Markov Decision Processes With Delays and Asynchronous Cost Collection

Konstantinos V. Katsikopoulos, *Member, IEEE,* and Sascha E. Engelbrecht

***Abstract*—Markov decision processes (MDPs) may involve three types of delays. First, state information, rather than being available instantaneously, may arrive with a delay (observation delay). Second, an action may take effect at a later decision stage rather than immediately (action delay). Third, the cost induced by an action may be collected after a number of stages (cost delay). We derive two results, one for constant and one for random delays, for reducing an MDP with delays to an MDP without delays, which differs only in the size of the state space. The results are based on the intuition that costs may be collected asynchronously, i.e., at a stage other than the one in which they are induced, as long as they are discounted properly.**

***Index Terms*—Asynchrony, delays, Markov decision processes (MDPs), neuro-dynamic programming.**

I. Introduction

HOWARD [1] popularized Markov decision processes (MDPs) as a general framework for treating control problems. An MDP is a quadruple $\langle S, A, P_A, g\rangle$. First, S is the finite state space of a discrete-time Markov chain that describes a system the decision maker wants to control optimally. Second, A is the set of options available to the decision maker for doing so. Third, P_A is a family of transition probability matrices, indexed by the set of actions, which is applied to the state–space. That is, for $a \in A$, the matrix P_a is defined so that the entry (s, s'), for s, $s' \in S$, equals the probability, $p_a(s'|s)$, that the system moves form state s to state s' when action a is selected. Note that here it is implicitly assumed that all actions $a \in A$ can be applied to all states $s \in S$. Fourth, $g : S \times A \rightarrow \Re$ is a real-valued, bounded function that represents the cost of selecting, when the process is in a certain state, a certain action. In summary, given a certain state of a system, the decision maker selects an action that brings the system to a new state and induces a cost, the new state is observed and the cost is collected, then the decision maker selects a new action, and so on.

We consider MDPs where actions are being selected for an infinite number of times, i.e., the decision-making horizon is infinite, and costs are being discounted by a factor γ, with $0 < \gamma < 1$. The objective is to minimize the total expected cost induced over the horizon. The expectation of the cost is taken over all possible sequences of states and actions following a certain initial state. Thus, the total cost, from any initial state, may be computed for any mapping $\pi : S \rightarrow A$ prescribing which action is selected at each state. Such mappings are termed policies and the major problem of MDP theory is to identify policies that minimize total cost for all possible initial states. A number of methods for doing so have been developed [2], and applied with much success to a variety of control problems [3].

However, the basic MDP framework, as portrayed above, makes a number of restrictive assumptions that may limit its applicability. In particular, it is assumed that 1) the system's current state is always available to the decision maker, 2) the decision maker's actions always take effect immediately, and 3) the cost induced by an action is always collected without delay. In many situations, these assumptions are not appropriate.

For example, in medical decision-making, results of laboratory tests are typically not available instantaneously. Additionally, even if they are available instantaneously, these results often refer only to the patient's condition at a previous rather than at the current time. Overall, with respect to state information, there exist observation delays. Furthermore, the effects of treatments on the patient's condition are typically not immediate. That is, the decision maker experiences action delays. Finally, treatments may also induce certain costs, e.g., side effects, which are collected at a later time. Thus, the decision maker may also experience cost delays.

MDPs with delays have been applied to a number of control problems, such as communication network design (Altman and Nain [4]), human motor control analysis (Zelevinsky [5]), and transportation information network design (Bander and White [6]). Theoretically, Brooks and Leondes [7], Kim [8], and Altman and Nain [4], as well as Kim and Jeong [9], White [10], and Bander and White [6] (for MDPs with partially observable states [11]), considered constant observation delays, and Berstekas [12] and Altman, Basar, and Srikant [13] considered constant action and cost delays. Constant observation delays have also been considered in decentralized control problems (Varaiya and Walrand [14], and Hsi and Marcus [15]).

The main result thus far is that an MDP with delays may be reformulated as an MDP without delays. Relative to the original MDP with delays, the MDP without delays has an augmented state space, the same state-transition structure, and a more complex cost structure involving a conditional expectation.

This work may be extended in two ways. First, it can be determined whether the equivalent MDP without delays may be

Manuscript received September 20, 2001; revised June 20, 2002. Recommended by Associate Editor X. Zhou. This work was supported in part by the McDonnell–Pew Foundation under Grant JSMF 96-25 to S. E. Engelbrecht.

K. V. Katsikopoulos was with the Department of Mechanical and Industrial Engineering, University of Massachusetts, Amherst, MA 01003–2210 USA. He is now with the Center for Adaptive Behavior and Cognition, Max Planck Institute for Human Development, 14195 Berlin, Germany (e-mail: katsikop@mpib-berlin.mpg.de).

S. E. Engelbrecht was with the Department of Computer Science, University of Massachusetts, Amherst, MA 01003-2210 USA. He is now with Dresdner Bank AG, Treasury Control IT, Frankfurt, Germany (Sascha.Engelbrecht@dresdner-bank.com).

Digital Object Identifier 10.1109/TAC.2003.809799

brought to a simpler form. Because the augmented states used in this MDP contain only the information necessary for optimal action selection, simplifications may be only made in the conditional expectation of the cost structure. Second, one may consider the case where the delays are random variables.

Here, we work out these two extensions. Specifically, the rest of the paper is organized as follows. In Section II, we show how the cost structure of the equivalent MDP without delays may be simplified when delays are constant. In Section III, we derive a more general result for random delays. Both results are based on the intuition that costs may be collected *asynchronously*, i.e., at a stage other than the one in which they are induced, as long as they are discounted properly. A summary is provided in Section IV.

