A Bayesian Network Approach to Control of Networked Markov Decision Processes

Sachin Adlakha, Sanjay Lall and Andrea Goldsmith

***Abstract*— We consider the problem of finding an optimal feedback controller for a networked Markov decision process. Specifically, we consider a network of interconnected subsystems, where each subsystem evolves as a Markov decision process (MDP). A subsystem is connected to its neighbors via links over which signals are delayed. We consider centralized control of such networked MDPs. The controller receives delayed state information from each of the subsystem, and it chooses control actions for all subsystems. Such networked MDPs can be represented as partially observed Markov decision processes (POMDPs). We model such a POMDP as a Bayesian network and show that an optimal controller requires only a finite history of past states and control actions. The result is based on the idea that given certain past states and actions, the current state of the networked MDP is independent of the earlier states and actions. This dependence on only the finite past states and actions makes the computation of controllers for networked MDPs tractable.**

I. Introduction and Prior Work

We consider a network of interconnected subsystems, where each subsystem evolves as a Markov decision process. 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 of the subsystem. We refer to such systems as *networked Markov decision processes*.

In networked MDPs, the controller receives delayed state information from each subsystem. 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). Optimal control design for POMDPs has been studied extensively in literature [1], [2], [3]. The separation theorem for POMDPs states that the optimal controller is a function of the posterior distribution of the current state given all past observations. The control of a single MDP with delayed state information was considered in [4]. It was shown that the optimal control action depends upon the last observed state and a finite number of previous actions. For distributed systems, the earliest result was obtained in [5], where the separation structure for *one-step delay sharing* pattern for general non-linear dynamics was obtained.

A general networked system with arbitrary delay pattern was first considered in [6]. It was shown that a centralized optimal controller for such systems need only store the past few states of each subsystem and past few actions. In [7], the authors extend this result to a case where control action is applied to every subsystem. Using the principle of dynamic programming, the authors show that for networked MDPs, the information state consists of only a finite past history of states and actions. In this paper, we show that the finite history of states and actions that was obtained in [7] is exactly same as the information required to estimate the current state of the system. This, along with the separation principle, provides an alternate proof and additional insights into the finite memory of the controllers for networked MDPs. It shows that the finiteness of the bands occurs because given the finite history of states and actions, the current state of the system is independent of the remaining states and actions.

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 x_t^1 denotes the state of the subsystem 1 at time t. We use z to denote a realization of the state x and use a to denote a realization of the control input u. We define $x^i_{t_1:t_2} := \big(x^i_{t_1}, \dots, x^i_{t_2}\big)$ to refer to the list of variables corresponding to the subsystem i from time t_1 to t_2. If $t_2 < t_1$, we interpret the list as empty. To denote the list of variables corresponding to all subsystems, we define $x_t := \big(x_t^1, \dots, x_t^n\big)$. Similarly, we denote $u_t := \big(u_t^1, \dots, u_t^n\big)$ as the control action applied to all subsystems at time t. We define $A^i_{0\cdots t}$ to be the product of the variables corresponding to times $0, \dots, t$, that is $A^i_{0\cdots t} := A^i_0 A^i_1 \dots A^i_t$. 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 $\mathbb{N}$ for the set of natural numbers.

II. Model and Definitions

A. Networked Markov Decision Processes

A networked Markov decision process is a weighted directed graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$, where $\mathcal{V} = \{1, \dots, n\}$ 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

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

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.

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

This work was supported by Stanford clean slate network design and DARPA ITMANET programs.

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

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

Associated with each subsystem $i \in \mathcal{V}$ is the finite set $\mathcal{X}^i$, such that the state of subsystem i at time t is $x_t^i \in \mathcal{X}^i$. The system dynamics are

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

for all $i \in \mathcal{V}$. Here we write the function f^i taking an argument which is a set, with the understanding that the elements of the set are associated with particular vertices (in a programming language we would say that f takes named arguments). Subsystem i has control action $u_t^i \in \mathcal{U}^i$ applied at time t, where $\mathcal{U}^i$ has finite cardinality. 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 each subsystem $i \in \mathcal{V}$ we have a nonnegative integer N_i. The observations received by the controller at time t consist of the state of the subsystem i delayed by N_i time steps. At time t the controller thus receives $x_{t-N_i}^i$ for all $i \in \mathcal{V}$. The controller chooses input $\{u_t^i \mid i \in \mathcal{V}\}$ at time t based on history of these observations and its previous actions.

