Proceedings of the
47th IEEE Conference on Decision and Control
Cancun, Mexico, Dec. 9-11, 2008

Information State for Markov Decision Processes with Network Delays

Sachin Adlakha[1] Sanjay Lall[2] Andrea Goldsmith[1]

Abstract

We consider a networked control system, where each subsystem evolves as a Markov decision process (MDP). Each subsystem is coupled to its neighbors via communication links over which the signals are delayed, but are otherwise transmitted noise-free. A controller receives delayed state information from each subsystem. Such a networked Markov decision process with delays can be represented as a partially observed Markov decision process (POMDP). We show that this POMDP has a sufficient information state that depends only on a finite history of measurements and control actions. Thus, the POMDP can be converted into an information state MDP, whose state does not grow with time. The optimal controller for networked Markov decision processes can thus be computed using dynamic programming over a finite state space. This result generalizes the previous results on Markov decision processes with delayed state information.

1 Introduction and Prior Work

We consider a network of interconnected subsystems, where each subsystem evolves as a Markov decision process (MDP). Each subsystem has a finite state space and its state evolution is affected by delayed state of its neighbors. A centralized controller receives delayed state measurements from each subsystem. We refer to such systems as *networked Markov decision processes*.

Although the controller receives state information from each subsystem, each of these states is delayed by different amounts. Since the current state of each subsystem is not available to the controller, this system can be represented as a partially observed Markov decision process (POMDP). A standard approach for solving POMDPs involves generating a policy in which the control action at any time depends explicitly on the complete history of observations. This history is called the *information state* and it grows without bound as time increases. In this paper, we show that for a certain class of POMDPs, a sufficient information state has finite memory, i.e., it depends only on a finite number of past states and actions.

The optimal control design for POMDPs has been studied extensively in literature [3, 7, 8]. It has been shown that the optimal controller has a separation structure and is a function of the posterior distribution of the current state given all the past observations. The control of a single MDP with delayed state information was considered in [2]. It was shown that the optimal control action depends upon the last observed state and a finite number of previous actions. In [6], the authors consider a single MDP with observation delays, action delays as well cost delays. They also extend the result to the case of random delays.

Among the earliest works in a networked system with delays is [9], where a separation structure for the *one-step delay sharing pattern* for a system with general nonlinear dynamics was obtained. Algorithms to compute the optimal controller for such a system were obtained in [5] by essentially reducing the problem to a centralized control problem. An optimal controller is then synthesized using standard algorithms. More general decentralized control of MDPs has been shown to be intractable in [4].

A general networked system with arbitrary delay pattern was considered in [1]. It was shown that a centralized optimal controller for such systems need only store the past few states of each subsystem. In this paper, we generalize the result given in [1] by considering a system where control inputs are applied to all subsystems. Furthermore, we show that for networked Markov decision processes with delays, the sufficient information state only depends on a finite history. Thus, the result of [1] is a special case of the result presented in this paper. Since the sufficient information state does not grow with time, the computational difficulties associated with computing the optimal policy can be significantly reduced for such problems. Moreover, the amount of history required to compute the optimal policy depends only on the underlying network graph structure and the delays in the network.

Notation In the remainder of the paper, we use the following notation. We use superscripts to denote particular subsystems and subscripts for the time index. Thus

[1]S. Adlakha and A. Goldsmith are with the Department of Electrical Engineering, Stanford University, 350 Serra Mall, Stanford, CA, USA. adlakha@stanford.edu, andrea@wsl.stanford.edu

[2]S. Lall is with the Department of Aeronautics and Astronautics, Stanford University, Stanford, CA 94305, USA. lall@stanford.edu

978-1-4244-3124-3/08/$25.00 ©2008 IEEE

x_t^1 denotes the state of the subsystem 1 at time t. For simplicity, we ignore the superscript 1 if there is only one subsystem. Similarly, we denote y_t^i to be the observation received from subsystem i at time t and u_t^i to be the control input applied to subsystem i at time t. We also denote z, s and a to be a realization of the state x, observation y and control action u. We define $x_{0:t}^i := \left(x_0^i, \ldots, x_t^i\right)$ to refer to the list of variables corresponding to the subsystem i from time 0 to t. If $t < 0$, we interpret the list as empty. The notation $x_{0:t} = z_{0:t}$ is interpreted as element wise equality, *i.e.*, $x_0 = z_0, x_1 = z_1$, etc. To denote the list of variables corresponding to all subsystems, we define $x_t := \left(x_t^1, \ldots, x_t^n\right)$. Similarly, we denote $u_t := \left(u_t^1, \ldots, u_t^n\right)$ as the control action applied to all subsystems at time t. We define $A_{0\ldots t}^i$ to be the product of the variables corresponding to times $0, \ldots, t$, that is $A_{0\ldots t}^i := A_0^i A_1^i \ldots A_t^i$. For a set $\mathcal{X}$, we denote $\mathcal{X}^n$ to be the n-fold cartesian product of the set, that is $\mathcal{X}^n = \mathcal{X} \times \cdots \times \mathcal{X}$ n-times, with the interpretation that $\mathcal{X}^0 = \phi$. We write $\mathbb{Z}^+$ for the set of non-negative integers.

2 Model and Definitions

2.1 Markov Decision Processes

A Markov decision process provides a framework for sequential decision making in a stochastic environment. The decision (also known as the action) taken at time t affects the evolution of the future system. The goal of the decision maker is to choose a sequence of actions to optimize a predetermined criterion. For the purposes of this paper, we assume that the decisions are made at discrete times $t \in \mathbb{Z}^+$.

