246

Working Paper

विद्याविनियोगाद्विकासः
IIM
AHMEDABAD

INDIAN INSTITUTE OF MANAGEMENT AHMEDABAD

MEAN VARIANCE OPTIMALITY CRITERIA FOR DISCOUNTED MARKOV DECISION PROCESSES

by

J.K. Satia

W P No. 246
Sep.1978

INDIAN INSTITUTE OF MANAGEMENT
AHMEDABAD

ABSTRACT

The criteria of maximizing expected rewards has been widely used in Markov decision processes following Howard [2]. Recently considerations related to higher moments of rewards have also been incorporated by Jaquette [4] and Goldwerger [1]. This paper considers mean variance criteria for discounted Markov decision processes. Variability in rewards arising both out of variability of rewards during each period and due to stochastic nature of transitions is considered.

It is shown that randomized policies need not be considered when a function of mean and variance (μ-aσ) is to be optimized. However an example illustrates that policies which will simultaneously minimize variances for all states may not exist. We, therefore, provide a dynamic programming formulation for optimizing μ_i-aσ_i for each state i. . An example is given to illustrate the procedure.

MEAN VARIANCE OPTIMALITY CRITERIA FOR DISCOUNTED MARKOV DECISION PROCESSES

The criterion used for optimization in Markov decision process is maximization of expected rewards following the pioneering work of Howard [2]. The total rewards are probabilistic in nature both due to the stochastic nature of process of transition and the possible variability in rewards during each period. The criteria for optimization, therefore, need to be modified to incorporate provision for risk arising out of the probabilistic nature of total rewards.

Howard and Matheson [3] use utility functions which are exponential in form (implying constant risk aversion). The objective is to maximize certain equivalent reward. They use value iteration technique to optimize possible time varying processes of finite duration. A policy iteration procedure is also developed to find a stationary policy with highest certain equivalent gain for the infinite duration time invariant case.

Jaquette [4] uses the criteria of moment optimality. A policy is moment optimal if it lexicographically maximizes the sequence of signed moments of total discounted rewards with a positive sign if the moment is odd and negative sign if the moment is even. In other

words, first all policies with highest expected rewards are found. From this set, a subset is formed which minimizes second moment. This procedure is repeated for successive higher moments until either only one policy remains or the subset of policies have all identical higher moments. He discusses both discrete [5] and continuous [6] time case. A policy iteration procedure is used for optimization. He also extends his results for small interest rates [7]. It is however assumed that rewards in each period are certain.

It follows that in lexicographic moment optimality criteria, if there is a unique policy which optimizes the expected rewards then considerations related to higher moments are ignored. This may often be the case. We need to, therefore, consider criteria for optimization which incorporate considerations related to higher moments in addition to the expected rewards.

Goldwerger [1] considers the case of discrete time finite state discounted Markov decision process where the rewards during each periods are probabilistic. He gives a dynamic programming formulation for optimizing mean-variability criteria, namely, mean/standard deviation of rewards. An expression for variance of reward is developed for this purpose. However, in computing variance it is assumed that for a given horizon T, the reward during period t and future period t+s (s = 1,...,T-t) are independently distributed. In a

Markov decision process, the rewards depend upon the realization of the process. The rewards during t+s depend upon the state occupancy at period t. The rewards during period t also depend upon the state occupied at period t. Therefore, the rewards during period t and period t+s are not independent. Indeed, we need to consider variability of total discounted rewards even when the rewards during each period are certain and depend only upon state of the process and decision made.

In Section II of this paper, we derive an expression for variance of reward when the above assumption of independence is not made. The criterion used for optimization is a function of mean μ and standard deviation σ of total discounted rewards. More specifically the functional form $\mu-a\sigma$ is used where a is a specified Constant. It is shown in Section III that randomized policies need not be considered. However, an example shows that there may not exist policies which will simultaneously minimize variances for all states. Section IV, therefore, gives a dynamic programming formulation to optimize $\mu_i-a\sigma_i$ for each state i. The computational procedure is illustrated by a numerical example in Section V. Section VI is a summary and conclusion of the paper.

