An Introduction to Probabilistic Graphical Models

Michael I. Jordan
University of California, Berkeley

June 30, 2003

Chapter 18

The HMM and State Space Model Revisited

In this chapter we revisit the Hidden Markov model and the Linear Gaussian model from the more general point of view of the previous two chapters. It is illuminating to see the relationship between the recursive algorithms that we developed in Chapters 12 and 15 and the general junction tree constructions in Chapter 17.

18.1 Hidden Markov models

Recall that the Hidden Markov model (HMM) can be represented as the chain-structured graphical model shown in Figure 18.1(a). The multinomial *state variables* q_t form the backbone of the model; from this backbone hang the observable *output variables* y_t.

We parameterize the model by endowing the first node with an *initial probability* π, where $\pi_i \triangleq P(q_1^i = 1)$, and each subsequent state node with a *transition matrix* A, where $a_{ij} \triangleq P(q_{t+1}^j = 1 | q_t^i = 1)$. The output nodes are assigned the local conditional probability $P(y_t|q_t)$. For concreteness

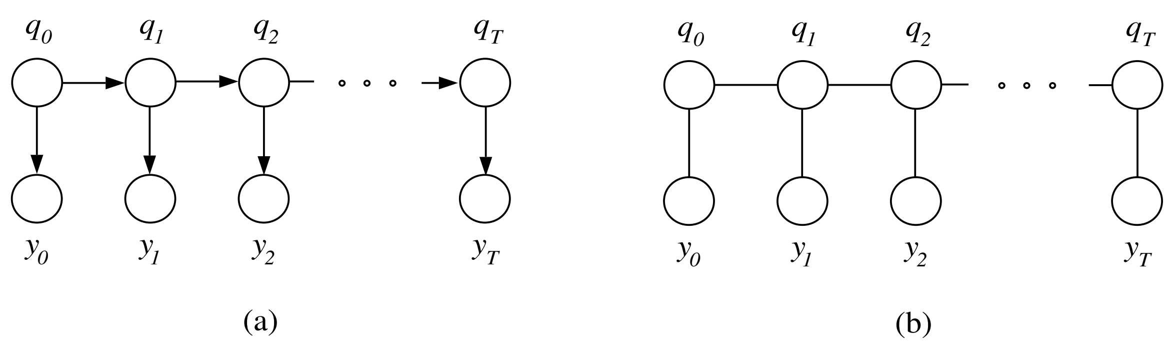

Figure 18.1: (a) The representation of a HMM as a graphical model. (b) The moralized, triangulated HMM graph.

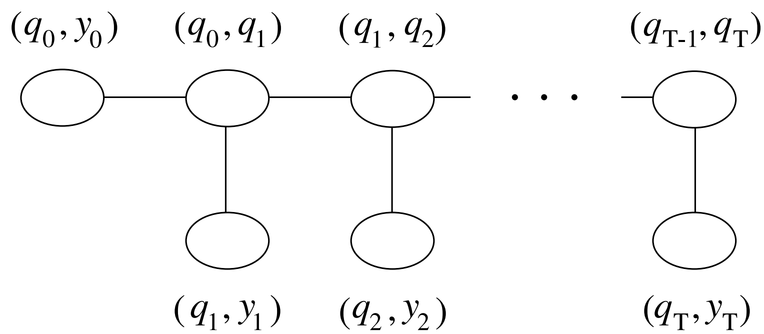

Figure 18.2: A maximal spanning tree on the cliques of the HMM graph.

we will assume that y_t is a multinomial node, so that $P(y_t|q_t)$ can be viewed as a matrix B, where $b_{ij} \triangleq P(y_t^j = 1|q_t^i = 1)$. However, given that y_t is always observed, this assumption will play no critical role in our discussion.

To convert the HMM into a junction tree, we proceed as outlined in the previous chapter. The moralization step is vacuous in this case, given that each node has at most a single parent. Moreover, the triangulation step is also vacuous, given that the graph has no cycles. We are left with the moralized, triangulated graph shown in Figure 18.1(b).

The cliques in the graph in Figure 18.1(b). are given by the pairs (q_t, q_{t+1}) and (q_t, y_t). There are several ways to connect these pairs so as to obtain a maximal spanning tree; with a bit of foresight we choose the maximal spanning tree shown in Figure 18.2.[1] This then is our junction tree.

Figure 18.3 shows the junction tree in which the separator sets have been made explicit and the potentials have been labeled. We make the following choice for the assignment of local conditional probabilities to potentials. The initial probability $P(q_0)$ as well as the conditional prob-

[1] In Exercise XXX we ask the reader to explore some of the alternative inference algorithms generated by the alternative choices of maximal spanning tree.

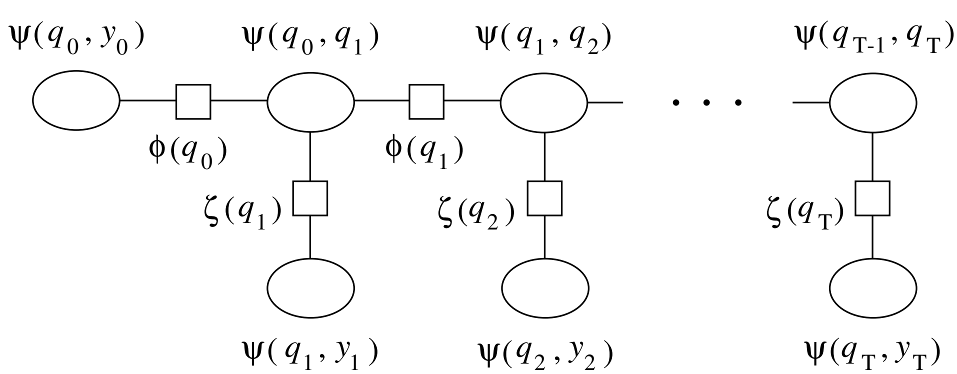

Figure 18.3: A junction tree for the HMM with the potentials labeled.

ability $P(y_0|q_0)$ is assigned to the potential $\psi(q_0, y_0)$. Note that this implies that this potential is initially set to the marginal $P(q_0, y_0)$. The state-to-state potentials are given the assignment $\psi(q_t, q_{t+1}) = P(q_{t+1}|q_t)$; note that these are conditional probabilities rather than marginals. Finally, the remaining output probabilities $P(y_t|q_t)$ are assigned to the potentials $\psi(q_t, y_t)$ and all separator potentials are initialized to one.