At each decision time t, the system occupies a *state*. We denote the set of all possible states by a finite set $\mathcal{X}$. At each time t, the decision maker choses a decision from a finite set denoted by $\mathcal{U}$. Formally,

Definition 1 (MDP)**.** *A Markov decision process is a tuple (A, g) where,*

1. *A is a sequence $A_0, A_1 \ldots$ with $A_0 : \mathcal{X} \to [0,1]$, such that $A_0(z) \geq 0$ for all $z \in \mathcal{X}$ and $\sum_z A_0(z) = 1$.*
   For $t \geq 1$, we have $A_t : \mathcal{X} \times \mathcal{X} \times \mathcal{U} \to [0,1]$, such that
$$A_t(z_1, z_2, a) \geq 0, \quad \forall\ z_1, z_2 \in \mathcal{X} \text{ and } a \in \mathcal{U},$$
$$\sum_{z_1} A_t(z_1, z_2, a) = 1, \quad \forall\ z_2 \in \mathcal{X} \text{ and } a \in \mathcal{U}.$$
2. *g is a sequence $g_0, g_1, \ldots$ with $g_t : \mathcal{X} \times \mathcal{U} \to \mathbb{R}$.*

As an example of an MDP, consider a discrete time dynamic system, where the state of the system at time $t \geq 0$ is denoted by x_t. The system dynamics are

$$x_{t+1} = f\left(x_t, u_t, w_t\right). \tag{1}$$

Here u_t is the control action or the decision taken at time t. The random variables w_t for $t \geq 0$ are independent noise processes. The initial state x_0 is chosen to be independent of the noise process w_t. Associated with this dynamic system is an MDP (A, g) defined as follows. For all $p \in \mathcal{X}$, let $A_0(p) = \text{Prob}(x_0 = p)$ be the probability mass function of the initial state of the system. For $t > 0$, let

$$A_t\left(z_t, z_{t-1}, a_{t-1}\right) = \text{Prob}\Big(x_t = z_t \mid x_{t-1} = z_{t-1}, u_{t-1} = a_{t-1}\Big), \tag{2}$$

be the conditional probability of state x_t given the previous state x_{t-1} and the applied input u_{t-1}. It is easy to verify that the sequence A satisfies all the properties as given in definition 1. The sequence $g_t\left(x_t, u_t\right)$ represents the cost at time t and it depends on the current state x_t of the system as well as the action u_t taken at time t.

As mentioned before, the decision maker (*i.e.*, the controller) needs to choose an action u_t at time t. This is chosen based upon the information available to the controller at that time. We define h_t^{mdp} to be the information available to the controller at time t, given by

$$h_t^{\text{mdp}} = \left(u_{0:t-1}, x_{0:t}\right).$$

We will also use i_t^{mdp} to denote a realization of h_t^{mdp} as

$$i_t^{\text{mdp}} = \left(a_{0:t-1}, z_{0:t}\right).$$

Here the sequences z and a specify the values of a realization of x and u, respectively. An MDP policy (also known as the control policy) specifies the decision or control action to be taken at each time t.

Definition 2 (MDP Policy)**.** *An MDP policy is a sequence $K = (K_0, K_1, \ldots)$ where $K_t : \mathcal{U} \times \mathcal{X}^{t+1} \times \mathcal{U}^t \to [0,1]$ for all $t \in \mathbb{Z}^+$ such that*

$$K_t(a, z, \tilde{a}) \geq 0, \quad \forall\ a \in \mathcal{U},\ z \in \mathcal{X}^{t+1} \text{ and } \tilde{a} \in \mathcal{U}^t,$$
$$\sum_a K_t(a, z, \tilde{a}) = 1, \quad \forall\ z \in \mathcal{X}^{t+1} \text{ and } \tilde{a} \in \mathcal{U}^t.$$

For the discrete time dynamic system given in equation (1), we can interpret the MDP policy as

$$K_t(a_t, i_t) = \text{Prob}(u_t = a_t \mid h_t^{\text{mdp}} = i_t).$$

The MDP policies as described above are called as *mixed policies* since the decision at time t is specified by a probability distribution which is a function of the information available to the controller.

Stochastic Process Generated by MDP Consider an MDP (A, g) and an MDP policy K. Associated with

(A, g) and K, is a stochastic process that is induced by the MDP and its policy. For MDPs evolving over a finite time horizon T, we can define the sample space of the stochastic process as

$$\Omega = \mathcal{X} \times \mathcal{U} \times \mathcal{X} \ldots \mathcal{U} \times \mathcal{X} = \{\mathcal{X} \times \mathcal{U}\}^{T-1} \times \mathcal{X}.$$

A typical element $\omega \in \Omega$ is given by a sequence of states and actions. For example, for infinite horizon model, a typical sample path would be given as

$$\omega = \{z_0, a_0, z_1, a_1 \ldots\}.$$

Definition 3 (MDP Stochastic Process)**.** *Suppose (A, g) is an MDP and K is an MDP policy. Define the state process $x_t(\omega)$ and the action process $u_t(\omega)$ by*

$$\text{Prob}\left(x_{0:t} = z_{0:t}, u_{0:t} = a_{0:t}\right) = A_0(z_0) \times \prod_{k=1}^{t} A_k\left(z_k, z_{k-1}, a_{k-1}\right) \times \prod_{k=0}^{t} K_k\left(a_k, z_{0:k}, a_{0:k-1}\right). \quad (3)$$

Note that this implies that for all t we have

$$\begin{aligned} \text{Prob}\left(x_t |\ x_{0:t-1}, u_{0:t-1}\right) &= \text{Prob}\left(x_t |\ x_{t-1}, u_{t-1}\right), \\ \text{Prob}\left(x_t |\ x_{t-1}, u_{t-1}\right) &= A_t\left(x_t, x_{t-1}, u_{t-1}\right), \\ \text{Prob}\left(u_t |\ x_{0:t}, u_{0:t-1}\right) &= K_t\left(u_t, x_{0:t}, u_{0:t-1}\right). \end{aligned}$$

The above equations show that the state X_t is conditionally independent of the past states and actions given the current state X_{t-1} and the current action U_{t-1}. The state evolution is thus Markov justifying the name.

