Hidden Markov Models

Tuan Van Pham

Advanced Signal Processing 2

30 May 2005, TU Graz

Outlines

- **Introduction.**
- **Discrete Markov Processes.**
- **Problems and Solutions for HMMs.**
- **Connections to Graphical Model.**
- **Kalman Filters.**
- **Conclusions.**

Introduction (1/2)

▶ Statistical Modeling Aspects

- Characterization of real-world signals in terms of signal models:
  –> Theoretical description; Learning ability.
- Choices for types of signal models:
  –> Deterministic models; Stochastic models (Poisson, HMM, ...).
- Why use HMMs ?
  –> Answer the question: "If I have a set of output symbols, what was the sequence of states & transitions that resulted in those output symbols ?"
- HMM is a powerful modern statistical technique. Why ?
- Identification & manipulation of conditional independence assumptions.

Introduction (2/2)

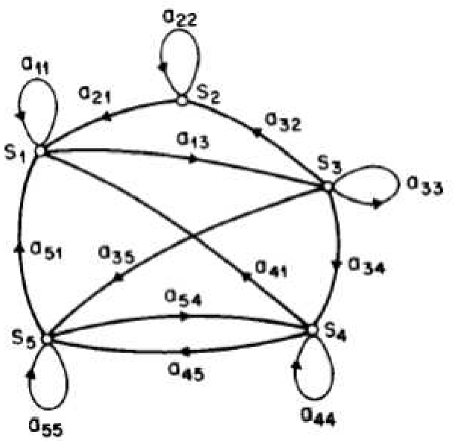
▶ Graphical Modeling Aspects

- Using of GRAPH to represent independent structure of probability models.
- Relationships between conditional independence in probability model & structural properties of graph.
- HMMs as DAGs:
  - Inference (forward-backward algorithm)
  - MAP (Viterbi algorithm)

  –> Graphical modeling provides an automatic method. How ?
  - Inference (Jensen, Lauritzen & Oleson's algorithm)
  - MAP (Dawid's algorithm)
- Kalman Filter as DAGs.

Discrete Markov Processes

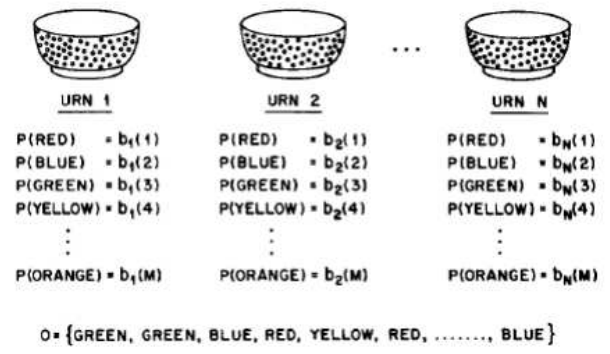
From Markov Chain to HMM



- Probabilistic description:

$$P(q_{t+1} = S_j \mid, q_t = S_i, q_{t-1} = S_k, \ldots)$$
$$= P(q_{t+1} = S_j \mid, q_t = S_i).$$

- –> Observable Markov Model since **output is set of states**.



- Markov model where observation is a probabilistic function of state.
- HMM: underlying stochastic process (**that is hidden**) can only be observed through another set of stochastic processes that produce the sequence of observations.

