J. Appl. Prob. **25**, 544–552 (1988)
Printed in Israel

THE OPTIMAL VALUE OF MARKOV STOPPING PROBLEMS WITH ONE-STEP LOOK AHEAD POLICY

MASAMI YASUDA,* *Chiba University*

Abstract

This paper treats stopping problems on Markov chains in which the OLA (one-step look ahead) policy is optimal. Its associated optimal value can be explicitly expressed by a potential for a charge function of the difference between the immediate reward and the one-step-after reward. As an application to the best choice problem, we shall obtain the value of three problems: the classical secretary problem, a problem with a refusal probability and a problem with a random number of objects.

OPTIMAL STOPPING PROBLEM; MARKOV POTENTIAL THEORY

1. Introduction

Let $\{x_n; n \geqq 0\}$ be a Markov chain over a state space S on the real line with stationary transition probabilities $P(x, dy)$, $x, y \in S$. The aim of the optimal stopping problem is to find a stopping time τ which maximizes the expectation of some payoff consisting of an immediate reward $r(x)$ and a sampling cost $c(x)$. The optimal value is denoted by

$$v(x) = \sup_{0 \leqq \tau < \infty} E\left[r(x_\tau) - \sum_{n=0}^{\tau-1} c(x_n) \,\middle|\, x_0 = x\right], \qquad x \in S \tag{1.1}$$

where the sup is taken over all finite stopping times. It is known that the value satisfies the optimality equation

$$v(x) = \max\{r(x), -c(x) + Pv(x)\}, \qquad x \in S. \tag{1.2}$$

The detailed analyses are discussed by many authors including Chow et al. [2].

We restrict ourselves to stopping problems in which the OLA (one-step look ahead) policy is optimal. Consider the set of states for which stopping immediately is at least as good as stopping after exactly one more period:

$$B = \{x \in S; r(x) \geqq Pr(x) - c(x)\} \tag{1.3}$$

where $Pr(x) = \int_S r(y)P(x, dy)$. The OLA policy is a rule that stops at the first time the process enters a state in this set. This policy is known to be optimal whenever the set is

Received 18 November 1986; revision received 24 July 1987.
* Postal address: College of General Education, Chiba University, Chiba, 260 Japan.

closed (Chow et al. [2], Ross [15], Cowan and Zabczyk [3] and Bojdecki [1]). This is the Markov version of the 'monotone case' in [2].

Our aim in this note is to obtain, by calculating a potential of the Markov chain, the optimal value associated with the OLA policy. We give in Section 3 the explicit solution of various versions for the best choice problem as application. The first is the classical problem and the second is a case with a refusal probability. The solution for a problem with a random number of objects is also calculated. The last case was reduced to a functional optimality equation by an *ad hoc* method in Yasuda [17], but the global form of the solution was not calculated explicitly.

The motivation for this approach has arisen in connection with the results of Darling [4] and Hordijk [10]. The former gave an upper bound for the optimal value by the potential operator and the latter gave a sufficient condition to find an optimal stopping time on the value by the potential.

2. The Markov potential and the OLA policy

For transition probabilities $P(x, dy)$, $x, y \in S$, a function $f(x)$, $x \in S$, is called a charge if $Nf(x) = \Sigma_{k=0}^{\infty} P^k f(x)$, $x \in S$, exists where $P^0 = I$ (the identity) and $P^{k+1} = PP^k$, $k = 1, 2, \cdots$. A function $g(x)$, $x \in S$ is a potential if there exists a charge f such that $g = Nf$. The support of a charge is the set on which the charge is not 0, the support of a potential is the support of its charge (refer to Kemeny et al. [11]). The relation between Markov potential theory and dynamic programming or Markov decision process is discussed by Hordijk [10].

The well-known condition for the optimality of the OLA policy is as follows.

Lemma 2.1. Assume that (i) the subset B of S defined in (1.3) is closed; i.e.

$$P(x, dy) = 0 \qquad \text{for } x \in B, dy \in \bar{B}, \tag{2.1}$$

(ii) the first hitting time $\tau(B)$ of the set B is finite with probability 1 for any initial state x_0. Then the OLA policy is optimal, that is, the first hitting time of B is the optimal stopping time whenever $r(x)$ is bounded above and $c(x)$ is non-negative.

Let $P_A f(x) = \int_A f(y) P(x, dy)$ for $A \subset S$ and let

$$N_B f(x) = \sum_{k=0}^{\infty} (P_B)^k f(x), \qquad x \in S.$$

We assume that

$$\lim_k [(P_{\bar{B}})^k r](x) = 0 \qquad \text{for } x \in \bar{B} \tag{2.2}$$

where $\bar{B}$ denotes the complement of the set B.

