Dynamic Programming for POMDP with Jointly Discrete and Continuous State-Spaces

Donghwan Lee, Niao He, and Jianghai Hu

Abstract— **In this work, we study dynamic programming (DP) algorithms for partially observable Markov decision processes with jointly continuous and discrete state-spaces. We consider a class of stochastic systems which have coupled discrete and continuous systems, where only the continuous state is observable. Such a family of systems includes many real-world systems, for example, Markovian jump linear systems and physical systems interacting with humans. A finite history of observations is used as a new information state, and the convergence of the corresponding DP algorithms is proved. In particular, we prove that the DP iterations converge to a certain bounded set around an optimal solution. Although deterministic DP algorithms are studied in this paper, it is expected that this fundamental work lays foundations for advanced studies on reinforcement learning algorithms under the same family of systems.**

I. Introduction

The goal of this paper is to study optimal control problems for partially observable Markov decision processes (POMDPs) with jointly continuous and discrete state-spaces. Optimal control designs for stochastic systems have been a fundamental research field for a long time [1]. Classical and popular approaches include, for example, the linear quadratic Gaussian control and stochastic model predictive control, where the stochastic model predictive control computes a suboptimal control policy with predictions of finite future trajectories. In this paper, we focus on stochastic systems with a special structure where the continuous state-space and discrete state-space coexist and interact with each other. The state in the discrete state-space evolves according to a Markov chain which depends on the state of the control system with the continuous state-space. The overall system can be viewed as a Markov decision process (MDP) [1] with coupled continuous and discrete state spaces. Such classes of systems arise in several stochastic control applications including

- vehicle path-planning [2], [3];
- Markovian jump linear systems with examples of macroeconomic model [4] and economic models of government expenditure [5];
- vehicle controls with driver's behavior models [6], [7];
- building climate control with human interactions [8]–[11];
- hybrid electric vehicle powertrain management [12], [13];

to name just a few.

The main contribution of this paper is the first formal study of dynamic programming formulation and its convergence results for the systems mentioned above with an additional assumption of the partial state observability, i.e., the discrete state is unobservable. An example includes control systems with human interactions, where unobservable human cognition and behaviors are modelled as discrete Markov chains or Markov decision processes. In particular, this class of systems arises in building management problems with human interactions [11]. To improve the control performance, a finite observation history is considered for the output-feedback control policy with its performance analysis. The use of finite observation histories for POMDPs is a common practice in the reinforcement learning literature [14], [15]. However, to our knowledge, there exists no attempt to analyze its sub-optimality so far. The proposed results build upon the previous work [11], [12]. Although deterministic DP algorithms are studied in this paper, this fundamental work lays foundations for advanced studies on reinforcement learning algorithms under the same family of systems.

A. Related Work

As long as a sufficient number of random samples of states can be obtained, the scenario-based (or sample-based) approximation approach, e.g., [2], [3], [16]–[19], can design a control policy for complex stochastic systems with generic probability distributions of uncertainties. It was successfully applied to robot path-planning problems in [2], [3] and aircraft conflict detection in [16]. In the scenario-based approach, the process related to uncertainties and the evolution of the control system trajectories are not fully coupled. For instance, in [2], [3], the uncertainties of the continuous system are affected by the discrete system's state, while the latter does not depend on the continuous system's state.

Fully coupled systems similar to those in this paper were studied in [12], [13] for hybrid electric vehicle powertrain management problems. They are fully coupled in the sense that both continuous and discrete system states affect each other. They considered approximate dynamic programming [20] (or reinforcement learning [21] from the machine learning context). Compared to [12], the control systems in this paper have additional continuous stochastic disturbances, while [12] only addresses discrete stochastic disturbances that depend on the discrete state of the Markov chain. Reinforcement learning algorithms were adopted in [10], [11]

This material is based upon work supported by the National Science Foundation under Grant No. 1539527

D. Lee and N. He are with Coordinated Science Laboratory (CSL), University of Illinois at Urbana-Champaign, IL 61801, USA donghwan@illinois.edu, niaohe@illinois.edu.

J. Hu is with the Department of Electrical and Computer Engineering, Purdue University, West Lafayette, IN 47906, USA jianghai@purdue.edu.

for building management systems with occupant interactions. Compared to [11], we consider more general Markov chain models for the discrete state-space evolution.

II. Preliminaries and Problem Formulation

A. Noataion

Throughout the paper, the following notations will be used: $\mathbb{N}$ and $\mathbb{N}_+$: sets of nonnegative and positive integers, respectively; $\mathbb{R}$: set of real numbers; $\mathbb{R}^n$: n-dimensional Euclidean space; $\mathbb{R}^{n\times m}$: set of all $n\times m$ real matrices; A^T: transpose of matrix A; $\mathbb{S}^n$ (resp. $\mathbb{S}^n_+$, $\mathbb{S}^n_{++}$): set of symmetric (resp. positive semi-definite, positive definite) $n\times n$ matrices; $|S|$: cardinality of a finite set S; $\mathbb{E}[\cdot]$: expectation operator; $\mathbb{P}[\cdot]$: probability of an event; $\mathrm{diam}(C)$: diameter of a set C in $\mathbb{R}^n$, i.e., $\mathrm{diam}(C) := \sup\{\|s-s'\|_2 : s,s' \in C\}$; for any vector x, $[x]_i$ is its i-th element; for any matrix P, $[P]_{ij}$ indicates its element in i-th row and j-th column; matrix $P\in\mathbb{R}^{n\times n}$ is called a (row) stochastic matrix if its row sums are one; $x\in\mathbb{R}^n$ is called a stochastic vector if its column sum is one; if $\mathbf{z}$ is a discrete random variable which has n values and $\mu\in\mathbb{R}^n$ is a stochastic vector, then $\mathbf{z}\sim\mu$ stands for $\mathbb{P}[\mathbf{z}=i]=[\mu]_i$ for all $i\in\{1,\ldots,n\}$; if $\mathbf{z}$ is a continuous random variable with the density function $\rho(\cdot)$, we denote $\mathbf{z}\sim\rho(\cdot)$; Δ_k with $k\in\mathbb{N}_+$: unit simplex defined as $\Delta_k := \{(\alpha_1,\ldots,\alpha_k) : \alpha_1+\cdots+\alpha_k=1, \alpha_i\geq 0, \forall i\in\{1,\ldots,k\}\}$; w.r.t: abbreviation for "with respect to." Throughout the paper, random variables will be highlighted by boldface fonts, while the corresponding realizations will be written by plain fonts.

B. Markov Decision Process

In this paper, we consider a discrete-time Markov decision process (MDP) [20] defined as a tuple $\langle X, S, U, p_{\mathbf{x}}, p_{\mathbf{s}}, p_{\mathbf{r}}, \mathbf{r}, \gamma\rangle$, where X is a continuous compact state-space, S is a discrete finite state-space, U is a continuous or discrete action-space, $p_{\mathbf{x}}(x'|x,s,u)$ defines a continuous state transition probability density function from the current state $x\in X$ to the next state $x'\in X$ under the action $u\in U$, and $s\in S$, $p_{\mathbf{s}}(s'|s,x)$ defines the discrete state transition probability mass function from the current state $s\in S$ to the next state $s'\in S$ under the current continuous state x, $\mathbf{r}: X\times S\times U\to\mathbb{R}$ is a stochastic reward function with its expectation $\mathbb{E}[\mathbf{r}(x,s,u)] = R(x,s,u)$ and density function $p_{\mathbf{r}}(\mathbf{r}|x,s,u)$ given (x,s,u), and $\gamma\in[0,1)$ is called the discount factor. We assume that the expected reward is bounded. For simplicity, we only consider the reward function $\mathbf{r}: X\times U\to\mathbb{R}$ which depends on the continuous state x and action u.

Assumption 1: The expected reward R satisfies $R(x,s)\in[0,M]$ for all $(x,s)\in X\times U$, where $M\in\mathbb{R}_{++}$.

