Appl. Math. J. Chinese Univ.
2020, 35(4): 417-436

Robust analysis of discounted Markov decision processes with uncertain transition probabilities

LOU Zhen-kai[1] HOU Fu-jun[1,*] LOU Xu-ming[2]

Abstract. Optimal policies in Markov decision problems may be quite sensitive with regard to transition probabilities. In practice, some transition probabilities may be uncertain. The goals of the present study are to find the robust range for a certain optimal policy and to obtain value intervals of exact transition probabilities. Our research yields powerful contributions for Markov decision processes (MDPs) with uncertain transition probabilities. We first propose a method for estimating unknown transition probabilities based on maximum likelihood. Since the estimation may be far from accurate, and the highest expected total reward of the MDP may be sensitive to these transition probabilities, we analyze the robustness of an optimal policy and propose an approach for robust analysis. After giving the definition of a robust optimal policy with uncertain transition probabilities represented as sets of numbers, we formulate a model to obtain the optimal policy. Finally, we define the value intervals of the exact transition probabilities and construct models to determine the lower and upper bounds. Numerical examples are given to show the practicability of our methods.

§1 Introduction

Markov decision process models have gained recognition in such diverse fields as ecology, economics, and communications engineering (Puterman, 2014). In most articles, transition probabilities are regarded as known information. Yet in some cases, part of these are unknown for decision makers, and an optimal policy may be sensitive to these transition probabilities.

More than forty years ago, Satia et al. (1973) discussed the MDPs with uncertain transition probabilities and showed a model based on game-theoretic and the Bayesian formulation. After that, numerical algorithms were proposed to obtain an optimal max-min strategy in this type of MDPs (White et al., 1994). Later, Kalyanasundaram et al. (2002) explored these issues

Received: 2018-11-19. Revised: 2019-12-25.
MR Subject Classification: 60J10, 90C40, 90C05.
Keywords: Markov decision processes, uncertain transition probabilities, robustness and sensitivity, robust optimal policy, value interval.
Digital Object Identifier(DOI): https://doi.org/10.1007/s11766-020-3664-1.
Supported by the National Natural Science Foundation of China (71571019).
*Corresponding author.

under a long-term average criterion and developed solution techniques to obtain a max-min optimal policy. Since Garud (2005) proposed a robust formulation to deal with the Markov decision problems with uncertain transition probabilities, this framework was adopted by many other scholars. Nilim et al. (2005) studied the robust control problem of a MDP with uncertain transition matrices, and provided a robust dynamic programming algorithm without adding extra computing cost. Li et al. (2007) showed a robust policy iteration by taking all initial states into account under a min-max criterion. Similarly, Delage et al. (2009) considered the tradeoff between optimistic and pessimistic point of views, and then proposed a set of percentile criterion to handle the MDPs with uncertain probabilities. Xu et al. (2012) proposed a decision criterion based on distributional robustness and found the optimal strategy under the most adversarial admissible probabilities distributions. In addition, Wiesemann et al. (2013) derived a confidence region of uncertain probabilities by taking advantage of an observation history of a discounted MDP, and obtained an optimal policy by the max-min criterion.

Despite the great progress of the MDPs with uncertain transition probabilities, there are still some important problems unsolved. Does an optimal policy remain unchanged when the uncertain transition probabilities change? What is the price to obtain the exact transition probabilities? In this paper, we intend to carry out a thorough study to address these problems.

Similar to Wiesemann et al. (2013), we first establish a programming model to estimate the unknown probabilities by using an observation history. Through analyzing the linear programming method, we deeply discuss the robustness of an optimal policy and the sensitivity of the highest expected total reward, and then develop solution techniques for them. Afterwards, we give a definition for a robust optimal policy with uncertain transition probabilities represented as sets of numbers, and formulate a model to obtain it. Furthermore, we prove that the highest expected total reward gained by a robust optimal policy cannot be greater than the one under any exact transition probabilities. Next, we describe the value of the exact transition probabilities. In order to obtain the value interval of the exact transition probabilities, we first prove that the highest expected total reward $U(i)$ is a continuous function of these transition probabilities. After that we propose two models to obtain the value interval with considering the feature of this problem. Clearly, the value interval is significant for both optimistic and pessimistic decision makers, namely, how much it is worth to obtain the exact transition probabilities.

§2 Notations description

In this paper we focus on a time-homogeneous MDP with some uncertain transition probabilities. The number of each state or action is finite. We assume that a partial observation history of the MDP is available. Under an expected discounted reward criterion, we search an optimal policy over an infinite horizon. For the sake of discussion, we make some descriptions for the discounted MDP. For more information, see Puterman (2014).

(1) Let T denote the set of decision epochs. Thus, the decision epochs of an infinite horizon MDP can be denoted by $T = \{0, 1, \cdots\}$.

(2) The set of all possible states is denoted by $\Omega = \{1, 2, \cdots, k\}$, where the number of state is finite. Let $S_n = i$be the state at period n, where $i \in \Omega$.

(3) Let A (i) be the set of allowable actions in state i. We denote by $\pi = (\pi_0, \pi_1, \cdots, \pi_t, \cdots)$ a sequence of decision rules of a MDP, and π_t is the decision rule at period$t$. Apparently, we have

$$\sum_{a \in A(i)} \pi_t(a|i) = 1.$$

We let Π be the set of stationary randomized policies and Π^d be the set of stationary deterministic policies. A deterministic policy is denoted by f

(4) At any decision point n, we denote by $p(j|\ i,\ a)$ the probability of one-step transition probability from state$i$ to state j by taking action a.

(5) At any decision point n, system would gain a reward $r(i, a)$ when it is in state$i$and takes action a.

To summarize, an infinite horizon MDP can be defined as a 5-tuple {T, Ω, A (i), $p(j|\ i,\ a)$, $r(i, a)$}. Sequential states and actions constitute a history. We denote by h_t:=(S_0, a_0, S_1, a_1, ... , S_{t-1}, a_{t-1}, S_t) the track of a MDP from period 0 to period t.

We use the notation $U_\beta(i,\pi)$ to denote the expected total reward of a MDP, where β is a discount factor. The expected total reward of a discounted MDP with initial state i can be defined as follows

$$U_\beta(i, \pi) \equiv \sum_{t=0}^{\infty} \beta^t E_\pi^i[r(X_t, \Delta_t)], i \in \Omega, 0 < \beta < 1,$$

where X_t is the random state and $_t$ the random action at period t Generally, they are both probability distributions.

A policy π^* is an optimal policy, if it satisfies

$$U_\beta^*(i) = U_\beta(i, \pi^*) \equiv \sup_{\pi \in \prod} U_\beta(i, \pi).$$

In Markov decision theory, the existence of optimal policies has been proved. Hence, we confirm that there definitely exists a deterministic stationary policy f^* that simultaneously is an optimal policy. The optimal policies we will mention below refer to deterministic stationary policies.

There are three standard methods to obtain the highest expected total reward: policy iteration, value iteration and linear programming (Kalyanasundaram et al., 2002). Each of the above algorithms has its feature, but the linear programming method seems better for dealing with the sensitivity of the parameters.

Based on the property of condensing mappings, the linear programming model of obtaining the highest expected total reward is given as follows:

$$\begin{aligned} &\min \sum_{i \in \Omega} \frac{1}{k} u(i) \\ &subjecttor(i, a) + \beta \sum_{j \in \Omega} p(j|i, a) u(j) \le u(i), a \in A(i), i \in \Omega. \end{aligned} \tag{1}$$

The value of each independent variable $u(i)$ in the optimal solution of model (1) actually is equal to the highest expected total reward $U_\beta^*(i)$.

§3 Estimation of the unknown transition probabilities

In most analyses of Markov decision problems, the transition probabilities are given. However, in many practical problems one may know only a partial set of the probabilities. There are many ways to estimate the unknown transition probabilities. For example, Baik et al. (2006) provided an ordered profit model.