As mentioned before, the goal of the Markov decision process formulation is to make sequential decisions in a stochastic environment. The controller's objective is to choose an MDP policy K so as to minimize a cost function. Typically, the cost function has the form

$$J_K\left(A, g\right) \triangleq \mathbb{E}\left(\sum_{t=0}^{T} g_t\left(x_t, u_t\right)\right).$$

Here, the expectation is taken over the noise processes and is with respect to the probability measure defined in equation (3). The notation $J_K\left(A, g\right)$ represents the cost of an MDP (A, g) under an MDP policy K. In this sense, the sequence g represents the cost function or the objective that the decision maker wishes to minimize.

2.2 Partially Observed Markov Decision Processes

A POMDP is an extension of an MDP, where the state of the system is not fully observable. Thus, the decision maker needs to make the decision with only *partial* knowledge of the state of the system. The set of all possible observations as seen by the decision maker is denoted by a finite set $\mathcal{Y}$.

Definition 4 (POMDP)**.** *A partially observed Markov decision process is a tuple (A, C, g) where,*

1. *(A, g) is a Markov decision process.*
2. *C is a sequence $C_0, C_1 \ldots$ with $C_t : \mathcal{Y} \times \mathcal{X} \to [0, 1]$, such that*

$$\begin{aligned} C_t\left(s, z\right) &\geq 0, \quad \forall\ s \in \mathcal{Y}, \forall\ z \in \mathcal{X}, \\ \sum_{s} C_t\left(s, z\right) &= 1, \quad \forall\ z \in \mathcal{X}. \end{aligned}$$

Akin to MDPs, the decision in the POMDPs is made based on the information available to the decision maker. We define h_t^{pomdp} to be the information available to the controller at time t, given by

$$h_t^{\text{pomdp}} = \left(u_{0:t-1}, y_{0:t}\right).$$

Also, we use i_t^{pomdp} to denote a realization of h_t^{pomdp} as

$$i_t^{\text{pomdp}} = \left(a_{0:t-1}, s_{0:t}\right).$$

Definition 5 (POMDP Policy)**.** *A POMDP policy is a sequence $K = (K_0, K_1, \ldots)$ where $K_t : \mathcal{U} \times \mathcal{Y}^{t+1} \times \mathcal{U}^t \to [0, 1]$ for all $t \in \mathbb{N}$ such that*

$$\begin{aligned} K_t(a, \tilde{s}, \tilde{a}) &\geq 0, \quad \forall\ a \in \mathcal{U},\ \tilde{s} \in \mathcal{Y}^{t+1} \text{ and } \tilde{a} \in \mathcal{U}^t, \\ \sum_{a} K_t(a, \tilde{s}, \tilde{a}) &= 1, \quad \forall\ \tilde{s} \in \mathcal{Y}^{t+1} \text{ and } \tilde{a} \in \mathcal{U}^t. \end{aligned}$$

For partially observed discrete time dynamic process, the POMDP policy gives the probability distribution over possible actions or controls as a function of the information available to the decision maker. That is

$$K_t(a_t, i_t) = \text{Prob}(u_t = a_t \mid h_t^{\text{pomdp}} = i_t).$$

Stochastic Process Generated by POMDP Consider a POMDP (A, C, g) and a POMDP policy K. Associated with every (A, C, g) and K is a stochastic process that is induced by it. For a POMDP evolving over a finite horizon T, we can define the sample space of the stochastic process as

$$\begin{aligned} \Omega_{pomdp} &= \mathcal{X} \times \mathcal{Y} \times \mathcal{U} \times \mathcal{X} \ldots \mathcal{Y} \times \mathcal{U} \times \mathcal{X} \\ &= \{\mathcal{X} \times \mathcal{Y} \times \mathcal{U}\}^{T-1} \times \mathcal{X} \end{aligned}$$

A typical sample path for the infinite horizon POMDP would be given as

$$\omega = \{z_0, s_0, a_0, z_1, s_1, a_1 \ldots\}.$$

Definition 6 (POMDP Stochastic Process)**.** *Consider a POMDP (A, C, g) along with a POMDP policy K. Define the state process $x_t(\omega)$, the observation process $y_t(\omega)$ and the action process $u_t(\omega)$ by*

$$\text{Prob}\left(x_{0:t} = z_{0:t}, y_{0:t} = s_{0:t}, u_{0:t} = a_{0:t}\right) = A_{0:t} C_{0:t} K_{0:t}. \quad (4)$$

Here we have suppressed the arguments for notational compactness. Note that this implies that for all t we have

$$\begin{aligned}\text{Prob}\left(x_t \mid x_{0:t-1}, u_{0:t-1}\right) &= \text{Prob}\left(x_t \mid x_{t-1}, u_{t-1}\right),\\ \text{Prob}\left(x_t \mid x_{t-1}, u_{t-1}\right) &= A_t\left(x_t, x_{t-1}, u_{t-1}\right),\\ \text{Prob}\left(y_t \mid x_t\right) &= C_t\left(y_t, x_t\right),\\ \text{Prob}\left(u_t \mid y_{0:t}, u_{0:t-1}\right) &= K_t\left(u_t, y_{0:t}, u_{0:t-1}\right).\end{aligned}$$

Similar to MDPs, the state evolution process X_t and the observation process Y_t are Markov. The POMDP policy only depends on the observation vector y and not on the actual state vector x, justifying the *partially observed* part of the name.

Similar to MDPs, we can define a random variable $f : \Omega \to \mathbb{R}$ on the sample space of the stochastic process induced by a POMDP. The cost function for POMDPs is given as

$$J_K\left(A, C, g\right) = \mathbb{E}\left(\sum_{t=0}^{T} g_t\left(x_t, u_t\right)\right).$$

where the expectation is taken with respect to the marginal probability measure derived from equation (4). The objective of a decision maker is find a POMDP policy which minimizes the expected cost.

