MARKOV DECISION MODEL FOR HUMAN HEALTH WITH POLICY ITERATION

Abubakar, U. Y.
Department of Mathematics/Statistics
Federal University of Technology Minna, Nigeria

Abstract

This paper focuses on a dynamic system which is reviewed at equidistant points of time and at each review, the system is classified into one possible number of states and subsequently a decision has to be made. The economic consequences of the decisions taken at the review times (decision epochs) are reflected in costs. These properties of Markov decision processes are employed to study the health condition of a human life. Consequently, the optimal cost of transition from a poor health condition to a good health condition and the long-run fraction of time that the man is in a poor health condition were obtained. The result could be used to study the health conditions of employees in both private and public sectors of the economy to determine productivity and retirement.

Keywords: Markov decision model, health condition, optimal cost and long-run fraction of time

Introduction

Consider a dynamic system which is reviewed at equidistant points of time (t = 0, 1, 2, ...) at each review the system is classified into one possible number of states and subsequently a decision has to be made. The set of possible states is denoted by I. for each state $i \in I$, a set A(i) of decisions or actions is given. The state space I and the action sets A(i) are assumed to be finite. The economic consequences of the decisions taken at the review times (decision epochs) are reflected in costs. This controlled dynamic system is referred to as a discrete-time Markov decision model. Markov Decision Models have been successfully applied in diverse industries from health Abubakar et al (2007), to a wide range of real life systems including inventory management, revenue management systems used nowadays by airlines and other industries Lautenbacher and Stidham(1999), Goto et a l(2004) and Samuelson(1969) respectively. The objective of this paper is to determine the optimal average cost and fraction of time that a man has a poor health condition in long-run, using Markov Decision model with policy iteration.

Markov Decision Process

According to Kurkani(1999), Puterman(1994), Goto et al (2004) and Hillier Lieberman(1980); we consider a Discrete Time Markov Chain (DTMC) {X_n, $n = 0, 1,...$}, whose transition probability matrix depends on the action taken A_n. Additionally, the system incurs in a cost $c(i, a)$ when an action a is chosen at some state i. Then the joint process {$(X_n, A_n), n=0, 1,...$}, is a Discrete Time Markov Decision Process (DTMDP).

Markov Decision Processes can be defined for finite horizon as well as for infinite horizon. The focus of this paper is infinite horizon problem.

Let X_n be the state of the system at any time n and its state space $S = \{1, 2,..., M\}$, assumed to be finite. After observing the state $X_n = i$, an action A_n is taken from the set of feasible actions $A(i)$ (i.e. the actions that can be taken at that state. which we assume to be finite) in any stage and $S(i, a)$ is the set of reachable states (i.e. the possible destinations when action *a* is taken in state *i*).

First of all, it is assumed that the system has the following Marko property,

$$P\{X_{n+1} = j|X_n = i, A_n = a\} = P\{X_{n+1} = j|X_n = i, A_n = a, X_{n-1}, A_{n-1}, \dots, X_0, A_0\}.$$

We assume that the process is homogeneous, in the sense that the previous probability does not depend on *n*, and denote

$P_{ij}(a) = P\{X_{n+1} = j|X_n = i, A_n = a\}$,

the probability that the system goes to state $j \in S\,(i, a)$ if action $a \in A(i)$, is chosen in state $i \in S$.

In order to give a description of how an action is chosen in the evolution of a DTMDP, it is necessary to define a *decision rule*, that is, a function that specifies the action $a \in A(i)$ that must be chosen when the system is in state $i \in S$ in every stage.

A policy is defined as a set of decision rules for each epoch of time. For infinite horizon problems under the homogeneity assumption stated above, and for any of the two criteria stated below, it can be shown that there exists an optimal policy that always chooses the same decision rule for every stage. Also it can be shown that there is an optimal policy depends only on the current state, and not on the previous history. This type of policy is called a Markov policy Diego et al(2006) *and* Derman(1970).

We define a *deterministic decision rule* will be denoted by $d(i), i \in S$ a map that assigns an action $a \in A(i)$ to every state and a *randomised policy*, denoted by $f(i, a)$, will be the probability that an action $a \in A_n$ is chosen if the state of the system is $i \in S_n$, i.e.

$f(i, a) = P\{A_n = a|X_n = i\}$.

Under a given policy *f*, the process $\{X_n, n \geq 0\}$ behaves like a discrete time Markov chain (DTMC) with transition probability matrix

$P^f = \left(P_{ij}^f\right)$, where $P_{ij}^f = P\{X_{n+1} = j|X_n = i\} = \sum_{a \in A} f(i, a) P_{ij}(a)$.

