NUMERICAL ENCLOSURE FOR THE OPTIMAL THRESHOLD PROBABILITY IN DISCOUNTED MARKOV DECISION PROCESSES

Toyonaga, Kenji
Graduate School of Mathematics, Kyushu University

Nakao, T. Mitsuhiro
Graduate School of Mathematics, Kyushu University

https://doi.org/10.5109/13494

出版情報：Bulletin of informatics and cybernetics. 32 (1), pp.81-90, 2000-06. Research Association of Statistical Sciences
バージョン：
権利関係：

Bulletin of Informatics and Cybernetics, Vol. 32, No. 1, 2000

NUMERICAL ENCLOSURE FOR THE OPTIMAL THRESHOLD PROBABILITY IN DISCOUNTED MARKOV DECISION PROCESSES

By

Kenji TOYONAGA* and Mitshiro T. NAKAO†

Abstract

There are various procedures to compute the optimal threshold probability in discounted Markov decision processes. In the actual numerical computation of an approximate optimal solution, the estimation of the discrepancy between the approximate solution and the exact solution is important. White(1993 b) derived such an error estimation for the value iteration method, however, this estimation is not actually in the computable form. In this paper, we present a numerical enclosure method to compute the optimal threshold probability, that guarantees a rigorous *a posteriori* error bound.

1. Introduction

In the study of Markov decision processes, there are several literatures related to problems for threshold probability, e.g., White(1993 a,b), Bouakiz and Kebir(1995), Sobel(1982), Van der Wal(1981), Wu and Lin(1999) etc.

The value iteration method is a method to calculate an optimal threshold probability. However, in actual applications, using this method requires a great deal of computing time, except for very simple models. For realistic models, it is not practical to use a large number of iterations to obtain an accurate approximation. For this reason, when applying this method, error estimation plays an important role in the actual calculation of the optimal threshold probability. White(1993 b) obtained an error estimation for the value iteration method. However, his estimaion method is not well suited to numerical evaluation.

In this paper, we attemped to obtain an error estimation for the value iteration method using numerical enclosure method, in which the exact solution is enclosed between two approximate solutions. With this method, we find that we can determine numerical error bounds even for complicated models. Next, we show a relation between our error estimation and that obtained White(1993 b). Then, we construct an actual

* Graduate School of Mathematics, Kyushu University, Fukuoka 812-8581, Japan
E-mail: toyonaga@math.kyushu-u.ac.jp

† Graduate School of Mathematics, Kyushu University,Fukuoka 812-8581, Japan
E-mail: mtnakao@math.kyushu-u.ac.jp

algorithm to numerically compute an enclosure of the optimal value function. Finally, we present a numerical example.

2. Notation and Formulation of Problem

The discounted Markov decision process with a discrete time space $N = \{1, 2, \cdots\}$ is defined as follows. The state space S is a finite set, and we denote the state at time $t \in N$ by X_t. The action space A is also a finite set, and $A(s)$ is an admissible action set, for each state $s \in S$, such that $\phi \neq A(s) \subset A$. We denote the action at time $t \in N$ by A_t and the random immediate reward function at time $t \in N$ by Y_t is defined by the following conditional probability for (X_{t+1}, Y_t) when (X_t, A_t) is given:

$$p^a(s', y|s) = P(X_{t+1} = s', Y_t \leq y | X_t = s, A_t = a).$$

Here, we have assumed that $0 \leq Y_t \leq H$, where H is some positive constant. We denote the discount factor by ρ, where $0 < \rho < 1$. Then, we define the new state space by $S \times R$, where $R = (-\infty, \infty)$. Let $H_1 := S \times R$, and let $H_{t+1} := H_t \times A \times S \times Y_t$, where $t \in N$. Thus H_t represents the set of all possible histories of the system up to time t. We denote such history by θ_t.