The property $\lim_k (P_{\bar{B}})^k v(x) = 0$ for the optimal value $v = v(x)$ is called equalizing in optimal gambling (Dubins and Savage [5], Hordijk [10]). One might say that here the earnings actually received in the time period up to k and the promised earnings equalize as k tends to ∞.

Theorem 2.2. Assume that the OLA policy is optimal by Lemma 2.1. Under (2.2), the optimal value is given by

$$v(x) = r(x) + N(Pr - r - c)^+(x) = \begin{cases} r(x) & \text{on } B \\ N_{\bar{B}}(P_B r - c)(x) & \text{on } \bar{B} \end{cases} \tag{2.3}$$

where $(\ )^+$ is the positive part of a function.

Proof. First we show that $v(x)$, $x \in S$, defined by the right hand side of (2.3) satisfies the optimality equation (1.2). On $x \in B$, for any $f = f(x)$, it holds that $Pf(x) = P_B f(x)$. From this and the closedness of the set B,

$$Pv(x) - c(x) = P_B v(x) - c(x) = P_B r(x) - c(x) = Pr(x) - c(x).$$

Hence

$$\max\{r(x), Pv(x) - c(x)\} = \max\{r(x), Pr(x) - c(x)\} = r(x) \qquad \text{for } x \in B.$$

On $x \in \bar{B}$, by substitution of (2.3), it holds that

$$\begin{aligned} Pv(x) - c(x) &= P_{\bar{B}} v(x) + P_B v(x) - c(x) \\ &= P_{\bar{B}}[N_{\bar{B}}(P_B r - c)](x) + P_B r(x) - c(x) \\ &= (P_{\bar{B}} N_{\bar{B}} - I)(P_B r - c)(x) \\ &= [N_{\bar{B}}(P_B r - c)](x). \end{aligned}$$

Simultaneously, on $x \in \bar{B}$,

$$r(x) < Pr(x) - c(x) = P_{\bar{B}} r(x) + P_B r(x) - c(x)$$

so $r(x) - P_{\bar{B}} r(x) < P_B r(x) - c(x)$. Applying the operator $N_{\bar{B}}$ to both sides of this inequality, we get, by assumption (2.2), $r(x) \leqq [N_{\bar{B}}(P_B r - c)](x)$ for $x \in \bar{B}$. Combining the two assertions, we have

$$\begin{aligned} &\max\{r(x), Pv(x) - c(x)\} \\ &\quad = \max\{r(x), [N_{\bar{B}}(P_B r - c)](x)\} = [N_{\bar{B}}(P_B r - c)](x) \quad \text{for } x \in \bar{B}. \end{aligned}$$

Since the hitting time of B is optimal by the assumption, $v(x)$ is the corresponding expected payoff and so it equals the optimal value.

Next, we calculate the potential $N(Pr - r - c)^+(x)$, $x \in S$. From the definition of the set B in (1.3), clearly we have $(Pr - r - c)^+(x) = 0$ on B. So the support of the charge is the complement of B and hence

$$N(Pr - r - c)^+(x) = 0, \qquad x \in B.$$

Since $(Pr - r - c)^+(x) = (Pr - r - c)(x) = [P_{\bar{B}} r + P_B r - r - c](x)$ for $x \in \bar{B}$, we have that

$$P(Pr - r - c)^+(x) = P_{\bar{B}}(Pr - r - c)(x)$$
$$= [(P_{\bar{B}})^2 r + P_{\bar{B}} P_B r - P_{\bar{B}} r - P_{\bar{B}} c](x), \qquad x \in \bar{B}.$$

Repeating this procedure to take the expectation up to k times, it implies that

$$\{I + P + P^2 + \cdots + P^k\}(Pr - r - c)^+(x)$$
$$= (P_{\bar{B}})^k r(x) + [\{(P_{\bar{B}})^{k-1} + (P_{\bar{B}})^{k-2} + \cdots + I_{\bar{B}}\}(P_B r - c)](x) - r(x), \qquad x \in \bar{B}.$$

Hence

$$N(Pr - r - c)^+(x) = [N_{\bar{B}}(P_B r - c)](x) - r(x), \qquad x \in \bar{B}$$

follows.

We remark that the upper estimated bound on the optimal value in Theorem 3.6 of Darling [4] equals exactly the optimal value in this case. That is, the bound holds with equality when the OLA policy is optimal and it is equalizing. The optimal stopping region is the complement of the support of the potential for the positive part of the difference between the reward for continuing for one step and the reward for stopping immediately.

