Bayesian reinforcement learning and partially observable Markov decision processes

Christos Dimitrakakis

EPFL

November 6, 2013
1 Introduction

2 Bayesian reinforcement learning
   - The expected utility
   - Stochastic branch and bound

3 Partially observable Markov decision processes
Summary of previous developments

- Probability and utility.
- Making decisions under uncertainty.
- Updating probabilities
- Optimal experiment design
- Markov decision processes
- Stochastic algorithms for Markov decision processes.
- MDP Approximations.
- Bayesian reinforcement learning
Summary of previous developments

- Probability and utility.
- Making decisions under uncertainty.
- Updating probabilities
- Optimal experiment design
- Markov decision processes
- Stochastic algorithms for Markov decision processes.
- MDP Approximations.
- Bayesian reinforcement learning
The reinforcement learning problem

Learning to act in an unknown environment, by interaction and reinforcement.
The reinforcement learning problem

Learning to act in an unknown environment, by interaction and reinforcement.

Markov decision processes (MDP)

We are in some environment \( \mu \), where at each time step \( t \):

- Observe state \( s_t \in S \).
- Take action \( a_t \in A \).
- Receive reward \( r_t \in \mathbb{R} \).
The reinforcement learning problem

Learning to act in an unknown environment, by interaction and reinforcement.

Markov decision processes (MDP)

We are in some environment $\mu$, where at each time step $t$:

- Observe state $s_t \in S$.
- Take action $a_t \in A$.
- Receive reward $r_t \in \mathbb{R}$.

The optimal policy for a given $\omega$

- Find policy $\pi : S \rightarrow A$ maximising the utility $U = \sum_t r_t$ in expectation.
The reinforcement learning problem

Learning to act in an unknown environment, by interaction and reinforcement.

Markov decision processes (MDP)

We are in some environment $\mu$, where at each time step $t$:

- Observe state $s_t \in S$.
- Take action $a_t \in A$.
- Receive reward $r_t \in \mathbb{R}$.

The optimal policy for a given $w$

- Find policy $\pi : S \rightarrow A$ maximising the utility $U = \sum_t r_t$ in expectation.
- When $w$ is known, use standard algorithms, such as value or policy iteration. However this is
The reinforcement learning problem

Learning to act in an unknown environment, by interaction and reinforcement.

Markov decision processes (MDP)

We are in some environment \( \mu \), where at each time step \( t \):

- Observe state \( s_t \in S \).
- Take action \( a_t \in A \).
- Receive reward \( r_t \in \mathbb{R} \).

The optimal policy for a given \( w \)

- Find policy \( \pi : S \rightarrow A \) maximising the utility \( U = \sum_t r_t \) in expectation.
- When \( w \) is known, use standard algorithms, such as value or policy iteration. However this is contrary to the problem definition!
The reinforcement learning problem

**Learning** to act in an **unknown** environment, by **interaction** and **reinforcement**.

Markov decision processes (MDP)

We are in some **environment** $\mu$, where at each time step $t$:
- Observe state $s_t \in S$.
- Take action $a_t \in A$.
- Receive reward $r_t \in \mathbb{R}$.

Bayesian RL: Use a subjective belief $\xi(\mu)$

$$\mathbb{E}(U \mid \pi, \xi)$$
The reinforcement learning problem

**Learning** to act in an **unknown** environment, by **interaction** and **reinforcement**.

Markov decision processes (MDP)

We are in some **environment** $\mu$, where at each time step $t$:

- Observe state $s_t \in S$.
- Take action $a_t \in A$.
- Receive reward $r_t \in \mathbb{R}$.

Bayesian RL: Use a subjective belief $\xi(\mu)$

$$
\mathbb{E}(U \mid \pi, \xi) = \sum_\mu \mathbb{E}(U \mid \pi, \mu) \xi(\mu)
$$
The reinforcement learning problem

Learning to act in an unknown environment, by interaction and reinforcement.

Markov decision processes (MDP)

We are in some environment \( \mu \), where at each time step \( t \):
- Observe state \( s_t \in S \).
- Take action \( a_t \in A \).
- Receive reward \( r_t \in \mathbb{R} \).

Bayesian RL: Use a subjective belief \( \xi(\mu) \)

Not actually easy as \( \pi \) must now map from complete histories to actions.

\[
U^*_\xi = \max_{\pi} \mathbb{E}(U \mid \pi, \xi) = \max_{\pi} \sum_{\mu} \mathbb{E}(U \mid \pi, \mu) \xi(\mu)
\]

Planning must take into account future learning.
Updating the belief

Example

When the number of MDPs is finite

Exercise

Another practical scenario is when we have an independent belief over the transition probabilities of each state-action pair. Consider the case where we have $n$ states and $k$ actions. Similar to the product-prior in the bandit exercise of exercise set 4, we assign a probability (density) $\xi_{s,a}$ to the probability vector $\theta_{(s,a)} \in \mathbb{S}^n$. We can then define our joint belief on the $(nk) \times n$ matrix $\Theta$ to be

$$\xi(\Theta) = \prod_{s \in S, a \in A} \xi_{s,a}(\theta_{(s,a)}).$$

Derive the updates for a product-Dirichlet prior on transitions and a product-Normal-Gamma prior on rewards.