Section II: Variance of Rewards

We assume that a Markov process can be in one of the state i in a finite set of states numbered 1,...,S. The process makes a

transition at discrete constant time intervals. Let p_{ij} denote the probability of transition from state i to state j. Let R_{ij} denote the associated probability distribution of rewards, and let the yields associated with the transition be denoted by x_{ij}.

The expected reward and the variance are assumed to be finite and are given by,

$$m_{ij} = \int x_{ij} dR_{ij}(x),$$

$$s_{ij}^2 = \int x_{ij}^2 dR_{ij}(x) - m_{ij}^2$$

The future rewards are discounted by a factor β. The total rewards when there are n periods left to go and current state is i is a stochastic variate denoted by $y_i(n)$. Let $\mu_i(n)$ and $\sigma_i^2(n)$ respectively be the mean and variance of $y_i(n)$. We assume $y_i(0) = 0$ for all .

If the process makes a transition from state i to state j then,

$$y_i(n) = x_{ij} + \beta y_j(n-1), \quad n = 1,2,.. \qquad ....(1)$$

Using conditional expectation,

$$\mu_i(n) = \Sigma p_{ij} m_{ij} + \beta \Sigma p_{ij} \mu_j(n-1) \qquad ....(2)$$

From equation (1) above, we can see that the variance in $y_i(n)$ is due to

1. the variance in rewards s_{ij}^2 during each period, and
2. the stochastic nature of transition from i to j.

Both of these need to be considered in computation of the variance.

Using the expression for unconditional variance Var (Y) in terms of conditional variance {Parzen [9]} , if $E(Y) < \infty$, then

$$\text{Var}(Y) = E\{\text{Var}(Y/\text{transition from } i \text{ to } j)\} + \text{Var}\{E(Y/\text{transition from } i \text{ to } j)\} \quad \ldots.(3)$$

In other words, the variance is equal to the mean of the conditional variance plus the variance of the conditional mean.

Using equations (1) and (3), letting n = 1, and conditioning on transitions from i to j, we get

$$\sigma_i^2(1) = E(s_{ij}^2) + \text{Var}(m_{ij}) = \Sigma p_{ij} s_{ij}^2 + \Sigma p_{ij} m_{ij}^2 - (\Sigma p_{ij} m_{ij})^2$$

For sake of brevity, let $V_{i.}(x)$ denote the variance of x with respect to the probability density function $p_{i.}$ Then, similarly, for any n,

$$\sigma_i^2(n) = E[\text{Var}\{x_{ij} + \beta y_j(n-1)\}] + \text{Var}[E\{r_{ij} + \beta y_j(n-1)\}]$$

$$= E\{s_{ij}^2 + \beta^2 \sigma_j^2(n-1)\} + V_i\{m_{ij} + \beta \mu_j(n-1)\}$$

$$\sigma_i^2(n) = \Sigma p_{ij} s_{ij}^2 + \beta^2 \Sigma p_{ij} \sigma_j^2(n-1) + V_i\{m_{ij} + \beta \mu_j(n-1)\} \quad \ldots.(4)$$

It can be shown that $\lim_{n \to \infty} \sigma_i^2(n) = \sigma_i^2$ exists for $0 \leq \beta < 1$ using a corresponding matrix notation

$$\sigma^2 = [I - \beta^2 P]^{-1}[Q]$$

where $Q_i = \Sigma p_{ij} s_{ij}^2 + V_i\{m_{ij} + \beta \mu_j(n-1)\}$

Alternatively, we could have computed Variance of $y_i(n)$ by first computing its second moment using conditional expectations.

$$
\begin{aligned}
E\{y_i^2(n)\} &= E[E\{x_{ij} + \beta y_j(n-1)\}^2] \\
&= E[E\{(x_{ij}^2) + \beta^2 E\{y_j^2(n-1)\} + 2\beta m_{ij}\mu_j(n-1)] \\
&= \Sigma p_{ij} E(x_{ij}^2) + \beta^2 \Sigma p_{ij} E\{y_j^2(n-1)\} + 2\beta \Sigma p_{ij} m_{ij} \mu_j(n-1)
\end{aligned}
$$

The above expression relates successive second moments of rewards and is an adaptation of expression used by Jaquette [4] to provide for variance in each period rewards. For sake of convenience, we would use equations (2) and (4) to find mean and variance of rewards.