The overall system can be expressed as

$$\begin{cases}\mathbf{x}(k+1)\sim p_{\mathbf{x}}(\cdot|\mathbf{x}(k),\mathbf{s}(k),\mathbf{u}(k)), & \mathbf{x}(0)\sim\rho_{\mathbf{x}}(\cdot)\\ \mathbf{s}(k+1)\sim p_{\mathbf{s}}(\cdot|\mathbf{s}(k),\mathbf{x}(k)), & \mathbf{s}(0)\sim\rho_{\mathbf{s}}(\cdot)\end{cases} \tag{1}$$

where $k\in\mathbb{N}$ is the time step, $\mathbf{x}(k)\in X$ is the continuous state at time k, $\mathbf{u}(k)\in U$ is the control input, and $\mathbf{s}(k)\in S$ is the discrete state, $\rho_{\mathbf{x}}$ is the initial distribution of $\mathbf{x}(0)$, and $\rho_{\mathbf{s}}$ is the initial distribution of $\mathbf{s}(0)$. In this paper, the discrete state transition $p_{\mathbf{s}}(s'|s,x)$ is represented by a Markov chain with and the state transition matrix $P(x)$ parameterized by $x\in X$. A visual description of the system is given in Figure 1. If we define the augmented state

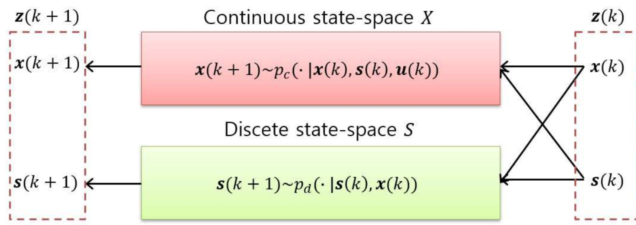

Fig. 1. Overall MDP structure.

$\mathbf{z}(k) := \begin{bmatrix}\mathbf{x}(k)^T & \mathbf{s}(k)^T\end{bmatrix}^T$, then (1) can be formulated by the single Markov decision process (MDP) with the transition probability density $p_{\mathbf{z}}(z'|z,u)$

$$\mathbf{z}(k+1)\sim p_{\mathbf{z}}(\cdot|\mathbf{z}(k),\mathbf{u}(k)),\quad \mathbf{z}(0)\sim\rho_{\mathbf{z}}(\cdot) \tag{2}$$

where the continuous and discrete state-spaces coexist and interact with each other.

C. Information State

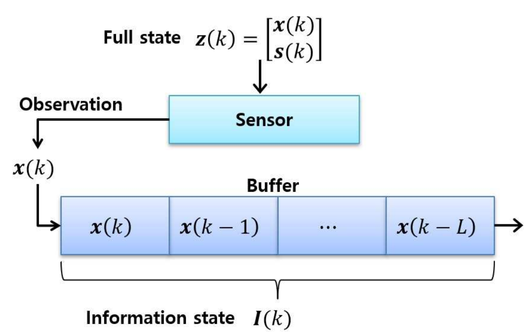

Fig. 2. An example of information state

In real applications, the full-state information is usually not available. In this paper, we adopt the following assumption.

Assumption 2: $\mathbf{x}(k)$ is measured in real time, while $\mathbf{s}(k)$ cannot be measured.

When the state is partially observable, the observation $\mathbf{x}(k)$ loses the Markov property [22, pp. 63] in general, i.e., its transition usually depends on all past history of $\mathbf{x}(k)$ and the current action $\mathbf{u}(k)$. We formally define this property.

Definition 1 (Markov property): The process $(\mathbf{I}(k),\mathbf{u}(k))_{k=1}^{\infty}$ is said to satisfy the Markov property if

$$\mathbf{I}(k+1)\sim\mathbb{P}[\mathbf{I}(k+1)=\cdot|\mathbf{I}(k)=I(k),\mathbf{u}(k)=u(k)]$$
$$=\mathbb{P}[\mathbf{I}(k+1)=\cdot|\mathbf{I}(i)=I(i), i=0,\ldots,k,\mathbf{u}(k)=u(k)].$$

Define the space of information $\mathcal{I}$ and let $\mathbf{I}(k)\in\mathcal{I}$ be an available information at time k, called the information state.

In other words, $\mathbf{I}(k)$ is a general form of the observation and can include artificially crafted information from the pure observation $\mathbf{x}(k)$. For example, the information is the current continuous state $\mathbf{x}(k)$, then the information space is X. Another example of the information state is the finite memory information state.

Definition 2 (Finite memory information state): The finite memory information state at time k is defined as

$$\mathbf{I}(k) =: (\mathbf{x}(k), \mathbf{x}(k-1), \dots, \mathbf{x}(k-L)) \in \mathcal{I} = X^{L+1}.$$

Figure 2 illustrates its concept and how to implement it in practice. It is known that the finite memory information structure in Definition 2 can alleviate the problem related to the POMDP [14]. In particular, roughly speaking, it can reduce the sensitivity of the next state's distribution with respect to the current state.

D. Output-Feedback Policy

A deterministic control policy $\pi : \mathcal{I} \to U$ is a map from the information space $\mathcal{I}$ to the control space U. In this paper, the set of all admissible control policies is denoted by Π. In addition, the sequence of control policies $(\pi_0, \pi_1, \dots) \in \Pi^\infty$ is denoted by $\bar{\pi}$. If $\pi_0 = \pi_1 = \dots = \pi$, then $\bar{\pi}$ or π is called a stationary control policy. In this case, $\bar{\pi}$ will be simply denoted by π if there exists no confusion. For the MDP, the episode is defined as a single realization of the state-action-reward trajectory.

Definition 3 (Episode): An episode of the MDP (1) under any policy $\bar{\pi}$ is defined as the process $(\mathbf{x}(k), \mathbf{s}(k), \mathbf{u}(k), \mathbf{r}(\mathbf{x}(k), \mathbf{u}(k)))_{k=0}^{\tau}$, where τ is the random stopping time.

In Definition 3, τ may be finite or not. In this paper, the stopping time is defined as the final time step before the first time instant that continuous state exists X.

Assumption 3 (Stopping time): The stopping time τ is defined as the first time step such that $\mathbf{x}(\tau + 1) \notin X$.

Assumption 3 is useful when we consider X which is a strict subset of $\mathbb{R}^n$, where n is the dimension of X. This is the case if we use a linear function approximation [23] which locally approximates the value function or consider a compact X to guarantee the stability of approximate dynamic programming algorithms. Since the stopping time depends on the initial state $\mathbf{I}(0)$ and the policy $\bar{\pi}$, it will be denoted by $\tau(\mathbf{I}(0); \bar{\pi})$ throughout the paper, while for brevity, τ will be used when it is clear from the context. When $\mathbf{x}(0) \notin X$, we set $\tau = -1$.

E. Problem Statement

For $I \in \mathcal{I}$, the value associated with a given $\bar{\pi} \in \Pi^\infty$ is defined as

$$J^{\bar{\pi}}(I) := \mathbb{E}\left[\left. \sum_{i=0}^{\tau(\mathbf{I}(0);\bar{\pi})} \gamma^i \mathbf{r}(\mathbf{x}(i), \pi_i(\mathbf{I}(i))) \right| \mathbf{I}(0) = I \right], \quad (3)$$

where $\tau(\mathbf{I}(0); \bar{\pi})$ is the stopping time given $\mathbf{I}(0)$ and $\bar{\pi}$, and the expectation is taken with respect to the episode. The optimal control design problem is stated as follows.

Problem 1 (Optimal Decision): Consider the finite memory information state in Definition 2. Find π such that

$$\pi^*(I) := \arg\inf_{\pi \in \Pi} J^\pi(I),$$

for all $I \in \mathcal{I}$.

In this paper, the optimal cost will be denoted by $J^*(I) := J^{\pi^*}(I)$.

III. Dynamic Programming with Markov Property

In this section, we study dynamic programming (DP) approaches to find the optimal cost function and its corresponding stationary control policy under the Markov assumption. In this section, we consider the finite memory information state in Definition 2.

Assumption 4: The process $(\mathbf{I}(k), \mathbf{u}(k))_{k=0}^{\tau}$ satisfies the Markov property.

In other words, $(\mathbf{I}(k), \mathbf{u}(k))_{k=0}^{\tau}$ is determined based on an MDP with the transition density function $p_{\mathbf{I}}(I'|I, u)$ from the current information I to I' given u. Under Assumption 1, the quantity (3) is always finite, and hence well defined. The first property of J^* is its boundedness on $\mathcal{I}$.

Proposition 1: $J^* \le M/(1-\gamma)$ on $\mathcal{I}$.

Proof: By using the definition (3) and Assumption 1, we have $J^*(I) \le \sum_{i=0}^{\infty} \gamma^i M = M/(1-\gamma)$. ■

Typical DP approaches [20] convert Definition 3 into a fixed point problem of a mapping called the Bellman operator. For a given $\pi \in \Pi$, we also define the following Bellman operator:

$$(T_\pi J_0)(I) := R(x(0), \pi(I)) + \gamma \mathbb{E}[\mathbb{I}_X(\mathbf{x}(1)) J_0(\mathbf{I}(1)) | \mathbf{I}(0) = I, \mathbf{u}(0) = \pi(I)], \quad (4)$$

where $\mathbb{I}_X : X \to \{0, 1\}$ is the indicator function,

$$\mathbb{I}_X(x) = \begin{cases} 1 & \text{if} \quad x \in X \\ 0 & \text{otherwise} \end{cases}$$

and

$$\begin{aligned} \mathbf{I}(0) &= (\mathbf{x}(0), \mathbf{x}(-1), \dots, \mathbf{x}(-L)), \\ \mathbf{I}(1) &= (\mathbf{x}(1), \mathbf{x}(0), \dots, \mathbf{x}(-L+1)), \\ I &= (x(0), x(-1), \dots, x(-L)), \\ I' &= (x(1), x(0), \dots, x(-L+1)). \end{aligned}$$

In (4), the indicator function is included to take into account the exit event. Then, the value function J^π in (3) corresponding to π satisfies $J^\pi = T_\pi J^\pi$, which is called the Bellman equation.