II. Constant Delays

We denote an MDP with a constant observation delay o, constant action delay a, and constant cost delay c as a seven-tuple $\langle S, A, P_A, g, o, a, c\rangle$. We refer to such a seven-tuple as a deterministic delayed MDP (DDMDP). We now discuss how a DDMDP may be reduced to an MDP.

Altman and Nain [4] considered the case of no action delay. They showed that the DDMDP $\langle S, A, P_A, g, o, 0, c\rangle$ is reducible to an MDP without delays, $\langle \mathrm{I}_o, A, P_A, g'\rangle$, with $\mathrm{I}_o = S \times A^o$, where A^o is the Cartesian product of A with itself for o times, and $g'(I_k, a_k) = E[g(s_k, a_k)|I_k]$. Note then that, in the equivalent MDP without delays, policies are defined as mappings $\pi : S \times A^o \rightarrow A$.

That is, the information necessary for optimal action selection at the k-stage is contained in $I_k = (s_{k-o}, a_{k-o}, \ldots, a_{k-1})$ where s_{k-o} is the most recently observed system state and $a_{k-o}, \ldots, a_{k-1}$ are the actions taken since. Thus, an action is selected each time a system state and the o actions taken after the state has been observed are available.

If action a_k is selected, the information necessary for optimal action selection at the $(k+1)$-stage is $I_{k+1} = (s_{k-o+1}, a_{k-o+1}, \ldots, a_k)$ with probability $p_{k-o}(s_{k-o+1}|s_{k-o})$, which is the probability that the system moves from state s_{k-o} to state s_{k-o+1} according to the transition probability matrix corresponding to action a_{k-o}.

Finally, the selection of action a_k induces the cost $g(s_k, a_k)$. Note however that, from the point of view of the decision-maker, this cost is a random variable since the state s_k that is necessary for evaluating the cost is not known with certainty, but only the conditional probability distribution $p(s_k|I_k)$ may be computed. Thus, it is intuitive that the decision-maker should consider the expectation of the cost, $E[g(s_k, a_k)|I_k]$.

Note now that in the aforementioned definition of the state I_k, no information about induced costs is included. In order to justify this, it is necessary to assume that policies do not depend on costs. In turn, this can be seen as a consequence of the following assumption. The available cost information does not provide system state information for after the most recent stage at which a state has been observed, i.e., for after the $(k-o)$-stage. Note that the latter assumption is sufficient but not necessary for ensuring that cost information does not enter the definition of I_k.

That is, it suffices to be assumed that the costs induced after the $(k-o)$-stage have not been collected by the decision maker. In other words, it is assumed that $k-c \leq k-o$, or $o \leq c$, i.e., that cost delay is greater than observation delay. This may not be an implausible assumption. For example, in medical decision-making, it is reasonable to expect that the side effects of a treatment may be assessed only after the patient's condition at the time of the treatment is known.

This result represents the state of the art in DDMDP solution methodology. That is, DDMDPs are solved via reformulation as a MDP with an augmented state space I_o, the same state-transition structure P_A, and a more complex cost structure g'.

This new cost structure is more complex because, as said, the state s_k that is necessary for evaluating the cost induced at the k-stage is not known with certainty, but instead the conditional probability distribution $p(s_k|I_k)$ should be computed. This computation has a complexity on the order of $\mathrm{O}(o|S|^2)$. Our result for constant delays is that it is not necessary to compute the conditional expectation $E[g(s_k, a_k)|I_k]$. It is more instructive to first consider the case where there are no action delays.

A. No Action Delays

We will show that instead of $g'(I_k, a_k)$, the simpler $g_o(I_k, a_k) = g(s_{k-o}, a_{k-o})$ which is essentially the same as the DDMDP cost structure, may be used in the equivalent MDP. That is, each cost may be collected o stages after it has been induced. The derivations are better understood if a concept often used in neuro-dynamic programming [16], [17] is used. Neuro-dynamic programming is an approach to solving sequential decision making problems as MDPs, which combines neural network methods used in artificial intelligence and dynamic programming methods used in operations research.

A Q-value [18] is a concept that bears an interesting relation to MDPs with delays. Specifically, a value $Q(s, a)$ is an approximation to the optimal total cost given that at state$s$ action a has been selected, $Q*(s, a)$.

Now, by extension, here we define the *generalized* $Q{*}_o(s, a_1, a_2, \ldots, a_o)$ value as the optimal total cost given that at state s, action a_1 was selected *first*, action a_2 was selected second, and overall o actions were selected in the *order* denoted by their indices. Then, by conditioning on the outcome of the action implemented first, a_1, we have

$$Q{*}_o(s, a_1, \ldots, a_o) = \min_a \left[g(s, a_1) + \gamma \Sigma_{s'} p_1(s'|s) Q{*}_o(s', a_2, \ldots, a_o, a)\right].$$

Consistent with the definition of P_A, $p_1(s'|s)$ is the probability that the system moves from state s to state s' according to the transition probability matrix corresponding to action a_1.

Now, note that the above recursion does not have the form that typically appears in dynamic programming applications, i.e., the minimum is taken over action a while the state transition is controlled by a different action, a_1. This is so exactly because of the observation delay: the action that is selected by the decision-maker, a, is not responsible for the most recently observed system transition from state s to state s'.

By appropriately changing indexes, it follows that the generalized $Q*_o$ values are the optimal total costs of the MDP $\langle \mathrm{I}_o, A, P_A, g_o \rangle$. Thus, it also follows that

$$\begin{aligned} Q*_o(I_k) &= Q*_o(s_{k-o}, a_{k-o}, \ldots, a_{k-1}) \\ &= E * \left[\Sigma_{l \geq k} \gamma^{l-k} g(s_{l-o}, a_{l-o}) | I_k \right] \end{aligned}$$

where the expectation is taken over all possible sequences of system states and actions that may occur under an optimal policy $\pi^* : \mathrm{I}_o \rightarrow A$ when I_k is given.

