Machine learning: lecture 18

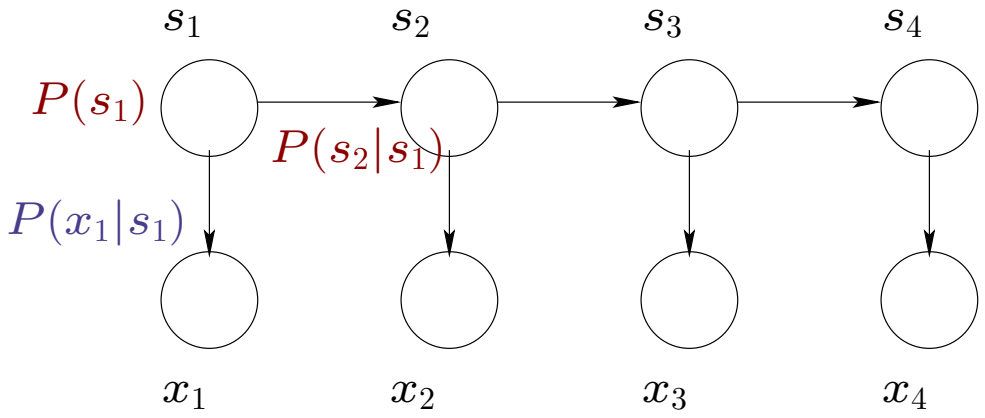
Tommi S. Jaakkola
MIT AI Lab
tommi@csail.mit.edu

Topics

- Hidden Markov models (HMMs)
  - examples, problems
  - forward-backward probabilities
  - EM algorithm

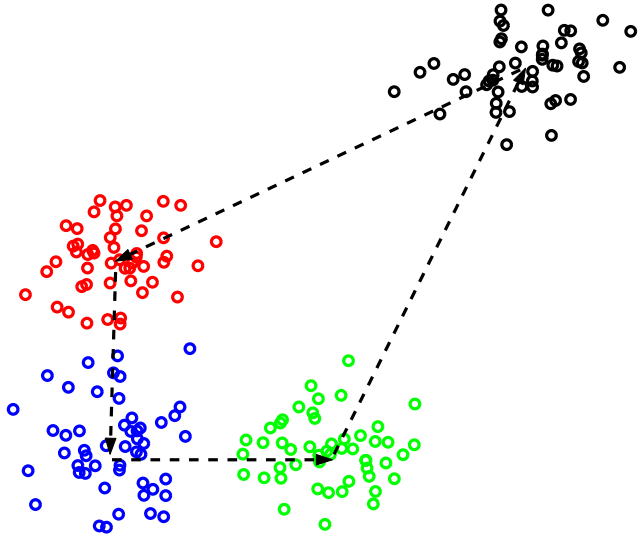
HMM

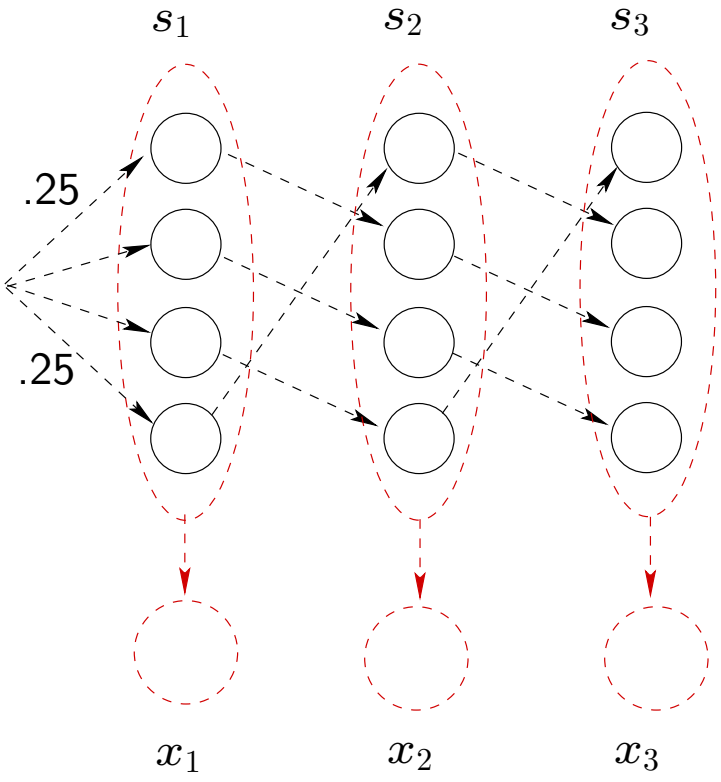
- HMMs are Markov models with observations: the state variables s_t are not observed directly, only via associated observations x_t