Discrete Markov Processes

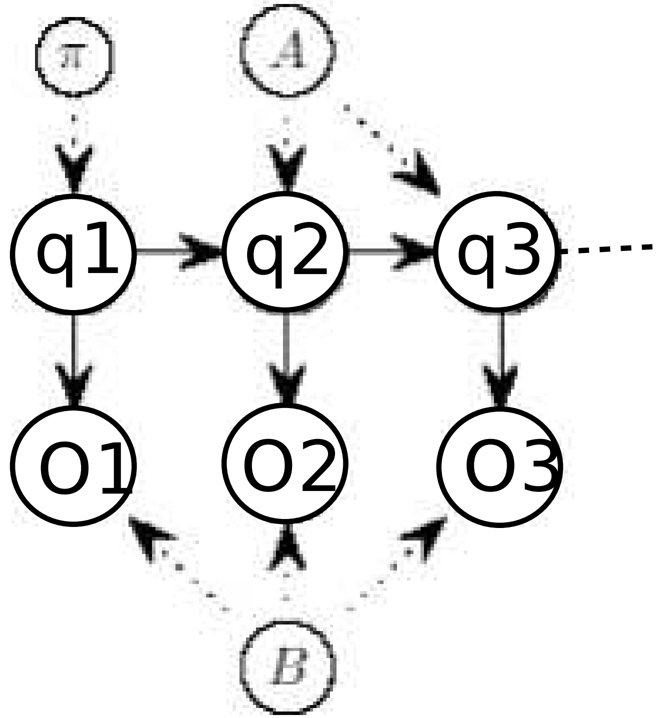
Elements of an HMM

- **N**: number of states in the model. (Individual states as $S = S_1, S_2, ..., S_N$. State at time t as q_t.)
- **M**: number of distinct observation symbols per state. (Individual symbols as $V = V_1, V_2, ..., V_M$.)
- $A = a_{ij}$: state transition probability distribution
$$a_{i,j} = P(q_{t+1} = S_j \mid q_t = S_i), \quad 1 \leq i, j \leq N.$$
- $B = b_j(k)$: observation symbol probability distribution in state j
$$b_j(k) = P(V_k \; at \; t \mid q_t = S_j), \quad 1 \leq \; j \leq N, 1 \leq \; k \leq M.$$
- $\pi = \pi_i$: initial state distribution
$$\pi_i = q_1 = S_i, \quad 1 \leq \; i \leq N.$$

Discrete Markov Processes

Generating observation sequence by HMM

1) Choose an initial state $q_1 = S_i$ according to the initial state distribution π.
2) Set $t = 1$.
3) Choose $O_t = v_k$ according to the symbol probability distribution in state S_i, i.e., $b_i(k)$.
4) Transit to a new state $q_{t+1} = S_j$ according to the state transition probability distribution for state S_i, i.e., a_{ij}.
5) Set $t = t + 1$; return to step 3)if $t < T$; otherwise terminate the procedure.



Problems & Solutions for HMM

Three basic problems of HMMs

Problems	Interpretation
Problem 1: Given $O = O_1 O_2 ... O_T$, and a model $\lambda = (A, B, \pi)$, compute $P(O \mid \lambda)$?	Evaluation / Scoring. **–> Forward-Backward**.
Problem 2: Given $O = O_1 O_2 ... O_T$, and a model $\lambda = (A, B, \pi)$, choose state sequence $Q = q_1 q_2 ... q_T$ which best explain O ?	Find the optimal state sequence / Decoding. **–> Viterbi**.
Problem 3: Adjust model parameters $\lambda = (A, B, \pi)$ to maximize $P(O \mid \lambda)$?	Reevaluation / Learning. **–> Baum-Welch (EM)**.

(Connection to Inference and MAP problems in Graphical Model ?)

Problems & Solutions for HMM

Assumptions in the theory of HMMs

- **Markov assumption:** *"The next state is dependent only upon the current state"*
$$a_{i,j} = P(q_{t+1} = S_j \mid q_t = S_i) \quad 1 \leq i, j \leq N.$$
- **Stationarity assumption:** *"The state transition probabilities are independent of the actual time at which the transitions takes place"*
$$P(q_{t_1+1} = S_j \mid q_{t_1} = S_i) = P(q_{t_2+1} = S_j \mid q_{t_2} = S_i)$$
- **Statistical independence assumption:** *"The current observation is statistically independent of the previous observations"*
$$O = O_1 O_2 ... O_T; \quad Q = q_1 q_2 ... q_T$$
$$P(O \mid Q, \lambda) = \Pi_{t=1}^{T} P(O_t \mid q_t, \lambda)$$

Problems & Solutions for HMM

Solution to Problem 1: Straightforward method (1/3)

- Accounting for every possible state sequence $Q = q_1 q_2 ... q_t$
- Probability of a state sequence Q is:

$$P(Q \mid \lambda) = \pi_{q_1} a_{q_1 q_2} a_{q_2 q_3} ... a_{q_{T-1} q_T}.$$

- Probability of the observation sequence O given state Q:

$$P(O \mid Q, \lambda) = \Pi_{t=1}^{T} P(O_t \mid q_t, \lambda) = b_{q_1}(O_1) b_{q_2}(O_2) ... b_{q_T}(O_T).$$

- Probability of O: summing joint probability $P(O, Q \mid \lambda)$ over Q:

$$P(O \mid \lambda) = \sum_{all\ Q} P(O, Q \mid \lambda) = \sum_{all\ Q} P(O \mid Q, \lambda) P(Q \mid \lambda).$$

$$P(O \mid \lambda) = \sum_{all\ Q} \pi_{q_1} b_{q_1}(O_1) a_{q_1 q_2} b_{q_2}(O_2) ... a_{q_{T-1} q_T} b_{q_T}(O_T).$$

- Complexity $O(2TN^T)$ –> computationally intractable.