In this section, we propose a method of estimating the unknown transition probabilities by taking advantage of an observation history. We assume that the actions taken before are visible, and at that least two states in the history are available. We disregard how the actions may be generated and simply regard them as known information.

As a first example, we investigate the case with two states available. We denote by $S_0 = i_0$ the initial state, and by $S_n = i_n$ the state at period n. Both states are visible. Given an action sequence a_0, $\cdots$, a_{n-1}, we formulate a recursive expression to derive the probability of each state at every period.

Given an initial condition $p(S_0 = i_0)$=1, thus $p(S_0 = i, i \neq i_0)$=0. We obtain the recurrence formula as follows:

$$p(S_{q+1} = j) = \sum_{i\in\Omega} p(S_q = i)p(S_{q+1} = j|i, a_q), q = 0, \cdots, n-1 \tag{2}$$

With formula (2) and the initial condition we can calculate the probability $p(S_n = i_n)$. Due to the partially unknown transition probabilities, the expression $p(S_n = i_n)$ is actually a function.

In order to be able to maximize $p(S_n = i_n)$, we propose a mathematical model to determine the unknown transition probabilities by the maximum likelihood method.

$$\begin{aligned} &\max p(S_n = i_n) \\ &\textit{subject to } \sum_{j\in\Omega} p(j|i, a) = 1, a \in A(i), i \in \Omega \\ &\qquad 0 \leq p(j|i, a) \leq 1, a \in A(i), i \in \Omega, j \in \Omega \end{aligned} \tag{3}$$

The objective function in model (3) is determined by the recurrence formula (2).

More generally, we indicate that the above model is appropriate for cases with more than two states visible. Specifically, if k states are visible, we simply regard the first k-1 states as S_0 and the last k-1 states as S_n. Thus, we can obtain the estimated values by means of a model similar to (3). The following example describes the process of obtaining the values.

Given a state space Ω={1,2}, the available action set of each state is A(1)=A(2)={a_1,a_2}. The transition probabilities while taking action a_1 are known, and we assume that $p(1|1,a_1) = 0.7$, $p(2|1,a_1) = 0.3$, $p(1|2,a_1) = 0.1$ and $p(2|2,a_1) = 0.9$. The transition probabilities of taking action a_2 in state 1 are estimated values, and we denote$p(1|1,a_2) = p$ and $p(2|1,a_2) = 1$-p. The remaining transition probabilities are known with certainty, and we assume that $p(1|2,a_2) = 0.2$ and $p(2|2,a_2) = 0.8$.

We assume that states at stage 0, stage 2 and stage 4 are known, and S_0=1, S_2=1 and S_4=2. The corresponding actions taken at each stage are known, and we assume that a_2, a_1,a_1 anda_2 correspond to stage 0, stage 1, stage 2 and stage 3, respectively.

We then estimate the unknown transition probabilities by utilizing the above information. According to the recurrence formula (2), $p(S_2{=}1) = 0.6p + 0.1$. We then consider $S_2{=}1$ as known information, and similarly obtain $p(S_4{=}2) = 0.94 - 0.7p$. Thus, we formulate a nonlinear programming model to perform the estimation based on model (3):

$$\max\{(0.6p + 0.1)(0.94 - 0.7p)\}$$

$$subject\ to\ 0 \le p \le 1$$

The process of acquiring the objective function of the above model can be described by figure 1.

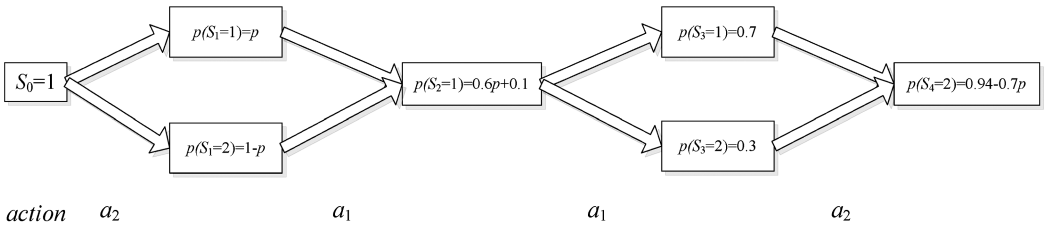

Figure 1. The process of acquiring the objective function.

By solving the above quadratic function we obtain the optimal value, i.e., $\max\{p(S_2 = 1) \cdot p(S_4 = 2)\} \approx 0.24$ when $p \approx 0.6$.

Estimates of the transition probabilities may nevertheless be inaccurate. In the next section we focus on whether or not an optimal policy remains unchanged when uncertain transition probabilities change.

§4 Robust analysis of the discounted MDPs

The scope of application for the policy iteration and the linear programming are nearly the same. Both require a finite action set and a finite state space. In contrast, the linear programming method needs more calculations for dealing with the simplex tableau. Moreover, an optimal policy cannot be obtained directly by linear programming.

However, the linear programming approach does has a unique advantage. Indeed, we can obtain an optimal policy by searching the equalities in the constraints of model (1). For example, assume that the highest expected total reward $U(i)$ associated with initial state $i$has been acquired, $i \in \Omega$, and we place this within the constraints of model (1). As a consequence, there must exist k equalities that satisfy

$$r(i, a) + \beta \sum_{j \in \Omega} p(j|i, a)U(j) = U(i)$$

The equality above means that the optimal action is a when given an initial state i. Furthermore, we are able to analyze the sensitivity of the parameters by using the linear programming algorithm.

4.1 Robust analysis of an optimal policy

Based on the above analysis, we further analyze the robustness of an optimal policy. Robustness here refers to the stability of an optimal policy when one or more uncertain transition probabilities change.

Theorem 4.1 The sufficient and necessary conditions for an optimal policy to be unchanged are that each of the following equalities holds before and after transition probabilities $p(j|i,\, a)$ change:

$$r(i,a)+\beta\sum_{j\in\Omega}p(j|i,a)u(j)=u(i), a\in A(i), i\in\Omega \tag{4}$$

Proof First we focus on the sufficiency of the condition. Given an initial state i despite the fact that the highest expected total reward may change with changes in the unknown transition probabilities, the action a corresponding to state i in equality (4) has not changed. According to the analysis above, we know that action $a$in equality (4) is the optimal action for initial state i. For any state $i\in\Omega$, we have an equality. This means that the optimal deterministic stationary policy remains unchanged. From the previous discussion we know that the optimal deterministic stationary policy of a discounted MDP is an optimal policy.

We propose a proof by contradiction to show the necessity condition. Suppose that an optimal policy remains unchanged when some transition probabilities change. There exists an initial state i for which we have the following equality before the transition probabilities change:

$$r(i,a)+\beta\sum_{j\in\Omega}p(j|i,a)U(j)=U(i)$$

Because of the reverse assumption, the above equality no longer makes sense after the transition probabilities change, namely

$$r(i,a)+\beta\sum_{j\in\Omega}p(j|i,a)U'(j)<U'(i)$$

According to the theory of linear programming, we know that there must exist another action b that makes the following equality true:

$$r(i,b)+\beta\sum_{j\in\Omega}p(j|i,b)U'(j)=U'(i)$$

The above equation means that the optimal policy changes, which contradicts the assumption. Hence, we affirm that the assumption is false. The necessity of the condition has been proved.

Definition 4.1 The robust range of a given optimal policy is described as follows: when uncertain transition probabilities change in this region, the optimal policy remains unchanged.

Next, we develop a solution technique to obtain the robust range of a given optimal policy.

Step **1** Let unknown numbers $p'(j|i,\, a)$ replace the estimated values of the transition probabilities$p(j|i,\, a)$. Obviously, $p'(j|i,\, a)$ still satisfies$\sum\limits_{j\in\Omega}p'(j|i,a)=1, a\in A(i), i\in\Omega$.

Step **2** Find the k equalities corresponding to an optimal policy f^*, and regard them as a system of equations with k variables$u(1),\cdots,\, u(k)$. Solve the system to obtain the expression for each $u(i)$. Generally $u(i)$ is a function of $p'(j|i,\, a)$.