Theorem 1: $J^\pi = T_\pi J^\pi$ holds.

Similarly, for any bounded function $J_0 : \mathcal{I} \to \mathbb{R}_+$, we can define the static operator

$$\begin{aligned} &(T J_0)(I) \\ &:= \inf_{u \in U} \{R(x(0), u) \\ &+ \gamma \mathbb{E}[\mathbb{I}_X(\mathbf{x}(1)) J_0(\mathbf{I}(1)) | \mathbf{I}(0) = I, \mathbf{u}(0) = u]\}, \end{aligned}$$

$$:= \inf_{u\in U}\left\{R(x(0),u)+\gamma\int_{x(1)\in X} J_0(I')p_{\mathbf{x}}(x(1)|I,u)dx(1)\right\}, \tag{5}$$

Define the space of bounded functions $J:\mathcal{I}\to\mathbb{R}_+$ as $\mathcal{M}:=\{J:\mathcal{I}\to\mathbb{R}_+ : J<\infty\}$. It can be proved that $(\mathcal{M},d)$ is a complete metric space [24, pp. 301] with the metric

$$d(J,J'):=\sup_{I\in\mathcal{I}}|J(I)-J'(I)|.$$

In the following, we prove that the optimal cost J^* uniquely satisfies $TJ^*=J^*$ called the Bellman's equation, and the sequence, $(J_k)_{k=0}^{\infty}$, generated by the DP algorithm, $J_{k+1}=TJ_k, J_0\equiv 0$ (called value iteration), converges to J^* under Assumption 1. We note that all proofs of this paper are contained in Appendix of the online supplemental material [25].

Theorem 2 (Convergence): The sequence $(J_k)_{k=0}^{\infty}$ generated by the DP algorithm, $J_{k+1}(I)=(TJ_k)(I)$, $I\in\mathcal{I}$ with $J_0\equiv 0$ uniformly converges to J^* w.r.t. the metric d.

Remark 1: For MDPs where continuous and discrete state-spaces coexist and are coupled, a convergence result of the DP was addressed in [12, Theorem 2,Theorem 3]. However, the proof in [12] cannot be directly applied to our case as the MDP has continuous stochastic disturbances for the system in (1).

If J^* is known, then the optimal control policy can be recovered by using

$$\pi^*(I):=\inf_{u\in U}\{R(x(0),u) + \gamma\mathbb{E}[\mathbb{I}_X(\mathbf{x}(1))J^*(\mathbf{I}(1))|\mathbf{I}(0)=I,\mathbf{u}(0)=u]\} \tag{6}$$

provided that the infimum is attained, and this is the case when U is a discrete and finite set or when U is compact. A disadvantage of the policy recovery form (6) lies in the fact that it relies on the model knowledge. A way to overcome this difficulty is to use the so-called action-value function or Q-function [1, pp. 192], [26]. In particular, for a given policy π and corresponding J^π, the Q-function, $Q^\pi:\mathcal{I}\times U\to\mathbb{R}_+$, is defined as

$$Q^\pi(I,u) := R(x(0),u)+\gamma\mathbb{E}[\mathbb{I}_X(\mathbf{x}(1))J^\pi(\mathbf{I}(1))|\mathbf{I}(0)=I,\mathbf{u}(0)=u].$$

Since $J^\pi=T_\pi J^\pi$, comparing the above equation with (4), one can prove $J^\pi(I)=Q^\pi(I,\pi(I))$ and

$$Q^\pi(I,u)=R(x(0),u) + \gamma\mathbb{E}[\mathbb{I}_X(\mathbf{x}(1))Q^\pi(\mathbf{I}(1),\pi(\mathbf{I}(1)))|\mathbf{I}(0)=I,\mathbf{u}(0)=u].$$

Similarly, the optimal Q-function is defined as

$$Q^*(I,u) := R(x(0),u)+\gamma\mathbb{E}[\mathbb{I}_X(\mathbf{x}(1))J^*(\mathbf{I}(1))|\mathbf{I}(0)=I,\mathbf{u}(0)=u]. \tag{7}$$

By comparing this definition with (6), the optimal policy can be expressed as $\pi^*(I):=\arg\inf_{u'\in U}Q^*(I,u')$. In addition, by using the definition of Q-function and $J^*=TJ^*$, the optimal cost can be represented by $J^*(I)=\inf_{u\in U}Q^*(I,u)$. Plugging it into (7) yields

$$\begin{aligned}Q^*(I,u)&=R(x(0),u)\\ &+\gamma\mathbb{E}[\mathbb{I}_X(\mathbf{x}(1))\inf_{u'\in U}Q^*(\mathbf{I}(1),u')|\mathbf{I}(0)=I,\mathbf{u}(0)u]\\ &=:(FQ^*)(I,u)\end{aligned} \tag{8}$$

where the expectation is with respect to $\mathbf{I}(1)$. Then, (7) can be written as the Q-Bellman equation $Q^*=FQ^*$, which is equivalent to the Bellman equation $J^*=TJ^*$. The Q-value iteration, $Q_{k+1}=FQ_k$ with $Q_0\equiv 0$, generates sequence $(Q_k)_{k=0}^{\infty}$ that converges to Q^* under the same condition as in the DP.

Corollary 1 (Convergence): The sequence, $(Q_k)_{k=0}^{\infty}$, generated by the DP algorithm, $Q_{k+1}(I,u)=(FQ_k)(I,u)$, $I\in\mathcal{I}$, $u\in U$, with $Q_0\equiv 0$ uniformly converges to Q^* w.r.t. the metric d.

An advantage of the Q-value iteration is that once found, the control policy can be recovered without the model knowledge.

Remark 2: In practice, the value function or Q-function can be represented by universal function approximators [20], for example, a deep neural network or radial basis functions. With such an approximator, the convergence proof should be modified for the specific approximator. However, this topic is out of the scope of this paper. Moreover, to implement DP algorithms in this paper, one needs to integrate over the entire information space in the definition of the Bellman operators, for instance, F in (5). In practice, reinforcement learning algorithms [21] can be applied to approximate the Bellman operators by their stochastic approximations, and perform the DP algorithms using the stochastic estimations.

IV. Dynamic Programming without Markov Property

If we consider the POMDP, then the operators in (5) and (4) are not well defined. In [27], the authors considered a behavior policy π_b and the corresponding limiting stationary distribution of the MDP (1)

$$\lim_{k\to\infty}p_{\mathbf{x}}(\cdot|\mathbf{x}(k),\mathbf{s}(k),\pi_b(\mathbf{I}(k)))=\xi_{\mathbf{x}}(\cdot;\pi_b),$$
$$\lim_{k\to\infty}p_{\mathbf{s}}(\cdot|\mathbf{s}(k),\mathbf{x}(k))=\xi_{\mathbf{s}}(\cdot;\pi_b),$$

where $\xi_{\mathbf{x}}(\cdot;\pi_b)$ is the stationary distribution of the continuous state and $\xi_{\mathbf{s}}(\cdot;\pi_b)$ is the stationary distribution of the discrete space under the behavior policy π_b provided that they exist.

Then, the probability density of the next continuous state $\mathbf{x}(k+1)$ given current state $\mathbf{x}(k)=x$ and action $\mathbf{u}(k)=u$ is

$$\begin{aligned}&\mathbb{P}[\mathbf{x}(k+1)=\cdot|\mathbf{x}(k)=x,\mathbf{u}(k)=u]\\ &=\sum_{s\in S}p_{\mathbf{x}}(\cdot|x,s,u)\xi_{\mathbf{s}}(s;\pi_b).\end{aligned}$$

Therefore, an information state transition density function $p_{\mathbf{I}}(I'|I, u; \pi_b)$ is well defined, and the operator corresponding to (8) can be defined in this setting. The DP solution under this condition is similar to how the standard Q-learning operates with POMDPs (see [27, Theorem 2]).

To consider a more generic scenario, a different approach is adopted in this paper. For the analysis, the concept of the belief state is introduced first. The belief state [28], $\mathbf{b}(k) \in \Delta_{|S|}$, at time k is defined as the probability distribution of $\mathbf{z}(k)$ at time k.

