A Dynamic Programming Algorithm for Decentralized Markov Decision Processes with a Broadcast Structure

Jeff Wu[1] Sanjay Lall[2]

In Proceedings of IEEE Conference on Decision and Control, pp. 6143–6148, 2010

Abstract

We give an optimal dynamic programming algorithm to solve a class of finite-horizon decentralized Markov decision processes (MDPs). We consider problems with a broadcast information structure that consists of a central node that only has access to its own state but can affect several outer nodes, while each outer node has access to both its own state and the central node's state, but cannot affect the other nodes. The solution to this problem involves a dynamic program similar to that of a centralized partially-observed Markov decision process.

1 Introduction and Prior Work

Decentralized control systems consist of several controllers having limited access to different sets of information. This is an attractive architecture for many types of interconnected systems. Unfortunately, solving for the optimal decentralized controller is in general very hard, even when only two controllers are involved [10, 6, 2]. Research on decentralized control has thus concentrated on finding special cases allowing for practical computation of the optimal controllers.

In this paper, we consider a class of Markov decision processes (MDPs), and we provide a method of dynamic programming for constructing the optimal decentralized controller. We call this class of problems *broadcast* MDPs. The solution technique turns out to be similar to methods used to solve centralized partially-observed MDPs (POMDPs).

Two related works deserve special mention. The first is a very recent paper by Swigart and Lall [9] that gives an explicit recursive solution for a two-player linear quadratic regulator, using spectral factorization techniques. This paper is important because it greatly reduces the complexity of solving for the optimal controllers, and also gives key insight on what the optimal controllers do. Our paper gives essentially the same type of result, except this time in a general broadcast MDP setting. Indeed, one can formally apply our MDP algorithm to the LQG setting to get the same results.

The second work is a paper by Mahajan, Nayyar, and Tenenketzis [5], which considers systems where each controller has an infinite memory to store common observations, but only a finite memory to store private observations. Like our paper, they then show that the problem can then be solved by centralized POMDP methods, although the size of the POMDP is exponential with the size of the finite memory. One of the main contributions of our paper is to show that for broadcast MDPs, we only need the most recent private observation to compute the optimal control, even if there is an infinite private memory available on each controller.

There are other classes of decentralized MDPs whose solution is significantly more tractable than the general case. These include systems where controllers share their information after a one-step time delay [4], and systems where the different subsystems evolve independently but whose cost function is coupled [1].

2 Notation

In this paper, we use a variant of MATLAB indexing notation that other authors have found convenient. We denote the $(x_t, x_{t+1}, \dots, x_\tau)$ as simply $x_{t:\tau}$. The subscripted indices will always refer to time.

We will also use superscripted indices, which will always denote player or node identity. Thus $x_t^{i:j} = (x_t^i, x_t^{i+1}, \dots, x_t^j)$ refers to some object or number given at time t across players i to j. To avoid confusion, we will *never* use a superscript to denote exponentiation of any kind.

When appropriate, we will omit parentheses around function arguments or drop the composition operator between functions. Thus if $F : X \to Y$, and $x \in X$, then we will sometimes denote $F(x)$ as simply Fx. This is especially handy when there are several compositions of functions, for example if $F : X \to Y$ and $G : Y \to Z$ and $H : Z \to W$, then we denote $H(G(F(x)))$ as simply $HGFx$, and the composition $H \circ G \circ F$ as HGF. Though this notation is like that of matrices, we emphasize that the functions F, G, or H need not be linear.

[1] J. Wu is with the Department of Electrical Engineering, Stanford University, Stanford, CA 94305, USA.
jeffwu78@stanford.edu

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

3 Broadcast MDPs : Definition

A **broadcast MDP** with M players is a collection of $2M$ stochastic processes $X = (X^1, \dots, X^M)$ and $U = (U^1, \dots, U^M)$, where X^i_t and U^i_t represent the **state** and **control action** of player i at time $t \in \mathbb{N}$. The distribution of (X, U) is determined by the following functions:

1. The **initial distributions** $p^i_0 : \mathcal{X}^i_0 \to \mathbb{R}$, where $p^i_0(x^i_0)$ represents the probability that initial state of player i is x^i_0.
2. The **transition laws** $p^1_{t+1} : \mathcal{X}^1_{t+1} \times \mathcal{X}^1_t \times \mathcal{U}^1_t \to \mathbb{R}$ for player 1, where
$$p^1_{t+1}(x^1_{t+1}|x^1_t, u^1_t)$$
represents the conditional probability that the next state is x^1_{t+1} given the current state and action are (x^1_t, u^1_t).
3. The **transition laws** $p^i_{t+1} : \mathcal{X}^i_{t+1} \times \mathcal{X}^1_t \times \mathcal{U}^1_t \times \mathcal{X}^i_t \times \mathcal{U}^i_t \to \mathbb{R}$ for players $i \geq 2$, where
$$p^i_{t+1}(x^i_{t+1}|x^1_t, u^1_t, x^i_t, u^i_t)$$
represents the conditional probability that the next state is x^i_{t+1} given the current state and action of player 1 is (x^1_t, u^1_t), and the current state and action for player i is (x^i_t, u^i_t).
4. The **control laws** $K^1_t : \mathcal{X}^1_0 \times \cdots \times \mathcal{X}^1_t \to \mathcal{U}^1_t$ for player 1, where $K^1_t(x^1_{0:t})$ represents the control action given the history of states is $x^1_{0:t}$.
5. The **control laws** $K^i_t : \mathcal{X}^1_0 \times \cdots \times \mathcal{X}^1_t \times \mathcal{X}^i_0 \times \cdots \times \mathcal{X}^i_t \to \mathcal{U}^i_t$ for players $i \geq 2$, where $K^i_t(x^1_{0:t}, x^i_{0:t})$ represents the control action given the history of states for player 1 and i are $(x^1_{0:t}, x^i_{0:t})$, respectively.

We call the sets $\mathcal{X}^i_t$ and $\mathcal{U}^i_t$ the **state** and **action spaces** for player i at time t, and assume the sets are all finite. The tuple $(M, \mathcal{X}, \mathcal{U}, p, K)$ is called the **parameters** of the broadcast MDP. We define the finite-dimensional distributions of X inductively as follows: The distribution of $X^{1:M}_0$ is

$$f_0(x^{1:M}_0) = \prod_{i=1}^{M} p^i_0(x^i_0) \tag{1}$$

Moreover, if f_t is the distribution of $X^{1:M}_{0:t}$, then the distribution of $X^{1:M}_{0:t+1}$ is

$$\begin{aligned} f_{t+1}(x^{1:M}_{0:t+1}) &= f_t(x^{1:M}_{0:t}) p^1_{t+1}(x^1_{t+1}|x^1_t, K^1_t(x^1_{0:t})) \\ &\times \prod_{i=2}^{M} p^i_{t+1}(x^i_{t+1}|x^1_t, K^1_t(x^1_{0:t}), x^i_t, K^i_t(x^1_{0:t}, x^i_{0:t})) \end{aligned} \tag{2}$$

It is easy to verify that these finite-dimensional distributions are consistent, and thus completely determine the distribution of X. The control action U^i_t is defined as

$$U^i_t = \begin{cases} K^1_t(X^1_{0:t}), & i = 1 \\ K^i_t(X^1_{0:t}, X^i_{0:t}), & i \geq 2 \end{cases}$$

so U is derived from X.

Let (X, U) be a broadcast MDP with parameters $(M, \mathcal{X}, \mathcal{U}, p, K)$. For each time t, let $c_t : \mathcal{X}^1_t \times \cdots \times \mathcal{X}^M_t \times \mathcal{U}^1_t \times \cdots \times \mathcal{U}^M_t \to \mathbb{R}$ be functions called **cost functions**. We then define the **expected cost** at the **time horizon** N to be

$$J = \sum_{t=0}^{N} E[c_t(X^{1:M}_t, U^{1:M}_t)]$$

The N**-horizon broadcast MDP problem** is then to find control laws $K^{1:M}_{0:N}$ that minimize J. The tuple $(N, M, \mathcal{X}, \mathcal{U}, p, c)$ specifies the **parameters** of the problem. Of course, when there is only $M = 1$ player, then a broadcast MDP reduces to a classic single-player MDP.

4 Optimal Controller Structure

We now prove a key structural result about the optimal control laws.

Theorem 1. *For any N-horizon broadcast MDP problem, there are optimal control laws where each control law for player $i \geq 2$ only depends on the history of player 1's states and player i's current state.*

Proof. Let $(N, M, \mathcal{X}, \mathcal{U}, p, c)$ be the parameters of the problem, and $K^{1:M}_{0:N}$ be a set of optimal control laws, which exist because there are only a finite number of control laws. We wish to transform K into the desired form without affecting optimality.

