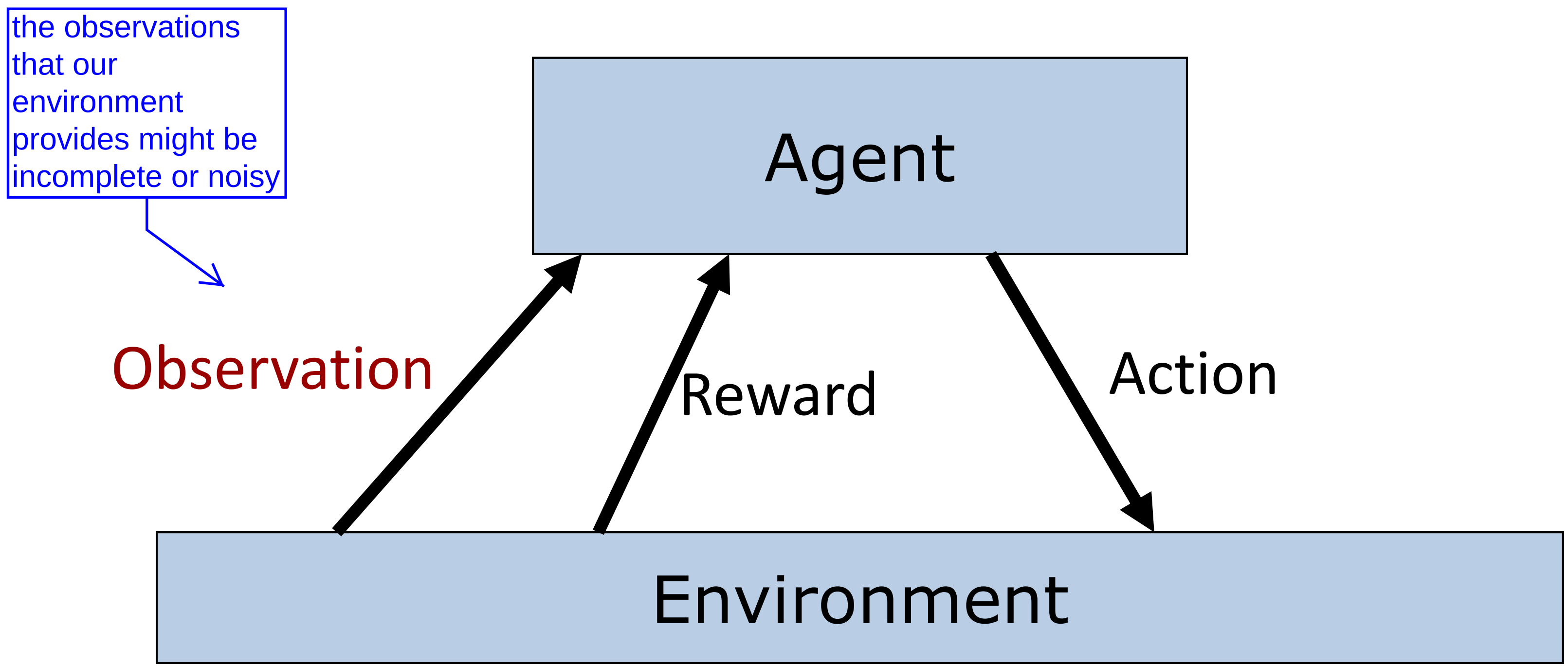
CS885 Reinforcement Learning
Lecture 11b: June 6, 2018