It is worth noting that if x_{ij} are normally distributed for all i,j, then the rewards associated with any realization would be sum of these rewards and would also be normally distributed. However, with each specific realization, a probability is associated. Therefore, the probability distribution of $y_i(n)$ would be a mixture of normal distributions.

Section III: Optimization using Mean Variance Criteria

We now assume that in each state i, a finite number of alternatives k (k = 1,...,K_i) are available. Superscript k will be used to denote dependence of parameters on decision k. We will use the maximization criteria $\mu_i - a\sigma_i$ to incorporate the risk following Markowitz [8]. It provides for differential weightage to be given to standard deviation of rewards as compared to expected rewards.

First we will show that randomized policies need not be considered. An example is given which illustrates that there may not exist a policy which will simultaneously maximize $\mu_i - a\sigma_i$ for all states i. We, therefore, need to redefine the optimization criteria so as to maximize $\mu_i - a\sigma_i$ for each state i.

Prop.1: Randomized policies need not be considered where criterion used for maximization is $\mu_i - a\sigma_i$.

It will suffice to show that a policy C which randomizes between any two policies A and B with probability p and q (= 1-p) respectively cannot do better than best of A and B. In other words it will suffice to show that

$$\mu_i^C - a\sigma_i^C \leq \max\{\mu_i^A - a\sigma_i^A, \mu_i^B - a\sigma_i^B \text{ for all } i$$

$$\text{But} \quad \mu_i^C = p\mu_i^A + q\mu_i^B$$

Therefore, it would suffice to show that

$$p\mu_i^A + q\mu_i^B - a\sigma_i^C \leq p\mu_i^A - ap\sigma_i^A + q\mu_i^B - aq\sigma_i^B$$

$$\text{i.e.} \quad \sigma_i^C \geq p\sigma_i^A + q\sigma_i^B$$

Since

$$pq\sigma_i^{2B} + pq\sigma_i^{2A} \geq 2pq\sigma_i^A\sigma_i^B ,$$

$$p\sigma_i^{2A} + q\sigma_i^{2B} \geq (p\sigma_i^A + q\sigma_i^B)^2 .$$

Now using equation (3) for conditional expectations,

$$\sigma_i^{2C} = p\sigma_i^{2A} + q\sigma_i^{2B} + \text{Var}\,(p\mu_i^A + q\mu_i^B)$$

$$\geq p\sigma_i^{2A} + q\sigma_i^{2B}$$

$$\geq (p\sigma_i^A + q\sigma_i^B)^2$$

$$\sigma_i^C \geq p\sigma_i^A + q\sigma_i^B$$

which completes the argument that randomized policies need not be considered.

The variance in any state i is also affected by the expected rewards μ_j for all possible states j to which transition from i can take place. The variance σ_i, therefore, may depend upon decisions made in other states. A question arises as to whether there exist policies which will simultaneously maximize $\mu_i - a\sigma_i$ for all states i.

The following example shows that there may not exist any such policy. If a is assumed to take sufficiently large values then considerations relating to μ can be ignored. Therefore, it will suffice to demonstrate the above for variance alone.

Example

Beta = .9

The rewards during each period are assumed to be deterministic.

State	Decision	Transition probabilities		Rewards
i	k	p_{ij}^k		
1	1	0.5	0.5	6
	2	0.9	0.1	4
2	1	0.4	0.6	-3

There are only two policies (1,1) and (2,1). The corresponding values of μ and σ^2 are given by

Policy	μ_1	μ_2	σ_1^2	σ_2^2
(1.1)	15.5	5.6	102.4	101.5
(2,1)	28.5	15.8	76.3	109.3

The variance of rewards when starting from state 2 is greater for policy (2,1) than when policy (1,1) is used. Therefore, policy (1,1) minimizes variance when the process is in state 2 and policy (2,1) minimizes variance when the process is in state 1. Thus we need to incorporate considerations relating to decision in state 1 for minimizing variance when the process is in state 2.