A decision rule δ_t for time $t \in N$ is the conditional probability $\delta_t(a_t|h_t) = P(A_t = a_t|\theta_t = h_t)$ for a given θ_t, where $h_t = (s_1, r, a_1, s_2, y_1, \ldots, a_{t-1}, s_t, y_{t-1}) \in H_t$ is a realization of $\theta_t = (X_1, r, A_1, X_2, Y_1, \cdots, A_{t-1}, X_t, Y_{t-1})$, and r is a given real number. It is assumed that $\delta_t(A_t \in A(s_t)|h_t) = 1$ for every history $h_t = (s_1, r, a_1, \cdots, s_t, y_{t-1}) \in H_t$, and that $\delta_t(a_t|\cdot)$ is a Lebesgue-Stieltjes measurable function on H_t. We denote by Δ the set of all decision rules. A policy π is an infinite sequence of decision rules $(\delta_1, \delta_2, \cdots, \delta_t, \cdots)$. We denote by C the set of all policies.

The random total discounted reward Z_n for a finite horizon is given by

$$Z_n = \sum_{t=1}^{n} \rho^{t-1} Y_t \qquad (n \geq 1),$$

and that for an infinite horizon case is given by

$$Z = \sum_{t=1}^{\infty} \rho^{t-1} Y_t.$$

Further, let w_t be a variable such that

$$W_t = (W_1 - Z_{t-1})/\rho^{t-1}, \qquad (t \geq 1).$$

Then, we define the new histories $h_t = (s_1, w_1, a_1, s_2, w_2, \cdots, a_{t-1}, s_t, w_t)$, which are realizations of $(X_1, W_1, A_1, X_2, \cdots, A_{t-1}, X_t, W_t)$. The decision rule and the policy are similarly defined for these new histories. If δ_t is a function of (s_t, w_t) for $X_t = s_t, W_t = w_t$, where s_t and w_t are realizations of X_t and Y_t, then δ_t is called a Markov decision rule. We denote the set of all Markov decision rules by Δ_M. If there exists $a_t \in A_t$ such that $\delta_t(A_t = a_t|s_t, w_t) = 1$ for $\delta_t \in \Delta_M$, then δ_t is called a deterministic decision rule. The set of such decision rules is denoted by Δ_D.

If all decision rules δ_t of a policy $\pi = (\delta_1, \delta_2, \cdots, \delta_t, \cdots)$ are Markov decision rules, the policy π is called a Markov policy. We denote the set of all Markov policies by C_M. If all decision rules δ_t of a policy π are a deterministic and $\delta_t = \delta_{t+1}$, then the policy π is called a deterministic Markov stationary policy. We denote the set of all such policies by C_D.

We denote the conditional probability of an event $Z \leq r$, for a given initial state $X_1 = s$ and policy π, by $P(Z^\pi \leq r|s)$, where the random variable Z depends on not only s and π but also on r. We define the criterion functions for the finite and infinite horizon cases by $F_n^{\pi(n)}(s,r) = P^\pi(Z_n \leq r|s)$ and $F^\pi(s,r) = P^\pi(Z \leq r|s)$, respectively, for all $(s,r) \in S \times R$ and $\pi \in C_M$. We also define the optimal value functions F_n^* and F^* for the finite and infinite horizon cases by

$$F_n^*(s,r) = \inf_{\pi(n) \in C_M} F_n^{\pi(n)}(s,r) \quad \text{and} \quad F^*(s,r) = \inf_{\pi \in C_M} F^\pi(s,r). \tag{2.1}$$

Then, let $\mathcal{F}$ be the set of functions F from $S \times R$ into $[0,1]$ such that for each $s \in S$, $F(s,r)$ is measureable on R,

$$F(s,r) = \begin{cases} 0 & (r < 0) \\ 1 & (r \geq H/1 - \rho) \end{cases},$$

and $F(s,r)$ is monotone nondecreasing and right continuous on R. In Wu and Lin(1999), it is shown that F_n^* and F^* are in $\mathcal{F}$.

We define the operators T^a, T^δ and T from $\mathcal{F}$ into itself as follows. For $F \in \mathcal{F}$, $a \in A(s)$ and $\delta \in \Delta$,

$$\begin{aligned} T^a F(s,r) &\equiv \int_{S \times R} F(s', (r-y)/\rho) dp^a(s', y \mid s), \\ T^\delta F(s,r) &\equiv \sum_{a \in A(s)} T^a F(s,r) \delta(a \mid (s,r)), \\ TF(s,r) &\equiv \inf_{\delta \in \Delta} T^\delta F(s,r) = \min_{a \in A(s)} T^a F(s,r). \end{aligned}$$

We denote the operator obtained from n repetitions of T by $(T)^n$. For $F, G \in \mathcal{F}$, if $F(s,r) \leq G(s,r)$ for all $(s,r) \in S \times R$, then we write $F \leq G$.

