Proceedings of the 2007 IEEE Symposium on Approximate
Dynamic Programming and Reinforcement Learning (ADPRL 2007)

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

Baohua Li
Department of Electrical Engineering
Arizona State University
Tempe, AZ 85287-5706
Email: Baohua.Li@asu.edu

Jennie Si
Department of Electrical Engineering
Arizona State University
Tempe, AZ 85287-5706
Telephone: (480) 965-6133
Fax: (480) 965-2811
Email: Si@asu.edu

Abstract— **In this paper, finite-state, finite-action, discounted infinite-horizon-cost Markov decision processes (MDPs) with uncertain stationary transition matrices are discussed in the deterministic policy space. Uncertain stationary parametric transition matrices are clearly classified into independent and correlated cases. It is pointed out in this paper that the optimality criterion of uniform minimization of the maximum expected total discounted cost functions for all initial states, or robust uniform optimality criterion, is not appropriate for solving MDPs with correlated transition matrices. A new optimality criterion of minimizing the maximum quadratic total value function is proposed which includes the previous criterion as a special case. Based on the new optimality criterion, robust policy iteration is developed to compute an optimal policy in the deterministic stationary policy space. Under some assumptions, the solution is guaranteed to be optimal or near-optimal in the deterministic policy space.**

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]. DP and approximate DP can solve the problems of Markov decision processes (MDPs) with accurate knowledge of transition probabilities to obtain an optimal or near-optimal policy by the algorithms such as value iteration, policy iteration, evolutionary policy iteration, approximate policy iteration based on Monte Carlo simulation [2]–[5]. However, in practice, accurate transition probabilities are very difficult to obtain. Thus, exact solutions via classic DP are not attainable, and those algorithms are not applicable to obtain solutions any more. Moreover, the estimated transition probabilities may be far away from the true values due to noise and other issues associated with the estimation process, or the estimation error may not be trivial that it can result in significant deviation of the solutions from the optimal values [6]. Hence, the idea of set estimation for transition matrices with high confidence can be used to alleviate some of the deficit from inaccurate point estimation. With those uncertain transition matrices, robust dynamic programming is desired to address the issue of designing approximation method with an appropriate robustness to extend the power of the Bellman Equation. Representative efforts in developing robust dynamic programming can be found in [6]–[10]. One commonly used principle of optimality criterion for robust algorithms is to minimize the maximum value functions for all initial states, which is referred to as robust uniform optimality criterion in this paper. Based on this optimality criterion, robust value iteration and robust policy iteration were proposed in [6] and [7], respectively, to obtain a deterministic, uniformly optimal policy. To deal with uncertain transition matrices, the notion of correlation in a transition matrix was introduced in [6] and [10] in the context of MDPs. The existing optimality criteria and associated robust algorithms were developed only for MDPs with independent transition matrices.

In this paper, finite-state, finite-action MDPs with discounted infinite-horizon cost is discussed. The optimal policy is constrained in the deterministic policy space. The transition matrices are assumed to be stationary, which is reasonable when systems are slowly varying. The contribution of this paper is as follows. First, mathematically clear and more tractable definitions of independent and correlated uncertain stationary transition matrices are provided. Using these newly formulated definitions, robust uniform optimality criterion is not applicable for MDPs with correlated transition matrices. Because of the correlation of uncertain transition matrices, for any given policy, there may be no optimal transition matrices such that the value functions of the policy reach maximum uniformly for all initial states. This makes the comparison among policies meaningless. Even if such transition matrices exist for all policies, there may not be an optimal policy such that its maximum value functions reach minimum uniformly for all initial states. Hence, a new optimality criterion of minimizing quadratic total value function is proposed. Based on this criterion, under a weak condition, there exists an optimal policy which can be optimal non-uniformly in initial states. The previous criterion becomes a special case of the new criterion. Note that the existing robust value iteration and robust policy iteration can not guarantee solutions for

MDPs with correlated transition matrices. A new robust policy iteration is developed to obtain an optimal policy in the stationary space. Under some assumptions, the solution is guaranteed to be optimal or near-optimal in the deterministic policy space.

The rest of the paper is organized as follows. In section 2, the optimality criterion of minimizing the maximum quadratic total value function is proposed and theorems are developed to show that this criterion makes robust uniform optimality criterion a special case and under some conditions, stationary optimal policies are proven to be optimal or near-optimal in the deterministic policy space. Base on this optimality criterion, robust policy iteration is developed in section 3. The paper concludes in section 4.

II. Optimality criterion Using Quadratic total value Functions

In this section, an MDP with uncertain parametric transition matrix is formulated. Based on this formulation, we present why an optimal solution may be sensitive to transition probabilities. After that, uncertain parametric transition matrices are classified into independent and correlated cases. Analysis demonstrates why an optimal policy may not exist under robust uniform optimality criterion. Consequently, the optimality criterion using the quadratic total value function is proposed. It is proven that robust uniform optimality criterion is a special case. Besides, two theorems and one corollary are provided to show that under some assumptions, stationary optimal policies are optimal or near-optimal in the deterministic policy space.

Finite-state, finite-action, infinite-horizon MDPs with stationary transition matrices are described as follows. Let T denote the discrete, infinite decision horizon, where $T = \{0, 1, 2, \cdots\}$. At each stage, the system occupies a state $i \in S$, where S is the state space with n states and denoted as $S = \{1, 2, \cdots, n\}$. With each state $i \in S$, a decision maker is allowed to choose an action a deterministically from a finite state-dependent set of allowable actions, denoted by $\mathcal{A}_i = \{1, 2, \cdots, m_i\}$. Let $\mathbf{M} = \sum_{i=1}^{n} m_i$, and let "$\mathbf{a}_t$" be a function mapping states into actions with $\mathbf{a}_t(i) \in \mathcal{A}_i$ at the time t, i.e.,

$$\mathbf{a}_t : \begin{pmatrix} 1 \\ 2 \\ \vdots \\ n \end{pmatrix} \rightarrow \begin{pmatrix} a_1 \\ a_2 \\ \vdots \\ a_n \end{pmatrix} \in \mathcal{A}_1 \times \mathcal{A}_2 \times \cdots \times \mathcal{A}_n. \quad (1)$$

Denote a policy by π

$$\pi = (\mathbf{a}_1, \mathbf{a}_2, \cdots), \quad (2)$$

and let Π represent the deterministic policy space. Denote a stationary policy by π_s

$$\pi_s = (\mathbf{a}, \mathbf{a}, \cdots), \quad (3)$$

and let Π_s represent the deterministic stationary policy space. Obviously, $\Pi_s \subset \Pi$. 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 nonnegative. The costs are time discounted by a factor γ $(0 < \gamma < 1)$. 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. All transition probabilities constitute an $\mathbf{M} \times n$ transition matrix P

$$P = \begin{pmatrix} \left(P_1^1\right)' & \cdots & \left(P_i^a\right)' & \cdots & \left(P_n^{m_n}\right)' \end{pmatrix}', \quad (4)$$

where the superscript "$'$" denotes the transpose and P_i^a represents a transition probability row under the state $i \in S$ and the action $a \in \mathcal{A}_i$.

