

# Robust Dynamic Programming for Discounted Infinite-Horizon Markov Decision Processes with Uncertain Stationary Transition Matrices

Baohua Li, and Jennie Si *Senior Member, IEEE*

## Abstract

In this paper, approximate dynamic programming (ADP) problems are modeled by discounted infinite-horizon Markov decision processes (MDPs) with uncertain stationary transition matrices. The uncertainty in the transition matrices signifies the realistic consideration that an accurate system model may not always be available for the controller design to achieve acceptable performance. In this paper, MDPs are considered under the reformulated definitions of independent or correlated transition matrices. Precisely existing results have focused on addressing optimality criteria and robust ADP algorithms for MDPs with independent transition matrices. Robust value iteration and robust policy iteration are such algorithms that lead to stationary deterministic optimal policies. Major contributions of this paper are two folds. It first provides sufficient conditions and algorithms to determine a stationary deterministic optimal policy for all independent transition matrices and some correlated transition matrices. As a consequence, the robust policy iteration algorithm is simplified. It then proposes a new, squared total value function to address the existence of a stationary deterministic optimal policy for all possible transition matrices, independent or correlated. A new robust policy iteration under total value function is developed, which degenerates to the previous robust policy iteration. The paper also reveals that the optimality criterion and the robust policy iteration under total value function can be used to solve MDPs with uncertain cost functions and weighted squared total value functions.

This work was supported by the National Science Foundation under ECS-0002098 and ECS-0233529.

The authors are with the Department of Electrical Engineering, Arizona State University, Tempe, AZ 85287-5706 (e-mail: Baohua.Li@asu.edu, Si@asu.edu).

**Keywords:** Markov decision processes, Uncertain transition matrix, Robust value iteration, Robust policy iteration, Value function, Total value function.

## I. INTRODUCTION

Dynamic programming (DP) is a computational approach to finding an optimal policy by employing the principle of optimality introduced by Richard Bellman [1]. Contemporary DP techniques make use of approximation and learning to address the “curse of dimensionality” when applying DP to large problems. These algorithms are now referred to as approximate dynamic programming (ADP). A number of the state-of-the-art ADP algorithms have been developed [2]. A natural tool for analyzing and designing DP and ADP algorithms is the use of Markov decision processes (MDPs). Under the setting of finite-state and finite-action MDPs, the algorithms of value iteration and policy iteration, which apply dynamic programming directly, are guaranteed to converge to an optimal policy. However, in practice, accurate knowledge represented in transition probabilities is difficult to obtain. Thus, exact solutions via classic dynamic programming are not attainable, and value iteration and policy iteration are not applicable anymore with guaranteed optimal solutions. Actually any small deviation from the estimation process of the transition probabilities may potentially generate significant impact on the solutions or may result in significant deviation of the solutions from the optimal values [5]. Hence, robust approximate dynamic programming is needed to address the question of how to use approximation method with an appropriate robustness built in to extend the power of the Bellman Equation.

In this paper, MDPs with uncertain transition matrices are discussed. Representative efforts in developing robust dynamic programming algorithms may be summarized in [5]- [9]. One commonly used principle of optimality criterion for robust algorithms is to minimize the maximum value function for any initial state, which generalizes the optimality criterion of minimizing the expected total discounted cost for MDPs with exact transition matrices. Based on this optimality criterion, the algorithms given in [5] and [6] are named as robust value iteration and robust policy iteration, respectively, since these algorithms follow similar steps to value iteration and policy iteration for MDPs with exact transition matrices. Note that robust policy iteration requires each transition probability row be estimated using a closed convex set.

To deal with uncertain transition matrices, the notion of correlation was first introduced in [5] and [9] in the context of MDPs. However, in this paper, mathematically clear and more tractable

definitions of independence and correlation of transition matrices are provided. Using these newly formulated definitions, existing optimality criteria and robust algorithms were developed only for MDPs with independent transition matrices. Based on the optimality criterion of minimizing the maximum value function for any initial state, the existence of a stationary deterministic optimal policy requires two conditions: (i) for any stationary deterministic policy, there is at least one among all possible stationary transition matrices such that the value function of the policy reaches maximum for any initial state; (ii) there is at least one stationary deterministic policy such that its maximum value function is less than or equal to the maximum value functions of all other policies for any initial state. These conditions are too strong to be satisfied for all MDPs with correlated transition matrices. Even if an optimal policy exists for an MDP with correlated transition matrices, robust value iteration and robust policy iteration may not be applicable to guarantee optimal solutions.

In this paper, using the reformulated definitions of independent transition matrices and correlated transition matrices, MDPs are classified into two types according to the properties of their corresponding transition matrices. Sufficient conditions are provided to guarantee the existence of a stationary deterministic optimal policy under the optimality criterion of minimizing the maximum value function for any initial state and the applicability of both robust value iteration and robust policy iteration, without the condition that the set estimation of each transition probability row is closed convex. Robust policy iteration is simplified in the policy evaluation step. An optimality criterion of minimizing the maximum squared total value function is proposed to guarantee the existence of a stationary deterministic optimal policy for MDPs with uncertain transition matrices under a weaker condition. This criterion generalizes the optimality criterion of minimizing the maximum value function for any initial state. Based on this new optimality criterion, a robust policy iteration under total value function is developed. It degenerates to robust policy iteration when the sufficient condition is satisfied. The optimality criterion and robust policy iteration under total value function can be used to solve problems of MDPs with uncertain cost functions and weighted squared total value functions.

The rest of the paper is organized as follows. Section 2 provides the problem formulation. In section 3, sufficient conditions are presented to guarantee that robust value iteration and robust policy iteration can be used. In section 4, an optimality criterion of minimizing the maximum squared total value function is proposed and it is proven that an optimal policy exists under a

weak condition and this optimality criterion generalizes the optimality criterion of minimizing the maximum value function for any initial state. Based on this optimality criterion, robust policy iteration under total value function is given in section 5. In section 6, two examples are given to illustrate how the optimality criterion and robust policy iteration under total value function work. The paper concludes in section 7.

## II. PROBLEM FORMULATION

A finite-state, finite-action, infinite-horizon MDP with stationary transition matrices is described as follows. Let  $T$  denote the discrete, infinite decision horizon, where  $T = \{0, 1, 2, \dots\}$ . At each stage, the system occupies a state  $i \in S$ , where  $S$  is the state space with  $n$  states and denoted as  $S = \{1, 2, \dots, n\}$ . A decision maker is allowed to choose an action  $a$  deterministically from a finite state-dependent set of allowable actions, denoted by  $\mathcal{A}_i = \{a_{1i}, a_{2i}, \dots, a_{m_i}\}$ . Let  $M = \sum_{i=1}^n m_i$ . Let “ $\mathbf{a}$ ” be a function mapping states into actions with  $\mathbf{a}(i) \in \mathcal{A}_i$ . Denote a stationary controller policy by  $\pi$ , i.e.,  $\pi = (\mathbf{a}, \mathbf{a}, \dots)$  and let  $\Pi$  represent the stationary deterministic controller policy space. Define the cost corresponding to state  $i \in S$  and action  $a \in \mathcal{A}_i$  by  $c(i, a)$ . Assume that  $c(i, a)$  is non-negative and finite. The costs are time discounted by the factor  $\gamma$  ( $0 < \gamma < 1$ ). The system starts from an initial state. The states make Markov transitions according to stationary transition probabilities  $p_{ij}^a$  from one state  $i$  to another state  $j$  under an action  $a$ . In a more general setting, let each transition probability  $p_{ij}^a$  ( $0 \leq p_{ij}^a \leq 1$ ) be represented by a function of a parameter vector denoted as  $\mathbf{U}_i^a$

$$p_{ij}^a \triangleq f_{ij}^a(\mathbf{U}_i^a) \quad \text{for any } i, j \in S, a \in \mathcal{A}_i, \quad (1)$$

where

$$\mathbf{U}_i^a = \left( u_{i_1}^a \quad \dots \quad u_{i_j}^a \quad \dots \quad u_{i_i}^a \right), \quad (2)$$

and

$$\sum_{j=1}^n p_{ij}^a = \sum_{j=1}^n f_{ij}^a(\mathbf{U}_i^a) = 1. \quad (3)$$

A special parameter vector  $\mathbf{U}_i^a$  is defined in [5]- [9], where

$$\mathbf{U}_i^a = \left( p_{i1}^a \quad \dots \quad p_{ij}^a \quad \dots \quad p_{i(n-1)}^a \right), \quad (4)$$

and thus

$$p_{ij}^a = f_{ij}^a(\mathbf{U}_i^a) = \begin{cases} p_{ij}^a & 1 \leq j \leq n-1 \\ 1 - \sum_{j=1}^{n-1} p_{ij}^a & j = n \end{cases}. \quad (5)$$

Let  $P_i^a$  be a transition probability row under the state  $i \in S$  and the action  $a \in \mathcal{A}_i$  of the transition matrix

$$P_i^a = (p_{i1}^a \cdots p_{ij}^a \cdots p_{in}^a) \triangleq (f_{i1}^a(\mathbf{U}_i^a) \cdots f_{ij}^a(\mathbf{U}_i^a) \cdots f_{in}^a(\mathbf{U}_i^a)). \quad (6)$$

Let  $\mathbf{f}_i^a(\mathbf{U}_i^a) = (f_{i1}^a(\mathbf{U}_i^a) \cdots f_{ij}^a(\mathbf{U}_i^a) \cdots f_{in}^a(\mathbf{U}_i^a))$ , and then

$$P_i^a \triangleq \mathbf{f}_i^a(\mathbf{U}_i^a). \quad (7)$$

Further, all transition probability rows constitute an  $M \times n$  transition matrix  $P$

$$P = \begin{pmatrix} P_1^{a_{11}} \\ \vdots \\ P_i^{a_{ji}} \\ \vdots \\ P_n^{a_{mn}} \end{pmatrix} \triangleq \begin{pmatrix} \mathbf{f}_1^{a_{11}}(\mathbf{U}_1^{a_{11}}) \\ \vdots \\ \mathbf{f}_i^{a_{ji}}(\mathbf{U}_i^{a_{ji}}) \\ \vdots \\ \mathbf{f}_n^{a_{mn}}(\mathbf{U}_n^{a_{mn}}) \end{pmatrix}. \quad (8)$$

Actually, let

$$\mathbf{U} = \begin{pmatrix} \mathbf{U}_1^{a_{11}} \\ \vdots \\ \mathbf{U}_i^{a_{ji}} \\ \vdots \\ \mathbf{U}_n^{a_{mn}} \end{pmatrix}. \quad (9)$$

Each transition probability is easily extended to be a function of  $\mathbf{U}$ ,

$$p_{ij}^a = \bar{f}_{ij}^a(\mathbf{U}) \triangleq f_{ij}^a(\mathbf{U}_i^a), \quad (10)$$

and then each transition probability row  $P_i^a$  and the transition matrix  $P$  can also be extended to be a function of  $\mathbf{U}$

$$P_i^a = \bar{\mathbf{f}}_i^a(\mathbf{U}) \triangleq (\bar{f}_{i1}^a(\mathbf{U}) \cdots \bar{f}_{ij}^a(\mathbf{U}) \cdots \bar{f}_{in}^a(\mathbf{U})), \quad (11)$$

$$P = \begin{pmatrix} P_1^{a_{11}} \\ \vdots \\ P_i^{a_{ji}} \\ \vdots \\ P_n^{a_{nn}} \end{pmatrix} \triangleq \begin{pmatrix} \bar{\mathbf{f}}_1^{a_{11}}(\mathbf{U}) \\ \vdots \\ \bar{\mathbf{f}}_i^{a_{ji}}(\mathbf{U}) \\ \vdots \\ \bar{\mathbf{f}}_n^{a_{nn}}(\mathbf{U}) \end{pmatrix}. \quad (12)$$

The  $n \times n$  transition matrix for a stationary controller policy  $\pi$  is denoted as

$$P^\pi = \begin{pmatrix} P_1^{\mathbf{a}(1)} \\ \vdots \\ P_i^{\mathbf{a}(i)} \\ \vdots \\ P_n^{\mathbf{a}(n)} \end{pmatrix} \triangleq \begin{pmatrix} \bar{\mathbf{f}}_1^{\mathbf{a}(1)}(\mathbf{U}) \\ \vdots \\ \bar{\mathbf{f}}_i^{\mathbf{a}(i)}(\mathbf{U}) \\ \vdots \\ \bar{\mathbf{f}}_n^{\mathbf{a}(n)}(\mathbf{U}) \end{pmatrix}. \quad (13)$$

Let the parameters in  $\mathbf{U}$  be uncertain, and let  $\mathbf{U}$  vary in a known subset of  $\mathfrak{R}^{|\mathbf{U}|}$  denoted as  $\mathcal{U}$ . The set estimation of the transition matrix  $P$  is expressed as

$$\mathcal{P} = \{P : \mathbf{U} \in \mathcal{U}\}. \quad (14)$$

The uncertain parameter vector  $\mathbf{U}$ , the uncertain transition matrix  $P$ , and the uncertain transition matrix for any stationary policy  $\pi$  denoted as  $P^\pi$  are regarded as variables.