The simple idea behind the proof is to choose any player $i \geq 2$ and fix the control laws for the remaining players. Then the problem of finding the optimal control laws for player i reduces to a single-player MDP problem. We omit the details, but one can verify that the parameters of this single-player problem are $(N, 1, \tilde{\mathcal{X}}, \tilde{\mathcal{U}}, \tilde{p}, \tilde{c})$, where the state and action spaces are

$$\tilde{\mathcal{X}}_t = \mathcal{X}^1_0 \times \cdots \times \mathcal{X}^1_t \times \mathcal{X}^i_t, \quad \tilde{\mathcal{U}}_t = \mathcal{U}^i_t$$

and the initial distributions are

$$\tilde{p}_0(x^1_0, x^i_0) = p^1_0(x^1_0) p^i_0(x^i_0)$$

and the transition functions are

$$\begin{aligned} &\tilde{p}_{t+1}(x^1_{0:t+1}, x^i_{t+1}|y^1_{0:t}, x^i_t, u^i_t) \\ &\quad = \begin{cases} p^1_{t+1}(x^1_{t+1}|x^1_t, K^1_t(x^1_{0:t})) & \\ \times p^i_{t+1}(x^i_{t+1}|x^1_t, K^1_t(x^1_{0:t}), x^i_t, u^i_t) \quad , & x^1_{0:t} = y^1_{0:t} \\ 0, & \text{otherwise} \end{cases} \end{aligned}$$

Finally, the cost functions are

$$\begin{aligned} \tilde{c}_t(x^1_{0:t}, x^i_t, u^i_t) &= \sum_{x^{2:i-1}_{0:t}, x^{i+1:M}_{0:t}} \Big(\prod_{j \neq 1, i} f^j_t(x^j_{0:t}|x^1_{0:t-1}) \Big) \\ \times\, c_t(x^{1:M}_t, &K^1_t(x^1_{0:t}), K^2_t(x^1_{0:t}, x^2_{0:t}), \dots, K^{i-1}_t(x^1_{0:t}, x^{i-1}_{0:t}), \\ &u^i_t, K^{i+1}_t(x^1_{0:t}, x^{i+1}_{0:t}), \dots, K^M_t(x^1_{0:t}, x^M_{0:t})) \end{aligned}$$

where we define

$$f_t^j(x_{0:t}^j|x_{0:t-1}^1) = p_0^j(x_0^j)\dots p_t^j(x_t^j|x_{t-1}^1, K_{t-1}^1(x_{0:t-1}^1),\\ x_{t-1}^j, K_{t-1}^j(x_{0:t-1}^1, x_{0:t-1}^j))$$

We then apply the standard result for single-player MDPs which states that optimal control laws only need to depend on the current state. Thus without affecting optimality, we can replace K_t^i with a control law that only depends on the history of player 1's state and player i's current state. Repeating the argument for the other players $i \geq 2$ completes the proof. ■

Thus for players $i \geq 2$, we will henceforth restrict ourselves to control laws of the form $K_t^i : \mathcal{X}_0^1 \times \dots \times \mathcal{X}_t^1 \times \mathcal{X}_t^i \to \mathcal{U}_t^i$.

5 The Transition and Cost Operators

Given sets X and Y, we define $\mathcal{F}(X, Y)$ to be the set of functions mapping X to Y. We abbreviate the set of real-valued functions $\mathcal{F}(X, \mathbb{R})$ as simply $\mathcal{F}(X)$.

Let $(M, \mathcal{X}, \mathcal{U}, p, K)$ be parameters to a broadcast MDP. We define the **joint initial distribution** as the function $b_0 : \mathcal{X}_0^1 \times \dots \times \mathcal{X}_0^M \to \mathbb{R}$, where

$$b_0(x_0^{1:M}) = \prod_{i=1}^M p_0^i(x_0^i)$$

We define the **joint control law** at time t as $K_t = (K_t^1, \dots, K_t^M)$. This tuple has a dual use as a function, where we define

$$K_t(x_{0:t}^1, x_t^{2:M})\\ = (K_t^1(x_{0:t}^1), K_t^2(x_{0:t}^1, x_t^2), \dots, K_t^M(x_{0:t}^1, x_t^M))$$

Now given the joint control law K_t, define the **transition operator**

$$P_{t+1}(K_t) : \mathcal{F}(\mathcal{X}_0^1 \times \dots \times \mathcal{X}_t^1 \times \mathcal{X}_t^2 \times \dots \times \mathcal{X}_t^M) \to\\ \mathcal{F}(\mathcal{X}_0^1 \times \dots \times \mathcal{X}_{t+1}^1 \times \mathcal{X}_{t+1}^2 \times \dots \times \mathcal{X}_{t+1}^M)$$

where $b_{t+1} = P_{t+1}(K_t)b_t$ iff

$$b_{t+1}(x_{0:t+1}^1, x_{t+1}^{2:M}) = p_{t+1}^1(x_{t+1}^1|x_t^1, K_t^1(x_{0:t}^1))\\ \times \sum_{x_t^{2:M}} \Big(\prod_{i=2}^M p_{t+1}^i(x_{t+1}^i|x_t^1, K_t^1(x_{0:t}^1), x_t^i, K_t^i(x_{0:t}^1, x_t^i))\\ \times b_t(x_{0:t}^1, x_t^{2:M}) \Big) \quad (3)$$

for all $(x_{0:t}^1, x_t^{2:M})$. The meaning of the transition operator is given by the following lemma:

Lemma 2. *Let (X, U) be a broadcast MDP with parameters $(M, \mathcal{X}, \mathcal{U}, p, K)$. For each t, define*

$$b_t = P_t(K_{t-1}) \dots P_1(K_0)b_0$$

Then b_t is the distribution of $(X_{0:t}^1, X_t^{2:M})$.