In a more general setting, let each transition probability p_{ij}^a be represented by a function of a parameter vector denoted as $\theta = \{\vartheta_1, \cdots, \vartheta_q\}$, i.e.,

$$p_{ij}^a \triangleq p_{ij}^a(\theta) \quad \forall i, j \in S, \quad (5)$$

where $0 \leq p_{ij}^a \leq 1$ and $\sum_{j=1}^{n} p_{ij}^a = 1$. In [6]–[10], the parameters are specified as the transition probabilities themselves, that is,

$$\theta = \{p_{11}^1, \cdots, p_{1j}^1, \cdots, p_{1(n-1)}^1, \cdots, p_{i1}^a, \cdots, p_{ij}^a, \cdots, p_{i(n-1)}^a, \cdots, p_{n1}^{m_n}, \cdots, p_{nj}^{m_n}, \cdots, p_{n(n-1)}^{m_n}\}. \quad (6)$$

The transition probability row P_i^a and the transition matrix P can also be represented as a function of θ, i.e., $P_i^a \triangleq P_i^a(\theta)$ and $P \triangleq P(\theta)$. Assume the parameter vector θ vary in a known subset $\Theta \subset \Re^{1\times q}$. And the transition matrix P is varying in the set $\mathcal{P} \triangleq \{P : \theta \in \Theta\}$.

It is not too difficult to see that optimal value functions may be very sensitive to the perturbation of the parameters in transition matrix. Consider a stationary policy π_s defined in (3). For any $P \in \mathcal{P}$,

$$v_P^{\pi_s} = \left(I - \gamma P^{\pi_s}\right)^{-1} C^{\pi_s}, \quad (7)$$

where $v_P^{\pi_s}$ is the expected total discounted cost function under the policy π_s and the trasition matrix P, P^{π_s} and C^{π_s} are the $n \times n$ transition matrix and the $n \times 1$ cost function vector respectively to π_s, i.e.,

$$P^{\pi_s} = \begin{pmatrix} P_1^{\mathbf{a}(1)} \\ \vdots \\ P_i^{\mathbf{a}(i)} \\ \vdots \\ P_n^{\mathbf{a}(n)} \end{pmatrix}, \quad C^{\pi_s} = \begin{pmatrix} c(1, \mathbf{a}(1)) \\ \vdots \\ c(i, \mathbf{a}(i)) \\ \vdots \\ c(n, \mathbf{a}(n)) \end{pmatrix}. \quad (8)$$

Obviously, $v_P^{\pi_s}$ is continuous in P [11]. However, P can be discontinuous in θ over Θ, which may lead to discontinuous value functions $v_P^{\pi_s}$ in θ. Even if P is continuous in θ and therefore $v_P^{\pi_s}$ is continuous in Θ, the function $v_P^{\pi_s}$ still may not be smooth enough, that is, a small perturbation in θ results in relatively large variation in $v_P^{\pi_s}$. Therefore there is a need to develop robust solutions under uncertainty.

We are now in a position to introduce the concepts of independence and correlation in P and P^{π_s}.

Definition (Correlated transition matrix, Independent

transition matrix):
The transition matrix P is correlated if

$$\mathcal{P} \subset \mathcal{P}_1^1 \times \cdots \times \mathcal{P}_i^a \times \cdots \times \mathcal{P}_n^{m_n}, \tag{9}$$

where $\mathcal{P}_i^a = \{P_i^a(\theta) : \theta \in \Theta\}$.
The transition matrix P is independent if

$$\mathcal{P} = \mathcal{P}_1^1 \times \cdots \times \mathcal{P}_i^a \times \cdots \times \mathcal{P}_n^{m_n}. \tag{10}$$

Definition (Correlated transition matrix for π_s, Independent transition matrix for π_s):
The transition matrix P^{π_s} is correlated if

$$\mathcal{P}^{\pi_s} \subset \mathcal{P}_1^{\mathbf{a}(1)} \times \cdots \times \mathcal{P}_i^{\mathbf{a}(i)} \times \cdots \times \mathcal{P}_n^{\mathbf{a}(n)}, \tag{11}$$

where $\mathcal{P}^{\pi_s} = \{P^{\pi_s}(\theta) : \theta \in \Theta\}$.
The transition matrix P^{π_s} is independent if

$$\mathcal{P}^{\pi_s} = \mathcal{P}_1^{\mathbf{a}(1)} \times \cdots \times \mathcal{P}_i^{\mathbf{a}(i)} \times \cdots \times \mathcal{P}_n^{\mathbf{a}(n)}. \tag{12}$$

Remarks:
(i) The set $\mathcal{P}_1^1 \times \cdots \times \mathcal{P}_i^a \times \cdots \times \mathcal{P}_n^{m_n}$ is an M-dimensional hyper-rectangle. Matrix P is correlated if $\mathcal{P}$ is a proper subset of this hyper-rectangle, and P is independent if $\mathcal{P}$ is equal to the hyper-rectangle. An exact transition matrix P is a special case of an independent transition matrix.
(ii) The set $\mathcal{P}_1^{\mathbf{a}(1)} \times \cdots \times \mathcal{P}_i^{\mathbf{a}(i)} \times \cdots \times \mathcal{P}_n^{\mathbf{a}(n)}$ is an n-dimensional hyper-rectangle. Matrix P^{π_s} is correlated if $\mathcal{P}^{\pi_s}$ is a proper subset of this hyper-rectangle, and P^{π_s} is independent if $\mathcal{P}^{\pi_s}$ is equal to the hyper-rectangle.
(iii) According to the above definitions, MDPs are classified into those with independent transition matrices and those with correlated transition matrices.