$$P(s_1, x_1, s_2, x_2, \ldots) = P(s_1)P(x_1|s_1)P(s_2|s_1)P(x_2|s_2)\cdots$$

HMM example

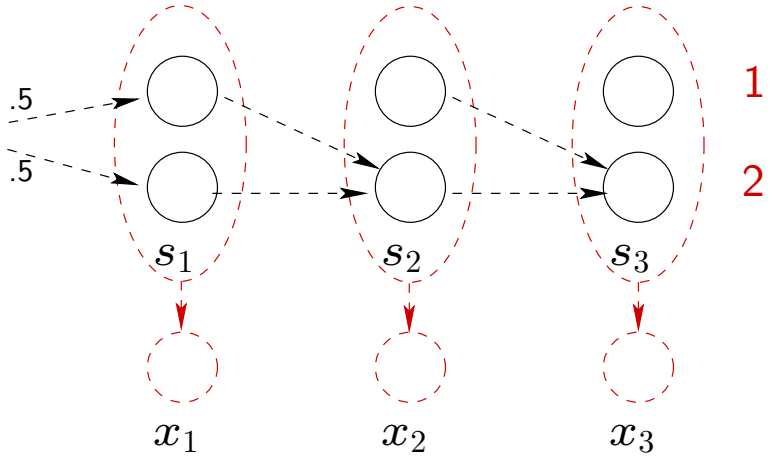




$$P(x_1|s_1 = i) = N(x_1; \mu_i, \Sigma_i)$$

HMM example

- Two states 1 and 2; observations are tosses of unbiased coins

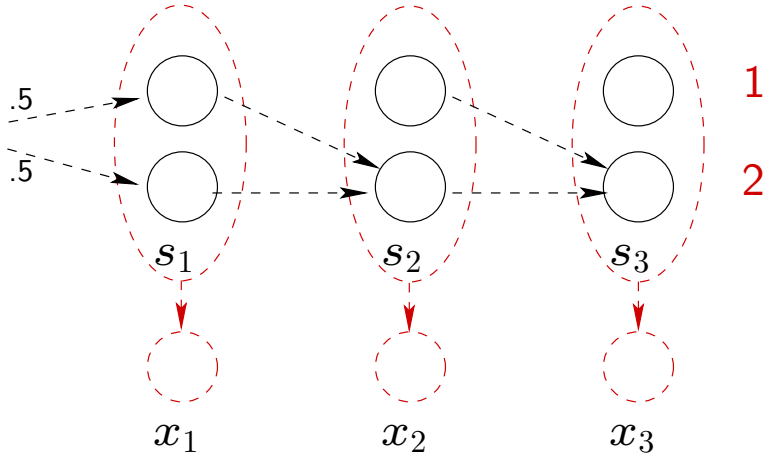


$$P_x(x = \text{heads}|s = 1) = 0.5, \quad P_x(x = \text{tails}|s = 1) = 0.5$$
$$P_x(x = \text{heads}|s = 2) = 0.5, \quad P_x(x = \text{tails}|s = 2) = 0.5$$

- This model is *unidentifiable*

HMM example: biased outputs

- Two states 1 and 2; outputs are tosses of *biased* coins



$$P_x(x = \text{heads}|s = 1) = 0.25, \quad P_x(x = \text{tails}|s = 1) = 0.75$$

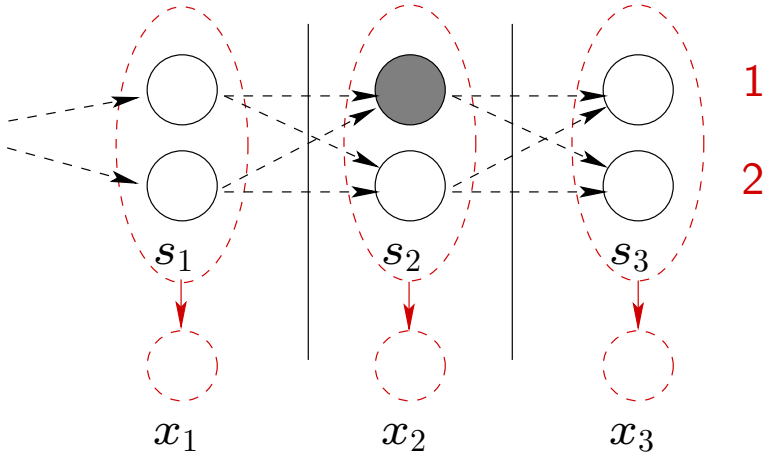
$$P_x(x = \text{heads}|s = 2) = 0.75, \quad P_x(x = \text{tails}|s = 2) = 0.25$$

- What type of output sequences do we get from this HMM model?

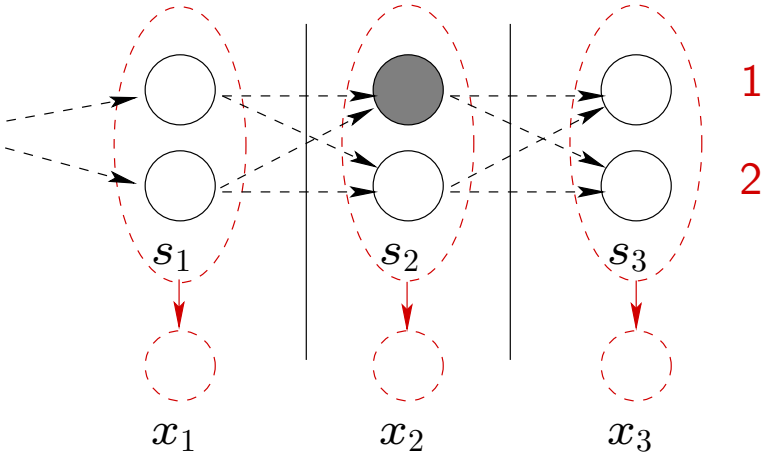
HMM problems

- There are several problems we have to solve
  1. How do we evaluate the probability of an observation sequence $\{\mathbf{x}_1, \ldots, \mathbf{x}_n\}$?
     - forward-backward algorithm
  2. How do we adapt the parameters of the HMM to better account for the observations?
     - the EM-algorithm
- How do we uncover the most likely hidden state sequence corresponding to the observations?
  - dynamic programming (Viterbi algorithm)

Forward-backward probabilities



Forward-backward probabilities



- Forward (predictive) probabilities $\alpha_t(i)$:

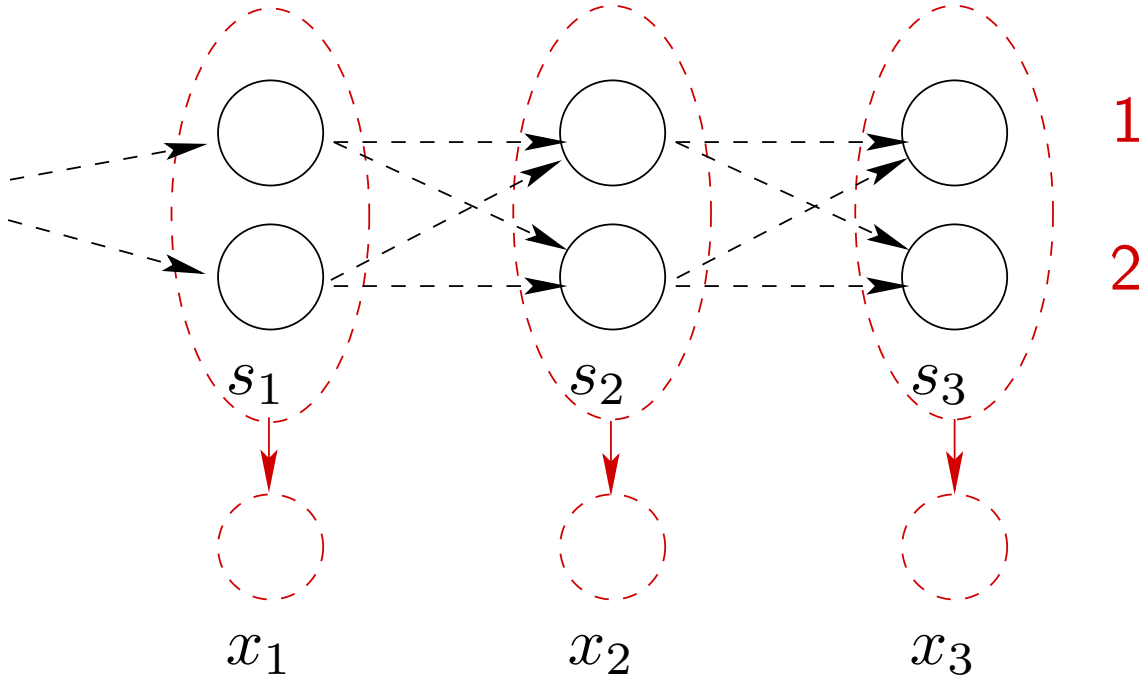
$$\begin{aligned} \alpha_t(i) &= P(\mathbf{x}_1, \ldots, \mathbf{x}_t, s_t = i) \\ \frac{\alpha_t(i)}{\sum_j \alpha_t(j)} &= P(s_t = i | \mathbf{x}_1, \ldots, \mathbf{x}_t) \end{aligned}$$

- Backward (diagnostic) propabilities $\beta_t(i)$:

$$\beta_t(i) = P(\mathbf{x}_{t+1}, \ldots, \mathbf{x}_n | s_t = i)$$

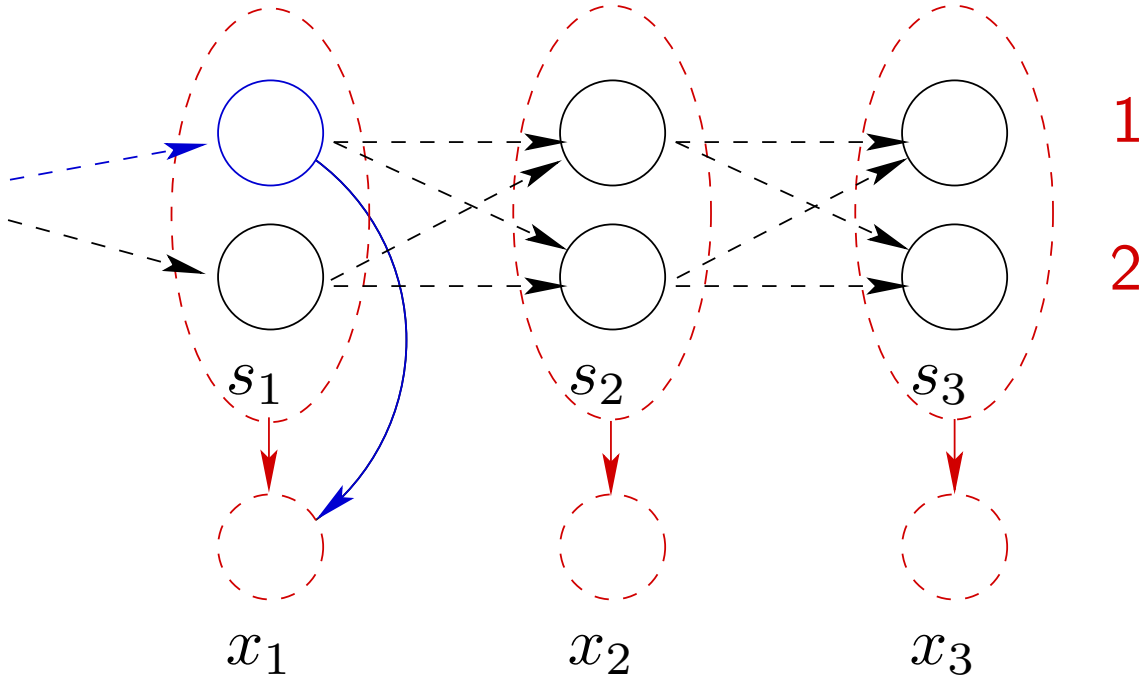
(evidence about the current state from future observations)