Transition probabilities. For $p \in \mathcal{X}^i$, let $A_0^i(p) = \mathrm{Prob}(x_0^i = p)$ define the probability mass function of the initial state of subsystem $i \in \mathcal{V}$. The initial states $x_0^1, \ldots, x_0^n$ are chosen independently. Let

$$\begin{aligned} A_t^i(z,p,q,a) = \mathrm{Prob}\Big(& x_t^i = z \mid x_{t-1}^i = p, \\ & \{x_{t-1-M_{ji}}^j = q^j \mid j \in \mathcal{I}^i\}, u_{t-1}^i = a\Big), \end{aligned} \qquad (2)$$

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. These probability mass functions are uniquely defined by equation (1) along with the statistics of the noise processes w_t^i. Also note that given these probability mass functions and the Markov assumption on the system, we can easily derive the functions f^i governing the system dynamics in equation (1). Thus, these mass functions are an equivalent representation of the system. The probability mass functions also encode the conditional dependence of the state x_t^i on the previous state of system i and past states of systems $j \in \mathcal{I}^i$.

Measurements available to the controller. We would like to consider the optimal performance achievable when the controller has access to the entire measurement history, and show that this level of performance may be achieved even if the controller only stores recent measurements. The complete history of measurements is defined as follows.

Definition 1: We define h_t to be the information available to the controller at time t, given by

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

Also denote i_t to be a realization of h_t as

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

Further, define the set $\mathcal{H}_t$ as

$$\mathcal{H}_t = \prod_{i=1}^n \left(\mathcal{X}^i\right)^{t+1-N_i} \times \prod_{i=1}^n \left(\mathcal{U}^i\right)^t$$

Here the sequences z and a specify the values of a realization of x and u, respectively. We consider general *mixed policies* for the controller input, *i.e.*, we consider controllers such that the control input at time t is specified by a probability distribution which is a function of the observations available to the controller. To do this, let the conditional probability for the control action u_t be K_t, so that

$$K_t(a_t, y_t) = \mathrm{Prob}(u_t = a_t \mid h_t = i_t)$$

Note that deterministic controllers are a special case of the above; a deterministic controller can be chosen by choosing all the densities K_t to be atomic. Also note that an optimal controller may always be found which is deterministic, and we explain how to construct it in this paper.

1) Example: Before we introduce the main result, we illustrate the main point of the paper via an example. Consider a networked MDP as shown in Figure 1. From the results proved in the paper, we would show that for such a networked system, the optimal controller is only required to store $b_i + 1$ values of the state of system i and d_i values of the past inputs to the subsystem i, where

$$\begin{aligned} b_1 &= \max\{0, N_2 + M_{12} - N_1\}, \\ b_2 &= \max\{0, N_1 + M_{21} - N_2\}, \\ d_1 &= \max\{N_1, N_2 - M_{21} - 1\}, \\ d_2 &= \max\{N_2, N_1 - M_{12} - 1\}. \end{aligned} \qquad (3)$$

In other words, an optimal controller exists for which the control action u_t is a memoryless function of previous control inputs $u_{t-d_i}^i, \ldots, u_{t-1}^i$ and measurements $x_{t-N_i-b_i}^i, \ldots, x_{t-N_i}^i$ only.

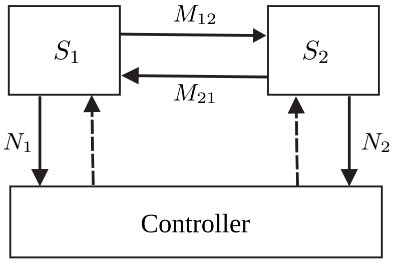

Fig. 1. A network of two interconnected subsystems with delays. Subsystem i is denoted by S_i, the network propagation delay from S_i to S_j is denoted by M_{ij} and the measurement delay from S_i to the controller is denoted N_i.

B. Bayesian Networks

A Bayesian network [8], $\mathcal{N}_b = \{\mathcal{G}_b, \mathcal{P}_b\}$ consists of

- A directed acyclic graph $\mathcal{G}_b = (\mathcal{V}_b, \mathcal{E}_b)$, and
- A set of conditional probability distributions $\mathcal{P}_b$.

Here the subscript b stands for Bayesian and is used to distinguish the Bayesian network graph from the networked MDP graph $\mathcal{G}$ as defined in the previous section. Associated with each vertex $v \in \mathcal{V}_b$ of the graph $\mathcal{G}_b$, is a random variable X_v taking values in a particular set. A directed edge $e \in \mathcal{E}_b$ between vertices describe the conditional dependence between the random variables corresponding to the vertices. If there is a directed edge from a vertex v_1 to v_2, we say that v_2 is a child of v_1 and that v_1 is a parent of v_2. The set of parent vertices of a vertex v is denoted by parent(v).