The optimality criterion of minimizing the maximum value function for any initial state is given as follows

$$\min_{\pi \in \Pi} \max_{P \in \mathcal{P}} v_P^\pi(i) \quad \forall i \in S, \tag{13}$$

where $v_P^\pi(i)$ is the expected total discounted cost function under the policy $\pi \in \Pi$ and the transition matrix $P \in \mathcal{P}$ at the initial state i, i.e.,

$$v_P^\pi(i) = \mathbf{E}_i^\pi\Big(\sum_{t=0}^{\infty} \gamma^t c(j,a)|P\Big). \tag{14}$$

According to the optimality criterion given in (13), an optimal policy is defined as follows.
Definition (Optimal policy): A policy π^* is optimal if there exists P^* such that

$$v_{P^*}^{\pi^*}(i) = \max_{P \in \mathcal{P}} v_P^{\pi^*}(i) = \min_{\pi \in \Pi} \max_{P \in \mathcal{P}} v_P^\pi(i) \quad \forall i \in S, \tag{15}$$

where there exists $\theta^* \in \Theta$ such that $P^* = P(\theta^*)$.

For MDPs with independent transition matrices, a deterministic optimal policy has been proven existent [6], [7]. However, for MDPs with correlated transition matrices, an optimal policy does not always exist. The definition of an optimal policy given by (15) can be interpreted as the conditions for the existence of an optimal policy: (i) for any given $\pi \in \Pi$, there is, at least one P^* such that $v_{P^*}^\pi(i) = \max_{P \in \mathcal{P}} v_P^\pi(i)$ for any $i \in S$, where from here on P^* depends on π; (ii) there is at least one $\pi^* \in \Pi$ such that $v_{P^*}^{\pi^*}(i) = \min_{\pi \in \Pi} v_{P^*}^\pi(i)$ for any $i \in S$. Such an optimal policy is uniformly optimal for initial states. For MDPs with correlated transition matrices, the above conditions are too strong to be satisfied, which can be shown in the following example.
Example: consider a two-state, two-action, infinite-horizon MDP. Let the state space be $S = \{1, 2\}$. The action spaces at state 1 and state 2 are $\mathcal{A}_1 = \{1, 2\}$ and $\mathcal{A}_2 = \{1, 2\}$. According to (1) and (3), all four stationary policies in Π_s are defined by $\pi_1, \pi_2, \pi_3, \pi_4$ as follows,

$$\pi_1 = (\begin{pmatrix} 1 \\ 1 \end{pmatrix}, \begin{pmatrix} 1 \\ 1 \end{pmatrix}, \cdots), \tag{16}$$

$$\pi_2 = (\begin{pmatrix} 1 \\ 2 \end{pmatrix}, \begin{pmatrix} 1 \\ 2 \end{pmatrix}, \cdots), \tag{17}$$

$$\pi_3 = (\begin{pmatrix} 2 \\ 1 \end{pmatrix}, \begin{pmatrix} 2 \\ 1 \end{pmatrix}, \cdots), \tag{18}$$

$$\pi_4 = (\begin{pmatrix} 2 \\ 2 \end{pmatrix}, \begin{pmatrix} 2 \\ 2 \end{pmatrix}, \cdots). \tag{19}$$

Let the transition matrix P have the following formulation

$$P = \begin{pmatrix} P_1^1 \\ P_1^2 \\ P_2^1 \\ P_2^2 \end{pmatrix} = \begin{pmatrix} \theta_1 & 1-\theta_1 \\ \theta_2 & 1-\theta_2 \\ 1-\theta_1^2 & \theta_1^2 \\ 1-\theta_3 & \theta_3 \end{pmatrix}, \tag{20}$$

and the cost functions are as follows,

$$c(1,1) = 1, c(1,2) = 2, c(2,1) = 3, c(2,2) = 4. \tag{21}$$

The discount factor γ is chosen at 0.9. Let $\Omega = \{0, 0.2, 0.4, 0.6, 0.8, 1\}$. Let $\theta = \{\theta_1, \theta_2, \theta_3\}$ and $\Theta = \{\theta : \theta_1, \theta_2, \theta_3 \in \Omega\}$. The transition matrix P is correlated, that is,

$$\mathcal{P} \subset \mathcal{P}_1^1 \times \mathcal{P}_1^2 \times \mathcal{P}_2^1 \times \mathcal{P}_2^2, \tag{22}$$

where

$$\mathcal{P}_1^1 = \mathcal{P}_1^2 = \mathcal{P}_2^2 = \{(0,1), (0.2, 0.8), (0.4, 0.6), (0.6, 0.4), (0.8, 0.2), (1, 0)\}, \tag{23}$$

$$\mathcal{P}_2^1 = \{(1,0), (0.96, 0.04), (0.84, 0.16), (0.64, 0.36), (0.36, 0.64), (0, 1)\}. \tag{24}$$

Actually, by (8),

$$P^{\pi_2} = \begin{pmatrix} \theta_1 & 1-\theta_1 \\ 1-\theta_3 & \theta_3 \end{pmatrix}, \tag{25}$$

$$P^{\pi_3} = \begin{pmatrix} \theta_2 & 1-\theta_2 \\ 1-\theta_1^2 & \theta_1^2 \end{pmatrix}, \tag{26}$$

$$P^{\pi_4} = \begin{pmatrix} \theta_2 & 1-\theta_2 \\ 1-\theta_3 & \theta_3 \end{pmatrix}. \tag{27}$$

By (12), P^{π_2}, P^{π_3} and P^{π_4} are independent and the maximum value functions are reachable. However, for π_1,

$$P^{\pi_1} = \begin{pmatrix} \theta_1 & 1-\theta_1 \\ 1-\theta_1^2 & \theta_1^2 \end{pmatrix}. \tag{28}$$

By (11), it is correlated and then there is no $P \in \mathcal{P}$ such that the maximum value functions for all initial states are reachable. Thus, by the optimality criterion defined in (13), an optimal policy does not exist.

