations for propagating beliefs along trees are reminiscent of methods for solving linear equations.

In the following we interpret belief propagation in trees in terms of solving a set of linear equations. We begin by presenting a motivating example for a simple Ising model case and show analytically that belief propagation operators can be constructed to compute the forward and backward messages in a Markov Chain and we then extend the topology to Markov trees. We conclude by recasting the issue of node clustering as a problem of describing the problem in matrix form and speculate about extensions to belief propagation to other graphs, such as poly-trees and higher dimensional structures.

2 Eigenvalues and Eigenstates in spin Ising Hamiltonians

The Ising model is often used as a physical analogy of graphical models. The Ising model is typically assumed to be a regular lattice of \( N \) sites, in \( d \) dimensions, with spin variables \( S_i \) and Hamiltonian

\[
H = -J \sum_{(i,j)} S_i S_j - 2\beta H_0 \sum_i S_i
\]

(1)

where \((i,j)\) denotes all neighbouring sites \( i \) and \( j \), \( \beta \) the Bohr magneton and \( H_0 \) is an external magnetic field which here we assume to be zero. For a 1-d Ising Model, the Hamiltonian (1) can be rewritten as

\[
H_{d=1} = -J \vec{S}_{N}^T A \vec{S}_N
\]

(2)

where \( \vec{S}_N = S_1, S_2, \cdots, S_N \) and \( A \) is a \( N \times N \) quasi-tridiagonal matrix with zeros along the main diagonal and ones in the bordering diagonals and in the top left and bottom right elements (to accommodate periodic boundary conditions).

The eigenvalues and vectors, respectively, of \( A \) in (2) were shown to be [1]

\[
\lambda_n = 2 \cos \left( \frac{2\pi}{N} (n - 1) \right) \quad \text{and} \quad \vec{v}_n = \frac{1}{\sqrt{N}} \sum_{s=1}^{N} e^{i2\pi (n-1)(s-1)/N} \vec{e}_n
\]

(3)

with \( \vec{e}_n \) denoting a vector of zeros except for the \( n \)-th entry, and \( i = \sqrt{-1} \). Noting that the matrix of eigenvectors \( \mathbf{V} \) is unitary, we can write \( H_{d=1} = \mathbf{V} \mathbf{D} \mathbf{V}^{-1} \) and readily evaluate the partition function.

In a 2-d case, if we assume a square net of \( N \) sites along the 2 dimensions, we can re-write the Hamiltonian as

\[
H_{d=2} = -J \vec{S}_{N^2}^T B \vec{S}_{N^2}
\]

(4)

with \( \vec{S}_{N^2} = \vec{S}_N \otimes \vec{S}_N \), and \( B \) a \( N^2 \times N^2 \) tri-block diagonal matrix with \( N \) matrices \( A \) along the main diagonal and identity matrices of order \( N \), \( I_N \), along the off block diagonals and at the top left and bottom right again to account for cyclic boundary conditions. Again \( H_{d=2} \) can be brought into diagonal form using its eigenvalues and eigenvectors [1].

If the Hamiltonian of a system is known all the system’s dynamics are known in principle. However, much detailed physical knowledge is needed to construct the Hamiltonian. Since there is an equivalence between the Hamiltonian description and an operator description, the same dynamics can be described by appropriately chosen operators. For the Ising model
the Hamiltonian described above is Hermitian and can be spectrally decomposed in the form

\[ H = \sum_n \lambda_n \vec{v}_n \vec{v}_n^h. \tag{5} \]

The eigen-solutions of the Hamiltonian are the stationary states, i.e.

\[ V e^{-i\Lambda t} V^h \equiv V U V^h \equiv \mathcal{P} \tag{6} \]

where we have define an operator \( \mathcal{P} \) which has the same spectral decomposition as the Hamiltonian. The eigen-decomposition of the operator \( \mathcal{P} \) will then, after exponentiation, give the same energy stationary states as those of the physical system.

Using the equivalence of the Ising model and graphical models, the aim is to construct an operator, the eigen-solution of which corresponds to the belief distributions of the graphical model. We will call this operator the global belief operator.

3 Operators for Markov Chains

Based on the observation that the marginal distributions of the nodes in a graphical model can be computed from the stationary distribution of a global belief operator, we attempt first to construct such an operator for a simple Markov chain. Operators are constructed depending on whether messages are passed forwards or backwards. To evaluate the method, we use a standard HMM-type graphical model, with stationary state transition probabilities and Gaussian observation model, but incorporate the observation nodes into the state transition probabilities so as to obtain a Markov chain. The Markov chain thus consists of locally perturbed state transitions, i.e.

\[ P_{t+1} \equiv P(S_{t+1}|S_t)P(O_{t+1}|S_{t+1}) = P(O_{t+1},S_{t+1}|S_t) \quad \forall t = 1 \cdots T \tag{7} \]

where \( S_t \) and \( O_t \) are the time-indexed states and observations, respectively.