This result is established just by conditioning on A_n.

Whenever action *a* is taken in state *i* the system incurs an expected cost $c(i, a)$. There are three types of costs criteria that are typically studied for an infinite horizon DTMDP: discounted cost, total cost and average cost; the latter was considered in this paper. The objective is to find the policy *f* that minimizes costs according to one of the given criteria Diego and German(2006).

Following Tijms(1988) the n-step transition probabilities of the Markov chain $\{X_n\}$ is given by $P_{ij}^{(n)}(R) = P(X_n = j \mid X_0 = i)$, i,j ∈ I and n = 1,2,...

These transition probabilities satisfy the recursion relation

$$P_{ij}^{(n+1)}(R) = \sum_{k \in I} P_{ik}^{(n)}(R)\, P_{kj}(R_k) \quad n = 1,2,\dots$$

Let us say that state j can be reached from state i under policy R if $P_{ij}^{(n)} > 0$ for some $n \geq 1$

Assumption 1: The Markov chain $\{X_n, n = 0, 1,2,\dots\}$ has some regeneration state r (say) such that $E(N|N_0 = i) < \infty$ for all $i \in I$. where N is the first time beyond epoch 0 that the process makes a transition into state r.

Assumption 2: The expectation of the total rewards earned up to the first time beyond epoch 0 at which the process makes a transition into state r is finite for each initial state X_0 = i. This assumption follows directly from assumption 1 in case the average reward per unit time that

$\lim_{k\to\infty}\frac{1}{n}\sum_{k=1}^{n} f(x_k) = \sum_{j\in I} f(j)\,\pi_j$ with probability 1 (1).

Independently of the initial state X_0 = i.

Where: $\lim_{n\to\infty} P_{ij}^{(n)} = \pi_j$ for all $i, j \in I$

Assumption 3: For each stationary policy R, a state r (that may depend on R) exists which can be reached from any other state under policy R.

Using finiteness of the state space, as the above assumption implies that for each stationary policy R the associated Markov chain $\{X_n\}$ satisfies the preceding two assumptions. Thus, for each stationary policy R, we have

$$\pi_j(R) = \lim_{m\to\infty} {}^{1}\!/_{m} \sum_{n=1}^{m} P_{ij}^{(n)}(R)$$

exists and is independent of the initial state X_0 = i. the equilibrium distribution $\{\pi_j(R) \in I\}$ satisfies the system of linear equations

$$\pi_j(R) = \sum_{k\in I} P_{kj}(R_k)\,\pi_k(R),\ j \in I \qquad (2)$$

$$\sum_{j\in I} \pi_j(R) = 1 \qquad (3)$$

This system of linear equations has a unique solution. Also by the ergodic relation equation (1), we have with probability 1. The long term average cost per unit time when using rule

$$R = \sum_{j\in I} C_j(R_j)\,\pi_j(R) \qquad (4)$$

independently of the initial state.

Let g(R) = $\sum_{j\in I} C_j(R_j)\,\pi_j(R)$

A stationary policy R* is said to be average cost optimal if g(R*) ≤ g(R) for each stationary policy R.It has been observed that it is computationally not feasible to find an average cost optimal policy by computing the associated average cost for each stationary policy separately from equations (2) to (4) Tijms(1988).

What is eventually used is the relative values associated with a policy R. Consider the relation

$\lim_{n\to\infty} {}^{V_n(i,R)}\!/_{n}$ = g(R) for all i

Where V_n(i, R) denotes the total expected costs over the first n decision epochs when the initial state is i and policy R is used.

Suppose that the values V_i(R), i ∈ I, exist such that for each i∈I

$$V_n(i, R) = ng(R) + V_i(R) \text{ for all n large} \qquad (5)$$

Note that $V_i(R) - V_j(R) \sim V_n(i, R) - V_n(j, R)$ for n large, so that

$V_i(R) - V_j(R)$ measures the difference in total costs when starting in state i rather than in state j, given that policy R is followed.