The set of probability distributions $\mathcal{P}_b$ contains one distribution $P\big(X_v|X_{\text{parent(v)}}\big)$ for every $v \in \mathcal{V}_b$. The joint distribution of all the variables $X_k, k = 1, \dots, n$ is given as

$$\mathrm{Prob}\big(X_1, \dots X_n\big) = \prod_{k=1}^{n} \mathrm{Prob}\big(X_k \mid \text{parents}(X_k)\big)$$

An example of a Bayesian network is shown in Figure 2. Here the graph $\mathcal{G}_b$ consists of vertices $\{A, B, C, D, E, F\}$ and edges $\{A \to C, B \to C, C \to D, C \to E, D \to F\}$. The set of probabilities is given as

$$\mathcal{P}_b = \{P(A), P(B), P(C|A, B), P(D|C), \\ P(E|C), P(F|D)\}.$$

Note that since the variables A and B have no parents, the probability set contains their unconditional probabilities.

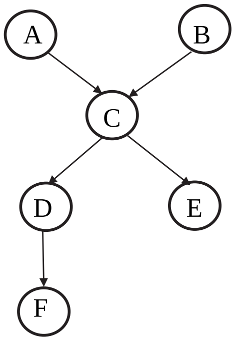

Fig. 2. A Bayesian network with 6 variables

d-Separation. As mentioned before, the graph $\mathcal{G}_b$ encodes the conditional dependencies between the variables. Conditional independence between variables is determined by the property of *d-separation*. If two variables X and Y are d-separated in the graph by a third variable Z, then the variables X and Y are conditionally independent given the variable Z.

Definition 2: A path π in the graph $\mathcal{G}_b = \{\mathcal{V}_b, \mathcal{E}_b\}$ is said to be d-separated by a set of nodes $Z \in \mathcal{V}_b$ if and only if one of the following holds

- π contains a chain $i \to z \to j$ such that $i, j \in \pi$ and $z \in Z$,
- π contains a fork $i \leftarrow z \to j$ such that $i, j \in \pi$ and $z \in Z$ and
- π contains an inverted fork (or a collider) $i \to z \leftarrow j$ such that $i, j \in \pi$ and neither z nor any of its children are in Z.

The concept of d-separation is closely tied to that of a Markov blanket. Before we define the Markov blanket, we introduce some notation.

Remark: Consider a set of variables $X = \{X_1, \dots, X_n\}$. Denote $\mathrm{P}(X)$ to be the set consisting of all parents of variables in the set X, not including the variables themselves. Similarly, we denote $\mathrm{CH}(X)$ (and $\mathrm{PCH}(X)$) to be the set consisting of all children (parents of children) of variables in the set X, not including the variables themselves.

Definition 3 (Markov Blanket): The Markov blanket of set of variables $X = \{X_1, \dots, X_n\}$(denoted by $\mathrm{MB}(X)$) is given as

$$\mathrm{MB}(X) = \mathrm{P}(X) \cup \mathrm{CH}(X) \cup \mathrm{PCH}(X) \tag{4}$$

The following theorem (see [8] for the proof) states that the variables in the set X are independent of the rest of the graph given its Markov blanket.

Theorem 4: Given a finite Bayesian network and two distinct variables X and $Y \notin \mathrm{MB}(X)$, we have

$$\mathrm{Prob}\big(X|\mathrm{MB}(X), Y\big) = \mathrm{Prob}\big(X|\mathrm{MB}(X)\big)$$

The Markov blanket of the set of variables shields the variables from the rest of the graph. Thus, the Markov blanket is the only knowledge required to predict the value of the variables. Furthermore, if all the variables in Markov blanket of X are known, then X is d-separated from the rest of the graph [8].

C. Networked MDPs as Bayesian Networks

In this subsection, we model networked Markov decision processes as Bayesian networks in a natural way. Consider a networked MDP given by a graph $\mathcal{G} = \{\mathcal{V}, \mathcal{E}\}$, where we let $\mathcal{V} = \{1, \dots, n\}$. As before, for each $i \in \mathcal{V}$. we have $x_t^i \in \mathcal{X}^i$. For the reminder of the paper, we would consider the evolution of the networked MDP over a finite horizon T. Associated with this networked MDP, we can construct a finite Bayesian network $\mathcal{N}_b = \{\mathcal{G}_b, \mathcal{P}_b\}$. The vertex set $\mathcal{V}_b$ is given as

$$\mathcal{V}_b = \big\{v_{i,t}^{\text{state}} \mid i \in \mathcal{V}, t = 0, 1, \dots, T\big\} \bigcup \\ \big\{v_{i,t}^{\text{action}} \mid i \in \mathcal{V}, t = 0, 1, \dots, T-1\big\}$$

Associated with a vertex $v_{i,t}^{\text{state}}$ is the random variable x_t^i, taking values in the finite set $\mathcal{X}^i$, that corresponds to the state of subsystem i at time t. Similarly, associated with a vertex $v_{i,t}^{\text{action}}$ is the random variable u_t^i, taking values in the finite set $\mathcal{U}^i$, that corresponds to the control action applied to subsystem i at time t. The edge set $\mathcal{E}_b$ consists of the following edges.