Problems & Solutions for HMM

Solution to Problem 1: F-B algorithm (2/3)

- Consider forward variable $\alpha_t(i)$:

$$\alpha_t(i) = P(O_1 O_2 ... O_t, q_t = S_i \mid \lambda).$$

(probability of the partial observation sequence O & state S_i at time t).

- Solving for $\alpha_t(i)$ inductively:

1) Initialization:

$$\alpha_1(i) = \pi_i b_i(O_1), \qquad 1 \leq i \leq N.$$

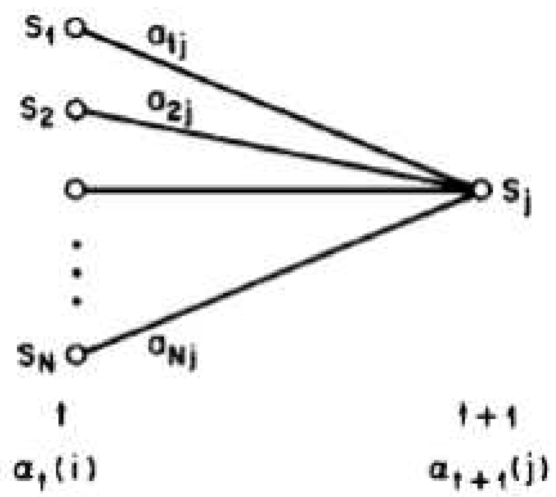
2) Induction:

$$\alpha_{t+1}(j) = \left[\sum_{i=1}^{N} \alpha_t(i) a_{ij} \right] b_j(O_{t+1}), \qquad 1 \leq t \leq T-1$$

$$1 \leq j \leq N.$$

3) Termination:

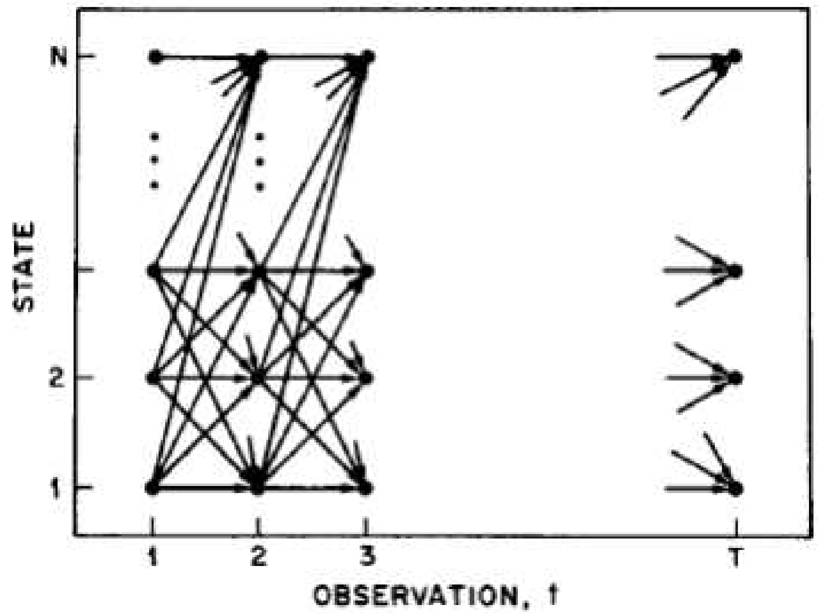
$$P(O|\lambda) = \sum_{i=1}^{N} \alpha_T(i).$$



Problems & Solutions for HMM

Solution to Problem 1: F-B algorithm (3/3)

- Requires complexity $O(N^2T)$ –> reduce computational load significantly.
- The Forward algorithm is based on trellis structure.
- With N states (N nodes at each time slot), all possible state sequences are formed without regarding to how long the observation sequence.



Problems & Solutions for HMM

Solution to Problem 2: Viterbi algorithm (1/3)

- There are several possible optimality criteria: **difficulty to select**.
- One possible criterion: choose the states q_t which are individually most likely.
- Probability of being in state S_i at time t given O, λ:

$$\gamma(i) = P(q_t = S_i \mid O, \lambda).$$

- Find the individually most likely state q_t at time t:

$$q_t = \underset{1\leq\ i\leq N}{argmax}[\gamma_t(i)] \quad 1 \leq\ t \leq T$$

- The solution determines the most likely state at every instant **without** regarding to **the probability of occurrence of sequence of states**.