Suppose also that the average cost g(R) and the relative values V_i(R), i∈I, satisfy a simultaneous system of linear equations

$$V_n(i, R) = C_i(R_i) + \sum_{j\in I} P_{ij}(R_i)\, V_{n-1}(j,R) \quad n \geq 1 \text{ and } i\in I$$

The recursion equation follows under the condition that the next state is j, the total expected costs over the remaining next n-1 decision epoch is $V_{n-1}(j, R)$. The next state is j with probability $P_{ij}(R_i)$ when the action a = R_i is used in the starting state i, by substituting 5 in the recursion equation, we find, after cancelling out common terms

$$g(R) + V_i(R) \approx C_i(R_i) + \sum_{j\in I} P_{ij}(R_i)\, V_j(R)\ i \in I$$

yielding the value-determination equations for policy R.

A rigorous way to introduce the relative values associated with a given stationary policy R is to consider the costs incurred until the first return to some regeneration state for policy R. we choose some state r such that for each initial state the Markov chain $\{X_n\}$ associated with policy R will visit state r with probability 1 after finitely many transitions.

We define for each state $i \in I$
$T_i(R)$ = the expected time until the first return to state r when starting in state i and using policy R

Let a cycle be the time elapsed between two consecutive visits to the regeneration state r under policy R, we have that $T_i(R)$ is the expected length of a cycle. Also define, for each $i\in I$

$K_i(R)$ = the cost incurred until the first return to state r when starting in state i and using policy R.

We suppose that $K_i(R)$ includes the cost incurred when starting in state i but excludes the cost incurred when returning to state r. The average cost per unit time equals the expected cost incurred in a cycle divided by the expected length of a cycle, thus

$$g(R) = \frac{K_i(R)}{T_i(R)} \tag{6}$$

We define the relative values

$$W_i(R) = K_i(R) - g(R)T_i(R),\ i\in I \tag{7}$$

as a consequence of (6), the normalization

$$W_i(R) = 0.$$

We state the following theorem without proof that the average cost per unit time and the relative values can be calculated simultaneously by solving a system of linear equations.

Theorem 1

Let R be a given stationary policy;

a) The average cost g(R) and the relative values $W_i(R)$, $i\in I$, satisfy the following system of linear equations in the unknown g and V_i, $i\in I$

$$V_i = C_i(R_i) - g + \sum_{j\in I} P_{ij}(R)\, V_j, i \in I \tag{8}$$

b) Let the numbers g and V_i, $i\in I$ be any solution to (8) then g = g(R). And for some constant C,

$V_i = W_i(R) + C$ for all $i\in I$

c) Let S be an arbitrary chosen state. Then the linear equations 8 together with the normalization equation $V_s = 0$ have a unique solution.

The economic interpretation of the relative value shows that for any solution $\{g(R), V_i(R)\}$ to the value determination equation 8 the numbers $V_i(R),\ i \in I$ are called the relative values of the various starting states when policy R is used. Assuming that the Markov chain $\{X_n\}$ is aperiodic, we have, for any two states $i, j \in I$

$V_i(R) - V_j(R) =$ the difference in total expected costs over an infinitely long period of time by starting in state i rather than in state j when using policy R.

In other words, $V_i(R) - V_j(R)$ is the maximum amount that a rational person is willing to pay to start the system in state j rather than in state i when the system is controlled by rule R.

Theorem 2

Let g and V$_i$, i∈I be given numbers. Suppose that the policy $\bar{R}$ has the property

$C_i(\bar{R}_i) - g + \sum_{j \in I} P_{ij}(\bar{R}_i)V_j \leq V_i \; for \; i \in I$ (9)

Then $g(\bar{R}) \leq g$. (10)

The theorem is also true when the inequality signs in (9) and (10) are reversed.

Proof: suppose that a control cost of $C_i(a) - g$ is incurred each time the action a is chosen in state i, while terminal cost of V$_j$ is incurred when the control of the system is stopped and the system is left behind in state j. Then 9 states that controlling the system for one step according to rule $\bar{R}$ and stopping next is preferable to stopping directly when the initial state is i. Since the property is true for each initial state, a repeated application of this property yields that controlling the system for m steps according to rule $\bar{R}$ and stopping after that is preferable to stopping directly. Thus using the notation of

$V_m(i,R) = \sum_{t=0}^{m-1} P_{ij}{}^{(t)}(R)C_j(R_j)$ (That is, the expected cost to be incurred at the decision epoch t given that X$_0$ = i and policy R is used.) with R replaced by $\bar{R}$, we have for each initial state i∈I,

$V_m(i,\bar{R}) - mg + \sum_{j \in I} P_{ij}{}^{(m)}(\bar{R})V_j \leq V_i \; for \; m = 1,2,\ldots$ (11)