3. The best choice problem

In this section we apply the previous method to the typical stopping problem known as the best choice problem or secretary problem. Refer to Freeman [8] for a review of the problem.

3.1. *The classical secretary problem.* The secretary problem is an optimal stopping problem based on relative ranks for objects arriving in a random fashion; the objective is to find the stopping rule that maximizes the probability of stopping at the best object of the sequence $\{1, 2, \cdots, n\}$. In the Markov formulation of the problem, by Dynkin and Yushkevitch [7], state i occurs when the ith arrival is relatively best, its reward equals i/n, i.e., the conditional probability that this relatively best arrival is best of all n (see also Chow et al. [2]). The optimality equation for the optimal value $v(i)$, $i \in S = \{1, 2, \cdots, n\}$, the maximal probability of win in the model, is then

$$v(i) = \max\left\{i/n, i \sum_{j=i+1}^{n} v(j)/((j-1)j)\right\}, \qquad i \in \{1, 2, \cdots, n-1\},$$

(3.1)

$$v(n) = 1.$$

Solving this equation, the optimal value when starting from the initial object, that is, the maximal probability of choosing the best object, $v(1)$ can be obtained. To consider the problem in the asymptotic form (infinite problem), we shall take the scale limit by $i/n \to x$ as $i, n \to \infty$. This reduces the equation (3.1) to

$$v(x) = \max\{x, Pv(x)\}, \qquad x \in [0, 1] \tag{3.2}$$

where the transition probability

$$P(x, dy) = \begin{cases} xy^{-2}dy & \text{for } 0 \leqq x < y \leqq 1, \\ 0 & \text{otherwise.} \end{cases} \tag{3.3}$$

The equation is a typical form of (1.2) with the reward function $r(x) = x$, no cost, $c(x) = 0$ and the state space $S = [0, 1]$. The solution of (3.2) is

$$v(x) = \begin{cases} e^{-1} & \text{on } [0, e^{-1}], \\ x & \text{on } [e^{-1}, 1]. \end{cases} \tag{3.4}$$

Hence the optimal value $v(0) = e^{-1}$ is the maximal probability of win in the infinite problem, as is well known from the literature (see Gianini and Samuels [9]).

Now we extend the problem by changing the reward function to a general reward $r(x)$ in order to expand application. The optimality equation (3.1) for the reward function $r(x)$, which we intend to consider, becomes

$$v(x) = \max\{r(x), Pv(x)\}, \qquad x \in [0, 1]. \tag{3.5}$$

However, the OLA rule does not hold — that is, the OLA policy is not always optimal — for a general reward case. To ensure the property, we assume, for a function $h(x)$, $x \in [0, 1]$ defined by

$$h(x) = r(x) - Pr(x), \tag{3.6}$$

that

(3.6i) $h(x)$ is finite-valued on $[0, 1]$,

(3.6ii) it changes its sign from $-$ to $+$ only once as x increases from 0 to 1.

So the equation $h(x) = 0$ has a unique solution α in the unit interval.

In conclusion, if (3.6) is assumed, then the unique solution of this equation (3.5) is given by

$$v(x) = \begin{cases} r(\alpha) & \text{on } [0, \alpha], \\ r(x) & \text{on } [\alpha, 1], \end{cases} \tag{3.7}$$

where α is defined by (3.6ii).

Although we can confirm it directly, we calculate the potential for this process and the solution (3.7) of the optimality equation (3.5) is given by applying the previous result (2.3) of Theorem 2.2. Immediately it is shown that $B = [\alpha, 1]$ by (2.1), and it is closed with respect to P. So $P_B r(x) = r(\alpha)(x/\alpha)$ and

$$(P_B)^n P_B r(x) = r(\alpha)(x/\alpha)\log^n(\alpha/x)/n!$$

for $x \in [0, \alpha]$, $n = 0, 1, 2, \cdots$. Hence we have

$$N_{\tilde{B}}P_{\tilde{B}}r(x) = r(\alpha)(x/\alpha)\{1 + \log(\alpha/x) + 2^{-1}\log^2(\alpha/x) + \cdots + (n!)^{-1}\log^n(\alpha/x) + \cdots\}$$
$$= r(\alpha)$$

for $x \in [0, \alpha]$. We can check the condition (2.2) by showing that

$$(P_{\tilde{B}})^n r(x) = x\int_x^\alpha \{r(y)y^{-2}\log^{n-1}(y/x)/(n-1)!\}dy$$

for $x \in [0, \alpha]$, tends to 0 as $n \to \infty$.