Definition 4 (Belief state): Consider the MDP (1). The belief state, $(\mathbf{b}(k))_{k=0}^{\tau}$, is defined by the recursion

$$\mathbf{b}(k+1) = P(\mathbf{x}(k))^T \mathbf{b}(k), \quad b(0) = \rho_d.$$

Note that $(\mathbf{b}(k))_{k=0}^{\tau}$ is also a stochastic process due to the randomness of $\mathbf{x}(k)$. In particular, each $\mathbf{b}(k)$ is defined on the probability space $(\Omega, \mathcal{F}, v)$ such that $\Omega = \Delta_{|S|}$, $v(F) = \mathbb{P}[\mathbf{b}(k) \in F], \forall F \in \mathcal{F}$, and $\mathcal{F}$ is a σ-algebra on $\Delta_{|S|}$. The sequence $(\mathbf{b}(k))_{k=0}^{\tau}$ corresponds to a single realization of $(\mathbf{x}(k))_{k=0}^{\tau}$ under a fixed policy π.

When the model and the current state are exactly known, then the next belief state can be computed at every time step in a deterministic fashion based on the current belief state. With the deterministic belief state propagation, a new DP is introduced in the next subsection.

A. Known Belief State

Assuming that the belief state is known, we consider the particular information structure in this subsection

$$\hat{\mathbf{I}}(k) =: (\mathbf{x}(k), \mathbf{b}(k)) \in \hat{\mathcal{I}} = X \times \Delta_{|S|}.$$

Then, $(\hat{\mathbf{I}}(k))_{k=0}^{\tau}$ is an MDP, i.e., its evolution can be expressed as $\hat{\mathbf{I}}(k+1) \sim \mathbb{P}[\hat{\mathbf{I}}(k+1) = \hat{I}(k+1)|\hat{\mathbf{I}}(k) = \hat{I}(k), \mathbf{u}(k) = u]$ as illustrated in Figure 3. Therefore, by Theorem 2, the DP

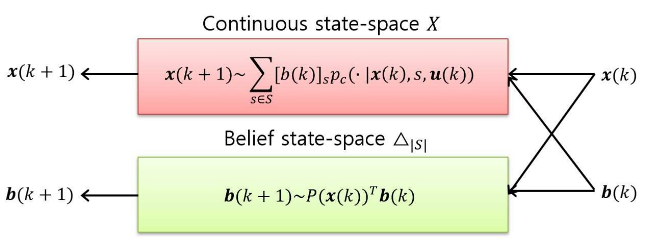

Fig. 3. MDP with belief state.

$$\hat{Q}_{k+1}(\hat{I}, u) = (\hat{F}\hat{Q}_k)(\hat{I}, u), \quad \hat{Q}_0(\hat{I}, u) \equiv 0, \tag{9}$$

converges to $\hat{Q}^*(\hat{I}, u)$, where

$$(\hat{F}\hat{Q}_0)(\hat{I}, u) = R(x, u) + \gamma \int_{x' \in X} \inf_{u' \in U} Q_0(\hat{I}', u') \sum_{s \in S} [b]_s p_{\mathbf{x}}(x'|x, s, u) dx', \tag{10}$$

$\hat{I} = (x, b)$, $\hat{I}' = (x', b')$, $[b]_s$ is the sth element of b, $b' = P(x)^T b$, and x' is the next state corresponding to $\hat{I}'$. The optimal solution of the DP (9) may give better performance compared to a DP solution, if exists, without the belief state information. In the next subsection, we introduce a DP-like algorithm without the belief state information, which may not have a fixed point solution. However, we establish a convergence of the algorithm to a set around the optimal solution $\hat{Q}^*$.

B. Unknown Belief State

In this subsection, we consider the case that the belief state is unknown. In this case, we define a sequence of operators $(\mathbf{F}^{(k)})_{k=0}^{\tau}$ associated with a sequence of the belief states $(\mathbf{b}(k))_{k=0}^{\tau}$.

Assumption 5: $(\mathbf{b}(k))_{k=0}^{\tau}$ is a sequence belief states corresponding to a single realization of the episode under a fixed (behavior) policy π.

Definition 5: For any bounded $Q_0 : \mathcal{I} \times U \to \mathbb{R}_+$, define a sequence of operators $(\mathbf{F}^{(k)})_{k=0}^{\tau}$ associated with the sequence of belief states, $(\mathbf{b}(k))_{k=0}^{\tau}$ in Assumption 5, as

$$\begin{aligned}(\mathbf{F}^{(k)}Q_0)(I, u) &= R(x, u) \\ &+ \gamma \int_{x' \in X} \inf_{u' \in U} Q_0(I', u') \\ &\times \sum_{s \in S} [\mathbf{b}(k)]_s p_{\mathbf{x}}(x(1)|x(0), s, u) dx(1),\end{aligned} \tag{11}$$

where

$$\begin{aligned}\mathbf{I}(0) &= (\mathbf{x}(0), \mathbf{x}(-1), \ldots, \mathbf{x}(-L)), \\ \mathbf{I}(1) &= (\mathbf{x}(1), \mathbf{x}(0), \ldots, \mathbf{x}(-L+1)), \\ I &= (x(0), x(-1), \ldots, x(-L)), \\ I' &= (x(1), x(0), \ldots, x(-L+1)).\end{aligned}$$

Note that the sequence of operators, $(\mathbf{F}^{(k)})_{k=0}^{\tau}$, is stochastic as each $\mathbf{F}^{(k)}$ depends on $\mathbf{b}(k)$, and the DP

$$\mathbf{Q}_{k+1} = \mathbf{F}^{(k)}\mathbf{Q}_k, \quad \mathbf{Q}_0 \equiv 0 \tag{12}$$

may not converge in general as $\mathbf{F}^{(k)}$ is time-varying. However, under certain conditions, we can obtain a bounded set around $\hat{Q}^*$ to which $(\mathbf{Q}_k)_{k=0}^{\tau}$ converges as $\tau \to \infty$. For convenience, new notations are adopted. Define

$$\begin{aligned}\beta(I, b) &:= \mathbb{P}[\mathbf{s}(0) = \cdot|\mathbf{I}(0) = I, \mathbf{b}(-L) = b] \\ &= P(x(0))^T \times \cdots \times P(x(-L))^T b.\end{aligned} \tag{13}$$

In (13), $b \in \Delta_{|S|}$ represents the belief state at time $-L$, and $\beta(I, b)$ implies the belief state at time 0 given $\mathbf{I}(0) = I$ and $\mathbf{b}(-L) = b$. The following result establishes the fact that $(\mathbf{Q}_k)_{k=0}^{\tau}$ converges to a bounded set around $\hat{Q}^*$ as $\tau \to \infty$.

Theorem 3: We assume that $\tau = \infty$. For any $L > 0$, let $l_L \in \mathbb{R}_+$ be a Lipschtz constant such that

$$\begin{aligned}&\|\beta(I, b) - \beta(I, b')\|_\infty \le l_L \|b - b'\|_\infty, \\ &\forall b, b' \in \Delta_{|S|}, \quad I \in X^{L+1}.\end{aligned} \tag{14}$$

Let l_L^* be the infimum over all such constants l_L. Moreover, let $(\mathbf{Q}_k)_{k=0}^{\infty}$ be the sequence generated by the DP (12).

Let $\hat{Q}^*$ be the optimal solution generated by the DP (9). For all $k \geq 0$, the worst error bound is given by

$$\sup_{u\in U, I\in X^{L+1}, b\in\Delta_{|S|}} |\mathbf{Q}_k(I,u) - \hat{Q}^*((x(0),\beta(I,b)),u)| \leq \frac{\gamma(1-\gamma^k)M|S|^2}{(1-\gamma)^2} l_L^* + \frac{\gamma^k M}{1-\gamma}$$

with probability one, where $x(0)$ is the first element of the tuple I, b corresponds to the belief state at time $-L$. In the limit $k \to \infty$, we obtain

$$\limsup_{k\to\infty} \sup_{u\in U, I\in X^{L+1}, b\in\Delta_{|S|}} |\mathbf{Q}_k(I,u) - \hat{Q}^*((x(0),\beta(I,b)),u)| \leq \frac{M\gamma|S|^2}{(1-\gamma)^2} l_L^*$$

with probability one.

In practice, the discount factor $\gamma \in [0,1)$ is close to one; thus the error bound may be large. However, l_L is often small as in many applications [11]. Then, under some conditions, we have $\lim_{L\to\infty} l_L^* = 0$. As a simple example, assume that $P(x)$ is a constant matrix P and that the Markov chain has a unique stationary distribution μ such that $\mu^T P = \mu^T$. This implies that $\lim_{L\to\infty}(P^L)^T b = \lim_{L\to\infty}(P^L)^T b' = \mu$ for any $b, b' \in \Delta_{|S|}$. Therefore, $\lim_{L\to\infty} \|(P^L)^T b - (P^L)^T b'\|_2 = 0$, meaning that $\lim_{L\to\infty} l_L^* = 0$.