Dividing both sides of (11) by m and let m→∞ it follows that $\square(\overline{\square}) - \square \leq \ell$, which is to be proved

Following Howard (1960), to improve a given policy R whose average cost *g*(*R*) and relative values V$_i$(R), i∈I have been computed, we apply the above theorem with

$\square = \square(\square)\ \square\square\square\ \square_\square = \square_\square(\square), \square \in \square.$

Thus, by constructing a new policy $\overline{\square}$ such that i∈I

$\square_\square(\square_\square) - \square(\square) + \sum_{\square \in \square} \square_{\square\square}(\overline{\square_\square})\square_\square \leq \square_\square$

(12)

We obtain an improved rule $\overline{\square}$ according to $g(\overline{\square}) \leq \square(R)$. In constructing such an improved policy $\overline{\square}$ it is important to realise that for each state i separately an action $\overline{\square_\square}$ satisfy (12) can be determined. A particular way to find for some $\square \in \square$ an action $\overline{\square_\square}$ satisfying (12) is to minimize

$$□_□(□) - □(□) + \sum_{□\in □} □_{□□}(□)□_□(□)$$
(13)

with respect to $□ \in □(□)$. Noting that (13) equals $□_□(□)$ for $□ = □_□$, it follows that (12) is satisfied for the action $\overline{□_□}$ which minimizes 13 with respect to $□ \in □(□)$.
The following presents the policy – iteration algorithm Tijm(1988).
Step 1: (initialization). Choose a stationary policy R.
Step 2: (value-determination step). For the current rule R, compute the unique solution $\{□(□), □_□(□)\}$ to the following system of linear equations

$$□_□ = □_□(□_□) - □ + \sum_{□\in □} □_{□□}(□_□)□_□, □ \in □$$
(14)

$$□_□ = 0 \qquad (15)$$

Where s is an arbitrary chosen state.

Step 3: (policy-improvement step). For each state $□ \in □$, determine an action $□(□)$ yielding the minimum in

$$\min_{a\in A(i)} \left\{□_□(□) - □(□) + \sum_{□\in □} □_{□□}(□)□_□(□)\right\}$$
(16)

The new stationary policy $\overline{□}$ is obtained by choosing $\overline{□_□} = □_□ □□□ □□□ □ \in □$ with the convection that $\overline{□_□}$ is chosen equal to the old action R_i when this action yields the minimum in (16)

Step 4: (convergence test). If the new policy $\overline{□}$ equals the old policy R_i, the algorithm is stopped with policy R. Otherwise go to step 2 with R replaced by $\overline{□}$.
The policy-iteration algorithm converges after a finite number of iterations. If the policy has converged to stationary policy $□^*$, then that policy is the average cost optimal.

The Model

Suppose that at the beginning of each day the health condition of a man is observed and classified as good health or poor health. If he is found to have poor health, he is given either a first aid/preventive treatment or curative treatment so that the health condition could change to good health and could attend to his usual activities

Suppose also that he could be found in good health conditions i = 1,2,... N. The good health condition i is better than i+1. That is the condition deteriorates in time. If the present condition is i and does not fall ill, then at the beginning of the next day then he has good health conditions j with probability p_{ij}. It is assumed that his body cannot improve on its own. That is $p_{ij} = 0$ for j<i so that $\sum j \geq i$, $p_{ij} = 1$. Let the health condition i = N represents a poor condition that requires treatment taking two days. For the intermediate states i with 1<i<N there is a choice for him to preventively take treatment so that he could remain in good health condition for the present day. Let a first aid/preventive treatment takes only one day at most and a change from poor health to a good health (after treatment) has a good health condition i=1. We wish to determine a rule which minimizes the long-term fraction of time the man is taking treatment.

Let us put the problem in the frame work of a discrete-time Markov decision model. We assume a cost of one for each day he takes treatment, the long-term average cost per day represent the long-term fraction of days that he takes treatment. Also, since a treatment for poor health condition N takes two days and in the discrete Markov decision model the state of the system has to be defined at the beginning of each day. We need auxiliary state for the situation in which a treatment is in progress. Thus the set of possible states of his health condition is chosen as

I = {1, 2, ... N, N+1}. Here the state i with 1 ≤ i ≤ N corresponds to the situation in which an observation reveals good health condition i, while state N+1 corresponds to the situation in which treatment is in progress already for one day. Denoting the two possible actions by

$$a = \begin{cases} 1 \ \square\square\ \square\square\square\ \square\square\square\square\square\square\square\square\square\ \square\square\ \square\square\square\square\ \square\square\square\square\square\square. \\ \qquad 0 \ \square\square\square\square\square\square\square\square\square \end{cases}$$