18.1.1 Unconditional inference

It is instructive to consider running the junction tree inference algorithm before any evidence has been observed. Suppose that we designate the node (q_{T-1}, q_T) as the root and collect to the root.

Consider first the operation of passing a message upward from a clique $(q_t, y_t) to its neighbor (q_{t-1}, q_t)$, for $t > 1$. The marginalization operation in this case yields $\sum_{y_t} \psi(q_t, y_t) = \sum_{y_t} P(y_t|q_t) = 1$; thus the separator potential $\zeta^*(q_t)$ remains set at one. This implies that the update factor $\zeta^*(q_t)\zeta(q_t)$ is one, and thus the potential $\psi(q_{t-1}, q_t)$ remains unchanged. In general, the messages that are passed upward from the leaves (q_t, y_t) have no effect when no evidence is observed.

Now consider the message from (q_0, y_0) to (q_0, q_1) (see Figure 18.3). We have:

$$\phi^*(q_0) = \sum_{y_0} \psi(q_0, y_0) = \sum_{y_0} P(q_0, y_0) = P(q_0) \tag{18.1}$$

$$\psi^*(q_0, q_1) = \psi(q_0, q_1)\phi^*(q_0) = P(q_1|q_0)P(q_0) = P(q_0, q_1). \tag{18.2}$$

This transformation propagates forward along the chain, changing the separator potentials on q_t into the marginals $P(q_t)$ and the clique potentials on (q_t, q_{t+1}) into the marginals $P(q_t, q_{t+1})$. Thus, all potentials along the backbone of the chain become marginals.

A subsequent pass of DistributeEvidence will have no effect on the potentials along the backbone of the chain (as the reader can verify), but it will convert the potentials $\zeta(q_t)$ into marginals $P(q_t)$ and the potentials $\psi(q_t, y_t)$ into marginals $P(q_t, y_t)$. Thus all potentials throughout the junction tree become marginal probabilities. This result is not surprising, given that it is an easy special case of Theorem 2, but it is reassuring.

Our result also helps to clarify the representation of the joint probability as the product of the clique potentials divided by the product of the separator potentials (cf. Eq. 17.11. While we would not expect to be able to represent the joint in general as the product of marginals such as $P(q_t, q_{t+1})$, we do get this representation if we divide by the separator potentials and those separator potentials are also marginals. Thus, for example, each pairing of a clique potential and a separator potential along the backbone contributes a factor $P(q_t, q_{t+1})/P(q_t)$ to the joint, which is nothing but the original local conditional $P(q_{t+1}|P(q_t)$.

18.1.2 Introducing evidence

We now suppose that the outputs y are observed. We wish to calculate the likelihood $P(y)$ as well as marginal posterior probabilities such as $P(q_t|y)$ and $P(q_t, q_{t+1}|y)$. We return to the original junction tree in which the separator potentials are initialized to unity.

The first step is to alter the potentials to reflect the introduction of the evidence. In the case that y_t is a multinomial node, recall that the potential $\psi(q_t, y_t)$ can be viewed as a matrix B,

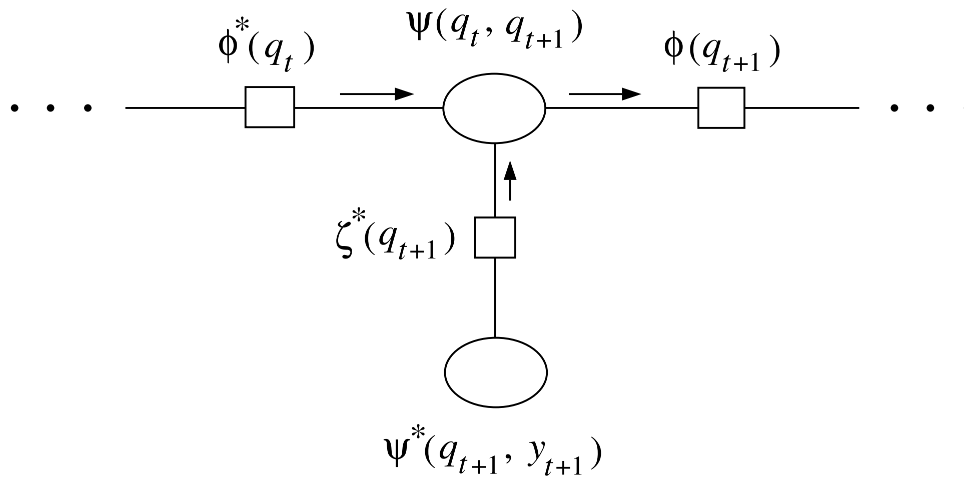

Figure 18.4: A fragment of the junction tree for the HMM.

with columns labeled by the possible values of y_t. Conceptually, observing y_t to be in its ith state corresponds to setting all other columns to zero, so that a subsequent marginalization over y_t simply picks out the ith column of the matrix. In practice we would simply set the separator potential to the desired column. Thus, we have:

$$\zeta^*(q_t) = P(y_t|q_t), \tag{18.3}$$

where y_t is viewed as a fixed constant.

18.1.3 Collecting to the root

We designate the clique (q_{T-1}, q_T) as the root of the junction tree and collect to the root.

Consider the update of clique (q_t, q_{t+1}), as shown in Figure 18.4. We suppose that the preceding separator potential $\phi^*(q_t)$ has already been updated and consider the computation of $\psi^*(q_t, q_{t+1})$ and $\phi^*(q_{t+1})$. Combining the update of clique (q_t, q_{t+1}) based on both of its neighbors (q_{t-1}, q_t) and (q_{t+1}, y_{t+1}), we have:

$$\begin{aligned}\psi^*(q_t, q_{t+1}) &= \psi(q_t, q_{t+1})\phi^*(q_t)\zeta^*(q_{t+1}) && (18.4)\\ &= a_{q_t,q_{t+1}}\phi^*(q_t)P(y_{t+1}|q_{t+1}), && (18.5)\end{aligned}$$

where, as in Chapter 12, we utilize the shorthand $a_{q_t,q_{t+1}} \triangleq P(q_{t+1}|q_t)$. Proceeding forward along the chain, we obtain:

$$\begin{aligned}\phi^*(q_{t+1}) &= \sum_{q_t} \psi^*(q_t, q_{t+1}) && (18.6)\\ &= \sum_{q_t} a_{q_t,q_{t+1}}\phi^*(q_t)P(y_{t+1}|q_{t+1}). && (18.7)\end{aligned}$$

Defining $\alpha(q_t) \triangleq \phi^*(q_t)$, we see that we have recovered exactly the alpha algorithm from Chapter 12 (cf. Eq. 12.22).