We are now in a position to introduce the concepts of independence and correlation for  $\mathbf{U}$ ,  $P$  and  $P^\pi$ .

**Definition (Projection, Correlation, Independence):**

Let

$$Y = \begin{pmatrix} y_1 \\ \vdots \\ y_i \\ \vdots \\ y_l \end{pmatrix} \quad (15)$$

where  $y_i \in \mathfrak{R}^{n_i}$  and

$$\{y_{i_1}, \dots, y_{i_j}, \dots, y_{i_r}\} \subseteq \{y_1, \dots, y_i, \dots, y_l\} \quad (1 \leq i_1 < \dots < i_j < \dots < i_l \leq l).$$

Let the set estimation of  $Y$  be  $\mathcal{Y}$ , where  $\mathcal{Y} \subseteq \mathbb{R}^{n_1} \times \cdots \times \mathbb{R}^{n_i} \times \cdots \times \mathbb{R}^{n_l}$ .

The projection of  $\mathcal{Y}$  in the direction of  $\{y_{i_1}, \cdots, y_{i_j}, \cdots, y_{i_r}\}$  is denoted as  $\mathcal{Y}_{\{i_1 \cdots i_j \cdots i_r\}}$ , where

$$\mathcal{Y}_{\{i_1 \cdots i_j \cdots i_r\}} \triangleq \left\{ \left( \begin{array}{c} y_{i_1} \\ \vdots \\ y_{i_j} \\ \vdots \\ y_{i_r} \end{array} \right) \in \mathbb{R}^{n_{i_1}} \times \cdots \times \mathbb{R}^{n_{i_j}} \times \cdots \times \mathbb{R}^{n_{i_r}} : \begin{array}{l} \text{there exists a } y \in \mathcal{Y} \\ \text{such that the } i_j\text{-th element} \\ \text{of } y \text{ is } y_{i_j} \quad (1 \leq j \leq r) \end{array} \right\}. \quad (16)$$

Under this definition, the projection of  $\mathcal{Y}$  in the direction of  $\{y_i\}$  or  $y_i$  is denoted as  $\mathcal{Y}_{\{i\}}$  or  $\mathcal{Y}_i$  ( $1 \leq i \leq l$ ).

$Y$  is correlated or  $\{y_1, \cdots, y_i, \cdots, y_l\}$  are correlated if

$$\mathcal{Y} \subset \mathcal{Y}_1 \times \cdots \times \mathcal{Y}_i \times \cdots \times \mathcal{Y}_l, \quad (17)$$

where “ $\subset$ ” indicates proper subset from here on.

$Y$  is independent or  $\{y_1, \cdots, y_i, \cdots, y_l\}$  are independent if

$$\mathcal{Y} = \mathcal{Y}_1 \times \cdots \times \mathcal{Y}_i \times \cdots \times \mathcal{Y}_l. \quad (18)$$

**Definition (Correlated transition matrix, Independent transition matrix):**

The transition matrix  $P$  is correlated if

$$\mathcal{P} \subset \mathcal{P}_1^{a_{11}} \times \cdots \times \mathcal{P}_i^{a_{ji}} \times \cdots \times \mathcal{P}_n^{a_{nn}}, \quad (19)$$

where  $\mathcal{P}_i^{a_{ji}}$  is the projection of  $\mathcal{P}$  in the direction of  $P_i^{a_{ji}}$  for any state  $i \in S$  and any action  $a_{ji} \in \mathcal{A}_i$ .

The transition matrix  $P$  is independent if

$$\mathcal{P} = \mathcal{P}_1^{a_{11}} \times \cdots \times \mathcal{P}_i^{a_{ji}} \times \cdots \times \mathcal{P}_n^{a_{nn}}. \quad (20)$$

**Definition (Correlated transition matrix for a stationary controller policy, Independent transition matrix for a stationary controller policy):**

The transition matrix  $P^\pi$  for a stationary controller policy  $\pi = (\mathbf{a}, \mathbf{a}, \cdots) \in \Pi$  is correlated if

$$\mathcal{P}^\pi \subset \mathcal{P}_1^{\mathbf{a}(1)} \times \cdots \times \mathcal{P}_i^{\mathbf{a}(i)} \times \cdots \times \mathcal{P}_n^{\mathbf{a}(n)}, \quad (21)$$

where  $\mathcal{P}^\pi$  is the projection of  $\mathcal{P}$  in the direction of  $\{P_1^{\mathbf{a}(1)}, \dots, P_i^{\mathbf{a}(i)}, \dots, P_n^{\mathbf{a}(n)}\}$  and  $\mathcal{P}_i^{\mathbf{a}(i)}$  is the projection of  $\mathcal{P}$  in the direction of  $P_i^{\mathbf{a}(i)}$  for any state  $i \in S$ .

The transition matrix  $P^\pi$  for a stationary controller policy  $\pi = (\mathbf{a}, \mathbf{a}, \dots) \in \Pi$  is independent if

$$\mathcal{P}^\pi = \mathcal{P}_1^{\mathbf{a}(1)} \times \dots \times \mathcal{P}_i^{\mathbf{a}(i)} \times \dots \times \mathcal{P}_n^{\mathbf{a}(n)}. \quad (22)$$

**Remark:** (i) The set  $\mathcal{P}_1^{\mathbf{a}(1)} \times \dots \times \mathcal{P}_i^{\mathbf{a}(i)} \times \dots \times \mathcal{P}_n^{\mathbf{a}(n)}$  is a  $M$ -dimension hyper-rectangle. When the set estimation of the transition matrix  $P$  denoted as  $\mathcal{P}$  is a proper subset of this hyper-rectangle,  $P$  is correlated. When  $\mathcal{P}$  is equal to the hyper-rectangle,  $P$  is independent. (ii) An exact transition matrix  $P$  is a special case of an independent transition matrix. (iii) When  $\mathbf{U}$  defined in (9) is independent, i.e.,  $\mathcal{U} = \mathcal{U}_1^{\mathbf{a}(1)} \times \dots \times \mathcal{U}_i^{\mathbf{a}(i)} \times \dots \times \mathcal{U}_n^{\mathbf{a}(n)}$ , the transition matrix  $P$  is also independent. (iv) The set  $\mathcal{P}_1^{\mathbf{a}(1)} \times \dots \times \mathcal{P}_i^{\mathbf{a}(i)} \times \dots \times \mathcal{P}_n^{\mathbf{a}(n)}$  is a  $n$ -dimension hyper-rectangle. When the set estimation of the transition matrix  $P^\pi$  for a controller policy  $\pi$  denoted as  $\mathcal{P}^\pi$  is a proper subset of this hyper-rectangle,  $P^\pi$  is correlated. When  $\mathcal{P}^\pi$  is equal to the hyper-rectangle,  $P^\pi$  is independent. (v) When  $\{\mathbf{U}_1^{\mathbf{a}(1)}, \dots, \mathbf{U}_i^{\mathbf{a}(i)}, \dots, \mathbf{U}_n^{\mathbf{a}(n)}\}$  is independent, i.e., the projection of  $\mathbf{U}$  in the direction of  $\{\mathbf{U}_1^{\mathbf{a}(1)}, \dots, \mathbf{U}_i^{\mathbf{a}(i)}, \dots, \mathbf{U}_n^{\mathbf{a}(n)}\}$  is a  $n$ -dimension hyper-rectangle,  $P^\pi$  is also independent. (vi) Example 1 in section 6 illustrates the concepts of the projection of  $\mathcal{P}$  in the direction of some transition probability row  $P_i^{\mathbf{a}}$ , independent transition matrix, correlated transition matrix, independent transition matrix for some controller policy, and correlated transition matrix for some controller policy.

According to the properties of transition matrices, MDPs are thus classified into MDPs with independent transition matrices and MDPs with correlated transition matrices.

**Definition (A stationary nature policy):** A stationary nature policy refers to a specific collection of time-independent transition matrices chosen by nature [5], denoted as  $\tau$ , i.e.,  $\tau = (P, P, \dots)$ .

The set of stationary admissible nature policies is denoted as  $\mathcal{T}$

$$\mathcal{T} \triangleq \{\tau = (P, P, \dots) : P \in \mathcal{P}\}. \quad (23)$$

The optimality criterion of minimizing the maximum value function for any initial state has the following three equivalent expressions:

$$\min_{\pi \in \Pi} \max_{\tau \in \mathcal{T}} v_\tau^\pi(i) = v^*(i) \quad \text{for any initial state } i \in S, \quad (24)$$

where  $v_\tau^\pi(i)$  is the value function under the stationary controller policy  $\pi$  and the stationary nature policy  $\tau$  at the initial state  $i$

$$v_\tau^\pi(i) = \mathbf{E}_i^\pi \left( \sum_{t=0}^{\infty} \gamma^t c(j, a) | \tau \right); \quad (25)$$

or

$$\min_{\pi \in \Pi} \max_{P \in \mathcal{P}} v_P^\pi(i) = v^*(i) \quad \text{for any initial state } i \in S, \quad (26)$$

where  $v_P^\pi(i)$  is the value function under the stationary controller policy  $\pi$  and the transition matrix  $P$  at the initial state  $i$

$$v_P^\pi(i) = \mathbf{E}_i^\pi \left( \sum_{t=0}^{\infty} \gamma^t c(j, a) | P \right); \quad (27)$$

or

$$\min_{\pi \in \Pi} \max_{\mathbf{U} \in \mathcal{U}} v_{\mathbf{U}}^\pi(i) = v^*(i) \quad \text{for any initial state } i \in S, \quad (28)$$

where  $v_{\mathbf{U}}^\pi(i)$  is the value function under the controller policy  $\pi$  and the parameter vector  $\mathbf{U}$  at the initial state  $i$

$$v_{\mathbf{U}}^\pi(i) = \mathbf{E}_i^\pi \left( \sum_{t=0}^{\infty} \gamma^t c(j, a) | \mathbf{U} \right). \quad (29)$$

Next, a stationary optimal policy pair is defined, according to the optimality criterion given in (24)-(29).