2.3 Information State for POMDPs

An information state for a POMDP represents all the information about the history of POMDP that is relevant to the selection of the optimal control. The POMDP can be reformulated as an MDP using the information state. For a POMDP, the information state consists of either a complete history of observations and actions or their corresponding sufficient statistics [7].

Definition 7. *Suppose (A, C, g) is a POMDP and define a sequence of functions*

$$\gamma_t : \mathcal{U}^t \times \mathcal{Y}^{t+1} \to \mathcal{Q}.$$

Let $\xi_t = \gamma_t\left(u_{0:t-1}, y_{0:t}\right)$. Then ξ_t is called a sufficient information state for the POMDP if there exists an MDP $(\tilde{A}, \tilde{g})$ over the state space $\mathcal{Q}$ and action space $\mathcal{U}$ such that, for all POMDP policies K, we have

1. *$\tilde{A}$ is a sequence such that*

$$\begin{aligned}&\tilde{A}_{t+1}\left(q_{t+1}, q_t, a_t\right) =\\ &\quad \text{Prob}\left(\xi_{t+1} = q_{t+1} \mid \xi_{0:t} = q_{0:t}, u_{0:t} = a_{0:t}\right). \quad (5)\end{aligned}$$

2. *$\tilde{g}$ is a sequence $\tilde{g}_0,\ \tilde{g}_1 \ldots$ such that*

$$\tilde{g}_t\left(q_t, a_t\right) = \mathbb{E}\left(g_t\left(x_t, a_t\right) \mid \xi_t = q_t, u_t = a_t\right). \quad (6)$$

3. *For all $t \geq 0$, we have*

$$\begin{aligned}&\text{Prob}\Big(x_t = z_t \mid \xi_t = \gamma_t\left(s_{0:t}, a_{0:t-1}\right), \ldots,\\ &\qquad \xi_0 = \gamma_0(s_0), u_{0:t-1} = a_{0:t-1}\Big) =\\ &\quad \text{Prob}\left(x_t = z_t \mid y_{0:t} = s_{0:t}, u_{0:t-1} = a_{0:t-1}\right). \quad (7)\end{aligned}$$

Note that $\tilde{A}$ in equation (5), $\tilde{g}_t$ in equation (6) and the conditional probability in equation (7) are independent of the POMDP policy K. Furthermore, equation (5) shows that the evolution of ξ_t is Markov.

3 Networked Markov Decision Processes

A networked Markov decision process (N-MDP) is a weighted directed graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$, where $\mathcal{V} = \{1, \ldots, n\}$ is a finite set of vertices and $\mathcal{E} \subset \mathcal{V} \times \mathcal{V}$ is a set of edges. Each vertex $i \in \mathcal{V}$ represents a Markov decision process. An edge $(i, j) \in \mathcal{E}$ if the MDP at vertex i directly affects the MDP at vertex j. Associated with each edge $(i, j) \in \mathcal{E}$ is a nonnegative integer weight, M_{ij}, which specifies the delay for the dynamics of vertex i to propagate to vertex j. We assume that $(i, i) \notin \mathcal{E}$.

Associated with each $j \in \mathcal{V}$, let $\mathcal{I}^j$ be the set of all vertices with an incoming edge to vertex j, specifically

$$\mathcal{I}^j = \{\, i \in \mathcal{V} \mid (i, j) \in \mathcal{E} \,\}.$$

Similarly, for each $j \in \mathcal{V}$, let $\mathcal{O}^j$ be the set of all vertices connected to by an edge outgoing from vertex j, specifically

$$\mathcal{O}^j = \{\, i \in \mathcal{V} \mid (j, i) \in \mathcal{E} \,\}.$$

At each time t, the state of the MDP at vertex i belongs to a finite set $\mathcal{X}^i$. The decision or the control action taken at vertex i is drawn out of a finite set $\mathcal{U}^i$.

Remark. In the remainder of the paper, we denote $\mathcal{X}^{-i} = \prod_{j \in \mathcal{I}^i} \mathcal{X}^j$. Also denote $\mathcal{X}^{(n)} = \prod_{i=1}^{n} \mathcal{X}^i$ as the cartesian product of state space corresponding to all vertices. Similarly, let $\mathcal{U}^{(n)} = \prod_{i=1}^{n} \mathcal{U}^i$.

Definition 8. *A networked Markov decision process is a tuple (A, g) where*

1. *A is a set of transition matrices $\{A_t^i,\ t \geq 0 \mid i \in \mathcal{V}\}$ with $A_0^i : \mathcal{X}^i \to [0, 1]$ for all $i \in \mathcal{V}$, such that for all $z \in \mathcal{X}^i$, we have*

$$A_0^i\left(z\right) \geq 0 \text{ and } \sum_z A_0^i\left(z\right) = 1.$$

For $t \geq 0$, we have $A_t : \mathcal{X}^i \times \mathcal{X}^i \times \mathcal{X}^{-i} \times \mathcal{U}^i \to [0, 1]$ such that, for all $i \in \mathcal{V}$ and for all $a \in \mathcal{U}^i$ and $\tilde{z} \in \mathcal{X}^{-i}$ we have

$$\begin{aligned}A_t^i(z_1, z_2, \tilde{z}, a) &\geq 0, \quad \forall\ z_1, z_2 \in \mathcal{X}^i,\\ \sum_{z_1} A_t^i(z_1, z_2, \tilde{z}, a) &= 1, \quad \forall\ z_2 \in \mathcal{X}.\end{aligned}$$

2. g is a sequence $g_0, g_1, \dots$ with $g_t : \mathcal{X}^{(n)} \times \mathcal{U}^{(n)} \to [0,1]$.

