Planning with Delayed State Information

Sascha E. Engelbrecht
Computer Science
University of Massachusetts
Amherst, MA 01003-4610
sascha@cs.umass.edu

Konstantinos V. Katsikopoulos
Mechanical and Industrial Engineering
University of Massachusetts
Amherst, MA 01003-2210
kkatsiko@ecs.umass.edu

Abstract

We consider a special case of partially observable Markov decision processes that arises when state information is perfect but arrives with a delay. We first formulate the decision process in its standard form and derive the Bellman equation that corresponds to it. We then introduce a second decision process that has a much simpler Bellman equation than the first, and is therefore, in general, much easier to solve. We demonstrate that even though the two decision processes have different optimal value functions, their optimal policies are the same. Exploitation of this result may lead to vast computational savings.

Introduction

Markov decision processes (MDPs) (Howard 1960) provide a useful framework for planning under uncertainty. In the standard MDP formulation, it is assumed that, at each stage of a decision process, the agent has access to perfect information about the system's state. However, many real-world decision processes violate the assumption of perfect state information. MDPs with imperfect state information are typically referred to as partially observable MDPs (POMDPs)(Sondik 1978). A special case of POMDPs arises when the agent receives state information that is perfect but arrives with a delay.

As discussed in (Zelevinsky 1998), the control of movement in biological systems may be thought of as planning with delayed state information because visual and proprioceptive information is relayed along neural pathways that have non-negligible transmission times[1]. Another domain in which decisions are frequently based on delayed state information is medical decision making. Here, state information may be delayed because the results of biochemical or other laboratory tests only become available after several days, or because test results only convey information relating to events much prior to testing (as, for instance, in AIDS tests).

[1]In humans, proprioceptive delays are roughly 30 ms, visual delays roughly 120 ms. Compared to the duration of a typical point-to-point arm movement, which is about 500 ms, these delay are quite significant.

Background

We are concerned with decision processes of the following form. We have a system with a finite number of states $\mathbf{s} \in \mathcal{S}$ and an agent that has available a finite number of actions $\mathbf{a} \in \mathcal{A}$. For each triple $(\mathbf{a}, \mathbf{s}, \mathbf{s}') \in \mathcal{A} \times \mathcal{S}^2$, we have a value $p_{\mathbf{a}}(\mathbf{s}'|\mathbf{s}) \in \mathcal{P}_A$ that specifies the probability of the system moving from $\mathbf{s}$ to $\mathbf{s}'$ if action $\mathbf{a}$ is implemented. For each pair $(\mathbf{s}, \mathbf{a})$, there is an expected immediate cost $g(\mathbf{s}, \mathbf{a})$, and the total expected cost along any system sample trajectory is defined as the infinite discounted sum $\sum_{k=0}^{\infty} \gamma^k g(\mathbf{s}_k, \mathbf{a}_k)$, with discount factor $\gamma \in (0, 1]$. The agent cannot observe the system's current state (and, by extension, it cannot observe the immediate cost $g(\mathbf{s}, \mathbf{a})$). Instead, it observes the state the system was in τ stages before. The history of the agent's observations and actions up to stage k, $\mathbf{H}_k = (\mathbf{s}_0, \mathbf{a}_0, \ldots, \mathbf{s}_{k-\tau}, \mathbf{a}_{k-\tau}, \ldots, \mathbf{a}_{k-1})$, defines a probability distribution over possible current states $s_k \in \mathcal{S}$. With regard to determining this probability distribution, some of the information in $\mathbf{H}_k$ is redundant. In particular, from the Markov property of the system's state transitions, it follows that

$$\begin{aligned} p(\mathbf{s}_k|\mathbf{H}_k) &= p(\mathbf{s}_k|\mathbf{s}_0, \mathbf{a}_0, \ldots, \mathbf{s}_{k-\tau}, \mathbf{a}_{k-\tau}, \ldots, \mathbf{a}_{k-1}) \\ &= p(\mathbf{s}_k|\mathbf{s}_{k-\tau}, \mathbf{a}_{k-\tau}, \ldots, \mathbf{a}_{k-1}). \end{aligned}$$