**Definition (Stationary optimal policy pair):** A stationary policy pair  $(\pi^*, \tau^*)$  is optimal if

$$v_{\tau^*}^{\pi^*}(i) = \max_{\tau \in \mathcal{T}} v_\tau^{\pi^*}(i) = \min_{\pi \in \Pi} \max_{\tau \in \mathcal{T}} v_\tau^\pi(i) \quad \text{for any initial state } i \in S; \quad (30)$$

or equivalently,  $(\pi^*, P^*)$  is optimal if

$$v_{P^*}^{\pi^*}(i) = \max_{P \in \mathcal{P}} v_P^{\pi^*}(i) = \min_{\pi \in \Pi} \max_{P \in \mathcal{P}} v_P^\pi(i) \quad \text{for any initial state } i \in S, \quad (31)$$

where  $\tau^* = (P^*, P^*, \dots)$ ; or equivalently,  $(\pi^*, \mathbf{U}^*)$  is optimal if

$$v_{\mathbf{U}^*}^{\pi^*}(i) = \max_{\mathbf{U} \in \mathcal{U}} v_{\mathbf{U}}^{\pi^*}(i) = \min_{\pi \in \Pi} \max_{\mathbf{U} \in \mathcal{U}} v_{\mathbf{U}}^\pi(i) \quad \text{for any initial state } i \in S, \quad (32)$$

where  $P^*$  is the value at  $\mathbf{U}^*$ .

For MDPs with independent transition matrices, the existence of a stationary optimal policy pair has been shown and it can be computed by robust value iteration [5]. For MDPs with independent transition matrices under the condition that the set estimation of each transition

probability row defined in (11) is closed convex, the existence of a stationary optimal policy pair has been shown and it can be computed by robust policy iteration [6].

However, for MDPs with correlated transition matrices, it is more complicated. There are three possible scenarios: (i) a stationary optimal policy pair exists and both robust value iteration and robust policy iteration can be used for obtaining an optimal policy pair; (ii) a stationary optimal policy exists but neither robust value iteration nor robust policy iteration is useful; (iii) a stationary optimal policy pair does not exist at all. Example 1 given in section 6 is used to illustrate those scenarios.

Next we shed some light on why sometimes an optimal policy pair does not exist under the optimality criterion defined in (24)-(29) for MDPs with correlated transition matrices. The definition of a stationary optimal policy pair given by (30)-(32) can be interpreted as the conditions for the existence of a stationary optimal policy pair:

for any given  $\pi \in \Pi$ , there is, at least one  $\tau^\pi$  such that

$$v_{\tau^\pi}^\pi(i) = \max_{\tau \in \mathcal{T}} v_\tau^\pi(i) \quad \text{for any } i \in S; \quad (33)$$

and there is at least one  $\pi^* \in \Pi$  such that

$$v_{\tau^{\pi^*}}^{\pi^*}(i) = \min_{\pi \in \Pi} v_{\tau^\pi}^\pi(i) \quad \text{for any } i \in S. \quad (34)$$

We can see that a stationary optimal policy pair  $(\pi^*, \tau^{\pi^*})$  is independent of initial states. However, for MDPs with correlated transition matrices, initial states may affect optimal controller policies or optimal nature policies. Different initial states can correspond to different optimal controller policies or optimal nature policies. That contradicts the definition of a stationary optimal policy pair which is independent of initial states.

Based on the three possible scenarios for MDPs with correlated transition matrices, three prominent issues are present: (i) under what sufficient conditions a stationary optimal policy pair exists subject to the optimality criterion defined in (24)-(29), and furthermore, both robust value iteration and robust policy iteration can be used for obtaining a stationary optimal policy pair; (ii) what new optimality criterion can be proposed to guarantee the existence of a stationary optimal policy pair under relatively weak conditions as compared with those given by (33)-(34) with regard to the optimality criterion defined in (24)-(29); (iii) what efficient algorithm can be developed to obtain a stationary optimal policy pair based on the new optimality criterion.

### III. SUFFICIENT CONDITIONS FOR ROBUST VALUE ITERATION AND ROBUST POLICY ITERATION

In this section, sufficient conditions are developed for MDPs with correlated transition matrices to guarantee that a stationary optimal policy pair exists and that both robust value iteration and robust policy iteration can be used for obtaining a stationary optimal policy pair. First, three lemmas are presented, where Lemma 1 was proven in [5]. Then, two theorems are given for robust value iteration and robust policy iteration, respectively. Details of the proofs of these results are given in Appendices.

*Lemma 1:* For any given  $\pi = (\mathbf{a}, \mathbf{a}, \dots) \in \Pi$  and any given non-negative row vector  $q \in \mathfrak{R}_+^{1 \times n}$ , consider

$$\max_{v \in \mathfrak{R}^{n \times 1}} qv : v(i) \leq (g^\pi(v))_i \quad (35)$$

$$(g^\pi(v))_i \triangleq c(i, \mathbf{a}(i)) + \gamma \max_{P_i^{\mathbf{a}(i)} \in \mathcal{P}_i^{\mathbf{a}(i)}} P_i^{\mathbf{a}(i)} v \quad i \in S, \quad (36)$$

where  $\mathcal{P}_i^{\mathbf{a}(i)}$  is the projection of  $\mathcal{P}$  in the direction of  $P_i^{\mathbf{a}(i)}$ . Let the feasible solution space be denoted as  $\Omega_1$ . For any given non-negative row vector  $q \in \mathfrak{R}_+^{1 \times n}$ , consider

$$\max_{v \in \mathfrak{R}^{n \times 1}} qv : v(i) \leq (g(v))_i \quad (37)$$

$$(g(v))_i \triangleq \min_{a \in \mathcal{A}_i} \left( c(i, a) + \gamma \max_{P_i^a \in \mathcal{P}_i^a} P_i^a v \right) \quad i \in S, \quad (38)$$

where  $\mathcal{P}_i^a$  is the projection of  $\mathcal{P}$  in the direction of  $P_i^a$ . Let the feasible solution space be denoted as  $\Omega_2$ . Then, the functions  $g^\pi$  and  $g$  are monotone non-decreasing and contractive. The problems (35) and (37) have the unique optimal solutions denoted as  $v_\infty^\pi$  and  $v_\infty$ , which are the unique solutions to the fixed-point equations  $v = g^\pi(v)$  and  $v = g(v)$ , respectively. The optimal

transition probability rows for (35) are given by

$$\left(P_i^{\mathbf{a}(i)}\right)^* \in \arg \max_{P_i^{\mathbf{a}(i)} \in \mathcal{P}_i^{\mathbf{a}(i)}} \left\{P_i^{\mathbf{a}(i)} v_\infty^\pi\right\} \quad i \in S, \quad (39)$$

where the optimal transition matrix for  $\pi$  denoted as  $(P^\pi)^*$  is defined by

$$(P^\pi)^* \triangleq \begin{pmatrix} \left(P_1^{\mathbf{a}(1)}\right)^* \\ \vdots \\ \left(P_i^{\mathbf{a}(i)}\right)^* \\ \vdots \\ \left(P_n^{\mathbf{a}(n)}\right)^* \end{pmatrix} \in \mathcal{P}_1^{\mathbf{a}(1)} \times \cdots \times \mathcal{P}_i^{\mathbf{a}(i)} \times \cdots \times \mathcal{P}_n^{\mathbf{a}(n)}. \quad (40)$$

The optimal transition probability rows for (37) are given by

$$\left(P_i^a\right)^* \in \arg \max_{P_i^a \in \mathcal{P}_i^a} \left\{P_i^a v_\infty\right\} \quad i \in S, \quad a \in \mathcal{A}_i, \quad (41)$$

where the optimal transition matrix  $P^*$  is defined by

$$P^* \triangleq \begin{pmatrix} \left(P_1^{a_{11}}\right)^* \\ \vdots \\ \left(P_i^{a_{ji}}\right)^* \\ \vdots \\ \left(P_n^{a_{mn}}\right)^* \end{pmatrix} \in \mathcal{P}^{a_{11}} \times \cdots \times \mathcal{P}_i^{a_{ji}} \times \cdots \times \mathcal{P}_n^{a_{mn}}. \quad (42)$$

According to the fixed-point theorem,  $v_\infty^\pi$  and  $v_\infty$  are computed by the iterative processes given in Algorithm 1 and 2, respectively.

---

**Algorithm 1** Iteration Algorithm for Optimal Solution  $v_\infty^\pi$  of Problem (35)

---

1. select  $v_0^\pi \in \mathfrak{R}^{n \times 1}$  and set  $k = 0$ ;

2. compute  $v_{k+1}^\pi$  by

$$v_{k+1}^\pi = g^\pi(v_k^\pi); \quad (43)$$

3. terminate if  $v_{k+1}^\pi = v_k^\pi$  and output  $v_\infty^\pi = v_k^\pi$  with  $(P^\pi)^*$  given in (40); otherwise, set  $k = k + 1$  and go to 2.

---

---

**Algorithm 2** Iteration Algorithm for Optimal Solution  $v_\infty$  of Problem (37)
 

---

1. select  $v_0 \in \mathfrak{R}^{n \times 1}$  and set  $k = 0$ ;

2. compute  $v_{k+1}$  by

$$v_{k+1} = g(v_k); \quad (44)$$

3. terminate if  $v_{k+1} = v_k$  and output  $v_\infty = v_k$  with  $P^*$  given in (42); otherwise, set  $k = k + 1$  and go to 2.

---

**Remark:** In principle,  $v_0^\pi$  and  $v_0$  can be selected arbitrarily in  $\mathfrak{R}^{n \times 1}$  for Algorithm 1 and Algorithm 2, respectively. However, some special choices of  $v_0^\pi$  and  $v_0$  can accelerate the iterative convergence process. Let  $\mathcal{G}^\pi = \{v : v \geq g^\pi(v)\}$  and  $\mathcal{G} = \{v : v \geq g(v)\}$ . For (35), when  $v_0^\pi \in \Omega_1$ , the sequence  $\{v_k^\pi\}$  is non-decreasing, or when  $v_0^\pi \in \mathcal{G}^\pi$ , the sequence  $\{v_k^\pi\}$  is non-increasing, and thus  $\{v_k^\pi\}$  converges to  $v_\infty^\pi$  faster than initial values which are not in  $\Omega_1$  or in  $\mathcal{G}^\pi$ . For (37), when  $v_0 \in \Omega_2$ , the sequence  $\{v_k\}$  is non-decreasing, or when  $v_0 \in \mathcal{G}$ , the sequence  $\{v_k\}$  is non-increasing, and thus, the sequence  $\{v_k\}$  converges to  $v_\infty$  faster than initial values which are not in  $\Omega_2$  or in  $\mathcal{G}$ .

*Lemma 2:* For any given  $\pi = (\mathbf{a}, \mathbf{a}, \dots) \in \Pi$  and any given non-negative  $q \in \mathfrak{R}_+^{1 \times n}$ , consider

$$\max_{\tau \in T, v} qv = \max_{P^\pi \in \mathcal{P}^\pi, v} qv : v(i) \leq c(i, \mathbf{a}(i)) + \gamma P_i^{\mathbf{a}(i)} v \quad i \in S, \quad (45)$$

where  $\mathcal{P}^\pi$  is the projection of  $\mathcal{P}$  in the direction of  $\{P_1^{\mathbf{a}(1)}, \dots, P_i^{\mathbf{a}(i)}, \dots, P_n^{\mathbf{a}(n)}\}$ . Let the feasible solution space be denoted as  $\Omega_3$ . Then, for  $\pi \in \Pi$ , if there exists at least one  $(P^\pi)^*$  defined in (40) such that  $(P^\pi)^* \in \mathcal{P}^\pi$ , Algorithm 1 for problem (35) can be used for problem (45) to obtain the optimal solution  $v_\infty^\pi \in \Omega_3$ .

*Lemma 3:* For any given non-negative  $q \in \mathfrak{R}_+^{1 \times n}$ , consider

$$\max_{\tau \in T, v} qv = \max_{P \in \mathcal{P}, v} qv : v(i) \leq c(i, a) + \gamma P_i^a v \quad i \in S, \quad a \in \mathcal{A}_i. \quad (46)$$

Let the feasible solution space be denoted as  $\Omega_4$ . Then, if there exists at least one  $P^*$  defined

in (42) such that  $P^* \in \mathcal{P}$ , Algorithm 2 for problem (37) can be used for problem (46) to obtain the optimal solution  $v_\infty \in \Omega_4$ .

