Module 14
Introduction to Partially Observable Markov Decision Processes

CS 886 Sequential Decision Making and Reinforcement Learning

University of Waterloo

Markov Decision Processes

- MDPs:
  - Fully Observable MDPs
  - Decision maker knows the state at each time step
- POMDPs:
  - Partially Observable MDPs
  - Decision does not know the state
  - But makes observations that are correlated with the underlying state
    - E.g. sensors provide noisy information about the state

Applications

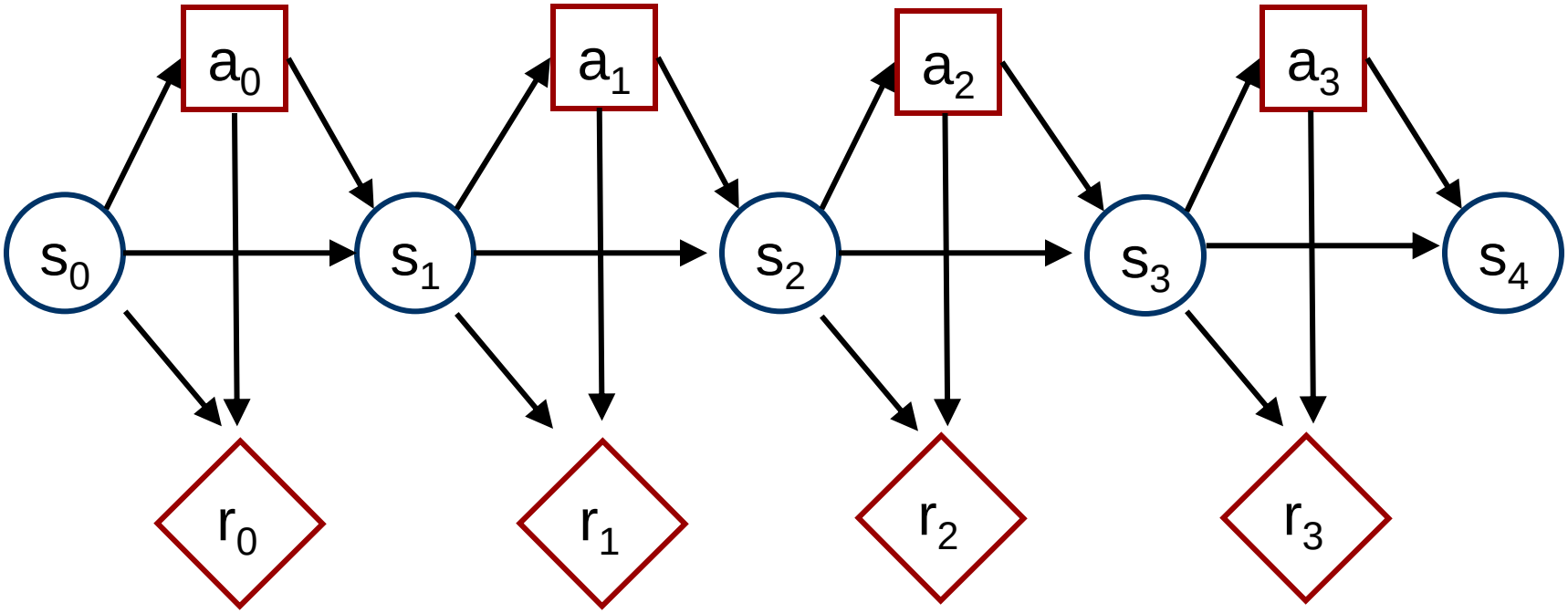
- Robotic control
- Dialog systems
- Assistive Technologies
- Operations Research

Model Description

- Definition
  - Set of states: S
  - Set of actions (i.e., decisions): A
  - Transition model: $\Pr(s_t|s_{t-1}, a_{t-1})$
  - Reward model (i.e., utility): $R(s_t, a_t)$
  - Discount factor: $0 \leq \gamma \leq 1$
  - Horizon (i.e., # of time steps): h
  - **Set of observations: O**
  - **Observation model: $\Pr(o_t|s_t, a_{t-1})$**