$$\mathcal{E}_b = \big\{v_{i,t}^{\text{state}} \to v_{i,t+1}^{\text{state}}, v_{j,t-M_{ji}}^{\text{state}} \to v_{i,t+1}^{\text{state}}, \\ v_{i,t}^{\text{action}} \to v_{i,t+1}^{\text{state}}, v_{i,0:t-N_i}^{\text{state}} \to v_{k,t}^{\text{action}}, \\ v_{i,0:t-1}^{\text{action}} \to v_{k,t}^{\text{action}} \mid j \in \mathcal{I}^i, i, k \in \mathcal{V}, t \in \mathbb{N}\big\}$$

Here $v^{\text{state}}_{i,0:t-N_i} \to v^{\text{action}}_{k,t}$ is interpreted as a directed edge between $v^{\text{state}}_{i,\tau} \to v^{\text{action}}_{k,t}$ for every $\tau = 0, \ldots, t-N_i$. An edge $v^{\text{state}}_{j,t-M_{ji}} \to v^{\text{state}}_{i,t+1}$ means that the random variable $x^j_{t-M_{ji}}$ affects the random variable x^i_{t+1}. Similar interpretations exist for other edges in the edge set $\mathcal{E}_b$. The set of conditional probability densities $\mathcal{P}_b$ consists of all the transition probabilities, that is

$$\mathcal{P}_b = \big\{A^i_t \mid i \in \mathcal{V},\ t = 0, \ldots, T\big\} \cup \big\{K_t \mid t = 0, \ldots, T-1\big\}$$

For a finite time horizon T, let $\mathcal{S}_T$ be the set of random variables given as

$$\mathcal{S}_T = \big\{x^i_t \mid i \in \mathcal{V},\ t = 0, 1, \ldots, T\big\} \quad \bigcup \\ \big\{u^i_t \mid i \in \mathcal{V},\ t = 0, 1, \ldots, T-1\big\}$$

The joint probability density function of all the variables in the set $\mathcal{S}_T$ can then be written as

$$\operatorname{Prob}(\mathcal{S}_T) = A^1_{0:T} A^2_{0:T} \ldots A^n_{0:T} K_{0:T-1}$$

As an example, consider again the networked system of Figure 1. The system dynamics equations are given as

$$\begin{aligned} x^1_{t+1} &= f^1(x^1_t, x^2_{t-M_{21}}, u^1_t, w^1_t), \\ x^2_{t+1} &= f^2(x^2_t, x^1_{t-M_{12}}, u^2_t, w^2_t). \end{aligned} \tag{5}$$

For the purpose of this example, we choose $M_{12} = 2$ and $M_{21} = 1$. Thus, the transition probability matrices are given as

$$A^1_t\Big(z^1_t, z^1_{t-1}, z^2_{t-2}, a^1_{t-1}\Big) = \operatorname{Prob}\Big(x^1_t = z^1_t \mid x^1_{t-1} = z^1_{t-1}, \\ x^2_{t-2} = z^2_{t-2}, u^1_{t-1} = a^1_{t-1}\Big), \tag{6}$$

and

$$A^2_t\Big(z^2_t, z^2_{t-1}, z^1_{t-3}, a^2_{t-1}\Big) = \operatorname{Prob}\Big(x^2_t = z^2_t \mid x^2_{t-1} = z^2_{t-1}, \\ x^1_{t-3} = z^1_{t-3}, u^2_{t-1} = a^2_{t-1}\Big), \tag{7}$$

Associated with this networked control system is a Bayesian network as shown in Figure 3. The directed acyclic graph $\mathcal{G}_b$ consists of a vertex for each state of the two systems and two control actions applied at time t. A directed edge between two vertices v_1 and v_2 exists if the variable corresponding to vertex v_1 affects the variable corresponding to vertex v_2. For example, a directed edge exists between the vertex corresponding to x^2_{t-2} and the vertex corresponding to x^1_t. Similarly, a directed edge exists between the vertex corresponding to control action u^2_{t-1} and the vertex corresponding to x^2_t. The set of probability distributions $\mathcal{P}_b$ consists of the transition probabilities A^1_t, A^2_t and K_t for all $t \geq 0$.

III. Main Results

Before we present the statement of the theorem, we make the following definitions.

Definition 5: Let

$$d_i = \max\{N_i, \max_{k \in \mathcal{I}^i}(N_k - M_{ki} - 1)\} \tag{8}$$

and define the integers b_i by

$$b_i = \max\{d_i, \max_{k \in \mathcal{O}^i}(d_k + M_{ik})\} - N_i \tag{9}$$

Define

$$h^{\text{mem}}_t = \big(x^1_{t-N_1-b_1:t-N_1}, u^1_{t-d_1:t-1}, \ldots, \\ x^n_{t-N_n-b_n:t-N_n}, u^n_{t-d_n:t-1}\big) \tag{10}$$

to the finite history of observations at time t and denote

$$i^{\text{mem}}_t = \big(z^1_{t-N_1-b_1:t-N_1}, a^1_{t-d_1:t-1}, \ldots, \\ z^n_{t-N_n-b_n:t-N_n}, a^n_{t-d_n:t-1}\big)$$

to be a realization of h^{mem}_t. Further define the set $\mathcal{H}^{\text{mem}}_t$ as

$$\mathcal{H}^{\text{mem}}_t = \prod_{i=1}^{n} \left(\mathcal{X}^i\right)^{b_i+1} \times \prod_{i=1}^{n} \left(\mathcal{U}^i\right)^{d_i}.$$