Forward probabilities



$$\alpha_1(1) = P(x_1, s_1 = 1)$$
$$\alpha_1(2) = P(x_1, s_1 = 2)$$

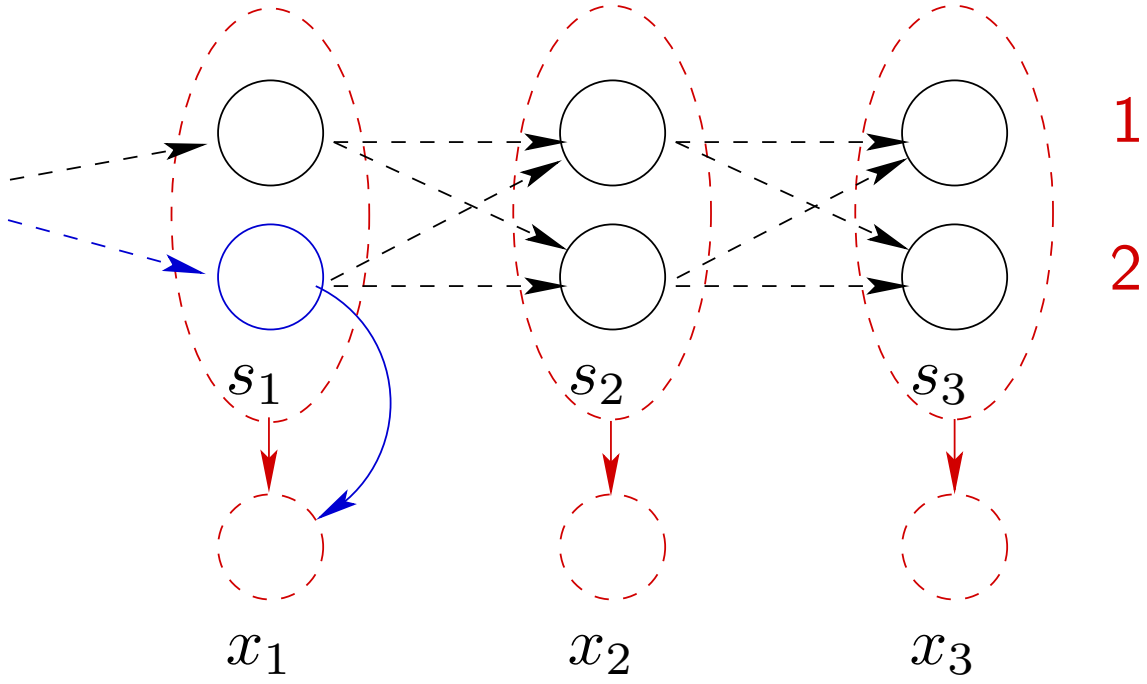
Forward probabilities



$$\alpha_1(1) = P(x_1, s_1 = 1)$$
$$\alpha_1(2) = P(x_1, s_1 = 2)$$

$$\alpha_1(1) = P(1)P_x(\mathbf{x}_1|1)$$

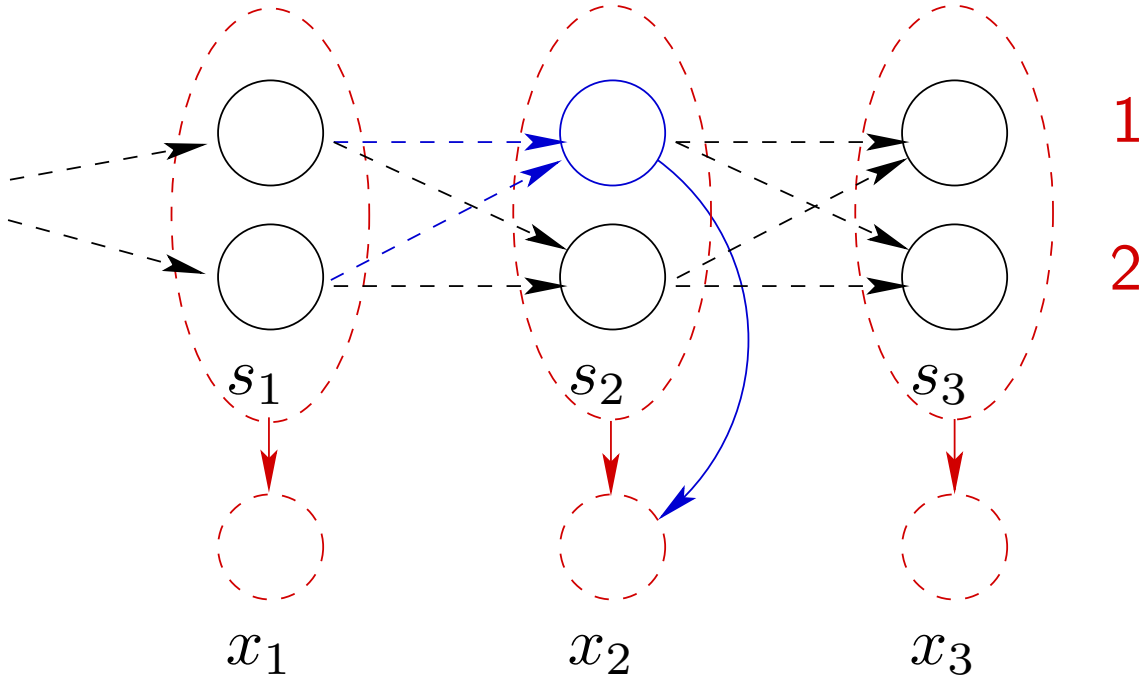
Forward probabilities



$$\alpha_1(1) = P(x_1, s_1 = 1)$$
$$\alpha_1(2) = P(x_1, s_1 = 2)$$

$$\alpha_1(1) = P(1)P_x(\mathbf{x}_1|1)$$
$$\alpha_1(2) = P(2)P_x(\mathbf{x}_1|2)$$