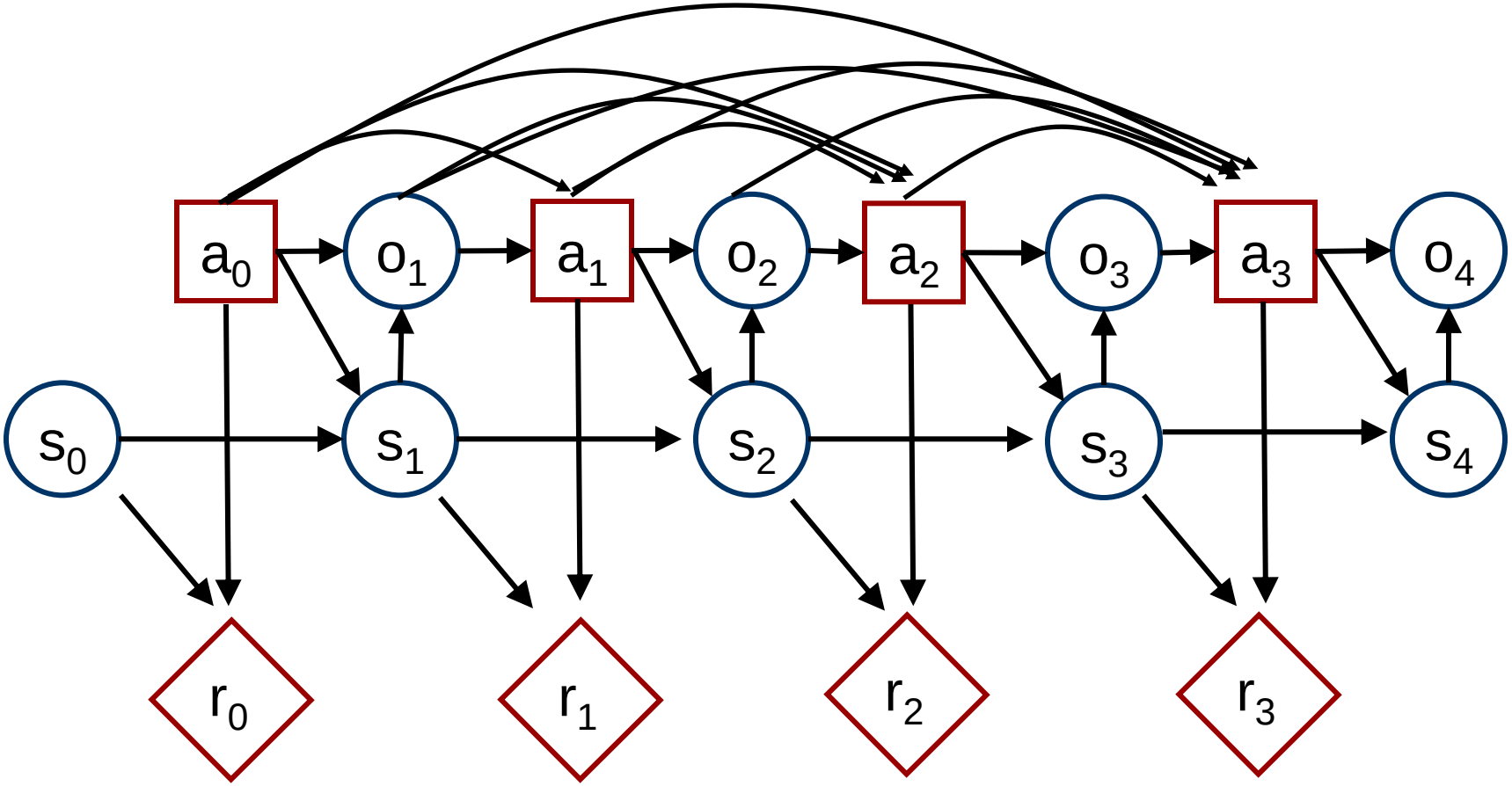
Graphical Model

- Fully Observable MDP



Graphical Model

- Partially Observable MDP



Policies

- MDP policies: $\pi: S \rightarrow A$
  - Markovian policy
- But state is unknown in POMDPs
- POMDP policies: $\pi: B_0 \times H_t \rightarrow A_t$
  - B_0 is the space of initial beliefs b_0