The truncated history

$$\mathbf{I}_k = (\mathbf{s}_{k-\tau}, \mathbf{a}_{k-\tau}, \ldots, \mathbf{a}_{k-1}) \in \mathcal{I} = S \times \mathcal{A}^{\tau}$$

therefore constitutes a sufficient statistic for the decision process, and we shall refer to it as the agent's information state. Note, also, that the information state transition $\mathbf{I}_k \rightarrow \mathbf{I}_{k+1}$ occurs with probability $p_{\mathbf{a}_{k-\tau}}(\mathbf{s}_{k-\tau+1}|\mathbf{s}_{k-\tau})$.

A policy for the decision process may now be defined as a function $\pi : \mathcal{I} \rightarrow \mathcal{A}$. For each $\mathbf{I} \in \mathcal{I}$, the expected total cost associated with policy π equals

$$V^{\pi}(\mathbf{I}) = E_{\pi}\left\{ \sum_{j=0}^{\infty} \gamma^j g(\mathbf{s}_j, \mathbf{a}_j) \middle| \mathbf{I}_0 = \mathbf{I} \right\}, \qquad (1)$$

where the expectation is taken with respect to the joint distribution of the random variables involved in the information-state transitions $\mathbf{I}_0 \rightarrow \mathbf{I}_1$, $\mathbf{I}_1 \rightarrow \mathbf{I}_2, \ldots$.

A decision process $P = <\mathcal{I}, \mathcal{A}, \mathcal{P}_A, g>$ is then defined as follows. Given the information-state space $\mathcal{I}$, action set A, transition-probability set $\mathcal{P}_A$, and cost function g, find a policy π that minimizes the expected total cost $V^\pi(\mathbf{I})$ for all $\mathbf{I} \in \mathcal{I}$.

The solution to this problem is typically obtained in two steps: First, for all $\mathbf{I} \in \mathcal{I}$, we solve the Bellman equation

$$\begin{aligned} V(\mathbf{I}) &= \min_{\mathbf{a}} \left[E\{g(\mathbf{s},\mathbf{a})|\mathbf{I}\} + \gamma \sum_{\mathbf{I}'} p(\mathbf{I}'|\mathbf{I}) V(\mathbf{I}') \right] \\ &= \min_{\mathbf{a}} \left[g(\mathbf{I},\mathbf{a}) + \gamma \sum_{\mathbf{I}'} p(\mathbf{s}'|\mathbf{s}) V(\mathbf{I}') \right], \end{aligned} \quad (2)$$

with

$$\begin{aligned} g(\mathbf{I},\mathbf{a}) &= E\{g(\mathbf{s},\mathbf{a})|\mathbf{I}\} \\ &= \sum_{\mathbf{s}_1,\ldots,\mathbf{s}_\tau} p_{a_0}(\mathbf{s}_1|\mathbf{s}_0) \cdot \ldots \cdot p_{a_{\tau-1}}(\mathbf{s}_\tau|\mathbf{s}_{\tau-1}) g(\mathbf{s}_\tau,\mathbf{a}), \\ &\qquad \mathbf{I}_\tau = \mathbf{I}. \end{aligned} \quad (3)$$

Second, using

$$\pi^* = \arg\min_{\mathbf{a}} \left[g(\mathbf{I},\mathbf{a}) + \gamma \sum_{\mathbf{I}'} p(\mathbf{s}'|\mathbf{s}) V(\mathbf{I}') \right], \quad (4)$$

we determine an optimal policy.

It should be clear that, for most decision processes of this type, the solution of the Bellman equation will require vast computational resources. This is not only so because of the size of $\mathcal{I}$, which is exponential in τ, but also because of the complexity of computing the expected immediate cost $g(\mathbf{I},\mathbf{a})$. From (3), it can be shown that the computation of $g(\mathbf{I},\mathbf{a})$ has a complexity of $\mathcal{O}(\tau|\mathcal{S}^2|)$. Fortunately, the latter complexity may be avoided if the decision process is appropriately reformulated.