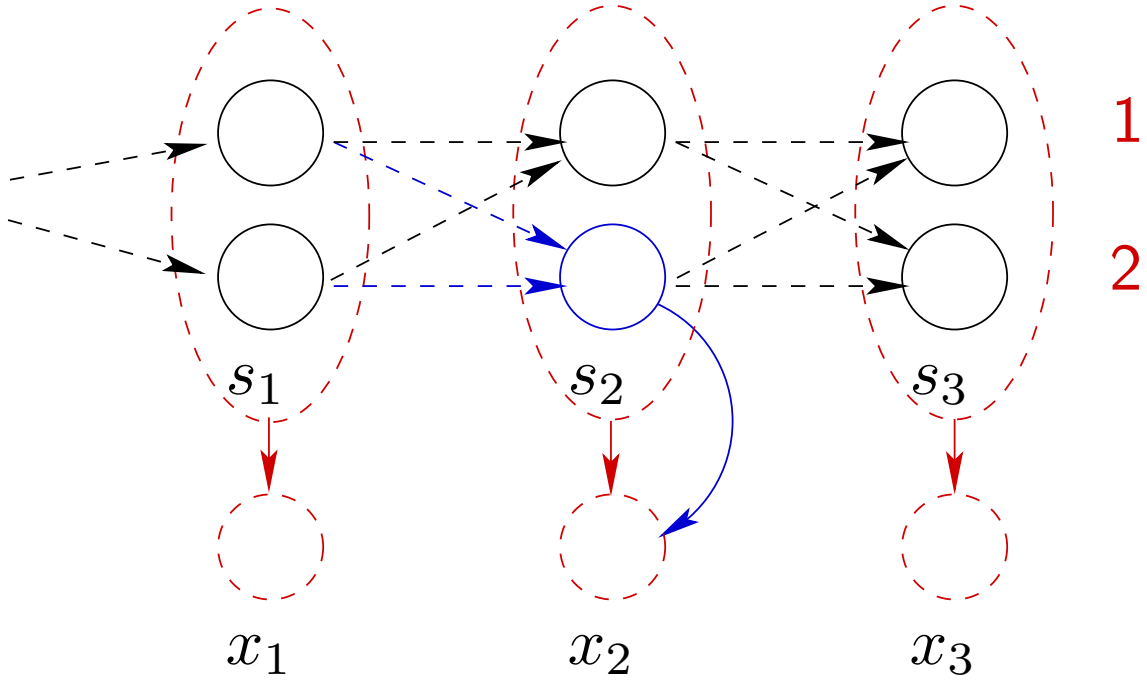
Forward probabilities



$$\alpha_2(1) = P(x_1, x_2, s_2 = 1)$$
$$\alpha_2(2) = P(x_1, x_2, s_2 = 2)$$

$$\alpha_2(1) = \left[\alpha_1(1)P_1(1|1) + \alpha_1(2)P_1(1|2)\right] P_x(\mathbf{x}_2|1)$$

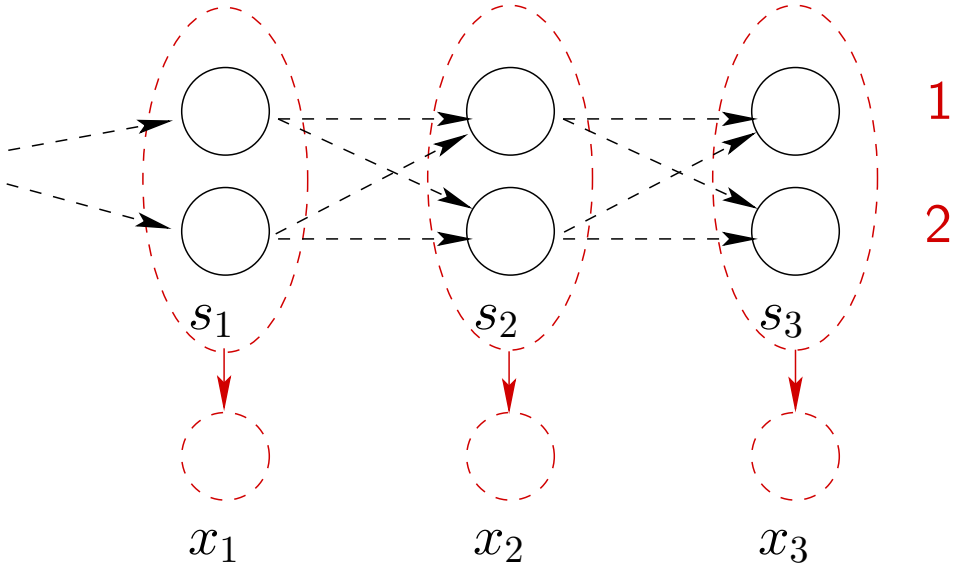
Forward probabilities



$$\alpha_2(1) = P(x_1, x_2, s_2 = 1)$$
$$\alpha_2(2) = P(x_1, x_2, s_2 = 2)$$

$$\alpha_2(1) = \left[\alpha_1(1)P_1(1|1) + \alpha_1(2)P_1(1|2)\right] P_x(\mathbf{x}_2|1)$$
$$\alpha_2(2) = \left[\alpha_1(1)P_1(2|1) + \alpha_1(2)P_1(2|2)\right] P_x(\mathbf{x}_2|2)$$

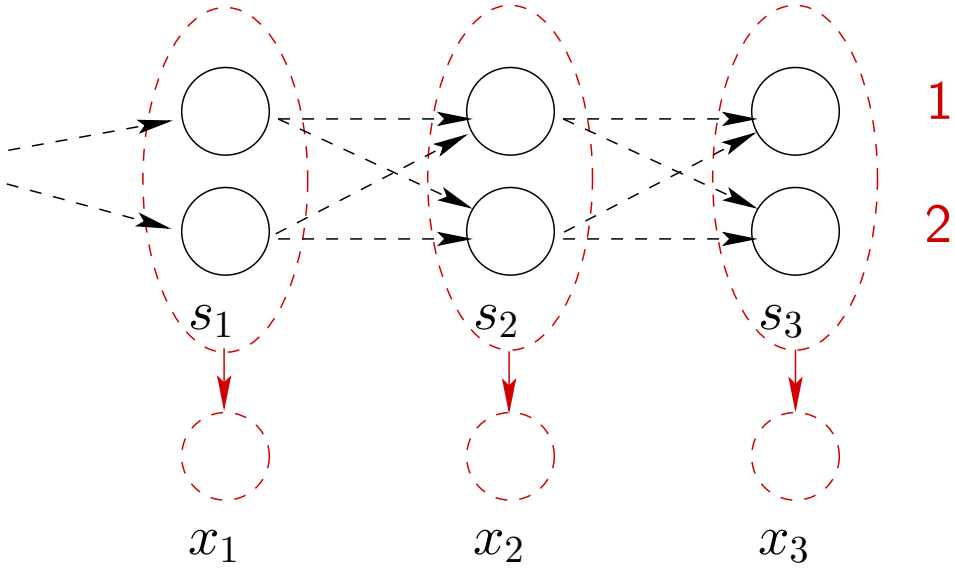
Forward probabilities



- We get the following recursive equation for calculating the forward probabilities $\alpha_t(i) = P(\mathbf{x}_1, \ldots, \mathbf{x}_t, s_t = i)$:

$$\begin{aligned} \alpha_1(i) &= P(i)P_x(\mathbf{x}_1|i) \\ \alpha_t(i) &= \left[\sum_j \alpha_{t-1}(j)P_1(i|j)\right] P_x(\mathbf{x}_t|i) \end{aligned}$$

Backward probabilities



- We can proceed analogously to derive a recursive equation for the backward probabilities $\beta_t(i) = P(\mathbf{x}_{t+1}, \ldots, \mathbf{x}_n | s_t = i)$:

$$\begin{aligned} \beta_n(i) &= 1 \\ \beta_t(i) &= \left[\sum_j P_1(j|i) P_x(\mathbf{x}_{t+1}|j) \beta_{t+1}(j) \right] \end{aligned}$$