Partially Observable RL

[RusNor] Sec. 17.3 [SigBuf] Chap. 7

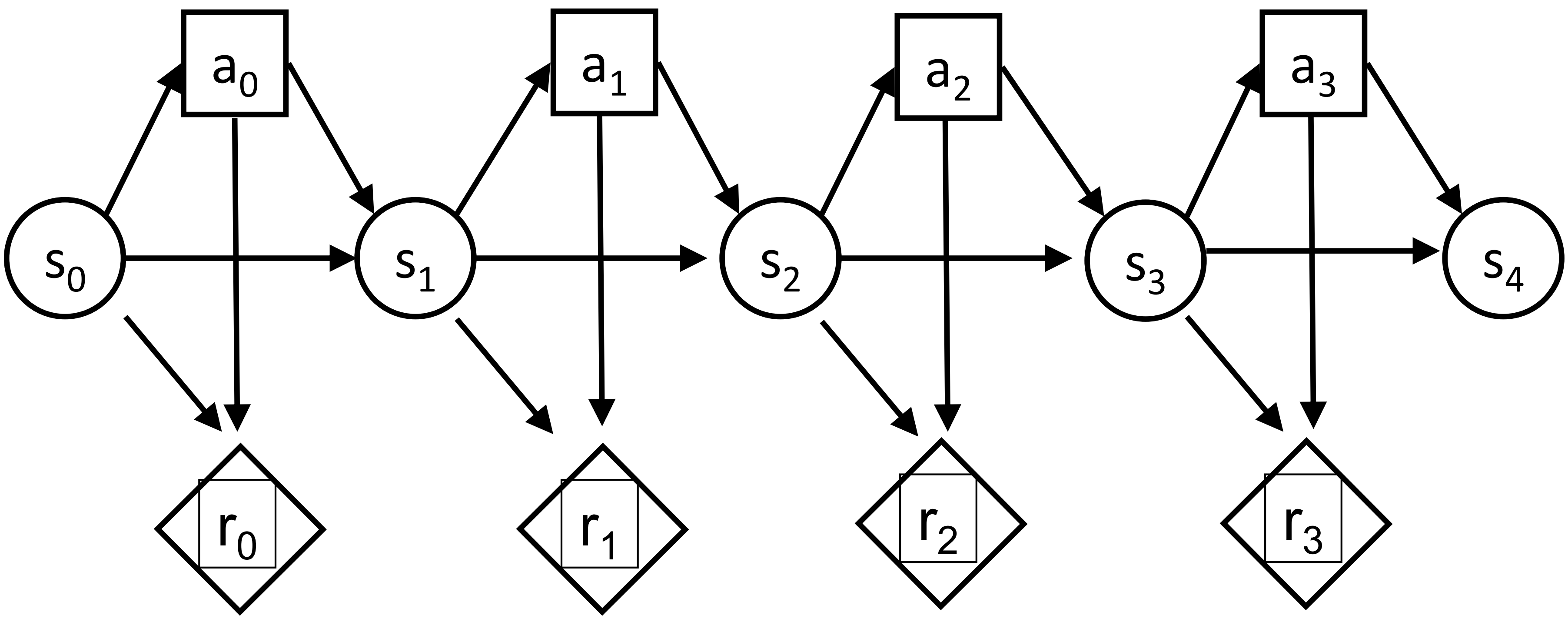
Reinforcement Learning Problem



Goal: Learn to choose actions that maximize rewards

(Fully Observable)
Markov Decision Process (MDP)

with the Markovian assumption, we can say that the current state (and action) is all we need to know to determine the future state. We don't care about any states in the past. Thus, we have all the information we need.