More generally, for an arbitrary but fixed policy $\pi : \mathrm{I}_o \rightarrow A$

$$Q^\pi{}_o(I_k) = E_\pi \left[\Sigma_{l \geq k} \gamma^{l-k} g(s_{l-o}, a_{l-0}) | I_k \right]$$

where the expectation is taken over all possible sequences of system states and actions under π, when I_k is given.

Also, for π, the total cost for the MDP $\langle \mathrm{I}_o, A, P_A, g' \rangle$ is

$$V^\pi{}_o(I_k) = E_\pi \left[\Sigma_{l \geq k} \gamma^{l-k} g(s_l, a_l) | I_k \right].$$

Now, we will show that, there exists an $r = r(I_k)$, so that, for all I_k,

$$Q^\pi{}_o(I_k) = r(I_k) + \gamma^0 V^\pi_o(I_k). \tag{1}$$

It is easy to see that (1) holds by rewriting $E_\pi[\Sigma_{l \geq k} \gamma^{l-k} g (s_{l-o}, a_{l-o}) | I_k]$ as $E_\pi[\Sigma_{l \geq k-o} \gamma^{l-k+o} g(s_l, a_l) | I_k]$, and then by noticing that the latter equals

$$\begin{aligned} &(E_\pi \left[g(s_{k-o}, a_{k-o}) + \gamma g(s_{k-o+1}, a_{k-o+1}) + \ldots \right. \\ &\quad \left. + \gamma^{o-1} g(s_{k-1}, a_{k-1}) | I_k \right]) + \gamma^o E_\pi \left[\Sigma_{l \geq k} \gamma^{l-k} g(s_l, a_l) | I_k \right]. \end{aligned}$$

It follows that the MDPs $\langle \mathrm{I}_o, A, P_A, g_o \rangle$ and $\langle \mathrm{I}_o, A, P_A, g' \rangle$ have the same optimal policies since $arg\ min_\pi[Q^\pi{}_o(I_k)] = arg\ min_\pi[V^\pi{}_o(I_k)]$, for all I_k. Thus, we have the following.

Lemma 1: The DDMDP $\langle S, A, P_A, g, o, 0, c \rangle$, with $o \leq c$, is reducible to the MDP $\langle \mathrm{I}_o, A, P_A, g_o \rangle$.

Note here that the fact that $o \leq c$ does not directly enter the proof. But, recall that it is used to justify the definition of I_k used.

More importantly, note that Lemma 1 is based on asynchronous cost collection: costs are *not* collected at the stage they are induced. But, each cost is collected once. Also, the number of stages between the collections of two costs equals the number of stages between the inducements of the two costs. In this sense, costs are discounted *properly*. Policies are also evaluated properly in the sense that the ordering of total expected costs of policies is retained although note that total expected costs per se change. For a numerical example, see [19].

From the proof of Lemma 1 it is also clear that Lemma 1 does not hold if the horizon is finite. This is so because, for finite horizons, not all costs would be collected.

Now, we consider the case where $a \neq 0$. This will allow deriving our first result.

B. Action Delays

An intuitive argument can be made why action delay is, from the point of view of the decision maker, functionally equivalent to observation delay. Let us assume, for the moment, that no observation delay exists. A decision maker experiencing an action delay a, even if he/she has access to the current system state, will find himself/herself in a position of having to select an action that is to be applied on a system state that will be observed after a stages. In other words, this decision maker has to decide based on system state information that is delayed by a stages, as also happens in the observation delay case.

In fact, Bertsekas [12] has shown that the DDMDP $\langle S, A, P_A, g, 0, a, c \rangle$ is reducible to an MDP without delays $\langle \mathrm{I}_a, A, P_A, g_a \rangle$, with $\mathrm{I}_a = S \times A^a$, $g_a(I_k, a_k) = g(s_k, a_{k-a})$. Note then that, in the equivalent MDP without delays, policies are defined as mappings $\pi : S \times A^a \rightarrow A$.

That is, the information necessary for optimal action selection at the k-stage is contained in $I_k = (s_k, a_{k-a}, \ldots, a_{k-1})$. Thus, an action is selected each time a system state and the a actions taken before the state has been observed are available.

If action a_k is selected, the information necessary for optimal action selection at the $(k+1)$-stage is $I_{k+1} = (s_{k+1}, a_{k-a+1}, \ldots, a_k)$ with probability $p_{k-a}(s_{k+1}|s_k)$, which is the probability that the system moves from state s_k to state s_{k+1} according to the transition probability matrix corresponding to action a_{k-a}.

Finally, the cost induced at stage k equals, by definition, $g(s_k, a_{k-a})$.

From Lemma 1, the above imply that the DDMDPs $\langle S, A, P_A, g, 0, a, c \rangle$ and $\langle S, A, P_A, g, o, 0, c \rangle$ are formally identical for $a = o$.

More generally, by applying the arguments used in the proof of Lemma 1, it is easy to see that the effects of observation and action delays on the structure of the equivalent MDP without delays are *additive* in the following sense.

When $o \neq 0$ and $a \neq 0$, the necessary information for optimal action selection at the k-stage is contained in $I_k = (a_{k-o-a}, \ldots, a_{k-o-1}, s_{k-o}, a_{k-o}, \ldots, a_{k-1})$.

If action a_k is selected, the information necessary for optimal action selection at the $(k+1)$-stage is $I_{k+1} = (a_{k-o-a+1}, \ldots, a_{k-o}, s_{k-o+1}, a_{k-o+1}, , \ldots, a_k)$ with probability $p_{k-o-a}(s_{k-o+1}|s_{k-o})$, which is the probability that the system moves from state s_{k-o} to state s_{k-o+1} according to the transition probability matrix corresponding to action a_{k-o-a}.