Step **3** Substitute the expression of each $u(i)$ into the surplus inequalities of the constraints

in model (1). Thus we obtain the inequalities that only contain several $p'(j|i,\ a)$. Taking into account $\sum_{j\in\Omega} p'(j|i,a) = 1, a \in A(i), i \in \Omega$, we obtain the value range of each variable$p'(j|i,\ a)$.

It is clear that an optimal policy remains unchanged when all the uncertain transition probabilities change within the value range obtained by the above method.

Theorem 4.2 The equation group in step 2 has a unique solution, namely, the expression of each $u(i)$ is unique.

Proof Without loss of generality, the system of equations mentioned above can be expressed as follows:

$$\begin{cases} r(1,a_1)+\beta p(1|1,a_1)u(1)+\beta p(2|1,a_1)u(2)+\cdots+\beta p(k|1,a_1)u(k)=u(1) \\ r(2,a_2)+\beta p(1|2,a_2)u(1)+\beta p(2|2,a_2)u(2)+\cdots+\beta p(k|2,a_2)u(k)=u(2) \\ \vdots \\ r(k,a_k)+\beta p(1|k,a_k)u(1)+\beta p(2|k,a_k)u(2)+\cdots+\beta p(k|k,a_k)u(k)=u(k). \end{cases}$$

As $u(i)$ is an independent variable here, we adjust the form of the system of equations as follows:

$$\begin{cases} (\beta p(1|1,a_1)-1)u(1)+\beta p(2|1,a_1)u(2)+\cdots+\beta p(k|1,a_1)u(k)=-r(1,a_1) \\ \beta p(1|2,a_2)u(1)+(\beta p(1|2,a_2)-1)u(2)+\cdots+\beta p(k|2,a_2)u(k)=-r(2,a_2) \\ \vdots \\ \beta p(1|k,a_k)u(1)+\beta p(2|k,a_k)u(2)+\cdots+(\beta p(k|k,a_k)-1)u(k)=-r(k,a_k). \end{cases} \tag{5}$$

In model (5), the terms on the right of the equal signs are constants. We express the coefficient matrix of model (5) as follows:

$$\begin{bmatrix} (\beta p(1|1,a_1)-1) & \beta p(2|1,a_1) & \cdots & \beta p(k|1,a_1) \\ \beta p(1|2,a_2) & (\beta p(1|2,a_2)-1) & \cdots & \beta p(k|2,a_2) \\ \vdots & \vdots & \vdots & \vdots \\ \beta p(1|k,a_k) & \beta p(2|k,a_k) & \cdots & (\beta p(k|k,a_k)-1) \end{bmatrix}$$

As $0<\beta<1$, we have $\beta\sum_{j\in\Omega} p(j|i,a_i) = \beta < 1$. Considering the elements in each row of the matrix above, we have the following result:

$$|\sum_{j\in\Omega, j\neq i} \beta p(j|i,a_i)| = |\beta - p(i|i,a_i)| < |1-p(i|i,a_i)|$$

Hence, the coefficient matrix of the linear equations is a strictly diagonally dominant matrix. As a consequence, the value of the determinant of the coefficient matrix cannot be zero. According to Cramer's rule in linear algebra we know that the equation group mentioned above has a unique solution.

In order to present the process of the robust analysis clearly, we provide an example.

We adopt the same conditions used in the above section. Given a state space Ω={1,2}, the available action set of each state is A(1)=A(2)={a_1a_2}. The discount factor is assumed to be β=0.9. The transition

probabilities while taking action a_1 are accurate, and we assume that $p(1|1,a_1) = 0.7$ $p(2|1,a_1) = 0.3$ $p(1|2,a_1) = 0.1$ and $p(2|2,a_1) = 0.9$ The transition probabilities of taking action a_2 in state 1 are estimated values, and from the above section we know that$p(1|1,a_2) = 0.6$ and $p(2|1,a_2) =$

0.4. The remaining probabilities are accurate, and we assume that $p(1|2,a_2) = 0.2$ and $p(2|2,a_2) = 0.8$ The corresponding rewards are assumed to be $r(1,a_1) = 6$ $r(1,a_2) = 8$ $r(2a_1) = 1$ and $r(2a_2) = 3$

According to equation (1), we model the above example of a discounted MDP as follows:

$$\begin{aligned} &\min \ \ \tfrac{1}{2}\{u(1)+u(2)\} \\ &\begin{cases} 6+0.9(0.7u(1)+0.3u(2)) \le u(1)(i) \\ 8+0.9(0.61u(1)+0.4u(2)) = u(1)(ii) \\ 1+0.9(0.1u(1)+0.9u(2)) \le u(2)(iii) \\ 3+0.9(0.2u(1)+0.8u(2)) = u(2)(iv) \end{cases} \end{aligned}$$

It is easy to obtain the solution for the above model by using the simplex method. This yields $U(1)$=$u(1)$=51.9, $U(2)$=$u(2)$=44.1. In the constraints, (ii) and (iv) hold equal signs. Hence, the optimal policy in this example is f^*= (a_2,a_2). Namely, no matter whether the current state is 1 or 2, the optimal action should be a_2.

According to the steps of the robust analysis, we then set about dealing with the follows inequalities:

$$\begin{cases} 6+0.9(0.7u(1)+0.3u(2)) \le u(1)(v) \\ 8+0.9(\lambda_1 u(1)+\lambda_2 u(2)) = u(1)(vi) \\ 1+0.9(0.1u(1)+0.9u(2)) \le u(2)(vii) \\ 3+0.9(0.2u(1)+0.8u(2)) = u(2)(viii) \\ \lambda_1+\lambda_2 = 1, \lambda_1 \ge 0, \lambda_2 \ge 0. \end{cases}$$

We obtain the expressions for $u(1)$ and$u(2)$ by the equalities (vi) and $(viii)$.

$$\begin{cases} u(1) = \dfrac{17.64-9.64\lambda_1}{0.42-0.32\lambda_1} \\ u(2) = \dfrac{15.84-9.63\lambda_1}{0.42-0.32\lambda_1}. \end{cases}$$

Substituting the expressions into inequalities (v) and (vii), we obtain the solutions $0.284 \le \lambda_1 \le 1, 0 \le \lambda_2 \le 0.716$. Clearly, the estimated values λ_1=0.6 and λ_2=0.4 are located in these intervals.

As the above intervals are wide, we can say that the optimal policy f^*=(a_2, a_2)^T in this example is strongly robust.

4.2 Sensitivity analysis of the highest expected total reward

In practical problems, we are concerned with how the highest expected total reward changes when the transition probabilities vary.

Sometimes we determine an optimal policy based on estimated transition probabilities. However, the highest expected total reward $U(i)$ may be sensitive when the exact transition probabilities are larger or smaller than their estimates. In this subsection we assume that the optimal policy remains unchanged.

Similar to the robust analysis of an optimal policy, we provide a method to handle the sensitivity of the highest expected total reward.

Step **1** Let unknown numbers $p'(j|i,\ a)$ replace the estimated values of the transition

probabilities$p(j|i,\ a)$. Clearly, $p'(j|i,\ a)$ still satisfies $\sum\limits_{j\in\Omega} p'(j|i,a)=1, a\in A(i), i\in\Omega$.

Step **2** Denote by $U(i)$ the highest expected total reward given an initial state i, and assume that the optimal action for state i is$a$. According to the optimal policy f^* and the following formula

$$U(i)=r(i,a)+\beta\sum_{j\in\Omega}p(j|i,a)=U(j)$$

we obtain k linear equations and $\sum\limits_{i\in\Omega}|A(i)|-k$ inequalities.

Step **3** Solve above system of linear equations to obtain the expression for each $U(i)$. After substituting the expressions into the inequalities, we solve the inequalities to obtain the value range for each variable$p'(j|i,\ a)$.

Step **4** Analyze the change in the highest expected total reward $U(i)$ when the considered uncertain transition probabilities vary within their value ranges.