What is the meaning of using a Normal-Wishart prior on rewards?
The expected MDP heuristic

1. For a given belief $\xi$, calculate the expected MDP:

$$\bar{\mu}_\xi \triangleq \mathbb{E}_\xi \mu.$$  

2. Calculate the optimal memoryless policy for $\bar{\mu}_\xi$:

$$\pi^*(\bar{\mu}_\xi) \in \arg \max_{\pi \in \Pi_1} V_{\bar{\mu}_\xi},$$

where $\Pi_1 = \{ \pi \in \Pi \mid \mathbb{P}_\pi(a_t \mid s^t, a^{t-1}) = \mathbb{P}_\pi(a_t \mid s_t) \}.$

3. Execute $\pi^*(\bar{\mu}_\xi)$.

**Problem**

Unfortunately, this approach may be far from the optimal policy in $\Pi_1$. 
Counterexample\(^1\)

\[ \mathcal{M} = \{ \mu_1, \mu_2 \}, \quad \xi(\mu_1) = \theta, \quad \xi(\mu_2) = 1 - \theta, \]  

deterministic transitions.

- For \( T \to \infty \), the \( \bar{\mu}_\xi \)-optimal policy is not optimal in \( \Pi_1 \) if:

\[ \epsilon < \frac{\gamma \theta (1 - \theta)}{1 - \gamma} \left( \frac{1}{1 - \gamma \theta} + \frac{1}{1 - \gamma (1 - \theta)} \right) \]

- In this example, \( \bar{\mu}_\xi \notin \mathcal{M} \).
- For smooth beliefs, \( \bar{\mu}_\xi \) is close to \( \hat{\mu}_\xi^* \).

\(^1\)Based on one by Remi Munos
Counterexample for $\hat{\mu}_\xi^* \triangleq \arg \max_\mu \xi(\mu)$

MDP set $\mathcal{M} = \{\mu_i \mid i = 1, \ldots, n\}$ with $\mathcal{A} = \{0, \ldots, n\}$. In all MDPs, $a_0$ gives you a reward of $\epsilon$ and the MDP terminates. In the $i$-th MDP, all other actions give you a reward of 0 apart from the $i$-th action which gives you a reward of 1.

- The $\xi$-optimal policy takes action $i$ iff $\xi(\mu_i) \geq \epsilon$, otherwise takes action 0.
- The $\hat{\mu}_\xi^*$-optimal policy takes $a = \arg \max_i \xi(\mu_i)$.
Policy evaluation

Expected utility of a policy $\pi$ for a belief $\xi$

$$V_{\xi}^{\pi} \triangleq \mathbb{E}(U \mid \xi, \pi) \quad (2.1)$$

$$= \int_{\mathcal{M}} \mathbb{E}(U \mid \mu, \pi) \, d\xi(\mu) \quad (2.2)$$

$$= \int_{\mathcal{M}} V_{\mu}^{\pi} \, d\xi(\mu) \quad (2.3)$$

Bayesian Monte-Carlo policy evaluation

**Input** policy $\pi$, belief $\xi$

**For** $k = 1, \ldots, K$ **do**

- $\mu_k \sim \xi$
- $v_k = V_{\mu_k}^{\pi}$

**End For**

$$u = \frac{1}{K} \sum_{k=1}^{K} v_k.$$  

**Return** $u$. 

Upper bounds on the utility for a belief $\xi$

$$V^*_\xi \triangleq \sup_\pi \mathbb{E}(U \mid \xi, \pi) = \sup_\pi \int_\mathcal{M} \mathbb{E}(U \mid \mu, \pi) \, d\xi(\mu)$$  \hspace{1cm} (2.4)$$

$$\leq \int_\mathcal{M} \sup_\pi V^*_\mu \, d\xi(\mu) = \int_\mathcal{M} V^*_\mu \, d\xi(\mu) \triangleq V^+_\xi$$ \hspace{1cm} (2.5)$$

Bayesian Monte-Carlo upper bound

**input** policy $\pi$, belief $\xi$

**for** $k = 1, \ldots, K$ **do**

- $\mu_k \sim \xi$.
- $\nu_k = V^*_{\mu_k}$

**end for**

$$u^* = \frac{1}{K} \sum_{k=1}^K \nu_k.$$ 

**return** $u^*$.
Bounds on $V_\xi^* \triangleq \max_\pi \mathbb{E}( U \mid \pi, \xi )$
Bounds on $V_\xi^* \triangleq \max_\pi \mathbb{E}(U \mid \pi, \xi)$

