ECEN 743: Reinforcement Learning

Partially Observed MDP (POMDP)

Dileep Kalathil
Assistant Professor
Department of Electrical and Computer Engineering
Texas A&M University

References

- [Chapter 6] P.R. Kumar and P. Varaiya, "Stochastic Systems: Estimation, Identification, and Adaptive Control", Prentice Hall, 1986.

Partial Observation

- One (implicit) core assumption of all MDP/RL algorithms we discussed so far is that perfect state observation is available at each time to select the control action

Partial Observation

- One (implicit) core assumption of all MDP/RL algorithms we discussed so far is that perfect state observation is available at each time to select the control action
- However, in real-world setting, state observations may be imperfect

Partial Observation

- One (implicit) core assumption of all MDP/RL algorithms we discussed so far is that perfect state observation is available at each time to select the control action
- However, in real-world setting, state observations may be imperfect
  - noisy sensors, state obtained from images, hidden states in games,

Partial Observation

- One (implicit) core assumption of all MDP/RL algorithms we discussed so far is that perfect state observation is available at each time to select the control action
- However, in real-world setting, state observations may be imperfect
  - noisy sensors, state obtained from images, hidden states in games,
- Algorithm may only get an imperfect observation $o_t \neq s_t$

Partial Observation

- One (implicit) core assumption of all MDP/RL algorithms we discussed so far is that perfect state observation is available at each time to select the control action
- However, in real-world setting, state observations may be imperfect
  - noisy sensors, state obtained from images, hidden states in games,
- Algorithm may only get an imperfect observation $o_t \neq s_t$
- How do we find the optimal policy for a system when the perfect state observation is not available?

POMDP

- Partially Observable MDP: $(\mathcal{S}, \mathcal{O}, \mathcal{A}, P, Q, r)$ where $\mathcal{O}$ is the observation space and Q is the observation probability function

POMDP

- Partially Observable MDP: $(\mathcal{S}, \mathcal{O}, \mathcal{A}, P, Q, r)$ where $\mathcal{O}$ is the observation space and Q is the observation probability function
  - State evolution is Markovian: $s_{t+1} \sim P(\cdot|s_t, a_t)$

POMDP

- Partially Observable MDP: $(\mathcal{S}, \mathcal{O}, \mathcal{A}, P, Q, r)$ where $\mathcal{O}$ is the observation space and Q is the observation probability function
  - State evolution is Markovian: $s_{t+1} \sim P(\cdot|s_t, a_t)$
  - At each time, algorithm gets an observation $o_t \in \mathcal{O}$

POMDP

- Partially Observable MDP: $(\mathcal{S}, \mathcal{O}, \mathcal{A}, P, Q, r)$ where $\mathcal{O}$ is the observation space and Q is the observation probability function
  - State evolution is Markovian: $s_{t+1} \sim P(\cdot | s_t, a_t)$
  - At each time, algorithm gets an observation $o_t \in \mathcal{O}$
  - Observation o_t depends only on s_t: $\mathbb{P}(o_t = o | s_{0:t}, o_{0:t-1}, a_{0:t-1}) = Q(o_t = o | s_t)$

POMDP

- Partially Observable MDP: $(\mathcal{S}, \mathcal{O}, \mathcal{A}, P, Q, r)$ where $\mathcal{O}$ is the observation space and Q is the observation probability function
  - State evolution is Markovian: $s_{t+1} \sim P(\cdot | s_t, a_t)$
  - At each time, algorithm gets an observation $o_t \in \mathcal{O}$
  - Observation o_t depends only on s_t: $\mathbb{P}(o_t = o | s_{0:t}, o_{0:t-1}, a_{0:t-1}) = Q(o_t = o | s_t)$
  - Reward function depends on the true (unobserved) state: $r(s_t, a_t)$

POMDP

- Partially Observable MDP: $(\mathcal{S}, \mathcal{O}, \mathcal{A}, P, Q, r)$ where $\mathcal{O}$ is the observation space and Q is the observation probability function
  - State evolution is Markovian: $s_{t+1} \sim P(\cdot|s_t, a_t)$
  - At each time, algorithm gets an observation $o_t \in \mathcal{O}$
  - Observation o_t depends only on s_t: $\mathbb{P}(o_t = o|s_{0:t}, o_{0:t-1}, a_{0:t-1}) = Q(o_t = o|s_t)$
  - Reward function depends on the true (unobserved) state: $r(s_t, a_t)$
- A control policy π_t specifies the action a_t to take at each time step t

POMDP

- Partially Observable MDP: $(\mathcal{S}, \mathcal{O}, \mathcal{A}, P, Q, r)$ where $\mathcal{O}$ is the observation space and Q is the observation probability function
  - State evolution is Markovian: $s_{t+1} \sim P(\cdot|s_t, a_t)$
  - At each time, algorithm gets an observation $o_t \in \mathcal{O}$
  - Observation o_t depends only on s_t: $\mathbb{P}(o_t = o|s_{0:t}, o_{0:t-1}, a_{0:t-1}) = Q(o_t = o|s_t)$
  - Reward function depends on the true (unobserved) state: $r(s_t, a_t)$
- A control policy π_t specifies the action a_t to take at each time step t
  - Control policy π_t can possibly select action a_t depending on the entire observed history $h_t = \{o_0, a_0, r_0, \ldots, o_{t-1}, a_{t-1}, r_{t-1}, o_t\}$

POMDP

- Partially Observable MDP: $(\mathcal{S}, \mathcal{O}, \mathcal{A}, P, Q, r)$ where $\mathcal{O}$ is the observation space and Q is the observation probability function
  - State evolution is Markovian: $s_{t+1} \sim P(\cdot|s_t, a_t)$
  - At each time, algorithm gets an observation $o_t \in \mathcal{O}$
  - Observation o_t depends only on s_t: $\mathbb{P}(o_t = o|s_{0:t}, o_{0:t-1}, a_{0:t-1}) = Q(o_t = o|s_t)$
  - Reward function depends on the true (unobserved) state: $r(s_t, a_t)$
- A control policy π_t specifies the action a_t to take at each time step t
  - Control policy π_t can possibly select action a_t depending on the entire observed history $h_t = \{o_0, a_0, r_0, \ldots, o_{t-1}, a_{t-1}, r_{t-1}, o_t\}$
- Objective is to select the optimal control policy which solves the problem

$$\max_{\pi} \mathbb{E}\left[\sum_{t=0}^{T-1} r(s_t, a_t)\right], \text{ where } a_t = \pi_t(h_t)$$

POMDP Control Policy

- Using a history dependent control policy is challenging

POMDP Control Policy

- Using a history dependent control policy is challenging
  - input dimension changes over the time, memory requirement increases linearly w.r.t. history length t, computational complexity increases exponentially w.r.t. history length t

POMDP Control Policy

- Using a history dependent control policy is challenging
  - input dimension changes over the time, memory requirement increases linearly w.r.t. history length t, computational complexity increases exponentially w.r.t. history length t
- In MDP, we were able to restrict our search to the class of Markov policies without loss of generality (recall the question from assignment 2)

POMDP Control Policy

- Using a history dependent control policy is challenging
  - input dimension changes over the time, memory requirement increases linearly w.r.t. history length t, computational complexity increases exponentially w.r.t. history length t
- In MDP, we were able to restrict our search to the class of Markov policies without loss of generality (recall the question from assignment 2)
- Can we use Markov policies for POMDP? What should be the input to the policy?