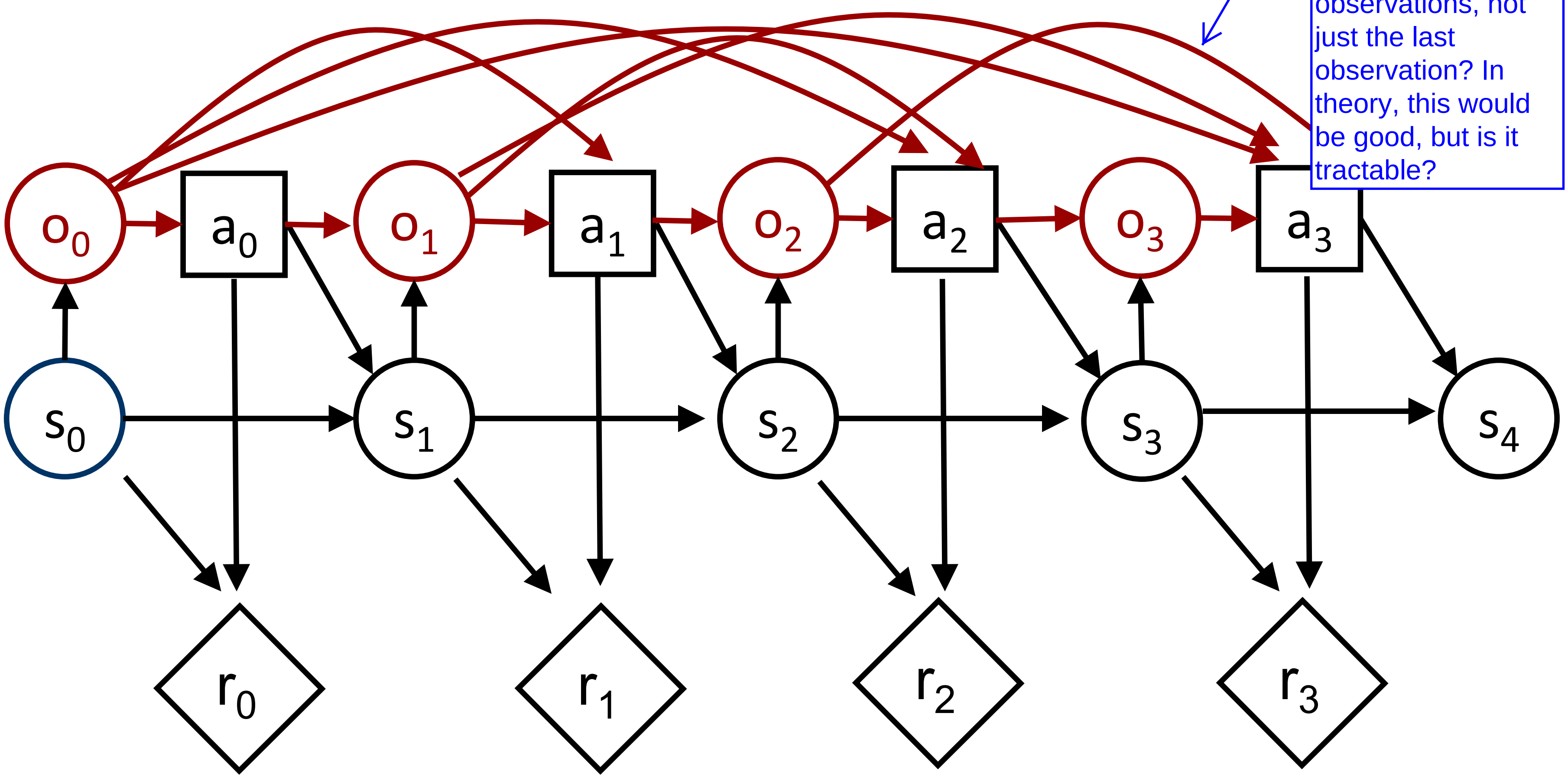
now we can't make decisions based on the states. Instead we have to make them based on observations. But the observations and states are highly correlated

we could just pretend its fully observable and pick our actions based on the last observation. But if that observation is incomplete or noisy, it might not be optimal.

Partially Observable Markov Decision Process (POMDP)

- MDP augmented with observations

what if we conditioned actions on all previous observations, not just the last observation? In theory, this would be good, but is it tractable?



Partially Observable RL

- Definition
  - States: $s \in S$
  - Observations: $o \in O$
  - Actions: $a \in A$
  - Rewards: $r \in \mathbb{R}$
  - Transition model: $\Pr(s_t|s_{t-1}, a_{t-1})$
  - Observation model: $\Pr(o_t|a_{t-1}, s_t)$
  - Reward model: $\Pr(r_t|s_t, a_t)$
  - Discount factor: $0 \leq \gamma \leq 1$
    - discounted: $\gamma < 1$ undiscounted: $\gamma = 1$
  - Horizon (i.e., # of time steps): h
    - Finite horizon: $h \in \mathbb{N}$ infinite horizon: $h = \infty$