*Corollary 1:* When the transition matrix for a controller policy  $\pi \in \Pi$ ,  $P^\pi$  defined by (13), is independent in the set  $\mathcal{P}^\pi$ , Algorithm 1 for problem (35) can be used for problem (45).

*Corollary 2:* When the transition matrix  $P$  defined by (12) is independent in the set  $\mathcal{P}$ , Algorithm 1 and Algorithm 2 for problems (35) and (37) can be used for problems (45) and (46), respectively.

*Theorem 1:* When for any  $\pi \in \Pi$ , there exists at least one  $(P^\pi)^*$  defined in (40) such that  $(P^\pi)^* \in \mathcal{P}^\pi$ , and there exists at least one  $P^*$  defined in (42) such that  $P^* \in \mathcal{P}$ , a stationary optimal policy pair exists under the optimality criterion of minimizing the maximum value function for any initial state defined in (24)-(29), and robust value iteration can be used for obtaining a stationary optimal policy pair.

*Theorem 2:* When for any  $\pi \in \Pi$ , there exists at least one,  $(P^\pi)^*$  defined in (40) such that  $(P^\pi)^* \in \mathcal{P}^\pi$ , and there exists at least one  $P^*$  defined in (42) such that  $P^* \in \mathcal{P}$ , a stationary optimal policy pair exists under the optimality criterion of minimizing the maximum value function for any initial state defined in (24)-(29), and robust policy iteration can be used for obtaining a stationary optimal policy pair in finite iterations.

Robust policy iteration in [6] requires the condition that the set estimation of each transition probability row defined by (11) is closed convex. Actually, it can be used without this condition, and it is proven in an easier way than in [6] as shown in the appendix. Furthermore, the policy evaluation step in robust policy iteration can be simplified as shown in Algorithm 1, i.e., for any  $\pi \in \Pi$ , if  $(P^\pi)^* \in \mathcal{P}^\pi$ , the maximum value function for any initial state can be computed by Algorithm 1.

Using the above two theorems, robust value iteration and robust policy iteration are ready to be applied to address MDPs with correlated transition matrices.

Specially, consider for any  $\pi \in \Pi$ ,  $(P^\pi)^* \in \mathcal{P}^\pi$ , and robust value iteration or robust policy iteration can be used to obtain an optimal controller policy  $\pi^*$  and all possible optimal nature policies  $P^*$ . This implies that with a one step assessment for the existence of at least one optimal nature policy  $P^*$  such that  $P^* \in \mathcal{P}$ , according to Theorem 1 and Theorem 2, robust value iteration and robust policy iteration algorithms will lead to an optimal solution. On the other hand, if there is no  $P^* \in \mathcal{P}$ , then there is no guarantee that robust value iteration and robust policy iteration algorithms will lead to an optimal solution. In Example 1 given in section 6, for case (i), any  $P^\pi$  and  $P$  are independent and thus both robust value iteration and robust policy iteration can be used. For cases (ii), (iii) and (iv), any  $P^\pi$  is independent, but  $P$  is correlated. Using robust value iteration and robust policy iteration, an optimal controller policy  $\pi^*$  and the unique optimal nature policy  $\tau^*$  with  $P^*$  are obtained. Since  $P^* \in \mathcal{P}$  for cases (ii) and (iii), both robust value and policy iterations can be used. However, for case (iv),  $P^* \notin \mathcal{P}$  and thus neither robust value iteration nor robust policy iteration can be used.

#### IV. OPTIMALITY CRITERION USING THE SQUARED TOTAL VALUE FUNCTION

An optimality criterion of minimizing the maximum squared total value function is proposed to deal with non-optimal solutions subject to the optimality criterion defined in (24)-(29). First, the optimality criterion and a stationary optimal policy pair based on this optimality criterion are defined. Then, two theorems are given to show that a stationary optimal policy pair exists under a weak condition and this optimality criterion generalizes the optimality criterion of minimizing the maximum value function for any initial state.

Given a transition matrix  $P$  or its parameter vector  $\mathbf{U}$ , the total value function for a controller policy  $\pi$  is defined as follows

$$\|v_P^\pi\| = \sqrt{(v_P^\pi)' v_P^\pi}, \quad (47)$$

where in terms of the value function defined in (27)

$$v_P^\pi = \left( v_P^\pi(1) \quad \cdots \quad v_P^\pi(i) \quad \cdots \quad v_P^\pi(n) \right)'; \quad (48)$$

or equivalently,

$$\|v_{\mathbf{U}}^\pi\| = \sqrt{(v_{\mathbf{U}}^\pi)' v_{\mathbf{U}}^\pi}, \quad (49)$$

where in terms of the value function defined in (29),

$$v_{\mathbf{U}}^{\pi} = \left( v_{\mathbf{U}}^{\pi}(1) \quad \cdots \quad v_{\mathbf{U}}^{\pi}(i) \quad \cdots \quad v_{\mathbf{U}}^{\pi}(n) \right)'. \quad (50)$$

The optimality criterion of minimizing the maximum squared total value function is

$$\min_{\pi \in \Pi} \max_{P \in \mathcal{P}} \|v_P^{\pi}\|^2, \quad (51)$$

or equivalently,

$$\min_{\pi \in \Pi} \max_{\mathbf{U} \in \mathcal{U}} \|v_{\mathbf{U}}^{\pi}\|^2. \quad (52)$$

**Definition (Stationary optimal policy pair):** A stationary policy pair  $(\pi^*, P^*)$  is optimal if

$$\|v_{P^*}^{\pi^*}\|^2 = \max_{P \in \mathcal{P}} \|v_P^{\pi^*}\|^2 = \min_{\pi \in \Pi} \max_{P \in \mathcal{P}} \|v_P^{\pi}\|^2, \quad (53)$$

where  $\tau^* = (P^*, P^*, \dots)$ ; or equivalently,  $(\pi^*, \mathbf{U}^*)$  is optimal if

$$\|v_{\mathbf{U}^*}^{\pi^*}\|^2 = \max_{\mathbf{U} \in \mathcal{U}} \|v_{\mathbf{U}}^{\pi^*}\|^2 = \min_{\pi \in \Pi} \max_{\mathbf{U} \in \mathcal{U}} \|v_{\mathbf{U}}^{\pi}\|^2, \quad (54)$$

where  $P^*$  is the value at  $\mathbf{U}^*$ .

The optimality criterion under the squared total value function defined in (51)-(52) no longer takes into account of initial states explicitly, or in other words, a stationary optimal policy pair defined in (53)-(54) can be regarded as being independent of initial states. In a stationary optimal policy pair, the optimal nature policy may depend on the optimal controller policy.

*Theorem 3:* Assume that for any  $\pi$ , there exists at least one  $\mathbf{U}^{\pi} \in \mathcal{U}$  such that  $\|v_{\mathbf{U}^{\pi}}^{\pi}\|^2 = \max_{\mathbf{U} \in \mathcal{U}} \|v_{\mathbf{U}}^{\pi}\|^2$ . Then a stationary optimal policy pair  $(\pi^*, P^*)$  exists with regard to the optimality criterion defined in (51)-(52).

By Theorem 3, a stationary optimal policy pair exists under a weaker condition than the one under the optimality criterion of minimizing the maximum value function for any initial state defined by (24)-(29). For Example 1 in section 6, a stationary optimal policy pair exists for all cases under the optimality criterion defined in (51)-(52). Actually, the optimality criterion defined by (24)-(29) is a special case of the optimality criterion defined by (51)-(52).

*Theorem 4:* When a stationary optimal policy pair exists with regard to the optimality criterion of minimizing the maximum value function for any initial state defined in (24)-(29), this optimality criterion is equivalent to the optimality criterion of minimizing the maximum squared total value function defined in (51)-(52), i.e., a stationary optimal policy pair denoted by  $(\pi_{pre}^*, P_{pre}^*)$  under the optimality criterion defined in (24)-(29) is equivalent to a stationary optimal policy pair denoted by  $(\pi_{new}^*, P_{new}^*)$  under the optimality criterion defined in (51)-(52).

## V. ROBUST POLICY ITERATION UNDER TOTAL VALUE FUNCTION

Based on the optimality criterion of minimizing the maximum squared total value function given in (51)-(52), MDPs can be solved by the method provided in the proof of Theorem 3 (refer to the Appendix). However, the computational complexity of this approach may be prohibiting. Computing the  $M^n$  values of  $U^\pi$  is demanding when  $M$  or  $n$  is large, where  $\|v_{U^\pi}^\pi\|^2 = \max_{U \in \mathcal{U}} \|v_U^\pi\|^2$ . A robust policy iteration under total value function is thus proposed to reduce the computation complexity.

Important steps in policy iteration include policy evaluation and policy improvement.

For policy evaluation of any policy  $\pi$ , there are two approaches: one is to compute the maximum squared total value function by inversion and maximization

$$\|v_{U^\pi}^\pi\|^2 = \max_{U \in \mathcal{U}} \|v_U^\pi\|^2 = \max_{U \in \mathcal{U}} (C^\pi)' \left( I - \gamma (P_U^\pi)' \right)^{-1} (I - \gamma P_U^\pi)^{-1} C^\pi, \quad (55)$$

which is referred to as the direct method in this paper; the other is to use Algorithm 1 to compute the maximum value function for any initial state denoted as  $v_{U^\pi}^\pi(i)$  ( $i \in S$ ), where  $U^\pi$  is found from  $(P^\pi)^*$  and then sum all squared  $v_{U^\pi}^\pi(i)$  to obtain the maximum squared total value function, which is referred to as the iterative method. The latter is preferred since it avoids computing an inversion and a maximum value. However, it requires the condition that  $(P^\pi)^* \in \mathcal{P}$  in Lemma 2 to be satisfied, which is not always guaranteed.

For policy improvement, in robust policy iteration, the policy can be improved easily by

$$\mathbf{a}_{k+1}(i) \in \arg \min_{a \in \mathcal{A}_i} \{c(i, a) + \gamma \max_{P_i^a \in \mathcal{P}_i^a} P_i^a v_k\}, \quad (56)$$

where  $v_k$  is a vector including all the maximum value function for any initial state at the  $k$ -th iteration (see robust policy iteration in Appendix). However, it requires the conditions stated in

Theorem 2. Hence, controller policy elimination method is considered as a means of improving policy. The controller policy elimination is that some suboptimal controller policies will be eliminated in each iteration by a necessary condition of being an optimal controller policy. The inequality  $\|v_{\mathbf{U}^{\pi_k}}^{\pi}\|^2 \leq \|v_{\mathbf{U}^{\pi_k}}^{\pi_k}\|^2$  given in (60) is such a necessary condition for a stationary controller policy  $\pi$  being optimal, where  $\|v_{\mathbf{U}^{\pi_k}}^{\pi_k}\|^2$  is the maximum squared total value function of  $\pi_k$  and  $\|v_{\mathbf{U}^{\pi_k}}^{\pi}\|^2$  is the squared total value function of  $\pi$  at the parameter vector  $\mathbf{U}^{\pi_k}$ .

Robust policy iteration under total value function is thus described in Algorithm 3.