Consider first message passing forward in time. One possible choice global belief operator is represented by a block-partitioned transition matrix with all \( P_t \) of the Markov chain on the lower block diagonal,

\[
\mathcal{P}_f = \begin{bmatrix}
0 & 0 & 0 & \cdots & P_1 \\
P_2 & 0 & 0 & \cdots & 0 \\
0 & P_3 & 0 & \cdots & 0 \\
0 & 0 & P_4 & \cdots & 0 \\
\vdots & \ddots & \ddots & \ddots & \ddots \\
0 & \cdots & 0 & P_T & 0
\end{bmatrix} \tag{8}
\]

Similarly, for the backward messages we construct a global block-partitioned transition matrix with all \( P_t \) on the upper block diagonal

\[
\mathcal{P}_b = \begin{bmatrix}
0 & 0 & P_1^T & 0 & \cdots & 0 \\
0 & 0 & P_2^T & \cdots & 0 \\
\vdots & \ddots & \ddots & \ddots & \ddots \\
0 & \cdots & 0 & P_T^T \\
P_1^T & \cdots & \cdots & \cdots & 0
\end{bmatrix} = \mathcal{P}_f^T \tag{9}
\]

The operator forms a global state transition probability (with cyclic boundary conditions). The global state vector is now formed by stacking the state variables of the original Markov chain’s states at each time step \( \vec{s} = [s_1^T \cdots s_T^T]^T \). Thus, assuming the \( P_t \) is \( K \times K \), \( \mathcal{P}_f \) is \( KT \times KT \).
It is quite possible to use the operator above and spectrally decompose it to obtain belief distributions. However, to compare it against standard forward-backward solutions of Markov chains, a change of boundary conditions is needed. The classical forward-backward recursions assume Dirichlet boundary conditions, i.e. the prior of the first state node and the likelihood of the last state node are permanently clamped to some values. The resulting belief operator is of the form

$$\mathbf{P}_f = \begin{bmatrix} P_0 & 0 & 0 & \cdots & 0 \\ P_1 & 0 & 0 & \cdots & 0 \\ 0 & P_2 & 0 & \cdots & 0 \\ 0 & 0 & P_3 & \cdots & 0 \\ \vdots & \ddots & \ddots & \ddots & \vdots \\ 0 & \cdots & 0 & P_T & 0 \end{bmatrix}$$

(10)

Similarly, for the backward recursions we construct a global block-partitioned transition matrix with all $P_k$ on the upper block diagonal

$$\mathbf{P}_b = \begin{bmatrix} 0 & P_1^T & 0 & \cdots & 0 \\ 0 & 0 & P_2^T & \cdots & 0 \\ \vdots & \ddots & \ddots & \ddots & \vdots \\ 0 & \cdots & 0 & P_T^T \\ 0 & \cdots & 0 & P_{\infty}^T \end{bmatrix}$$

(11)

The boundary conditions are now specified in the matrices $P_0$ and $P_{\infty}$, for the forward and backward operators respectively. In particular, $P_0$ is chosen such that its eigen-solution is equivalent to the prior distribution of the Markov chain and $P_{\infty} = \mathbf{1}_K (\mathbf{1}_K)^T$, where $\mathbf{1}_K$ is a $K$-element column vector of ones, corresponding to the unity likelihood $P(O_{\infty}|S_T)$ at the chain’s end.

It is now easy to prove that the operators $\mathbf{P}_f$ and $\mathbf{P}_b$ have the desired eigen-solution. By construction, the characteristic equation of $\mathbf{P}_f$ is

$$0 = |\mathbf{P}_f - \lambda \mathbf{I}_{KT}| = |P_0 - \lambda \mathbf{I}_K||\lambda \mathbf{I}_{K-T}-K| .$$

(12)

This has $KT-K$ multiple solutions of $\lambda = 0$ and the eigen-solutions of $P_0$ of which one, by construction, has $\lambda = 1$. The eigenvector $\mathbf{v}$ associated with $\lambda = 1$ is the ground energy solution we seek. Writing $\mathbf{v} = \text{vec}(\mathbf{v}_0, \mathbf{v}_2, \cdots, \mathbf{v}_T)$, and where vec($\cdot$) is just the vector, or stacking, operator, the eigenvector’s $T$ sub-vectors relate as

$$\mathbf{v}_0 = \mathbf{P}_0 \mathbf{v}_0$$

(13)

$$\mathbf{v}_t = \mathbf{P}_{t-1} \mathbf{v}_{t-1} \quad \forall t = 1, \cdots, T .$$