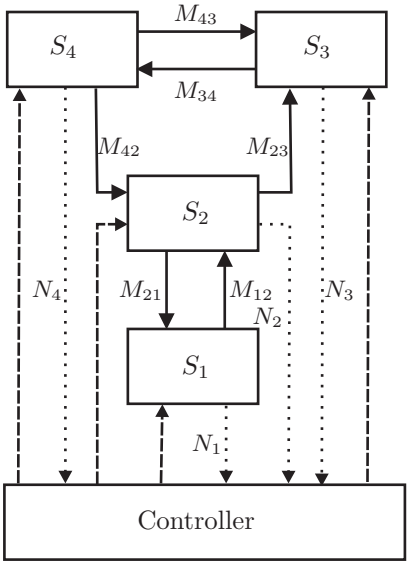

Figure 1: A network of interconnected subsystems with delays. Subsystem i is denoted by S_i, the network propagation delay from S_i to S_j is denoted by M_{ij} and the measurement delay from S_i to the controller is denoted by N_i.

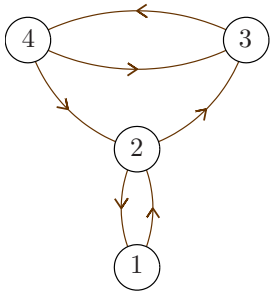

Figure 2: Directed graph for the network of Figure 1.

As an example of a networked Markov decision process, consider a networked system consisting of four subsystems as shown in Figure 1. The corresponding directed graph is shown in Figure 2. The system dynamics are

$$x_{t+1}^i = f^i\big(x_t^i, \{x_{t-M_{ji}}^j \mid j \in \mathcal{I}^i\}, u_t^i, w_t^i\big), \qquad (8)$$

for all $i \in \mathcal{V}$. Here $u_t^i \in \mathcal{U}^i$ is the control action applied to subsystem i at time t. The random variables x_0^i, w_t^i for $t \geq 0$ and $i \in \mathcal{V}$ are independent, *i.e.*, the noise processes are independent across both time and subsystems.

Associated with this system is a networked MDP (A, g) as defined below. For $p \in \mathcal{X}^i$, let $A_0^i(p) = \text{Prob}(x_0^i = p)$ define the probability mass functions of the initial states of subsystem $i \in \mathcal{V}$. The initial states $x_0^1, \dots, x_0^n$ are chosen independently. For $t > 0$, let

$$A_t^i(z,p,q,a) = \text{Prob}\Big(x_t^i = z \mid x_{t-1}^i = p, \{x_{t-1-M_{ji}}^j = q^j \mid j \in \mathcal{I}^i\}, u_{t-1}^i = a\Big),$$

be the conditional probability mass function of state x_t^i given the previous states x_{t-1}^i and $\{x_{t-1-M_{ji}}^j \mid j \in \mathcal{I}^i\}$ and the applied input u_{t-1}^i. It is easy to verify that the sequence A satisfies the properties in Definition 8. The sequence $g_t(x_t, u_t)$ represents the cost at time t and depends on the state of the system $x_t = (x_t^1, \dots, x_t^n)$ as well as the action $u_t = (u_t^1, \dots, u_t^n)$ applied at time t.

In a networked MDP, the controller needs to choose a control action corresponding to each vertex $i \in \mathcal{V}$. The actions are chosen based on the information available to the controller at time t. Associated with each vertex $i \in \mathcal{V}$ of a networked MDP, we have a nonnegative integer N_i which specifies the delay in receiving the state measurement from system i. We define $h_t^{\text{n-mdp}}$ to be the information available to the decision maker at time t, given by

$$h_t^{\text{n-mdp}} = \big(x_{0:t-N_1}^1, u_{0:t-1}^1, \dots, x_{0:t-N_n}^n, u_{0:t-1}^n\big).$$

Also define $i_t^{\text{n-mdp}}$ to be a realization of $h_t^{\text{n-mdp}}$ as

$$i_t^{\text{n-mdp}} = \big(z_{0:t-N_1}^1, a_{0:t-1}^1, \dots, z_{0:t-N_n}^n, a_{0:t-1}^n\big).$$

Thus, the observations received by the decision maker at time t consist of the state of the subsystem i delayed by N_i time steps. A networked MDP policy specifies the decisions taken at time t.

Definition 9 (Networked-MDP Policy)**.** *A networked MDP policy is a sequence $K = (K_0, K_1, \dots)$ where*

$$K_t : \mathcal{U}^{(n)} \times \prod_{i=1}^n \left(\mathcal{X}^i\right)^{t+1-N_i} \times \prod_{i=1}^n \left(\mathcal{U}^i\right)^t \to [0,1],$$

for all $t \in \mathbb{Z}^+$ such that for all $\tilde{z} \in \prod_{i=1}^n \left(\mathcal{X}^i\right)^{t+1-N_i}$ and $\tilde{a} \in \prod_{i=1}^n \left(\mathcal{U}^i\right)^t$ we have

$$K_t\left(a, \tilde{z}, \tilde{a}\right) \geq 0 \ \forall a \in \mathcal{U}^{(n)},$$
$$\sum_a K_t\left(a, \tilde{z}, \tilde{a}\right) = 1.$$

For the networked systems as given in equation (8), a general mixed control policy is defined as a sequence of transition matrices K_t $t \geq 0$ given by

$$K_t(a_t, i_t) = \text{Prob}(u_t = a_t \mid h_t^{\text{n-mdp}} = i_t).$$

3.1 Networked MDP as a POMDP

In networked MDPs, although the controller receives state information from the subsystems, these states are delayed by different amounts. Thus, a networked MDP can be written as a POMDP. Consider a networked MDP as given in Definition 8. Let us define a new state $\hat{x}_t = \left\{x^i_{0:t} \mid i \in \mathcal{V}\right\}$. The state $\hat{x}$ is chosen such that in the resulting system the observation at time t is only a function of the current state at time t. It is easy to check that there exists a function $\hat{f}$ such that