CONCLUSION

In this paper, we have studied a DP framework for POMDPs with jointly continuous and discrete state-spaces. A finite observation history has been used as an information structure of an output-feedback control policy. We have established a convergence of the DP algorithm to a set round an optimal solution. Developments of reinforcement learning algorithms based on the current analysis can be potential future research directions.

REFERENCES

[1] D. P. Bertsekas, *Dynamic programming and optimal control*. Athena Scientific Belmont, MA, 1995, vol. 1, no. 2.

[2] L. Blackmore, A. Bektassov, M. Ono, and B. C. Williams, "Robust, optimal predictive control of jump markov linear systems using particles," in *International Workshop on Hybrid Systems: Computation and Control*, 2007, pp. 104–117.

[3] L. Blackmore, M. Ono, A. Bektassov, and B. C. Williams, "A probabilistic particle-control approximation of chance-constrained stochastic predictive control," *IEEE Transactions on Robotics*, vol. 26, no. 3, pp. 502–517, 2010.

[4] P. Patrinos, P. Sopasakis, H. Sarimveis, and A. Bemporad, "Stochastic model predictive control for constrained discrete-time markovian switching systems," *Automatica*, vol. 50, no. 10, pp. 2504–2514, 2014.

[5] O. L. V. Costa, E. Assumpção Filho, E. K. Boukas, and R. Marques, "Constrained quadratic state feedback control of discrete-time markovian jump linear systems," *Automatica*, vol. 35, no. 4, pp. 617–626, 1999.

[6] M. Bichi, G. Ripaccioli, S. Di Cairano, D. Bernardini, A. Bemporad, and I. V. Kolmanovsky, "Stochastic model predictive control with driver behavior learning for improved powertrain control," in *49th IEEE Conference on Decision and Control (CDC)*, 2010, pp. 6077–6082.

[7] S. Di Cairano, D. Bernardini, A. Bemporad, and I. V. Kolmanovsky, "Stochastic mpc with learning for driver-predictive vehicle control and its application to hev energy management," *IEEE Transactions on Control Systems Technology*, vol. 22, no. 3, pp. 1018–1031, 2014.

[8] J. Page, D. Robinson, N. Morel, and J.-L. Scartezzini, "A generalised stochastic model for the simulation of occupant presence," *Energy and buildings*, vol. 40, no. 2, pp. 83–98, 2008.

[9] J. R. Dobbs and B. M. Hencey, "Model predictive hvac control with online occupancy model," *Energy and Buildings*, vol. 82, pp. 675–684, 2014.

[10] D. Lee, S. Lee, P. Karava, and J. Hu, "Simulation-based policy gradient and its building control application," in *Proc. American control conference (ACC2018)*, Milwaukee, USA, 2018.

[11] ——, "Approximate dynamic programming for building control problems with occupant interactions," in *Proc. American control conference (ACC2018)*, Milwaukee, USA, 2018.

[12] I. V. Kolmanovsky, L. Lezhnev, and T. L. Maizenberg, "Discrete-time drift counteraction stochastic optimal control: Theory and application-motivated examples," *Automatica*, vol. 44, no. 1, pp. 177–184, 2008.

[13] L. Johannesson, M. Asbogard, and B. Egardt, "Assessing the potential of predictive control for hybrid vehicle powertrains using stochastic dynamic programming," *IEEE Transactions on Intelligent Transportation Systems*, vol. 8, no. 1, pp. 71–83, 2007.

[14] V. Mnih, K. Kavukcuoglu, D. Silver, A. A. Rusu, J. Veness, M. G. Bellemare, A. Graves, M. Riedmiller, A. K. Fidjeland, G. Ostrovski *et al.*, "Human-level control through deep reinforcement learning," *Nature*, vol. 518, no. 7540, pp. 529–533, 2015.

[15] M. Hausknecht and P. Stone, "Deep recurrent Q-learning for partially observable MDPSs," *arXiv preprint arXiv:1507.06527*, 2015.

[16] M. Prandini, J. Hu, J. Lygeros, and S. Sastry, "A probabilistic approach to aircraft conflict detection," *IEEE Transactions on intelligent transportation systems*, vol. 1, no. 4, pp. 199–220, 2000.

[17] G. C. Calafiore and L. Fagiano, "Stochastic model predictive control of LPV systems via scenario optimization," *Automatica*, vol. 49, no. 6, pp. 1861–1866, 2013.

[18] G. Schildbach, L. Fagiano, C. Frei, and M. Morari, "The scenario approach for stochastic model predictive control with bounds on closed-loop constraint violations," *Automatica*, vol. 50, no. 12, pp. 3009–3018, 2014.

[19] G. C. Calafiore and L. Fagiano, "Robust model predictive control via scenario optimization," *IEEE Transactions on Automatic Control*, vol. 58, no. 1, pp. 219–224, 2013.

[20] D. P. Bertsekas and J. N. Tsitsiklis, *Neuro-dynamic programming*. Athena Scientific Belmont, MA, 1996.

[21] R. S. Sutton and A. G. Barto, *Reinforcement learning: An introduction*. MIT Press, 1998.

[22] S. I. Resnick, *Adventures in stochastic processes*. Springer Science & Business Media, 2013.

[23] A. Geramifard, T. J. Walsh, S. Tellex, G. Chowdhary, N. Roy, J. P. How *et al.*, "A tutorial on linear function approximators for dynamic programming and reinforcement learning," *Foundations and Trends® in Machine Learning*, vol. 6, no. 4, pp. 375–451, 2013.

[24] E. S. Şuhubi, "Functional analysis." Springer, 2003.

[25] D. Lee, N. He, and J. Hu, "Dynamic programming for stochastic control systems with jointly discrete and continuous state-spaces," *arXiv preprint arXiv:1803.08876*, 2018.

[26] C. Watkins, "Learning from delayed rewards," *PhD thesis, King's College, University of Cambridge*, 1989.

[27] S. P. Singh, T. Jaakkola, and M. I. Jordan, "Learning without state-estimation in partially observable markovian decision processes," in *Machine Learning Proceedings 1994*, 1994, pp. 284–292.

[28] M. Hauskrecht, "Value-function approximations for partially observable markov decision processes," *Journal of Artificial Intelligence Research*, vol. 13, pp. 33–94, 2000.

[29] R. Durrett, *Probability: theory and examples*. Cambridge university press, 2010.

APPENDIX

Throughout this section, we use the notation

$$\mathbf{I}(0) = (\mathbf{x}(0), \mathbf{x}(-1), \ldots, \mathbf{x}(-L)),$$
$$\mathbf{I}(1) = (\mathbf{x}(1), \mathbf{x}(0), \ldots, \mathbf{x}(-L+1)),$$
$$I = (x(0), x(-1), \ldots, x(-L)),$$
$$I' = (x(1), x(0), \ldots, x(-L+1)).$$

A. Proof of Theorem 1

By the definition, we have

$$\begin{aligned}
J^\pi(I) &:= \mathbb{E}\left[\sum_{i=0}^{\tau(\mathbf{I}(0);\pi)} \gamma^i \mathbf{r}(\mathbf{x}(i), \pi(\mathbf{I}(i)))\middle| \mathbf{I}(0) = I\right] \\
&= \mathbb{E}\left[\mathbf{r}(\mathbf{x}(0), \pi(\mathbf{I}(0))) + \sum_{i=1}^{\tau(\mathbf{I}(0);\pi)} \gamma^i \mathbf{r}(\mathbf{x}(i), \pi(\mathbf{I}(i)))\middle| \mathbf{I}(0) = I\right] \\
&= R(x(0), \pi(I)) + \gamma\mathbb{E}\left[\mathbb{I}_X(\mathbf{x}(1)) \sum_{i=1}^{\tau(\mathbf{I}(0);\pi)} \gamma^{i-1} \mathbf{r}(\mathbf{x}(i), \pi(\mathbf{I}(i)))\middle| \mathbf{I}(0) = I\right] \\
&= R(x(0), \pi(I)) + \gamma\mathbb{E}\left[\mathbb{I}_X(\mathbf{x}(1)) \sum_{i=1}^{\tau(\mathbf{I}(1);\pi)+1} \gamma^{i-1} \mathbf{r}(\mathbf{x}(i), \pi(\mathbf{I}(i)))\middle| \mathbf{I}(0) = I\right] \\
&= R(x(0), \pi(I)) + \gamma\mathbb{E}\left[\mathbb{I}_X(\mathbf{x}(1)) \sum_{i=0}^{\tau(\mathbf{I}(1);\pi)} \gamma^{i} \mathbf{r}(\mathbf{x}(i+1), \pi(\mathbf{I}(i+1)))\middle| \mathbf{I}(0) = I\right] \\
&= R(x(0), \pi(I)) + \gamma\mathbb{E}[\mathbb{I}_X(\mathbf{x}(1)) J^\pi(\mathbf{I}(1))| \mathbf{I}(0) = I] \\
&= T_\pi J^\pi(I).
\end{aligned}$$

B. Proof of Theorem 2

To prove Theorem 2, one needs to first prove that $J_k \in \mathcal{M}$ for all $k \in \mathbb{N}$ and that T is a contraction map.