Proof. The $t = 0$ case is clearly true by (1). Now suppose b_t is the distribution of $(X_{0:t}^1, X_t^{2:M})$. Let f_t be the distribution of $X_{0:t}^{1:M}$, so that distribution of $X_{0:t+1}^{1:M}$ is f_{t+1} given in (2), i.e.

$$f_{t+1}(x_{0:t+1}^{1:M}) = f_t(x_{0:t}^{1:M})p_{t+1}^1(x_{t+1}^1|x_t^1, K_t^1(x_{0:t}^1))\\ \times \prod_{i=2}^M p_{t+1}^i(x_{t+1}^i|x_t^1, K_t^1(x_{0:t}^1), x_t^i, K_t^i(x_{0:t}^1, x_t^i))$$

Now take the sum of both sides over the variables $x_{0:t}^{2:M}$. By definition of $P_{t+1}(K_t)$ given in (3), this gives

$$\sum_{x_{0:t}^{2:M}} f_{t+1}(x_{0:t+1}^{1:M}) = P_{t+1}(K_t)b_t = b_{t+1}$$

so b_{t+1} is the distribution of $(X_{0:t+1}^1, X_{t+1}^{2:M})$. The result follows by induction. ■

Now if c_t is the cost function and K_t are the control laws at time t, then we define the **cost operator**

$$C_t(K_t) : \mathcal{F}(\mathcal{X}_0^1 \times \dots \times \mathcal{X}_t^1 \times \mathcal{X}_t^2 \times \dots \times \mathcal{X}_t^M) \to \mathbb{R}$$

where

$$C_t(K_t)b_t\\ = \sum_{x_{0:t}^1, x_t^{2:M}} c_t(x_t^{1:M}, K_t(x_{0:t}^1, x_t^{2:M}))b_t(x_{0:t}^1, x_t^{2:M}) \quad (4)$$

for all b_t.

The transition and cost operators give us a classic expression for the expected cost.

Theorem 3. *Let $(M, \mathcal{X}, \mathcal{U}, p, K)$ be parameters of a broadcast MDP (X, U), and $c_0, \dots, c_N$ be associated cost functions. Then the expected cost at time horizon N is*

$$J = \sum_{t=0}^N C_t(K_t)P_t(K_{t-1}) \dots P_1(K_0)b_0 \quad (5)$$

Proof. By Lemma 2, $P_t(K_{t-1}) \dots P_1(K_0)b_0$ is the distribution of $(X_{0:t}^1, X_t^{2:M})$. Thus

$$E[c_t(X_t^{1:M}, U_t^{1:M})] = E[c_t(X_t^{1:M}, K_t(X_{0:t}^1, X_t^{2:M}))]\\ = C_t(K_t)P_t(K_{t-1}) \dots P_1(K_0)b_0$$

and the theorem follows. ■

We remark with the transition and cost operators appropriately defined, the formula for the cost given in (5) holds even for the most general decentralized MDPs.

6 Dynamic Programming

The formula given in (5) suggests the following recursive structure for the expected cost.

Theorem 4. *Let $(N, M, \mathcal{X}, \mathcal{U}, p, c)$ be parameters of a finite-horizon broadcast MDP problem. Then given the control laws K, define the* ***value functions***

$$V_t : \mathcal{F}(\mathcal{X}_0^1 \times \cdots \times \mathcal{X}_t^1 \times \mathcal{X}_t^2 \times \cdots \times \mathcal{X}_t^M) \to \mathbb{R}$$

by the backward recursion

$$V_N = C_N(K_N) \tag{6}$$
$$V_t = V_{t+1}P_{t+1}(K_t) + C_t(K_t) \tag{7}$$

Then the expected cost at the time horizon N is $J = V_0 b_0$.