$$\hat{x}_{t+1} = \hat{f}\left(\hat{x}_t, u_t, w_t\right).$$

Associated with this function is a transition probability mass function $\hat{A}_t\left(\hat{z}_{t+1}, \hat{z}_t, a_t\right)$, where $\hat{z}_t$ is the realization of the state $\hat{x}_t$. The observation at any time t is given as

$$\hat{y}_t = \hat{h}(\hat{x}_t).$$

Corresponding to this observation process is a probability mass function $\hat{C}_t\left(\hat{s}_t, \hat{z}_t\right)$, where $\hat{s}_t$ is the realization of the observation $\hat{y}_t$ and is given as

$$\hat{s}_t = \left\{z^i_{t-N_i} \mid i \in \mathcal{V}\right\}$$

The cost function is given as

$$\hat{g}_t\left(\hat{x}_t, a_t\right) = g_t\left(x_t, a_t\right) \tag{9}$$

It is easy to check that the functions $\hat{A}_t$, $\hat{C}_t$ and $\hat{g}_t$ satisfy the properties given in the definition 4. The networked MDP can thus be written as a POMDP $\left(\hat{A}, \hat{C}, \hat{g}\right)$.

4 Information State for Networked Markov Decision Processes

Before we present the main result of the paper, we make the following definitions.

Definition 10. *Let*

$$d_i = \max\{N_i, \max_{k \in \mathcal{I}^i}(N_k - M_{ki} - 1)\} \tag{10}$$

and define the integers b_i by

$$b_i = \max\{d_i, \max_{k \in \mathcal{O}^i}(d_k + M_{ik})\} - N_i \tag{11}$$

Remark. In the remainder of the paper, we use the following additional notation. We define a new function P_t for $t \geq 0$ by

$$P_t = A^1_{0:t} A^2_{0:t} \ldots A^n_{0:t}.$$

Define

$$\alpha_t = \{z^i_{0:t-N_i}, a^i_{0:t-1} \mid i \in \mathcal{V}\},$$
$$\beta_t = \{z^i_{t-N_i-b_i:t-N_i}, a^i_{t-d_i:t} \mid i \in \mathcal{V}\}.$$

Furthermore, the notation $z \notin \alpha_t$ means the set

$$\{z \mid z \notin \alpha_t\} = \{z^i_{t-N_i+1:t} \mid i \in \mathcal{V}\},$$

and the notation $z \notin \beta t$ and $a \notin \beta_t$ mean the sets

$$\{z \mid z \notin \beta_t\} = \{z^i_{0:t-N_i-b_i-1} \mid i \in \mathcal{V}\}.$$
$$\{a \mid a \notin \beta_t\} = \{a^i_{0:t-d_i-1} \mid i \in \mathcal{V}\}.$$

The following theorem is the main result in this paper. It defines the sufficient information state for the networked Markov decision processes. It shows that the networked MDPs can be converted into a fully observable MDP with a state that is bounded and does not grow with time. Note that the networked MDP can be written as a POMDP $\left(\hat{A}, \hat{C}, \hat{g}\right)$, with state $\hat{x}$.

Theorem 11. *Consider a networked Markov decision process. Then,*

$$\xi_t = \left\{u^i_{t-d_i:t-1}, x^i_{t-N_i-b_i:t-N_i} \mid i \in \mathcal{V}\right\}. \tag{12}$$

is the sufficient information state for the networked MDP.

To prove this theorem, we check the conditions of the sufficient information state as given in definition 7. The following key lemma shows that ξ_t as defined in equation (12) satisfies the first condition of the sufficient information state as given in equation (5).

Lemma 12. *Consider a networked Markov decision process (A, g) and a networked MDP policy K. Define*

$$\tilde{A}_{t+1} \triangleq \text{Prob}\left(\xi_{t+1} = q_{t+1} \mid \xi_t = q_t, u_t = a_t\right).$$

Then, ξ_t satisfies the following Markov property

$$\tilde{A}_{t+1} = \text{Prob}\left(\xi_{t+1} = q_{t+1} \mid \xi_{0:t} = q_{0:t}, u_{0:t} = a_{0:t}\right),$$

and $\tilde{A}$ is independent of the policy K.

Proof. Using Bayes' rule, we can write

$$L = \text{Prob}\left(\xi_{t+1} \mid \xi_{0:t}, u_{0:t}\right) = \frac{\text{Prob}\left(\xi_{0:t+1}, u_{0:t}\right)}{\text{Prob}\left(\xi_{0:t}, u_{0:t}\right)}. \tag{13}$$

Note that the sequence $\xi_{0:t}$ consists of the variables $\{x^i_{0:t-N_i}, u^i_{0:t-1} \mid i \in \mathcal{V}\}$. Let us denote the denominator of equation (13) by L_{den}. Then,

$$L_{\text{den}} = \sum_{z \notin \alpha_t} P_t K_{0:t}, \tag{14}$$

where we have used the definition of $\xi_{0:t}$ and the notation that $P_t = A^1_{0:t} \ldots A^n_{0:t}$. Note that the transition kernel A^i_t has arguments

$$z^i_t, z^i_{t-1}, a^i_{t-1}, \{z^k_{t-1-M_{ki}} \mid k \in \mathcal{I}^i\}.$$

We first show that some of the A_t^i are independent of the variables being summed over. Consider an arbitrary $s \geq 0$, and suppose A_{t-s}^i depends upon at least one of $z^1_{t-N_1+1:t}, \ldots, z^n_{t-N_n+1:t}$. Then, we must have

$$\begin{aligned} &t - N_i + 1 \leq t - s && \text{or} \\ &t - N_i + 1 \leq t - s - 1 && \text{or} \\ &t - N_k + 1 \leq t - s - 1 - M_{ki} && \text{for some } k \in \mathcal{I}^i \end{aligned}$$

where each inequality arises from the corresponding argument of A_{t-s}^i. This implies that

$$s \leq N_i - 1 \quad \text{or} \quad s \leq \max\{N_k - 1 - M_{ki} \mid k \in \mathcal{I}^i\} - 1.$$