Proposition 2: Assume $J_0 \equiv 0$. The following statements hold true:

1) $J_k \in \mathcal{M}$ for all $k \in \mathbb{N}$. Especially, $J_k \leq M/(1-\gamma)$;
2) T is a map from $\mathcal{M}$ to $\mathcal{M}$;
3) T is a contraction map.

Proof: To prove 1), note that $J_1 \leq M$ by Assumption 1. Moreover, by the definition of T, $J_2 \leq M + \gamma M$, and repeating it yields $J_k \leq M\sum_{i=0}^{k-1}\gamma^i \leq M\sum_{i=0}^{\infty}\gamma^i = \frac{M}{1-\gamma}$, implying $J_k \in \mathcal{M}$ for all $k \in \mathbb{N}$. 2) is proved directly from 1). To prove 3), consider any $J, J' \in \mathcal{M}$. Then, we have $d(TJ, TJ') = \sup_{I\in\mathcal{I}} |TJ(I) - TJ'(I)| \leq \sup_{I\in\mathcal{I}, u\in U} \gamma|\mathbb{E}[J(\mathbf{I}(1)) - J'(\mathbf{I}(1))|\mathbf{I}(0) = I, \mathbf{u}(0) = u]| \leq \gamma d(J, J')$. Therefore, T is a contraction map. ∎

Proposition 3: Assume $J_0 \equiv 0$. The following statements hold true:

1) T is monotone, i.e., if $J \geq J'$, then $TJ \geq TJ'$;
2) For any given $\pi \in \Pi$, T_π is monotone;
3) $(J_k)_{k=0}^\infty$ is a monotonically non-decreasing sequence;
4) $(J_k)_{k=0}^\infty$ uniformly converges to a unique fixed point $J_\infty \in \mathcal{M}$, i.e., $TJ_\infty = J_\infty$, w.r.t. the metric d.

Proof: To prove 1), assume $J \geq J'$ and recall the definition in (5)

$$\begin{aligned}
(TJ_0)(I) &:= \inf_{u\in U}\{R(x(0), u) + \gamma\mathbb{E}[\mathbb{I}_X(\mathbf{x}(1))J_0(\mathbf{I}(1))|\mathbf{I}(0) = I, \mathbf{u}(0) = u]\}, \\
&:= \inf_{u\in U}\left\{R(x(0), u) + \gamma\int_{x(1)\in X} J_0(I')p_{\mathbf{x}}(x(1)|I, u)dx(1)\right\},
\end{aligned}$$

Then, for any $I \in \mathcal{I}$, we have

$(TJ)(I) - (TJ')(I)$

$$
\begin{aligned}
&\geq \inf_{u\in U}\left\{R(x(0),u)+\gamma\int_{x(1)\in X}J(I')p_{\mathbf{x}}(x(1)|I,u)dx(1)-R(x(0),u)-\gamma\int_{x(1)\in X}J'(I')p_{\mathbf{x}}(x(1)|I,u)dx(1)\right\}\\
&=\gamma\inf_{u\in U}\left\{\int_{x(1)\in X}(J(I')-J'(I'))p_{\mathbf{x}}(x(1)|I,u)dx(1)\right\}\\
&\geq 0,
\end{aligned}
$$

where the last inequality follows from the hypothesis $J \geq J'$. This completes the proof of 1). The statement 2) can be proved in a similar way, so omitted. The proof of 3) is completed by an induction argument. Since $J_0 \equiv 0$ and $J_1(I) = (TJ_0)(I) := \inf_{u\in U} R(x(0), u) \geq 0$ by Assumption 1, $J_1 \geq J_0$ holds. By the monotonicity of the operator T in 1), we have $T^i J_1 \geq T^i J_0, \forall i \in \mathbb{N}_+$, meaning $J_{k+1} \geq J_k, \forall i \in \mathbb{N}$, which concludes the proof of 3). By Proposition 2, $T : \mathcal{M} \to \mathcal{M}$ is a contraction map on the complete metric space $(\mathcal{M}, d)$. By the Banach fixed point theorem [24, Theorem 5.6.1], $(J_k)_{k=0}^{\infty}$ converges to a unique fixed point $J_\infty \in \mathcal{M}$, i.e., $J_\infty = TJ_\infty$, w.r.t. the metric d. The convergence is uniform w.r.t. the metric d, which is proved directly from the definition of the convergence in the metric space as discussed in [24, pp. 301]. ■

The last step for the proof of Theorem 2 is to prove $J_\infty = J^*$, where $J_\infty := \lim_{k\to\infty} J_k$ and the sequence $(J_k)_{k=0}^{\infty}$ is generated by the DP algorithm, $J_{k+1} = (TJ_k)$ with $J_0 \equiv 0$. To this end, we need to prove that for any given $\bar{\pi} \in \Pi^\infty$, we have $\lim_{k\to\infty} T_{\pi_0}T_{\pi_1}\cdots T_{\pi_k}J_0 = J^{\bar{\pi}}$, which will be proved using two intermediate lemmas below.

Lemma 1: Assume $J_0 \equiv 0$. For any given $\bar{\pi} \in \Pi^\infty$ and $k \geq 1$, $T_{\pi_0}T_{\pi_1}\cdots T_{\pi_k}J_0$ is described as

$$
T_{\pi_0}T_{\pi_1}\cdots T_{\pi_k}J_0(I) = \mathbb{E}\left[\left.\sum_{i=0}^{\min\{\tau(\mathbf{I}(0);\bar{\pi}),k\}}\gamma^i\mathbf{r}(\mathbf{x}(i),\pi_i(\mathbf{I}(i)))\right|\mathbf{I}(0)=I\right] \tag{15}
$$

for any $I \in \mathcal{I}$.

Proof: The claim will be proved by an induction argument. Let $k = 0$. Since $J_0 \equiv 0$, by the definition in (4), $T_{\pi_0}J_0(I)$ is given by

$$
T_{\pi_0}J_0(I) = \mathbb{E}[\mathbf{r}(\mathbf{x}(0),\pi_0(\mathbf{I}(0)))|\mathbf{I}(0)=I] = \mathbb{E}\left[\left.\sum_{i=0}^{\min\{\tau(\mathbf{I}(0);\bar{\pi}),0\}}\gamma^i\mathbf{r}(\mathbf{x}(i),\pi_i(\mathbf{I}(i)))\right|\mathbf{I}(0)=I\right],
$$

where we use the fact that $\tau(\mathbf{I}(0); \bar{\pi}) \geq 0$ as $I \in \mathcal{I}$ and $x(0) \in X$.

Now, for an induction argument, suppose for $k \geq 1$

$$
T_{\pi_0}T_{\pi_1}\cdots T_{\pi_{k-1}}J_0(I) = \mathbb{E}\left[\left.\sum_{i=0}^{\min\{\tau(\mathbf{I}(0);\bar{\pi}),k-1\}}\gamma^i\mathbf{r}(\mathbf{x}(i),\pi_i(\mathbf{I}(i)))\right|\mathbf{I}(0)=I\right]
$$

holds. Then, shifting the time index of the control policy by one in the above equation yields

$$
T_{\pi_1}T_{\pi_2}\cdots T_{\pi_k}J_0(I) = \mathbb{E}\left[\left.\sum_{i=0}^{\min\{\tau(\mathbf{I}(0);\bar{\pi}_{1:\infty}),k-1\}}\gamma^i\mathbf{r}(\mathbf{x}(i),\pi_{i+1}(\mathbf{I}(i))\right|\mathbf{I}(0)=I\right], \tag{16}
$$

where $\bar{\pi}_{1:\infty} := (\pi_1, \pi_2, \ldots)$. We apply the operator T_{π_0} to (16) to have

$$
\begin{aligned}
T_{\pi_0}T_{\pi_1}\cdots T_{\pi_k}J_0(I) &= \mathbb{E}\left[\mathbf{r}(x(0),\pi_0(I))+\gamma\mathbb{E}[\mathbb{I}_X(\mathbf{x}(1))T_{\pi_1}\cdots T_{\pi_k}J_0(\mathbf{I}(1))]\right|\mathbf{I}(0)=I]\\
&= \mathbb{E}\left[\mathbf{r}(x(0),\pi_0(I))+\mathbb{I}_X(\mathbf{x}(1))\times\left.\sum_{i=0}^{\min\{\tau(\mathbf{I}(1);\bar{\pi}_{1:\infty}),k-1\}}\gamma^{i+1}\mathbf{r}(\mathbf{x}(i+1),\pi_{i+1}(\mathbf{I}(i+1)))\right|\mathbf{I}(0)=I\right],
\end{aligned} \tag{17}
$$

where the second equation is obtained by (16) and $\tau(\mathbf{I}(1); \bar{\pi}_{1:\infty}) \geq -1$ is the first time instant the trajectory $\mathbf{x}(k)$ starting from $\mathbf{x}(1)$ exits X given $\mathbf{I}(1)$ and $\bar{\pi}_{1:\infty}$.