3. Numerical Enclosure for the Optimal Threshold Probability

LEMMA 3.1. *Let $F, G \in \mathcal{F}$. Then, if $F \leq G$ we have $TF \leq TG$.*

This lemma is an immediate consequence of the definition of T.

The value iteration method is a method to determine the optimal threshold probability F^* defined by (2.1). In this section we present a numerical enclosure method for the optimal threshold probability using approximate solutions obtained from the value iteration method.

The value iteration method consists of the following.

VALUE ITERATION METHOD

(1) Choose an initial value $L_0 \in \mathcal{F}$.
(2) Compute $L_n := TL_{n-1} \equiv (T)^n L_0$.

There exists the following result regarding the pointwise convergence of L_n to F^* for this iterative method.

THEOREM 3.2. *In the value iteration procedure, if the iteration starts with $L_0 := F_0$, where $F_0(s,r) = 0(r < 0)$ and $F_0(s,r) = 1(r \geq 0)$, then $L_n = F_n^*$, and L_n converges pointwise to F^*. Furthermore, we have $TF^* = F^*$.*

PROOF. See White(1993 b) [Theorem 2].

THEOREM 3.3. *If we start the value iteration procedure with $L_0 := G_0$, where*

$$G_0(s,r) = \begin{cases} 0 & (r < \frac{H}{1-\rho}) \\ 1 & (r \geq \frac{H}{1-\rho}) \end{cases},$$

then L_n pointwise converges to G^*, where G^* is the left continuous transformation of F^* defined by $G^*(s,r) \equiv F^*(s, r-0)$.

PROOF. First, by induction we show that for any $n \in N$,

$$(T)^n G_0(s,r) = F_n^*(s, r - \frac{\rho^n H}{1-\rho}). \tag{3.1}$$

For $n = 1$, we have

$$\begin{aligned} T[G_0(s,r)] &= \inf_{k \in K(s)} \int G_0(s', \frac{r-y}{\rho}) dp^k(s', y \mid s) \\ &= \inf_{k \in K(s)} \int_{S \times (-\infty, r - \frac{H}{1-\rho}]} dp^k(s', y \mid s) \\ &= \inf_{\pi \in C(M)} F_1^{\pi(1)}(s, r - \frac{\rho H}{1-\rho}) = F_1^*(s, \frac{\rho H}{1-\rho}). \end{aligned}$$

Hence (3.1) is true for $n = 1$.
Next, assume that (3.1) holds up to n. Then

$$\begin{aligned} F_{n+1}^* & (s, r - \frac{\rho^{n+1} H}{1-\rho}) \\ = & \inf_{\pi(n+1)} P(Z_{n+1}^\pi \leq r - \frac{\rho^{n+1} H}{1-\rho} \mid s) \\ = & \inf_{k \in K(s), \pi(n)} \int_I P(\sum_{t=2}^{n+1} \rho^{t-2} Y_t^\pi \leq \frac{r-y}{\rho} - \frac{\rho^n H}{1-\rho} \mid X_2 = s') dp^k(s', y \mid s) \\ = & TF_n^*(s, r - \frac{\rho^n H}{1-\rho}), \end{aligned}$$

which implies that (3.1) holds for any $n \in N$.

Then, since from Theorem 3.2 we know that $F_n^*(s,r)$ converges pointwise to $F^*(s,r)$, it follows that $(T)^n G_0(s,r)$ converges to $G^*(s,r)$.

Generally, as the number of states or actions increases, the time required to numerically compute L_n also increases. In fact, in some cases we found that it took dozens of hours to calculate only up to n=10, even when the number of actions and states was not large.