Finally, the cost collected at stage k may be assumed to equal $g_o + a(s_k, a_k) = g(s_{k-o}, a_{k-o-a})$. Overall, the following holds.

Result 1: The DDMDP $\langle S, A, P_A, g, o, a, c \rangle$, with $o \leq c$, is reducible to the MDP $\langle \mathrm{I}_{o+a}, A, P_A, g_{o+a} \rangle$.

Note again that the fact that $o \leq c$ does not directly enter the proof, but it is used to justify the definition of I_k used. Also, the comments on asynchronous cost collection made after the statement of Lemma 1 apply here as well. Again, Result 1 also holds only for an infinite horizon.

Moreover, Result 1 makes clear the special nature of a partially observable MDP [11] that arises when the information available for action selection is incomplete due to observation and action delays. While such a MDP is partially observable with respect to the current decision stage, it is completely observable with respect to an appropriately chosen past decision stage. This fact, by itself, guarantees that the state-transition structure of the original MDP may be used. And, when this fact is coupled with asynchronous cost collection, it guarantees that the cost structure of the original MDP may be used.

III. Random Delays

We now discuss stochastic delayed MDPs (SDMDPs). First, for any decision stage $k = 1, 2, \ldots$ consider the pair of successive states s_k and s_{k+1}, the pair of successive actions a_k and a_{k+1}, and the pair of successive cost collections $g(s_k, a_k)$ and $g(s_{k+1}, a_{k+1})$.

Now, note that in DDMDPs the number of stages between the observations of such successive system states, between the effects of such successive actions, and between the collections of such successive costs is constant, in fact it equals exactly zero stages. The defining difference between a DDMDPs and SDMDPs is then that in SDMDPs we assume that the number of stages between the observations of successive states, the effects of successive actions, and the collections of successive costs are discrete, finite-valued, nonnegative random variables.

Thus, we define an SDMDP as an eight-tuple $\langle S, A, P_A, g, O, AC, C, \gamma\rangle$, with the random variables O, AC, and C denoting the number of stages between the observations of successive states s_k and s_{k+1}, the effects of successive actions a_k and a_{k+1}, and the collections of successive costs $g(s_k, a_k)$ and $g(s_{k+1}, a_{k+1})$, respectively. The reason for specifically including the discounting factor γ in this definition will become clear later.

Note that in this it is implicitly assumed that, for all stages k, it is possible to observe state s_{k+1} *only* after state s_k has been observed, and that action a_{k+1} can take effect only after action a_k has taken effect. This may be reasonable in situations where the mediums for observing system states and applying actions are *serial* in the sense that processing of a state observation and an action application can start only after the processing of the preceding state and action, respectively, has been completed. Finally, due to this assumption, the cost induced at the $(k+1)$-stage also can be collected *only* at a stage after the cost induced at the k-stage has been collected.

We now present some arguments similar to the ones presented in Section II to prove a lemma that generalizes Lemma 1 and a result that generalizes Result 1.

A. No Action Delays

We assume that successive system state observation delays are independent random variables, following a common probability distribution $P(O = o)$,$o = 0, 1, \ldots, o_{\max}$, and are also independent of state identity. There are no action delays.

We also assume that the available cost information does not provide system state information for after the most recent stage where a state was observed. This is accomplished by assuming that there is a zero probability that a cost is collected before the system state that results from the action that induced this cost is observed, i.e., $Pr(O \leq C) = 1$. Note that, as in the deterministic case, this assumption is sufficient but not necessary for ensuring that cost information does not enter the definition of I_k.

It is natural to define the state vector at stage k as $I_k = (s_{k-o}, k-r, a_{k-o}, \ldots, a_{k-1})$ where s_{k-o} is the most recently observed system state which was observed for the first time at stage $k-r$, with $r \leq o$, and $a_{k-o}, \ldots, a_{k-1}$ are the actions taken since stage $k-o$. Note here that it is necessary that r is included in the definition of as I_k since it affects the transition probabilities as explained below.

First, let $q(r) = Pr(O = r)/Pr(O \geq r)$ be the probability that the system state s_{k-o+1} is observed for the first time at stage $k+1$, given that the system state s_{k-o} was observed for the first time at stage $k-r$.

Then, note that the dimension of I_k is not constant. Specifically, if action a_k is selected, $I_{k+1} = (s_{k-o+1}, k+1, a_{k-o+1}, \ldots, a_k)$ with probability $p_{k-o}(s_{k-o+1}|s_{k-o})q(r)$, but $I_{k+1} = (s_{k-o}, k-r, a_{k-o}, \ldots, a_k)$ with probability $1 - q(r)$.

In other words, state vector dimension may remain constant or may increase by 1 at each stage. It follows that for an infinite horizon it will become infinity with probability 1. We assume then that there exists a maximum allowable state vector dimension, n. When this dimension is reached, it is assumed that the decision-making process *freezes* in the following sense.

First, the decision maker is assumed to take no actions until the most recent system state is observed. During this time period no new costs are induced since no actions are taken. Note however that although the underlying system does not make any new state transitions during that time period, the decision-maker does continue to observe the previous system state transitions.

Formally, let $I_k = (s_{k-n+2}, k-r, a_{k-n+2}, \ldots, a_{k-1})$ be a vector state with the maximum allowable dimension, n. The next events that occur are, in order, the observations of the following system states: $s_{k-n+3}, s_{k-n+4}, \ldots, s_k$. When all these events have occurred, the vector state is $I_k = (s_k)$ and then the decision-making process starts anew. Finally, when the maximum allowable state dimension is reached again, the decision-making process freezes again, and so on.

Overall, the maximum allowable vector state dimension assumption may be plausible in situations where making decisions with very old state information is highly undesirable.