Note that $\mathbf{x}(1)$ is a random variable and $\tau(\mathbf{I}(1); \bar{\pi}_{1:\infty}) = -1$ when $\mathbf{x}(1) \notin X$. In this case, we define $\sum_{i=1}^{-1}\cdot = 0$. By conditioning on the stopping time $\tau(\mathbf{I}(0); \bar{\pi})$, the expectation in (17) is expressed as

$$
\begin{aligned}
&\mathbb{E}[\mathbf{r}(x(0),\pi_0(I))|\,\tau(I;\bar{\pi})=0]\mathbb{P}[\tau(I;\bar{\pi})=0]\\
&+\mathbb{E}\left[\mathbf{r}(x(0),\pi_0(I))+\sum_{i=0}^{\min\{\tau(\mathbf{I}(1);\bar{\pi}_{1:\infty}),k-1\}}\gamma^{i+1}\mathbf{r}(\mathbf{x}(i+1),\pi_{i+1}(\mathbf{I}(i+1)))\,|\mathbf{I}(0)=I,\tau(I;\bar{\pi})\geq 1\right]\mathbb{P}[\tau(I;\bar{\pi})\geq 1].
\end{aligned} \tag{18}
$$

In the second expectation, $\tau(I;\bar\pi) \geq 1$ implies $\tau(\mathbf{I}(1);\bar\pi_{1:\infty}) \geq 0$. Noting that $\mathbb{P}[\tau(\mathbf{I}(0);\bar\pi) = j] = \mathbb{P}[\tau(\mathbf{I}(1);\bar\pi_{1:\infty}) = j-1]$ for all $j \geq 1$, the quantity $\min\{\tau(\mathbf{I}(1);\bar\pi_{1:\infty}), k-1\}$ can be rewritten as

$$\min\{\tau(\mathbf{I}(1),\bar\pi_{1:\infty}), k-1\} = \min\{\tau(\mathbf{I}(0),\bar\pi) - 1, k-1\}.$$

Plugging it into the original formulation in (18) results in

$$\begin{aligned}
&\mathbb{E}[\mathbf{r}(x(0),\pi_0(I))|\,\tau(I;\bar\pi)=0]\mathbb{P}[\tau(I;\bar\pi)=0]\\
&+\mathbb{E}\left[\mathbf{r}(x(0),\pi_0(I)) + \sum_{i=0}^{\min\{\tau(\mathbf{I}(0);\bar\pi_{1:\infty})-1,k-1\}} \gamma^{i+1}\mathbf{r}(\mathbf{x}(i+1),\pi_{i+1}(\mathbf{I}(i+1)))\right|\mathbf{I}(0)=I,\tau(I;\bar\pi)\geq 1\Bigg]\mathbb{P}[\tau(I;\bar\pi)\geq 1]\\
&=\mathbb{E}[\mathbf{r}(x(0),\pi_0(I))|\,\tau(I;\bar\pi)=0]\mathbb{P}[\tau(I;\bar\pi)=0]\\
&+\mathbb{E}\left[\mathbf{r}(x(0),\pi_0(I)) + \sum_{i=1}^{\min\{\tau(\mathbf{I}(0);\bar\pi_{1:\infty}),k\}} \gamma^{i}\mathbf{r}(\mathbf{x}(i),\pi_{i}(\mathbf{I}(i)))\right|\mathbf{I}(0)=I,\tau(I;\bar\pi)\geq 1\Bigg]\mathbb{P}[\tau(I;\bar\pi)\geq 1]\\
&=\mathbb{E}\left[\sum_{i=0}^{\min\{\tau(\mathbf{I}(0);\bar\pi_{0:\infty}),k\}} \gamma^{i}\mathbf{r}(\mathbf{x}(i),\pi_{i}(\mathbf{I}(i)))\right|\mathbf{I}(0)=I\Bigg],
\end{aligned}$$

which is the desired result. ■

Lemma 2: Assume $J_0 \equiv 0$. For any given $\bar\pi \in \Pi^\infty$, we have $\lim_{k\to\infty} T_{\pi_0}T_{\pi_1}\cdots T_{\pi_k}J_0 = J^{\bar\pi}$.

Proof: Define $T_{\pi_0}T_{\pi_1}\cdots T_{\pi_k}J_0 =: J_k^{\bar\pi}$. Using Lemma 1 and following the proof of Proposition 2, it is easy to prove that $J_k^{\bar\pi}$ is bounded on $\mathcal{I}$, i.e., $J_k^{\bar\pi} \leq M/(1-\gamma)$. Moreover, from the definition, $J_k^{\bar\pi}$ is non-decreasing in k. Therefore, the point-wise limit $\lim_{k\to\infty} J_k^{\bar\pi} =: J_\infty^{\bar\pi}$ exists. Noting that the function inside the expectation operator is bounded, we apply the dominated convergence theorem [29, Theorem 1.5.6] to have

$$\begin{aligned}
J_\infty^{\bar\pi}(I) &= \lim_{k\to\infty} J_k^{\bar\pi}(I) = \lim_{k\to\infty} T_{\pi_0}T_{\pi_1}\cdots T_{\pi_k}J_0(I)\\
&= \lim_{k\to\infty}\mathbb{E}\left[\sum_{i=0}^{\min\{\tau(\mathbf{I}(0);\bar\pi),k\}}\gamma^i\mathbf{r}(\mathbf{x}(i),\pi_i(\mathbf{I}(i)))\right|\mathbf{I}(0)=I\Bigg]\\
&= \mathbb{E}\left[\lim_{k\to\infty}\sum_{i=0}^{\min\{\tau(\mathbf{I}(0);\bar\pi),k\}}\gamma^i\mathbf{r}(\mathbf{x}(i),\pi_i(\mathbf{I}(i)))\right|\mathbf{I}(0)=I\Bigg]\\
&= J^{\bar\pi}(I)
\end{aligned}$$

for all $I \in \mathcal{I}$, where the third equality is due to the dominated convergence theorem and the last equality is from the definition of $J^{\bar\pi}$. Therefore, $J^{\bar\pi} = J_\infty^{\bar\pi}$, and the desired result is obtained. ■

Now, we are in position to prove that $J_\infty = J^*$ holds with $J_0 \equiv 0$.

Proposition 4: Assume $J_0 \equiv 0$. Then, $J_\infty = J^*$.

Proof: We follow the proof of [20, Prop. 2.1]. Define π_∞ as

$$\pi_\infty(I) := \inf_{u\in U}\{R(x(k),u) + \gamma\mathbb{E}[J_\infty(\mathbf{I}(1))|\mathbf{I}(0)=I,\mathbf{u}(0)=u]\}.$$

The above quantity is well defined because $J_\infty \in \mathcal{M}$ by 4) of Proposition 3 and Assumption 1. From the definition (4), it also implies $T_{\pi_\infty}J_0 = TJ_\infty = J_\infty$. Then, since $J_0 \leq J_\infty$ by the monotonicity in Proposition 3, we have $J^{\pi_\infty} = \lim_{k\to\infty} T_{\pi_\infty}^k J_0 \leq \lim_{k\to\infty} T_{\pi_\infty}^k J_\infty = J_\infty$, implying $J^{\pi_\infty} \leq J_\infty$, where Lemma 2 is used in the first equality, the last equality follows from $T_{\pi_\infty}J_\infty = J_\infty$, and the inequality comes from $J_0 \leq J_\infty$ and the monotonicity of T_{π_∞} in Proposition 3. On the other hand, by the definition and the monotonicity of T, we have that for any policy $\bar\pi = (\pi_0,\pi_1,\ldots)$, $J_\infty = \lim_{k\to\infty} T^k J_0 \leq \lim_{k\to\infty} T_{\pi_0}T_{\pi_1}\cdots T_{\pi_k}J_0 = J^{\bar\pi}$, meaning $J_\infty \leq J^{\bar\pi}$ for all $\bar\pi \in \Pi^\infty$, where the last equality follows from Lemma 2. Combining both inequalities results in $J^{\pi_\infty} \leq J_\infty \leq J^{\bar\pi}, \forall\bar\pi \in \Pi^\infty$, yielding $J^{\pi_\infty} = J_\infty = J^*$. ■

Now, Theorem 2 is directly proved by using Proposition 3 and Proposition 4.