$$b_0 = \Pr(s_0)$$

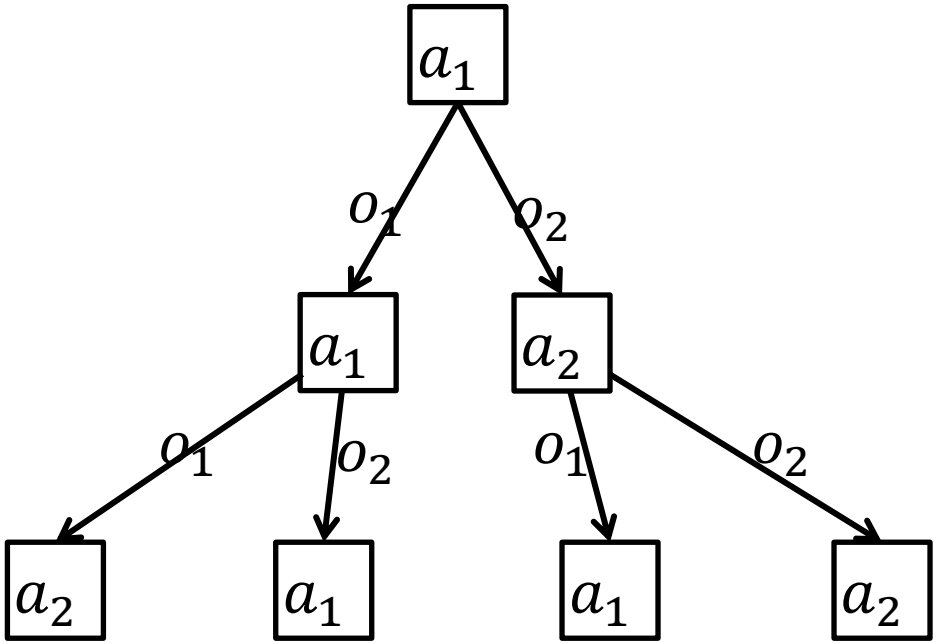
  - H_t is the space histories h_t of observables up to time t

$$h_t \stackrel{\text{def}}{=} \langle a_0, o_1, a_1, o_2, \ldots, a_{t-1}, o_t \rangle$$