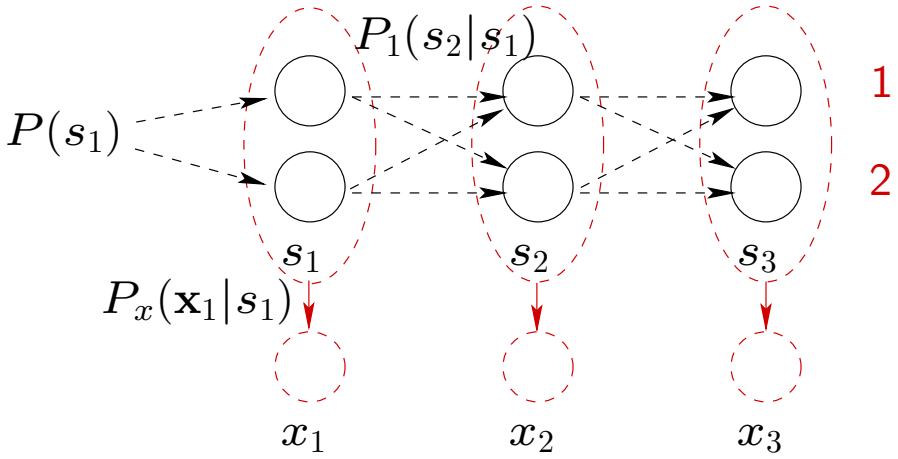
Uses of forward/backward probabilities

- The complementary forward/backward probabilities

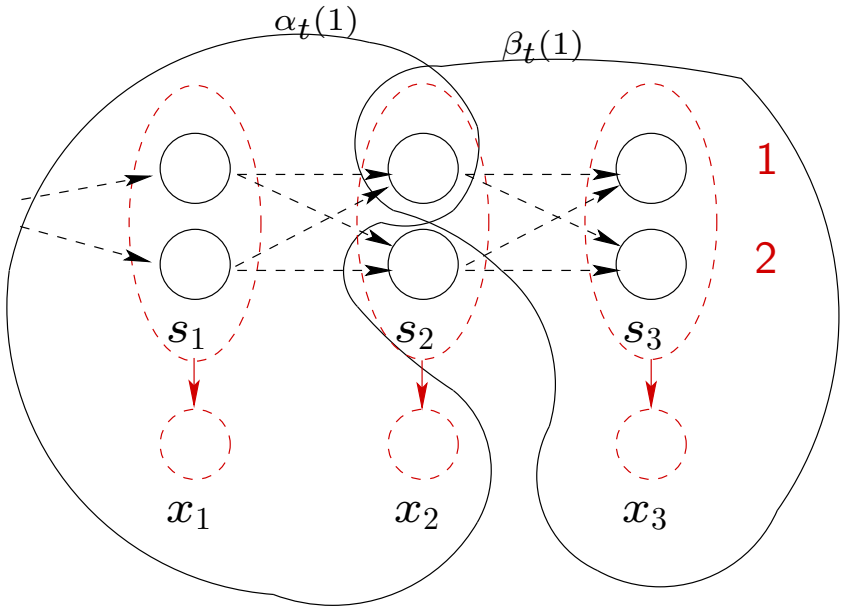
$$\alpha_t(i) = P(\mathbf{x}_1, \ldots, \mathbf{x}_t, s_t = i)$$
$$\beta_t(i) = P(\mathbf{x}_{t+1}, \ldots, \mathbf{x}_n | s_t = i)$$

permit us to evaluate various probabilities:

1. $P(\mathbf{x}_1, \ldots, \mathbf{x}_n)$
2. $\gamma_t(i) = P(s_t = i | \mathbf{x}_1, \ldots, \mathbf{x}_n)$
3. $\xi_t(i, j) = P(s_t = i, s_{t+1} = j | \mathbf{x}_1, \ldots, \mathbf{x}_n)$



Uses of forward/backward probabilities



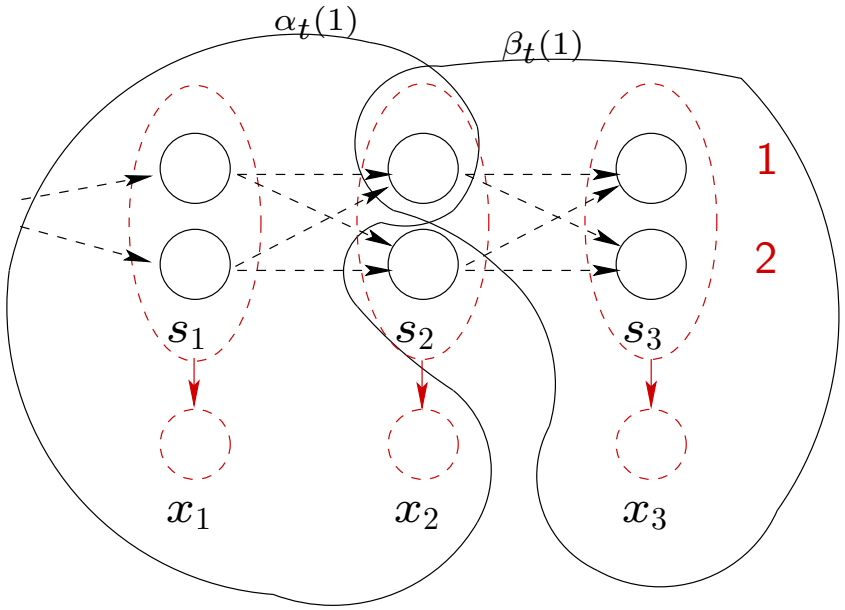
Probability of the observation sequence:

$$P(\mathbf{x}_1, \ldots, \mathbf{x}_n) = \sum_i P(\mathbf{x}_1, \ldots, \mathbf{x}_n, s_t = i)$$

$$= \sum_i P(\mathbf{x}_1, \ldots, \mathbf{x}_t, s_t = i) P(\mathbf{x}_{t+1}, \ldots, \mathbf{x}_n | s_t = i)$$