we don't have access to this stuff in grey anymore

unknown model

- Goal: find optimal policy π^* such that

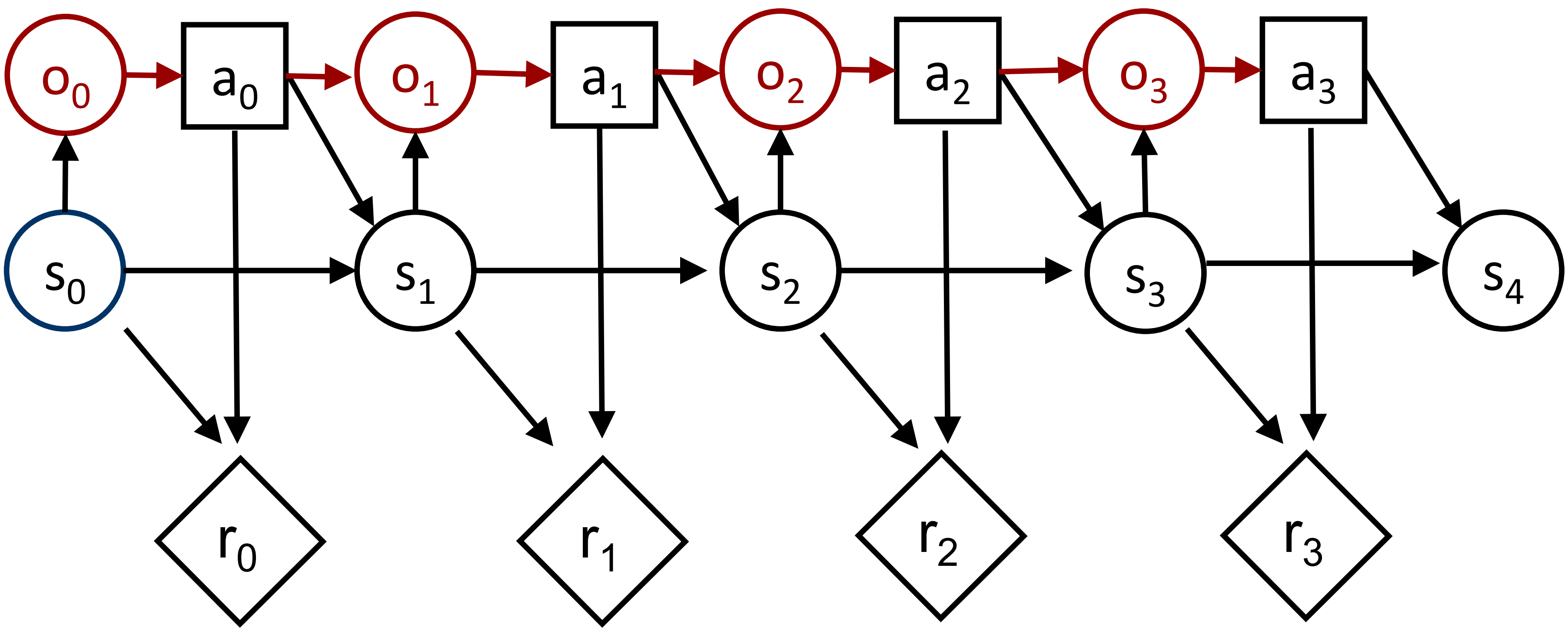
$$\pi^* = argmax_\pi \sum_{t=0}^{h} \gamma^t E_\pi[r_t]$$

let's just choose our action based on the last observation.

we use previous observations too because they might provide us with additional information that might not be capture by our current observation.