Problems & Solutions for HMM

Solution to Problem 2: Viterbi algorithm (2/3)

- Optimality criterion: find the single best state sequence Q given O.
- Need to determine:

$$\delta_t(i) = \max_{q_1, q_2, \dots, q_{t-1}} P[q_1, q_2, ..., q_t = S_i, O_1, O_2, ..., O_t \mid \lambda]$$

(The best score along a single path, at time t, which accounts for the first t observations & ends in state S_i)

- By induction, we get for time $t+1$:

$$\delta_{t+1}(j) = [\max_i \delta_t(i) a_{ij}] b_j(O_{t+1})$$

- The state sequence is gotten by tracking the argument $\psi_t(j)$.
- Difference is the Maximization instead of Summing procedure (Forward)

Problems & Solutions for HMM

Solution to Problem 2: Viterbi algorithm (3/3)

1) Initialization:

$$\delta_1(i) = \pi_i b_i(O_1), \qquad 1 \le i \le N$$

$$\psi_1(i) = 0.$$

2) Recursion:

$$\delta_t(j) = \max_{1 \le i \le N} [\delta_{t-1}(i) a_{ij}] b_j(O_t), \qquad 2 \le t \le T, \quad 1 \le j \le N$$

$$\psi_t(j) = \operatorname*{argmax}_{1 \le i \le N} [\delta_{t-1}(i) a_{ij}], \qquad 2 \le t \le T, \quad 1 \le j \le N.$$

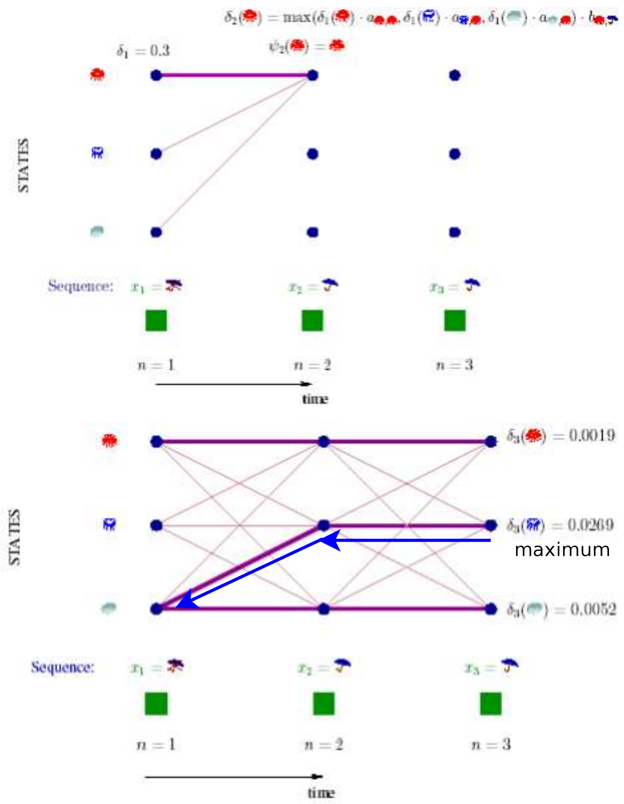
3) Termination:

$$P^* = \max_{1 \le i \le N} [\delta_T(i)]$$

$$q_T^* = \operatorname*{argmax}_{1 \le i \le N} [\delta_T(i)].$$

4) Path (state sequence) backtracking:

$$q_t^* = \psi_{t+1}(q_{t+1}^*), \qquad t = T-1, T-2, \cdots, 1.$$



- Idea: find the most likely path for each intermediate state.
- At each time t, only the most likely path leading to each state S_j survives.

Problems & Solutions for HMM

Solution to Problem 3: Baum-Welch (1/3)

- Locally optimize λ to best describe O –> iterative procedure Baum-Welch.
- Consider backward variable $\beta_t(i)$:

$$\beta_t(i) = P(O_{t+1}O_{t+2}...O_T \mid q_t = S_i, \lambda).$$

(probability of the partial observation sequence from $t+1$ to the end).