Although this definition of $\phi^*(q_t)$ achieves a formal equivalence of the update formulas, is it a reasonable definition? Is $\phi^*(q_t)$ equal to $P(y_0, \ldots, y_t, q_t)$? Indeed, we have $\phi^*(q_0) = P(y_0, q_0)$ by definition, and recursively:

$$\begin{aligned}
\phi^*(q_{t+1}) &= \sum_{q_t} a_{q_t,q_{t+1}} \phi^*(q_t) P(y_{t+1}|q_{t+1}) && (18.8)\\
&= \sum_{q_t} P(q_{t+1}|q_t) P(y_0, \ldots, y_t, q_t) P(y_{t+1}|q_{t+1}) && (18.9)\\
&= \sum_{q_t} P(y_0, \ldots, y_t, y_{t+1}, q_t, q_{t+1}) && (18.10)\\
&= P(y_0, \ldots, y_t, y_{t+1}, q_{t+1}), && (18.11)
\end{aligned}$$

so the definition is justified.

It is also possible to develop an alternative approach to the forward inference problem by specifying a recurrence on the (q_t, q_{t+t}) clique potentials. In fact, given that $\phi^*(q_t) = \sum_{q_{t-1}} \psi^*(q_{t-1}, q_t)$, substitution in Eq. 18.5 yields:

$$\psi^*(q_t, q_{t+1}) = a_{q_t,q_{t+1}} \sum_{q_{t-1}} \psi^*(q_{t-1}, q_t) P(y_{t+1}|q_{t+1}), \tag{18.12}$$

The reader can verify that $\psi^*(q_t, q_{t+1}) = P(y_0, \ldots, y_{t+1}, q_t, q_{t+1})$. Thus, defining the variable $\rho(q_t, q_{t+1}) \triangleq P(y_0, \ldots, y_{t+1}, q_t, q_{t+1})$, we have established the recurrence relation:

$$\rho(q_t, q_{t+1}) = a_{q_t,q_{t+1}} \sum_{q_{t-1}} \rho(q_{t-1}, q_t) P(y_{t+1}|q_{t+1}), \tag{18.13}$$

which is an alternative to the traditional alpha algorithm. Indeed, in Section 18.1.5 we will establish a backward recurrence involving the cliques (q_t, q_{t+1}) which, together with Eq. 18.13 will yield an alternative approach to HMM inference that produces the "xi" variables directly and the "alpha/gamma" variables indirectly.

The collect phase of the algorithm terminates with the update of $\psi(q_{T-1}, q_T)$. The updated potential will equal $P(y_0, \ldots, y_T, q_t, q_{t+1})$, and thus by marginalization:

$$P(y) = \sum_{q_{T-1},q_T} \psi^*(q_{T-1}, q_T) \tag{18.14}$$

we obtain the likelihood $P(y)$.

18.1.4 An alternative root

Suppose that instead of designating clique (q_{T-1}, q_T) as the root node of the junction tree, we instead utilize (q_0, q_1) as the root. Exercise XXX studies this case, showing that the result is the beta algorithm. Thus, we find that $\phi^*(q_t) = P(y_{t+1}, \ldots, y_T|q_t)$ and, moreover, the junction tree recursion linking $\phi^*(q_t)$ and $\phi^*(q_{t+1})$ is exactly the beta recursion of Eq. 12.30.

It is not necessary, however, to change the root of the junction tree to derive the beta algorithm. As we show in Section 18.1.5, the beta algorithm arises during the DistributeEvidence pass when utilizing clique (q_{T-1}, q_T) as the root. This is the preferred way to map the traditional HMM inference algorithms onto the junction tree machinery.

A final comment regarding the forward/backward algorithms and the notion of *filtering*; that of obtaining the conditional expectation of a state given a partial observation sequence. Note that the beta variables are conditionals $P(y_{t+1}, \ldots, y_T|q_t)$, involving the probability of a partial evidence sequence given the state. The alpha variables, on the other hand, are marginals $P(y_0, \ldots, y_t, q_t)$, involving the probability of a partial evidence sequence *and* the state at time t. Converting the latter variables to filtered quantities of the form $P(q_t|y_0, \ldots, y_t)$ is straightforward; one simply normalizes. Converting the beta variables to (backward) filtered quantities of the form $P(q_t|y_{t+1}, \ldots, y_T)$, on the other hand, is not so straightforward. Obtaining this conditional requires us to know $P(q_t)$, which is not available in the junction tree.

Suppose, however, that we consider a pre-initialized junction tree in which the inference algorithm has been performed without evidence. We saw in Section 18.1.1 that this procedure converts the potentials throughout the tree, including the separator potentials, into marginal probabilities. Thus, a marginal such as $P(q_t)$ is available in this tree, and we might expect to obtain a marginal version of the beta algorithm if we subsequently collect to the root. Indeed, in Exercise XXX we verify that this procedure yields $\phi^*(q_t) = P(y_{t+1}, \ldots, y_T, q_t)$. This quantity is readily converted to the filtered estimate $P(q_t|y_{t+1}, \ldots, y_T)$ by normalization.

18.1.5 Distributing from the root

We now return to our main thread, and consider running the DistributeEvidence algorithm from the root (q_{T-1}, q_T). We assume that we have already collected to this root, and thus the potentials at the outset of the DistributeEvidence are those discussed in Section 18.1.3.

The DistributeEvidence phase proceeds backward along the backbone of the state-to-state cliques as well as downward into the state-to-output cliques. Given that we have already obtained the likelihood, which is the only information regarding the probability of the outputs that is generally of interest, we restrict our attention to the updates along the backbone.