Thus, a new optimality criterion is needed and its corresponding optimal policy also need to be defined. For any fixed transition matrix P, the quadratic total value function for a policy π is defined as follows,

$$\|v_P^\pi\|^2 = (v_P^\pi)' v_P^\pi, \tag{29}$$

where in terms of the value function defined in (14)

$$v_P^\pi = \begin{pmatrix} v_P^\pi(1) & \cdots & v_P^\pi(i) & \cdots & v_P^\pi(n) \end{pmatrix}'. \tag{30}$$

The new optimality criterion is to minimize the maximum quadratic total value function, i.e,

$$\min_{\pi\in\Pi} \max_{P\in\mathcal{P}} \|v_P^\pi\|^2. \tag{31}$$

Definition (Optimal policy): A policy π^* is optimal if

$$\|v_{P^*}^{\pi^*}\|^2 = \max_{P\in\mathcal{P}} \|v_P^{\pi^*}\|^2 = \min_{\pi\in\Pi} \|v_{P^*}^\pi\|^2 = \min_{\pi\in\Pi} \max_{P\in\mathcal{P}} \|v_P^\pi\|^2. \tag{32}$$

Remarks:

(i) Based on the optimality criterion using the quadratic total value function, an optimal policy can be non-uniformly optimal in initial states.

(ii) The existence of such an optimal policy is guaranteed under a weak condition. For any π, P^* is easily satisfied to make the maximum value reachable. For example, the set $\mathcal{P}$ is compact and closed and the function $\|v_P^\pi\|^2$ is continuous over $\mathcal{P}$. Even if the maximum and minimum values are not reachable, "max" and "min" can be replaced by "inf" and "sup", respectively, and an near-optimal policy can be easily proposed.

(iii) If the cost function is also parameterized in θ, the optimality criterion can be modified using the quadratic value function with the cost function represented as $c(i, a, \theta)$.

(iv) If for any policy π, the probability distribution of the initial states is non-uniform, or the value functions for different initial states have different weights, the optimality criterion can be modified using weighted quadratic total value function $\|v_P^\pi\|_W^2 = (v_P^\pi)' W v_P^\pi$.

Actually, the optimality criterion defined by (31) generalizes the uniform optimality criterion defined by (13), which is shown in the following Theorem 1.

Theorem 1: If an optimal policy exists with regard to the optimality criterion defined in (13), then this optimality criterion is equivalent to the one defined in (31). That is, if an optimal policy exists, a policy under the optimality criterion defined in (13) is optimal if and only if it is optimal under the optimality criterion defined in (31).

Proof: Since an optimal policy exists under the optimality criterion defined in (13), for any $\pi \in \Pi$, there exists $P^* \in \mathcal{P}$ such that

$$v_{P^*}^\pi(i) = \max_{P\in\mathcal{P}} v_P^\pi(i) \quad \forall i \in S \tag{33}$$

and let π_1 be an optimal policy such that

$$v_{P^*}^{\pi_1}(i) = \max_{P\in\mathcal{P}} v_P^{\pi_1}(i) = \min_{\pi\in\Pi} \max_{P\in\mathcal{P}} v_P^\pi(i) \quad \forall i \in S. \tag{34}$$

Since $v_P^\pi(i) \geq 0$, by (33),

$$\max_{P\in\mathcal{P}} \sum_{i\in S} \left(v_P^\pi(i)\right)^2 = \sum_{i\in S} \max_{P\in\mathcal{P}} \left(v_P^\pi(i)\right)^2 = \|v_{P^*}^\pi\|^2, \tag{35}$$

and then by (34),

$$\min_{\pi\in\Pi} \max_{P\in\mathcal{P}} \sum_{i\in S} \left(v_P^\pi(i)\right)^2 = \sum_{i\in S} \min_{\pi\in\Pi} \max_{P\in\mathcal{P}} \left(v_P^\pi(i)\right)^2 = \|v_{P^*}^{\pi_1}\|^2. \tag{36}$$

" $\Rightarrow$ " : By (35) and (36),

$$\|v_{P^*}^{\pi_1}\|^2 = \max_{P\in\mathcal{P}} \sum_{i\in S} \left(v_P^{\pi_1}(i)\right)^2 = \min_{\pi\in\Pi} \max_{P\in\mathcal{P}} \sum_{i\in S} \left(v_P^\pi(i)\right)^2. \tag{37}$$

Hence, π_1 is an optimal policy under the optimality criterion defined in (31).

" $\Leftarrow$ " : Let π_2 be an optimal policy under the optimality criterion defined in (31). Then, for any $\pi \in \Pi$, there exists $Q^* \in \mathcal{P}$ such that

$$\|v_{Q^*}^\pi\|^2 = \sum_{i\in S} \left(v_{Q^*}^\pi(i)\right)^2 = \max_{P\in\mathcal{P}} \sum_{i\in S} \left(v_P^\pi(i)\right)^2, \tag{38}$$

and for π_2,

$$\|v_{Q^*}^{\pi_2}\|^2 = \sum_{i\in S} \left(v_{Q^*}^{\pi_2}(i)\right)^2 = \min_{\pi\in\Pi} \max_{P\in\mathcal{P}} \sum_{i\in S} \left(v_P^\pi(i)\right)^2. \tag{39}$$

By (35), for any $\pi \in \Pi$,

$$\|v_{Q^*}^\pi\|^2 = \sum_{i\in S} \max_{P\in\mathcal{P}} \left(v_P^\pi(i)\right)^2, \tag{40}$$

and then by (36),

$$\|v_{Q^*}^{\pi_2}\|^2 = \sum_{i\in S} \min_{\pi\in\Pi} \max_{P\in\mathcal{P}} \left(v_P^\pi(i)\right)^2. \tag{41}$$