White(1993 b) obtained the error estimation given in Theorem 3 for the value iteration procedure using the norm

$$\|F\| \equiv \sup_{(s,r)\in S\times R} \mid F(s,r) \mid \ .$$

THEOREM 3.4. *For any $l, m \in N$ and $n := lm$,defining*

$$\lambda(m) = \sup_{(s,r)\in I, \pi\in C_D(m)} \mid F_m^{\pi(m)}(s,r) - F_m^{\pi(m)}(s, r - \frac{\rho^m H}{1-\rho}) \mid ,$$

where $F_m^{\pi(m)}(s,r)$ is the threshold probability for $\pi(m)$ using policy π,we have

$$\|L_n - F^*\| \leq \{\lambda(m)\}^l .$$

PROOF. See White(1993 b) [Theorem 5].

Theorem 3.4 is significant as an *a priori* error estimation. However, because the decision rules of a policy $\pi(m) \in C_D(m)$ depend on (s,r), it is difficult to calculate $\lambda(m)$ in general using this theorem. Hence we need a more practical means of error estimation. In the following theorem, we compute two approximate solutions which enclose the optimal value function F^*. This enables us to estimate an *a posteriori* error bound. In what follows we also derive relation between our estimation and the above estimation involving $\lambda(m)$.

THEOREM 3.5. *Let $L_n := (T)^n F_0$ and $L'_n := (T)^n G_0$. Then*

$$L'_n \leq F^* \leq L_n$$

and

$$\|L_n - F^*\| \leq \|L_n - L'_n\| \leq 2\lambda(m)^l ,$$

where

$$F_0(s,r) = \begin{cases} 0 & (r < 0) \\ 1 & (r \geq 0) \end{cases} \quad \textit{and} \quad G_0(s,r) = \begin{cases} 0 & (r < \frac{H}{1-\rho}) \\ 1 & (r \geq \frac{H}{1-\rho}) \end{cases}$$

for each $s \in S$.

PROOF. First, by definition, we have

$$G_0 \leq F^* \leq F_0.$$

Therefore, from Lemma 3.1, it holds that

$$(T)^n G_0 \leq (T)^n F^* \leq (T)^n F_0.$$

Hence, by Theorem 3.2,

$$L'_n \leq F^* \leq L_n.$$

Thus, we have

$$\|L_n - F^*\| \leq \|L_n - L'_n\|.$$

Then, using an argument similar to that in the proof of Theorem 3.4, it follows that $\|F^* - L'_n\| \leq \{\lambda(m)\}^l$. By using the triangle inequality, we obtain

$$\|L_n - L'_n\| \leq \|L_n - F^*\| + \|F^* - L'_n\| \leq 2\{\lambda(m)\}^l.$$

4. Algorithm

In this section we present a numerical algorithm for computer to enclose the optimal value function for the infinite horizon case.

From Theorem 3.5 it is insured that there exists the optimal value function F^* between L_n and L'_n.

In the value iteration procedure, since we choose $L_0 := F_0$, L_1 is a step function. The following algorithm is based on determining the positions of the discontinuous points of L_n and L'_n.

ALGORITHM

1. Setting the model system:
   Let I and K be the number of states and actions, respectively. Then, set the transition probability p^k_{ij}, reward w^k_{ij} $(1 \leq i, j \leq I,\ 1 \leq k \leq K)$, and discount factor ρ, where $0 < \rho < 1$.

2. Computation of L_n :

   (1) Set initial function $L_0(s_i, r) := F_0(s_i, r)$ for each i $(1 \leq i \leq I)$.
   We set the discontinuous point (x, y) of the function $F_0(s_i, r)$ with respect to r, as $(x, y) := (0, 1)$.

   (2) Computation of $TL_{n-1}(s_i, r)$ for each $n \geq 1$ and i $(1 \leq i \leq I)$:

   (a)Each value r at which the function $T^{a_k} L_{n-1}(s_i, r)$ is discontinuous is determined by using the information regarding $L_{n-1}(s_i, r)$ and the relation $T^{a_k} L_{n-1}(s_i, r) = \sum_{j=1}^{I} p^k_{ij} L_{n-1}(s_j, (r - w^k_{ij})/\rho)$.

(b)We compute and store the minimum value $y := \min_{1 \leq k \leq K} T^{a_k} L_{n-1}\,(s_i, r)$ for each discontinuous point r.

Namely, the pair $(r, L_n(s_i, r))$ is computed at each discontinuous point r of L_n.