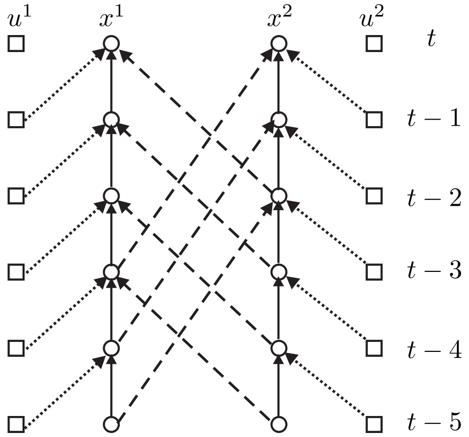

Fig. 3. The Bayesian network associated with the 2-subsystems networked MDP of Figure 1. Here the circle represents the state of the two subsystems and the square represents the control input. For this Bayesian network, we chose $M_{21} = 1$ and $M_{12} = 2$. The edges from state variables to control inputs have been omitted for visual clarity.

From the separation principle [1], we know that the optimal control action is a function of the belief state. We define the set of belief states at time t as follows.

Definition 6: Let $\mathcal{M}_t$ be a set defined as

$$\mathcal{M}_t = \big\{\Lambda_t : \mathcal{X}^{(n)} \times \mathcal{H}_t \to [0,1] \mid \Lambda_t(z_t, i_t) \geq 0, \\ \sum_{z_t} \Lambda_t(z_t, i_t) = 1\big\},$$

where we denote $\mathcal{X}^{(n)} = \prod_{i=1}^n \mathcal{X}^i$ to be the cartesian product of the state space corresponding to all vertices. Here, $\Lambda_t(z_t, i_t)$ is interpreted as the conditional probability density of the current state of the system given the entire observation history at time t. That is

$$\Lambda_t(z_t, i_t) = \operatorname{Prob}(x_t = z_t \mid h_t = i_t)$$

Let $\mathcal{F}_t : \mathcal{H}_t \to \mathcal{M}_t$ be a operator that maps the entire observation history at time t to an element in $\mathcal{M}_t$. That is, the operator $\mathcal{F}_t$ maps the observation history to a belief state. Furthermore, let $\mathcal{T}_t : \mathcal{M}_t \to \mathcal{A}$ be the operator that maps belief state to a control action. From the separation principle [1], we know that the optimal control K_t^*, as function of the observation history i_t, is given as

$$K_t^* = \mathcal{T}_t \circ \mathcal{F}_t$$

That is, $K_t^*(a_t, i_t) = \mathcal{T}_t\left(a_t, \Lambda_t(\cdot, i_t)\right)$.

The main result of the paper shows that for networked MDPs, there exists an optimal controller that depends only on i_t^{mem}. Let $\mathcal{P} : \mathcal{H}_t \to \mathcal{H}_t^{\text{mem}}$ be the projection operator that projects the entire observation history to a truncated history as defined in equation (10). The following theorem shows that there exists an operator $\mathcal{F}_t^{\text{mem}} : \mathcal{H}_t^{\text{mem}} \to \mathcal{M}_t$ such that

$$\mathcal{F}_t = \mathcal{F}_t^{\text{mem}} \circ \mathcal{P}$$

Theorem 7: For a networked Markov decision process, there exists $\Lambda_0^*, \dots \Lambda_T^*$ such that

$$\Lambda_t\left(z_t, i_t\right) = \Lambda_t^*\left(z_t, i_t^{\text{mem}}\right) \;\forall\; t = 0, 1, \dots T. \tag{11}$$

Thus, there exists an optimal controller $K_0^*, \dots, K_{T-1}^*$ such that

$$\begin{aligned} K_t^*\left(a_t, i_t\right) &= \mathcal{T}_t\left(a_t, \Lambda_t^*\left(\cdot, i_t^{\text{mem}}\right)\right) \\ &= \hat{K}_t\left(a_t, i_t^{\text{mem}}\right) \;\forall\; t = 0, 1, \dots T-1. \end{aligned} \tag{12}$$

Thus, b_i's are the bounds on the length of the observation history that an optimal estimator needs to maintain beyond it current observation.

Before we present the proof of the Theorem 7, we first prove a key lemma.