Since $\left(v_{Q^*}^{\pi_2}(i)\right)^2 \leq \max_{P\in\mathcal{P}} \left(v_P^{\pi_2}(i)\right)^2$,

$$\left(v_{Q^*}^{\pi_2}(i)\right)^2 = \max_{P\in\mathcal{P}} \left(v_P^{\pi_2}(i)\right)^2, \tag{42}$$

and then

$$v_{Q^*}^{\pi_2}(i) = \max_{P\in\mathcal{P}} v_P^{\pi_2}(i). \tag{43}$$

Since $\max_{P\in\mathcal{P}} \left(v_P^{\pi_2}(i)\right)^2 \geq \min_{\pi\in\Pi} \max_{P\in\mathcal{P}} \left(v_P^\pi(i)\right)^2$,

$$\left(v_{Q^*}^{\pi_2}(i)\right)^2 = \min_{\pi\in\Pi} \max_{P\in\mathcal{P}} \left(v_P^\pi(i)\right)^2, \tag{44}$$

and then

$$v_{Q^*}^{\pi_2}(i) = \min_{\pi\in\Pi} \max_{P\in\mathcal{P}} v_P^\pi(i). \tag{45}$$

Hence,

$$v_{Q^*}^{\pi_2}(i) = \max_{P\in\mathcal{P}} v_P^{\pi_2}(i) = \min_{\pi\in\Pi} \max_{P\in\mathcal{P}} v_P^{\pi_2}(i) \quad \forall i \in S. \tag{46}$$

That is to say, π_2 is an optimal policy under the optimality criterion defined in (13). ■

Generally, an optimal policy under the optimality criterion using the quadratic total value function may be non-stationary. However, under some assumptions, stationary optimal policies are optimal or ϵ-optimal in the policy space Π.

Theorem 2: If

$$\min_{\pi\in\Pi_s}\max_{P\in\mathcal{P}}\|v_P^\pi\|^2=\max_{P\in\mathcal{P}}\min_{\pi\in\Pi_s}\|v_P^\pi\|^2, \tag{47}$$

stationary optimal policies are optimal in the policy space Π.

Proof: For any fixed $P\in\mathcal{P}$, since $\min_{\pi\in\Pi_s}\|v_P^\pi\|^2=\min_{\pi\in\Pi}\|v_P^\pi\|^2$, by (47),

$$\min_{\pi\in\Pi_s}\max_{P\in\mathcal{P}}\|v_P^\pi\|^2=\max_{P\in\mathcal{P}}\min_{\pi\in\Pi_s}\|v_P^\pi\|^2=\max_{P\in\mathcal{P}}\min_{\pi\in\Pi}\|v_P^\pi\|^2. \tag{48}$$

Since by weak duality, $\max_{P\in\mathcal{P}}\min_{\pi\in\Pi}\|v_P^\pi\|^2\le\min_{\pi\in\Pi}\max_{P\in\mathcal{P}}\|v_P^\pi\|^2$,

$$\min_{\pi\in\Pi_s}\max_{P\in\mathcal{P}}\|v_P^\pi\|^2\le\min_{\pi\in\Pi}\max_{P\in\mathcal{P}}\|v_P^\pi\|^2. \tag{49}$$

Since $\min_{\pi\in\Pi_s}\max_{P\in\mathcal{P}}\|v_P^\pi\|^2\ge\min_{\pi\in\Pi}\max_{P\in\mathcal{P}}\|v_P^\pi\|^2$,

$$\min_{\pi\in\Pi_s}\max_{P\in\mathcal{P}}\|v_P^\pi\|^2=\min_{\pi\in\Pi}\max_{P\in\mathcal{P}}\|v_P^\pi\|^2. \tag{50}$$

That is to say, stationary optimal policies are also optimal in Π. ■

Assume that Θ is a compact and closed subspace of the vector space $\Re^{1\times q}$ and for any $\pi\in\Pi$, the function $\|v_{P(\theta)}^\pi\|^2$ is continuous in θ. Then, given an arbitrarily small constant $\epsilon>0$, for any $\theta_0\in\Theta$, there exists $\delta_{\theta_0}^\pi>0$, such that

$$\left|\|v_{P(\theta)}^\pi\|^2-\|v_{P(\theta_0)}^\pi\|^2\right|<\epsilon,\quad\forall\theta\in\mathcal{B}_{\theta_0}^\pi\cap\Theta \tag{51}$$

where $\mathcal{B}_{\theta_0}^\pi=\{\theta:\|\theta-\theta_0\|<\delta_{\theta_0}^\pi\}$. Since

$$\bigcup_{\theta_0\in\Theta}\mathcal{B}_{\theta_0}^\pi\supseteq\Theta, \tag{52}$$

by Heine-Borel theorem, there exist finite balls to cover Θ. Let those selected balls be $\mathcal{B}_{\theta_1}^\pi,\cdots,\mathcal{B}_{\theta_j}^\pi,\cdots,\mathcal{B}_{\theta_r}^\pi$ and $\mathcal{P}_d^\pi=\{P(\theta_j):1\le j\le r\}$, where θ_j and r depend on π. And then

$$\mathcal{P}_d=\bigcup_{\pi\in\Pi}\mathcal{P}_d^\pi. \tag{53}$$

Theorem 3: With the above assumptions and notations, if for any $\pi\in\Pi_s$,

$$\min_{\pi\in\Pi_s}\max_{P(\theta)\in\mathcal{P}_d}\|v_{P(\theta)}^\pi\|^2=\max_{P(\theta)\in\mathcal{P}_d}\min_{\pi\in\Pi_s}\|v_{P(\theta)}^\pi\|^2, \tag{54}$$

then stationary optimal policies on $\mathcal{P}_d$ are ϵ-optimal and stationary optimal policies on $\mathcal{P}$ are also ϵ-optimal in the policy space Π.