Alternative Formulation

Let us consider a decision process $P_\tau = <\mathcal{I}, \mathcal{A}, \mathcal{P}_A, g_\tau>$ that is identical to P, except that the expected total cost associated with a policy π is now defined as

$$V_\tau^\pi(\mathbf{I}) = E_\pi \left\{ \sum_{j=0}^{\infty} \gamma^j g(\mathbf{s}_{j-\tau}, \mathbf{a}_{j-\tau}) \middle| \mathbf{I}_0 = \mathbf{I} \right\}. \quad (5)$$

This is a time-shifted version of cost function (1), from which it is obtained by replacing $g(\mathbf{s}_k, \mathbf{a}_k)$ with $g(\mathbf{s}_{k-\tau}, \mathbf{a}_{k-\tau})$. Conceptually, this means that costs are assigned based on where the system was τ stages before rather than were it is currently.

The Bellman equation for P_τ,

$$\begin{aligned} V_\tau(\mathbf{I}) &= \min_{\mathbf{a}} \Big[E\{g(\mathbf{s}_{-\tau}, \mathbf{a}_{-\tau})|\mathbf{I}_0 = \mathbf{I}\} \\ &\qquad + \gamma \sum_{\mathbf{I}'} p(\mathbf{I}'|\mathbf{I}) V_\tau(\mathbf{I}') \Big], \quad \forall \mathbf{I} \in \mathcal{I}, \end{aligned} \quad (6)$$

is, in general, much easier to solve than the one for P. Since $\mathbf{s}_{-\tau}$ and $\mathbf{a}_{-\tau}$ are known when $\mathbf{I}_0$ is given, $E\{g(\mathbf{s}_{-\tau}, \mathbf{a}_{-\tau})|\mathbf{I}_0\}$ reduces to the single term $g(\mathbf{s}_{-\tau}, \mathbf{a}_{-\tau})$.

From the above, it is clear that, if a decision process P can be reformulated as a decision process P_τ, vast computational savings are possible. Of course, such a reformulation can only make sense if a solution to P_τ is also a solution to P. As we shall see shortly, this is indeed the case.

Theorem (Time-Shift Equivalence). Let P_τ be the decision process $<\mathcal{I}, \mathcal{A}, \mathcal{P}_A, g_\tau>$ defined above, and let $\pi_\tau^* : \mathcal{I} \to \mathcal{A}$ be an optimal policy for P_τ. Then π_τ^* is also an optimal policy for the decision process P.

Proof. We start by noticing that P and P_τ are identical except for their definition of cost. In P, a given policy π has an associated total cost

$$V^\pi(\mathbf{I}) = E_\pi \left\{ \sum_{j=0}^{\infty} \gamma^j g(\mathbf{s}_j, \mathbf{a}_j) \middle| \mathbf{I}_0 = \mathbf{I} \right\},$$

while, in P_τ, the same π has an associated cost

$$\begin{aligned} V_\tau^\pi(\mathbf{I}) &= E_\pi \left\{ \sum_{j=0}^{\infty} \gamma^j g(\mathbf{s}_{j-\tau}, \mathbf{a}_{j-\tau}) \middle| \mathbf{I}_0 = \mathbf{I} \right\} \\ &= E_\pi \left\{ \sum_{j=-\tau}^{\infty} \gamma^{j+\tau} g(\mathbf{s}_j, \mathbf{a}_j) \middle| \mathbf{I}_0 = \mathbf{I} \right\}. \end{aligned}$$