Proof. Expanding the recurrence gives

$$V_t = \sum_{\tau=t}^{N} C_\tau(K_\tau)P_\tau(K_{\tau-1})\dots P_{t+1}(K_t)$$

for each t. Thus $J = V_0 b_0$ by Theorem 3. ■

The value functions V_t have a natural partial ordering. We say that $V_t \preceq W_t$ if $V_t b \leq W_t b$ for all nonnegative functions b. Those familiar with Markov decision theory know that for *single-player MDPs*, it is possible to choose control laws to minimize the value functions V_t over this partial order by applying the recursion

$$V_N = \min_{K_N} C_N(K_N)$$
$$V_t = \min_{K_t}[V_{t+1}P_{t+1}(K_t) + C_t(K_t)]$$

with the optimal control laws are those achieving these minimums. The minimums are possible because there is a separate variable in K_t corresponding to each coordinate in the representation of the linear function V_t.

In the decentralized or partially observed case, however, some players that do not have access to the global state of the system, so in general we *cannot* minimize the value functions in this way. We must settle for something much weaker and computationally harder.

Theorem 5. *Let $(N, M, \mathcal{X}, \mathcal{U}, p, c)$ be parameters of a finite-horizon broadcast MDP problem. Define the* ***optimal value functions*** *V_t^* by the backward recursion*

$$V_N^* b_N = \min_{K_N}(C_N(K_N)b_N), \quad \text{for all } b_N \tag{8}$$
$$V_t^* b_t = \min_{K_t}(V_{t+1}^* P_{t+1}(K_t)b_t + C_t(K_t)b_t), \quad \text{for all } b_t \tag{9}$$

and define the ***Q-functions*** *$Q_t(K_t)$ by the equations*

$$Q_N(K_N) = C_N(K_N)$$
$$Q_t(K_t) = V_{t+1}^* P_{t+1}(K_t) + C_t(K_t)$$

Let b_0 be the joint initial distribution, and choose control laws $K_0^*, \dots, K_N^*$ according to the forward recursion

$$K_t^* \in \operatorname*{argmin}_{K_t} Q_t(K_t)b_t$$
$$b_{t+1} = P_{t+1}(K_t^*)b_t$$

Then the control laws K^* are optimal, i.e. they minimize the expected cost over the time horizon N.

Proof. By Theorem 4, the expected cost given the control laws $K_0, \dots, K_N$ is $V_0 b_0$, where V_t is defined by the recursion

$$V_N = C_N(K_N)$$
$$V_t = V_{t+1}P_{t+1}(K_t) + C_t(K_t)$$

Let Φ_t be the map from $(K_t, \dots, K_N)$ to V_t, so that $\Phi_0(K_0, \dots, K_N)b_0$ is the expected cost. We claim that

$$V_t^* b_t = \min_{K_t,\dots,K_N} \Phi_t(K_t, \dots, K_N)b_t$$

This is clearly true for $t = N$, since $\Phi_N(K_N) = C_N(K_N)$. Moreover, if the claim is true for time $t + 1$, then

$$\begin{aligned}
&\min_{K_t,\dots,K_N} \Phi_t(K_t, \dots, K_N)b_t \\
&\quad = \min_{K_t}\Big(\min_{K_{t+1},\dots,K_N}(\Phi_{t+1}(K_{t+1}, \dots, K_N)P_{t+1}(K_t)b_t) \\
&\qquad + C_t(K_t)b_t\Big) \\
&\quad = \min_{K_t}(V_{t+1}^* P_{t+1}(K_t)b_t + C_t(K_t)b_t) \\
&\quad = V_t^* b_t
\end{aligned}$$

so it is claim is true for time t. The claim follows by induction.

The above equations also show that if $t < N$ and we wish to minimize $\Phi_t(K_t, \dots, K_N)b_t$, we can first choose K_t to minimize

$$V_{t+1}^* P_{t+1}(K_t)b_t + C_t(K_t)b_t = Q_t(K_t)b_t$$

and then choose $K_{t+1}, \dots, K_N$ to minimize

$$\Phi_{t+1}(K_{t+1}, \dots, K_N)P_{t+1}(K_t)b_t$$

Of course, when $t = N$, we just need to choose K_N to minimize

$$\Phi_N(K_N)b_N = C_N(K_N)b_N = Q_N(K_N)b_N$$

Thus to minimize the expected cost $\Phi_0(K_0, \dots, K_N)b_0$, first choose K_0 to minimize $Q_0(K_0)b_0$. Then choose $(K_1, \dots, K_N)$ to be a minimum of $\Phi_1(K_1, \dots, K_N)b_1$, where $b_1 = P_1(K_0)b_0$, and keep going until we reach time N. But this precisely describes the control laws $(K_0^*, \dots, K_N^*)$, and so K^* minimizes the expected cost as desired. ■

The last theorem actually holds for quite general decentralized Markov decision processes if the transition and cost operators are appropriately defined. It is important to note that the optimal value functions V_t^* are highly nonlinear, and they are in fact piecewise-linear concave (i.e. a pointwise minimum of a finite number of linear functions). Techniques for how to effectively compute such representations via linear programming can be found in [8, 3]. Despite these techniques, however, the problem of computing a representation of the optimal value functions is still very difficult [7].