3.2. *A problem with refusal probability*. A variant of the best choice problem is a case with a refusal probability discussed by Smith [16]. There is a fixed probability $1 - p$ which allows the applicant to refuse an offer of employment in the model. When $p = 1$ it reduces to the classical secretary problem discussed in Section 3.1. Hence the optimality equation is given by

$$v(x) = \max\left\{pr(x) + (1-p)x\int_x^1 y^{-2}v(y)dy,\ x\int_x^1 y^{-2}v(y)dy\right\}, \quad x \in [0, 1] \tag{3.8}$$

in the asymptotic form (infinite problem). This is because when one makes the decision to stop, there is a probability p of attaining success $r(x)$ and $1 - p$ of failure and consequent obligation to continue the observation. Smith [16] had calculated the optimal value as $v(0) = p^{1/(1-p)}$ when $r(x) = x$.

The difficulty with the optimality equation (3.8) is that it is not in the form of (3.5). Let us therefore define the following new transition probability:

$$P(x, dy) = \begin{cases} py^{-1}(x/y)^p dy & \text{for } 0 \leqq x < y \leqq 1, \\ 0 & \text{otherwise,} \end{cases} \tag{3.9}$$

and consider the optimality equation (3.5) for this transition probability (3.9). We claim that (3.8) is equivalent to (3.5). This follows from the fact that the second terms of the right-hand sides of both (3.5) and (3.8), satisfy the same differential equation:

$$df(x)/dx = -px^{-1}(r(x) - f(x))^+, \qquad 0 < x < 1, \quad f(1-) = 0.$$

Although the transition probability (3.9) is more general than (3.3), it is possible to obtain the solution in the same way. We define, as in (3.6), a function $h_p(x)$ by

$$h_p(x) = r(x) - Pr(x), \qquad x \in [0, 1]. \tag{3.10}$$

If h_p satisfies (3.6i) and (3.6ii), so that the OLA policy is optimal for (3.9), then we obtain the solution by applying Theorem 2.2, namely

$$v(x) = \begin{cases} r(\alpha) & \text{on } [0, \alpha], \\ r(x) & \text{on } [\alpha, 1], \end{cases} \tag{3.11}$$

where α is the unique solution of $h_p(x) = 0$ in (3.10). We note that, when $r(x) = x$, then

$v(x) = p^{1/(1-p)}$ for $x \in [0, p^{1/(1-p)}]$, $= x$ for $x \in [p^{1/(1-p)}, 1]$ which is consistent with Smith [16].

3.3. *A problem with random number of objects.* In the classical version, the number of objects to be observed is fixed. We now consider the problem in which the number is random with known probability function. Presman and Sonin [13] formulated the probability model as a Markov chain and derived the optimality equation: that is, let $w(i), i = 1, 2, \cdots$ be the optimal value and p_i be a probability that the number of objects equals i, then

$$w(i) = \max\left\{\sum_{k=i}^{\infty} ik^{-1}p_k\pi_i^{-1}, \sum_{k=i+1}^{\infty} i\pi_k(k(k-1))^{-1}\pi_i^{-1}w(k)\right\}, \tag{3.12}$$

where $\pi_i = \Sigma_{k=i}^{\infty} p_k$. If we assume that the random number is bounded with probability 1,

$$n = \sup\{k; p_k > 0\} < \infty, \tag{3.13}$$

we can obtain the asymptotic form (infinite problem) in the unit interval by taking a scaling limit $i/n \to x$ as $i, n \to \infty$ of appropriate sequences of finite problems.

Let $\Phi(x)$, $x \in [0, 1]$ be the limiting distribution function of $\Sigma_{j=1}^{i+1} p_j = 1 - \pi_i$; $i \in \{1, 2, \cdots, n\}$. The optimality equation (3.12) is reduced to

$$w(x) = \max\{R(x), Pw(x)\}, \qquad x \in [0, 1] \tag{3.14}$$

where

$$R(x) = x(1 - \Phi(x))^{-1} \int_x^1 y^{-1} d\Phi(y)$$

and

$$P(x, dy) = \begin{cases} x(1 - \Phi(x))^{-1} y^{-2}(1 - \Phi(y))dy & \text{for } 0 \leqq x < y \leqq 1, \\ 0 & \text{otherwise.} \end{cases} \tag{3.15}$$

We could not apply the previous method to this transition probability. However a simple scheme on the equation (3.14) yields the reduction to the classical problem (3.5) with transition function given by (3.3), and a general reward.