Section IV: A Dynamic Programming Formulation

We assume that the decision in a state i will be selected so as to maximize $\mu_i - a\sigma_i$ with respect to that state i and for the remaining horizon period. This assumption implies that for an infinite horizon period, the optimal policy discovered may not simultaneously maximize $\mu_i - a\sigma_i$ for all states i. In the example of the previous section, therefore, policy (2,1) may be considered optimal in the individual state sense. A dynamic programming formulation is possible with this assumption.

Let $f_i(n) = \max_k\{\mu_i^k(n) - a\sigma_i^k(n)\}$ be the optimal risk adjusted return function when the current state of the process is i and n periods are remaining. Then using equations (2) and (4), and using principles of optimality,

$$f_i(n) = \max_k [\Sigma p_{ij}^k m_{ij}^k + \beta \Sigma p_{ij}^k \mu_j(n-1) - a\{\Sigma p_{ij}^k s_{ij}^{2k} + \beta^2 \Sigma p_{ij}^k \sigma_j^2(n-1) + V_i^k\{m_{ij}^k + \beta \mu_j(n-1)\}\}^{1/2}]$$

for all i, ... (5)

The procedure for solution of the above functional equation is as follows:

For each n, n = 1,2,...,

1) Find best decision k for each state i using equation (5)

2) Compute expected value of rewards using equation (2) and

3) Compute variance of rewards using equation (4)

The following proposition shows that the solution of the above functional equations converge.

Prop. 2: $\lim_{n\to\infty} f_i(n)$ exists where $f_i(n)$ are successive solutions to the functional equations 5.

Given any sequence of decisions, it can be shown that $\lim_{n\to\infty} \mu_i(n)$ exists. Then using equation (5) it can be shown that

$$|\sigma_i^2(n+1) - \sigma_i^2(n)| \leq \beta^2 |\sigma_i^2(n) - \sigma_i^2(n-1)| + T_n$$

Where $\lim_{n\to\infty} T_n = 0$. Therefore $\lim_{n\to\infty} \sigma_i^2(n)$ exists.

Since $f_i(n) = \mu_i(n) - a\sigma_i(n)$, $\lim_{n\to\infty} f_i(n) = f_i$ exists.

The optimal stationary policy (as defined in the beginning of this section) is given by solution to the following functional equations.

Let $f_i = \lim_{n\to\infty} f_i(n)$

Then
$$f_i = \max_k [\Sigma p_{ij}^k (m_{ij}^k + \beta\mu_j) - a\{\Sigma p_{ij}^k (s_{ij}^{2k} + \beta^2\sigma_j^2) + V_i^k (m_{ij}^k + \beta\mu_j)\}^{1/2}] \text{ for all } i \quad \ldots\ (6)$$

Where if $A = A_1,\ldots,A_s$ is the optimal policy then,

$$\mu_i = \Sigma p_{ij}^{A_i} m_{ij}^{A_i} + \beta\Sigma p_{ij}^{A_i}\mu_j \text{ , and}$$

$$\sigma_i^2 \quad \Sigma p_{ij}^{A_i} s_{ij}^{2A_i} + \beta\Sigma p_{ij}^{A_i}\sigma_j^2 + V_i^{A_i}(m_{ij}^{A_i} + \beta\mu_j)$$

Section V: An Illustrative Example

In this section we illustrate the dynamic programming procedure using the example of Goldwerger. The data for example is as follows:

Discount factor = 0.5

State	Decision	Transition probabilities		Expected rewards		Variance of rewards	
i	k	p_{ij}^k		m_{ij}^k		s_{ij}^{2k}	
1	1	.5	.5	9	3	5	2
	2	.8	.2	4	4	2	1
2	1	.4	.6	3	-7	2	3
	2	.7	.3	1	-19	.5	2

The dynamic programming equations were solved for values of a = 0 +0.2, +1.0, and the results are given below.

a = 0

n	$\mu_1(n)$	$\sigma_1^2(n)$	$f_1(n)$	$\mu_2(n)$	$\sigma_2^2(n)$	$f_2(n)$	k_1	k_2
1	6.00	12.50	6.00	-3.00	26.60	-3.00	1	1
2	6.75	35.95	6.75	-2.70	58.30	-2.70	1	1
3	7.01	44.03	7.01	-2.46	66.98	-2.46	1	1
4	7.14	46.19	7.14	-2.34	69.17	-2.34	1	1
5	7.20	46.74	7.20	-2.27	69.72	-2.27	1	1