- Solving for $\beta_t(i)$ inductively:

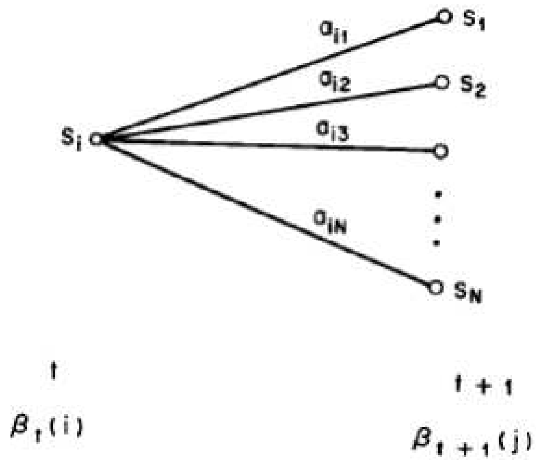
1) Initialization:

$$\beta_T(i) = 1, \qquad 1 \leq i \leq N.$$

2) Induction:

$$\beta_t(i) = \sum_{j=1}^{N} a_{ij} b_j(O_{t+1})\, \beta_{t+1}(j),$$

$$t = T-1, T-2, \cdots, 1, \quad 1 \leq i \leq N.$$



Problems & Solutions for HMM

Solution to Problem 3: Baum-Welch (2/3

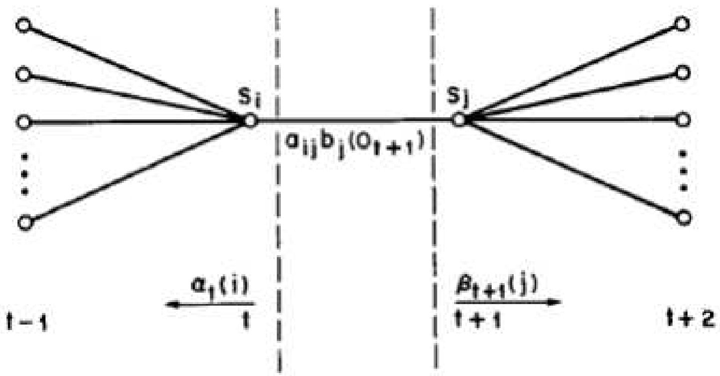
- To describe procedure for reestimation, define variable $\xi_t(i,j)$, the probability of being in state S_i at time t & state S_j at time $t+1$:

$$\xi_t(i,j) = P(q_t = S_i, q_{t+1} = S_j \mid O, \lambda).$$

▷ Rewrite $\xi_t(i,j)$ in form of F-B variables:

$$\xi_t(i,j) = \frac{\alpha_t(i)\, a_{ij} b_j(O_{t+1})\, \beta_{t+1}(j)}{P(O|\lambda)} = \frac{\alpha_t(i)\, a_{ij} b_j(O_{t+1})\, \beta_{t+1}(j)}{\sum_{i=1}^{N}\sum_{j=1}^{N} \alpha_t(i)\, a_{ij} b_j(O_{t+1})\, \beta_{t+1}(j)}$$

▷ The sequence of operations to compute joint event $\xi_t(i,j)$:



Problems & Solutions for HMM

Solution to Problem 3: Baum-Welch (3/3)

$$\bar{\pi}_i = \text{expected frequency (number of times) in state } S_i \text{ at time } (t = 1) = \gamma_1(i)$$

$$\bar{a}_{ij} = \frac{\text{expected number of transitions from state } S_i \text{ to state } S_j}{\text{expected number of transitions from state } S_i}$$

$$= \frac{\sum_{t=1}^{T-1} \xi_t(i, j)}{\sum_{t=1}^{T-1} \gamma_t(i)}$$

$$\bar{b}_j(k) = \frac{\text{expected number of times in state } j \text{ and observing symbol } v_k}{\text{expected number of times in state } j}$$

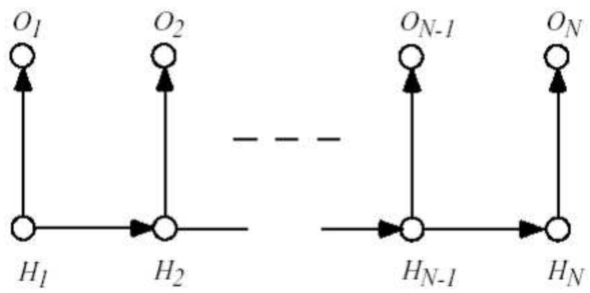
$$= \frac{\sum_{\substack{t=1 \\ \text{s.t. } O_t = v_k}}^{T} \gamma_t(j)}{\sum_{t=1}^{T} \gamma_t(j)}.$$

- Model $\overline{\lambda}$ is more likely than model λ. ($P(O \mid \overline{\lambda}) > P(O \mid \lambda)$).
- Maximizing $Q(\lambda, \overline{\lambda}) = \Sigma_Q P(Q \mid O, \lambda) log[P(O, Q \mid \overline{\lambda}]$ –> increase likelihood.
- Equivalence to EM algorithm: E (estimation) step is calculation of $Q(\lambda, \overline{\lambda})$, M (modification) step is the maximization over $\overline{\lambda}$.