By using similar arguments as in [4] and [12], it can be shown that the SDMDP $\langle S, A, P_A, g, O, 0, C, \gamma\rangle$ is reducible to the MDP $\langle \mathrm{I}_O, A, {P_A}^O, g'\rangle$ with the same discounting factor γ, with I_O and ${P_A}^O$ being the state space and state-transition structures previously described, and $g'(I_k, a_k) = E[g(s_k, a_k)|I_k]$.

We will use asynchronous cost collection to simplify the cost structure of the equivalent MDP without delays. Recall that for constant observation delays, the cost $g(s_{k-o}, a_{k-o})$ referring to the most recently observed system state was used at the k-stage. However, this choice is problematic for random observation delays because the same cost $g(s_{k-o}, a_{k-o})$ would be collected repeatedly until the next state s_{k-o+1} is observed.

We need to make sure that each cost is collected once. A reasonable way of doing this is to assume that $g(s_{k-o}, a_{k-o})$ is collected *only* in the first stage in which the effect of action a_{k-o} on state s_{k-o} is observed, i.e., only in the first stage where state s_{k-o+1} is observed. That is, in all decision stages after a_{k-o} has been applied and until the state s_{k-o+1} is observed, a zero cost is collected.

We label this cost structure g_O. Formally, let $I_k = (s_{k-o}, k-r, a_{k-o}, \ldots, a_{k-1})$, and let the state s_{k-o+1} be observed for the first time at the $(k+u)$-stage, with $u > 0$. Then, the following holds:

$$\begin{aligned} g_O(s_{k-r}, a_{k-r}) &= g(s_{k-o-1}, a_{k-o-1}) \\ g_O(s_{k-r+1}, a_{k-r+1}) &= g_O(s_{k-r+2}, a_{k-r+2}) = \cdots \\ &= g_O(s_{k+u-1}, a_{k+u-1}) = 0 \\ g_O(s_{k+u}, a_{k+u}) &= g(s_{k-o}, a_{k-o}). \end{aligned}$$

We expect to be able to repeat the arguments of the proof of Lemma 1. However, there exists an important subtlety. The number of stages intervening between successive cost collections using g_O is not constant across costs and, thus, the discounting factor is also not constant across costs. Thus, we anticipate that the optimal costs of the MDP $\langle \mathrm{I}_O, A, {P_A}^O, g_O \rangle$ use a different discounting factor from the optimal costs of the original SDMDP. Specifically, we will show the following.

Lemma 2: The SDMDP $\langle S, A, P_A, g, O, 0, C, \gamma \rangle$, with $Pr(O \leq C) = 1$, is reducible to the MDP $\langle \mathrm{I}_O, A, {P_A}^O, g_O \rangle$, with discounting factor $\gamma_O : E_O[{\gamma_O}^{O+1}] = \gamma$.

Proof: For an arbitrary but fixed policy $\pi : \mathrm{I}_O \rightarrow A$, let

$$\begin{aligned} {Q^\pi}_O(I_k) &= E_\pi \left[\Sigma_{l \geq k} {\gamma_O}^{l-k} g_O(I_l, a_l) | I_k \right] \\ &= E_{\pi,O} \left[\Sigma_{l \geq k} {\gamma_O}^{l-k} g(s_{l-o}, a_{l-o}) | I_k \right] \end{aligned}$$

where the second expectation is taken not only over all possible sequences of system states and actions under π given that I_k is known, but also over all possible sequences of observation delays. In the notation of the second expectation, it is also implied that the costs $g(s_{l-o}, a_{l-o}), l \geq k$, are collected only at the stage where the state s_{l-o+1} is observed first.

Also, for π, the total cost for MDP $\langle \mathrm{I}_O, A, {P_A}^O, g\prime \rangle$ with discounting factor γ is

$${V^\pi}_O(I_k) = E_\pi \left[\Sigma_{l \geq k} \gamma^{l-k} g(s_l, a_l) | I_k \right].$$

It suffices to find $r_O = r_O(I_k)$, and $u_O = u_O(\gamma, o, \gamma_O, O) > 0$ so that, for all I_k

$${Q^\pi}_O(I_k) = r_O(I_k) + u_O(\gamma, o, \gamma_O, O){V^\pi}_O(I_k)$$

However,

$$\begin{aligned} {Q^\pi}_O(I_k) &= E_{\pi,O} \left[\Sigma_{l \geq k} {\gamma_O}^{l-k} g(s_{l-o}, a_{l-o}) | I_k \right] \\ &= E_{\pi,O} \left[\Sigma_{l \geq k} {\gamma_O}^{l-k+o} g(s_l, a_l) | I_k \right] \\ &= (E_{\pi,O} [g(s_{k-o}, a_{k-o}) + \gamma_O g(s_{k-1}, a_{k-1}) \ldots \\ &\quad + {\gamma_O}^{o-1} g(s_{k-1}, a_{k-1}) | I_k]) \\ &\quad + {\gamma_O}^o E_{\pi,O} \left[\Sigma_{l \geq k} {\gamma_O}^{l-k} g(s_l, a_l) | I_k \right]. \end{aligned}$$

The first term is independent of π and, thus, it may be directly written as $r_O(I_k)$.

Thus, it suffices to find ${u_O}' = {u_O}'(\gamma, o, \gamma_O, O) > 0$ so that, for all I_k

$$E_{\pi,O} \left[\Sigma_{l \geq k} {\gamma_O}^{l-k} g(s_l, s_l) | I_k \right] = {u_O}'(\gamma, o, \gamma_O, O){V^\pi}_O(I_k). \tag{2}$$

To show (2), we first compute $E_{\pi,O}[g(s_{k-i}, a_{k-i}) | I_k]$, for $i = 1, \ldots, o$.