POMDP Control Policy

- Using a history dependent control policy is challenging
  - input dimension changes over the time, memory requirement increases linearly w.r.t. history length t, computational complexity increases exponentially w.r.t. history length t
- In MDP, we were able to restrict our search to the class of Markov policies without loss of generality (recall the question from assignment 2)
- Can we use Markov policies for POMDP? What should be the input to the policy?
- The observation o_t need not be Markovian

POMDP Control Policy

- Using a history dependent control policy is challenging
  - input dimension changes over the time, memory requirement increases linearly w.r.t. history length t, computational complexity increases exponentially w.r.t. history length t
- In MDP, we were able to restrict our search to the class of Markov policies without loss of generality (recall the question from assignment 2)
- Can we use Markov policies for POMDP? What should be the input to the policy?
- The observation o_t need not be Markovian
  - Consider a hidden Markov model with $\mathcal{S} = \{-1, 1\}$ and $o_t = s_t + w_t$, where w_t is zero mean Gaussian noise

POMDP Control Policy

- Using a history dependent control policy is challenging
  - input dimension changes over the time, memory requirement increases linearly w.r.t. history length t, computational complexity increases exponentially w.r.t. history length t
- In MDP, we were able to restrict our search to the class of Markov policies without loss of generality (recall the question from assignment 2)
- Can we use Markov policies for POMDP? What should be the input to the policy?
- The observation o_t need not be Markovian
  - Consider a hidden Markov model with $\mathcal{S} = \{-1, 1\}$ and $o_t = s_t + w_t$, where w_t is zero mean Gaussian noise
  - Assume that $P(s_{t+1} = 1 | s_t = 1) = P(s_{t+1} = -1 | s_t = -1) = 1 - \epsilon$. So, s_t tends to stick in the same state for a while

POMDP Control Policy

- Using a history dependent control policy is challenging
  - input dimension changes over the time, memory requirement increases linearly w.r.t. history length t, computational complexity increases exponentially w.r.t. history length t
- In MDP, we were able to restrict our search to the class of Markov policies without loss of generality (recall the question from assignment 2)
- Can we use Markov policies for POMDP? What should be the input to the policy?
- The observation o_t need not be Markovian
  - Consider a hidden Markov model with $\mathcal{S} = \{-1, 1\}$ and $o_t = s_t + w_t$, where w_t is zero mean Gaussian noise
  - Assume that $P(s_{t+1} = 1 | s_t = 1) = P(s_{t+1} = -1 | s_t = -1) = 1 - \epsilon$. So, s_t tends to stick in the same state for a while
  - Do we get better prediction by using history for o_{t+1} if $o_t = 4$? if $o_t = 0$?

Belief State

- The belief state at time t, denoted as $b_t \in \mathcal{R}^{|\mathcal{S}|}$, is defined as $b_t(s) = \mathbb{P}(s_t = s | h_t)$

Belief State

- The belief state at time t, denoted as $b_t \in \mathcal{R}^{|\mathcal{S}|}$, is defined as $b_t(s) = \mathbb{P}(s_t = s | h_t)$
- b_{t+1} can be computed using only b_t, a_t and o_{t+1}; we do not need to store history

Belief State

- The belief state at time t, denoted as $b_t \in \mathcal{R}^{|\mathcal{S}|}$, is defined as $b_t(s) = \mathbb{P}(s_t = s | h_t)$
- b_{t+1} can be computed using only b_t, a_t and o_{t+1}; we do not need to store history

Lemma

$$b_{t+1}(s) = \Phi(b_t, a_t, o_{t+1})(s) = \frac{Q(o_{t+1}|s_{t+1} = s) \sum_{s_t \in \mathcal{S}} P(s_{t+1} = s | s_t, a_t) b_t(s_t)}{\sum_{s_{t+1} \in \mathcal{S}} Q(o_{t+1}|s_{t+1}) \sum_{s_t \in \mathcal{S}} P(s_{t+1} | s_t, a_t) b_t(s_t)}$$

Belief State

- The belief state at time t, denoted as $b_t \in \mathcal{R}^{|\mathcal{S}|}$, is defined as $b_t(s) = \mathbb{P}(s_t = s | h_t)$
- b_{t+1} can be computed using only b_t, a_t and o_{t+1}; we do not need to store history

Lemma

$$b_{t+1}(s) = \Phi(b_t, a_t, o_{t+1})(s) = \frac{Q(o_{t+1}|s_{t+1} = s) \sum_{s_t \in \mathcal{S}} P(s_{t+1} = s | s_t, a_t) b_t(s_t)}{\sum_{s_{t+1} \in \mathcal{S}} Q(o_{t+1}|s_{t+1}) \sum_{s_t \in \mathcal{S}} P(s_{t+1}|s_t, a_t) b_t(s_t)}$$

$$\text{Prediction} : b_{t+1|t}(s) := \mathbb{P}(s_{t+1} = s | h_t, a_t) = \sum_{s_t \in \mathcal{S}} P(s_{t+1} = s | s_t, a_t) b_t(s_t)$$

$$\text{Correction} : b_{t+1}(s) = \frac{b_{t+1|t}(s) Q(o_{t+1}|s_{t+1} = s)}{\sum_s b_{t+1|t}(s) Q(o_{t+1}|s_{t+1} = s)}$$

Belief State Update

Belief State Update

Belief State Update

$$\mathbb{P}(s_{t+1}|h_{t+1}) = \frac{\mathbb{P}(s_{t+1}, h_t, a_t, o_{t+1})}{\mathbb{P}(h_{t+1})}$$

Belief State Update

$$\mathbb{P}(s_{t+1}|h_{t+1}) = \frac{\mathbb{P}(s_{t+1}, h_t, a_t, o_{t+1})}{\mathbb{P}(h_{t+1})} = \frac{\mathbb{P}(o_{t+1}|s_{t+1}, h_t, a_t)\mathbb{P}(s_{t+1}, h_t, a_t)}{\mathbb{P}(h_{t+1})}$$

Belief State Update

$$\mathbb{P}(s_{t+1}|h_{t+1}) = \frac{\mathbb{P}(s_{t+1}, h_t, a_t, o_{t+1})}{\mathbb{P}(h_{t+1})} = \frac{\mathbb{P}(o_{t+1}|s_{t+1}, h_t, a_t)\mathbb{P}(s_{t+1}, h_t, a_t)}{\mathbb{P}(h_{t+1})}$$
$$= \frac{Q(o_{t+1}|s_{t+1})\mathbb{P}(s_{t+1}, h_t, a_t)}{\mathbb{P}(h_{t+1})}$$