(14)

By comparison, the forward iterations in the Markov chain are given as [3]

$$\alpha_{t+1} = P(O_{t+1}^*, S_{t+1}) = P(O_{t+1}|S_{t+1})\alpha_t \quad \forall t = 1, \cdots, T .$$

(15)

The proof for the backward iterations is similarly trivial, resulting in the backwards eigenvector’s $T$ sub-vectors, $\mathbf{b} = \text{vec}(\mathbf{b}_1, \mathbf{b}_2, \cdots, \mathbf{b}_\infty)$, to be

$$\mathbf{b}_{\infty} = \mathbf{P}_{\infty}^T \mathbf{b}_{\infty}$$

(16)

$$\mathbf{b}_t = \mathbf{P}_{t+1}^T \mathbf{b}_{t+1} \quad \forall t = 1, \cdots, T .$$

(17)

The complete belief of the individual states, $\gamma_t$ for $t = 1, 2, \cdots, T$, can be computed [3] by

$$\gamma_t \propto \mathbf{v}_t \odot \mathbf{b}_t,$$

(18)

with $\odot$ denoting the Hadamard product.
3.1 Example

We demonstrate the algorithm of the previous section and compare the results with those obtained by the standard HMM forward backward recursions. The HMM has a state space dimension of 2 and a Gaussian observation model. The Gaussian kernels have means

\[ \mu_1 = \begin{bmatrix} -0.3728 \\ -0.0190 \\ 0.3017 \\ 0.4101 \end{bmatrix}, \quad \mu_2 = \begin{bmatrix} -0.0455 \\ -0.0000 \\ 0.0852 \\ 0.0434 \end{bmatrix} \]

and Covariance matrices

\[ \Sigma_1 = \begin{bmatrix} 0.0765 & 0 & 0 & 0 \\ 0 & 0.0821 & 0 & 0 \\ 0 & 0 & 0.1130 & 0 \\ 0 & 0 & 0 & 0.0440 \end{bmatrix}, \quad \Sigma_2 = \begin{bmatrix} 0.1066 & 0 & 0 & 0 \\ 0 & 0.0908 & 0 & 0 \\ 0 & 0 & 0.0594 & 0 \\ 0 & 0 & 0 & 0.0776 \end{bmatrix} \]

The HMM state transition matrix is given as

\[ P = \begin{bmatrix} 0.9242 & 0.0758 \\ 0.0325 & 0.9675 \end{bmatrix} \]

which has as its stationary distribution \( \pi = [0.3, 0.7]^T \).

Figure (1) shows the solutions of the forward messages, backward messages and marginal beliefs computed using the Baum-Welch iterations. The equivalent variables obtained by eigen-decomposition are shown in the same figures above the message passing solutions. Note that in HMMs the backward recursions are usually scaled by a scaling factor obtained during the forward recursions. This is not done in the example here, but backward variables are scaled independently from the forward variables.

4 Operators for Markov Trees

The solution for the Markov chain case suggests that it would should, in principle, be possible to apply the same method to any tree structure. Consider, for example, the top-down belief operator for Markov trees. Assuming a probabilistic graphical model with
nodes \( X_V = \{X_1, X_2, \cdots, X_N\} \), its joint probability distribution can be expressed as \[4\]
\[
Pr\{X_V\} = \prod_i Pr\{X_i|X_{\pi(i)}\} \quad \forall i = 1, \cdots, N
\]
(19)

where \( \pi(i) \) denotes the parents of node \( i \). By labelling the nodes in increasing order, the belief operator of any tree structured graph can be brought into a block-triangular form

\[
P_f = \begin{bmatrix}
P_{11} & 0 & 0 & \cdots & 0 \\
P_{21} & 0 & 0 & \cdots & 0 \\
P_{31} & P_{32} & 0 & \cdots & 0 \\
P_{41} & P_{42} & P_{43} & \cdots & 0 \\
\vdots & \ddots & \ddots & \ddots & \vdots \\
0 & \cdots & 0 & P_{NN-1} & 0
\end{bmatrix}
\]
(20)

with \( P_{ij} \) equal to the null matrix \( 0_K \) of order \( K \) if nodes \( i \not\sim j \), i.e. nodes \( i \) and \( j \) are not neighbours.