3. Setting $L_0'(s_i, r) := G_0(s_i, r)$ for each i $(1 \leq i \leq I)$:
   We set the discontinuous point (x, y) of the graph for the function $G_0(s_i, r)$ with respect to r, as $(x, y) = (H/(1-\rho), 1)$.
4. Computation of L_n':
   F is replaced by G in step 2 and procedures similar to (2.1) and (3.1) are carried out.

From Theorem 3.5, it is insured that the optimal value function exists between L_n and L_n' computed using the above algorithm. Thus, the actual errors are bounded by $\|L_n - L_n'\|$.

5. A Numerical Example

For this example we computed an optimal value function verified numerically using the value iteration algorithm presented in the previous section.

In studies reported up to this time, only very simple models, e.g. two states and two actions have been considered. We treat a more complicated model here.

As the numbers of states and actions become larger, it becomes difficult to carry out many iterations. In fact, even if the numbers of states and actions are both fewer than ten, it often takes on the order of a dozen hours to iterate ten times. For this reason, *a posteriori* error estimation and verified numerical computation become important. In this example, we adopted a model which has three states and three actions. This model may appear very simple, but even in such a case, it is not feasible to compute the optimal value function by hand with sufficient accuracy. If we use the algorithm developed in this paper, it is also possible to enclose the optimal value function for more complicated models.

EXAMPLE

Let $S = \{s_1, s_2, s_3\}$ be the state space, and let $A \equiv A(s_i) = \{a_1, a_2, a_3\}$ $(i = 1, 2, 3)$ be the action space. We choose the discount factor ρ=0.05. Then we assume that stochastic behavior is given as follows:

$$p_{11}^1 = 0.5, p_{12}^1 = 0.2, p_{13}^1 = 0.3$$
$$p_{21}^1 = 0.4, p_{22}^1 = 0.1, p_{23}^1 = 0.5$$
$$p_{31}^1 = 0.2, p_{32}^1 = 0.3, p_{33}^1 = 0.5$$
$$p_{11}^2 = 0.5, p_{12}^2 = 0.25, p_{13}^2 = 0.25$$
$$p_{21}^2 = 0.6, p_{22}^2 = 0.2, p_{23}^2 = 0.2$$
$$p_{31}^2 = 0.5, p_{32}^2 = 0.3, p_{33}^2 = 0.2$$
$$p_{11}^3 = 0.2, p_{12}^3 = 0.3, p_{13}^3 = 0.5$$

$$p^3_{21} = 0.3, p^3_{22} = 0.5, p^3_{23} = 0.2$$
$$p^3_{31} = 0.2, p^3_{32} = 0.5, p^3_{33} = 0.3$$

$$w^1_{11} = 0, w^1_{12} = 10, w^1_{13} = 20$$
$$w^1_{21} = 40, w^1_{22} = 5, w^1_{23} = 20$$
$$w^1_{31} = 10, w^1_{32} = 5, w^1_{33} = 2$$
$$w^2_{11} = 0, w^2_{12} = 10, w^2_{13} = 30$$
$$w^2_{21} = 20, w^2_{22} = 5, w^2_{23} = 10$$
$$w^2_{31} = 5, w^2_{32} = 5, w^2_{33} = 10$$
$$w^3_{11} = 5, w^3_{12} = 10, w^3_{13} = 15$$
$$w^3_{21} = 10, w^3_{22} = 0, w^3_{23} = 3$$
$$w^3_{31} = 5, w^3_{32} = 10, w^3_{33} = 2.$$

We set $L_0 := F_0$, where $F_0(s,r) = 0$ for $r < 0$ and $F_0(s,r) = 1$ for $r \geq 0$. Using the algorithm described above, we computed L_8 and L'_8. We found to be there 10857 discontinuous points with respect to r for both $L_8(s_1,r)$ and $L'_8(s_1,r)$. Theorem 3.5 guarantees that the optimal value function which we seek exists between L_8 and L'_8. For example, the actual error for $L_8(s_1,r)$ $r \in [0,\infty)$ is bounded as follows:

$$\mid L_8(s_1,r) - F^*(s_1,r) \mid\leq\mid L_8(s_1,r) - L'_8(s_1,r) \mid\leq 3 \times 10^{-3}.$$

The approximate solution $L_8(s_1,r)$ we computed is shown in Figure 1, and a partial enlargements of $L_8(s_1,r)$ and $L'_8(s_1,r)$ corresponding to part A in Figure 1 are shown in Figure 2. The solid line represents $L_8(s_1,r)$, and the dotted represents $L'_8(s_1,r)$. $F^*(s_1,r)$ exists between $L_8(s_1,r)$ and $L'_8(s_1,r)$.