Belief State Update

$$\mathbb{P}(s_{t+1}|h_{t+1}) = \frac{\mathbb{P}(s_{t+1}, h_t, a_t, o_{t+1})}{\mathbb{P}(h_{t+1})} = \frac{\mathbb{P}(o_{t+1}|s_{t+1}, h_t, a_t)\mathbb{P}(s_{t+1}, h_t, a_t)}{\mathbb{P}(h_{t+1})}$$

$$= \frac{Q(o_{t+1}|s_{t+1})\mathbb{P}(s_{t+1}, h_t, a_t)}{\mathbb{P}(h_{t+1})} = \frac{Q(o_{t+1}|s_{t+1})\mathbb{P}(s_{t+1}|h_t, a_t)\mathbb{P}(h_t, a_t)}{\mathbb{P}(h_{t+1})}$$

Belief State Update

$$\mathbb{P}(s_{t+1}|h_{t+1}) = \frac{\mathbb{P}(s_{t+1}, h_t, a_t, o_{t+1})}{\mathbb{P}(h_{t+1})} = \frac{\mathbb{P}(o_{t+1}|s_{t+1}, h_t, a_t)\mathbb{P}(s_{t+1}, h_t, a_t)}{\mathbb{P}(h_{t+1})}$$

$$= \frac{Q(o_{t+1}|s_{t+1})\mathbb{P}(s_{t+1}, h_t, a_t)}{\mathbb{P}(h_{t+1})} = \frac{Q(o_{t+1}|s_{t+1})\mathbb{P}(s_{t+1}|h_t, a_t)\mathbb{P}(h_t, a_t)}{\mathbb{P}(h_{t+1})}$$

Next,

$$\mathbb{P}(h_{t+1}) = \sum_{s_{t+1} \in \mathcal{S}} \mathbb{P}(s_{t+1}, h_{t+1})$$

Belief State Update

$$\mathbb{P}(s_{t+1}|h_{t+1}) = \frac{\mathbb{P}(s_{t+1}, h_t, a_t, o_{t+1})}{\mathbb{P}(h_{t+1})} = \frac{\mathbb{P}(o_{t+1}|s_{t+1}, h_t, a_t)\mathbb{P}(s_{t+1}, h_t, a_t)}{\mathbb{P}(h_{t+1})}$$

$$= \frac{Q(o_{t+1}|s_{t+1})\mathbb{P}(s_{t+1}, h_t, a_t)}{\mathbb{P}(h_{t+1})} = \frac{Q(o_{t+1}|s_{t+1})\mathbb{P}(s_{t+1}|h_t, a_t)\mathbb{P}(h_t, a_t)}{\mathbb{P}(h_{t+1})}$$

Next,

$$\mathbb{P}(h_{t+1}) = \sum_{s_{t+1}\in\mathcal{S}} \mathbb{P}(s_{t+1}, h_{t+1}) = \sum_{s_{t+1}\in\mathcal{S}} \mathbb{P}(s_{t+1}, h_t, a_t, o_{t+1})$$

Belief State Update

$$\mathbb{P}(s_{t+1}|h_{t+1}) = \frac{\mathbb{P}(s_{t+1}, h_t, a_t, o_{t+1})}{\mathbb{P}(h_{t+1})} = \frac{\mathbb{P}(o_{t+1}|s_{t+1}, h_t, a_t)\mathbb{P}(s_{t+1}, h_t, a_t)}{\mathbb{P}(h_{t+1})}$$
$$= \frac{Q(o_{t+1}|s_{t+1})\mathbb{P}(s_{t+1}, h_t, a_t)}{\mathbb{P}(h_{t+1})} = \frac{Q(o_{t+1}|s_{t+1})\mathbb{P}(s_{t+1}|h_t, a_t)\mathbb{P}(h_t, a_t)}{\mathbb{P}(h_{t+1})}$$

Next,

$$\mathbb{P}(h_{t+1}) = \sum_{s_{t+1}\in\mathcal{S}} \mathbb{P}(s_{t+1}, h_{t+1}) = \sum_{s_{t+1}\in\mathcal{S}} \mathbb{P}(s_{t+1}, h_t, a_t, o_{t+1}) = \sum_{s_{t+1}\in\mathcal{S}} \mathbb{P}(o_{t+1}|s_{t+1}, a_t, h_t)\mathbb{P}(s_{t+1}, a_t, h_t)$$

Belief State Update

$$\begin{aligned}\mathbb{P}(s_{t+1}|h_{t+1}) &= \frac{\mathbb{P}(s_{t+1}, h_t, a_t, o_{t+1})}{\mathbb{P}(h_{t+1})} = \frac{\mathbb{P}(o_{t+1}|s_{t+1}, h_t, a_t)\mathbb{P}(s_{t+1}, h_t, a_t)}{\mathbb{P}(h_{t+1})} \\ &= \frac{Q(o_{t+1}|s_{t+1})\mathbb{P}(s_{t+1}, h_t, a_t)}{\mathbb{P}(h_{t+1})} = \frac{Q(o_{t+1}|s_{t+1})\mathbb{P}(s_{t+1}|h_t, a_t)\mathbb{P}(h_t, a_t)}{\mathbb{P}(h_{t+1})}\end{aligned}$$

Next,

$$\begin{aligned}\mathbb{P}(h_{t+1}) &= \sum_{s_{t+1}\in\mathcal{S}} \mathbb{P}(s_{t+1}, h_{t+1}) = \sum_{s_{t+1}\in\mathcal{S}} \mathbb{P}(s_{t+1}, h_t, a_t, o_{t+1}) = \sum_{s_{t+1}\in\mathcal{S}} \mathbb{P}(o_{t+1}|s_{t+1}, a_t, h_t)\mathbb{P}(s_{t+1}, a_t, h_t) \\ &= \sum_{s_{t+1}\in\mathcal{S}} Q(o_{t+1}|s_{t+1})\mathbb{P}(s_{t+1}, a_t, h_t)\end{aligned}$$

Belief State Update

$$\mathbb{P}(s_{t+1}|h_{t+1}) = \frac{\mathbb{P}(s_{t+1}, h_t, a_t, o_{t+1})}{\mathbb{P}(h_{t+1})} = \frac{\mathbb{P}(o_{t+1}|s_{t+1}, h_t, a_t)\mathbb{P}(s_{t+1}, h_t, a_t)}{\mathbb{P}(h_{t+1})}$$
$$= \frac{Q(o_{t+1}|s_{t+1})\mathbb{P}(s_{t+1}, h_t, a_t)}{\mathbb{P}(h_{t+1})} = \frac{Q(o_{t+1}|s_{t+1})\mathbb{P}(s_{t+1}|h_t, a_t)\mathbb{P}(h_t, a_t)}{\mathbb{P}(h_{t+1})}$$