The existence of the eigen-solution is identical to the Markov Chain case, resulting in eigenvectors for each variable \( i \in V \)

\[
\vec{v}_0 = P_{11} \vec{v}_0
\]
(21)

\[
\vec{v}_i = \sum_{j=1}^{i-1} P_{ij} \vec{v}_j.
\]
(22)

Taking to account the tree topology of the graph, i.e. \( P_{ij} = 0_K \) if \( i \not\sim j \), it is now easy to see that, using the eigenvectors (22), the top-down messages from node \( j \) arriving at node \( i \), \( \mu_i \), can be expressed as

\[
\vec{v}_i = \prod_{(k,l) \in i \rightarrow j} P_{kl} \vec{v}_j
\]
(23)

where \((k,l) \in i \rightarrow j\) denotes all the node pairs along the path from node \( i \) to node \( j \). Thus, the messages in the eigen-solution are not messages between two neighbouring nodes, as in the message passing case, but between any two nodes which are connected.

In a tree structure, top down propagated messages are basically cloned at every node which has multiple children. This results in the simple operator structure seen before. Bottom-up messages, in contrast, need to be fused. This, inherently makes the operator structure more complex.

4.1 Example

For example of a graphical model’s, shown in figure (2), conditional probabilities, adapted from [4], is shown in table (1). Propagating the prior probability of the root node \( A \) down, the tree, the resultant node probabilities obtained by eigen-decomposition are given in the left column of table (1).

5 Conclusion

The aim of this work has been to show that Pearl’s message passing algorithms can be viewed, not only as a fixed points of Bethe’s free energy [6], but also as solutions to a highly sparse linear set of equations. The solutions obtained by spectral decomposition and belief propagation are identical (or very similar if different boundary conditions are chosen). Actual numerical differences can also be attributed to the numerical accuracies of the spectral decomposition algorithm.
Figure 2: Example graphical model with tree topology.

Table 1: Graphical Model with Conditional Probabilities adapted from [4] (left) and computed beliefs (right)

<table>
<thead>
<tr>
<th>CPT Specifications</th>
<th>Top Down: Eigen-solution (left), Message Passing (right)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Pr(A=0)</td>
</tr>
<tr>
<td>Pr(A=0)</td>
<td>0.7</td>
</tr>
<tr>
<td>Pr(B=0</td>
<td>0.1</td>
</tr>
<tr>
<td>Pr(C=0</td>
<td>0.8</td>
</tr>
<tr>
<td>Pr(D=0</td>
<td>0.5</td>
</tr>
<tr>
<td>Pr(E=0</td>
<td>0.3</td>
</tr>
<tr>
<td>Pr(F=0</td>
<td>0.7</td>
</tr>
<tr>
<td>Pr(G=0</td>
<td>0.9</td>
</tr>
<tr>
<td>Pr(H=0</td>
<td>1.0</td>
</tr>
<tr>
<td>Pr(I=0</td>
<td>0.2</td>
</tr>
<tr>
<td>Pr(J=0</td>
<td>0.7</td>
</tr>
</tbody>
</table>

Embedding the entire graph structure into a what is effectively a global transition probability in the Markov chain case, implies that the stationary distribution is sought along a “time” or iteration dimension perpendicular to the time dimension of the original Markov chain. This also makes clear that the search is done over all possible paths, since the matrix exponential is simply the limit of the sum of increasing powers of the of this global transition probability. In fact, this relationship to Feynman path integrals suggests approximations to them as methods for computing marginal beliefs.

The computational efficiency of the message passing algorithm has been lost by re-generalling the problem in Ising form. However, there are two important aspects worth noting. Apart from phrasing the problem in terms of classical linear algebra and thus opening up a completely new arsenal of tools, much more is known about the behaviour of linear sets of equations subject to constraints, such as boundary problems. For instance, limit cycle phenomena, as observed in loopy belief propagation, may also have their origin in unsatisfiable constraints. Thus, the analysis of algorithms are placed on a concrete mathematical footing. The question also arises, whether new loopy belief propagation algorithms can be derived by simply looking for the largest eigenvalue solution, or whether the global transition matrix can be sampled and thus fusing graph structure search and belief propagation.
At this stage, several issues need to be solved first. In the example above we have made sure that the stationary distribution of the HMM transition probability matrix is not symmetrical. So the transpose operation, akin to a time reversal operation, is the general case. Several issues are worth noting at this stage. In the example above we have made sure that the stationary distribution of the HMM transition probability matrix is not symmetrical and that the transpose operation is also a time reversal operation [2]. Yet to be studied are time reversal for tree structures and belief operators for poly-trees.

References