In subsection 4.1 we obtained the unique optimal policy f^*= (a_2,a_2) with the estimated transition probabilities $p(j|i,\ a)$, and we also derived the highest expected total reward $U(1)=51.9$, $U(2)=44.1$. We denote by $U'(1)$ and $U'(2)$ the highest expected total rewards when the estimated transition probabilities change. According to the above method, we have

$$\begin{cases} 8+0.9(\lambda_1U'(1)+\lambda_2U'(2))=U'(1)\\ 3+0.9(0.2U'(1)+0.8U'(2))=U'(2). \end{cases}$$

From subsection 4.1 we know that the expressions for $U'(1)$ and $U'(2)$ are as follows:

$$\begin{cases} U'(1)=\dfrac{17.64-9.64\lambda_1}{0.42-0.32\lambda_1}\\ U'(1)=\dfrac{17.64-9.64\lambda_1}{0.42-0.32\lambda_1} \end{cases}$$

and the intervals in which the optimal policy remains unchanged are $0.284\leq\lambda_1\leq1$, $0\leq\lambda_2\leq0.716$.

Take, for example, $U'(1)$. We analyze its sensitivity with respect to the inaccurate transition probabilities. Let $\lambda_1$be 0.5, and as a consequence $U'(1)$ is equal to 49.3, a value that has been decreased by 5%. When λ_1 is equal to 0.4, the corresponding $U'(1)$ is 47.2. Apparently, the sensitivity of the highest expected total reward in this example is low. Of course, the result may be related to other parameters, for example the reward value $r(i,\ a)$.

In this subsection we have examined the sensitivity of the highest expected total reward when transition probabilities vary. Next, we will demonstrate that the highest expected total reward is a continuous function of $p(j|i,\ a)$ whether or not the optimal policy changes.

§5 Values of the exact transition probabilities

In the previous section we dealt with the robustness of a given optimal policy and the sensitivity of the highest expected total reward. In other words, when uncertain transition probabilities vary in different regions, the corresponding optimal policy may be different. Next, we disregard how to determine these regions. Following Reis et al. (2019), we simply provide an interval for each $p(j|i,\ a)$ directly.

In this section, we continue using linear programming to further explore the above consider-

ations. Although other methods such as policy iteration are suited to robust analysis problems (Kalyanasundaram et al. 2002, Wiesemann et al. 2013), from section 4 we know that in general they are incapable of acquiring the robust range of an optimal policy. In addition, it seems that the linear programming has unique advantages to analyze the mathematical properties of the highest expected total reward.

5.1 Robust optimal policy

As some transition probabilities are uncertain, we need to analyze the effect caused by the variation of the transition probabilities when choosing a policy. Actually, this is a robust decision-making problem. The sup-inf model has been widely adopted to cope with this type of issue since the work of Garud in 2005. In this subsection we also take advantage of this principle to find a robust optimal policy.

Let $\Phi_{ij}(a)$ be the interval of a certain transition probability $p(j|i,\ a)$. We denote by $\Phi = \{\Phi_{ij}(a)|\Phi_{ij}(a) = [p^m_{ij(a)}, p^M_{ij(a)}], p^m_{ij(a)} \le p(j|i,a) \le p^M_{ij(a)}, i \in \Omega, j \in \Omega, a \in A(i)\}$ the interval set of all transition probabilities. $\Phi_{ij}(a)$ becomes a single point when $p(j|i,\ a)$ is an exact value, i.e., $p^m_{ij(a)} = p(j|i,a) = p^M_{ij(a)}$. Under the above conditions, the problem of finding a robust optimal policy for an infinite horizon discounted MDP with initial state i can be modeled as follows:

$$\begin{aligned} &\sup_{f\in\Pi^d} \inf_{p(j|i,a)\in\Phi_{ij}(a)} u(i) \\ &subject\ to \sum_{j\in\Omega} p(j|i,a) = 1, a \in A(i), i \in \Omega \\ &\qquad 0 \le p(j|i,a) \le 1, a \in A(i), i, j \in \Omega \end{aligned} \tag{6}$$

In model (6), $u(i)$ is derived from the objective function of model (1). Let $U^*(i)$ be the optimal solution of model (6). Apparently, $U^*(i)$ is the robust highest expected total reward. We remark that the reason why we use supremum and infimum here is that some intervals may be open.

Next we describe a method of solving model (6). Denote by Π^d=$\{f_1, \cdots, f_m\}$ the stationary policy set, where $m = \sum_{i\in\Omega} |A(i)|$.

Step **1** Set j:=1.

Step **2** Take f_j out of Π^d, and establish $k$linear equations similar to equations (5). According to Cramer's rule we obtain the expression of each $u(i)$. By theorem 4.2 we know that the expression is unique.

Step **3** The expression of each $u(i)$ is a continuous function of some transition probabilities $p(j|i,\ a)$. We then find the minimum $u_j(i)$ when each $p(j|i,\ a)$ changes on $\Phi_{ij}(a)$ and record $(f_j u_j(i))$.

Step **4** If $j = m$, the algorithm terminates. Otherwise, replace j by j+1 and return to Step 2.

We obtain $U^*(i) = \max_{j\in1,\cdots,m} u_j(i)$ by examining the value of every $u_j(i)$ recorded in step 3, and denote the corresponding robust optimal policy by f^*, where $f^* = f_g$, $g \in \arg\max_{j\in\{1,\cdots,m\}} u_j(i)$. Considering theorem 4.1, we draw the following conclusion.

Proposition 5.1 For a certain policy f_l, we get the expression of each $u(i)$, and then substitute the expressions into the $m-k$inequalities of the constraints in model (1) to obtain a value range of each $p(j|i,\ a)$, which we denote by$\phi_{ij}(a)$. If the intersection$\phi_{ij}(a) \cap \Phi_{ij}(a)$ is an empty set, f_l would have no chance to be an optimal policy.

Removing a group of transition probabilities $p'(j|i,\ a)$ from Φ that satisfy the constraints of model (6), we thus obtain the highest expected total reward $U(i)$ under these transition probabilities. We draw a conclusion about $U(i)$ and $U^*(i)$ as follows.

Theorem 5.1 $\square p(j|i,\ a) \in \Phi_{ij}(a),\ U(i) \quad \geq \quad U^*(i)$.

Proof For any initial state i, we denote by f^* a robust optimal policy. Based on the sup-inf criterion we have the following relationship: $\inf\limits_{p(j|i,a)\in\Phi_{ij}(a)} U_{f*}(i) \leq U_{f*}^{p(j|i,a)}(i) \leq U(i)$ where $U_{f*}^{p(j|i,a)}(i)$ represents the expected total reward gained by adopting policy f^* under a given group $p(j|i,\ a)$. According to the definition of f^* we know that $U^*(i) = \inf\limits_{p(j|i,a)\in\Phi_{ij}(a)} U_{f*}(i)$ So $U^*(i) \leq U(i)$ always holds.

Actually, if we denote by $\inf\limits_{p(j|i,a)\in\Phi_{ij}(a)} \sup\limits_{f\in\Pi^d} u(i)$ the highest expected total reward under the worst case scenario, according to theorem 5.1 we have the following conclusion:

$$\inf_{p(j|i,a)\in\Phi_{ij}(a)} \sup_{f\in\Pi^d} u(i) \geq \sup_{f\in\Pi^d} \inf_{p(j|i,a)\in\Phi_{ij}(a)} u(i)$$

Namely, no matter how unfavorable the condition is, it is better than knowing nothing. There exists a similar theorem in game theory.

Definition 5.1 Let $V_{\min}(i) = \inf\limits_{p(j|i,a)\in\Phi_{ij}(a)} (U(i) - U^*(i))$ be the lowest value of the exact transition probabilities for a discounted MDP with an initial state i under some uncertain transition probabilities, and $V_{\max}(i) = \inf\limits_{p(j|i,a)\in\Phi_{ij}(a)} (U(i) - U^*(i))$ the highest value of the exact transition probabilities.

In practice, it is always cost-effective to spend less than $V_{min}(i)$ to acquire the exact transition probabilities. Similarly, we have no need to obtain the exact transition probabilities when the cost is higher than $V_{max}(i)$.