7 Simplifying the Dynamic Program

While very useful conceptually, the dynamic programming algorithm in Theorem 5 does not take advantage of the structure within a broadcast MDP. As we shall see in this section, the dynamic program can be greatly simplified if we factor out player 1's state history, which is available to all players.

Let $(N, M, \mathcal{X}, \mathcal{U}, p, c)$ be parameters of a broadcast MDP problem. We call any element of the set

$$\mathcal{U}_t^1 \times \mathcal{F}(\mathcal{X}_t^2, \mathcal{U}_t^2) \times \cdots \times \mathcal{F}(\mathcal{X}_t^M, \mathcal{U}_t^M)$$

a **partial control law** at time t. Like the joint control law, a partial control law $k_t = (k_t^1, k_t^2, \dots, k_t^M)$ has a dual use as a function, where we define

$$k_t(x_t^{2:M}) = (k_t^1, k_t^2(x_t^2), \dots, k_t^M(x_t^M))$$

Given $x_t^1 \in \mathcal{X}_t^1$ and a partial control law k_t at time t, we define the **partial transition operator**

$$\mathcal{P}_{t+1}(x_t^1, k_t) : \mathcal{F}(\mathcal{X}_t^2 \times \cdots \times \mathcal{X}_t^M) \to \mathcal{F}(\mathcal{X}_{t+1}^2 \times \cdots \times \mathcal{X}_{t+1}^M)$$

where $\beta_{t+1} = \mathcal{P}_{t+1}(x_t^1, k_t)\beta_t$ iff

$$\beta_{t+1}(x_{t+1}^{2:M}) = \sum_{x_t^{2:M}} \prod_{i=2}^{M} p_{t+1}^i(x_{t+1}^i | x_t^1, k_t^1, x_t^i, k_t^i(x_t^i))\beta_t(x_t^{2:M})$$

We also define the **partial cost operator**

$$\mathcal{C}_t(x_t^1, k_t) : \mathcal{F}(\mathcal{X}_t^2 \times \cdots \times \mathcal{X}_t^M) \to \mathbb{R}$$

where for all $\beta_t \in \mathcal{F}(\mathcal{X}_t^2 \times \cdots \times \mathcal{X}_t^M)$, we have

$$\mathcal{C}_t(x_t^1, k_t)\beta_t = \sum_{x_t^{2:M}} c_t(x_t^{1:M}, k_t(x_t^{2:M}))\beta_t(x_t^{2:M})$$

The partial transition operator and cost operators are related to the full transition and cost operators by the following lemma.

Lemma 6. *Let $(N, M, \mathcal{X}, \mathcal{U}, p, c)$ be parameters of a broadcast MDP problem. Let K_t be any joint control law at time t. Then $b_{t+1} = P_{t+1}(K_t)b_t$ iff*

$$b_{t+1}(x_{0:t+1}^1, \cdot) = p_{t+1}^1(x_{t+1}^1 | x_t^1, K_t^1(x_{0:t}^1)) \times \mathcal{P}_{t+1}(x_t^1, K_t(x_{0:t}^1, \cdot))b_t(x_{0:t}^1, \cdot)$$

for all $x_{0:t+1}^1$. Moreover, we have

$$C_t(K_t)b_t = \sum_{x_{0:t}^1} \mathcal{C}_t(x_t^1, K_t(x_{0:t}^1, \cdot))b_t(x_{0:t}^1, \cdot)$$

Proof. Follows immediately from the definitions. ■

Given $x_t^1 \in \mathcal{X}_t^1$, define the **partial value functions** $\mathcal{V}_t^*(x_t^1) : \mathcal{F}(\mathcal{X}_t^2 \times \cdots \times \mathcal{X}_t^M) \to \mathbb{R}$ where

$$\mathcal{V}_N^*(x_N^1)\beta_N = \min_{k_N} \mathcal{C}_N(x_N^1, k_N)\beta_N, \quad \text{for all } \beta_N$$

and for $t < N$,

$$\mathcal{V}_t^*(x_t^1)\beta_t = \min_{k_t}\Big[\sum_{x_{t+1}^1}\Big(p_{t+1}^1(x_{t+1}^1 | x_t^1, k_t^1) \times \mathcal{V}_{t+1}^*(x_{t+1}^1)\mathcal{P}_{t+1}(x_t^1, k_t)\beta_t\Big) + \mathcal{C}_t(x_t^1, k_t)\beta_t\Big], \quad \text{for all } \beta_t$$