  - Non-Markovian policy

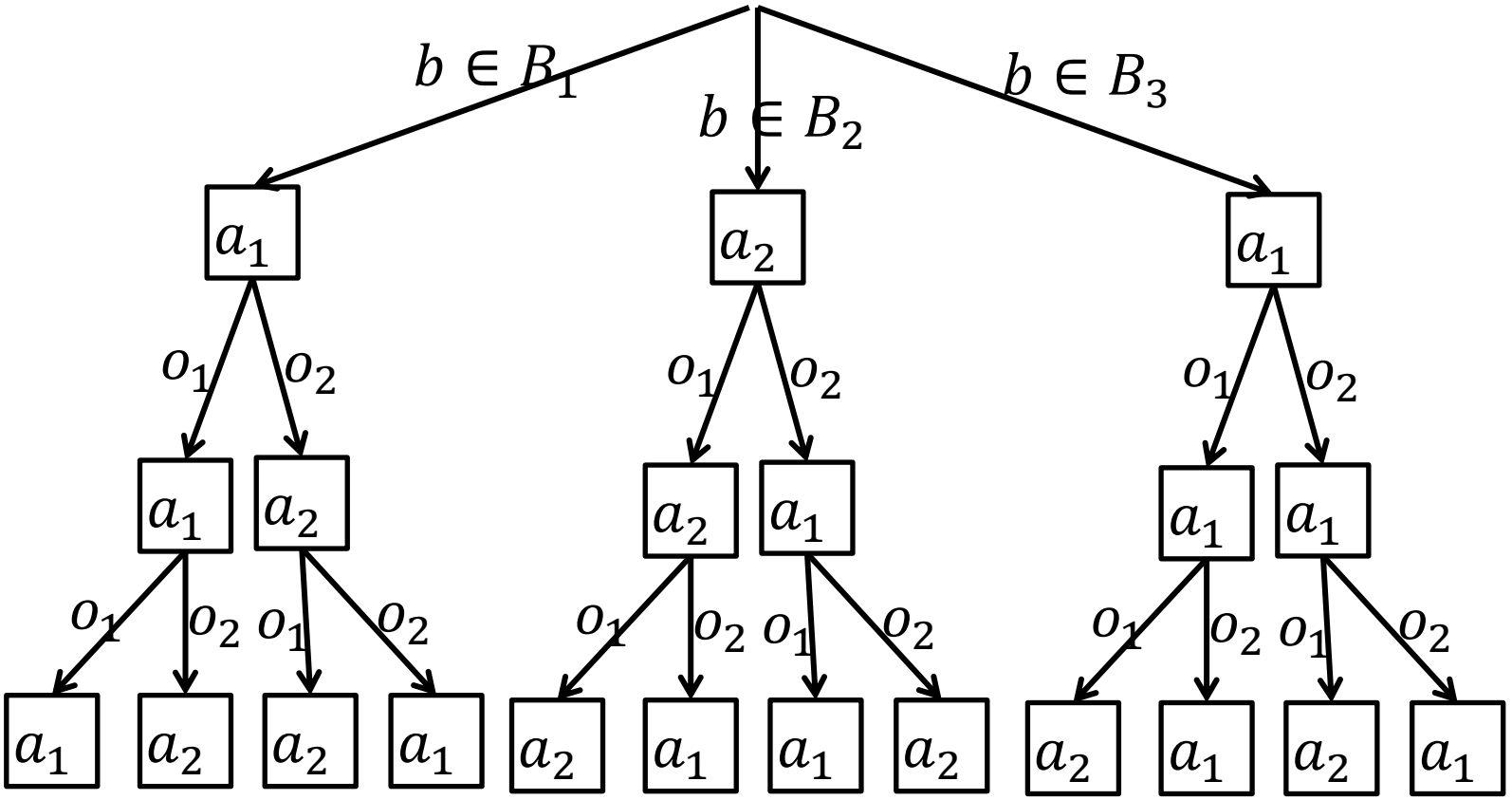
Policy Trees

- Policy $\pi: B \times H_t \rightarrow A_t$
- Consider a single initial belief b
- Then π can be represented by a tree



Policy Trees (continued)

- Policy $\pi: B \times H_t \rightarrow A_t$
  - Set of trees



Beliefs

- Belief $b_t(s) = \Pr(s_t)$
  - Distribution over states at time t

- Belief about the underlying state based on history h_t

$$b_t(s) = \Pr(s_t | h_t, b_0)$$

Belief Update

- Belief update: $b_t, a_t, o_{t+1} \rightarrow b_{t+1}$

$$b_{t+1}(s_{t+1}) = \Pr(s_{t+1}|h_{t+1}, b_0)$$

$$= \Pr(s_{t+1}|o_{t+1}, a_t, h_t, b_0) \qquad h_{t+1} \equiv o_{t+1}, a_t, h_t$$

$$= \Pr(s_{t+1}|o_{t+1}, a_t, b_t) \qquad b_t \equiv b_0, h_t$$

$$= \frac{\Pr(s_{t+1}, o_{t+1}|a_t, b_t)}{\Pr(o_{t+1}|a_t, b_t)} \qquad \text{Bayes' theorem}$$

$$= \frac{\Pr(o_{t+1}|s_{t+1}, a_t)\,\Pr(s_{t+1}|a_t, b_t)}{\Pr(o_{t+1}|a_t, b_t)} \qquad \text{chain rule}$$

$$= \frac{\Pr(o_{t+1}|s_{t+1}, a_t)\,\sum_{s_t}\Pr(s_{t+1}|s_t, a_t) b_t(s_t)}{\Pr(o_{t+1}|a_t, b_t)} \qquad \text{belief definition}$$

$$\propto \Pr(o_{t+1}|s_{t+1}, a_t)\,\sum_{s_t}\Pr(s_{t+1}|s_t, a_t)\, b_t(s_t)$$