Next,

$$\mathbb{P}(h_{t+1}) = \sum_{s_{t+1}\in\mathcal{S}} \mathbb{P}(s_{t+1}, h_{t+1}) = \sum_{s_{t+1}\in\mathcal{S}} \mathbb{P}(s_{t+1}, h_t, a_t, o_{t+1}) = \sum_{s_{t+1}\in\mathcal{S}} \mathbb{P}(o_{t+1}|s_{t+1}, a_t, h_t)\mathbb{P}(s_{t+1}, a_t, h_t)$$
$$= \sum_{s_{t+1}\in\mathcal{S}} Q(o_{t+1}|s_{t+1})\mathbb{P}(s_{t+1}, a_t, h_t) = \sum_{s_{t+1}\in\mathcal{S}} Q(o_{t+1}|s_{t+1})\mathbb{P}(s_{t+1}|a_t, h_t)\mathbb{P}(a_t, h_t)$$

Belief State Update

$$\mathbb{P}(s_{t+1}|h_{t+1}) = \frac{\mathbb{P}(s_{t+1}, h_t, a_t, o_{t+1})}{\mathbb{P}(h_{t+1})} = \frac{\mathbb{P}(o_{t+1}|s_{t+1}, h_t, a_t)\mathbb{P}(s_{t+1}, h_t, a_t)}{\mathbb{P}(h_{t+1})}$$
$$= \frac{Q(o_{t+1}|s_{t+1})\mathbb{P}(s_{t+1}, h_t, a_t)}{\mathbb{P}(h_{t+1})} = \frac{Q(o_{t+1}|s_{t+1})\mathbb{P}(s_{t+1}|h_t, a_t)\mathbb{P}(h_t, a_t)}{\mathbb{P}(h_{t+1})}$$

Next,

$$\mathbb{P}(h_{t+1}) = \sum_{s_{t+1}\in\mathcal{S}} \mathbb{P}(s_{t+1}, h_{t+1}) = \sum_{s_{t+1}\in\mathcal{S}} \mathbb{P}(s_{t+1}, h_t, a_t, o_{t+1}) = \sum_{s_{t+1}\in\mathcal{S}} \mathbb{P}(o_{t+1}|s_{t+1}, a_t, h_t)\mathbb{P}(s_{t+1}, a_t, h_t)$$
$$= \sum_{s_{t+1}\in\mathcal{S}} Q(o_{t+1}|s_{t+1})\mathbb{P}(s_{t+1}, a_t, h_t) = \sum_{s_{t+1}\in\mathcal{S}} Q(o_{t+1}|s_{t+1})\mathbb{P}(s_{t+1}|a_t, h_t)\mathbb{P}(a_t, h_t)$$

Combining, we get,

$$\mathbb{P}(s_{t+1}|h_{t+1}) = \frac{Q(o_{t+1}|s_{t+1})\mathbb{P}(s_{t+1}|a_t, h_t)}{\sum_{s_{t+1}\in\mathcal{S}} Q(o_{t+1}|s_{t+1})\mathbb{P}(s_{t+1}|a_t, h_t)}$$

Belief State Update

$$\mathbb{P}(s_{t+1}|h_{t+1}) = \frac{\mathbb{P}(s_{t+1}, h_t, a_t, o_{t+1})}{\mathbb{P}(h_{t+1})} = \frac{\mathbb{P}(o_{t+1}|s_{t+1}, h_t, a_t)\mathbb{P}(s_{t+1}, h_t, a_t)}{\mathbb{P}(h_{t+1})}$$
$$= \frac{Q(o_{t+1}|s_{t+1})\mathbb{P}(s_{t+1}, h_t, a_t)}{\mathbb{P}(h_{t+1})} = \frac{Q(o_{t+1}|s_{t+1})\mathbb{P}(s_{t+1}|h_t, a_t)\mathbb{P}(h_t, a_t)}{\mathbb{P}(h_{t+1})}$$

Next,

$$\mathbb{P}(h_{t+1}) = \sum_{s_{t+1}\in\mathcal{S}} \mathbb{P}(s_{t+1}, h_{t+1}) = \sum_{s_{t+1}\in\mathcal{S}} \mathbb{P}(s_{t+1}, h_t, a_t, o_{t+1}) = \sum_{s_{t+1}\in\mathcal{S}} \mathbb{P}(o_{t+1}|s_{t+1}, a_t, h_t)\mathbb{P}(s_{t+1}, a_t, h_t)$$
$$= \sum_{s_{t+1}\in\mathcal{S}} Q(o_{t+1}|s_{t+1})\mathbb{P}(s_{t+1}, a_t, h_t) = \sum_{s_{t+1}\in\mathcal{S}} Q(o_{t+1}|s_{t+1})\mathbb{P}(s_{t+1}|a_t, h_t)\mathbb{P}(a_t, h_t)$$

Combining, we get,

$$\mathbb{P}(s_{t+1}|h_{t+1}) = \frac{Q(o_{t+1}|s_{t+1})\mathbb{P}(s_{t+1}|a_t, h_t)}{\sum_{s_{t+1}\in\mathcal{S}} Q(o_{t+1}|s_{t+1})\mathbb{P}(s_{t+1}|a_t, h_t)}$$

Now,

$$\mathbb{P}(s_{t+1}|a_t, h_t) = \sum_{s_t\in\mathcal{S}} \mathbb{P}(s_{t+1}, s_t|a_t, h_t) = \sum_{s_t\in\mathcal{S}} \mathbb{P}(s_{t+1}|s_t, a_t, h_t)\mathbb{P}(s_t|a_t, h_t)$$

Belief State Update

$$\mathbb{P}(s_{t+1}|h_{t+1}) = \frac{\mathbb{P}(s_{t+1}, h_t, a_t, o_{t+1})}{\mathbb{P}(h_{t+1})} = \frac{\mathbb{P}(o_{t+1}|s_{t+1}, h_t, a_t)\mathbb{P}(s_{t+1}, h_t, a_t)}{\mathbb{P}(h_{t+1})}$$
$$= \frac{Q(o_{t+1}|s_{t+1})\mathbb{P}(s_{t+1}, h_t, a_t)}{\mathbb{P}(h_{t+1})} = \frac{Q(o_{t+1}|s_{t+1})\mathbb{P}(s_{t+1}|h_t, a_t)\mathbb{P}(h_t, a_t)}{\mathbb{P}(h_{t+1})}$$