From the point of view of the decision maker at stage k, the cost $g(s_{k-o}, a_{k-o})$ will be collected when state s_{k-o+1} is observed first. If O is not *memoryless*, i.e., if it is not geometrically distributed; see also [20], the probability of this occurring on some stage after stage k depends on how many stages before k was state s_{k-o} observed first, i.e., it depends on r (if O is memoryless, it does not). We first consider the case $r = 0$. It will then become clear how those arguments may be applied for $r \neq 0$.

If $r = 0$, then state s_{k-o} was observed for the first time at stage k. Then, the cost $g(s_{k-o}, a_{k-o})$ will be collected with a delay of t stages with probability $P(O = t)$ and, at that stage it will be worth ${\gamma_O}^t g(s_{k-o}, a_{k-o})$. Thus, the expected worth of this cost is

$$\begin{aligned} E_{\pi,O} \left[g(s_{k-o}, a_{k-o}) \right] &= \Sigma_t P(O = t) {\gamma_O}^t g(s_{k-o}, a_{k-o}) \\ &= \left(\Sigma_t P(O = t) {\gamma_O}^t \right) g(s_{k-o}, a_{k-o}) \\ &= E_O \left[{\gamma_O}^O \right] g(s_{k-o}, a_{k-o}). \end{aligned}$$

Similarly, the cost $g(s_{k-o+1}, a_{k-o+1})$ will be collected with a delay of t stages with probability $P(O_1 + O_2 = t)$, where O_i, $i = 1, 2$ is the observation delay for state s_{k-o+i}.

If follows that the expected worth of this cost is

$$\begin{aligned} E_{\pi,O} \left[g(s_{k-o+1}, a_{k-o+1}) | I_k \right] &= \Sigma_t P(O_1 + O_2 = t) \\ &\quad \times {\gamma_O}^t g(s_{k-o+1}, a_{k-o+1}) \\ &= \left(\Sigma_t P(O_1 + O_2 = t) {\gamma_O}^t \right) \\ &\quad \times g(s_{k-o+1}, a_{k-o+1}) \\ &= \left(E_O \left[{\gamma_O}^{O1+O2} \right] \right) \\ &\quad \times g(s_{k-o+1}, a_{k-o+1}) \\ &= \left(E_O \left[{\gamma_O}^{O1} \right] \left[{\gamma_O}^{O2} \right] \right) \\ &\quad \times g(s_{k-o+1}, a_{k-o+1}) \\ &= \left(E_O \left[{\gamma_O}^{O1} \right] \right) \left(E_O \left[{\gamma_O}^{O2} \right] \right) \\ &\quad \times g(s_{k-o+1}, a_{k-o+1}) \\ &= \left(E_O \left[{\gamma_O}^{O} \right] \right)^2 \\ &\quad \times g(s_{k-o+1}, a_{k-o+1}). \end{aligned}$$

More generally, for $i = 1, \ldots, o$, we have that

$$E_{\pi,O} \left[g(s_{k-i}, a_{k-i}) | I_k \right] = \left(E_O \left[{\gamma_O}^O \right] \right)^{o+1+i} g(s_{k-i}, a_{k-i}).$$

Similarly, for $l \geq k$, we have that

$$E_{\pi,O} \left[g(s_l, a_l) | I_k \right] = \left(E_O \left[{\gamma_O}^O \right] \right)^{o+1+l-k} E_\pi \left[g(s_l, a_l) | I_k \right].$$

It follows that

$$\begin{aligned}
&E_{\pi,O}\left[\Sigma_{l\geq k}{\gamma_O}^{l-k}g(s_l,a_l)|I_k\right]\\
&\quad=\Sigma_{l\geq k}\left({\gamma_O}^{l-k}E_{\pi,O}\left[g(s_l,a_l)|I_k\right]\right)\\
&\quad=\Sigma_{l\geq k}\left({\gamma_O}^{l-k}\left(E_O\left[{\gamma_O}^{O}\right]\right)^{o+1+l-k}E_\pi\left[g(s_l,a_l)|I_k\right]\right)\\
&\quad=\left(E_O\left[{\gamma_O}^{O}\right]\right)^{o+1+l-k}E_\pi\left[\Sigma_{l\geq k}\,{\gamma_O}^{l-k}g(s_l,a_l)|I_k\right]\\
&\quad=\left(E_O\left[{\gamma_O}^{O}\right]\right)^{o+1}E_\pi\Big[\Sigma_{l\geq k}\left(E_O\left[{\gamma_O}^{O}\right]\right)^{l-k}\\
&\qquad\times{\gamma_O}^{l-k}g(s_l,a_l)|I_k\Big]\\
&\quad=\left(E_O\left[\frac{{\gamma_O}^{O+1}}{\gamma_O}\right]\right)^{o+1}E_\pi\Big[\Sigma_{l\geq k}\left(E_O\left[{\gamma_O}^{O}\right]\right)^{l-k}\\
&\qquad\times{\gamma_O}^{l-k}g(s_l,a_l)|I_k\Big]\\
&\quad=\left(\frac{\gamma^{o+1}}{{\gamma_O}^{o+1}}\right)E_\pi\Big[\Sigma_{l\geq k}\left(E_O\left[{\gamma_O}^{O}\right]\right)^{l-k}\\
&\qquad\times{\gamma_O}^{l-k}g(s_l,a_l)|I_k\Big]\\
&\quad=\left(\frac{\gamma^{o+1}}{{\gamma_O}^{o+1}}\right)E_\pi\left[\Sigma_{l\geq k}\left(E_O\left[{\gamma_O}^{O+1}\right]\right)^{l-k}g(s_l,a_l)|I_k\right]\\
&\quad=\left(\frac{\gamma^{o+1}}{{\gamma_O}^{o+1}}\right)E_\pi\left[\Sigma_{l\geq k}\gamma^{l-k}g(s_l,a_l)|I_k\right]\\
&\quad=\left(\frac{\gamma^{o+1}}{{\gamma_O}^{o+1}}\right){V^\pi}_O(I_k).
\end{aligned}$$

Thus, when $r=0$, (2) holds.