Connections to Graphical Model

HMMs as DAGs

- **Goal**: Inference (F-B alg.) & MAP (Viterbi alg.) for HMMs are special cases of more general Inference algorithms for GMs.
- HMM is a probability model & has a direct representation as a simple GM.



- –> These problems can be solved by standard algorithms of GM :

▶ Inference alg. for DAGs: **JLO's alg.** (developed by Jensen, Lauritzen, Oleson (1990)).

▶ MAP alg. for DAGs: **Dawid's alg.** (developed by Dawid (1992)).

Connections to Graphical Model

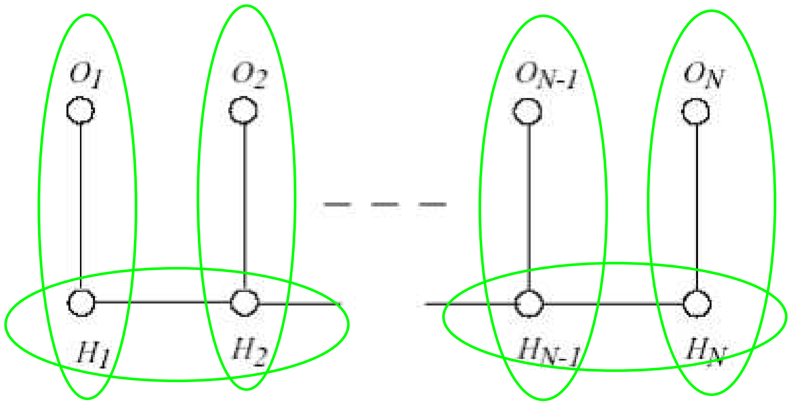
Review Exact Inference

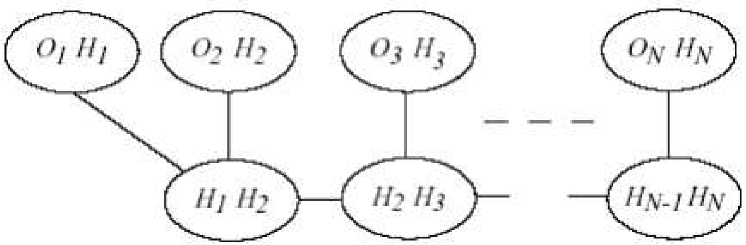
The JLO and Dawid algorithms operate as a two-step process:

1. Construction step: The directed graph is moralized, triangulated, then a junction tree is formed.

2. Propagation step: Junction tree is used in a local message-passing manner to propagate the effects of observed evidence.

–> Resulted junction tree for HMM (final clique (H_{N-1}, H_N) is the root clique):

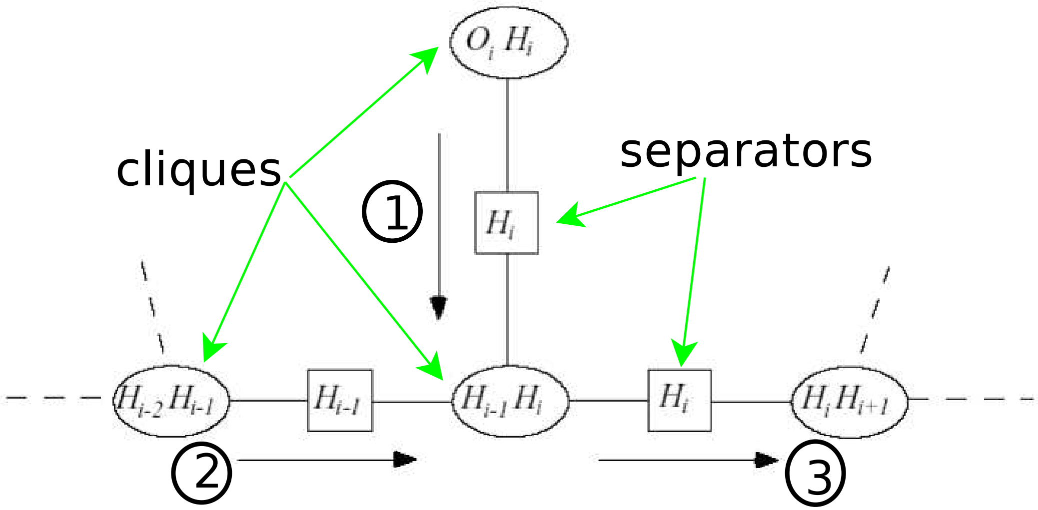




Connections to Graphical Model

Relationship between F-B & JLO (1/3)