Next,

$$\mathbb{P}(h_{t+1}) = \sum_{s_{t+1}\in\mathcal{S}} \mathbb{P}(s_{t+1}, h_{t+1}) = \sum_{s_{t+1}\in\mathcal{S}} \mathbb{P}(s_{t+1}, h_t, a_t, o_{t+1}) = \sum_{s_{t+1}\in\mathcal{S}} \mathbb{P}(o_{t+1}|s_{t+1}, a_t, h_t)\mathbb{P}(s_{t+1}, a_t, h_t)$$
$$= \sum_{s_{t+1}\in\mathcal{S}} Q(o_{t+1}|s_{t+1})\mathbb{P}(s_{t+1}, a_t, h_t) = \sum_{s_{t+1}\in\mathcal{S}} Q(o_{t+1}|s_{t+1})\mathbb{P}(s_{t+1}|a_t, h_t)\mathbb{P}(a_t, h_t)$$

Combining, we get,

$$\mathbb{P}(s_{t+1}|h_{t+1}) = \frac{Q(o_{t+1}|s_{t+1})\mathbb{P}(s_{t+1}|a_t, h_t)}{\sum_{s_{t+1}\in\mathcal{S}} Q(o_{t+1}|s_{t+1})\mathbb{P}(s_{t+1}|a_t, h_t)}$$

Now,

$$\mathbb{P}(s_{t+1}|a_t, h_t) = \sum_{s_t\in\mathcal{S}} \mathbb{P}(s_{t+1}, s_t|a_t, h_t) = \sum_{s_t\in\mathcal{S}} \mathbb{P}(s_{t+1}|s_t, a_t, h_t)\mathbb{P}(s_t|a_t, h_t)$$
$$= \sum_{s_t\in\mathcal{S}} P(s_{t+1}|s_t, a_t)b_t(s_t)$$

Belief State Evolution

- We have proved that the belief state b_t evolves according to the equation $b_{t+1} = \Phi(b_t, a_t, o_{t+1})$:

$$b_{t+1}(s) = \frac{b_{t+1|t}(s)Q(o_{t+1}|s_{t+1} = s)}{\sum_s b_{t+1|t}(s)Q(o_{t+1}|s_{t+1} = s)}, \quad \text{where, } b_{t+1|t}(s) = \sum_{s_t \in \mathcal{S}} P(s_{t+1} = s|s_t, a_t)b_t(s_t)$$

Belief State Evolution

- We have proved that the belief state b_t evolves according to the equation $b_{t+1} = \Phi(b_t, a_t, o_{t+1})$:

$$b_{t+1}(s) = \frac{b_{t+1|t}(s)Q(o_{t+1}|s_{t+1} = s)}{\sum_s b_{t+1|t}(s)Q(o_{t+1}|s_{t+1} = s)}, \quad \text{where, } b_{t+1|t}(s) = \sum_{s_t \in \mathcal{S}} P(s_{t+1} = s|s_t, a_t)b_t(s_t)$$

- Note that $b_t \in \mathcal{B}$ where $\mathcal{B}$ is a probability simplex of dimension $|\mathcal{S}|$

Belief State Evolution

- We have proved that the belief state b_t evolves according to the equation $b_{t+1} = \Phi(b_t, a_t, o_{t+1})$:

$$b_{t+1}(s) = \frac{b_{t+1|t}(s) Q(o_{t+1}|s_{t+1} = s)}{\sum_s b_{t+1|t}(s) Q(o_{t+1}|s_{t+1} = s)}, \quad \text{where, } b_{t+1|t}(s) = \sum_{s_t \in \mathcal{S}} P(s_{t+1} = s|s_t, a_t) b_t(s_t)$$

- Note that $b_t \in \mathcal{B}$ where $\mathcal{B}$ is a probability simplex of dimension $|\mathcal{S}|$

Lemma

The belief state process $\{b_t\}_{t \geq 0}$ is a controlled Markov chain. Let $B \subset \mathcal{B}$. Then,

$$\mathbb{P}(b_{t+1} \in B | b_{0:t}, a_{0:t}) = \mathbb{P}(b_{t+1} \in B | b_t, a_t) = \sum_{o \in \mathcal{O}} \mathbb{1}\{\Phi(b_t, a_t, o) \in B\} \sum_s Q(o|s) b_{t+1|t}(s)$$

Belief State Evolution

- We have proved that the belief state b_t evolves according to the equation $b_{t+1} = \Phi(b_t, a_t, o_{t+1})$:

$$b_{t+1}(s) = \frac{b_{t+1|t}(s)Q(o_{t+1}|s_{t+1} = s)}{\sum_s b_{t+1|t}(s)Q(o_{t+1}|s_{t+1} = s)}, \quad \text{where, } b_{t+1|t}(s) = \sum_{s_t \in \mathcal{S}} P(s_{t+1} = s|s_t, a_t)b_t(s_t)$$

- Note that $b_t \in \mathcal{B}$ where $\mathcal{B}$ is a probability simplex of dimension $|\mathcal{S}|$

Lemma

The belief state process $\{b_t\}_{t \geq 0}$ is a controlled Markov chain. Let $B \subset \mathcal{B}$. Then,

$$\mathbb{P}(b_{t+1} \in B|b_{0:t}, a_{0:t}) = \mathbb{P}(b_{t+1} \in B|b_t, a_t) = \sum_{o \in \mathcal{O}} \mathbb{1}\{\Phi(b_t, a_t, o) \in B\} \sum_s Q(o|s)b_{t+1|t}(s)$$

Proof:

$$\mathbb{P}(o_{t+1} = o|b_{0:t}, a_{0:t}) = \sum_s \mathbb{P}(s_{t+1} = s, o_{t+1} = o|b_{0:t}, a_{0:t}) = \sum_s \mathbb{P}(o_{t+1} = o|s_{t+1} = s)\mathbb{P}(s_{t+1} = s|b_{0:t}, a_{0:t})$$

Belief State Evolution

- We have proved that the belief state b_t evolves according to the equation $b_{t+1} = \Phi(b_t, a_t, o_{t+1})$:

$$b_{t+1}(s) = \frac{b_{t+1|t}(s) Q(o_{t+1}|s_{t+1} = s)}{\sum_s b_{t+1|t}(s) Q(o_{t+1}|s_{t+1} = s)}, \quad \text{where, } b_{t+1|t}(s) = \sum_{s_t \in \mathcal{S}} P(s_{t+1} = s|s_t, a_t) b_t(s_t)$$

- Note that $b_t \in \mathcal{B}$ where $\mathcal{B}$ is a probability simplex of dimension $|\mathcal{S}|$

Lemma

The belief state process $\{b_t\}_{t \geq 0}$ is a controlled Markov chain. Let $B \subset \mathcal{B}$. Then,

$$\mathbb{P}(b_{t+1} \in B | b_{0:t}, a_{0:t}) = \mathbb{P}(b_{t+1} \in B | b_t, a_t) = \sum_{o \in \mathcal{O}} \mathbb{1}\{\Phi(b_t, a_t, o) \in B\} \sum_s Q(o|s) b_{t+1|t}(s)$$