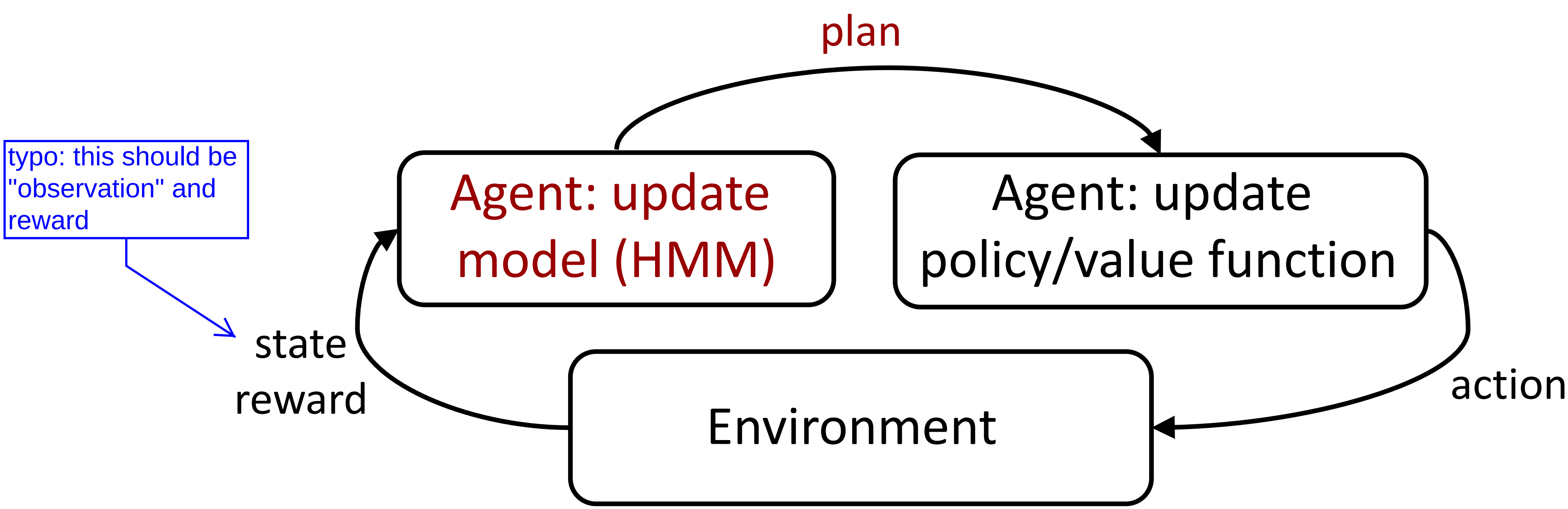
Simple Heuristic

- Approximate s_t by o_t (or finite window of previous observations: $o_{t-k}, o_{t-k+1}, \dots, o_t$



Model-based Partially Observable RL

- Model-based RL
  - Learn HMM from data
  - Plan by optimizing POMDP policy
    - Value iteration, Monte Carlo tree search

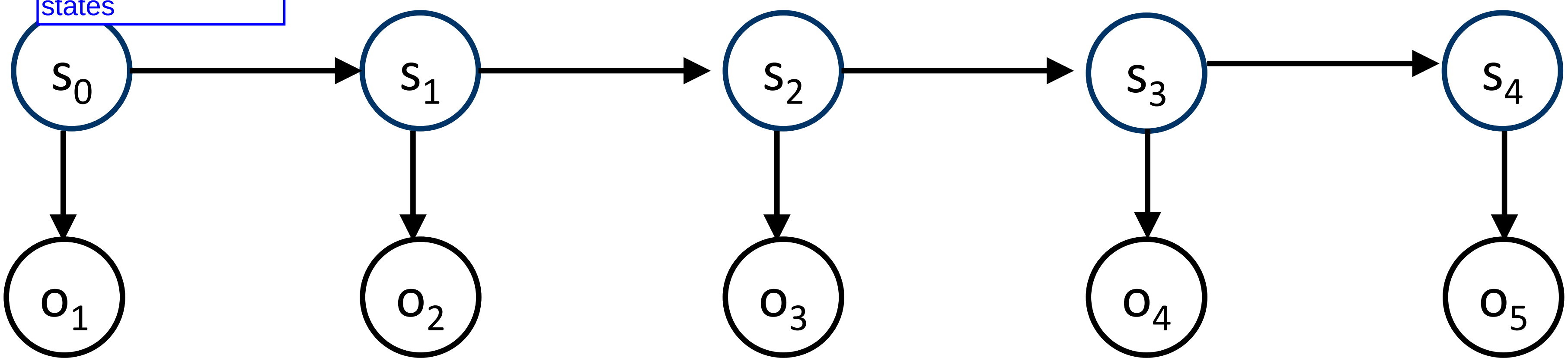


HMM Parameters

- Let $s_t \in \{c_1, c_2\}$ and $o_t \in \{v_1, v_2\}$
- Parameters
  - Initial state distribution: $\psi_c = \Pr(s_0 = c)$
  - Transition probabilities: $\theta_{c'|c} = \Pr(s_{t+1} = c' | s_t = c)$
  - Observation probabilities: $\phi_{v|c} = \Pr(o_t = v | s_t = c)$