Similarly as before, define a function $h_\Phi(x)$ for $x \in [0, 1]$ by

$$h_\Phi(x) = g(x) - \int_x^1 y^{-1} g(y) dy \tag{3.16}$$

where

$$g(x) = \int_x^1 y^{-1} d\Phi(y).$$

We assume that

(3.16i) $R(x)$ and $h_\Phi(x)$ are well defined on $x \in [0, 1]$,

(3.16ii) $h_\Phi(x)$ changes its sign only once from $-$ to $+$ as x varies from 0 to 1.

To solve (3.14), let $v(x) = (1 - \Phi(x))w(x)$. So it becomes

$$v(x) = \max\left\{xg(x), x\int_x^1 y^{-2}v(y)dy\right\}. \tag{3.17}$$

Then, using the assumption (3.16), the solution (3.17) is immediately obtained by the result of (3.7) as

$$v(x) = \begin{cases} \alpha g(\alpha), & x \in [0, \alpha], \\ xg(x), & x \in [\alpha, 1]. \end{cases}$$

Hence

$$w(x) = \begin{cases} \alpha g(\alpha)(1 - \Phi(x))^{-1}, & x \in [0, \alpha], \\ R(x), & x \in [\alpha, 1]. \end{cases} \tag{3.18}$$

This accords with the result of Yasuda [15] but this method is simpler than the *ad hoc* treatment of the functional equation approach adapted there. To give an example, if the the random number of objects is uniformly distributed $\Phi(x) = x$, $x \in [0, 1]$, then the assumption (3.16) is satisfied and $\alpha = e^{-2}$. The solution is derived by (3.18) as

$$w(x) = \begin{cases} 2e^{-2}(1 - x)^{-1}, & x \in [0, e^{-2}], \\ -x(1 - x)^{-1}\log(x), & x \in [e^{-2}, 1]. \end{cases}$$

Since the optimal value starting at the initial object is $w(0)$, it equals $2e^{-2}$ and this value is well known from, for instance, Rasmussen and Robbins [14].

References

[1] Bojdecki, T. (1978) On optimal stopping of a sequence of independent random variables — Probability maximizing approach. *Stoch. Proc. Appl.* **6**, 153–163.

[2] Chow, Y. S., Robbins, H. and Siegmund, D. (1971) *Great Expectations: The Theory of Optimal Stopping*. Houghton Mifflin, Boston.

[3] Cowan, R. and Zabczyk, J. (1978) An optimal selection problem associated with the Poisson process. *Theory Prob. Appl.* **23**, 584–592.

[4] Darling, D. A. (1985) Contribution to the optimal stopping problem, *Z. Wahrscheinlichkeitsth.* **70**, 525–533.

[5] Dubins, L. E. and Savage, L. J. (1965) *How to Gamble if You Must: Inequalities for Stochastic Processes*. McGraw-Hill, New York.

[7] Dynkin, E. B. and Yushkevitch, A. A. (1969) *Theorems and Problems on Markov Processes*. Plenum Press, New York.

[8] Freeman, P. R. (1983) The secretary problem and its extensions: a review. *Int. Statist. Rev.* **51**, 189–206.

[9] Gianini, J. P. and Samuels, S. M. (1976) The infinite secretary problem. *Ann. Prob.* **13**, 418–432.

[10] Hordijk, A. (1974) *Dynamic Programming and Markov Potential Theory*. Mathematisch Centrum, Amsterdam.

[11] Kemeny, J. G., Snell, J. L. and Knapp, A. W. (1976) *Denumerable Markov Chains*, 2nd edn. Springer-Verlag, Berlin.

[12] MUCCI, A. G. (1973) Differential equations and optimal choice problem. *Ann. Statist.* **1**, 104–113.

[13] PRESMAN, E. L. AND SONIN, I. M. (1972) The best choice problem for a random number of objects. *Theory Prob. Appl.* **17**, 657–668.

[14] RASMUSSEN, W. T. AND ROBBINS, H. (1975) The candidate problem with unknown population size. *J. Appl. Prob.* **12**, 692–701.

[15] ROSS, S. M. (1970) *Applied Probability Models with Optimization Applications*. Holden Day, San Francisco.

[16] SMITH, M. H. (1975) A secretary problem with uncertain employment. *J. Appl. Prob.* **12**, 620–624.

[17] YASUDA, M. (1984) Asymptotic results for the best-choice problem with a random number of objects. *J. Appl. Prob.* **21**, 521–536.