Proof:

$$\mathbb{P}(o_{t+1} = o | b_{0:t}, a_{0:t}) = \sum_s \mathbb{P}(s_{t+1} = s, o_{t+1} = o | b_{0:t}, a_{0:t}) = \sum_s \mathbb{P}(o_{t+1} = o | s_{t+1} = s) \mathbb{P}(s_{t+1} = s | b_{0:t}, a_{0:t})$$

$$= \sum_s Q(o_{t+1} = o | s_{t+1} = s) b_{t+1|t}(s)$$

Belief State Evolution

- We have proved that the belief state b_t evolves according to the equation $b_{t+1} = \Phi(b_t, a_t, o_{t+1})$:

$$b_{t+1}(s) = \frac{b_{t+1|t}(s) Q(o_{t+1}|s_{t+1} = s)}{\sum_s b_{t+1|t}(s) Q(o_{t+1}|s_{t+1} = s)}, \quad \text{where, } b_{t+1|t}(s) = \sum_{s_t \in \mathcal{S}} P(s_{t+1} = s|s_t, a_t) b_t(s_t)$$

- Note that $b_t \in \mathcal{B}$ where $\mathcal{B}$ is a probability simplex of dimension $|\mathcal{S}|$

Lemma

The belief state process $\{b_t\}_{t \geq 0}$ is a controlled Markov chain. Let $B \subset \mathcal{B}$. Then,

$$\mathbb{P}(b_{t+1} \in B|b_{0:t}, a_{0:t}) = \mathbb{P}(b_{t+1} \in B|b_t, a_t) = \sum_{o \in \mathcal{O}} \mathbb{1}\{\Phi(b_t, a_t, o) \in B\} \sum_s Q(o|s) b_{t+1|t}(s)$$

Proof:

$$\mathbb{P}(o_{t+1} = o|b_{0:t}, a_{0:t}) = \sum_s \mathbb{P}(s_{t+1} = s, o_{t+1} = o|b_{0:t}, a_{0:t}) = \sum_s \mathbb{P}(o_{t+1} = o|s_{t+1} = s) \mathbb{P}(s_{t+1} = s|b_{0:t}, a_{0:t})$$
$$= \sum_s Q(o_{t+1} = o|s_{t+1} = s) b_{t+1|t}(s)$$

we also have, $\mathbb{P}(b_{t+1} \in B, o_{t+1} = o|b_{0:t}, a_{0:t}) = \mathbb{1}\{\Phi(b_t, a_t, o) \in B\} \mathbb{P}(o_{t+1} = o|b_{0:t}, a_{0:t})$

Belief State MDP

- Belief State MDP: $(\mathcal{B}, \mathcal{A}, \bar{P}, \bar{r})$
- $\bar{P}$ specifies the probability of transition to b_{t+1} given b_t and a_t
- $\bar{r}(b, a) = \sum_s b(s) r(s, a)$

Belief State MDP

- Belief State MDP: $(\mathcal{B}, \mathcal{A}, \bar{P}, \bar{r})$
- $\bar{P}$ specifies the probability of transition to b_{t+1} given b_t and a_t
- $\bar{r}(b, a) = \sum_s b(s) r(s, a)$
- We can then show that for any history dependent control policy $\pi = \{\pi_t\}_{t \geq 0}$, we have

$$\mathbb{E}_\pi[r(s_t, a_t)] = \mathbb{E}_\pi[\bar{r}(b_t, a_t)]$$

Belief State MDP

- Belief State MDP: $(\mathcal{B}, \mathcal{A}, \bar{P}, \bar{r})$
- $\bar{P}$ specifies the probability of transition to b_{t+1} given b_t and a_t
- $\bar{r}(b, a) = \sum_s b(s) r(s, a)$
- We can then show that for any history dependent control policy $\pi = \{\pi_t\}_{t \geq 0}$, we have

$$\mathbb{E}_\pi [r(s_t, a_t)] = \mathbb{E}_\pi [\bar{r}(b_t, a_t)]$$

$$\mathbb{E}_\pi [r(s_t, a_t)] = \mathbb{E}_\pi [\mathbb{E}_\pi [r(s_t, a_t) | h_t]]$$

Belief State MDP

- Belief State MDP: $(\mathcal{B}, \mathcal{A}, \bar{P}, \bar{r})$
- $\bar{P}$ specifies the probability of transition to b_{t+1} given b_t and a_t
- $\bar{r}(b, a) = \sum_s b(s) r(s, a)$
- We can then show that for any history dependent control policy $\pi = \{\pi_t\}_{t \geq 0}$, we have

$$\mathbb{E}_\pi[r(s_t, a_t)] = \mathbb{E}_\pi[\bar{r}(b_t, a_t)]$$

$$\mathbb{E}_\pi[r(s_t, a_t)] = \mathbb{E}_\pi[\mathbb{E}_\pi[r(s_t, a_t)|h_t]] = \mathbb{E}_\pi[\mathbb{E}_\pi[r(s_t, \pi_t(h_t))|h_t]]$$

Belief State MDP

- Belief State MDP: $(\mathcal{B}, \mathcal{A}, \bar{P}, \bar{r})$
- $\bar{P}$ specifies the probability of transition to b_{t+1} given b_t and a_t
- $\bar{r}(b, a) = \sum_s b(s) r(s, a)$
- We can then show that for any history dependent control policy $\pi = \{\pi_t\}_{t \geq 0}$, we have

$$\mathbb{E}_\pi[r(s_t, a_t)] = \mathbb{E}_\pi[\bar{r}(b_t, a_t)]$$

$$\begin{aligned}\mathbb{E}_\pi[r(s_t, a_t)] &= \mathbb{E}_\pi[\mathbb{E}_\pi[r(s_t, a_t)|h_t]] = \mathbb{E}_\pi[\mathbb{E}_\pi[r(s_t, \pi_t(h_t))|h_t]] \\ &= \mathbb{E}_\pi[\sum_s \mathbb{P}(s_t = s|h_t) r(s, \pi_t(h_t))]\end{aligned}$$

Belief State MDP

- Belief State MDP: $(\mathcal{B}, \mathcal{A}, \bar{P}, \bar{r})$
- $\bar{P}$ specifies the probability of transition to b_{t+1} given b_t and a_t
- $\bar{r}(b, a) = \sum_s b(s) r(s, a)$
- We can then show that for any history dependent control policy $\pi = \{\pi_t\}_{t \geq 0}$, we have

$$\mathbb{E}_\pi[r(s_t, a_t)] = \mathbb{E}_\pi[\bar{r}(b_t, a_t)]$$

$$\begin{aligned}
\mathbb{E}_\pi[r(s_t, a_t)] &= \mathbb{E}_\pi[\mathbb{E}_\pi[r(s_t, a_t)|h_t]] = \mathbb{E}_\pi[\mathbb{E}_\pi[r(s_t, \pi_t(h_t))|h_t]] \\
&= \mathbb{E}_\pi[\sum_s \mathbb{P}(s_t = s|h_t) r(s, \pi_t(h_t))] = \mathbb{E}_\pi[\sum_s b_t(s) r(s, \pi_t(h_t))]
\end{aligned}$$