In the computations we used double-precision floating point arithmetic, and thus rounding error was introduced. If we wish to avoid the influence of such errors, we need make use of interval arithmetic software(e.g. PROFIL(Knüppel,1994)).

6. Conclusions

(1)In this note, we presented a numerical method of enclosing the optimal threshold probability F^*. We proved the validity of this method by showing that two kinds of approximations asymptotically converge to F^*, one from the upper side and one from the lower side.

(2)We confirmed that our enclosing algorithm actually works with sufficient accuracy for a particular model.

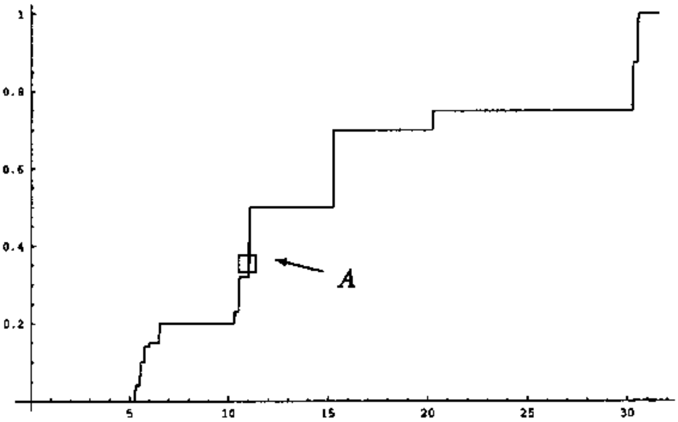

Figure 1: A approximate solution $L_8(s_1, r)$

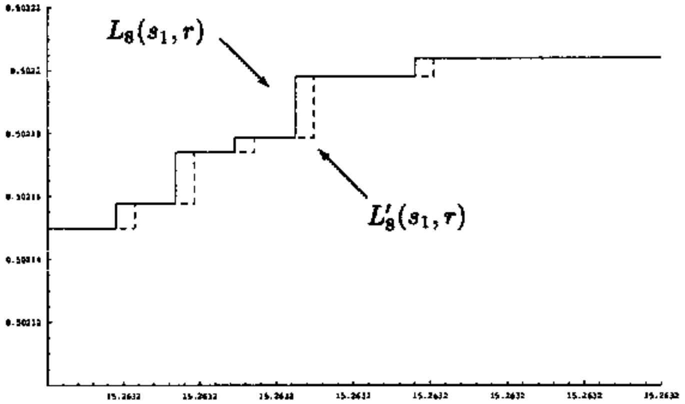

Figure 2: Enclosure of the optimal value function $F^*(s_1, r)$ corresponding to part A in Figure 1.($L'_8 \leq F^* \leq L_8$)

References

Bouakiz, M and Kebir, Y. (1995),Target-level criterion in Markov decision processes, *J. Opt. Th. Appl.* **86**, 1-15.

Knüppel, O. (1994), PROFIL/BIAS - A fast interval liblary, Computing **53**, 277-288.

Sobel, M. (1982), The variance of discounted Markov decision processes, *J. Appl. Probab,* **19**, 794-802.

Van der Wal, J. (1981), Stochastic Dynamic Programming,*Mathematical Centre Tracts*, **139**, Mathematisch Centrum, Amsterdam.

White, D. J. (1993 a), *Markov Decision Processes*, John Wiley, New York.

White, D. J. (1993 b), Minimising a threshold probability in discounted Markov decision processes, *J. Math. Anal. Appl,* **173**, 634-646.

Wu, C. and Lin, Y. (1999), Minimizing risk models in Markov decision processes with policies depending on target values, *J. Math. Anal. Appl.* **231**, 47-67.

Received June 14,2000
Revised September 7,2000