Lemma 8: Suppose there exists an optimal $K_j^*, j = t+1, \dots, T-1$ such that

$$K_j^*\left(a_j, i_j\right) = \hat{K}_j\left(a_j, i_j^{\text{mem}}\right)$$

for all a_j. Then

$$K_t^*\left(a_t, i_t\right) = \hat{K}_t\left(a_t, i_t^{\text{mem}}\right)$$

for all a_t.

Proof. From the separation principle [1], we know that

$$K_t^*\left(a_t, i_t\right) = T\left(a_t, \Lambda_t\left(\cdot, i_t\right)\right)$$

Thus, to prove the lemma it suffices to show that $\Lambda_t\left(z_t, i_t\right) = \Lambda_t^*\left(a_t, i_t^{\text{mem}}\right)$. At time t, the controller knows $i_t = \{z^i_{0:t-N_i}, a^i_{0:t-1} \mid i \in \mathcal{V}\}$. Let

$$\mathcal{S}_t^u = \left(x^1_{t-N_1+1:t}, \dots, x^n_{t-N_n+1:t}\right)$$

be the states that are unknown at the controller at time t. Here the superscript u is used to indicate that these states are unknown to the controller at time t. Note that states of subsystem i are part of $\mathcal{S}_t^u$ if and only if $N_i \geq 1$. This is because if $N_i = 0$, then the current state of subsystem i is known to controller. Let

$$\mathcal{Z}_t^u = \left(z^1_{t-N_1+1:t}, \dots, z^n_{t-N_n+1:t}\right)$$

be a realization of $\mathcal{S}_t^u$. Let $L_t\left(\mathcal{Z}_t^u, i_t\right)$ be the joint conditional probability of the variables in the set $\mathcal{S}_t^u$ given i_t. That is,

$$L_t\left(\mathcal{Z}_t^u, i_t\right) = \operatorname{Prob}\left(\mathcal{S}_t^u = \mathcal{Z}_t^u \mid h_t = i_t\right)$$

Define

$$L_t^*\left(\mathcal{Z}_t^u, i_t^{\text{mem}}\right) = \operatorname{Prob}\left(\mathcal{S}_t^u = \mathcal{Z}_t^u \mid h_t^{\text{mem}} = i_t^{\text{mem}}\right).$$

If we can show that there exists L_t^* such that

$$L_t\left(\mathcal{Z}_t^u, i_t\right) = L_t^*\left(\mathcal{Z}_t^u, i_t^{\text{mem}}\right), \tag{13}$$

then it follows that

$$\begin{aligned} \Lambda_t\left(z_t, i_t\right) &= \sum_{\{z^i_{t-N_i+1:t-1} \mid i \in \mathcal{V}\}} L_t\left(\mathcal{Z}_t^u, i_t\right) \\ &= \sum_{\{z^i_{t-N_i+1:t-1} \mid i \in \mathcal{V}\}} L_t^*\left(\mathcal{Z}_t^u, i_t^{\text{mem}}\right) \\ &= \Lambda_t^*\left(z_t, i_t^{\text{mem}}\right) \end{aligned} \tag{14}$$

Thus, to prove the lemma it suffices to find an L^* satisfying equation (13). To prove the existence of an L^*, we show that the Markov blanket of the set $\mathcal{S}_t^u$ consists of the variables i_t^{mem}. Theorem 4 would then prove the existence of L^*.

Note that $\mathcal{S}_t^u$ contains $x^j_{t-\tau_j}$ for $\tau_j = 0, 1, \dots, N_j - 1$ and $j = 1, 2, \dots, n$. From equation (4), we know that the Markov blanket of $\mathcal{S}_t^u$ consists of parents, children and parents of children of the variables in the set $\mathcal{S}_t^u$. We focus on a single variable $x^j_{t-\tau_j}$ and find its parents, its children and all the parents of its children.

To find the parents of $x^j_{t-\tau_j}$, we look at the transition probability of this variable. From equation (2), we note that $x^j_{t-\tau_j}$ depends on

$$\mathrm{P}\left(x^j_{t-\tau_j}\right) = \left\{x^j_{t-\tau_j-1}, u^j_{t-\tau_j-1}, x^s_{t-(\tau_j+1+M_{sj})} \mid s \in \mathcal{I}^j\right\}, \tag{15}$$

and hence these variables are the parents of $x^j_{t-\tau_j}$.

To find the children of $x^j_{t-\tau_j}$, consider the set $\mathcal{O}^j$ of outgoing vertices of subsystem j and let $p \in \mathcal{O}^j$. Consider $A^p_{t-t'}$ and note that this transition probability contains $x^j_{t-t'-1-M_{jp}}$. Thus, $x^j_{t-\tau_j}$ would be a parent of $x^p_{t-t'}$ for all $p \in \mathcal{O}^j$, if $t - t' - 1 - M_{jp} = t - \tau_j$, which gives that $t' = \tau_j - 1 - M_{jp}$.