Belief State MDP

- Belief State MDP: $(\mathcal{B}, \mathcal{A}, \bar{P}, \bar{r})$
- $\bar{P}$ specifies the probability of transition to b_{t+1} given b_t and a_t
- $\bar{r}(b, a) = \sum_s b(s) r(s, a)$
- We can then show that for any history dependent control policy $\pi = \{\pi_t\}_{t \geq 0}$, we have

$$\mathbb{E}_\pi[r(s_t, a_t)] = \mathbb{E}_\pi[\bar{r}(b_t, a_t)]$$

$$\begin{aligned}
\mathbb{E}_\pi[r(s_t, a_t)] &= \mathbb{E}_\pi[\mathbb{E}_\pi[r(s_t, a_t)|h_t]] = \mathbb{E}_\pi[\mathbb{E}_\pi[r(s_t, \pi_t(h_t))|h_t]] \\
&= \mathbb{E}_\pi[\sum_s \mathbb{P}(s_t = s|h_t) r(s, \pi_t(h_t))] = \mathbb{E}_\pi[\sum_s b_t(s) r(s, \pi_t(h_t))] \\
&= \mathbb{E}_\pi[\sum_s b_t(s) r(s, a_t)] = \mathbb{E}_\pi[\bar{r}(b_t, a_t)]
\end{aligned}$$

Belief State MDP/DP

- Using the previous result, we can show that for any history dependent control policy $\pi = \{\pi_t\}_{t\geq 0}$,

$$\mathbb{E}_{\pi}[\sum_{t=0}^{T-1} r(s_t, a_t)] = \mathbb{E}_{\pi}[\sum_{t=0}^{T-1} \bar{r}(b_t, a_t)]$$

Belief State MDP/DP

- Using the previous result, we can show that for any history dependent control policy $\pi = \{\pi_t\}_{t\geq 0}$,

$$\mathbb{E}_\pi[\sum_{t=0}^{T-1} r(s_t, a_t)] = \mathbb{E}_\pi[\sum_{t=0}^{T-1} \bar{r}(b_t, a_t)]$$

- So, we can consider only the belief-based policies where $a_t = \pi_t(b_{0:t}, a_{0:t-1})$

Belief State MDP/DP

- Using the previous result, we can show that for any history dependent control policy $\pi = \{\pi_t\}_{t\geq 0}$,

$$\mathbb{E}_\pi[\sum_{t=0}^{T-1} r(s_t, a_t)] = \mathbb{E}_\pi[\sum_{t=0}^{T-1} \bar{r}(b_t, a_t)]$$

- So, we can consider only the belief-based policies where $a_t = \pi_t(b_{0:t}, a_{0:t-1})$
- Now, using the same arguments as we used for the MDP, we can restrict to the class of belief-based Markov policies where $a_t = \pi_t(b_t)$

Belief State MDP/DP

- Using the previous result, we can show that for any history dependent control policy $\pi = \{\pi_t\}_{t\geq 0}$,
$$\mathbb{E}_\pi[\sum_{t=0}^{T-1} r(s_t, a_t)] = \mathbb{E}_\pi[\sum_{t=0}^{T-1} \bar{r}(b_t, a_t)]$$
- So, we can consider only the belief-based policies where $a_t = \pi_t(b_{0:t}, a_{0:t-1})$
- Now, using the same arguments as we used for the MDP, we can restrict to the class of belief-based Markov policies where $a_t = \pi_t(b_t)$
- So, we can use dynamic programming methods to compute the optimal policy for the Belief State MDP: $(\mathcal{B}, \mathcal{A}, \bar{P}, \bar{r})$

Belief State MDP/DP

- Using the previous result, we can show that for any history dependent control policy $\pi = \{\pi_t\}_{t \geq 0}$,

$$\mathbb{E}_\pi[\sum_{t=0}^{T-1} r(s_t, a_t)] = \mathbb{E}_\pi[\sum_{t=0}^{T-1} \bar{r}(b_t, a_t)]$$

- So, we can consider only the belief-based policies where $a_t = \pi_t(b_{0:t}, a_{0:t-1})$
- Now, using the same arguments as we used for the MDP, we can restrict to the class of belief-based Markov policies where $a_t = \pi_t(b_t)$
- So, we can use dynamic programming methods to compute the optimal policy for the Belief State MDP: $(\mathcal{B}, \mathcal{A}, \bar{P}, \bar{r})$
- By the previous argument, this will be the optimal policy for the original POMDP problem

Belief State MDP/DP

- Using the previous result, we can show that for any history dependent control policy $\pi = \{\pi_t\}_{t\geq 0}$,
$$\mathbb{E}_\pi[\sum_{t=0}^{T-1} r(s_t, a_t)] = \mathbb{E}_\pi[\sum_{t=0}^{T-1} \bar{r}(b_t, a_t)]$$
- So, we can consider only the belief-based policies where $a_t = \pi_t(b_{0:t}, a_{0:t-1})$
- Now, using the same arguments as we used for the MDP, we can restrict to the class of belief-based Markov policies where $a_t = \pi_t(b_t)$
- So, we can use dynamic programming methods to compute the optimal policy for the Belief State MDP: $(\mathcal{B}, \mathcal{A}, \bar{P}, \bar{r})$
- By the previous argument, this will be the optimal policy for the original POMDP problem
- Caveat: Belief State DP is computationally intractable

Belief State MDP/DP

- Using the previous result, we can show that for any history dependent control policy $\pi = \{\pi_t\}_{t\geq 0}$,

$$\mathbb{E}_\pi[\sum_{t=0}^{T-1} r(s_t, a_t)] = \mathbb{E}_\pi[\sum_{t=0}^{T-1} \bar{r}(b_t, a_t)]$$

- So, we can consider only the belief-based policies where $a_t = \pi_t(b_{0:t}, a_{0:t-1})$
- Now, using the same arguments as we used for the MDP, we can restrict to the class of belief-based Markov policies where $a_t = \pi_t(b_t)$
- So, we can use dynamic programming methods to compute the optimal policy for the Belief State MDP: $(\mathcal{B}, \mathcal{A}, \bar{P}, \bar{r})$
- By the previous argument, this will be the optimal policy for the original POMDP problem
- Caveat: Belief State DP is computationally intractable
  - Even if the original POMDP has N finite states, the belief state DP has an N-dimensional continuous state space

Control with Unknown Model

- Objective is to select the optimal control policy which solves the problem

$$\max_{\pi} \mathbb{E}\left[\sum_{t=0}^{T-1} r(s_t, a_t)\right], \text{ where } a_t = \pi_t(h_t), s_{t+1} \sim P_\theta(\cdot|s_t, a_t), \theta \in \Theta = \{\theta_1, \theta_2\}$$