    or $N(v|\mu_c, \sigma_c) = \Pr(o_t = v | s_t = c)$

how the observations are correlated with states



Maximum Likelihood

- Supervised Learning: o's are known
- Objective: $argmax_{\psi,\theta,\phi} \Pr(o_{1..t}, s_{1..t} | \psi, \theta, \phi)$
  - Set derivative to 0
  - Isolate parameters ψ, θ, ϕ

- Data (multinomial observations)
  - Let $\#c_i^{start}$ be # of times that process **starts** in class c_i
  - Let $\#c_i$ be # of times that process is in class c_i
  - Let $\#(c_i, c_j)$ be # of times that c_i follows c_j
  - Let $\#(v_i, c_j)$ be # of times that v_i occurs with c_j

Multinomial observations

- Maximum likelihood solution: relative frequency counts

$$\psi_{c_1^{start}} = \#c_1^{start}/(\#c_1^{start} + \#c_2^{start})$$

$$\theta_{c_1|c_1} = \#(c_1, c_1)/(\#(c_1, c_1) + \#(c_2, c_1))$$

$$\theta_{c_1|c_2} = \#(c_1, c_2)/(\#(c_1, c_2) + \#(c_2, c_2))$$

$$\phi_{v_1|c_1} = \#(v_1, c_1)/(\#(v_1, c_1) + \#(v_2, c_1))$$

$$\phi_{v_1|c_2} = \#(v_1, c_2)/(\#(v_1, c_2) + \#(v_2, c_2))$$

Gaussian Observations

- Maximum likelihood solution

$$\psi_{c_1^{start}} = \#c_1^{start}/(\#c_1^{start} + \#c_2^{start})$$

$$\theta_{c_1|c_1} = \#(c_1, c_1)/(\#(c_1, c_1) + \#(c_2, c_1))$$

$$\theta_{c_1|c_2} = \#(c_1, c_2)/(\#(c_1, c_2) + \#(c_2, c_2))$$

$$\mu_{c_1} = \frac{1}{\#c_1}\sum_{\{t|s_t=c_1\}} o_t, \qquad \sigma_{c_1}^2 = \frac{1}{\#c_1}\sum_{\{t|s_t = c_1\}}\left(o_t - \mu_{c_1}\right)^2$$

$$\mu_{c_2} = \frac{1}{\#c_2}\sum_{\{t|s_t=c_2\}} o_t, \qquad \sigma_{c_2}^2 = \frac{1}{\#c_2}\sum_{\{t|s_t = c_2\}}\left(o_t - \mu_{c_2}\right)^2$$

empirical average

empirical variance

Planning

- Idea: summarize previous observations into a distribution about the current unobserved state called **belief**

this gives a way to use all previous observations while also being tractable!

we're keeping the equations simple here, but beliefs should actually be based on actions too...See the last slide for more complete equations

- Belief: $b_t(s_t) = \Pr(s_t|o_{1..t})$
  - Sufficient statistic: $b_t \equiv o_{1..t}$

can use "forward algorithm". After estimating the HMM, we can use this to do inference

- Belief monitoring:

$$\Pr(s_t|o_{1..t}) \propto \Pr(o_t|s_t) \sum_{s_{t-1}} \Pr(s_t|s_{t-1}) \Pr(s_{t-1}|o_{1..t-1})$$