Hence for each i, the largest such s is exactly equal to $d_i - 1$ where d_i is defined by equation (10). Thus if $s \geq d_i$ then A_{t-s}^i does not depend on any of $z^1_{t-N_1+1:t}, \ldots, z^n_{t-N_n+1:t}$. In other words, $A^i_{0:t-d_i}$ are independent of all the variables of summation. Further note that $K_{0:t}$ only depend on the variables in $\{\alpha_t\}$ and hence are independent of the variables of summation. Thus, we can write the denominator of equation (13) as

$$L_{\text{den}} = A^1_{0:t-d_1} \cdots A^n_{0:t-d_n} K_{0:t} \sum_{z \notin \alpha_t} A^1_{t-d_1+1:t} \cdots A^n_{t-d_n+1:t}. \tag{15}$$

Let us denote the numerator of equation (13) as L_{num}. Then,

$$L_{\text{num}} = \sum_{z \notin \alpha_{t+1}} P_{t+1} K_{0:t}. \tag{16}$$

Following the same argument as above, it is easy to verify that if $s \geq d_i - 1$, then A_{t-s}^i does not depend on any of $z^1_{t-N_1+2:t+1}, \ldots, z^n_{t-N_n+2:t+1}$. Thus, $A^i_{0:t-d_i+1}$ are independent of the variables of summation of L_{num}. We can thus write L_{num} as

$$L_{\text{num}} = A^1_{0:t-d_1} \cdots A^n_{0:t-d_n} K_{0:t} \sum_{z \notin \alpha_{t+1}} A^1_{t-d_1+1:t+1} \cdots A^n_{t-d_n+1:t+1}.$$

Canceling the common factors from the numerator and denominator gives

$$L = \frac{\sum_{z \notin \alpha_{t+1}} A^1_{t-d_1+1:t+1} \cdots A^n_{t-d_n+1:t+1}}{\sum_{z \notin \alpha_t} A^1_{t-d_1+1:t} \cdots A^n_{t-d_n+1:t}}. \tag{17}$$

Using Bayes' rule, we can write

$$R = \text{Prob}\left(\xi_{t+1} \mid \xi_t, u_t\right) = \frac{\text{Prob}\left(\xi_{t+1}, \xi_t, u_t\right)}{\text{Prob}\left(\xi_t, u_t\right)}. \tag{18}$$

Let R_{den} denote the denominator of equation (18). Using the definition of ξ_t, we can write the denominator as,

$$R_{\text{den}} = \sum_{a \notin \beta_t} \sum_{z \notin \beta_t} \sum_{z \notin \alpha_t} P_t K_{0:t}.$$

As before $A^i_{t-d_i}$ and $K_{0:t}$ are independent of the variables of summation $\{z \notin \alpha_t\}$ and hence we can write R_{den} as

$$R_{\text{den}} = \sum_{a \notin \beta_t} \sum_{z \notin \beta_t} A^1_{0:t-d_1} \cdots A^n_{0:t-d_n} K_{0:t} \times \underbrace{\sum_{z \notin \alpha_t} A^1_{t-d_1+1:t} \cdots A^n_{t-d_n+1:t}}_{\hat{R}_{\text{den}}}.$$

Let us determine explicitly what variables $\hat{R}_{\text{den}}$ depends on. For notational convenience, let us denote

$$T = A^1_{t-d_1+1:t} \cdots A^n_{t-d_n+1:t}.$$

If T depends on z_s^i then we must have

$$\begin{aligned} &t - d_i \leq s && \text{or} \\ &t - d_k - M_{ik} \leq s && \text{for some } k \in \mathcal{O}^i. \end{aligned}$$

The first inequality holds if z_s^i occurs in $A^i_{t-d_i+1\cdots t}$ and the second holds if it occurs in $A^k_{t-d_k+1\cdots t}$. If $\hat{R}_{\text{den}}$ depends on on $z^i_{t-N_i-r}$ then,

$$\begin{aligned} &t - d_i \leq t - N_i - r && \text{or} \\ &t - d_k - M_{ik} \leq t - N_i - r && \text{for some } k \in \mathcal{O}^i, \end{aligned}$$

and these conditions imply that

$$\begin{aligned} &r \leq d_i - N_i && \text{or} \\ &r \leq \max\{d_k + M_{ik} \mid k \in \mathcal{O}^i\} - N_i. \end{aligned}$$

Using the definition of b_i in equation (11), these two inequalities imply that $r \leq b_i$. Thus $\hat{R}_{\text{den}}$ depends on $\{a^i_{t-d_i:t-1} \mid \in \mathcal{V}\}$ and $\{z^i_{t-N_i-b_i:t-N_i} \mid i \in \mathcal{V}\}$ and hence is independent of variables $\{a \notin \beta_t\}$ and $\{z \notin \beta_t\}$. Thus, we can write

$$R_{\text{den}} = \left(\sum_{a \notin \beta_t} \sum_{z \notin \beta_t} A^1_{0:t-d_1} \cdots A^n_{0:t-d_n} K_{0:t}\right) \times \left(\sum_{z \notin \alpha_t} A^1_{t-d_1+1:t} \cdots A^n_{t-d_n+1:t}\right). \tag{19}$$

Let R_{num} denote the numerator of the equation (18). Then,

$$R_{\text{num}} = \sum_{a \notin \beta_t} \sum_{z \notin \beta_t} \sum_{z \notin \alpha_{t+1}} P_{t+1} K_{0:t}.$$

Using the same argument as above we can write the numerator as

$$R_{\text{num}} = \left(\sum_{a \notin \beta_t} \sum_{z \notin \beta_t} A^1_{0:t-d_1} \cdots A^n_{0:t-d_n} K_{0:t}\right) \times \left(\sum_{z \notin \alpha_{t+1}} A^1_{t-d_1+1:t+1} \cdots A^n_{t-d_n+1:t+1}\right). \tag{20}$$

From equation (19) and equation (20) we have

$$R = \frac{\sum_{z \notin \alpha_{t+1}} A^1_{t-d_1+1:t+1} \cdots A^n_{t-d_n+1:t+1}}{\sum_{z \notin \alpha_t} A^1_{t-d_1+1:t} \cdots A^n_{t-d_n+1:t}}. \quad (21)$$

The result follows from equations (17) and (21). ∎

The next lemma evaluates the cost function $\tilde{g}_t$ for the induced MDP and shows that it is independent of the POMDP policy chosen.