5.2 Value interval of exact transition probabilities

In this subsection we focus on developing a technique to obtain $V_{min}(i)$ and$V_{max}(i)$. Actually, $U^*(i)$ is a certain value, so we only need to obtain $\inf\limits_{p(j|i,a)\in\Phi_{ij}(a)} U(i)$ and $\sup\limits_{p(j|i,a)\in\Phi_{ij}(a)} U(i)$.

In order to prove that the highest expected total reward$U(i)$ is a continuous function of $p(j|i,\ a)$, we first assume that $U(i)$ is bounded. Given a finite value M, for any initial state i we have $|U(i)| \quad \leq$ M. Apparently, the above assumption is reasonable. In fact, we can obtain the infimum of M by the following model:

$$\begin{aligned} &\sup_{f\in\prod^d} \sup_{p(j|i,a)\in\Phi_{ij}(a)} u(i) \\ &\textit{subject to } \sum_{j\in\Omega} p(j|i,a) = 1, a \in A(i), i \in \Omega \\ &\qquad 0 \leq p(j|i,a) \leq 1 \in A(i), ij \in \Omega. \end{aligned}$$

For a given policy $f_l \in \Pi^d=\{f_1, \cdots, f_m\}$, we obtain a unique expression for each $u(i)$ according to Theorem 4.2. Taking $u(1)$ for example, we derive the corresponding expression as follows:

$$u(1) = \frac{\begin{vmatrix} -r(1,a_1) & \beta p(2|1,a_1) & \cdots & \beta p(k|1,a_1) \\ -r(2,a_2) & (\beta p(2|2,a_2)-1) & \cdots & \beta p(k|2,a_2) \\ \vdots & \vdots & \vdots & \vdots \\ -r(k,a_k) & \beta p(2|k,a_k) & \cdots & (\beta p(k|k,a_k)-1) \end{vmatrix}}{\begin{vmatrix} (\beta p(1|1,a_1)-1) & \beta p(2|1,a_1) & \cdots & \beta p(k|1,a_1) \\ \beta p(1|2,a_2) & (\beta p(2|2,a_2)-1) & \cdots & \beta p(k|2,a_2) \\ \vdots & \vdots & \vdots & \vdots \\ \beta p(1|k,a_k) & \beta p(2|k,a_k) & \cdots & (\beta p(k|k,a_k)-1) \end{vmatrix}} \tag{7}$$

As mentioned above, we substitute k expressions for all $u(i)$ into the $m-k$inequalities. Thus we obtain inequalities for $p(j|i, a)$. As long as each $p(j|i, a)$ changes in $\Phi_{ij}(a)$ and guarantees that the inequalities hold, f_l is always an optimal policy.

Lemma 5.1 For an optimal policy f^*, if there exist a group of transition probabilities $p'(j|i, a)$ within the robust range that make at least one of the inequalities to be an equality, then there exists another simultaneous optimal policy.

Proof Without loss of generality, we denote by $f^*= (a_1,a_2, \cdots,a_k)$ a current optimal policy. For any $i \in \Omega$ we have

$$r(i,a_i) + \beta \sum_{j\in\Omega} p(j|i,a_i)u(j)=u(i), \tag{8}$$

Through the k equations we obtain the expression of each $u_p(i)$, where $u_p(i)$ is a rational function of $p\ (j|i, a)$ as in expression (7).

According to the given condition, we know that there exists at least one equality in the previous inequalities when each $p'(j|i, a)$ is substituted into the inequalities. Without loss of generality, we assume that the equality corresponds to state 1, i.e.,

$$r(1,a_1') + \beta \sum_{j\in\Omega} p(j|1,a_1')u_p(j)=u_p(1). \tag{9}$$

There is a similar relationship when the equality corresponds to any other state.

In the equality constraints of f^*, the equality corresponding to state 1 is

$$r(1,a_1) + \beta \sum_{j\in\Omega} p(j|1,a_1)u_p(j)=u_p(1). \tag{10}$$

Replacing equality (10) by equality (9), we obtain a new system of equalities. Apparently, $f'=(a_1',a_2, \cdots ,a_k)$ is also an optimal policy at this moment.

Corollary 5.1 For a given group of transition probabilities $p(j|i, a)$, if there exist two simultaneous optimal policies f_h andf_l, then, $\square i \in \Omega \quad U_{f_h}(i) = U_{f_l}(i)$.

Proof We solve the systems of linear equations corresponding to policy f_h and policyf_l. For any $i \in \Omega$, we obtain the unique expressions $u_{f_h}(i)$ and $u_{f_l}(i)$ based on theorem 4.2. As f_h and f_l are both optimal policies, their inequalities hold when the expressions for $p'(j|i, a)$ are substituted into the relevant equations.

On the basis of MDP theory we know that the highest expected total reward is unique for

a certain group of $p'(j|i,\ a)$. Hence, we have $U_{f_h}(i)$=$u_{f_h}(i)$=$u_{f_l}(i) = U_{f_l}(i)$.

Lemma 5.2 The optimal policy would shift at most once when each uncertain transition probability changes within a small enough interval.

Proof First we obtain the expressions for each $u_p(i)$ and then substitute these into the $m-k$ inequalities corresponding to a given optimal policy f^*. Next, we demonstrate the correctness of this lemma according to the results obtained when substituting a given group of $p(j|i,\ a)$ into the inequalities.

(i). If each inequality satisfies the following relationship:

$$r(1, a_1) + \beta \sum_{j\in\Omega} p(j|1, a)u_p(j)=u_p(1).$$

then we turn our attention to the sufficient condition for which each inequality still holds when one or more transition probabilities change:

$$|\beta \sum j \in \Omega p' u'_p(j) - u'_p(i) - \beta \sum j \in \Omega p u_p(j) + u_p(i)| < u_p(i) - r(i, a) - \beta \sum j \in \Omega p u_p(j) \quad (11)$$

In the above inequality, the expressions for$u_p(i)$ and $u'_p(i)$ are the same. Apparently, $u_p(i)$ is an elementary function of $p\ (j|i,\ a)$, so it is continuous in $\Phi_{ij}(a)$. We rewrite $p(j|i,\ a)$ as $p$for convenience. From the properties of continuous functions we know that there exists dp_i, and inequality (11) holds when $dp \leq dp_i$, $\square i,\ j \in \Omega,\ a \in \mathrm{A}_i$.

Let$dp^* = \min\{dp_i, i = 1, \cdots, m-k\}$; thus, the $m-k$ inequalities similar to (11) will hold when $|dp| \leq |dp^*|$. Consequently, the optimal policy remains unchanged when each dp varies within the interval [-$|dp^*|$, $|dp^*|$].

(ii). If at least one inequality turns into an equality after substituting a given group of $p(j|i,\ a)$ into the inequalities corresponding to the optimal policy f^*, through lemma 5.1 we know that there is at least one other optimal policy for which the inequalities hold as well. We now consider all of the inequalities corresponding to these optimal policies. Using a similar process as with (i), we aim to find such intervals of the transition probabilities for which optimal policy will shift no more than once.

a. We first consider the inequalities with less-than signs. Similar to (i), we can find the dp^s that guarantees that the inequalities with less-than signs hold when $|dp| \leq |dp^s|$.

b. Afterwards, we consider the equalities in the previous inequalities after substituting a given set of $p(j|i,\ a)$. Without loss of generality, we rewrite each of these in following form:

$$f(p) = r(i, a) + \beta \sum_{j\in\Omega} p(j|i, a)u_p(j) - u_p(i), \quad (12)$$

where $f(p)$ is a function of several variables of $p(j|i,\ a)$. In function (12), all $u_p(i)$ are rational functions whose denominators are the same and not zero. For function (12), we reduce the fractions to a common denominator and denote by $g(p)$ the numerator of the rewritten $f(p)$. Apparently, $g(p)$ is a polynomial function of $p(j|i,\ a)$. Hence, $f(p)$ is not equal to zero if and only if $g(p)$ is not equal to zero. We then take the partial derivative of $g(p)$.