Control with Unknown Model

- Objective is to select the optimal control policy which solves the problem

$$\max_{\pi} \; \mathbb{E}\left[\sum_{t=0}^{T-1} r(s_t, a_t)\right], \text{ where } a_t = \pi_t(h_t), s_{t+1} \sim P_\theta(\cdot|s_t, a_t), \theta \in \Theta = \{\theta_1, \theta_2\}$$

- Assume that full state observation s_t is available

Control with Unknown Model

- Objective is to select the optimal control policy which solves the problem

$$\max_{\pi} \mathbb{E}\left[\sum_{t=0}^{T-1} r(s_t, a_t)\right], \text{ where } a_t = \pi_t(h_t), s_{t+1} \sim P_\theta(\cdot|s_t, a_t), \theta \in \Theta = \{\theta_1, \theta_2\}$$

- Assume that full state observation s_t is available
- Assume that we know P_{θ_1} and P_{θ_2}. So, we can compute the optimal control policy for $\pi^*_{\theta_i}$ for P_{θ_i} by solving dynamic programming

Control with Unknown Model

- Objective is to select the optimal control policy which solves the problem

$$\max_{\pi} \mathbb{E}\left[\sum_{t=0}^{T-1} r(s_t, a_t)\right], \text{ where } a_t = \pi_t(h_t), s_{t+1} \sim P_\theta(\cdot|s_t, a_t), \theta \in \Theta = \{\theta_1, \theta_2\}$$

- Assume that full state observation s_t is available
- Assume that we know P_{θ_1} and P_{θ_2}. So, we can compute the optimal control policy for $\pi^*_{\theta_i}$ for P_{θ_i} by solving dynamic programming
- We don't know the "true" θ. We, however, may have some prior over Θ, denoted as $f_0(\cdot)$

Control with Unknown Model

- Objective is to select the optimal control policy which solves the problem

$$\max_{\pi} \mathbb{E}\left[\sum_{t=0}^{T-1} r(s_t, a_t)\right], \text{ where } a_t = \pi_t(h_t), s_{t+1} \sim P_\theta(\cdot|s_t, a_t), \theta \in \Theta = \{\theta_1, \theta_2\}$$

- Assume that full state observation s_t is available
- Assume that we know P_{θ_1} and P_{θ_2}. So, we can compute the optimal control policy for $\pi^*_{\theta_i}$ for P_{θ_i} by solving dynamic programming
- We don't know the "true" θ. We, however, may have some prior over Θ, denoted as $f_0(\cdot)$
- What should be the optimal control policy?

Control with Unknown Model

- Model is P_θ where $\theta \in \Theta$ with a prior f_0

Control with Unknown Model

- Model is P_θ where $\theta \in \Theta$ with a prior f_0
- State s_t is fully observed; History of observation $h_t = (s_{0:t}, a_{0:t-1})$

Control with Unknown Model

- Model is P_θ where $\theta \in \Theta$ with a prior f_0
- State s_t is fully observed; History of observation $h_t = (s_{0:t}, a_{0:t-1})$
- The true model parameter θ is unobserved

Control with Unknown Model

- Model is P_θ where $\theta \in \Theta$ with a prior f_0
- State s_t is fully observed; History of observation $h_t = (s_{0:t}, a_{0:t-1})$
- The true model parameter θ is unobserved
- Belief $b_t(\theta) = \mathbb{P}(\theta | h_t)$

Control with Unknown Model

- Model is P_θ where $\theta \in \Theta$ with a prior f_0
- State s_t is fully observed; History of observation $h_t = (s_{0:t}, a_{0:t-1})$
- The true model parameter θ is unobserved
- Belief $b_t(\theta) = \mathbb{P}(\theta | h_t)$
- We can show that

$$b_{t+1}(\theta) = \frac{P_\theta(s_{t+1}|s_t, a_t) b_t(\theta)}{\sum_{\theta \in \Theta} P_\theta(s_{t+1}|s_t, a_t) b_t(\theta)}$$

Control with Unknown Model

- Model is P_θ where $\theta \in \Theta$ with a prior f_0
- State s_t is fully observed; History of observation $h_t = (s_{0:t}, a_{0:t-1})$
- The true model parameter θ is unobserved
- Belief $b_t(\theta) = \mathbb{P}(\theta | h_t)$
- We can show that

$$b_{t+1}(\theta) = \frac{P_\theta(s_{t+1}|s_t, a_t) b_t(\theta)}{\sum_{\theta \in \Theta} P_\theta(s_{t+1}|s_t, a_t) b_t(\theta)}$$

- We can then consider the belief MDP with belief $\bar{b}_t = (s_t, b_t)$

Control with Unknown Model

- Model is P_θ where $\theta \in \Theta$ with a prior f_0
- State s_t is fully observed; History of observation $h_t = (s_{0:t}, a_{0:t-1})$
- The true model parameter θ is unobserved
- Belief $b_t(\theta) = \mathbb{P}(\theta | h_t)$
- We can show that

$$b_{t+1}(\theta) = \frac{P_\theta(s_{t+1}|s_t, a_t) b_t(\theta)}{\sum_{\theta \in \Theta} P_\theta(s_{t+1}|s_t, a_t) b_t(\theta)}$$

- We can then consider the belief MDP with belief $\bar{b}_t = (s_t, b_t)$
- Dual Control:

Control with Unknown Model

- Model is P_θ where $\theta \in \Theta$ with a prior f_0
- State s_t is fully observed; History of observation $h_t = (s_{0:t}, a_{0:t-1})$
- The true model parameter θ is unobserved
- Belief $b_t(\theta) = \mathbb{P}(\theta | h_t)$
- We can show that

$$b_{t+1}(\theta) = \frac{P_\theta(s_{t+1}|s_t, a_t) b_t(\theta)}{\sum_{\theta \in \Theta} P_\theta(s_{t+1}|s_t, a_t) b_t(\theta)}$$

- We can then consider the belief MDP with belief $\bar{b}_t = (s_t, b_t)$
- Dual Control:
    - Should you take control action to estimate the model quickly?

Control with Unknown Model

- Model is P_θ where $\theta \in \Theta$ with a prior f_0
- State s_t is fully observed; History of observation $h_t = (s_{0:t}, a_{0:t-1})$
- The true model parameter θ is unobserved
- Belief $b_t(\theta) = \mathbb{P}(\theta | h_t)$
- We can show that

$$b_{t+1}(\theta) = \frac{P_\theta(s_{t+1}|s_t, a_t) b_t(\theta)}{\sum_{\theta \in \Theta} P_\theta(s_{t+1}|s_t, a_t) b_t(\theta)}$$

- We can then consider the belief MDP with belief $\bar{b}_t = (s_t, b_t)$
- Dual Control:
  - Should you take control action to estimate the model quickly?
  - Should you take the control action to maximize the reward given the current estimate of the model?

Temporary page!

LaTeX was unable to guess the total number of pages correctly. As there was some unproces
that should have been added to the final page this extra page has been added to receive it.
If you rerun the document (without altering it) this surplus page will go away, because LaTeX
how many pages to expect for this document.