The optimal policy for a = +0.2 remains same as above and therefore, value of $f_1(n)$ and $f_2(n)$ only change. However, when a = +1.00, the optimal policy changes to (2,1) and the results are given below.

n	$\mu_1(n)$	$\sigma_1^2(n)$	$f_1(n)$	$\mu_2(n)$	$\sigma_2^2(n)$	$f_2(n)$	k_1	k_2
1	4.00	1.80	2.66	-3.00	26.60	- 8.16	2	1
2	5.30	5.45	2.97	-3.10	50.51	-10.21	2	1
3	5.81	8.24	2.94	-2.87	59.12	-10.56	2	1
4	6.04	9.42	2.97	-2.70	61.64	-10.55	2	1
5	6.15	9.82	3.01	-2.60	62.33	-10.50	2	1
6	6.20	9.94	3.04	-2.55	62.52	-10.46	2	1
7	6.22	9.98	3.06	-2.53	62.56	-10.44	2	1
8	6.24	9.99	3.08	-2.51	62.58	-10.42	2	1
9	6.24	9.99	3.08	-2.51	62.58	-10.42	2	1
10	6.25	9.99	3.09	-2.50	62.58	-10.41	2	1

The above example shows, that the optimal policy will depend upon the penalty assigned for risk, thus emphasizing the importance of consideration of variance. When a = +1, the policy selected has lower variance and lower mean compared to the optimal policy for a = 0. It should also be noted that the convergence of $f_i(n)$ need not necessarily be monotonic, unlike the case where the optimization criteria considers only the expected rewards.

Section VI: Summary and Conclusion

In this paper we have derived an expression for variance of total discounted rewards when each period rewards are uncertain. The variance in total rewards arises from two sources, variability of rewards during each period and the stochastic nature of the process of transition. So the risk considerations are important even when each period rewards are certain.

Goldwerger [1] had assumed independence of rewards during different periods and thus ignored the full effect of stochastic nature of process. The expression derived in this paper does not assume such independence. The optimality criteria used is $\mu_i - a\sigma_i$. By varying a, the weightage given to variability in rewards for different policies can be altered.

An example was given which demonstrated that an optimal policy which will simultaneously maximize this criteria for all states i may not exist unlike the expected rewards case. This interaction among states implies that a joint optimality criteria may need to be defined as in multiple-criteria objective function. An obvious choice for a joint criteria does not seem to emerge.

We, therefore, have assumed that $\mu_i - a\sigma_i$ would be maximized for each individual state i ignoring the simultaneous consideration of

decisions in other states. A dynamic programming formulation is provided for this purpose. Only discounted Markov decision process was considered here. The approach needs to be extended to the case where there is no discounting.

The approach suggested in this paper is applicable wherever variance considerations become important. A few examples are manpower planning, maintenance-replacement and disease treatment models.

References

1 Goldwerger,J., "Dynamic Programming for a Stochastic Markovian Process with an Application to the Mean Variance Models", Management Science, Vol. 23, No. 6, Feb. 1977

2 Howard, R.A., Dynamic Programming and Markov Processes, The M.I.T. Press, 5th Printing, 1969

3 Howard, R.A. and Matheson, J.E., "Risk/Sensitive Markov Decision Processes", Management Science, Vol. 18, No. 7, March 1972

4 Jaquette, S.C., Markov decision processes with a new optimality criterion, Technical Report No. 15, Department of Operations Research, Stanford University, 1971

5 Jaquette, S.C., "Markov decision processes with a new optimality criterion: discrete time" The Annals of Statistics, Vol. 1, 1973, pp. 496-505

6 Jaquette, S.C., "Markov decision processes with a new optimality criterion: continuous time", The Annals of Statistics, Vol. 3 No. 2, 1975, pp. 547-553

7 Jaquette, S.C., "Markov decision processes with a new optimality criterion: small interest rates", Annals of Mathematical Statistics, Vol. 43, No. 6, 1972, pp. 1894-1901

8 Markowtizh, H., Portfolio Selection, New York, 1959

9 Parzen, E., Stochastic Processes, Holdenday Inc., 1962