Note that the children of $x^j_{t-\tau_j}$ also consist of all the control variables that depend on $x^j_{t-\tau_j}$. From the assumption in the lemma, we know that the $K^*_{t+1:T-1}$ are only a function of the finite past history of states given by i^{mem}. Thus, a directed edge exists between $x^j_{t-\tau_j}$ and $u_{t-t'}$ for all $t' = \tau_j - N_j - b_j : \tau_j - N_j$. Thus, the children of $x^j_{t-\tau_j}$ consists of

$$\begin{aligned} \mathrm{CH}\left(x^j_{t-\tau_j}\right) = &\left\{x^j_{t-\tau_j+1}, x^p_{t-\tau_j+M_{jp}+1} \mid p \in \mathcal{O}^j\right\} \bigcup \\ &\left\{u^k_{t-\tau_j+N_j:t-\tau_j+N_j+b_j} \mid k \in \mathcal{V}\right\} \end{aligned} \tag{16}$$

To find the parents of children of $x^j_{t-\tau_j}$, we find the parents of the variables given in equation (16). From transition probability equation (2), we note that the parents of $x^p_{t-\tau_j+M_{jp}+1}$ include

$$\left\{x^p_{t-\tau_j+M_{jp}}, u^p_{t-\tau_j+M_{jp}}, x^r_{t-\tau_j+M_{jp}-M_{rp}} \mid r \in \mathcal{I}^p\right\}$$

To find the parents of $\{u^k_{t-\tau_j+N_j:t-\tau_j+N_j+b_j} \mid k \in \mathcal{V}\}$, we note that from the assumption in the lemma, these control inputs only depend on i^{mem}_t. Thus, the parents of $\{u^k_{t-\tau_j+N_j:t-\tau_j+N_j+b_j} \ k \in \mathcal{V}\}$ consist of

$$\Big\{x^i_{t-\tau_j+N_j-b_i-N_i:t-\tau_j+N_j+b_j-N_i}, \\ u^i_{t-\tau_j+N_j-d_i:t-\tau_j+N_j+b_j-1} \mid i \in \mathcal{V}\Big\}$$

Thus we have

$$\begin{aligned}\text{PCH}\big(x^j_{t-\tau_j}\big) = \Big\{&x^s_{t-\tau_j-M_{sj}}, u^j_{t-\tau_j}, x^p_{t-\tau_j+M_{jp}}, u^p_{t-\tau_j+M_{jp}}, \\ &x^r_{t-\tau_j+M_{jp}-M_{rp}} \mid \ s \in \mathcal{I}^j, r \in \mathcal{I}^p, \ p \in \mathcal{O}^j\Big\} \bigcup \\ &\Big\{x^i_{t-\tau_j+N_j-b_i-N_i:t-\tau_j+N_j+b_j-N_i}, \\ &u^i_{t-\tau_j+N_j-d_i:t-\tau_j+N_j+b_j-1} \mid i \in \mathcal{V}\Big\} \quad (17)\end{aligned}$$

Let us denote the set of parents, the children and the parents of children of $x^j_{t-N_j+1:t}$ by $\mathcal{M}_j$. From equations (15), (16), (17), we get that the set $\mathcal{M}_j$ contains

$$\begin{aligned}\mathcal{M}_j = \Big\{&x^j_{t-N_j:t+1}, x^s_{t-(N_j+M_{sj}):t-M_{sj}}, \\ &x^p_{t-(N_j-1-M_{jp}):t+M_{jp}+1}, x^r_{t-(N_j-1-M_{jp}+M_{rp}):t-(M_{rp}-M_{jp})}, \\ &x^i_{t-N_i-b_i+1:t-N_i+b_j+N_j} \mid s \in \ \mathcal{I}^j, p \in \mathcal{O}^j, \ r \in \mathcal{I}^p, \ i \in \mathcal{V}\Big\} \bigcup \\ &\Big\{u^j_{t-N_j:t}, u^k_{t+1:t+N_j+b_j}, u^p_{t-(N_j-1-M_{jp}):t+M_{jp}}, \\ &u^i_{t-(d_i-1):t+N_j+b_j-1} \mid p \in \mathcal{O}^j, \ k, i \in \mathcal{V}\Big\}\end{aligned}$$

Let us denote $\mathcal{M} = \cup_{j\in\mathcal{V}}\mathcal{M}_j$. Note that $u^k_{t-s_k} \in \mathcal{M}$ if $s_k \geq N_k$ or $s_k \geq N_j - 1 - M_{jk}$ for all $j \in \mathcal{I}^k$ or $s_k \geq d_k - 1$. From definition 5, this implies that

$$\begin{aligned}s_k &= \max\big\{N_k, d_k - 1, N_j - M_{jk} - 1 \mid j \in \mathcal{I}^k\big\} \\ &= d_k\end{aligned}$$

Similarly, $x^k_{t-q_k} \in \mathcal{S}$ if and only if $x^k_{t-q_k} \in \mathcal{M}$. This happens if one of the following conditions holds.