Proof: Without loss of generality, for any $\pi\in\Pi$, let $P(\theta^*)\in\mathcal{P}$, $P(\theta_1)\in\mathcal{P}_d$ and $P(\theta_2)\in\mathcal{P}_d^\pi$ such that

$$\|v_{P(\theta^*)}^\pi\|^2=\max_{P(\theta)\in\mathcal{P}}\|v_{P(\theta)}^\pi\|^2, \tag{55}$$

$$\|v_{P(\theta_1)}^\pi\|^2=\max_{P(\theta)\in\mathcal{P}_d}\|v_{P(\theta)}^\pi\|^2, \tag{56}$$

$$0\le\|v_{P(\theta^*)}^\pi\|^2-\|v_{P(\theta_2)}^\pi\|^2<\epsilon. \tag{57}$$

Since $\|v_{P(\theta_2)}^\pi\|^2\le\|v_{P(\theta_1)}^\pi\|^2\le\|v_{P(\theta^*)}^\pi\|^2$,

$$0\le\max_{P(\theta)\in\mathcal{P}}\|v_{P(\theta)}^\pi\|^2-\max_{P(\theta)\in\mathcal{P}_d}\|v_{P(\theta)}^\pi\|^2<\epsilon. \tag{58}$$

Then,

$$0\le\min_{\pi\in\Pi}\max_{P(\theta)\in\mathcal{P}}\|v_{P(\theta)}^\pi\|^2-\min_{\pi\in\Pi}\max_{P(\theta)\in\mathcal{P}_d}\|v_{P(\theta)}^\pi\|^2\le\epsilon. \tag{59}$$

By (54) and Theorem 2,

$$\min_{\pi\in\Pi_s}\max_{P(\theta)\in\mathcal{P}_d}\|v_P^\pi\|^2=\min_{\pi\in\Pi}\max_{P(\theta)\in\mathcal{P}_d}\|v_P^\pi\|^2. \tag{60}$$

Hence,

$$0\le\min_{\pi\in\Pi}\max_{P(\theta)\in\mathcal{P}}\|v_{P(\theta)}^\pi\|^2-\min_{\pi\in\Pi_s}\max_{P(\theta)\in\mathcal{P}_d}\|v_{P(\theta)}^\pi\|^2\le\epsilon, \tag{61}$$

that is to say, stationary optimal policies on $\mathcal{P}_d$ are ϵ-optimal in Π. Since

$$0\le\min_{\pi\in\Pi_s}\max_{P(\theta)\in\mathcal{P}}\|v_{P(\theta)}^\pi\|^2-\min_{\pi\in\Pi_s}\max_{P(\theta)\in\mathcal{P}_d}\|v_{P(\theta)}^\pi\|^2\le\epsilon, \tag{62}$$

$$\begin{aligned}0&\le\min_{\pi\in\Pi_s}\max_{P(\theta)\in\mathcal{P}}\|v_{P(\theta)}^\pi\|^2-\min_{\pi\in\Pi}\max_{P(\theta)\in\mathcal{P}}\|v_{P(\theta)}^\pi\|^2\\&=\min_{\pi\in\Pi_s}\max_{P(\theta)\in\mathcal{P}}\|v_{P(\theta)}^\pi\|^2-\min_{\pi\in\Pi_s}\max_{P(\theta)\in\mathcal{P}_d}\|v_{P(\theta)}^\pi\|^2\\&+\min_{\pi\in\Pi_s}\max_{P(\theta)\in\mathcal{P}_d}\|v_{P(\theta)}^\pi\|^2-\min_{\pi\in\Pi}\max_{P(\theta)\in\mathcal{P}}\|v_{P(\theta)}^\pi\|^2\\&\le\epsilon.\end{aligned} \tag{63}$$

Hence, stationary optimal policies on $\mathcal{P}$ are ϵ-optimal in Π. ■

Remarks:

(i) If for any fixed $\pi\in\Pi$, the function $\|v_{P(\theta)}^\pi\|^2$ is Lipschitz with the constant L^π that depends on π, that is,

$$\left|\|v_{P(\theta_1)}^\pi\|^2-\|v_{\theta_2}^\pi\|^2\right|\le L^\pi\|\theta_1-\theta_2\|,\quad\forall\theta_1,\theta_2\in\Theta \tag{64}$$

then the parameter $\delta_{\theta_0}^\pi$ can be chosen as follows,

$$\delta_{\theta_0}^\pi=\delta^\pi\triangleq\frac{\epsilon}{L^\pi+1},\quad\forall\theta_0\in\Theta. \tag{65}$$

That is to say, $\delta_{\theta_0}^\pi$ is independent of θ_0.

(ii) If the function $\|v_{P(\theta)}^\pi\|^2$ is continuously differentiable with respect to θ, it is also Lipschitz and the constant L^π can be chosen as follows,

$$\infty>L^\pi\ge\max_{\theta\in\Theta}\left\|\frac{\partial\|v_{P(\theta)}^\pi\|^2}{\partial\theta}\right\|. \tag{66}$$

(iii) The smaller the L^π is, the larger the δ^π should be, and then the smaller the cardinality of $\mathcal{P}_d^\pi$ is, which can make the condition (54) relatively easy to satisfy.

(iv) If for some policies, the constants L^π are equal, the same balls $\mathcal{B}_{\theta_j}^\pi$ can be selected to cover Θ and also their sets $\mathcal{P}_d^\pi$ are equal, which may reduce the cardinality of $\mathcal{P}_d$ and make the condition (54) satisfied more easily.

When for all $\pi\in\Pi$, the functions $\|v_{P(\theta)}^\pi\|^2$ are Lipschitz with the constant L that is independent of π, then the parameters $\delta_{\theta_0}^\pi$ can be independent of not only θ_0 but also of π, that is,

$$\delta_{\theta_0}^\pi=\delta\triangleq\frac{\epsilon}{L+1},\quad\forall\theta_0\in\Theta,\pi\in\Pi, \tag{67}$$

and consequently, the balls $\mathcal{B}^{\pi}_{\theta_0}$ are independent of π , that is,

$$\mathcal{B}^{\pi}_{\theta_0} = \mathcal{B}_{\theta_0} \stackrel{\Delta}{=} \{\theta : \|\theta - \theta_0\| < \delta\}, \quad \forall \pi \in \Pi. \tag{68}$$

Hence, for all π, the same balls $\mathcal{B}^{\pi}_{\theta_j}$ can be selected to cover Θ and also all $\mathcal{P}^{\pi}_d$ are equal. Thus, the set $\mathcal{P}_d$ becomes finite, that is, $\mathcal{P}_d = \mathcal{P}_f \stackrel{\Delta}{=} \{P(\theta_j) : 1 \le j \le r\}$, where θ_j and r are independent of π.

Corollary 1: With the above assumptions and notations, if

$$\min_{\pi\in\Pi_s} \max_{P(\theta)\in\mathcal{P}_f} \|v^{\pi}_P\|^2 = \max_{P(\theta)\in\mathcal{P}_f} \min_{\pi\in\Pi_s} \|v^{\pi}_P\|^2, \tag{69}$$

then stationary optimal policies on $\mathcal{P}_f$ are ϵ-optimal and stationary optimal policies on $\mathcal{P}$ are also ϵ-optimal in the policy space Π.