**Remarks:**

- (i) Robust policy iteration under total value function terminates in finite iterations since  $|\Pi_0|$  is finite and  $\Pi_0 \supset \Pi_1 \supset \Pi_2 \supset \Pi_3 \cdots$ ;
- (ii) The sequence  $\{\pi_k\}$  converges to an optimal controller policy;
- (iii) Robust policy iteration under total value function can reduce computation complexity since it requires at most  $(1 + |\Pi_1|)$  values of  $\mathbf{U}^{\pi_k}$ ;
- (iv) A good initial controller policy  $\pi_0$ , which has relatively small total value function, can reduce the number of  $|\Pi_1|$  denoted as  $|\Pi_1|$  and accelerate the convergence;
- (v) When the sufficient conditions given in Theorem 2 are satisfied, robust policy iteration under total value function degenerates to the simplified robust policy iteration;
- (vi) All cases given in Example 1 in section 6 can be solved by robust policy iteration under total value function;
- (vii) When cost functions contain uncertain parameters  $\mathbf{U}$  and they are denoted as  $c(i, a, \mathbf{U})$ , robust policy iteration under total value function can be modified using the squared total value function

$$\|v_{\mathbf{U}}^{\pi}\|^2 = (C_{\mathbf{U}}^{\pi})' (I - \gamma (P_{\mathbf{U}}^{\pi})')^{-1} (I - \gamma P_{\mathbf{U}}^{\pi})^{-1} C_{\mathbf{U}}^{\pi}, \quad (62)$$

where  $C_{\mathbf{U}}^{\pi} = (c(1, \mathbf{a}(1), \mathbf{U}), \dots, c(n, \mathbf{a}(n), \mathbf{U}))'$ ;

- (viii) When for any policy  $\pi$ , the probability distribution of the initial states is non-uniform, or the value functions for different initial states have different weights, a semi-positive definite real matrix related to  $\pi$ ,  $Q^{\pi}$ , is introduced as the weighted matrix for  $n$  value functions, robust policy iteration under total value function can be modified using the weighted squared total value

---

**Algorithm 3** Robust Policy Iteration Under Total Value Function
 

---

**1.** Initialization: set  $k = 0$ ,  $\Pi_k = \Pi$ ,  $\mathcal{O} = +\infty$  and select  $\pi_k = \{\mathbf{a}_k, \mathbf{a}_k, \dots\}$ ;

**2.** Policy evaluation:

if sufficient condition given in Lemma 2 for  $\pi_k$  is satisfied,

(a) use the iterative method to compute  $v_{\mathbf{U}^{\pi_k}}^{\pi_k}$ ,  $\mathbf{U}^{\pi_k}$  and  $\|v_{\mathbf{U}^{\pi_k}}^{\pi_k}\|^2$  such that

$$v_{\mathbf{U}^{\pi_k}}^{\pi_k}(i) = \max_{\mathbf{U} \in \mathcal{U}} v_{\mathbf{U}}^{\pi_k}(i) \quad \text{for any initial state } i, \quad (57)$$

$$\|v_{\mathbf{U}^{\pi_k}}^{\pi_k}\|^2 = \max_{\mathbf{U} \in \mathcal{U}} \|v_{\mathbf{U}}^{\pi_k}\|^2 = \sum_{i=1}^n (v_{\mathbf{U}^{\pi_k}}^{\pi_k}(i))^2; \quad (58)$$

else

(b) use the direct method to compute  $v_{\mathbf{U}^{\pi_k}}^{\pi_k}$ ,  $\mathbf{U}^{\pi_k}$  and  $\|v_{\mathbf{U}^{\pi_k}}^{\pi_k}\|^2$  such that

$$\|v_{\mathbf{U}^{\pi_k}}^{\pi_k}\|^2 = \max_{\mathbf{U} \in \mathcal{U}} \|v_{\mathbf{U}}^{\pi_k}\|^2; \quad (59)$$

**3.** Policy improvement:

(a) eliminate controller policies to obtain  $\Pi'_k$

$$\Pi'_k = \{\pi \in \Pi_k : \|v_{\mathbf{U}^{\pi_k}}^{\pi}\|^2 \leq \|v_{\mathbf{U}^{\pi_k}}^{\pi_k}\|^2\}; \quad (60)$$

if  $|\Pi'_k| > 1$

if sufficient conditions given in Theorem 2 are satisfied

(b) set  $\Pi_{k+1} = \Pi'_k$  and  $\mathcal{O} = \|v_{\mathbf{U}^{\pi_k}}^{\pi_k}\|^2$  and select  $\pi_{k+1} = \{\mathbf{a}_{k+1}, \mathbf{a}_{k+1}, \dots\} \in \Pi_{k+1}$  by

$$\mathbf{a}_{k+1}(i) \in \arg \min_{a \in \mathcal{A}_i} \{c(i, a) + \gamma \max_{P_i^a \in \mathcal{P}_i^a} P_i^a v_{\mathbf{U}^{\pi_k}}^{\pi_k}\}; \quad (61)$$

if  $\pi_{k+1} = \pi_k$ , go to **4**; otherwise, set  $k = k + 1$  and go to **2**;

else

(c) if  $\|v_{\mathbf{U}^{\pi_k}}^{\pi_k}\|^2 < \mathcal{O}$ , set  $\mathcal{O} = \|v_{\mathbf{U}^{\pi_k}}^{\pi_k}\|^2$  and  $\Pi_{k+1} = \Pi'_k$ , and then select  $\pi_{k+1} \neq \pi_k \in \Pi_{k+1}$  and set  $k = k + 1$  and go to **2**; otherwise, select  $\pi'_k \in \Pi'_k - \{\pi_k\}$  and set  $\Pi_k = \Pi'_k - \{\pi_k\}$  and  $\pi_k = \pi'_k$  and go to **2**;

else

(d) go to **4**;

**4.** Termination: compute  $P^*$  with  $\mathbf{U}^{\pi_k}$  and output the stationary optimal policy pair  $(\pi_k, P^*)$  and the corresponding maximum squared total value function  $\mathcal{O}$ .

---

function

$$\|v_{\mathbf{U}}^{\pi}\|_{Q^{\pi}}^2 = (v_{\mathbf{U}}^{\pi})' Q^{\pi} v_{\mathbf{U}}^{\pi} = (C_{\mathbf{U}}^{\pi})' \left( I - \gamma (P_{\mathbf{U}}^{\pi})' \right)^{-1} Q^{\pi} \left( I - \gamma P_{\mathbf{U}}^{\pi} \right)^{-1} C_{\mathbf{U}}^{\pi}. \quad (63)$$

## VI. EXAMPLES

In this section, two examples are given. Example 1 illustrates the scenario of MDPs with independent transition matrices and three possible scenarios of MDPs with uncertain transition matrices under the optimality criterion defined by (24)-(29). Example 2 shows how the optimality criterion defined by (51)-(52) and robust policy iteration under total value function given by Algorithm 3 work.

**Example 1:** Consider a two-state, two-action, infinite-horizon MDP with the optimality criterion defined by (24)-(29). Let the state space be  $S = \{1, 2\}$ . The action spaces at state 1 and state 2 are the same, i.e.,  $\mathcal{A}_1 = \mathcal{A}_2 = \{a_1, a_2\}$ . Hence, there are four possible stationary deterministic controller policies, i.e.,

$$\Pi = \{\pi_1 = (\mathbf{a}_1, \mathbf{a}_1, \dots), \pi_2 = (\mathbf{a}_2, \mathbf{a}_2, \dots), \pi_3 = (\mathbf{a}_3, \mathbf{a}_3, \dots), \pi_4 = (\mathbf{a}_4, \mathbf{a}_4, \dots)\}, \quad (64)$$

where

$$\mathbf{a}_1(1) = a_1, \mathbf{a}_1(2) = a_1; \quad (65)$$

$$\mathbf{a}_2(1) = a_1, \mathbf{a}_2(2) = a_2; \quad (66)$$

$$\mathbf{a}_3(1) = a_2, \mathbf{a}_3(2) = a_1; \quad (67)$$

$$\mathbf{a}_4(1) = a_2, \mathbf{a}_4(2) = a_2. \quad (68)$$

The transition matrix  $P$  has the following formulation

$$P = \begin{pmatrix} P_1^{a_1} \\ P_1^{a_2} \\ P_2^{a_1} \\ P_2^{a_2} \end{pmatrix} = \begin{pmatrix} u_1 & 1 - u_1 \\ u_3 & 1 - u_3 \\ 1 - u_2^2 & u_2^2 \\ 1 - u_4 & u_4 \end{pmatrix} \quad (69)$$

The cost functions are as follows

$$c(1, a_1) = 1, \quad c(1, a_2) = 2, \quad c(2, a_1) = 3, \quad c(2, a_2) = 4 \quad (70)$$

The discount factor  $\gamma$  is 0.9. Let  $\mathbf{U}$  be the unknown parameter vector,

$$\mathbf{U} = (u_1 \quad u_2 \quad u_3 \quad u_4), \quad (71)$$

where  $\mathcal{U}$  is the set estimation of  $\mathbf{U}$ . Let  $\mathcal{W} = \{0, 0.2, 0.4, 0.6, 0.8, 1\}$ .

(i) When  $\mathcal{U}_1 = \{\mathbf{U} : u_1, u_2, u_3, u_4 \in \mathcal{W}\}$ , the projection of  $\mathcal{U}_1$  in the direction of  $u_i$  ( $i = 1, 2, 3, 4$ ) is  $\mathcal{W}$ .  $\mathcal{U}_1 = \mathcal{W} \times \mathcal{W} \times \mathcal{W} \times \mathcal{W}$ , i.e.,  $\mathbf{U}$  is independent. Hence,  $P$  is independent. The set estimation of  $P$ , denoted by  $\mathcal{P}_1$ , has the property

$$\mathcal{P}_1 = \mathcal{P}_1^{a_1} \times \mathcal{P}_1^{a_2} \times \mathcal{P}_2^{a_1} \times \mathcal{P}_2^{a_2}, \quad (72)$$

where  $\mathcal{P}_1^{a_1}, \mathcal{P}_1^{a_2}, \mathcal{P}_2^{a_1}$  and  $\mathcal{P}_2^{a_2}$  are the corresponding projections

$$\mathcal{P}_1^{a_1} = \mathcal{P}_1^{a_2} = \mathcal{P}_2^{a_2} = \{(0, 1) \ (0.2, 0.6) \ (0.4, 0.6) \ (0.6, 0.4) \ (0.8, 0.2) \ (1, 0)\} \quad (73)$$

$$\mathcal{P}_2^{a_1} = \{(1, 0) \ (0.96, 0.04) \ (0.84, 0.16) \ (0.64, 0.36) \ (0.36, 0.64) \ (0, 1)\}. \quad (74)$$

The transition matrices for four policies  $P^{\pi_1}, P^{\pi_2}, P^{\pi_3}$  and  $P^{\pi_4}$  are independent. A stationary optimal policy pair exists and both robust value iteration and robust policy iteration can be used for a stationary optimal policy pair, where stationary optimal policy pairs are unique. Using these algorithms, the optimal controller policy  $\pi^*$  is

$$\pi^* = (\mathbf{a}^*, \mathbf{a}^*, \dots) \quad (75)$$

where  $\mathbf{a}^*(1) = a_1, \mathbf{a}^*(2) = a_1$ , and the optimal nature policy  $\tau^*$  is

$$\tau^* = (P^*, P^*, \dots), \quad (76)$$

where

$$P^* = \begin{pmatrix} 0 & 1 \\ 0 & 1 \\ 0 & 1 \\ 0 & 1 \end{pmatrix}. \quad (77)$$

(ii) When  $\mathcal{U}_2 = \{\mathbf{U} : u_1 = u_3, u_1, u_2, u_4 \in \mathcal{W}\}$ , the projection of  $\mathcal{U}_2$  in the direction of  $u_i$  ( $i = 1, 2, 3, 4$ ) is  $\mathcal{W}$ . Since  $\mathcal{U}_2 \subset \mathcal{W} \times \mathcal{W} \times \mathcal{W} \times \mathcal{W}$ ,  $\mathbf{U}$  is correlated. This results in correlated  $P$ . The set estimation of  $P$ , denoted by  $\mathcal{P}_2$ , has the property

$$\mathcal{P}_2 \subset \mathcal{P}_1^{a_1} \times \mathcal{P}_1^{a_2} \times \mathcal{P}_2^{a_1} \times \mathcal{P}_2^{a_2}. \quad (78)$$

The transition matrices for four policies  $P^{\pi_1}, P^{\pi_2}, P^{\pi_3}$  and  $P^{\pi_4}$  are independent. A stationary optimal policy pair exists and both robust value and policy iterations can be used. The optimal controller policy and the optimal nature policy are the same as in case (i).