Separating past and future costs, we may rewrite the latter expression in the form

$$\begin{aligned} V_\tau^\pi(\mathbf{I}) &= E_\pi \left\{ \left(\sum_{j=-\tau}^{-1} \gamma^{j+\tau} g(\mathbf{s}_j, \mathbf{a}_j) \right. \right. \\ &\qquad \left. \left. + \sum_{j=0}^{\infty} \gamma^{j+\tau} g(\mathbf{s}_j, \mathbf{a}_j) \right) \middle| \mathbf{I}_0 = \mathbf{I} \right\} \\ &= K(\mathbf{I}) + \gamma^\tau E \left\{ \sum_{j=0}^{\infty} \gamma^j g(\mathbf{s}_j, \mathbf{a}_j) \middle| \mathbf{I}_0 = \mathbf{I} \right\} \\ &= K(\mathbf{I}) + \gamma^\tau V^\pi(\mathbf{I}), \end{aligned}$$

with

$$K(\mathbf{I}) = E_\pi \left\{ \sum_{j=-\tau}^{-1} \gamma^{j+\tau} g(\mathbf{s}_j, \mathbf{a}_j) \middle| \mathbf{I}_0 = \mathbf{I} \right\}.$$

An optimal policy for P_τ is therefore a policy that, for all $\mathbf{I} \in \mathcal{I}$, satisfies

$$\begin{aligned} \pi_\tau^*(\mathbf{I}) &= \arg\min_{\pi} V_\tau^\pi(\mathbf{I}) \\ &= \arg\min_{\pi} \Big[K(\mathbf{I}) + \gamma^\tau V^\pi(\mathbf{I}) \Big]. \end{aligned}$$

Now note that $K(\mathbf{I})$ is a function only of $\mathbf{s}_{-\tau}$, $\mathbf{a}_{-\tau}, \ldots, \mathbf{a}_{-1}$, which are fixed when $\mathbf{I}$ is given. Hence, $K(\mathbf{I})$ is independent of π, and the above equation reduces to

$$\begin{aligned} \pi^*_\tau(\mathbf{I}) &= \arg\min_\pi \gamma^\tau V^\pi(\mathbf{I}) \\ &= \arg\min_\pi V^\pi(\mathbf{I}). \end{aligned}$$

This establishes that an optimal policy for P_τ is also an optimal policy for P (and vice versa).

Example

To illustrate the relationship between P and P_τ, we consider the following medical decision process. A person's hormone level may be in one of five different states, $\mathcal{S} = \{0, 1, 2, 3, 4\}$. The hormone level may be altered by application of one of two drugs, each of which may be given in dosages of 0–4 pills. One drug lowers the hormone level, the other increases it, so that $\mathcal{A} = \{-4, -3, -2, -1, 0, 1, 2, 3, 4\}$. Drug application alters the hormone level according to the equation $s' = B(s + a + w_a)$, where w_a is random variable that follows the discrete uniform distribution on the integers $-|a|, \ldots, |a|$, and $B(x)$ is a function defined as follows: $B(x) = 0$ if $x \leq 0$, $B(x) = x$ if $0 < x < 4$, and $B(x) = 4$ if $x \geq 4$. Drugs are given once a day. Also once a day, a blood sample is obtained from the patient and is submitted to biochemical analysis. The analysis results, which become available a day later ($\tau = 1$), give accurate information about the patient's hormone level at the day of testing. It is desirable to have a hormone level of 2. Deviation from the desired level leads to a cost of 1, and drug usuage leads to side-effects with cost $|a|$, so that $g(s, a) = |a|$ if $s = 2$, and $g(s, a) = |a| + 1$ otherwise. Long-term and short-term costs are given equal weight ($\gamma = 1$). Even though there is no discounting, it should be clear that there exist policies whose infinite-horizon costs are less than infinity. (Note that, if the hormone level is 2 and no drug is taken, the level remains at 2 and no cost is incurred.)