Referring to Figure 18.5, we suppose that the preceding separator potential $\phi^{**}(q_{t+1})$ has already been updated and consider the update of $\psi^{**}(q_t, q_{t+1})$ and $\phi^{**}(q_t)$. We have:

$$\psi^{**}(q_t, q_{t+1}) = \psi^*(q_t, q_{t+1}) \frac{\phi^{**}(q_{t+1})}{\phi^*(q_{t+1})}, \tag{18.15}$$

and proceeding backward a further step:

$$\begin{aligned} \phi^{**}(q_t) &= \sum_{q_{t+1}} \frac{\psi^*(q_t, q_{t+1})}{\phi^*(q_{t+1})} \phi^{**}(q_{t+1}) && (18.16) \\ &= \sum_{q_{t+1}} \frac{\psi^*(q_t, q_{t+1})}{\sum_{q_t} \psi^*(q_t, q_{t+1})} \phi^{**}(q_{t+1}). && (18.17) \end{aligned}$$

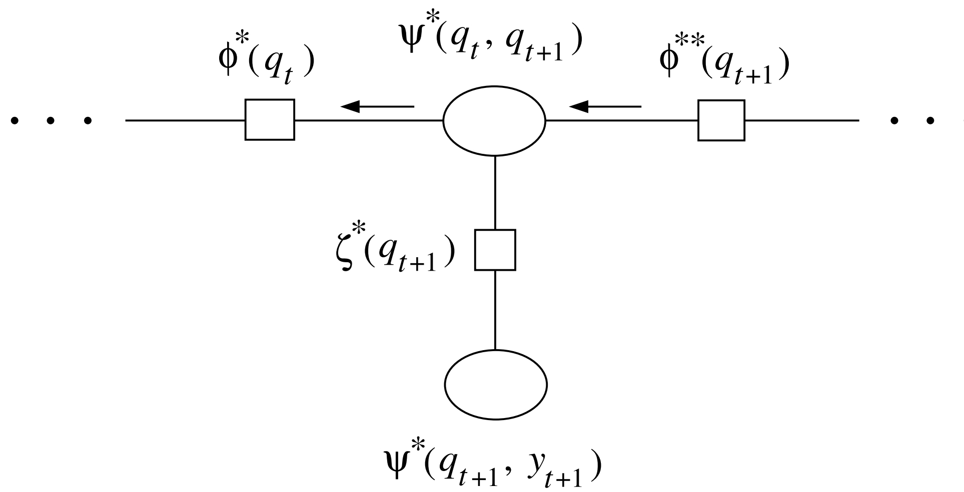

Figure 18.5: A fragment of the junction tree for the HMM with backward-going messages.

Substituting from Eq. 18.5 we have:

$$\phi^{**}(q_t) = \sum_{q_{t+1}} \frac{a_{q_t,q_{t+1}}\phi^*(q_t)}{\sum_{q_t} a_{q_t,q_{t+1}}\phi^*(q_t)}\phi^{**}(q_{t+1}) \tag{18.18}$$

$$= \sum_{q_{t+1}} \frac{a_{q_t,q_{t+1}}\alpha(q_t)}{\sum_{q_t} a_{q_t,q_{t+1}}\alpha(q_t)}\phi^{**}(q_{t+1}). \tag{18.19}$$

Defining $\gamma(q_t)$ as equal to $\phi^{**}(q_t)$, up to a constant of proportionality, we see that we have recovered the gamma recursion from Chapter 12 (cf. Eq. 12.41).

Once again we can reassure ourselves that $\gamma(q_t)$ defined this way is indeed proportional to the posterior probability $P(q_t|y_0,\ldots,y_T)$. This can be done by direct calculation (cf. Exercise XXX). Alternatively, we can trust the general theory of Chapter 17, which assures us that the potentials produced by the junction tree algorithm must be proportional to the posterior probabilities. Indeed, it is easy to verify that $\phi^{**}(q_t) = P(y_0,\ldots,y_T,q_t)$ and thus the proportionality constant is the likelihood $P(y)$. Once the junction tree algorithm has run, we can obtain the likelihood by normalizing any of the potentials.

In summary, we have identified the alpha and the gamma variables in the junction tree algorithm, and have derived the "alpha-gamma" recursion discussed in Chapter 12. We can also derive the "alpha-beta" recursion via the junction tree algorithm. Consider in particular the update factor $\phi^{**}(q_t)/\phi^*(q_t)$:

$$\frac{\phi^{**}(q_t)}{\phi^*(q_t)} = \frac{\sum_{q_{t+1}} \psi^{**}(q_t, q_{t+1})}{\phi^*(q_t)} \tag{18.20}$$

$$= \sum_{q_{t+1}} \frac{\psi^*(q_t, q_{t+1})}{\phi^*(q_t)} \frac{\phi^{**}(q_{t+1})}{\phi^*(q_{t+1})} \tag{18.21}$$

$$= \sum_{q_{t+1}} \frac{a_{q_t,q_{t+1}}\phi^*(q_t)P(y_{t+1}|q_{t+1})}{\phi^*(q_t)} \frac{\phi^{**}(q_{t+1})}{\phi^*(q_{t+1})} \tag{18.22}$$

$$= \sum_{q_{t+1}} a_{q_t,q_{t+1}} P(y_{t+1}|q_{t+1}) \frac{\phi^{**}(q_{t+1})}{\phi^*(q_{t+1})}. \tag{18.23}$$

where we have used Eq. 18.5 in the third equality. Defining $\beta(q_t) \triangleq \phi^{**}(q_t)/\phi^*(q_t)$, we have recovered the beta recursion from Chapter 12 (cf. Eq. 12.30). Moreover, putting together our definitions, we have:

$$\alpha(q_t)\beta(q_t) = \phi^*(q_t)\frac{\phi^{**}(q_t)}{\phi^*(q_t)} = \phi^{**}(q_t) \propto \gamma(q_t) \tag{18.24}$$

and thus we have also recovered the original definition of $\gamma(q_t)$ in Eq. 12.13.

Finally, it is also of interest to note that the junction tree algorithm gives us an explicit recursion in the variables $\xi(q_t, q_{t+1})$ (cf. Eq. 12.42). Starting from Eq. 18.15, we have:

$$\psi^{**}(q_{t-1}, q_t) = \psi^*(q_{t-1}, q_t)\frac{\phi^{**}(q_t)}{\phi^*(q_t)} \tag{18.25}$$

$$= \frac{\psi^*(q_{t-1}, q_t)}{\phi^*(q_t)} \sum_{q_{t+1}} \psi^{**}(q_t, q_{t+1}) \tag{18.26}$$

$$= \frac{\rho(q_{t-1}, q_t)}{\sum_{q_{t-1}} \rho(q_{t-1}, q_t)} \sum_{q_{t+1}} \psi^{**}(q_t, q_{t+1}). \tag{18.27}$$

Defining $\xi(q_t, q_{t+1})$ to be equal to $\psi^{**}(q_t, q_{t+1})$, again up to a constant of proportionality which turns out to the likelihood, we have obtained an explicit recursion for the $\xi(q_t, q_{t+1})$ variables:

$$\xi(q_{t-1}, q_t) = \frac{\rho(q_{t-1}, q_t)}{\sum_{q_{t-1}} \rho(q_{t-1}, q_t)} \sum_{q_{t+1}} \xi(q_t, q_{t+1}). \tag{18.28}$$

This "rho-xi algorithm" is the analog of the "alpha-gamma algorithm."