Proof: By (69) and Theorem 3, stationary optimal policies on $\mathcal{P}_f$ are ϵ-optimal and stationary optimal policies on $\mathcal{P}$ are also ϵ-optimal in Π. ■

Remarks:

(i) If the continuous derivatives of the functions $\|v^{\pi}_{P(\theta)}\|^2$ are uniformly bounded for all $\pi \in \Pi$, the Lipschitz constant L can be chosen as follows,

$$\infty > L \ge \sup_{\pi\in\Pi} \max_{\theta\in\Theta} \left\| \frac{\partial \|v^{\pi}_{P(\theta)}\|^2}{\partial\theta} \right\|. \tag{70}$$

(ii) The smaller the L is, the larger the δ should be, and consequently the smaller the cardinality of $\mathcal{P}_f$ is, which can make the condition (69) relatively easy to satisfy.

III. Robust Policy Iteration under quadratic total value function

Based on the optimality criterion of minimizing the maximum quadratic total value function given in (31), a robust policy iteration is developed to find a stationary optimal policy for MDPs. Even though this solution is generally suboptimal in the deterministic policy space Π, according to Theorems 2, 3, and Corollary 1, this stationary optimal policy can be guaranteed to be optimal or ϵ-optimal in Π. In this section, we consider stationary policies. For simplicity, the subscript "s" which indicates association with stationary policies, is dropped out.

Policy iteration includes policy evaluation and policy improvement steps. Under the policy evaluation step, for any policy $\pi = (\mathbf{a}, \mathbf{a}, \cdots)$, the maximum quadratic total value function is expressed as follows in terms of (7)

$$\begin{aligned} \|v^{\pi}_{P^*}\|^2 &= \max_{P\in\mathcal{P}} \|v^{\pi}_P\|^2 \\ &= \max_{P\in\mathcal{P}} (C^{\pi})' \left(I - \gamma (P^{\pi})'\right)^{-1} (I - \gamma P^{\pi})^{-1} C^{\pi}, \end{aligned} \tag{71}$$

The maximum value $\|v^{\pi}_{P^*}\|^2$ and the optimal P^* are to be calculated. Such approaches are available to compute these values. Actually, if the transition matrix for the policy π is independent, an efficient iterative method given in Algorithm 1 can be used to compute $\|v^{\pi}_{P^*}\|^2$.

Remarks:

(i) Given a policy π, for the maximum point P^*, the transition probability rows for all states i and the corresponding actions $\mathbf{a}(i)$, denoted by $\left(P_i^{\mathbf{a}(i)}\right)^*$, are computed as follows,

$$\left(P_i^{\mathbf{a}(i)}\right)^* = \arg \max_{P_i^{\mathbf{a}(i)} \in \mathcal{P}_i^{\mathbf{a}(i)}} \left\{ P_i^{\mathbf{a}(i)} v^{\pi}_k \right\}, \quad i \in S, \tag{73}$$

and there are no special constraints for other rows.

(ii) A proof of convergence of the iterative algorithm under the total value function can be obtained similar to that in [6], [7].

(iii) In principle, the initial value v^{π}_0 can be selected arbitrarily in $\Re^{n\times 1}$. However, a carefully selected v^{π}_0 can accelerate the convergence of the iterative process. To see that, let

$$\mathcal{G}_1 = \{v : v(i) \le c(i, \mathbf{a}(i)) + \gamma \max_{P_i^{\mathbf{a}(i)} \in \mathcal{P}_i^{\mathbf{a}(i)}} P_i^{\mathbf{a}(i)} v\} \tag{74}$$

and

$$\mathcal{G}_2 = \{v : v(i) \ge c(i, \mathbf{a}(i)) + \gamma \max_{P_i^{\mathbf{a}(i)} \in \mathcal{P}_i^{\mathbf{a}(i)}} P_i^{\mathbf{a}(i)} v\}. \tag{75}$$

If $v^{\pi}_0 \in \mathcal{G}_1$, the sequence $\{v^{\pi}_k\}$ is non-decreasing. If $v^{\pi}_0 \in \mathcal{G}_2$, the sequence $\{v^{\pi}_k\}$ is non-increasing. Thus, the sequence $\{v^{\pi}_k\}$ converges to $v^{\pi}_{P^*}$ faster than those with $v^{\pi}_0 \notin \mathcal{G}_1 \cup \mathcal{G}_2$.

For the policy improvement step, the policy can be improved easily for MDPs with independent transition matrices using robust policy iteration [7]. However, such improvement method does not guarantee solutions for MDPs with correlated transition matrices. Hence, the policy elimination method is considered herein as a means of improving policy. The policy elimination process entails elimination of some non-optimal policies during each iteration by a necessary condition that a stationary policy is optimal. The inequality $\|v^{\mu}_{P^*}\|^2 \le \|v^{\pi_k}_{P^*}\|^2$ given in (77) is such a necessary condition for a stationary policy μ being optimal, where $\|v^{\mu}_{P^*}\|^2$ is the quadratic total value function of μ at P^*, $\|v^{\pi_k}_{P^*}\|^2$ is the maximum quadratic total value function of π_k and P^* is the corresponding optimal matrix. Robust policy iteration under quadratic total value function is described in details in Algorithm 2.

Remarks:

(i) Robust policy iteration under quadratic total value function terminates in finite iterations when the cardinality of the set Π_k, denoted by $|\Pi_k|$ in Algorithm 2, is equal to one, since the cardinality of Π_0 is finite and $\Pi_{k+1} \subset \Pi_k (k = 0, 1, 2, \cdots)$. The sequence $\{\pi_k\}$ converges to a stationary optimal policy.

(ii) A good initial policy π_0, which has relatively small total value function, can accelerate the convergence process.