- Notation: subscript indicate used variables to derive potential functions.
- Consider the portion of the junction tree, flow from (O_i, H_i) to (H_{i-1}, H_i)
- Collect phase: Local message passing in junction tree



1a. Updated potential on H_i:

$$f^*_{O_i}(h_i) = p(h_i, o^*_i)$$

1b. Update factor from H_i into clique (H_{i-1}, H_i):

$$\lambda_{O_i}(h_i) = \frac{p(h_i, o^*_i)}{p(h_i)} = p(o^*_i \mid h_i)$$

1c. It is absorbed into (H_{i-1}, H_i) :

$$f^*_{O_i}(h_{i-1}, h_i) = p(h_{i-1}, h_i)\lambda_{O_i}(h_i) = p(h_{i-1}, h_i)p(o^*_i \mid h_i)$$

Connections to Graphical Model

Relationship between F-B & JLO (2/3)

2a. Updated potential on H_{i-1}: $f^*_{\Phi_{1,i-1}}(h_{i-1}) = p(h_i, \phi^*_{1,i-1})$

2b. Update factor from H_{i-1} into clique (H_{i-1}, H_i):

$$\lambda_{\Phi_{1,i-1}}(h_{i-1}) = \frac{p(h_i, \phi^*_{1,i-1})}{p(h_{i-1})}$$

2c. It is absorbed into (H_{i-1}, H_i) :

$$f^*_{\Phi_{1,i}}(h_{i-1}, h_i) = f^*_{O_i}(h_{i-1}, h_i)\lambda_{\Phi_{1,i-1}}(h_{i-1}) = p(o^*_i \mid h_i)p(h_i \mid h_{i-1})p(h_i, \phi^*_{1,i-1})$$

3. New potential on H_i for the flow from clique (H_{i-1}, H_i) to (H_i, H_{i+1}):

$$f^*_{\Phi_{1,i}}(h_i) = \sum_{h_{i-1}} f^*_{\Phi_{1,i}}(h_{i-1}, h_i) = p(o^*_i \mid h_i) \sum_{h_{i-1}} p(h_i \mid h_{i-1}) f^*_{\Phi_{1,i-1}}(h_{i-1})$$

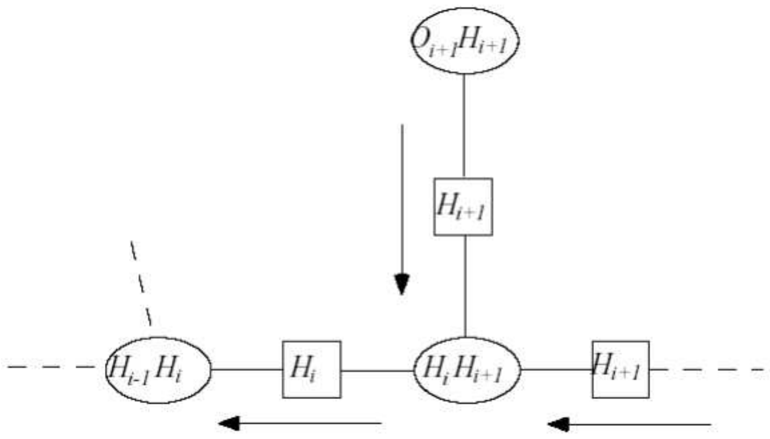
Comparing with: $\alpha_{t+1}(j) = b_j(O_{t+1})\Sigma_{i=1}^N \alpha_t(i) a_{ij}$ (Forward alg.)

▶ Proceeding recursively to obtain result at the root clique.

Connections to Graphical Model

Relationship between F-B & JLO (3/3)

- Distribution phase: Local message passing in junction tree



- By the similar method, achieve equivalence between Backward & JLO.
- Get the update factor on separator H_i:

$$\lambda^*_{\Phi_{i+1,N}}(h_i) = \sum_{h_{i-1}} p(h_i \mid h_{i+1}) p(o^*_{i+1} \mid h_{i+1}) \lambda^*_{\Phi_{i+2,N}}(h_{i+1})$$