Figure: A geometric view of the bounds
Bounds on $V^*_\xi \triangleq \max_{\pi} E(U | \pi, \xi)$

**Figure:** A geometric view of the bounds
Bounds on $V_\xi^* \triangleq \max_\pi \mathbb{E}(U \mid \pi, \xi)$
Bounds on $V_{\xi}^* \triangleq \max_\pi \mathbb{E}(U \mid \pi, \xi)$

Figure: A geometric view of the bounds
Bounds on $V_\xi^* \triangleq \max_\pi \mathbb{E}(U | \pi, \xi)$

Figure: A geometric view of the bounds
Better lower bounds [? ]

Main idea: maximisation in memoryless policies

- Then we can assume a fixed belief.
- Backwards induction on $n$ MDPs
- This improves the naive lower bound.
Bayesian reinforcement learning

The expected utility

\[ Q_{\pi, t}(s, a) \triangleq \int_{\mathcal{M}} \left\{ \bar{R}_\mu(s, a) + \gamma \int_{\mathcal{S}} V_{\mu, t+1}(s') dT_{\mu}^{s,a}(s') \right\} d\xi(\mu) \]  

(2.6)

Multi-MDP Backwards Induction

1: \text{MMBMI} \ M, \xi, \gamma, T
2: Set \( V_{\mu, t+1}(s) = 0 \) for all \( s \in \mathcal{S} \).
3: \textbf{for} \( t = T, T - 1, \ldots, 0 \) \textbf{do}
4: \quad \textbf{for} \( s \in \mathcal{S}, a \in \mathcal{A} \) \textbf{do}
5: \quad \quad Calculate \( Q_{\xi, t}(s, a) \) from (2.6) using \( \{ V_{\mu, t+1} \} \).
6: \quad \textbf{end for}
7: \textbf{for} \( s \in \mathcal{S} \) \textbf{do}
8: \quad \( a_{\xi, t}(s) \in \arg \max_{a \in \mathcal{A}} Q_{\xi, t}(s, a) \).
9: \quad \textbf{for} \( \mu \in \mathcal{M} \) \textbf{do}
10: \quad \quad \( V_{\mu, t}(s) = Q_{\mu, t}(s, a_{\xi, t}(s)) \).
11: \quad \textbf{end for}
12: \quad \textbf{end for}
13: \textbf{end for}
Bayesian reinforcement learning

The expected utility

MCBRL: Application to Bayesian RL

1. For $i = 1, \ldots$
2. At time $t_i$, sample $n$ MDPs from $\xi_{t_i}$.
3. Calculate best memoryless policy $\pi_i$ wrt the sample.
4. Execute $\pi_i$ until $t = t_{i+1}$.

Relation to other work

- For $n = 1$, this is equivalent to the Thompson sampling used by Strens \[8\].
- Unlike BOSS \[8\] it does not become more optimistic as $n$ increases.
- BEETLE\[8\] is a belief-sampling approach.
- Furmston and Barber \[8\] use approximate inference to estimate policies.
Generalisations

- Policy search for improving lower bounds.
- Search enlarged class of policies
- Examine all history-based policies.
The augmented MDP

The optimal policy for the augmented MDP is the $\xi$-optimal for the original problem.

$$P(s_{t+1} \in S \mid \xi_t, s_t, a_t) \triangleq \int_S P_\mu(s_{t+1} \in S \mid s_t, a_t) \, d\xi_t(\mu)$$  \hspace{1cm} (2.7)

$$\xi_{t+1}(\cdot) = \xi_t(\cdot \mid s_{t+1}, s_t, a_t)$$  \hspace{1cm} (2.8)
Consider an MDP family $\mathcal{M}$ with $\mathcal{A} = \{a^1, a^2\}$, $\mathcal{S} = \{s^1, s^2\}$.

$$\omega_t = (s_t, \xi_t)$$
Belief-augmented MDP tree structure

Consider an MDP family $\mathcal{M}$ with $\mathcal{A} = \{a^1, a^2\}, \mathcal{S} = \{s^1, s^2\}$.