Algorithm 1 Iterative Algorithm for maximum quadratic total value function of a stationary policy with its independent transition matrix

1. select $v^{\pi}_0 \in \Re^{n\times 1}$ and set $k = 0$;

2. compute v^{π}_{k+1} by

$$v^{\pi}_{k+1}(i) = c(i, \mathbf{a}(i)) + \gamma \max_{P_i^{\mathbf{a}(i)} \in \mathcal{P}_i^{\mathbf{a}(i)}} P_i^{\mathbf{a}(i)} v^{\pi}_k, \quad i \in S; \tag{72}$$

3. terminate if $v^{\pi}_{k+1} = v^{\pi}_k$; otherwise, increment k by 1 and go to **2**;

4. output $\|v^{\pi}_k\|^2$ as $\|v^{\pi}_{P^*}\|^2$.

Algorithm 2 Robust Policy Iteration Under Quadratic Total Value Function

1. Initialization: set $k = 0$, $\Pi_0 = \Pi_s$, $\mathcal{O}_{-1} = +\infty$, $\{\pi_{-1}\} = \emptyset$, and select $\pi_0 \in \Pi_0$;

2. Policy evaluation: compute $\|v_{P^*}^{\pi_k}\|^2$ and P^* such that

$$\|v_{P^*}^{\pi_k}\|^2 = \max_{P \in \mathcal{P}} \|v_P^{\pi_k}\|^2; \tag{76}$$

3. Policy improvement:

if $|\Pi_k| > 1$

(a) eliminate policies to obtain $\widetilde{\Pi}_k$ such that

$$\widetilde{\Pi}_k = \{\mu \in \Pi_k : \|v_{P^*}^{\mu}\|^2 \leq \|v_{P^*}^{\pi_k}\|^2\}; \tag{77}$$

if $|\widetilde{\Pi}_k| > 1$

if $\|v_{P^*}^{\pi_k}\|^2 < \mathcal{O}_{k-1}$

(b) set $\mathcal{O}_k = \|v_{P^*}^{\pi_k}\|^2$, $\widehat{\Pi}_k = \widetilde{\Pi}_k - \{\pi_{k-1}\}$;

(c) if $|\widehat{\Pi}_k| = 1$, set $\Pi_{k+1} = \widehat{\Pi}_k = \{\pi_k\}$, $\pi_{k+1} = \pi_k$, $\mathcal{O}_{k+1} = \mathcal{O}_k = \|v_{P^*}^{\pi_k}\|^2$, increment k by 1, and go to **4**; otherwise, go to (d);

(d) set $\Pi_{k+1} = \widehat{\Pi}_k$, select $\pi_{k+1} \in \Pi_{k+1} - \{\pi_k\}$, increment k by 1, and return to **2**;

else

(e) if $|\widetilde{\Pi}_k - \{\pi_k\} - \{\pi_{k-1}\}| \geq 1$, select $\widetilde{\pi}_k \in \widetilde{\Pi}_k - \{\pi_k\} - \{\pi_{k-1}\}$, set $\Pi_k = \widetilde{\Pi}_k - \{\pi_k\}$, $\pi_k = \widetilde{\pi}_k$, and return to **2**; otherwise, set $\Pi_k = \widetilde{\Pi}_k - \{\pi_k\} = \{\pi_{k-1}\}$, $\pi_k = \pi_{k-1}$, $\mathcal{O}_k = \mathcal{O}_{k-1} = \|v_{P^*}^{\pi_{k-1}}\|^2$;

else

(f) set $\Pi_{k+1} = \widetilde{\Pi}_k = \{\pi_k\}$, $\pi_{k+1} = \pi_k$, $\mathcal{O}_{k+1} = \|v_{P^*}^{\pi_k}\|^2$, increment k by 1, and go to **4**;

else

(g) go to **4**;

4. Termination: output π_k as the stationary optimal policy with the corresponding optimal transition matrix P^* and maximum quadratic total value function $\mathcal{O}_k$.

(iii) Robust policy iteration under quadratic total value function can be extended to solving problems with parameterized cost functions and weighted quadratic total value functions.
(iv) Assume that the set Θ is compact and closed in $\Re^{1\times q}$, and if for any fixed π and $P(\theta)$, the quadratic total value function can be computed within arbitrarily small error bound, then Algorithm 2 can be extended to obtain a stationary ϵ-optimal policy.

IV. Conclusion

A theoretical framework is proposed to study MDPs with uncertainties in the transition probability matrices. As a result, the paper concludes with a robust policy iteration procedure under a newly proposed total value function, which guarantees optimal or near-optimal solutions. Such a robust construct may be especially useful in practical applications when there is too limited amount of experimental data to obtain an accurate transition probability matrix, or when an accurate estimation of the probability transition matrix is unrealistic. The proposed framework is straightforward to apply in practical problems.

References

[1] P. Bellman and R. kalaba, *Dynamic Programming and Modern Control Theory*, New York: Academic Press, 1965.

[2] M. Putterman, *Markov Decision Processes: Discrete Stochastic Dynamic Programming*, Wiley-Interscince, New York, 1994.

[3] Hyeong Soo Chang, Hong-Gi Lee, Michael C. Fu, and Steven I. Marcus, "Evolutionary Policy Iteration for Solving Markov Decision Processes," *IEEE Transactions on Automatic Cntrol*, vol. 50, n. 11, pp. 1804-1808, 2005.

[4] D. Berstsekas and J. Tsitsiklis, *Neuro-Dynamic Programming*, Athena Scientific Massachusetts, 1996.

[5] J. Si, A. G. Barto, W. B. Powell and D. Wunsch, *Handbook of Learning and Approximate Dynamic Programming*, Wiley-IEEE Press, 2004.

[6] 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.

[7] J. K. Satia and R. E. Lave, Jr., "Markov Decision Processes with Uncertain Transition Probabilities," *Operations Research*, vol. 21, pp. 728-740, 1973.

[8] C. C. White and H. K. Eldeib, "Markov Decision Processes with Imprecise Transition Probabilities," *Operations Research*, vol. 42, pp. 739-749, 1994.

[9] R. Givan , S. Leach, and T. Dean, "Bounded Parameter Markov Decision Processes," *Artificial Intelligence*, vol. 122, no. 1-2, pp. 71-109, 2000.

[10] 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.

[11] M. Abbad, and J. A. Filar, "Perturbation and Stability Theory for Markov Control Problems," *IEEE transations on Automatic Control*, vol. 37, pp. 1415-1420, 1992.