Once the rho-xi algorithm has run, $\alpha(q_t)$ and $\gamma(q_t)$ can be obtained via:

$$\alpha(q_t) = \sum_{q_{t-1}} \rho(q_{t-1}, q_t) \tag{18.29}$$

$$\gamma(q_t) = \sum_{q_{t-1}} \xi(q_{t-1}, q_t), \tag{18.30}$$

which are verified by plugging in the corresponding junction tree definitions.

18.2 Linear Gaussian models

We now turn to the linear Gaussian model (the "LG-HMM"). We rederive two of the LG-HMM inference algorithms—a forward (filtering) algorithm and a backward (smoothing) algorithm—from Chapter 15 in order to exemplify the relationships between the junction tree and the earlier material.

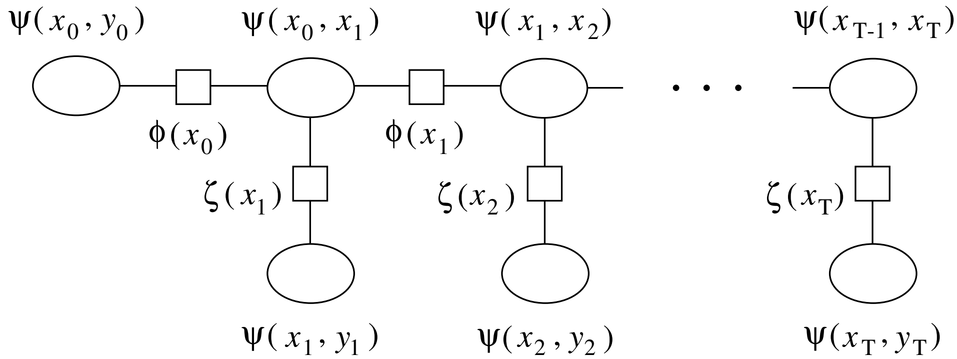

Figure 18.6: A junction tree for the LG-HMM with the potentials labeled.

Recall from Chapter 15 that the LG-HMM has the same graphical structure as the HMM, the difference being the node types and the parameterization. Thus, as before, Figure 18.1(a) is the graphical model for the LG-HMM (substituting "x" for "q") and Figure 18.1(b) is the moralized, triangulated graph. Also, the junction tree for the LG-HMM, shown in Figure 18.6, is identical to that for the HMM.[2] Figure 18.6

The potentials in the junction tree are Gaussian potentials on pairs of nodes (the cliques) or singletons (the separators). We set up an assignment of local conditional probabilities to clique potentials that parallels that of the HMM. In particular, clique (x_0, y_0) is assigned the initial state probability $P(x_0)$ and the conditional $P(y_0|x_0)$, and thus $\psi(x_0, y_0)$ is initialized to the marginal $P(x_0, y_0)$. All of the other clique potentials are conditionals; in particular the state-to-state clique potential $\psi(x_t, x_{t+1})$ is set equal to $P(x_{t+1}|x_t)$ and the state-to-output potential $\psi(x_t, y_t)$ is set equal to $P(y_t|x_t)$.

Consider first the potential $\psi(x_t, x_{t+1})$. Given the dynamical equation:

$$x_{t+1} = Ax_t + Gw_t, \tag{18.31}$$

the conditional probability $P(x_{t+1}|x_t)$ is a multivariate Gaussian with mean Ax_t and covariance matrix GQG^T, where Q is the covariance matrix of w_t. Letting $H \triangleq GQG^T$, we have:

$$P(x_{t+1}|x_t) \propto \exp\left\{-\frac{1}{2}(x_{t+1}) - Ax_t)^T H^{-1}(x_{t+1}) - Ax_t)\right\}. \tag{18.32}$$

Expressing the quadratic form as a function of two variables, we obtain:

$$\psi(x_{t+1}|x_t) = \exp\left\{-\frac{1}{2}\begin{pmatrix} x_t \\ x_{t+1} \end{pmatrix}^T \begin{bmatrix} A^T H^{-1} A & -A^T H^{-1} \\ H^{-1} A & H^{-1} \end{bmatrix} \begin{pmatrix} x_t \\ x_{t+1} \end{pmatrix}\right\} \tag{18.33}$$

Note that in this representation, and in all subsequent potential function representations of Gaussians, we carry along only the exponential factor and ignore the Gaussian normalization factor.

[2] As in the case of the HMM, it is a worthwhile exercise to investigate the algorithms that arise from alternative choices for the junction tree.

The reason for this is as follows. At the end of the junction tree algorithm we are guaranteed that the potentials are equal to marginal probabilities, up to a normalization factor. Now, all of the potentials in the LG-HMM are Gaussian, both before and after the running of the inference algorithm. This implies that we can simply read off the normalization factors from the final form of the potentials; we are not required to explicitly normalize before or during the inference procedure.

From the output equation of the LG-HMM:

$$y_t = Cx_t + v_t, \tag{18.34}$$

we have that $P(y_t|x_t)$ is a multivariate Gaussian with mean Cx_t and covariance matrix R, where R is the covariance matrix of v_t. This yields:

$$P(y_t|x_t) \propto \exp\left\{-\frac{1}{2}(y_t - Cx_t)^T R^{-1}(y_t - Cx_t)\right\}. \tag{18.35}$$

We can proceed as before and convert this to a bivariate representation for $\psi(x_t, y_t)$. Note, however, that y_t is an observed constant in applications of the LG-HMM. This implies that the marginalization step that yields $\zeta^*(x_t)$ simply reproduces Eq. 18.35 for that observed value of y_t (i.e., we integrate $\psi(x_t, y_t)$ against a delta function). We thus leave Eq. 18.35 unexpanded in anticipation of its later role as $\zeta^*(x_t)$.