The above decision process has the general form P, and its solution may be computed using equations (2) and (4). It may also be reformulated as a decision process P_τ. We computed optimal cost functions and optimal policies for both formulations. The set of optimal policies for P is shown in Table 1, and, as expected, we found that the set of optimal policies for P_τ was the same as for P. From the proof of the above theorem, we know that the optimal cost function for P_τ should differ from the one for P by $K(\mathbf{I})$ (recall that $\gamma = 1$), which in the present example is equal to $g(s_{-1}, a_{-1})$. The optimal cost functions for P_τ and for P (values in parentheses) are shown in Table 2, and it may be verified that the differences between the two optimal cost functions are as expected.

Previous Dosage	Previous Cell Count 0	1	2	3	4
-4	{1}	{1}	{0, 1}	{0, 1}	{0, 1}
-3	{1}	{1}	{0, 1}	{0, 1}	{0, 1}
-2	{1}	{1}	{0, 1}	{0, 1}	{0}
-1	{1}	{1}	{0}	{0}	{0}
0	{1}	{1}	{0}	{-1}	{-1}
1	{0}	{0}	{0}	{-1}	{-1}
2	{0}	{-1, 0}	{-1, 0}	{-1}	{-1}
3	{-1, 0}	{-1, 0}	{-1, 0}	{-1}	{-1}
4	{-1, 0}	{-1, 0}	{-1, 0}	{-1}	{-1}

Table 1: Optimal policies

Previous Dosage	Previous Cell Count 0	1	2	3	4
-4	13.00 (8.00)	13.00 (8.00)	7.78 (7.78)	13.00 (8.00)	13.00 (8.00)
-3	12.00 (8.00)	12.00 (8.00)	7.71 (7.71)	11.71 (7.71)	11.71 (7.71)
-2	11.00 (8.00)	11.00 (8.00)	7.20 (7.20)	10.20 (7.20)	10.20 (7.20)
-1	10.00 (8.00)	10.00 (8.00)	6.00 (6.00)	8.00 (6.00)	8.00 (6.00)
0	9.00 (8.00)	9.00 (8.00)	0.00 (0.00)	9.00 (8.00)	9.00 (8.00)
1	8.00 (6.00)	8.00 (6.00)	6.00 (6.00)	10.00 (8.00)	10.00 (8.00)
2	10.20 (7.20)	10.20 (7.20)	7.20 (7.20)	11.00 (8.00)	11.00 (8.00)
3	11.71 (7.71)	11.71 (7.71)	7.71 (7.71)	12.00 (8.00)	12.00 (8.00)
4	13.00 (8.00)	13.00 (8.00)	7.78 (7.78)	13.00 (8.00)	13.00 (8.00)

Table 2: Optimal value functions for P_τ and for P (values in parentheses)

Summary

When an observation delay is introduced into a fully observable MDP, we obtain a new decision process that is only partially observable. If the delay has a constant duration of τ stages, the truncated history $\mathbf{I} \in \mathcal{I} = S \times \mathcal{A}^{\tau}$ constitutes a sufficient statistic for the decision process, and a policy that minimizes total expected discounted cost may be found from (2) and (4). However, vast computational savings are possible if, rather than using (2) and (4), the problem is first reformulated by replacing cost function (1) with its time-shifted version (5). This results in a new decision process that has a different optimal value function than the original one, but has the same optimal policies and is much easier to solve.

Acknowledgments. This research has been supported by grant JSMF 96-25 from the James S. McDonnell Foundation to the first author. We would like to thank Andy Barto, Rich Sutton, and Leo Zelevinsky (whose M.S. thesis motivated this work) for helpful discussions.

References

Howard, R. A. 1960. *Dynamic Programming and Markov Processes.* Cambridge, MA: MIT Press.

Sondik, E. J. 1978. The Optimal Control of Partially Observable Markov Processes Over the Infinite Horizon: Discounted Costs. *Operations Research* 26: 282-304.

Zelevinsky, L. 1998. Does Time-Optimal Control of a Stochastic System with Sensory Delay Produce Movement Units? M.S. thesis, Dept. of Computer Science, University of Massachusetts, Amherst.