- Comparing with Backward alg. :

$$\beta_t(j) = \Sigma_{j=1}^N a_{ij} b_j(O_{t+1}) \beta_{t+1}(j)$$

Connections to Graphical Model

Relationship between Viterbi & Dawid

- Similarly, applying the collection phase, followed by distribution phase.
- **Change**: Marginalization operations are replaced by Maximization.
- –> Obtain the new potential on separator from (H_{i-1}, H_i) to (H_i, H_{i+1}):

$$\widehat{f}_{\Phi_{1,i}}(h_i) = \max_{h_{i-1}} \widehat{f}_{\Phi_{1,i}}(h_{i-1}, h_i) =$$

$$p(o_i^* \mid h_i) \max_{h_{1,i-1}} \left[p(h_i \mid h_{i-1}) p(h_{i-1}, h_{1,i-2}, \phi_{1,i-1}^*) \right]$$

- Comparing with δ in Viterbi alg. :

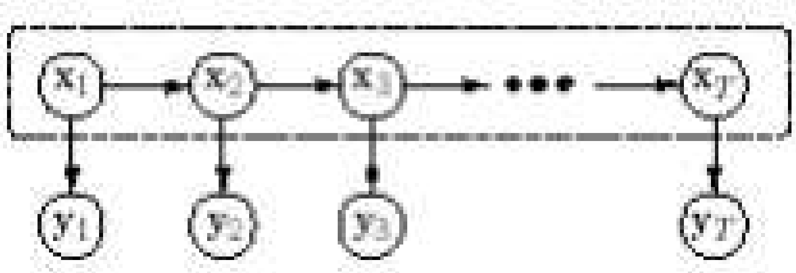
$$\delta_t(j) = \max_{1\leq\ j\leq N} b_j(O_t)[\delta_{t-1}(i)a_{ij}]$$

- Proceeding recursively untill root clique, one can get the likelihood of obervation given the most likely state sequence.

Kalman Filter (LGMs)

Linear Dynamic System (LDS)

- State Space Model (SSM): hidden state variables are continuous.



- LDS is the special case of SSM with the linear functions & the noise term are Gaussian.

$$x_t = Ax_{t-1} + \omega_t$$
$$y_t = Ax_t + \omega_t$$
$$\omega_t \sim N(0, Q)$$
$$\upsilon_t \sim N(0, R)$$

Kalman Filter (LGMs)

Kalman Filter Models (KFMs)

- KFM is also known as LDS, SSMs.
- The transition & observation functions are linear-Gaussian:

$$P(X_t = x_t \mid X_{t-1} = x_{t-1}, U_t = u) \sim N(x_t; Ax_{t-1} + Bu + \mu x, Q)$$

$$P(Y_t = y \mid X_t = x, U_t = u) \sim N(y; Cx + Du + \mu y, R)$$

- Represent as linear functions:

$$X_t = Ax_{t-1} + Bu + V_t$$

where $V_t \sim N(\mu_x, Q)$ is a Gaussian noise term.

$$Y_t = CX_t + DU_t + W_t$$

where $W_t \sim N(\mu_y, R)$ is another Gaussian noise term assumed independent of V_t

Kalman Filter (LGMs)

Kalman Inference

- Kalman filter to perform exact online inference in LDS.
- Equivalence to the forward alg. for HMMs:

$$P(X_t = i \mid y_{1:t}) = \alpha_t(i) \propto$$
$$P(y_t \mid X_t = i)\Sigma_j P(X_t = i \mid X_{t-1} = j)P(X_{t-1} = j \mid y_{1:t-1}).$$

- The Rauch-Tung-Strievel smoother to perform exact offline inference in LDS.
- Equivalence to the F-B alg. for HMMs:

$$P(X_t = i \mid y_{1:T}) = \gamma_t(i) \propto \alpha_t(i)\beta_t(i).$$

Conclusions

- Structure of Hidden Markov Model.
- Three basic problems of HMM.
- Solutions: Forward-Backward, Viterbi, Baum-Welch algorithms.
- Relationships between HMM & Graphical Models in term of Inference problems: JLO & Dawid algorithms.
- Short introduction about Kalman fi lter.

References

[1] Lawrence R. Rabiner, "A Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition", *IEEE Proc.*, Vol. 77, No. 2, pp. 257-286, 1989.

[2] P. Smyth, D. Heckerman, M.I. Jordan, "Probabilistic Independence Networks for Hidden Markov Probability Models", *Technical Report, Microsoft Research*, June, 1996.

[3] P. Smyth, "Belief networks, hidden Markov models, and Markov random fi elds: a unifying view", *Pattern Recognition Letters*, 1998.

[4] K. P. Murphy, "Dynamic Bayesian Networks: Representation, Learning and Inference", *PhD. thesis*, University of California, Berkeley, 2002.