$$= \sum_i \alpha_t(i) \beta_t(i)$$

Forward/backward probabilities cont'd



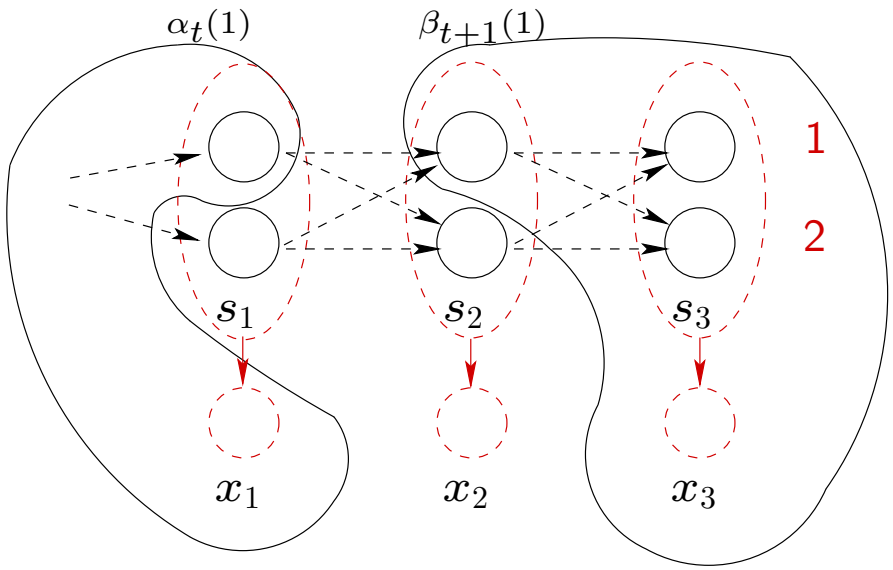
- We can evaluate the posterior probability that the HMM was in a particular state i at time t

$$P(s_t = i|\mathbf{x}_1, \ldots, \mathbf{x}_n) = \frac{P(\mathbf{x}_1, \ldots, \mathbf{x}_n, s_t = i)}{P(\mathbf{x}_1, \ldots, \mathbf{x}_n)}$$

$$= \frac{\alpha_t(i)\beta_t(i)}{\sum_j \alpha_t(j)\beta_t(j)} \stackrel{def}{=} \gamma_t(i)$$

Forward/backward probabilities cont'd

- We can also compute the posterior probability that the system was in state i at time t AND transitioned to state j at time $t+1$:



$$P(s_t = i, s_{t+1} = j | \mathbf{x}_1, \ldots, \mathbf{x}_n)$$

$$= \frac{\alpha_t(i) \overbrace{P_1(s_{t+1} = j | s_t = i) P_x(\mathbf{x}_{t+1} | s_{t+1} = j)}^{\text{fixed } i \to j \text{ transition, one observation}} \beta_{t+1}(j)}{\sum_j \alpha_t(j)\beta_t(j)}$$

$$\stackrel{def}{=} \xi_t(i,j),$$

The EM algorithm for HMMs

Assume we have L observation sequences $\mathbf{x}_1^{(l)}, \ldots, \mathbf{x}_{n_l}^{(l)}$

E-step: compute the posterior probabilities

$$\gamma_t^{(l)}(i) \quad \text{for all } l,\ i, \text{ and } t\ (t = 1, \ldots, n_l)$$

$$\xi_t^{(l)}(i,j) \quad \text{for all } l,\ i, \text{ and } t\ (t = 1, \ldots, n_l - 1)$$

M-step: First, the initial state distribution can be updated according to the expected fraction of times the sequences started from a specific state i

$$\hat{P}(i) \leftarrow \frac{1}{L} \sum_{l=1}^{L} \gamma_1^{(l)}(i)$$

M-step cont'd

Second, the transition probabilities can be updated on the basis of the posterior counts:

$$\hat{P}_1(j|i) \leftarrow \frac{\hat{n}(i,j)}{\sum_{j'} \hat{n}(i,j')}$$

where

$$\hat{n}(i,j) = \sum_{l=1}^{L} \sum_{t=0}^{n-1} \xi_t^{(l)}(i,j)$$

defines the expected number of transitions from i to j

M-step cont'd

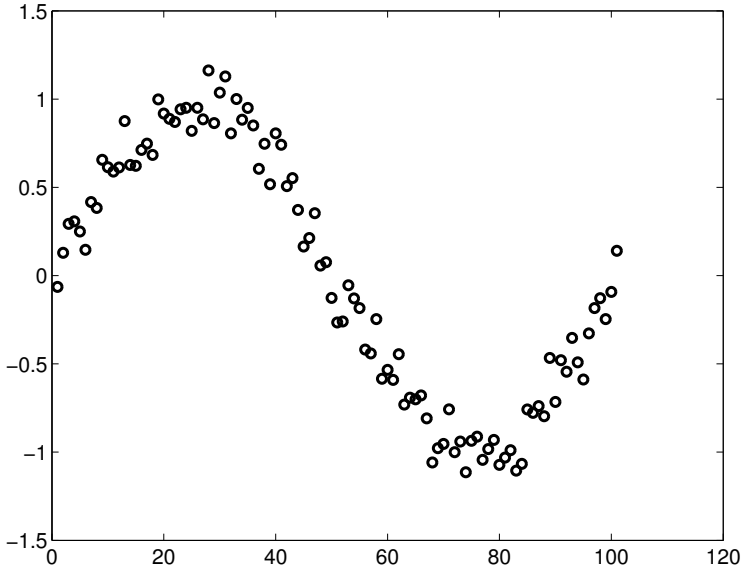
- Lastly, for the outputs we have to (in general) solve a weighted maximum likelihood estimation problem:

  Separately for each state i we maximize:

$$J(\theta_i) = \sum_{l=1}^{L} \sum_{t=1}^{n_l} \gamma_t^{(l)}(i) \log P(\mathbf{x}_t^{(l)} | \theta_i)$$

  with respect to the parameters θ_i (e.g, the mean and the covariance of a Gaussian).