1) $q_k \geq N_k$.
2) $q_k \geq N_j + M_{kj}$ such that $k \in \mathcal{I}^j$ for some $j \in \mathcal{V}$. This happens for all $j \in \mathcal{O}^k$.
3) $q_k \geq (N_j - 1 - M_{jk})$ such that $k \in \mathcal{O}^j$ for some $j \in \mathcal{V}$. That is if $q_k = (N_j - 1 - M_{jk})$ for all $j \in \mathcal{I}^k$.
4) For the last term, we need to find all $j \in \mathcal{V}$ such that for all $p \in \mathcal{O}^j$, we have $k \in \mathcal{I}^p$. This happens for all $j \in \mathcal{I}^p$, such that $p \in \mathcal{O}^k$. Thus we have $q_k \geq N_j - 1 - M_{jp} + M_{kp}$ for all $p \in \mathcal{O}^k$ and all $j \in \mathcal{I}^p$.
5) $q_k \geq b_k + N_k - 1$

Thus, we get that

$$\begin{aligned}q_k = \max\Big\{&N_k, N_s + M_{ks}, N_r - 1 - M_{rk}, \\ &N_p - 1 - M_{ps} + M_{ks}, b_k + N_k - 1 \mid p \in \mathcal{I}^s, s \in \mathcal{O}^k, r \in \mathcal{I}^k\Big\}\end{aligned}$$

Using the definition of b_k and d_k, it is easy to verify that $q_k = b_k + N_k$. This proves that the Markov blanket of the variables $\mathcal{S}^u_t$ consists of only i^{mem}_t. Thus, there exists L^*_t such that equation (13) is satisfied. The lemma then follows from equation (14). ■

Proof of Theorem 7. To prove the main theorem, we first show that at time $T-1$, the belief state is only a function of i^{mem}_{T-1}. To see this, note that at time $T-1$, the set of unknown states at the controller $\mathcal{S}^u_T$ has no children. Thus, using a simplified version of the argument given in the proof of lemma 8, it is easy to verify that there exists Λ^*_{T-1} such that

$$\Lambda_{T-1}\left(a_{T-1}, i_{T-1}\right) = \Lambda^*_{T-1}\left(a_{T-1}, i^{\text{mem}}_{T-1}\right).$$

Thus, there exists an optimal controller K^*_{T-1} such that

$$\begin{aligned}K^*_{T-1}\left(a_{T-1}, i_{T-1}\right) &= \mathcal{T}\left(a_{T-1}, \Lambda^*_{T-1}\left(\cdot, i^{\text{mem}}_{T-1}\right)\right) \\ &= \hat{K}_{T-1}\left(a_{T-1}, i^{\text{mem}}_{T-1}\right)\end{aligned}$$

The proof of the theorem then follows from the inductive argument using Lemma 8. ■

IV. Conclusions

We studied centralized control of networked Markov decision processes with delays using a Bayesian network approach. Each subsystem in the networked MDP transmits its state to a controller via a link with an associated delay. Since the controller does not have access to the current state of the system, this networked MDP can be modeled as a POMDP. For this POMDP, the optimal control action at time t is a function of the belief state at time t which is the conditional distribution of the current state of the system given the entire past observation history. We show that the for networked MDPs with delays, the belief state is a function of only the finite history of observations. In particular, the optimal controller depends on only the most recent $b_i + 1$ states received from subsystem i and d_i control inputs applied to subsystem i. The bands b_i and d_i depend on the inter-subsystem delays, the measurement delays and the underlying graph structure of the networked MDP.

Acknowledgments

We thank Prof. Matthew West for helpful discussions.

References

[1] K. J. Astrom, “Optimal control of Markov processes with incomplete state estimation,” *Journal of Mathematical Analysis and Applications*, vol. 10, pp. 174–205, 1965.

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

[3] R. D. Smallwood and E. J. Sondik, “The optimal control of partially observable Markov processes over a finite horizon,” *Operations Research*, vol. 21, no. 5, pp. 1071–1088, 1973.

[4] E. Altman and P. Nain, “Closed-loop control with delayed information,” *Performance Evaluation Review*, vol. 20, pp. 193–204, 1992.

[5] P. Varaiya and J. Walrand, “On delayed sharing patterns,” *IEEE Transactions on Automatic Control*, vol. 23, pp. 443–445, 1978.

[6] 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*, pp. 3308–3314, 2007.

[7] S. Adlakha, S. Lall, and A. Goldsmith, “Information state for Markov decision processes with network delays,” *Accepted to IEEE Conference on Decision and Control*, 2008.

[8] F. V. Jensen, *Bayesian Networks and Decision Graphs*. Springer, 2001.