Markovian Policies

- Beliefs are sufficient statistics equivalent to histories (with the initial belief)

$$b_0, h_t \Leftrightarrow b_t$$

- Policies:
  - Based on histories: $\pi: B_0 \times H_t \rightarrow A_t$
    - Non-Markovian
  - Based on beliefs: $\pi: B \rightarrow A$
    - Markovian

Belief State MDPs

- POMDPs can be viewed as belief state MDPs
  - States: B (beliefs)
  - Actions: A
  - Transitions:

$$\Pr(b_{t+1}|b_t, a_t) = \begin{cases} \Pr(o_{t+1}|b_t, a_t) & \text{if } b_t, a_t, o_{t+1} \rightarrow b_{t+1} \\ 0 & \text{otherwise} \end{cases}$$

  - Rewards: $R(b, a) = \sum_s b(s)R(s, a)$

- Belief state MDPs
  - Fully observable
  - Continuous belief space

Policy Evaluation

- Value V^{π} of a POMDP policy π
  - Expected sum of rewards:

$$V^{\pi}(b) = E\left[\sum_{t} \gamma^{t} R(b_t, \pi(b_t))\right]$$

  - Policy evaluation: Bellman's equation

$$V^{\pi}(b) = R(b, \pi(b)) + \gamma \sum_{b'} \Pr(b' | b, \pi(b))\, V^{\pi}(b') \quad \forall b$$

  - Equivalent equation

$$V^{\pi}(b) = R\big(b, \pi(b)\big) + \gamma \sum_{o'} \Pr(o' | b, a)\, V^{\pi}(b^{a,o'}) \quad \forall b$$

Policy Tree Value Function

- Theorem: The value function $V^{\pi}(b)$ of a policy tree is linear in b
  - i.e. $V^{\pi}(b) = \sum_s \alpha(s) b(s)$
- Proof by induction:
  - Base case: at the leaves
    - $V_0(b) = R\big(b, \pi(b)\big) = \sum_s b(s) R(s, \pi(s))$
      - Hence $\alpha(s) = R(s, \pi(s))$
  - Assumption: for all trees of depth n, there exists an α-vector such that $V_n(b) = \sum_s b(s)\alpha(s)$

Proof continued

- Induction

$$V_{n+1}(b) = R\big(b,\pi(b)\big) + \gamma \sum_{o'} \Pr(o'|b,\pi(b))\, V_n(b^{\pi(b),o'})$$

$$= R\big(b,\pi(b)\big) + \gamma \sum_{o'} \Pr(o'|b,\pi(b)) \sum_{s'} b^{\pi(b),o'}(s')\alpha^{o'}(s')$$

$$= R\big(b,\pi(b)\big) + \gamma \sum_{o'} \Pr(o'|b,\pi(b)) \sum_{s,s'} \frac{b(s)\Pr\big(s'|s,\pi(b)\big)\Pr\big(o'|s',\pi(b)\big)}{\Pr(o'|b,\pi(b))}\alpha^{o'}(s')$$

$$= \sum_s b(s)\Big[R\big(s,\pi(b)\big) + \gamma \sum_{o',s'} \Pr(s'|s,\pi(b))\Pr(o'|s',\pi(b))\,\alpha^{o'}(s')\Big]$$