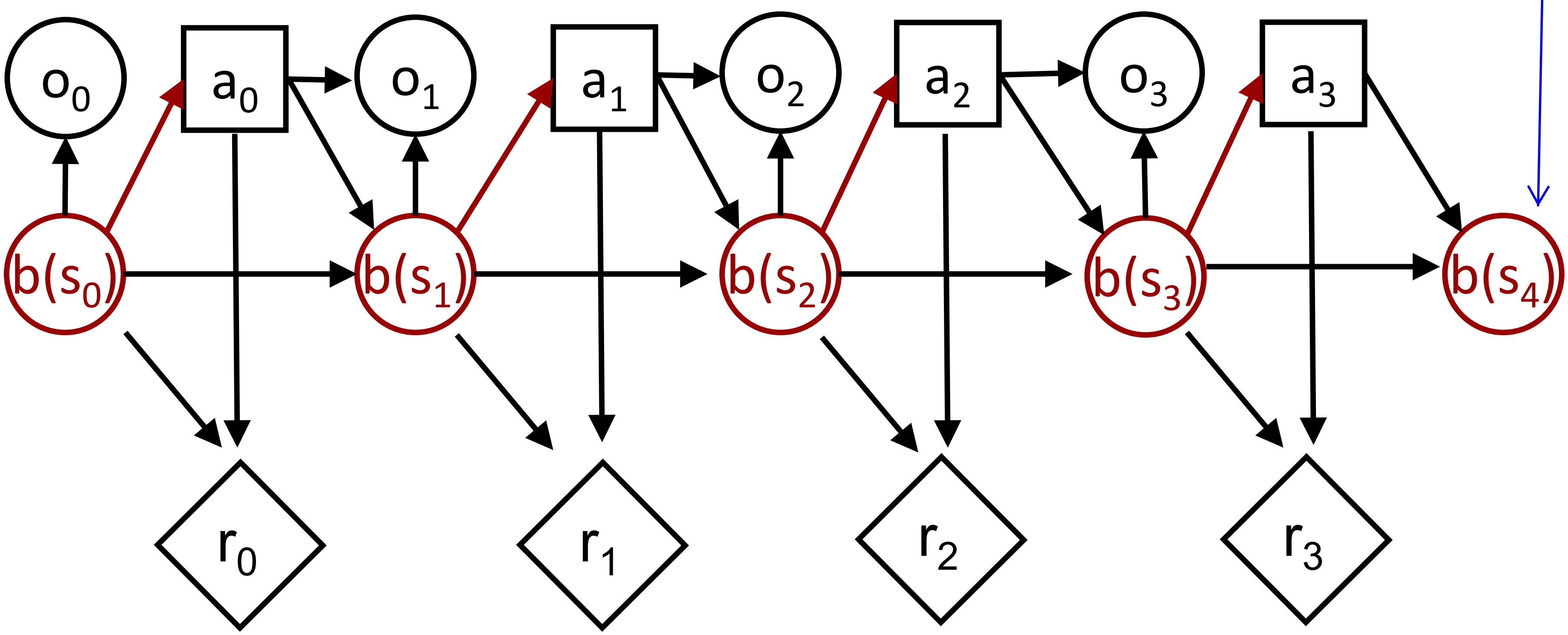
$$b_t(s_t) \propto \Pr(o_t|s_t) \sum_{s_{t-1}} \Pr(s_t|s_{t-1})\, b_{t-1}(s_{t-1})$$

now we've reformulated a POMDP as a Belief MDP. So instead of conditioning on all previous observations, we condition our actions on beliefs. The belief sufficiently captures all previous observations. The belief is really a distribution.

Belief MDP

so we estimate the distribution over the hidden state

- Replace s_t by $b(s_t)$
- Action depends only on previous belief



Value Iteration Algorithm

this is the updated belief

valueIteration(beliefMDP)

$V_0^*(b) \leftarrow \max_a R(b,a) \;\; \forall s$

all we do is replace states with beliefs

For $t = 1$ to h do

$$V_t^*(b) \leftarrow \max_a R(b,a) + \gamma \sum_{o'} \Pr(o'|b,a)\, V_{t-1}^*(b^{ao'}) \;\; \forall s$$

Return V^*

expectation of the rewards w.r.t the beliefs

the belief at the next timestep after executing action a and being in observation o'

Where

$R(b,a) = \sum_s b(s) R(s,a)$

$\Pr(o'|b,a) = \sum_{s'} \Pr(o'|s',a) \sum_s \Pr(s'|s,a)\, b(s)$

$b^{ao'}(s') = \Pr(s'|b,a,o') \propto \Pr(o'|s',a) \sum_s \Pr(s'|s,a)\, b(s)$