Define also the **partial Q-functions** $\mathcal{Q}_t(x_t^1, k_t)$, where

$$\mathcal{Q}_N(x_N^1, k_N) = \mathcal{C}_N(x_N^1, k_N)$$

and for $t < N$,

$$\mathcal{Q}_t(x_t^1, k_t) = \mathcal{C}_t(x_t^1, k_t) + \sum_{x_{t+1}^1} p_{t+1}^1(x_{t+1}^1 | x_t, k_t^1)\mathcal{V}_{t+1}^*(x_{t+1}^1)\mathcal{P}_{t+1}(x_t^1, k_t)$$

It is easy to show that partial value and Q-functions are also piecewise-linear concave. Most importantly, the domain of the partial value functions or Q-functions does not *grow* as time progresses, which makes representing them more practical. Moreover, we can relate to the optimal value functions defined in Theorem 5 by the following lemma.

Lemma 7. *Let $(N, M, \mathcal{X}, \mathcal{U}, p, c)$ be parameters of a broadcast MDP problem. Then*

$$Q_t(K_t)b_t = \sum_{x_{0:t}^1} \mathcal{Q}_t(x_t^1, K_t(x_{0:t}^1, \cdot))b_t(x_{0:t}^1, \cdot) \tag{10}$$

$$V_t^* b_t = \sum_{x_{0:t}^1} \mathcal{V}_t^*(x_t^1)b_t(x_{0:t}^1, \cdot) \tag{11}$$

for any K_t and b_t.

Proof. First note that if (10) holds, then so does (11), since

$$\begin{aligned} V_t^* b_t &= \min_{K_t} Q_t(K_t)b_t \\ &= \min_{K_t} \sum_{x_{0:t}^1} \mathcal{Q}_t(x_t^1, K_t(x_{0:t}^1, \cdot))b_t(x_{0:t}^1, \cdot) \\ &= \sum_{x_{0:t}^1} \min_{K_t(x_{0:t}^1, \cdot)} \mathcal{Q}_t(x_t^1, K_t(x_{0:t}^1, \cdot))b_t(x_{0:t}^1, \cdot) \\ &= \sum_{x_{0:t}^1} \mathcal{V}_t^*(x_t^1)b_t(x_{0:t}^1, \cdot) \end{aligned}$$

Now by Lemma 6, we have

$$\begin{aligned} Q_N(K_N)b_N &= C_N(K_N)b_N \\ &= \sum_{x_{0:N}^1} \mathcal{C}_N(x_N^1, K_N(x_{0:N}^1, \cdot))b_N(x_{0:N}^1, \cdot) \\ &= \sum_{x_{0:N}^1} \mathcal{Q}_N(x_N^1, K_N(x_{0:N}^1, \cdot))b_N(x_{0:N}^1, \cdot) \end{aligned}$$

so the result is true for time $t = N$. Moreover, if the result is true for time $t+1$, then by Lemma 6 again, we have

$$\begin{aligned} & Q_t(K_t)b_t \\ &\quad = V^*_{t+1}P_{t+1}(K_t)b_t + C_t(K_t)b_t \\ &\quad = \sum_{x^1_{0:t+1}} \Big(p^1_{t+1}(x^1_{t+1}|x^1_t, K^1_t(x^1_{0:t})) \\ &\qquad \times \mathcal{V}^1_{t+1}(x^1_t, K_t(x^1_{0:t},\cdot))\mathcal{P}_{t+1}(x^1_t, K_t(x^1_{0:t},\cdot))b_t(x^1_{0:t},\cdot)\Big) \\ &\qquad + \sum_{x^1_{0:t}} \mathcal{C}_t(x^1_t, K_t(x^1_{0:t},\cdot))b_t(x^1_{0:t},\cdot) \\ &\quad = \sum_{x^1_{0:t}} \mathcal{Q}_t(x^1_t, K_t(x^1_{0:t},\cdot))b_t(x^1_{0:t},\cdot) \end{aligned}$$

so the result is true for time t. The result follows by induction. ■

We now can present the dynamic programming algorithm for broadcast MDPs in its simplified form. This version uses only the partial Q-functions, which are much easier to compute than the original Q-functions.

Theorem 8. *Let $(N, M, \mathcal{X}, \mathcal{U}, p, c)$ be parameters of a broadcast MDP problem. Let $K_0, \dots, K_N$ be control laws such that given any $x^1_{0:N} \in \mathcal{X}^1_0 \times \cdots \times \mathcal{X}^1_N$, there is a sequence $\beta_0, \dots, \beta_N$ satisfying the recursion*

$$\begin{aligned} \beta_0(x^{2:M}_0) &= \prod_{i=2}^{M} p^i_0(x^i_0) \quad \textit{for all } x^{2:M}_0 \\ K_t(x^1_{0:t},\cdot) &\in \operatorname*{argmin}_{k_t} \mathcal{Q}_t(x^1_t, k_t)\beta_t \\ \beta_{t+1} &= \mathcal{P}_{t+1}(x^1_t, K_t(x^1_{0:t},\cdot))\beta_t \end{aligned}$$

Then $K_0, \dots, K_N$ is optimal.