If there exists a function $g(p)$ whose partial derivatives are all zero after substituting into a given group of $p(j|i,\ a)$, we can find a dp° such that at least one of the partial derivatives is not zero when $|dp| \in (0, |dp^\circ|]$. Apparently, such a dp° exists as long as $g(p)$ is not identically equal to a constant. In that case, for any $|dp| \in (0, |dp^\circ|]$, $g(p)$ cannot be zero. Otherwise, there

would be an extreme point for $g(p)$ within the deleted neighborhood, and all partial derivatives of $g(p)$ at this point would be zero.

For other $g(p)$, there always exists at least one partial derivative that is not zero. Similarly, we can find a dp^t such that all partial derivatives of each $g(p)$ preserve their signs when $|dp| \in (0, |dp^t|]$. Clearly, such a dp^t is attainable.

Given the above, we denote by $|dp^*|=\min\{|dp^s|, |dp^\circ|, |dp^t|\}$ the range of variation of each $p(j|i, a)$. When each $p(j|i, a)$ changes in the deleted neighborhood $[p-|dp^*|, p+|dp^*|]$, function (12) cannot be zero. Hence, the optimal policy would not shift when $|dp| \in (0, |dp^*|]$. In conclusion, the optimal policy will shift no more than once when $|dp| \in [0, |dp^*|]$.

Actually, lemma 5.2 is an extension of subsection 4.1, and $[p-|dp^*|, p+|dp^*|]$ can be viewed as a robust range.

Theorem 5.2 $U(i)$ is a continuous function of $p(j|i, a)$.

Proof Through lemma 5.1 and corollary 5.1 we have demonstrated the correctness of the above theorem from the perspective of algebra. Now we give a rigorous proof based on mathematical analysis. We concentrate our attention on whether a given optimal policy will shift when each uncertain transition probability changes on a small enough interval.

(i). If the optimal policy remains unchanged when each uncertain transition probability changes on a small enough interval, through theorem 4.1 we know that the equalities in model (1) would hold. Namely, for any one of these equalities we have

$$U(i) = r(i, a) + \beta \sum_{j\in\Omega} (p(j|i, a)U(j) \tag{13}$$

Give each $p(j|i, a)$ a change of no more than $|dp^*|$. Equality (13) becomes

$$U'(i) = r(i, a) + \beta \sum_{j\in\Omega} (p(j|i, a) + dp(j|i, a))U'(j), \tag{14}$$

where $U(i)$ and $U'(i)$ represent the respective highest expected total reward before and after each change in $p(j|i, a)$. Subtracting equality (13) from both sides of equality (14) we have

$$dU(i) = \beta \sum_{j\in\Omega} (pdU(j) + U(j)dp + dU(j)dp),$$

where $p$is a logogram of $p(j|i, a)$. Let $dp\to 0$, and notice that

$$\lim_{dp\to 0} |\sum_{j\in\Omega} dU(j)dp| = \lim_{dp\to 0} |\sum_{j\in\Omega} (U'(j) - U(j))dp| \le \lim_{dp\to 0} \sum_{j\in\Omega} |U'(j)| + |U(j)|)dp \le 2M \lim_{dp\to 0} \sum_{j\in\Omega} dp,$$

we obtain the following inequality

$$\lim_{dp\to 0} |\sum_{j\in\Omega} (U(j)dp + dU(j)dp)| \le 3M \lim_{dp\to 0} \sum_{j\in\Omega} dp = 0.$$

Therefore, we have

$$\lim_{dp\to 0} |dU(i)| = \lim_{dp\to 0} \beta |\sum_{j\in\Omega} (pdU(j))| \le \lim_{dp\to 0} \beta dU_{\max},$$

where $dU_{max}=\max\{|dU(1)|, \cdots, |dU(k)|\}$. We assume that $U(m)$ has a biggest change among all $U(i)$, then owing to the arbitrariness of $U(i)$ we have

$$\lim_{dp\to 0} dU_{\max} = \lim_{dp\to 0} |dU(m)| = \lim_{dp\to 0} \beta |\sum_{j\in\Omega} (pdU(j))| \le \lim_{dp\to 0} \beta dU_{\max},$$

According to the previous assumption we know that $0 < \beta < 1$. Hence, we obtain $\lim_{dp\to 0} dU_{\max} = 0$. We thus conclude that $\square i \in \Omega$, $\lim_{dp\to 0} dU(i) = 0$.

(ii). If the optimal policy shifts when each $p(j|i, a)$ changes on a small enough interval, one or more of the equalities (14) will no longer hold. We focus on the equality corresponding to $dU_{\max}$.

a. If the equality corresponding to dU_{max} holds when each $p(j|i, a)$ changes on a certain small enough interval, then through (i) we know that $\lim_{dp\to 0} dU(i) = 0$ for any $i \in \Omega$.

b. If the equality corresponding to dU_{max} no longer holds when $p(j|i, a)$ changes on a small enough interval, then we have the following set of equalities and inequalities:

$$\begin{cases} (a).U(m) = r(m, a_x) + \beta \sum_{j\in\Omega} p(j|m, a_x)U(j) \\ (b).U(m) \geq r(m, a_y) + \beta \sum_{j\in\Omega} p(j|m, a_y)U(j) \end{cases},$$

$$\begin{cases} (c).U'(m) \geq r(m, a_x) + \beta \sum_{j\in\Omega} p(j|m, a_x)U(j) \\ (d).U'(m) = r(m, a_y) + \beta \sum_{j\in\Omega} p(j|m, a_y)U(j) \end{cases},$$

If $dU(m) \leq 0$, we take (c) minus (a) and let $dp \to 0$; we then have

$$-\lim_{dp\to 0} dU_{\max} = \lim_{dp\to 0} |dU(m) \geq \lim_{dp\to 0} \beta \sum_{j\in\Omega} (p(j|i, a_x)dU(j)) \geq -\lim_{dp\to 0} \beta dU_{\max},$$

By examining the absolute values of both sides in inequality (15) we have $\lim_{dp\to 0} dU_{\max} \leq \lim_{dp\to 0} \beta dU_{\max}$.

If $dU(m) \geq 0$, we take (d) minus (b) and let $dp \to 0$, then we have

$$\lim_{dp\to 0} dU_{\max} = \lim_{dp\to 0} dU(m) \leq \lim_{dp\to 0} \beta \sum_{j\in\Omega} (p(j|i, a_y)dU(j)) \leq \beta dU_{\max}.$$

Hence, we have the result that $\lim_{dp\to 0} dU(i) = 0$ as well.

Apparently, the above proof makes sense for any $p(j|i, a) \in \Phi_{ij}(a)$. Through (i) and (ii) we conclude that $U(i)$ is a continuous function of $p(j|i, a)$.

Afterwards, we develop solution techniques for searching for the upper and lower bounds of the highest expected total reward$U(i)$. The mathematical programming model used to determine the lower bound of $U(i)$ is given as follows:

$$\begin{aligned} &\inf_{p(j|i,a)\in\Phi_{ij}(a)} \sup_{f\in\prod^d} u(i) \\ &subject\ to \sum_{j\in\Omega} p(j|i, a) = 1, a \in A(i), i \in \Omega \\ &\qquad 0 \leq p(j|i, a) \leq 1, a \in A(i), i, j \in \Omega. \end{aligned} \tag{15}$$

The model used to find the upper bound of $U(i)$ is given as follows

$$\begin{aligned} &\sup_{p(j|i,a)\in\Phi_{ij}(a)} \sup_{f\in\prod^d} u(i) \\ &subject\ to \sum_{j\in\Omega} p(j|i, a) = 1, a \in A(i), i \in \Omega \\ &\qquad 0 \leq p(j|i, a) \leq 1, a \in A(i), i, j \in \Omega. \end{aligned} \tag{16}$$

$u(i)$ in model (16) and model (17) is derived from the objective function of model (1).