Similarly, when $r\neq 0$, it is easy to see that

$$\begin{aligned}
&E_{\pi,O}\left[\Sigma_{l\geq k}{\gamma_O}^{l-k}\,g(s_l,a_l)|I_k\right]\\
&\qquad=\left(\frac{E_O\left[{\gamma_O}^{O}|O\geq r\right]}{E_O[{\gamma_O}^{O}]}\right)\left(\frac{\gamma^{o+1}}{\gamma 0^{o+1}}\right){V^\pi}_O(I_k).
\end{aligned}$$

Thus, (2) holds again and this completes the proof. ■

For constant observation and cost delays ($O=C=0$), Lemma 2 implies $\gamma_O=\gamma$. AdditionallyI$_O=I_o, {P_A}^O={P_A}^o$, and $g_O=g_o$, and thus Lemma 2 generalizes Lemma 1.

Note that the arguments used in the above proof can be also used if the maximum allowable state vector dimension is reached. That is, the time period until the most recent system state is observed can be treated just like any other period in which the maximum allowable state vector dimension is not reached. This is so since although the underlying system does not make any new state transitions during that time period, the decision-maker does continue to observe the previous system state transitions. Thus, the cost structure g_O can be applied.

Note also that, in analogy to the proof of Lemma 1, the fact that $Pr(O\leq C)=1$ does not directly enter the proof but it is used to justify the definition of I_k used.

Note that, again as in Lemma 1, policies are evaluated properly in the sense that the ordering of total expected costs of policies is retained although total expected costs per se change. Briefly, note that a given policy does not have the same total costs in the two formulations that use g and g_O. However, (2) shows that for all policies total costs in the two formulations are connected through the same monotonic relationship and, thus, the ordering of policies is also the same across formulations.

Finally, from the proof of Lemma 2 it is also clear that Lemma 2 does not hold if the horizon is finite. This is so because, for finite horizons, not all costs would be collected.

Next, we show that it is always possible to apply Lemma 2. That is, we show that for fixed but arbitrary scalar γ with $0<\gamma<1$ and discrete random variable O, a γ_O exists such that $0<\gamma_O<1$ and $E_O[{\gamma_O}^{O+1}]=\gamma$.

This follows from Bolzano's theorem if we let $f(\gamma_O)=E_O[{\gamma_O}^{O+1}]-\gamma$, and notice that f is continuous and $f(0)=-\gamma<0$, and $f(1)=1-\gamma>0$.

For example, if O is geometrically distributed with parameter p, it is easy to see that the equation $E_O[{\gamma_O}^{O+1}]=\gamma$ reduces to the binomial equation $p\gamma_O^2+(1-p)\gamma\gamma_O-\gamma=0$.

Then, assuming that $0<\gamma<1$, it can be shown that $(1-p)^2\gamma^2+4\gamma p>0$, and that the solution $\gamma_O=[-(1-p)\gamma+\sqrt{(1-p)^2\gamma^2+4\gamma p}]/(2p)$ lies between 0 and 1.

Now, we consider the case with $AC\neq 0$. This will allow deriving our second result.

B. Action Delays

We assume that successive action delays are independent random variables, following a common probability distribution $P(AC=a), a=0,1,\ldots,a_{max}$, and are also independent of state identity.

Naturally, a state cannot be observed if the action that is applied at the preceding state has not taken effect. Thus, we also assume that *given* an action has taken effect, the resulting state may be observed with a delay following a probability distribution $P(O=o), o=1,2,\ldots,o_{max}$. Formally, the observation delay distribution for state s_{k+1}, for all k, is conditioned on the event that action a_k has taken effect, but for simplicity we do not state this. Finally, all action delays are assumed to be independent of all observation delays.

With these assumptions, it is easy to see that, as for constant delays, the effects of observation and action delays on the structure of the equivalent MDP without delays are additive. That is, the random variable $O+AC$ now plays the role that O played in the case of no action delays.

In other words, by repeating the arguments used in the proofs of Result 1 and Lemma 2, the following holds.

Result 2: The SDMDP $\langle S,A,P_A,g,O,AC,C,\gamma\rangle$, with $Pr(O\leq C)=1$ is reducible to the MDP $\langle \mathrm{I}_{O+AC},A,{P_A}^{O+AC},g_{O+AC}\rangle$, with discounting factor $\gamma_{O+AC}:E_{O+AC}[{\gamma_O}^{O+AC+1}]=\gamma$.

Note that Result 2 also generalizes Result 1 for constant delays.

Note again that the above result also holds if the maximum allowable state dimension is reached. And note that, as in Lemma 2, the fact that $Pr(O\leq C)=1$ does not directly enter the proof but it is used to justify the definition of I_k used. Note also that, again as in Lemma 2, policies are evaluated properly in the sense that the ordering of total expected costs of policies is retained although total expected costs *per se* change. Again, Result 2 also holds only for an infinite horizon.

Result 2, as Result 1, shows that a partially observable MDP that arises when the information available for action selection is incomplete due to random observation and action delays, is completely observable with respect to an appropriately chosen past decision stage. For random delays, complete observability is possible if the decision maker concentrates on an MDP *embedded* in the original SDMDP. This embedded MDP consists of the stages where actions take effect and system states are observed. If asynchronous cost collection is applied to this embedded MDP, the state-transition and cost structures of the original MDP may be used.