Proof of Theorem 2: By (4) of Proposition 3, $(J_k)_{k=0}^\infty$ uniformly converges to a unique fixed point $J_\infty \in \mathcal{M}$, i.e., $TJ_\infty = J_\infty$, w.r.t. the metric d. By Proposition 4, $J_\infty = J^*$, and hence, $(J_k)_{k=0}^\infty$ uniformly converges to J^* w.r.t. the metric d. This completes the proof.

C. Proof of Theorem 3

To prove Theorem 3, we first introduce the following basic relation lemma.

Lemma 3 (Basic relation): Let $(\mathbf{Q}_k)_{k=0}^{\infty}$ be the sequence generated by the DP

$$\mathbf{Q}_{k+1} = \mathbf{F}^{(k)}\mathbf{Q}_k, \quad \mathbf{Q}_0 \equiv 0,$$

where $\mathbf{F}^{(k)}$ is defined in (11). Let $\hat{Q}^*$ be the optimal solution of $\hat{Q}^* = \hat{F}\hat{Q}^*$, where $\hat{F}$ is defined in (10). Then, we have

$$\begin{aligned}&\sup_{u\in U, I\in X^{L+1}, b\in\Delta_{|S|}} |\mathbf{Q}_{k+1}(I,u) - \hat{Q}^*((x(0),\beta(I,b)),u)|\\ &\le \frac{\gamma l_L^* M|S|^2}{1-\gamma} + \gamma \sup_{u\in U, I\in X^{L+1}, b\in\Delta_{|S|}} |\mathbf{Q}_k(I,u) - Q^*((x(0),\beta(I,b)),u)|\end{aligned}$$

for all $k \ge 0$ with probability one, where $x(0) \in X$ is the first element of I.

Proof: Using the definitions of $\mathbf{F}^{(k)}$ and $\hat{F}$, we have

$$\begin{aligned}\mathbf{Q}_{k+1}(I,u) - \hat{Q}^*((x(0),\beta(I,b)),u) &= (\mathbf{F}^{(k)}\mathbf{Q}_k)(I,u) - (\hat{F}\hat{Q}^*)((x(0),\beta(I,b)),u)\\ &= \gamma \int_{x(1)\in X} \inf_{u'\in U} \mathbf{Q}_k(I',u') \sum_{s\in S} [\mathbf{b}(k)]_s p_{\mathbf{x}}(x(1)|x(0),s,u)dx(1)\\ &\quad - \gamma \int_{x(1)\in X} \inf_{u'\in U} \hat{Q}^*((x(1),\beta(I',b')),u') \sum_{s\in S} [\beta(I,b)]_s p_{\mathbf{x}}(x(1)|x(0),s,u)dx(1),\\ &= \gamma \int_{x(1)\in X} \inf_{u'\in U} \mathbf{Q}_k(I',u') \sum_{s\in S} [\beta(I,\mathbf{b}(k-L))]_s p_{\mathbf{x}}(x(1)|x(0),s,u)dx(1)\\ &\quad - \gamma \int_{x(1)\in X} \inf_{u'\in U} \hat{Q}^*((x(1),\beta(I',b')),u') \sum_{s\in S} [\beta(I,b)]_s p_{\mathbf{x}}(x(1)|x(0),s,u)dx(1),\end{aligned}$$

where $b' = P(x)^T b$. Adding and subtracting

$$\gamma \int_{x(1)\in X} \inf_{u'\in U} \mathbf{Q}_k(I',u') \sum_{s\in S} [\beta(I,b)]_s p_{\mathbf{x}}(x(1)|x(0),s,u)dx(1)$$

to the last equation, we have

$$\begin{aligned}&\mathbf{Q}_{k+1}(I,u) - \hat{Q}^*((x,\beta(I,b)),u)\\ &\le \gamma|S| \int_{x(1)\in X} \inf_{u'\in U} \mathbf{Q}_k(I',u') \times \sum_{s\in S} |[\beta(I,\mathbf{b}(k-L))]_s - [\beta(I,b)]_s| \frac{1}{|S|} p_{\mathbf{x}}(x(1)|x(0),s,u)dx(1)\\ &+ \gamma \int_{x(1)\in X} \sup_{u'\in U} (|\mathbf{Q}_k(I',u') - \hat{Q}^*((x(1),\beta(I',b')),u')) \times \sum_{s\in S} [\beta(I,b)]_s p_{\mathbf{x}}(x(1)|x(0),s,u)dx(1)\\ &\le \gamma|S| \int_{x(1)\in X} \frac{M|S|}{1-\gamma} \|\beta(I,\mathbf{b}(k-L)) - \beta(I,b)\|_\infty \times \sum_{s\in S} \frac{1}{|S|} p_{\mathbf{x}}(x(1)|x(0),s,u)dx(1)\\ &+ \gamma \int_{x(1)\in X} \sup_{u'\in U} |\mathbf{Q}_k(I',u') - \hat{Q}^*((x(1),\beta(I',b')),u')| \times \sum_{s\in S} [\beta(I,b)]_s p_{\mathbf{x}}(x(1)|x(0),s,u)dx(1),\end{aligned}$$

where in the second inequality, we use the definition of $\|\cdot\|_\infty$ and the bound $\mathbf{Q}_k(I',u') \le M/(1-\gamma), \forall I' \in \mathcal{I}, u' \in U$. Using (14), we have

$$\begin{aligned}&\mathbf{Q}_{k+1}(I,u) - \hat{Q}^*((x(0),\beta(I,b)),u)\\ &\le \frac{\gamma M|S|^2}{1-\gamma} l_L^* \|\mathbf{b}(k-L) - b\|_\infty \int_{x(1)\in X} \sum_{s\in S} \frac{1}{|S|} p_{\mathbf{x}}(x(1)|x(0),s,u)dx(1)\\ &+ \gamma \sup_{u'\in U, I'\in X^{L+1}} |\mathbf{Q}_k(I',u') - \hat{Q}^*((x(1),\beta(I',b')),u')|\\ &\le \gamma \frac{M|S|^2}{1-\gamma} l_L^* + \gamma \sup_{u'\in U, I'\in X^{L+1}} (|\mathbf{Q}_k(I',u') - \hat{Q}^*((x(1),\beta(I',b')),u')|),\end{aligned}$$

where in the last inequality, we use the fact that $\|\mathbf{b}(k-L)-b\|_\infty \le 1$ for any $\mathbf{b}(k-L), b$ inside the unit simplex $\Delta_{|S|}$ and $\sum_{s\in S}\frac{1}{|S|}p_{\mathbf{x}}(x(1)|x(0),s,u)$ is a probability density function. The desired result follows by taking the superimum over $I\in\mathcal{I}, u\in U$, and $b\in\Delta_{|S|}$ on the left-hand side of the last inequality. ■

Proof of Theorem 3: Combining the inequalities in Lemma 3 for $k=0$ and $k=1$ yields

$$\begin{aligned}&\frac{1}{\gamma^2}\sup_{u\in U, I\in X^{L+1}, b\in\Delta_{|S|}}|\mathbf{Q}_2(I,u)-\hat{Q}^*((x(0),\beta(I,b)),u)|\\&\le(1+\gamma^{-1})\frac{M|S|^2}{1-\gamma}l_L^* + \sup_{u\in U, I\in X^{L+1}, b\in\Delta_{|S|}}|\mathbf{Q}_0(I,u)-\hat{Q}^*((x(0),\beta(I,b),u)|.\end{aligned}$$

Repeating this $k-2$ times, one gets

$$\begin{aligned}&\sup_{u\in U, I\in X^{L+1}, b\in\Delta_{|S|}}|\mathbf{Q}_k(I,u)-\hat{Q}^*((x(0),\beta(I,b)),u)|\\&\le\gamma^k\left(\sum_{i=0}^{k-1}\frac{1}{\gamma^i}\right)\frac{M|S|^2}{1-\gamma}l_L^* + \gamma^k\sup_{u\in U, I\in X^{L+1}, b\in\Delta_{|S|}}|\mathbf{Q}_0(I,u)-\hat{Q}^*((x(0),\beta(I,b),u)|.\end{aligned}$$

Using $\mathbf{Q}_0(I,u)\equiv 0$, $\hat{Q}^*((x(0),\beta(I,b),u)\le M/(1-\gamma)$, and $\sum_{i=0}^{k-1}(1/\gamma^i)=(1-(\gamma^{-1})^k)/(1-\gamma^{-1})$, the first inequality in Theorem 3 follows. The second inequality can be obtained by taking $\limsup_{k\to\infty}$ on both sides of the inequality. This completes the proof.