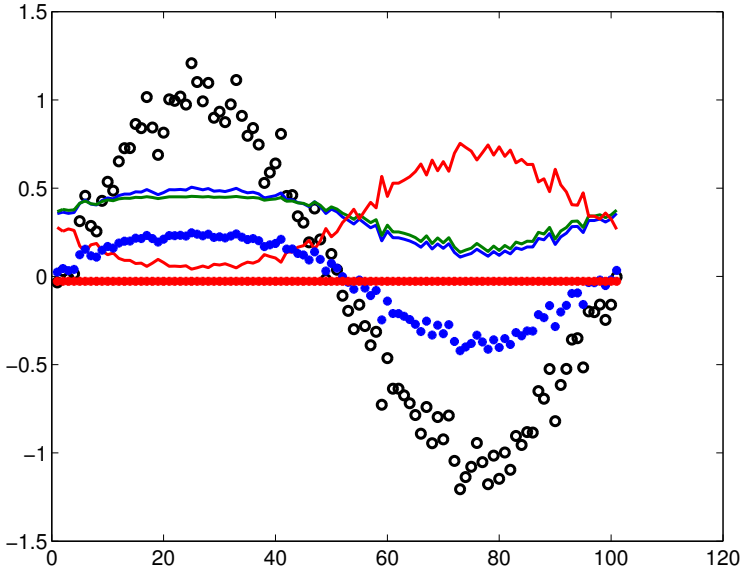
HMM example



Observed output as a function of time

- We will try to model this with a 3-state HMM with Gaussian outputs $p(x|s=i) = p(x|\mu_i, \sigma_i^2)$, $i = 1, 2, 3$.

HMM example cont'd



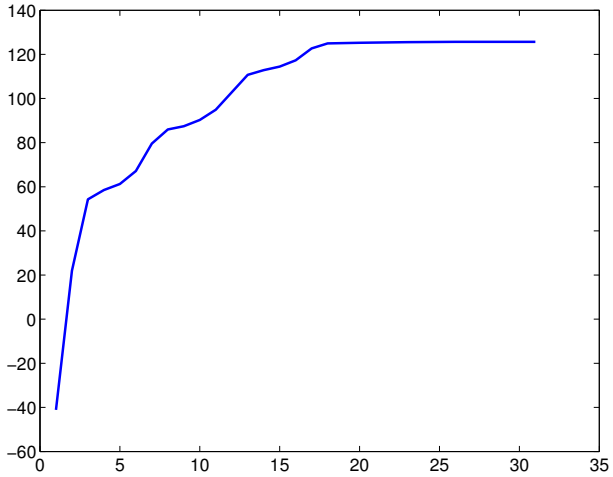
prior/posterior means and $\gamma_t(\cdot)$

$$\text{prior mean}(t) = \sum_i P_t(i)\hat{\mu}_i \quad (\text{'*'})$$

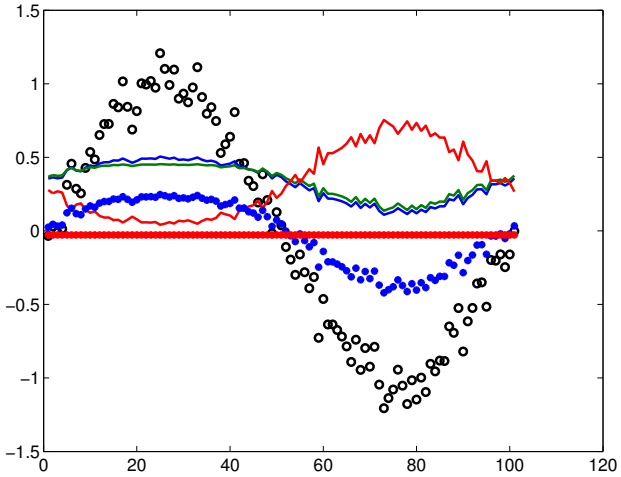
$$\text{posterior mean}(t) = \sum_i \gamma_t(i)\hat{\mu}_i \quad (\text{'*'})$$

where $P_t(i)$ is the probability of being in state i after t steps without observations; $\hat{\mu}_i$ is the mean output from the i^{th} state

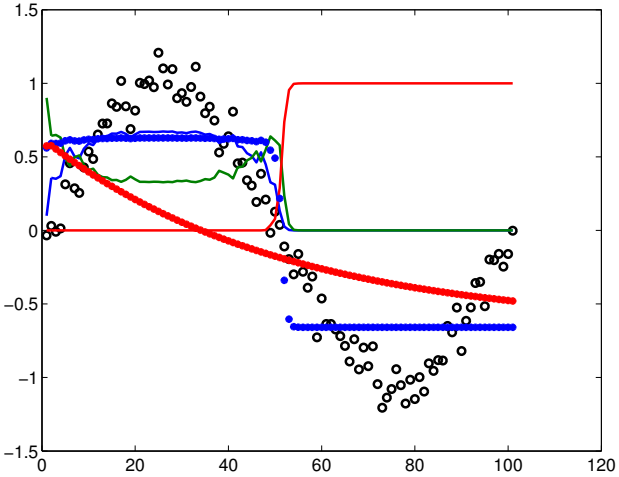
HMM example cont'd



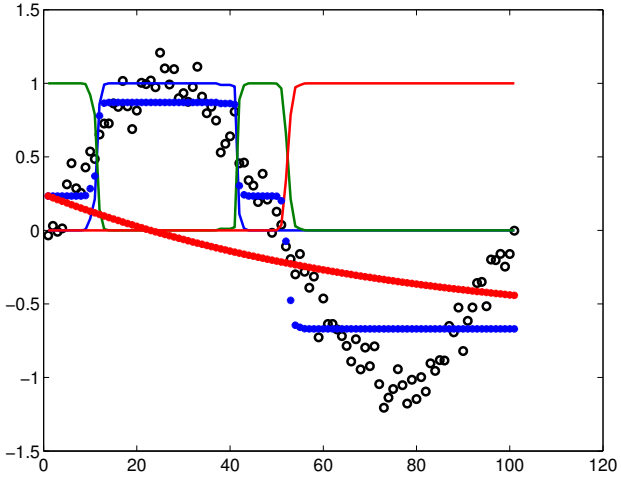
Log-prob. of data



after 0 iterations



after 7 iterations



final