(iii) When  $\mathcal{U}_3 = \{\mathbf{U} : u_1 = u_3, u_2 = u_4, u_1, u_4 \in \mathcal{W}\}$ , the projection of  $\mathcal{U}_3$  in the direction of  $u_i$  ( $i = 1, 2, 3, 4$ ) is  $\mathcal{W}$ . Since  $\mathcal{U}_3 \subseteq \mathcal{U}_2$ ,  $P$  is correlated. The set estimation of  $P$ , denoted by  $\mathcal{P}_3$ , has the property

$$\mathcal{P}_3 \subseteq \mathcal{P}_2. \quad (79)$$

The transition matrices for four policies  $P^{\pi_1}$ ,  $P^{\pi_2}$ ,  $P^{\pi_3}$  and  $P^{\pi_4}$  are independent. A stationary optimal policy pair exists and both robust value and policy iterations can be used. The optimal controller policy and the optimal nature policy are the same as in case (i).

(iv) When

$$\mathcal{U}_4 = \left\{ \mathbf{U} : u_1, u_2 \in \mathcal{W}, u_3 = \begin{cases} 0 & u_1 = 1 \\ u_1 + 0.2 & u_1 \neq 1 \end{cases}, u_4 = \begin{cases} 0 & u_2 = 1 \\ u_2 + 0.2 & u_2 \neq 1 \end{cases} \right\}, \quad (80)$$

the projection of  $\mathcal{U}_4$  in the direction of  $u_i$  ( $i = 1, 2, 3, 4$ ) is  $\mathcal{W}$ . Since  $\mathcal{U}_4 \subset \mathcal{W} \times \mathcal{W} \times \mathcal{W} \times \mathcal{W}$ ,  $\mathbf{U}$  is correlated. This results in correlated  $P$ . The set estimation of  $P$ , denoted by  $\mathcal{P}_4$ , has the property

$$\mathcal{P}_4 \subset \mathcal{P}_1^{a_1} \times \mathcal{P}_1^{a_2} \times \mathcal{P}_2^{a_1} \times \mathcal{P}_2^{a_2}. \quad (81)$$

The transition matrices for four policies  $P^{\pi_1}$ ,  $P^{\pi_2}$ ,  $P^{\pi_3}$  and  $P^{\pi_4}$  are independent. A stationary optimal policy pair exists. There are 36 possible transition matrices since  $|\mathcal{U}_4| = 36$ . Computing all value functions for four stationary policies and two initial states under these transition matrices by the equation (125) and comparing them, optimal policy pairs can be obtained. An optimal controller policy is

$$\pi^* = (\mathbf{a}^*, \mathbf{a}^*, \dots) \quad (82)$$

where  $\mathbf{a}^*(1) = a_1$ ,  $\mathbf{a}^*(2) = a_1$  and the corresponding optimal nature policy  $\tau^*$  is

$$\tau^* = (P^*, P^*, \dots), \quad (83)$$

where

$$P^* = \begin{pmatrix} 0 & 1 \\ 0.2 & 0.8 \\ 0 & 1 \\ 1 & 0 \end{pmatrix}. \quad (84)$$

However, because

$$\begin{pmatrix} 0 & 1 \\ 0 & 1 \\ 0 & 1 \\ 0 & 1 \end{pmatrix} \notin \mathcal{P}_4, \quad (85)$$

neither robust value iteration nor robust policy iteration can be used to obtain an optimal policy pair under the optimality criterion given by (24)-(29).

(v) When

$$\mathcal{U}_5 = \{\mathbf{U} : u_1 = u_2, u_1, u_3, u_4 \in \mathcal{W}\}, \quad (86)$$

the projection of  $\mathcal{U}_5$  in the direction  $u_i$  ( $i = 1, 2, 3, 4$ ) is  $\mathcal{W}$ . Since  $\mathcal{U}_5 \subset \mathcal{W} \times \mathcal{W} \times \mathcal{W} \times \mathcal{W}$ ,  $\mathbf{U}_5$  is correlated. This results in correlated  $P$ . The set estimation of  $P$ , denoted by  $\mathcal{P}_5$ , has the property

$$\mathcal{P}_5 \subset \mathcal{P}_1^{a_1} \times \mathcal{P}_1^{a_2} \times \mathcal{P}_2^{a_1} \times \mathcal{P}_2^{a_2}. \quad (87)$$

The transition matrix for  $\pi_1$  denoted as  $P^{\pi_1}$  is correlated and the others are independent. All value functions for four stationary policies and two initial states under these transition matrices can be computed by the equation (125). Since for  $\pi_1$ , there is no  $P \in \mathcal{P}_5$  such that the maximum value function for any initial state is reachable, According to the condition given in (33) of the existence of optimal policy pairs, a stationary optimal policy pair does not exist under the optimality criterion given by (24)-(29).

(vi) When

$$\mathcal{U}_6 = \{\mathbf{U} : u_1 = u_2, u_3 = u_4, u_1, u_4 \in \mathcal{W}\}, \quad (88)$$

the projection of  $\mathcal{U}_6$  in the direction of  $u_i$  ( $i = 1, 2, 3, 4$ ) is  $\mathcal{W}$ . Since  $\mathcal{U}_6 \subseteq \mathcal{U}_5$ ,  $P$  is correlated. The set estimation of  $P$ , denoted by  $\mathcal{P}_6$ , has the property

$$\mathcal{P}_6 \subseteq \mathcal{P}_5. \quad (89)$$

the transition matrices for  $\pi_1$  and  $\pi_4$  denoted as  $P^{\pi_1}$  and  $P^{\pi_4}$  are correlated and the others are independent. All value functions for four stationary policies and two initial states under these transition matrices can be computed by the equation (125). Since for  $\pi_1$  and  $\pi_4$ , there is no  $P \in \mathcal{P}_6$  such that the maximum value function for any initial state is reachable, according to the condition given in (33) of the existence of optimal policy pairs, a stationary optimal policy

pair does not exist under the optimality criterion given by (24)-(29).

(vii) When

$$\mathcal{U}_7 = \{\mathbf{U} : u_1 = u_2 = u_3 = u_4, u_1 \in \mathcal{W}\}, \quad (90)$$

the projection of  $\mathcal{U}_7$  in the direction of  $u_i$  ( $i = 1, 2, 3, 4$ ) is  $\mathcal{W}$ , and  $\mathcal{U}_7 \subseteq \mathcal{U}_6 \subseteq \mathcal{U}_5$ . The set estimation of  $P$ , denoted by  $\mathcal{P}_7$ , has the property

$$\mathcal{P}_7 \subseteq \mathcal{P}_6 \subseteq \mathcal{P}_5. \quad (91)$$

The transition matrices for four policies are also correlated. All value functions for four stationary policies and two initial states under these transition matrices can be computed by the equation (125). Since for  $\pi_1$  and  $\pi_2$ , there is no  $P \in \mathcal{P}_7$  such that the maximum value function for any initial state is reachable, according to the condition given in (33) of the existence of optimal policy pairs, a stationary optimal policy pair does not exist under the optimality criterion given by (24)-(29).

However, under the optimality criterion under total value function defined by (51)-(52), all of cases given in Example 1 have a stationary optimal policy pair and can be solved by robust policy iteration under total value function given by Algorithm 3. The stationary optimal policy pairs for case (i), (ii), (iii) and (iv) are the same as the ones given in Example 1. For cases (v), (vi) and (vii), a stationary optimal controller policy  $\pi^*$  is

$$\pi^* = (\mathbf{a}^*, \mathbf{a}^*, \dots) \quad (92)$$

where  $\mathbf{a}^*(1) = a_1, \mathbf{a}^*(2) = a_1$ , and one corresponding optimal nature policy  $\tau^*$  is

$$\tau^* = (P^*, P^*, \dots), \quad (93)$$

where

$$P^* = \begin{pmatrix} 1 & 0 \\ 1 & 0 \\ 0 & 1 \\ 0 & 1 \end{pmatrix}. \quad (94)$$

**Example 2:** Consider an  $M/G/1/N$  queue arrival rate control problem shown in Fig. 1 [11]-[12], where the arrival process is Poisson, the storage capacity is  $N$ , the number of services is 1. The service distribution is a Coxian distribution [10] consisting of two stages, each with an

exponential distribution with mean  $s_i$  ( $i = 1, 2$ ). After receiving service at stage  $i$ , a customer enters stage  $i + 1$  with probability  $u_i$ , and leaves the station with probability  $1 - u_i$ . Let  $n$  be the number of customers in the queue. When an arriving customer reaches  $n = N$ , the customer is rejected.

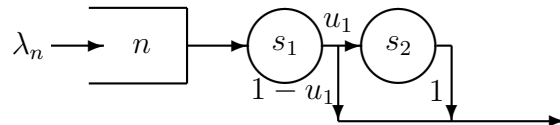


Fig.1 An M/G/1/N Queue Arrival Rate Control Problem

The arrival rate  $\lambda_n$  as an action may take values from  $\bar{\mathcal{A}} = \{0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9\}$ . The action space is  $\mathcal{A} \subseteq \bar{\mathcal{A}}$ . The state space is  $\chi = \{0, (n, s) : n = 1, \dots, N, s = 1, 2\}$ , with  $s$  denoting the stage of the customer being served. We rank the states and denote them by  $S = \{1, 2, \dots, 2^N + 1\}$ , where  $1 \in S$  denotes the state  $0 \in \chi$ ;  $1 \in S$  denotes the state  $(1, 1) \in \chi$ ,  $2 \in S$  denoted the state  $(1, 2) \in \chi$  and so on. The cost function is

$$c(i, \lambda_n) = \begin{cases} 9 \times (1/\lambda_n)^2 & 0 \\ n^2 + 9 \times (1/\lambda_n) & n < N \\ 5 \times n^2 & n = N \end{cases} \quad (95)$$

In the numerical calculation, suppose that  $N = 2$ ,  $s_1 = s_2 = 7/6$ ,  $u_2 = 0$ ,  $S = \{1, 2, 3, 4, 5\}$ ,  $\mathcal{A} = \{0.1, 0.3, 0.5, 0.7, 0.9\}$  for any state in  $S$ , the discount factor  $\gamma$  is 0.9, and the set estimation of  $u_1$  is  $\mathcal{U}_1 = [0.7, 0.9]$ .