$$\alpha(s)$$

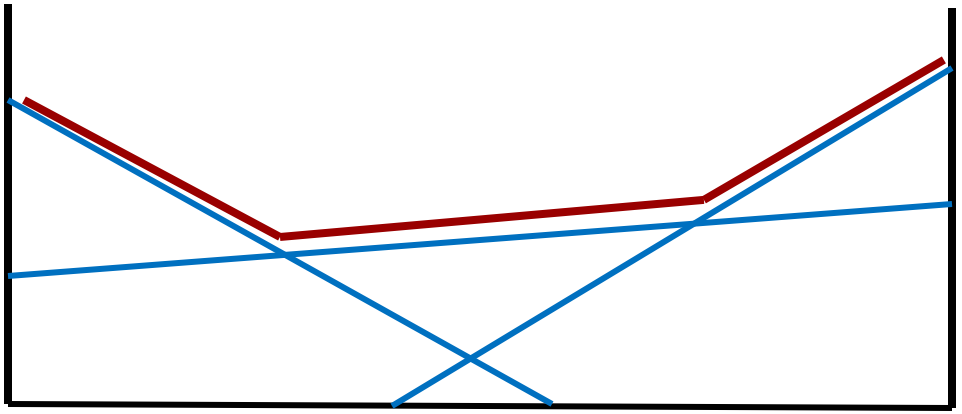
Value Function

- Corollary: A policy made up of a set of trees is piece-wise linear

- Proof:
  - Each tree leads to a linear piece for a region of the belief space
  - Hence the value function is made up of several linear pieces.

Optimal Value Function

- Theorem: Optimal value function $V^*(b)$ for finite horizon is piece-wise linear and convex in b
- Proof:
  - There are finitely many trees of finite depth
  - Each tree gives rise to a linear piece α
  - At each belief, select the highest linear piece



Value Iteration

- Bellman's Equation:

$$V^*(b) = \max_a R(b,a) + \gamma \sum_{o'} \Pr(o'|b,a)\, V^*(b^{a,o'})$$

- Value Iteration:
  - Idea: repeat

$$V^*(b) \leftarrow \max_a R(b,a) + \gamma \textstyle\sum_{o'} \Pr(o'|b,a)\, V^*(b^{a,o'}) \quad \forall b$$

  - But we can't enumerate all beliefs
  - Instead compute linear pieces α for a subset of beliefs

Point-Base Value Iteration

- Let $B = \{b_1, b_2, \ldots, b_k\}$ be a subset of beliefs
- Let $\Gamma = \{\alpha_1, \alpha_2, \ldots, \alpha_k\}$ be a set of α-vectors such that α_i is associated with b_i

- Point-based value iteration:
  - Repeatedly improve $V(b_i)$ at each b_i

$$V(b_i) = \max_a R(b_i, a) + \gamma \sum_{o'} \Pr(o'|b, a) \max_{\alpha \in \Gamma} \alpha(b^{a,o'})$$

  - Find $\alpha_i(b)$ such that $V(b_i) = \sum_s b_i(s)\alpha(s)$
    - $\alpha^{a,o'} \leftarrow argmax_{\alpha \in \Gamma} \sum_{s'} b_i^{a,o'}(s')\alpha(s')$
    - $a^* \leftarrow argmax_a R(b_i, a) + \gamma \sum_{o'} \Pr(o'|b_i, a)\, \alpha^{a,o'}(b_i^{a,o'})$
    - $\alpha_i(s) \leftarrow R(s, a^*) + \gamma \sum_{s',o'} \Pr(s'|s, a^*) \Pr(o'|s', a^*)\, \alpha^{a^*,o'}(s')$

Algorithm

Point-base Value Iteration(B,h)

Let B be a set of beliefs

$\alpha_{init}(s) = \min_{a,s} \frac{R(s,a)}{1-\gamma} \quad \forall s$

$\Gamma_0 \leftarrow \{\alpha_{init}\}$

For $n = 1$ to h do

For each $b_i \in B$ do

$\alpha^{a,o'} \leftarrow argmax_{\alpha \in \Gamma_n} \sum_{s'} b_i^{a,o'}(s')\alpha(s')$

$a^* \leftarrow argmax_a \, R(b_i, a) + \gamma \sum_{o'} \Pr(o'|b_i, a)\, \alpha^{a,o'}(b_i^{a,o'})$

$\alpha_i(s) \leftarrow R(s, a^*) + \gamma \sum_{s',o'} \Pr(s'|s, a^*) \Pr(o'|s', a^*)\, \alpha^{a^*,o'}(s')$

$\Gamma_n \leftarrow \{\alpha_i\}_{\forall i}$

Return Γ_n