Multivariate Gaussian potentials can be represented in terms of either moments (μ, Σ), or canonical parameters (ξ, Λ). In Chapter 13 we derived the formulas for marginalization and conditioning in both representations. In the context of the junction tree algorithm we also need to multiply potentials. This is significantly easier in the canonical parameterization. Thus, if we have potentials $\psi_1(x) = \exp \xi_1^T x - 1/2 x^T K_1 x$ and $\psi_2(x) = \exp \xi_2^T x - 1/2 x^T K_2 x$, the product is a potential $\psi(x) = \exp \xi^T x - 1/2 x^T K x$, where:

$$K = K_1 + K_2 \tag{18.36}$$

$$\xi = \xi_1 + \xi_2. \tag{18.37}$$

Expressing this operation in terms of moments requires an application of the inverse matrix lemma. On the other hand, if our interest is in obtaining filtered or smoothed estimates of the states—i.e., moments—then we if we develop an algorithm in terms of canonical parameters we will require an application of the inverse matrix lemma at the end. The situation is "pay now or pay later."

Moreover, in the context of linear Gaussian systems, moments behave particularly nicely and calculations on moments can save substantial labor. If we take the junction tree algorithm literally and require ourselves to multiply the Gaussian potentials then we miss out on this opportunity. Indeed, in doing the necessary inverse matrix operations we are essentially rederiving the fact that the parameters μ and Σ are moments of the Gaussian distribution. It is also possible to replace the multiplication step of the junction tree algorithm with a step that calculates the moments of the product directly, utilizing the probabilistic interpretation of the potentials. For example, suppose that we know that $\phi(x)$ is proportional to the marginal of x and we have that $\psi(x, y)$ is proportional to the conditional of y given x. Moreover, let $y = Cx + v$. In this case we can directly compute $E[y]$, $\mathrm{Var}[y]$ and $\mathrm{Cov}[x, y]$ in terms of the moment parameterization of $\phi(x)$, thereby obtaining the moment parameterization of $\psi^*(x, y) = \psi(x, y)\phi(x)$.

In this chapter we generally take the more literal interpretation of the junction tree algorithm and work in the canonical representation (with the notable exception of the following section, where we exploit moment calculations). Exercise XXX asks the reader to develop the moment-based approach more generally, in particular relating this approach to the derivation of the Kalman filter that we presented in Chapter 15.

18.2.1 Unconditional inference

As in the HMM case it is useful to consider running the junction tree inference algorithm before any evidence has been observed. Suppose that we designate the node (x_{T-1}, x_T) as the root and collect to the root.

The derivation that we carried out in Section 18.1.1 for the HMM was entirely generic, and (replacing the sums with integrals) we obtain the same results for the LG-HMM. In particular, once again the operation of passing a message upward from the leaves (x_t, y_t) to the cliques (x_t, x_{t+1}), for $t > 1$, has no effect. The operation of passing messages from (x_0, y_0) to (x_0, x_1) and subsquently along the backbone of cliques (x_t, x_{t+1}) has the effect of changing all of those clique potentials, as well as the separator potentials, to marginal probabilities.

The form that these marginal probabilities take is readily obtained via moment calculations. Let (μ_t, Σ_t) denote the mean and covariance matrix of x_t. Given the dynamical equation $x_{t+1} = Ax_t + Gw_t$, and given our assumption that the initial state has mean zero, we see that $\mu_t = 0$ for all t. As for the covariance matrix, we obtain:

$$\Sigma_{t+1} = A\Sigma_t A^T + H, \tag{18.38}$$

which is the *Lyapunov equation* of Chapter 15 (cf. Eq. 15.6). Thus after the CollectEvidence phase, the separator potentials can be represented by $(0, \Sigma_t)$, or by $(0, \Sigma_t^{-1})$ in canonical parameters. The clique potential $\psi(x_t, x_{t+1})$ is represented in moment parameters by:

$$\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \qquad \begin{bmatrix} \Sigma_t & -\Sigma_t A \\ -A^T\Sigma_t & A\Sigma_t A^T + H \end{bmatrix} \tag{18.39}$$

The representation in terms of canonical parameters can be obtained by inverting the covariance matrix (using the partitioned matrix inverse theorem, Eq. 13.16):

$$\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \qquad \begin{bmatrix} A^T H^{-1} A + \Sigma_t^{-1} & -A^T H^{-1} \\ -H^{-1}A & H^{-1} \end{bmatrix} \tag{18.40}$$

18.2.2 Introducing evidence

We return to the original junction tree in which the separator potentials are initialized to unity and now suppose that the outputs y are observed.

As in the case of the HMM we are not interested in the potential on (x_t, y_t) beyond its effect on the separator potential $\zeta^*(x_t)$. Moreover, $\zeta^*(x_t)$ is obtained by simply evaluating $P(y_t|x_t)$ at the observed value y_t:

$$\zeta^*(x_t) = \exp\left\{-\frac{1}{2}(y_t - Cx_t)^T R^{-1}(y_t - Cx_t)\right\}. \tag{18.41}$$

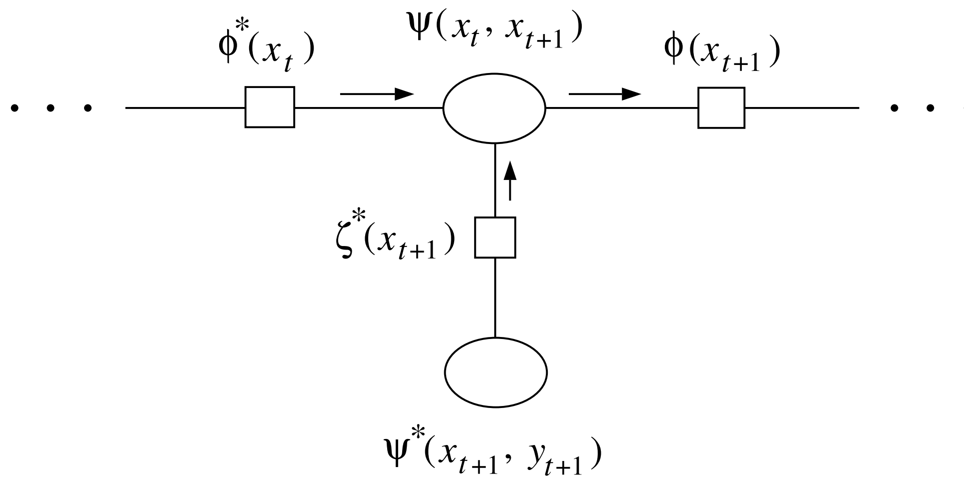

Figure 18.7: A fragment of the junction tree for the LG-HMM.