Lemma 13. *The cost function as defined in equation* (6) *is independent of the POMDP policy K.*

Proof. From equation (6), we have that

$$\tilde{g}_t(q_t, a_t) = \mathbb{E}\left(\hat{g}_t(\hat{x}_t, a_t) \mid \xi_t = q_t, u_t = a_t\right),$$

Using the definition of $\hat{g}_t$ from equation (9), we get that

$$\begin{aligned}\tilde{g}_t(q_t, a_t) &= \mathbb{E}\left(g_t(x_t, a_t) \mid \xi_t = q_t, u_t = a_t\right), \\ &= \sum_{z_t} g_t(z_t, a_t) \frac{\text{Prob}(z_t, q_t, a_t)}{\text{Prob}(q_t, a_t)}.\end{aligned}$$

Using the definition of ξ_t we get that

$$\text{Prob}(q_t, a_t) = \sum_{a \notin \beta_t} \sum_{z \notin \beta_t} \sum_{z \notin \alpha_t} P_t K_{0:t}.$$

Thus, we get that $\tilde{g}_t$ is given as

$$\begin{aligned}\tilde{g}_t(q_t, a_t) &= \frac{\sum_{a \notin \beta_t} \sum_{z \notin \beta_t} \sum_{z \notin \alpha_t} g_t(z_t, a_t) P_t K_{0:t}}{\sum_{a \notin \beta_t} \sum_{z \notin \beta_t} \sum_{z \notin \alpha_t} P_t K_{0:t}}, \\ &= \frac{\sum_{z \notin \alpha_t} g_t(z_t, a_t) A^1_{t-d_1+1:t} \cdots A^n_{t-d_n+1:t}}{\sum_{z \notin \alpha_t} A^1_{t-d_1+1:t} \cdots A^n_{t-d_n+1:t}}.\end{aligned}$$

where the last equality follows from a similar argument as given for equation (19). Thus the cost function is independent of the POMDP policy K. ∎

The following lemma shows that the conditional probability density function for the state at time t is same for the induced MDP and the original POMDP.

Lemma 14. *For all $t \geq 0$, we have*

$$\begin{aligned}&\text{Prob}(\hat{x}_t = \hat{z}_t \mid \xi_{0:t} = q_{0:t}, u_{0:t-1} = a_{0:t-1}) = \\ &\quad \text{Prob}(\hat{x}_t = \hat{z}_t \mid \hat{y}_{0:t} = \hat{s}_{0:t}, u_{0:t-1} = a_{0:t-1}), \quad (22)\end{aligned}$$

where we have used the notation $\gamma_t(s_{0:t}, a_{0:t-1}) = q_t$.

Proof. Note that the sequence $\xi_{0:t}$ consists of the variables $\{x^i_{0:t-N_i}, u^i_{0:t-1} \mid i \in \mathcal{V}\}$. Also for the subsection 3.1, we know that $\hat{y}_t = \left\{x^i_{t-N_i} \mid i \in \mathcal{V}\right\}$. The lemma follows trivially from these two facts. ∎

Proof of Theorem 11. From lemmas 12, 13, and 14, we get that ξ_t as defined in equation (12) is the sufficient information state for the networked MDPs. ∎

5 Conclusions

We studied networked Markov decision processes with network delays between subsystems. Each subsystem transmits its state to a centralized controller via a link with an associated delay. Since the controller does not have access to the current state of the system, these systems are a special case of partially observed Markov decision processes. We show that for this special class of POMDPs, the sufficient information state is a function of finite history of the system state and the past controller inputs. The number of past states as well as the past inputs depends only on the underlying graph structure of the networked Markov decision process. This result shows that the controller synthesis can be achieved at substantially lower computational cost. A dynamic programming algorithm based on the finite information state can be effectively used to compute the optimal controller for such systems.

References

[1] S. Adlakha, R. Madan, S. Lall, and A. Goldsmith. Optimal control of distributed Markov decision processes with network delays. *IEEE Conference on Decision and Control*, pages 3308–3314, 2007.

[2] E. Altman and P. Nain. Closed-loop control with delayed information. *Performance Evaluation Review*, 20:193–204, 1992.

[3] K. J. Astrom. Optimal control of Markov processes with incomplete state estimation. *Journal of Mathematical Analysis and Applications*, 10:174–205, 1965.

[4] D. S. Bernstein, R. Givan, N. Immerman, and S. Zilberstein. The complexity of decentralized control of Markov decision processes. *Mathematics of Operations Research*, 27(4):819–840, 2002.

[5] K. Hsu and H. I. Marcus. Decentralized control of finite state Markov processes. *IEEE Transactions on Automatic Control*, 2:426–431, 1982.

[6] K. V. Katsikopoulos and S. E. Engelbrecht. Markov decision processes with delays and asynchronous cost collection. *IEEE Transactions on Automatic Control*, 48(4):568–574, 2003.

[7] P. R. Kumar and P. Varaiya. *Stochastic Systems: Estimation, Identification and Adaptive Control*. Prentice Hall, 1986.

[8] R. D. Smallwood and E. J. Sondik. The optimal control of partially observable Markov processes over a finite horizon. *Operations Research*, 21(5):1071–1088, 1973.

[9] P. Varaiya and J. Walrand. On delayed sharing patterns. *IEEE Transactions on Automatic Control*, 23:443–445, 1978.