The  $M/G/1/N$  queue with observable stages is essentially an MDP. This MDP can be discretized by uniformization [11]. The uniformization parameter is 2.1. The corresponding transition matrices of the discrete MDP are expressed as follows

$$P^{\lambda_n} = \begin{pmatrix} P_1^{\lambda_n} \\ P_2^{\lambda_n} \\ P_3^{\lambda_n} \\ P_4^{\lambda_n} \\ P_5^{\lambda_n} \end{pmatrix} = \begin{pmatrix} 1 - \frac{\lambda_n}{2.1} & \frac{\lambda_n}{2.1} & 0 & 0 & 0 \\ \frac{(1-u_1)s_1}{2.1} & 1 - \frac{\lambda_n+s_1}{2.1} & \frac{u_1s_1}{2.1} & \frac{\lambda_n}{2.1} & 0 \\ \frac{s_2}{2.1} & 0 & 1 - \frac{\lambda_n+s_2}{2.1} & 0 & \frac{\lambda_n}{2.1} \\ 0 & \frac{(1-u_1)(\lambda_n+s_1)}{2.1} & 0 & 1 - \frac{\lambda_n+s_1}{2.1} & \frac{u_1(\lambda_n+s_1)}{2.1} \\ 0 & \frac{\lambda_n+s_2}{2.1} & 0 & 0 & 1 - \frac{\lambda_n+s_2}{2.1} \end{pmatrix} \quad \forall \lambda_n \in \bar{\mathcal{A}} \quad (96)$$

The transition matrices are uncertain and correlated. As in case (vii) of Example 1, robust value iteration and robust policy iteration can not be used with a guarantee of a stationary optimal policy pair under the optimality criterion given by (24)-(29). Hence, robust policy iteration under total value function given by Algorithm 3 is used by the direct method for policy evaluation and the policy elimination for policy improvement. In the direct method, since  $\mathcal{U}_1$  is compact and  $\|v_{u_1}^\pi\|^2$  is differentiable, the maximum squared total value function for the policy  $\pi$  is obtained by computing  $u_1'$  such that

$$\left. \frac{\partial \|v_{u_1}^\pi\|^2}{\partial u_1} \right|_{u_1=u_1' \in \mathcal{U}_1} = 0 \quad (97)$$

and comparing the values of  $\|v_{u_1}^\pi\|^2$  at  $u_1', 0.7$  and  $0.9$ . The initial controller policy  $\pi_0$  is

$$\pi_0 = (\mathbf{a}_0, \mathbf{a}_0, \dots) \quad (98)$$

where  $\mathbf{a}_0(i) = 0.1$  for any state  $i \in S$ . A stationary optimal controller policy  $\pi^*$  is

$$\pi^* = (\mathbf{a}^*, \mathbf{a}^*, \dots), \quad (99)$$

where

$$\mathbf{a}^*(i) = 0.9 \quad \text{for any state } i \in S. \quad (100)$$

The corresponding stationary optimal nature policy  $\tau^*$  is

$$\tau^* = (P^*, P^*, \dots), \quad (101)$$

where  $P^*$  is the value at  $u_1^* = 0.7$ .

## VII. CONCLUSION

In this paper, discounted, finite-state, finite-action, infinite-horizon Markov decision processes are classified into two types: MDPs with independent transition matrices and MDPs with correlated transition matrices, where MDPs with exact transition matrices are special cases of MDPs with independent transition matrices. With this formulation of transition matrices, we have developed sufficient conditions for obtaining stationary deterministic optimal policies under the optimality criterion of minimizing the maximum value function for any initial state. Furthermore, both robust value iteration and robust policy iteration can be applied to obtain such an optimal policy. Our analysis concludes that all MDPs with independent transition matrices, and some of the MDPs with correlated transition matrices actually satisfy these sufficient conditions. To address general MDPs with correlated transition matrices, a new notion of squared total value function is introduced. We have proved that robust policy iteration under total value function can be used to obtain a stationary deterministic optimal policy. The results developed in this paper have demonstrated the feasibility of applying approximate dynamic programming methods to realistic problems where the system dynamics reflected in the transition matrices are uncertain. We have shown specific guarantees of optimal solutions to the robust dynamic programming problems.

## APPENDIX I

### ROBUST VALUE ITERATION AND ROBUST POLICY ITERATION

#### Robust Value Iteration

1. Select  $v_0 \in \Re^{n \times 1}$  and set  $k = 0$ ;
2. Compute  $v_{k+1}$  by

$$v_{k+1}(i) = \min_{a \in \mathcal{A}_i} (c(i, a) + \gamma \max_{P_i^a \in \mathcal{P}_i^a} P_i^a v_k); \quad (102)$$

3. If  $v_{k+1} = v_k$ , then go to **4**; otherwise increment  $k$  by 1 and go to **2**;
4. Compute an optimal stationary controller policy  $\pi^* = (\mathbf{a}^*, \mathbf{a}^*, \dots)$  and an optimal stationary nature policy  $\tau^* = (P^*, P^*, \dots)$  by

$$\mathbf{a}^*(i) \in \arg \min_{a \in \mathcal{A}_i} \{c(i, a) + \gamma \max_{P_i^a \in \mathcal{P}_i^a} P_i^a v_k\} \quad (103)$$

$$(P^*)_i^a \in \arg \max_{P_i^a \in \mathcal{P}_i^a} \{P_i^a v_k\} \quad (104)$$

### Robust Policy Iteration

1. Initialization: select  $\pi_0 = (\mathbf{a}_0, \mathbf{a}_0, \dots) \in \Pi$  and set  $k = 0$ ;
2. Policy evaluation: select any feasible transition probability row  $P_i^{\mathbf{a}_k(i)} \in \mathcal{P}_i^{\mathbf{a}_k(i)}$  and constitute a transition matrix of  $\pi_k, P^{\pi_k}$ ;

(a) compute  $v_k$  by

$$v_k = C^{\pi_k} + \gamma P^{\pi_k} v_k \quad (105)$$

where  $C^{\pi_k} = (c(1, \mathbf{a}_k(1)) \cdots (c(i, \mathbf{a}_k(i)) \cdots c(n, \mathbf{a}_k(n))))'$ ;

(b) find the transition probability row  $\tilde{P}_i^{\mathbf{a}_k(i)}$  for each initial state  $i$

$$\tilde{P}_i^{\mathbf{a}_k(i)} \in \arg \max_{P_i^{\mathbf{a}_k(i)} \in \mathcal{P}_i^{\mathbf{a}_k(i)}} \{P_i^{\mathbf{a}_k(i)} v_k\}; \quad (106)$$

(c) if  $c(i, \mathbf{a}_k(i)) + \gamma \tilde{P}_i^{\mathbf{a}_k(i)} v_k = v_k(i)$  for each  $i$ , go to **3**; otherwise, set  $P_i^{\mathbf{a}_k(i)} = \tilde{P}_i^{\mathbf{a}_k(i)}$  for each  $i$  and go to (a) of **2**;

3. Policy improvement: find  $\pi_{k+1} = (\mathbf{a}_{k+1}, \mathbf{a}_{k+1}, \dots)$

$$\mathbf{a}_{k+1}(i) \in \arg \min_{a \in \mathcal{A}_i} \{c(i, a) + \gamma \max_{P_i^a \in \mathcal{P}_i^a} P_i^a v_k\} \quad (107)$$

4. Termination: stop if  $\pi_{k+1} = \pi_k$ , go to **5**; otherwise, increment  $k$  by 1 and go to **2**;
5. Output: take  $\pi_k$  as a stationary optimal controller policy  $\pi^*$  and compute a stationary optimal nature policy  $\tau^* = (P^*, P^*, \dots)$  by

$$(P^*)_i^a \in \arg \max_{P_i^a \in \mathcal{P}_i^a} \{P_i^a v_k\} \quad \forall i \in S \quad a \in \mathcal{A}_i. \quad (108)$$

**Remark:** robust value iteration and robust policy iteration shows that the selected optimal nature policy is independent of an optimal controller policy.

## APPENDIX II

### PROOFS OF LEMMA 2 AND LEMMA 3

*Proof:* Because  $(P^\pi)^* \in \mathcal{P}^\pi$ ,  $v_\infty^\pi \in \Omega_3$ . Because  $\Omega_1 \supseteq \Omega_3$ , Algorithm 1 for problems (35) can be used for obtaining optimal solutions for problems (45). ■

*Proof:* Because  $(P)^* \in \mathcal{P}$ ,  $v_\infty \in \Omega_4$ . Because and  $\Omega_2 \supseteq \Omega_4$ , Algorithm 2 for (37) can be used for obtaining optimal solutions for (46). ■

## APPENDIX III

## PROOF OF THEOREM 1

*Proof:* In [3], the problem

$$\min_{\pi \in \Pi} \mathbf{E}_i^\pi \left( \sum_{t=0}^{\infty} \gamma^t c(j, a) \right) \quad \text{for any given exact transition matrix } P \text{ and any initial state } i \in S \quad (109)$$

is expressed with the linear program

$$\max_{v \in \mathbb{R}^{n \times 1}} qv : v(i) \leq c(i, a) + \gamma P_i^a v \quad i \in S, a \in \mathcal{A}_i. \quad (110)$$

The problem

$$\mathbf{E}_i^\pi \left( \sum_{t=0}^{\infty} \gamma^t c(j, a) \right) \quad \text{for any given } \pi \in \Pi, \text{ any given exact transition matrix } P, \text{ and any initial state } i \in S \quad (111)$$

is also expressed with the linear program

$$\max_{v \in \mathbb{R}^{n \times 1}} qv : v(i) \leq c(i, a) + \gamma P_i^{\mathbf{a}(i)} v \quad i \in S. \quad (112)$$

Hence, for any given  $\pi \in \Pi$ , the problem

$$\max_{\tau \in T} v_\tau^\pi(i) = \max_{P^\pi \in \mathcal{P}^\pi} v_{P^\pi}^\pi(i) \quad \text{for any initial state } i \in S \quad (113)$$

can be expressed as the optimal solution to the problem (45) by letting  $P^\pi$  vary in  $\mathcal{P}^\pi$ , and the problem

$$\max_{\tau \in T} \min_{\pi \in \Pi} v_\tau^\pi(i) = \max_{P \in \mathcal{P}} \min_{\pi \in \Pi} v_P^\pi(i) \quad \text{for any initial state } i \in S \quad (114)$$

can be expressed as the optimal solution to the problem (46) by letting  $P$  vary in  $\mathcal{P}$ .

Because for any  $\pi \in \Pi$ ,  $(P^\pi)^* \in \mathcal{P}^\pi$ , and  $P^* \in \mathcal{P}$ , by Lemma 2 and Lemma 3, the optimal solutions for the problems given in (45) and (46) are  $v_\infty^\pi$  and  $v_\infty$ , and  $v_\infty^\pi$  and  $v_\infty$  are computed by Algorithm 1 and Algorithm 2 for problems (35) and (37), respectively. For (113),

$$\max_{\tau \in T} v_\tau^\pi(i) = \max_{P^\pi \in \mathcal{P}^\pi} v_{P^\pi}^\pi(i) = v_\infty^\pi(i) \quad \text{for any initial state } i \in S \quad (115)$$

For (114),

$$\max_{\tau \in T} \min_{\pi \in \Pi} v_\tau^\pi(i) = \max_{P \in \mathcal{P}} \min_{\pi \in \Pi} v_P^\pi(i) = v_\infty(i) \quad \text{for any initial state } i \in S \quad (116)$$

Let  $\pi^* = (\mathbf{a}^*, \mathbf{a}^*, \dots)$ , where

$$\mathbf{a}^*(i) \in \arg \min_{a \in \mathcal{A}_i} \left\{ c(i, a) + \gamma \max_{P_i^a \in \mathcal{P}_i^a} P_i^a v_\infty \right\}, \quad (117)$$

or

$$\mathbf{a}^*(i) \in \arg \min_{a \in \mathcal{A}_i} \{ c(i, a) + \gamma (P_i^a)^* v_\infty \}. \quad (118)$$

Obviously,  $v_\infty = v_\infty^{\pi^*}$  and  $\max_{\tau \in \mathcal{T}} \min_{\pi \in \Pi} v_\tau^\pi(i) = \max_{\tau \in \mathcal{T}} v_\tau^{\pi^*}(i) = v_{P^*}^{\pi^*}(i)$ . By weak duality,

$$\min_{\pi \in \Pi} \max_{\tau \in \mathcal{T}} v_\tau^\pi(i) \geq \max_{\tau \in \mathcal{T}} \min_{\pi \in \Pi} v_\tau^\pi(i). \quad (119)$$

Because  $\max_{\tau \in \mathcal{T}} v_\tau^{\pi^*}(i) \geq \min_{\pi \in \Pi} \max_{\tau \in \mathcal{T}} v_\tau^\pi(i)$ ,

$$\min_{\pi \in \Pi} \max_{\tau \in \mathcal{T}} v_\tau^\pi(i) \geq \max_{\tau \in \mathcal{T}} \min_{\pi \in \Pi} v_\tau^\pi(i) = \max_{\tau \in \mathcal{T}} v_\tau^{\pi^*}(i) = v_{P^*}^{\pi^*}(i) \geq \min_{\pi \in \Pi} \max_{\tau \in \mathcal{T}} v_\tau^\pi(i). \quad (120)$$