IV. Summary

We considered how a (discrete-time, total-expected-cost, infinite-horizon) Markov decision process with observation, action, and cost delays is reduced to a Markov decision process without delays. First, we drew connections among the three delay types. Then, previous work on constant delays was extended: it was shown that the cost structure of the process without delays is the same as that of the original process with delays.

The topic of random delays was introduced and it was shown that, by considering an embedded process that behaves similar to a process with constant delays, again the cost structure of the process without delays is the same with that of the original process with random delays. However, a different discounting factor needs to be computed.

Both results are based on the intuition of asynchronous cost collection. That is, costs may be induced and collected at different decision stages and policies can still be compared properly as long as costs are discounted properly. The derivations are better understood through a concept often used in neuro-dynamic programming.

Practically, this work may be used to solve MDPs with delays more efficiently. That is, the increase in computational effort with delays is due only to state space augmentation. Theoretically, this work may be used as preparation to considering a new dimension of asynchrony in MDPs beyond the one related to dynamic programming [21]. Finally, it would be interesting and useful to consider continuous-time and/or average-cost MDPs.

Acknowledgment

The authors would like to thank A. Barto, D. Bernstein, O. Simsek, J. McGregor-Smith, R. Sutton, and L. Zelevinsky for helpful discussions.

References

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

[2] M. L. Putterman, *Markov Decision Processes: Discrete Stochastic Dynamic Programming*. Chichester, U.K.: Wiley, 1994.

[3] D. J. White, “A survey of applications of Markov decision processes,” *J Opl. Res.*, vol. 44, pp. 1073–1096, 1993.

[4] E. Altman and P. Nain, “Closed-loop control with delayed information,” *Perf. Eval. Rev.*, vol. 14, pp. 193–204, 1992.

[5] L. Zelevinsky, “Does time-optimal control of a stochastic system with sensory delay produce movement units?,” M.S. thesis, Dept. Comput. Sci., Univ. Mass., Amherst, MA, 1998.

[6] J. L. Bander and C. C. White III, “Markov decision processes with noise-corrupted and delayed state observations,” *J. Opl. Res. Soc.*, vol. 50, pp. 660–668, 1999.

[7] D. Brooks and C. Leondes, “Markov decision processes with state-information lag,” *Opns. Res.*, vol. 20, pp. 904–907, 1972.

[8] S. H. Kim, “State information lag Markov decision process with control limit rule,” *Naval Res. Log. Q.*, vol. 32, pp. 491–196, 1985.

[9] S. H. Kim and B. H. Jeong, “A partially observable Markov decision process with lagged information,” *J. Opl. Res. Soc.*, vol. 38, pp. 439–446, 1987.

[10] C. C. White III, ““Note on” a partially observable Markov decision process with lagged information,” *J. Opl. Res. Soc.*, vol. 39, pp. 217–218, 1988.

[11] E. J. Sondik, “The optimal control of partially observable Markov processes over the infinite horizon: Discounted cost,” *Opns. Res.*, vol. 26, pp. 282–304, 1978.

[12] D. P. Bertsekas, *Dynamic Programming: Deterministic and Stochastic Models*. Upper Saddle River, NJ: Prentice-Hall, 1987.

[13] E. Altman, T. Basar, and R. Srikant, “Congestion control as a stochastic control problem with action delays,” *Automatica*, vol. 12, pp. 1937–1950, 1999.

[14] P. Varaiya and J. Walrand, “On delayed sharing patterns,” *IEEE Trans. Automat. Contr.*, vol. AC-23, pp. 443–445, 1978.

[15] K. Hsi and S. I. Marcus, “Decentralized control of finite state Markov processes,” *IEEE Trans. Automat. Contr.*, vol. AC-27, pp. 426–431, 1982.

[16] D. J. Bertsekas and J. N. Tsitsiklis, *Neuro-Dynamic Programming*. Belmont, MA: Athena, 1996.

[17] R. Sutton and A. Barto, *Reinforcement Learning: An Introduction*. Cambridge, MA: MIT Press, 1998.

[18] C. J. C. H. Watkins and P. Dayan, “Q-learning,” *Mach. Learn.*, vol. 8, pp. 279–292, 1992.

[19] S. E. Engelbrecht and K. V. Katsikopoulos, “Planning with delayed state information,” Dept. Comput. Sci., Univ. Mass., Amherst, MA, Tech. Rep. 99-30, 1999.

[20] S. M. Ross, *Stochastic Processes*. New York:]Wiley, 1996.

[21] D. J. Bertsekas, “Distributed dynamic programming,” *IEEE Trans. Automat. Contr.*, vol. AC-27, pp. 610–616, 1982.

Konstantinos V. Katsikopoulos (M’00) was born in Athens, Greece. He received the Diploma and Postgraduate degrees from the University of Athens, and the Ph.D. degree from the University of Massachusetts, Amherst, in 1992, 1994, and 1999, respectively.

He has been a Visiting Faculty Member at the University of Massachusetts and at the Naval Postgraduate School, Monterey, California. Since August 2002, he has been a Postdoctoral Research Fellow with the Max Planck Institute for Human Development, Berlin, Germany. His research interests include descriptive models of human performance, especially decision-making, and their relation to normative models.

Dr. Katsikopoulos is an Associate Editor for IEEE Transactions on Systems, Man, and Cybernetics: Systems and Humans.

Sascha E. Engelbrecht was born in Wolfhagen, Germany, in 1969. He received the Diplom. degree from Philipps University, Marburg, Germany, and the M.Sc. and Ph.D. degrees from the University of Massachusetts, Amherst.

From November 1996 to January 2000, he was a Postdoctoral Research Associate at the University of Massachusetts Adaptive Networks Laboratory. Since February 2000, he has been with Dresdner Bank, Frankfurt, Germany, where he manages the groupwide development of treasury risk systems. His research interests include machine learning, biocybernetics, and decision making under uncertainty.