Proof. By Theorem 5, if

$$\begin{aligned} b_0(x^{1:M}_0) &= \prod_{i=1}^{M} p^i_0(x^i_0), \text{ for all } x^{1:M}_0 \\ K^*_t &\in \operatorname*{argmin}_{K_t} Q_t(K_t)b_t \\ b_{t+1} &= P_{t+1}(K^*_t)b_t \end{aligned}$$

for each t, then $K^*_0, \dots, K^*_N$ is optimal. We now show that if the control laws satisfy the conditions of this theorem, they will satisfy the above conditions as well, thus guaranteeing optimality.

We first note that for any x^1_0, $b_t(x^1_0,\cdot)$ is a nonnegative scalar multiple of β_0. Now suppose that for each $x^1_{0:t}$, $b_t(x^1_{0:t},\cdot)$ is a nonnegative scalar multiple of the computed β_t. Now by Lemma 7, we have

$$Q_t(K_t)b_t = \sum_{x^1_{0:t}} \mathcal{Q}_t(x^1_t, K_t(x^1_{0:t},\cdot))b_t(x^1_{0:t},\cdot)$$

so if for each $x^1_{0:t}$, we choose $K_t(x^1_{0:t},\cdot)$ so that it is a minimum of $\mathcal{Q}_t(x^1_t, K_t(x^1_{0:t},\cdot))b_t(x^1_{0:t},\cdot)$, then this will also minimize the function $Q_t(K_t)b_t$. Moreover, by positive homogeneity of $\mathcal{Q}_t(x^1_t, K_t(x^1_{0:t},\cdot))$, any $K_t(x^1_{0:t},\cdot)$ that minimizes $\mathcal{Q}_t(x^1_t, K_t(x^1_{0:t},\cdot))\beta_t$ will also minimize $\mathcal{Q}_t(x^1_t, K_t(x^1_{0:t},\cdot))b_t(x^1_{0:t},\cdot)$.

Finally, from Lemma 6 we have $b_{t+1} = P_{t+1}(K_t)b_t$ iff

$$\begin{aligned} b_{t+1}(x^1_{0:t+1},\cdot) = {} & p^1_{t+1}(x^1_{t+1}|x^1_t, K^1_t(x^1_{0:t})) \\ & \times \mathcal{P}_{t+1}(x^1_t, K_t(x^1_{0:t},\cdot))b_t(x^1_{0:t},\cdot) \end{aligned}$$

Thus if the $b_t(x^1_{0:t},\cdot)$ is a nonnegative scalar multiple of β_t, then $b_{t+1}(x^1_{0:t+1},\cdot)$ is a nonnegative scalar multiple of $\beta_{t+1} = \mathcal{P}_{t+1}(x^1_t, K_t(x^1_{0:t},\cdot))\beta_t$. The theorem follows by induction. ■

References

[1] R. Becker, S. Zilberstein, V. Lesser, and C. V. Goldman. Solving transition independent decentralized Markov decision processes. *Journal of Artificial Intelligence Research*, 22:423–455, 2004.

[2] D. 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.

[3] A. R. Cassandra, M. L. Littman, and N. L. Zhang. Incremental pruning: A simple, fast, exact method for partially observable Markov decision processes. *Proceedings of the 13th Conference on Uncertainty in Artificial Intelligence*, 2008.

[4] K. Hsu and S. I. Marcus. Decentralized control of finite state Markov processes. *IEEE Transactions on Automatic Control*, 27(2):426–431, 1982.

[5] A. Mahajan, A. Nayyar, and D. Tenenketzis. Identifying tractable decentralized control problems on the basis of information structure. *Proceedings of the 46th Allerton Conference*, pages 1440–1449, 2008.

[6] C. H. Papadimitriou and J. N. Tsitsiklis. Intractable problems in control theory. *SIAM Journal of Control and Optimization*, 24(4):639–654, 1986.

[7] C. H. Papadimitriou and J. N. Tsitsiklis. The complexity of Markov decision processes. *Mathematics of Operations Research*, 12(3):441–450, 1987.

[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] J. Swigart and S. Lall. An explicit state-space solution for a decentralized two-player linear-quadratic regulator. *2010 American Control Conference*, submitted.

[10] H. S. Witsenhausen. A counterexample in stochastic optimal control. *SIAM Journal of Control*, 6(1):131–147, 1968.