We then propose a method to solve model (16). Through corollary 5.1 we know that the highest expected total reward is unique when several policies are simultaneously optimal. Hence, we only need to find the solution by considering every policy. As $U(i)$ is a continuous function of $p(j|i, a)$, we transform all possible open intervals into closed intervals and propose programming models to address this problem.

For each $f_l \in \Pi^d$=$\{f_1, \cdots, f_m\}$, we obtain k equalities and $m-k$ inequalities. Regard k equalities as a linear system of equations and solve the system to obtain the unique expression for each $u(i)$. Substituting all the derived expressions into the $m-k$ inequalities, we obtain the value range of $p(j|i, a)$, which we denote by$\phi_{ij}^l(a)$. Let $\phi_{ij}^l(a) \quad \cap \Phi_{ij}^l(a)$ be the new interval of $p(j|i, a)$. We denote by $U_h(i)$ the highest expected total reward when f_m is an optimal policy. We then search for the minimum in $\phi_{ij}^h(a) \cap \Phi_{ij}^h(a)$:

$$\begin{aligned} &\min U_h(i) \\ &\textit{subject to} \sum_{j\in\Omega} p(j|i,a) = 1, a \in A(i), i \in \Omega \\ &\qquad p(j|i,a) \in \Phi_{ij}(a) \cap \phi_{ij}^h(a), a \in A(i), i \in \Omega \\ &\qquad 0 \le p(j|i,a) \le 1, a \in A(i), i, j \in \Omega. \end{aligned} \tag{17}$$

Let $h$be $1,\cdots, m$ in sequence; we then obtain the m solutions using model (18). Apparently, the minimum of the m values is the solution of model (16).

The method for solving model (17) is similar. We need to solve $m$models as follows:

$$\begin{aligned} &\max U_h(i) \\ &\textit{subject to} \sum_{j\in\Omega} p(j|i,a) = 1, a \in A(i), i \in \Omega \\ &\qquad p(j|i,a) \in \Phi_{ij}(a) \cap \phi_{ij}^h(a), a \in A(i), i \in \Omega \\ &\qquad 0 \le p(j|i,a) \le 1, a \in A(i), i, j \in \Omega. \end{aligned} \tag{18}$$

Similarly, the maximum of the m values is the solution of model (17).

Given the above, the solution of model (16) is $\inf\limits_{p'(j|i,a)\in\Phi_{ij}(a)} U(i)$ and the solution of model (17) is $\sup\limits_{p'(j|i,a)\in\Phi_{ij}(a)} U(i)$.

5.3 A numerical example

In this subsection we provide a numerical example to verify the feasibility of the models and methods proposed above. We follow the conditions used in the above section. Given a state space Ω=$\{1,2\}$, the available action set of each state is assumed to be A(1)=A(2)=$\{a_1a_2\}$. The discount factor β is assumed to be 0.9. The values of the transition probabilities when we take action a_1 are accurate, and we assume that $p(1|1,a_1) = 0.7$, $p(2|1,a_1) = 0.3$, $p(1|2,a_1) = 0.1$ and $p(2|2,a_1) = 0.9$. The values of the transition probabilities when we take action a_2 are uncertain, and we denote by $p(1|1,a_2) = p_1$, $p(2|1,a_2)$ =1-p_1, $p(1|2,a_2) = p_2$ and $p(2|2,a_2)$ =1-p_2. We assume that $p_1 \in$[0,0.6] and $p_2 \in$[0.2,0.8]. The corresponding rewards are assumed to be $r(1,a_1) = 6$ $r(1,a_2) = 8$ $r(2a_1) = 1$ and $r(2a_2) = 3$

We consider initial state to be 1 in the following analysis. According to model (6) and the method given in subsection 5.1, we intend to find a robust optimal policy. Through the assumptions we know that Π^d=$\{f_1,\ f_2,\ f_3,\ f_4\}$, where f_1= $(a_1,\ a_1)^T$,f_2= $(a_1,\ a_2)^T$,f_3= $(a_2,\ a_1)^T$, and f_4= $(a_2,\ a_2)^T$.

If we choosef_1= $(a_1,\ a_1)^T$, the corresponding equations are as follows:

$$\begin{cases} 6{+}0.9(0.7U(1)+0.3U(2))=U(1) \\ 1+0.9(0.1U(1)+0.9U(2))=U(2). \end{cases}$$

Solving the system, we obtain $U(1)$=30.65. In this case, $U(1)$ has nothing to do with uncertain transition probabilities, thus $\inf U(1)=U(1)=30.65$.

If we adopt f_2= $(a_1,\ a_2)^T$, the corresponding equations are

$$\begin{cases} 6{+}0.9(0.7U(1)+0.3U(2))=U(1) \\ 3+0.9(p_2U(1)+(1-p_2)U(2))=U(2) \\ 0.2\le p_2\le 0.8. \end{cases}$$

By solving the equations, we obtain $U(1)=\frac{5.4p_2+1.41}{0.09p_2+0.037}$. As $U'_{p_2}(1)>0$, we obtain the infimum of $U(1)$ when p_2=0.2. Thus, we have $\inf U(1)=45.27$.

If we choose f_3= $(a_2,\ a_1)^T$, the corresponding equations are

$$\begin{cases} 8{+}0.9(p_1U(1)+(1-p_1)U(2))=U(1) \\ 1+0.9(0.1U(1)+0.9U(2))=U(2) \\ 0\le p_1\le 0.6. \end{cases}$$

Solve the equations, we obtain $U(1)=\frac{2.42-0.9p_1}{0.109-0.09p_1}$. Because $U'_{p_1}(1)>0$, we obtain the infimum of $U(1)$ when p_1=0. We have $\inf U(1)=22.20$ under f_3.

If we adopt f_4= $(a_2,\ a_2)^T$, the corresponding equations are

$$\begin{cases} 8{+}0.9(p_1U(1)+(1-p_1)U(2))=U(1) \\ 3+0.9(p_2U(1)+(1-p_2)U(2))=U(2) \\ 0\le p_1\le 0.6, 0.2\le p_2\le 0.8. \end{cases}$$

By solving the above equations we obtain $U(1)=\frac{-2.7p_1+7.2p_2+3.5}{-0.09p_1+0.09p_2+0.1}$. On the given intervals $p_1\in$[0,0.6] and$p_2\in$[0.2,0.8] we have $U'_{p_1}(1)>0$ and $U'_{p_2}(1)>0$. Hence, we obtain the infimum of $U(1)$ when p_1=0 and p_2=0.2. Thus, we have $\inf U(1)=41.86$ under f_4.

Given the above, $\sup\limits_{f_l\in\prod^d}\inf\limits_{p(j|i,a)\in\Phi_{ij}(a)} U(1)=45.27$, and the unique robust optimal policy is f_2.

We next concentrate on determining the value interval of exact transition probabilities.

Let f_1= $(a_1,\ a_1)^T$ be an optimal policy, then we have the following constraints:

$$\begin{cases} 6{+}0.9(0.7u(1)+0.3u(2))=u(1) \\ 1+0.9(0.1u(1)+0.9u(2))=u(2) \\ 8{+}0.9(p_1u(1)+(1-p_1)u(2))\le u(1) \\ 3+0.9(p_2u(1)+(1-p_2)u(2))\le u(2) \\ 0\le p_1\le 1, 0\le p_2\le 1. \end{cases}$$

By solving the equations we obtain $u(1)$ =30.65, $u(2)$ = 19.78. After substituting these values into the inequalities, we discover that there is no solution, i.e., $\phi_{2,1}(a_2)=\emptyset$ Through

proposition 5.1 we know that f_1 cannot be an optimal policy.