That is,

$$v_{P^*}^{\pi^*}(i) = \max_{\tau \in \mathcal{T}} v_\tau^{\pi^*}(i) = \min_{\pi \in \Pi} \max_{\tau \in \mathcal{T}} v_\tau^\pi(i). \quad (121)$$

Hence,  $(\pi^*, P^*)$  is a stationary optimal policy pair and they are computed by robust value iteration. ■

## APPENDIX IV

### PROOF OF THEOREM 2

*Proof:* Because  $g^\pi$  is monotone non-decreasing, for any given  $v \in \mathcal{G}^\pi$ ,  $v^0 = v \geq v^{(1)} = g^\pi(v) \geq v^{(2)} = g^\pi(g^\pi(v)) \geq \dots$ . Because  $g^\pi$  is contractive, the non-increasing sequence  $\{v^{(k)}\}$  converges to  $v_\infty^\pi$  and  $v^{(k)} \geq v_\infty^\pi \geq v_\infty$ . Because in step 3 of robust policy iteration,  $v_k \geq g(v_k) = g^{\pi_{k+1}}(v_k)$ ,  $v_k \in \mathcal{G}^{\pi_{k+1}}$  and so  $v_k \geq g(v_k) \geq v_\infty^{\pi_{k+1}} = v_{k+1} \geq v_\infty$ . That is to say, step 3 improves the controller policy by forming the sequence

$$v_1 \geq g(v_1) \geq v_2 \geq g(v_2) \geq \dots \geq v_\infty. \quad (122)$$

The sequence  $\{v_k\}$  is included in  $\mathcal{G}$  and converges to  $v_\infty$  in finite iterations because of the existence of an optimal controller policy in the finite space  $\Pi$ . When  $\pi_{k+1} = \pi_k$ ,  $g(v_k) = v_k$  and so  $v_k = v_\infty$ .  $\pi_k$  is a stationary optimal controller policy and the corresponding stationary nature policy  $\tau^*$  is computed in step 5. ■

## APPENDIX V

## PROOF OF THEOREM 3

*Proof:* For any  $\pi = (\mathbf{a}, \mathbf{a}, \dots) \in \Pi$  and any  $\mathbf{U} \in \mathcal{U}$ , it is proven in [3] that

$$v_{\mathbf{U}}^{\pi} = C^{\pi} + \gamma P_{\mathbf{U}}^{\pi} v_{\mathbf{U}}^{\pi}, \quad (123)$$

where

$$P_{\mathbf{U}}^{\pi} = \begin{pmatrix} P_1^{\mathbf{a}(1)} \\ \vdots \\ P_i^{\mathbf{a}(i)} \\ \vdots \\ P_n^{\mathbf{a}(n)} \end{pmatrix} = \begin{pmatrix} \bar{\mathbf{f}}_1^{\mathbf{a}(1)}(\mathbf{U}) \\ \vdots \\ \bar{\mathbf{f}}_i^{\mathbf{a}(i)}(\mathbf{U}) \\ \vdots \\ \bar{\mathbf{f}}_n^{\mathbf{a}(n)}(\mathbf{U}) \end{pmatrix} \quad C^{\pi} = \begin{pmatrix} c(1, \mathbf{a}(1)) \\ \vdots \\ c(i, \mathbf{a}(i)) \\ \vdots \\ c(n, \mathbf{a}(n)) \end{pmatrix} \quad (124)$$

Because  $(I - \gamma P_{\mathbf{U}}^{\pi})$  is invertible,

$$v_{\mathbf{U}}^{\pi} = (I - \gamma P_{\mathbf{U}}^{\pi})^{-1} C^{\pi} \quad (125)$$

$$\|v_{\mathbf{U}}^{\pi}\|^2 = (v_{\mathbf{U}}^{\pi})' v_{\mathbf{U}}^{\pi} \quad (126)$$

$$= \left( (I - \gamma P_{\mathbf{U}}^{\pi})^{-1} C^{\pi} \right)' (I - \gamma P_{\mathbf{U}}^{\pi})^{-1} C^{\pi} \quad (127)$$

$$= (C^{\pi})' \left( I - \gamma (P_{\mathbf{U}}^{\pi})' \right)^{-1} (I - \gamma P_{\mathbf{U}}^{\pi})^{-1} C^{\pi} \quad (128)$$

Because there exists  $\mathbf{U}^{\pi} \in \mathcal{U}$  such that

$$\|v_{\mathbf{U}^{\pi}}^{\pi}\|^2 = \max_{\mathbf{U} \in \mathcal{U}} \|v_{\mathbf{U}}^{\pi}\|^2 = \max_{\mathbf{U} \in \mathcal{U}} (C^{\pi})' \left( I - \gamma (P_{\mathbf{U}}^{\pi})' \right)^{-1} (I - \gamma P_{\mathbf{U}}^{\pi})^{-1} C^{\pi},$$

a stationary optimal controller policy is

$$\pi^* \in \arg \min_{\pi} \{ \|v_{\mathbf{U}^{\pi}}^{\pi}\|^2 : \pi \in \Pi \}. \quad (129)$$

The corresponding stationary optimal nature policy  $\tau^*$  is

$$\tau^* = (P^*, P^*, \dots) \quad (130)$$

where  $P^*$  is the value at  $\mathbf{U}^*$ . ■

## APPENDIX VI

## PROOF OF THEOREM 4

*Proof:* Because a stationary optimal policy pair exists under the optimality criterion defined in (24)-(29) and it is denoted as  $(\pi_{pre}^*, P_{pre}^*)$ , for any  $\pi \in \Pi$ , there exists  $P_\pi^* \in \mathcal{P}$  such that

$$\bar{v}_{P_\pi^*}^\pi(i) = \max_{P \in \mathcal{P}} v_P^\pi(i) \quad \text{for any initial state } i \in S \quad (131)$$

and

$$v_{P_{pre}^*}^{\pi_{pre}^*}(i) = \max_{P \in \mathcal{P}} v_P^\pi(i) = \min_{\pi \in \Pi} \max_{P \in \mathcal{P}} v_P^\pi(i) \quad \text{for any initial state } i \in S. \quad (132)$$

Because  $v_P^\pi(i) \geq 0$ ,

$$\max_{P \in \mathcal{P}} \sum_{i \in S} (v_P^\pi(i))^2 = \sum_{i \in S} \max_{P \in \mathcal{P}} (v_P^\pi(i))^2 = \|v_{P_\pi^*}^\pi\|^2, \quad (133)$$

$$\min_{\pi \in \Pi} \max_{P \in \mathcal{P}} \sum_{i \in S} (v_P^\pi(i))^2 = \sum_{i \in S} \min_{\pi \in \Pi} \max_{P \in \mathcal{P}} (v_P^\pi(i))^2 = \|v_{P_{pre}^*}^{\pi_{pre}^*}\|^2. \quad (134)$$

“ $\Rightarrow$ ” : based on (133) and (134),

$$\|v_{P_{pre}^*}^{\pi_{pre}^*}\|^2 = \max_{P \in \mathcal{P}} \sum_{i \in S} (v_P^{\pi_{pre}^*}(i))^2 = \min_{\pi \in \Pi} \max_{P \in \mathcal{P}} \sum_{i \in S} (v_P^\pi(i))^2 \quad (135)$$

Hence,  $(\pi_{pre}^*, P_{pre}^*)$  is a stationary optimal policy pair under the optimality criterion defined in (51)-(52).

“ $\Leftarrow$ ” : based on (133) and (134),

$$\|v_{P_{new}^*}^{\pi_{new}^*}\|^2 = \sum_{i \in S} (v_{P_{new}^*}^{\pi_{new}^*}(i))^2 = \max_{P \in \mathcal{P}} \sum_{i \in S} (v_P^{\pi_{new}^*}(i))^2 = \sum_{i \in S} \max_{P \in \mathcal{P}} (v_P^{\pi_{new}^*}(i))^2, \quad (136)$$

$$\|v_{P_{new}^*}^{\pi_{new}^*}\|^2 = \sum_{i \in S} (v_{P_{new}^*}^{\pi_{new}^*}(i))^2 = \min_{\pi \in \Pi} \max_{P \in \mathcal{P}} \sum_{i \in S} (v_P^\pi(i))^2 = \sum_{i \in S} \min_{\pi \in \Pi} \max_{P \in \mathcal{P}} (v_P^\pi(i))^2. \quad (137)$$

Because  $(v_{P_{new}^*}^{\pi_{new}^*}(i))^2 \leq \max_{P \in \mathcal{P}} (v_P^{\pi_{new}^*}(i))^2$ ,

$$(v_{P_{new}^*}^{\pi_{new}^*}(i))^2 = \max_{P \in \mathcal{P}} (v_P^{\pi_{new}^*}(i))^2. \quad (138)$$

Because  $(v_{P_{new}^*}^{\pi_{new}^*}(i))^2 \geq \min_{\pi \in \Pi} \max_{P \in \mathcal{P}} (v_P^\pi(i))^2$ ,

$$(v_{P_{new}^*}^{\pi_{new}^*}(i))^2 = \min_{\pi \in \Pi} \max_{P \in \mathcal{P}} (v_P^\pi(i))^2. \quad (139)$$

Hence,

$$v_{P_{new}^*}^{\pi_{new}^*}(i) = \max_{P \in \mathcal{P}} v_P^{\pi_{new}^*}(i) = \min_{\pi \in \Pi} \max_{P \in \mathcal{P}} v_P^{\pi_{new}^*}(i) \quad \text{for any initial state } i \in S. \quad (140)$$

That is to say,  $(\pi_{new}^*, P_{new}^*)$  is a stationary optimal policy pair under the optimality criterion defined in (24)-(29). ■

## REFERENCES

- [1] P. Bellman and R. Kalaba, *Dynamic Programming and Modern Control Theory*, New York: Academic Press, 1965.
- [2] J. Si, A. G. Barto, W. B. Powell and D. Wunsch, *Handbook of Learning and Approximate Dynamic Programming*, Wiley-IEEE Press, 2004.
- [3] M. Putterman, *Markov Decision Processes: Discrete Stochastic Dynamic Programming*, Wiley-Interscience, New York, 1994.
- [4] D. Berstsekas and J. Tsitsiklis, *Neuro-Dynamic Programming*, Athena Scientific Massachusetts, 1996.
- [5] A. Nilim and L. E. Ghaoui, "Robust control of Markov decision processes with uncertain transition matrices," *Operations Research*, vol. 53, n. 5, pp. 780-798, 2005.
- [6] J. K. Satia and R. E. Lave, Jr., "Markov decision processes with uncertain transition probabilities," *Operations Research*, vol. 21, pp. 728-740, 1973.
- [7] C. C. White and H. K. Eldeib, "Markov decision processes with imprecise transition probabilities," *Operations Research*, vol. 43, pp. 739-749, 1994.
- [8] R. Givan, S. Leach, and T. Dean, "Bounded parameter Markov decision processes," *Artificial Intelligence*, vol. 122, no. 1-2, pp. 71-109, 2000.
- [9] S. Kalyanasundaram, E. K. P. Chong, and N. B. Shroff, "Markov decision processes with uncertain transition rates: sensitivity and robust control," *Proceedings of the 41st IEEE Conference on Decision and Control*, vol. 4, pp. 3799-3804, 2002.
- [10] D. R. Cox, "A use of complex probabilities in the theory of stochastic processes," *Proc. Cambridge Philosophical Society*, vol. 51, pp. 313-319, 1955.
- [11] C. G. Cassandras and S. Lafortune, *Introduction to discrete event systems*, Boston : Kluwer Academic, 1999.
- [12] X. R. Cao and H. T. Tang, "Gradient-based policy iteration: An example," *Proceedings of the 41st IEEE Conference on Decision and Control*, pp. 3367-3371, 2002.