In canonical parameters this potential can be represented as: $(-C^T y_t, C^T R^{-1} C)$.

18.2.3 A forward algorithm

We designate the clique (x_{T-1}, x_T) as the root of the junction tree and collect to the root.

Let $(\hat{\xi}_{t|t}, S_{t|t})$ be the canonical parameters of the distribution of x_t conditioned on $(y_0, \ldots, y_t)$. Similarly, let $(\hat{\xi}_{t+1|t}, S_{t+1|t})$ be the canonical parameters of the distribution of x_t conditioned on $(y_0, \ldots, y_{t+1})$. Our goal is to establish recursions for these quantities.

Consider the junction tree fragment shown in Figure 18.7. We suppose that the preceding separator potential Suppose that we have already obtained an updated $\phi^*(x_t)$ and wish to update $\psi(x_t, x_{t+1})$. We know from our general theory and (our experience with the HMM) that $\phi^*(x_t)$ must be proportional to $P(x_t|y_0, \ldots, y_t)$; thus, we immediately identify the canonical parameters of $\phi^*(x_t)$ with $(\hat{\xi}_{t|t}, S_{t|t})$. Using Eq. 18.33 we now update $\psi(x_t, x_{t+1})$ by multiplying by $\phi^*(x_t)$. This yields the following canonical parameters for the updated potential:

$$\begin{bmatrix} \hat{\xi}_{t|t} \\ 0 \end{bmatrix}, \qquad \begin{bmatrix} S_{t|t} + A^T H^{-1} A & -A^T H^{-1} \\ H^{-1} A & H^{-1} \end{bmatrix}. \tag{18.42}$$

If we now marginalize this potential with respect to x_t, we should expect to obtain a representation for $P(x_{t+1}|y_0, \ldots, y_t)$. Indeed, applying Eq. 13.29 and Eq. 13.29, we obtain:

$$\hat{\xi}_{t+1|t} = H^{-1} A (S_{t|t} + A^T H^{-1} A)^{-1} \hat{\xi}_{t|t} \tag{18.43}$$

$$S_{t+1|t} = H^{-1} - H^{-1} A (S_{t|t} + A^T H^{-1} A)^{-1} A^T H^{-1}. \tag{18.44}$$

These two equations are identical to the time updates for the information filter in Chapter 15 (cf. Eq. 15.49 and Eq. 15.39).

The next step is to incorporate the evidence y_{t+1} by multiplying the updated clique potential on (x_t, x_{t+1}) by $\zeta^*(x_{t+1})$. This yields the following canonical parameterization for (x_t, x_{t+1}):

$$\begin{bmatrix} \hat{\xi}_{t|t} \\ C^T R^{-1} y_{t+1} \end{bmatrix}, \qquad \begin{bmatrix} S_{t|t} + A^T H^{-1} A & -A^T H^{-1} \\ H^{-1} A & H^{-1} + C^T R^{-1} C \end{bmatrix}. \tag{18.45}$$

Once again we marginalize this potential with respect to x_t. The result is the canonical parameterization of $\phi^*(x_{t+1})$:

$$\hat{\xi}_{t+1|t+1} = C^T R^{-1} y_{t+1} + H^{-1} A (S_{t|t-1} + A^T H^{-1} A)^{-1} \hat{\xi}_{t|t} \tag{18.46}$$

$$= \hat{\xi}_{t+1|t} + C^T R^{-1} y_{t+1} \tag{18.47}$$

$$S_{t+1|t+1} = H^{-1} + C^T R^{-1} C - H^{-1} A (S_{t|t} + A^T H^{-1} A)^{-1} A^T H^{-1} \tag{18.48}$$

$$= S_{t+1|t} + C^T R^{-1} C. \tag{18.49}$$

These results are identical to the measurement updates for the information filter in Chapter 15 (cf. Eq. 15.54 and Eq. 15.43).

18.2.4 A backward algorithm

In Section15.7.2 we derived a backward algorithm for the LG-HMM by explicitly inverting the dynamics of the linear-Gaussian model and applying the information filter to the inverted dynamics. Recall that this approach yielded filtered estimates, i.e., conditional probabilities $P(x_t|y_{t+1}, \ldots, y_T)$, rather than the usual "beta variables" $P(y_{t+1}, \ldots, y_T|x_t)$.

In this section we see that this algorithm emerges in a straightforward way from the junction tree algorithm. In particular, suppose that we pre-initialize the junction tree by running a forward pass without evidence. From Section 18.2.1 we know that this pre-initialization pass leaves marginal probabilities on both the clique and the separator potentials. We would expect that a subsequent backward pass in the pre-initialized tree should yield filtered estimates $P(x_t|y_{t+1}, \ldots, y_T)$.

Note in particular that whereas in Section15.7.2 we had to invert the dynamics explicitly, this is not necessary in the current approach. The junction tree algorithm effectively inverts the dynamics for us. (In particular the derivation does not assume that A is an invertible matrix).

The derivation of the algorithm is similar to that of the previous section; the main difference being that we must remember to divide by the pre-initialized separator potentials. These potentials have the canonical representation $(0, \Sigma_t^{-1})$. Recall also that the initial values of the clique potentials are given by Eq. 18.40.

To obtain a two-step procedure in the backwards direction, it is useful to utilize an alternative junction tree in which the clique (x_t, y_t) is attached to the clique (x_t, x_{t+1}) rather than (x_{t-1}, x_t). This junction tree is shown in Figure 18.8 where we see that it is the final clique and not the initial clique that has two state-to-output neighbors. Note that in this junction tree the evidence node (x_t, y_t) is to the left of the separator (x_{t+1}), just as the evidence node (x_{t+1}, y_{t+1}) was to the to the right of the separator (x_t) in the earlier junction tree (Figure 18.2). As we will see, this allows us to obtain the usual two-step time and measurement updates from the junction tree.