Let f_2= $(a_1, a_2)^T$ be an optimal policy. We then have the following constraints:

$$\begin{cases} 6+0.9(0.7u(1)+0.3u(2)) = u(1) \\ 3+0.9(p_2u(1)+(1-p_2)u(2)) = u(2) \\ 8+0.9(p_1u(1)+(1-p_1)u(2)) \le u(1) \\ 1+0.9(0.1u(1)+0.9u(2)) \le u(2) \\ 0 \le p_1 \le 1, 0 \le p_2 \le 1. \end{cases}$$

Similarly, we first obtain the expressions for $u(1)$ and $u(2)$ as in expression (7) and then substitute these into the inequalities. Solve the two inequalities, we have $0.27p_1$+$0.18p_2$ ≤0.115 and p_2 ≥0.636. Clearly, f_2 may be an optimal policy. For $U(1) = \frac{5.4p_2+1.41}{0.09p_2+0.037}$ and $U_{p_2}^{'}(1) > 0$, we obtain the minimum of $U(1)$ whenp_2=0.636, i.e., min$U(1)$=51.40. Similarly, we obtain the maximum of $U(1)$ whenp_2=0.639, and max$U(1)$=51.43.

Likewise, let $f_3 = (a_2, a_1)^T$ be an optimal policy. We obtain p_1 ≥0.496 and $0.18p_1$-$0.63p_2$ ≥ 0.155. We find that there is no solution for p_1, i.e., $\phi_{1,1}(a_2) \quad \cap \quad \Phi_{1,1}\ (a_2) = \emptyset$ Therefore, f_3 has no chance to be an optimal policy.

Let $f_4 = (a_2, a_2)^T$ be an optimal policy. We obtain $0.27p_1$+$0.18p_2$ ≥0.115, $0.18p_1$-$0.63p_2$ ≤ 0.155. There exist a feasible region for f_4 to be an optimal policy. For the highest expected total reward $U(1) = \frac{-2.7p_1+7.2p_2+3.5}{-0.09p_1+0.09p_2+0.1}$, we take its partial derivatives and obtain $U_{p_1}^{'}(1) > 0$ and $U_{p_2}^{'}(1) > 0$ when p_1 ∈[0,0.6] andp_2 ∈[0.2,0.8]. Hence, the minimum of $U(1)$ would only be obtained on the line $0.27p_1$+$0.18p_2$=0.115. Eliminate p_1 and we have

$$U(1) = \frac{27p_2+7.05}{0.45p_2+0.185}, U'(1) = \frac{1.8225}{(0.45p_2+0.185)^2} > 0.$$

Apparently, we obtain the minimum of $U(1)$ whenp_2=0.2, and min$U(1)$=45.27. We obtain the maximum of $U(1)$ when p_1=1 andp_2=1, and max$U(1)$=80.

Given the above, we obtain $\inf\limits_{p'(j|i,a)\in\Phi_{ij}(a)} U(i) = 45.27$ and $\sup\limits_{p'(j|i,a)\in\Phi_{ij}(a)} U(i) = 80$. Further, we have $V_{min}(i)$=0, $V_{max}(i)$=34.73, and the value interval [0, 34.73].

Actually, when $0.27p_1$+$0.18p_2$=0.115 and p_2 ∈[143/225, 23/36], f_2= $(a_1, a_2)^T$ and f_4= $(a_2, a_2)^T$ are both optimal policies, and the highest expected total rewards under the two policies are equal:

$$U_{f_2}(1) = \frac{5.4p_2+1.41}{0.09p_2+0.037} = \frac{27p_2+7.05}{0.45p_2+0.185} = U_{f_4}(1),$$

which coincides with the conclusions of lemma 5.1 and corollary 5.1.

The above result in this example illustrates that obtaining exact transition probabilities may not improve the highest expected total reward under the most unfavorable situation. It is important to provide the value intervals for both pessimistic and optimistic decision makers.

§6 Conclusion

In this paper we addressed the problem of discounted MDPs with uncertain transition probabilities. Our research has made several significant contributions to MDP theory. We summarize these as follows:

(1) A method for estimating uncertain transition probabilities was proposed. By using an observation history, we derived a programming model to estimate the uncertain probabilities and provided a numerical example.

(2) The robustness of discounted MDPs was studied. We considered the robustness of an optimal policy and the sensitivity of the highest expected total reward and provided methods of analysis.

(3) The value intervals of exact transition probabilities were given. We first proved that the highest expected total reward obtained by a robust optimal policy cannot be greater than the reward obtained under accurate transition probabilities. Afterwards, we proposed a method to find a robust optimal policy. Finally, we presented several models for determining the value intervals of the exact transition probabilities.

(4) During the process of determining the value intervals, we considered the shift of optimal policies and the continuity of the highest expected total reward, and we drew several significant conclusions.

Apparently, the methods proposed in this paper are appropriate for other Markov decision models with finite actions and states. Nevertheless, the proposed solution techniques will require large computations when the number of states or actions increases. We also did not consider the situations in which the number of states or actions is infinite.

In practice, the distributionally robust framework is also adopted to deal with Markov decision issues with uncertain parameters when the uncertain parameters are random variables following an unknown distribution (Yu et al. 2016). Besides, the change of parameter plays a critical rule for the robustness of stochastic systems (see, e.g., Zhu (2018) and Zhu (2019)). When involving continuous-time problems, semi-Markov models are always applied (Wang et al. 2018). In the future work, we are about to consider the continuous-time Markov models with uncertain parameters.

References

[1] H S Baik, H S Jeong, D M Abraham. *Estimating transition probabilities in Markov chain-based deterioration models for management of wastewater systems*, Journal of Water Resources Planning and Management, 2006, 132(1): 15-24.

[2] E Delage, S Mannor. *Percentile optimization for Markov decision processes with parameter uncertainty*, Operations Research, 2009, 58(1): 203-213.

[3] N I Garud. *Robust dynamic programming*, Mathematics of Operations Research, 2005, 30(2): 257-280.

[4] S Kalyanasundaram, E K P Chong, N B Shroff. *Markov decision processes with uncertain transition rates: sensitivity and max hyphen min control*, Asian Journal of Control, 2004, 6(2): 253-269

[5] B H Li, J Si. *Robust dynamic programming for discounted infinite-horizon Markov decision processes with uncertain stationary transition matrice*, Proceedings of the 2007 IEEE Symposium on Approximate, 2007, 96-102.

[6] A Nilim, L E Ghaoui. *Robust control of Markov decision processes with uncertain transition matrices*, Operations Research, 2005, 53(5): 780-798.

[7] M L Puterman. *Markov decision processes: discrete stochastic dynamic programming*, John Wiley & Sons, New Jersey, 2014.

[8] W A S Reis, L N Barros, K V Delgado. *Robust topological policy iteration for infinite horizon bounded Markov Decision Processes*, International Journal of Approximate Reasoning, 2019, 105: 287-304.

[9] J K Satia, R E Lave. *Markovian decision processes with uncertain transition probabilities*, Operations Research, 1973, 21(3): 728-740.

[10] B Wang, Q X Zhu. *Stability analysis of semi-Markov switched stochastic systems*, Automatic, 2018, 94: 72-80.

[11] C C White, H K Eldeib. *Markov decision processes with imprecise transition probabilities*, Operations Research, 1994, 42(4): 739-749.

[12] W Wiesemann, D Kuhn, B Rustem. *Robust Markov decision processes*, Mathematics of Operations Research, 2013, 38(1): 153-183.

[13] H Xu, S Mannor. *Distributionally robust Markov decision processes*, Mathematics of Operations Research, 2012, 37(2): 288-300.

[14] P Q Yu, H Xu. *Distributionally robust counterpart in Markov decision processes*, IEEE Transactions on Automatic Control, 2016, 61(9): 2538-2543.

[15] Q X Zhu. *Stability analysis of stochastic delay differential equations with Lévy noise* Systems & Control Letters, 2018, 118: 62-68.

[16] Q X Zhu. *Stabilization of stochastic nonlinear delay systems with exogenous disturbances and the event-triggered feedback control*, IEEE Transactions on Automatic Control, 2019, 64(9): 3764-3771.

[1]School of Management and Economics, Beijing Institute of Technology, Beijing 100081, China.
[2]School of Economics and Management, Xi'an University of Posts and Telecommunications, Xi'an 710121, China.
Email: louzk@bit.edu.cn