$\omega_t = (s_t, \xi_t)$
Consider an MDP family $\mathcal{M}$ with $\mathcal{A} = \{a^1, a^2\}$, $\mathcal{S} = \{s^1, s^2\}$.

$$\omega_t = (s_t, \xi_t)$$
Branch and bound

Value bounds

Let upper and lower bounds $q^+$ and $q^-$ such that:

$$ q^+(\omega, a) \geq Q^*(\omega, a) \geq q^-(\omega, a) \quad (2.9) $$

$$ v^+(\omega) = \max_{a \in \mathcal{A}} Q^+(\omega, a), \quad v^-(\omega) = \max_{a \in \mathcal{A}} Q^-(\omega, a). \quad (2.10) $$

$$ q^+(\omega, a) = \sum_{\omega'} p(\omega' | \omega, a) \left[ r(\omega, a, \omega') + V^+(\omega') \right] \quad (2.11) $$

$$ q^-(\omega, a) = \sum_{\omega'} p(\omega' | \omega, a) \left[ r(\omega, a, \omega') + V^-(\omega') \right] \quad (2.12) $$

Remark

If $q^-(\omega, a) \geq q^+(\omega, b)$ then $b$ is sub-optimal at $\omega$. 
Stochastic branch and bound for belief tree search

- (Stochastic) Upper and lower bounds on the values of nodes (via Monte-Carlo sampling)
- Use upper bounds to expand tree, lower bounds to select final policy.
- Sub-optimal branches are quickly discarded.
Partially observable Markov decision processes (POMDP)

When acting in $\mu$, each time step $t$:

- The system state $s_t \in S$ is not observed.
- We receive an observation $x_t \in \mathcal{X}$ and a reward $r_t \in \mathcal{R}$.
- We take action $a_t \in \mathcal{A}$.
- The system transits to state $s_{t+1}$.

Definition

Partially observable Markov decision process (POMDP) A POMDP $\mu \in \mathcal{M}_P$ is a tuple $(\mathcal{X}, S, A, P)$ where $\mathcal{X}$ is an observation space, $S$ is a state space, $A$ is an action space, and $P$ is a conditional distribution on observations, states and rewards. The following Markov property holds:

$$P_{\mu}(s_{t+1}, r_t, x_t | s_t, a_t, \ldots) = P(s_{t+1} | s_t, a_t)P(x_t | s_t)P(r_t | s_t) \quad (3.1)$$
When acting in $\mu$, each time step $t$:

- The system state $s_t \in S$ is not observed.
- We receive an observation $x_t \in X$ and a reward $r_t \in R$.
- We take action $a_t \in A$.
- The system transits to state $s_{t+1}$.

**Definition**

Partially observable Markov decision process (POMDP) A POMDP $\mu \in \mathcal{M}_P$ is a tuple $(X, S, A, P)$ where $X$ is an observation space, $S$ is a state space, $A$ is an action space, and $P$ is a conditional distribution on observations, states and rewards. The following Markov property holds:

$$P_{\mu}(s_{t+1}, r_t, x_t \mid s_t, a_t, \ldots) = P(s_{t+1} \mid s_t, a_t)P(x_t \mid s_t)P(r_t \mid s_t) \quad (3.1)$$
Belief state in POMDPs when $\mu$ is known

If $\mu$ defines starting state probabilities, then the belief is not subjective

Belief $\xi$

For any distribution $\xi$ on $S$, we define:

$$\xi(s_{t+1} \mid a_t, \mu) \triangleq \int_S P_\mu(s_{t+1} \mid s_t a_t) \, d\xi(s_t) \quad (3.2)$$

Belief update

$$\xi_t(s_{t+1} \mid x_{t+1}, r_{t+1}, a_t, \mu) = \frac{P_\mu(x_{t+1}, r_{t+1} \mid s_{t+1}) \xi_t(s_{t+1} \mid a_t, \mu)}{\xi_t(x_{t+1} \mid a_t, \mu)} \quad (3.3)$$

$$\xi_t(s_{t+1} \mid a_t, \mu) = \int_S P_\mu(s_{t+1} \mid s_t, a_t, \mu) \, d\xi_t(s_t) \quad (3.4)$$

$$\xi_t(x_{t+1} \mid a_t, \mu) = \int_S P_\mu(x_{t+1} \mid s_{t+1}) \, d\xi_t(s_{t+1} \mid a_t, \mu) \quad (3.5)$$
Example

If $S, A, X$ are finite, and then we can define

- $\partial_t(j) = P(x_t \mid s_t = j)$
- $A_t(i, j) = P(s_{t+1} = j \mid s_t = i, a_t)$.
- $b_t(i) = \xi_t(s_t = i)$

We can then use Bayes theorem:

$$b_{t+1} = \frac{\text{diag}(p_{t+1}) A_t b_t}{p_t^T A_t b_t}, \quad (3.6)$$
When the POMDP $\mu$ is unknown

\[ \xi(\mu, s^t \mid x^t, a^t) \propto P_{\mu}(x^t \mid s^t, a^t)P_{\mu}(s^t \mid a^t)\xi(\mu) \]  

(3.7)

Cases

- Finite $\mathcal{M}$.
- Finite $\mathcal{S}$
- General case
Strategies for POMDPs

- Bayesian RL on POMDPs? EXP inference and planning
- Approximations and stochastic methods.
- Policy search methods.