The set of possible actions in state i is chosen as
A (1) = {0}, A(i) ={0,1} for 1<i<N, A(N) = A(N+1) = {1}

We find that the one step transition probabilities P_{ij} (a) are given by

P_{ij} (1) =1 for 1<i<N
$P_{N,\,N+1}(1) = 1 = P_{N+1,\,1}$ (1)
P_{ij} (0) = P_{ij} for 1≤i≤N and j≥i
P_{ij} (a) = 0 otherwise

Further, the one step costs C_i (a) are given by
C_i (1) =1 and C_i (0) = 0.

A rule or policy for controlling the health condition is a prescription for taking actions at each decision epoch.

In view of Markovian assumption, and the fact that the planning horizon is infinitely long, we shall therefore consider stationary policies. A stationary policy R is a rule that always prescribes a single action R_i whenever the system is found in state i at a decision epoch.

The rule prescribing a treatment or poor health condition only when he has a good health condition for at least 5 working days is given by R_i = 0 for 1≤i<5 and R_i = 1 for 5 ≤ i ≤ N+1 **.**

Illustration

The average cost optimal when the number of possible working conditions equals N = 5 and the deterioration probabilities of the health conditions of staff in a company is given below

$$\square = \begin{pmatrix} 0.80 & 0.15 & 0.05 & 0 & 0 \\ 0 & 0.60 & 0.20 & 0.10 & 0.10 \\ 0 & 0 & 0.40 & 0.35 & 0.25 \\ 0 & 0 & 0 & 0.50 & 0.50 \end{pmatrix}$$

The policy – iteration algorithm is initialized with the policy which prescribes treatment, be it a first aid or curative action a=1 in each state except state 1
From equations (14) to (16), after some iterations, we obtain the minimum fraction of days that the staff is in a poor health condition equals 0.214; and to have assumed a cost of one unit for

each time of treatment we therefore have that value as the average cost optimal for the treatment.

Conclusion

The relative value associated with the policy obtained represent both the fraction of time in the long-run that the staff could be in a poor condition of health and perhaps absent from work, and the minimal cost incurred in the treatment. This could be determined for each staff so that for the staff whose value is a large contrast to that of the staff of the company could be considered as being in poor health condition quite often and therefore unproductive and may be retired. The cost obtained is not very realistic; it could be determined by other methods.

References

Abubakar, U. Y., Reju, S. A. & Awojoyogbe, B. O. (2007). Markov decision model and the application to the cost of treatment of leprosy disease. *Leonardo Journal of Sciences 11, 69-78. Available online: http://www ljs.acamicdiret.org.*

Derman, C. (1970). *Finite state Markovian decision processes.* New York: Academic Press

Diego, B. & German, R. (2006). *Linear programming solvers for Markov decision Processes.*www.sys.virgina.edu/sieds06/paper/FMorningSession5.1. Date accessed: Jan.3rd 2011.

Goto, G. H., Lewis, M. E. & Puterman, M. L. (2004). Coffee, tea, or ...? A Markov decision process Model for airline meal provisioning. *Transportation Science, 38 (1) 107-118.*

Hillier & Lieberman (1980). *Introduction to operations research.* Holden Day.

Howard, R. A. (1960). *Dynamic Programming and Markov Processes*. New York: Wiley.

Kulkani, V. G. (1999). *Modelling, analysis, design, and control of stochastic system.* Springer.

Lautedbacher, C. J. & Stidham, S. Jr. (1999). The underlying Markov decision process in the single-leg airline yield management problem. *Transportation Science, 33(2) 136-146.*

Puterman, M. L. (1990). *Handbooks in operations research and management science.* Amsterdam Elsevier science Publishers, 1(2), Markov decision Processes, 331-434.

Puterman, M. (1994). *Markov decision processes: Discrete stochastic dynamic programming.* New York: John Wiley.

Samuelson, P.A. (1969). Lifetime portfolio selection by dynamic stochastic programming. *The review of Economics and Statistics 51 (3), 239-246.*

Stidham, S. Jr. (2000). Optimal control of Markov chains; in computational probability. W.K. Grassman, Ed. Massachusetts, USA; *Kluwer's International Series in Operations Research and Management Science.*

Tijm, H.C. (1988). *Stochastic modelling and analysis: A computational approach.* New York: John Wiley & Sons.