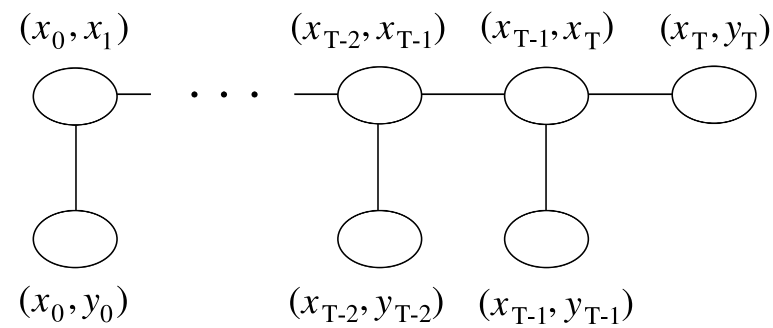

Figure 18.8: An alternative maximal spanning tree for the LG-HMM.

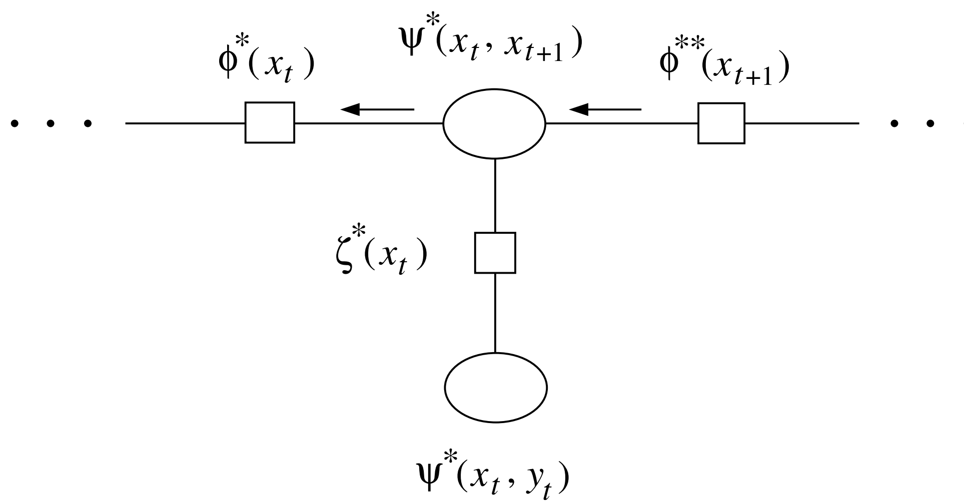

Figure 18.9: A fragment of the alternative junction tree for the LG-HMM.

We initially set the potentials of all cliques and separators to be proportional to the corresponding (unconditional) marginal probabilities. In particular, the potential $\psi(x_T, y_T)$ at the end of the chain is set proportional to $P(x_T, y_T) = P(y_T|x_T)P(x_T)$. We can conceptually view this assignment of potentials as resulting from a forward pass in the unconditioned graph (cf. Section 18.2.1).

Let us denote the canonical parameters of $\phi^*(x_{t+1})$ by $(\hat{\xi}_{t+1|t+1}, S_{t+1|t+1})$; note that this potential reflects the evidence from y_{t+1} onward. We now update $\psi(x_t, x_{t+1})$ by multiplying by $\phi^*(x_{t+1})$ and dividing by $\phi(x_{t+1})$, where the latter refers to the potential that was obtained unconditionally (see Figure 18.9). The parameters of the updated potential are:

$$\begin{bmatrix} 0 \\ \hat{\xi}_{t+1|t+1} \end{bmatrix}, \qquad \begin{bmatrix} A^T H^{-1} A + \Sigma_t^{-1} & -A^T H^{-1} \\ H^{-1} A & S_{t+1|t+1} + H^{-1} - \Sigma_{t+1}^{-1} \end{bmatrix}, \tag{18.50}$$

where the $S_{t+1|t+1}$ in the lower-right-hand corner is due to multiplication by $\psi(x_t, x_{t+1})$ and the Σ_{t+1}^{-1} is due to division by $\phi(x_{t+1})$.

We now marginalize with respect to x_{t+1} to obtain the canonical representation of $P(x_t|y_{t+1},\ldots,y_T)$:

$$\begin{aligned}\hat{\xi}_{t|t+1} &= A^T H^{-1}(S_{t+1|t+1} + H^{-1} - \Sigma_{t+1}^{-1})^{-1}\hat{\xi}_{t+1|t+1} && (18.51)\\ S_{t|t+1} &= A^T H^{-1}(S_{t+1|t+1} + H^{-1} - \Sigma_{t+1}^{-1})^{-1}H^{-1}A. && (18.52)\end{aligned}$$

These updates are identical to Eq. 15.98 and Eq. 15.96.

We now update the clique potential on (x_t, x_{t+1}) to reflect the evidence y_t. This amounts to adding the term $C^T R^{-1} C$ to the inverse covariance matrix and $C^T R^{-1} y_t$ to the linear term:

$$\begin{bmatrix} C^T R^{-1} y_t \\ \hat{\xi}_{t+1|t+1} \end{bmatrix}, \qquad \begin{bmatrix} A^T H^{-1} A + \Sigma_t^{-1} + C^T R^{-1} C & -A^T H^{-1} \\ H^{-1} A & S_{t+1|t+1} + H^{-1} - \Sigma_{t+1}^{-1} \end{bmatrix}. \tag{18.53}$$

Marginalizing with respect to x_{t+1} yields the canonical representation of $P(x_t|y_t,\ldots,y_T)$:

$$\begin{aligned}\hat{\xi}_{t|t} &= \hat{\xi}_{t|t+1} + C^T R^{-1} y_t && (18.54)\\ S_{t|t+1} &= S_{t|t+1} + C^T R^{-1} C, && (18.55)\end{aligned}$$

which are identical to Eq. 15.98 and Eq. 15.97.

18.3 Summary

In this chapter we have shown how to derive many of the classical algorithms associated with the HMM and the LG-HMM from the point of view of the junction tree framework. Although we have not derived all possible algorithms, we have provided a representative sampling that shows several of the tricks of the trade.

A virtue of the junction tree formalism is that it displays all of relevant dependencies and their interrelationships; in particular, the pairwise clique potentials in the case of HMMs and LG-HMMs. This is a useful display for deriving algorithms, whether or not one uses the junction tree update formulas literally as we have done in this chapter, or uses the junction tree structure to derive alternative update formulas (e.g., based on moments).

18.4 Historical remarks and bibliography