Operations Research

Publication details, including instructions for authors and subscription information:
http://pubsonline.informs.org

Technical Note—An Equivalence Between Continuous and Discrete Time Markov Decision Processes

Richard F. Serfozo,

To cite this article:
Richard F. Serfozo, (1979) Technical Note—An Equivalence Between Continuous and Discrete Time Markov Decision Processes. Operations Research 27(3):616-620. https://doi.org/10.1287/opre.27.3.616

Technical Notes

An Equivalence Between Continuous and Discrete Time Markov Decision Processes

RICHARD F. SERFOZO
Bell Laboratories, Holmdel, New Jersey

(Received December 1976; accepted May 1978)

A continuous time Markov decision process with uniformly bounded transition rates is shown to be equivalent to a simpler discrete time Markov decision process for both the discounted and average reward criteria on an infinite horizon. This result clarifies some earlier work in this area.

THE EQUIVALENCE we shall discuss for Markov decision processes is based on the following well known equivalence for Markov processes. Let $Y = \{Y(t): t \geq 0\}$ be a continuous time Markov process with a countable state space whose jumps are determined by transition probabilities $Q(i, j)$, and whose sojourns in state i have an exponential distribution with parameter $0 \leq \lambda_i < \infty$ ($\lambda_i = 0$ means that i is an absorbing state). See Cinlar [1] or Gihman and Skorohod [2]. The infinitesimal generator of Y is given by

$$A(i, j) = \lim_{t \to 0} \frac{d}{dt} P(Y(t) = j \mid Y(0) = i) = \begin{cases} -p_i\lambda_i & \text{if } i = j \\ \lambda_i Q(i, j) & \text{if } i \neq j, \end{cases}$$

where $p_i = 1 - Q(i, i)$. Assume that $c = \sup_i p_i\lambda_i < \infty$, that is, the transition rates of Y are bounded. Now consider another continuous time Markov process $Y' = \{Y'(t): t \geq 0\}$ with transition matrix

$$Q'(i, j) = \begin{cases} 1 - p_i\lambda_i/c & \text{if } i = j \\ \lambda_i Q(i, j)/c & \text{if } i \neq j \end{cases}$$

and the exponential sojourn parameters $\lambda_i' = c$ for all i.

An easy check shows that the infinitesimal generator of Y' is the same as that for Y. From this it follows that Y is equal in distribution to Y'. That is, their finite dimensional distributions are equal when they have the same initial distribution. This equivalence between Y and Y' is useful

Operations Research
Vol. 27, No. 3, May–June 1979

0030-364X/79/2703-0616 $01.25

Published in Operations Research on June 01, 1979 as DOI: 10.1287/opre.27.3.616.
This article has not been copyedited or formatted. The final version may differ from this version.

for simplifying certain computations for Y, see [1]: the Y' is a Markov chain subordinated to a Poisson process and is more tractable than Y.

The above equivalence was used by Howard [4] and Veinott [11] to obtain an equivalence between continuous and discrete time Markov decision processes. They considered continuous time processes with finite state spaces and discounted rewards, where rewards are received continuously over time. Two related works concerning algorithms are Porteus [8] and Schweitzer [9].

Lippman [6] apparently was the first to recognize the usefulness of this equivalence for establishing the existence of optimal control policies with certain monotonicity properties. He considers queuing processes with countable state spaces and discounted rewards, where lump rewards are received at transition times as well as continuously over time. In his general discussion, Lippman states that the one-step reward functions are the same for the two equivalent processes. This is true for continuous rewards, but it is not true when there are lump rewards. His specific applications are correct, however, since he adjusts the rewards implicitly in his analyses.

In this article we present a formal description of this equivalence which clarifies the form of the one-step reward functions for the two equivalent processes. We consider average rewards as well as discounted rewards. We also consider processes that may take bogus jumps from a state back to itself and may have absorbing states. The results herein are used in Serfozo [9] to analyze controlled birth and death processes and queues in terms of equivalent random walks.

THE EQUIVALENCE

We shall consider a continuous time Markov decision process which moves as follows (see [4], [5] and [7]). Upon arriving at a state $i \in S$, an action $a \in A$ is chosen, and a reward $r(i, a)$ is received. For simplicity we take the sets S and A to be countable (our results readily extend, with appropriate measurability conditions to more general spaces, as in [3]). The process remains in state i for a random sojourn time which is exponentially distributed with parameter $\lambda(i, a) \geq 0$, and then it jumps to state $j \in S$ with probability $p(i, a, j)$. This series of events is repeated indefinitely. Note that $\lambda(i, a) = 0$ means that i is an absorbing state under action a. Also we allow $p(i, a, i) > 0$, which means that the process may take a "bogus" jump from i back to itself.

We let f denote the stationary policy which chooses action $f(i)$ when the process is in state i. We assume that the discounted and average rewards under f exist. We denote these by $W_f(i) = E_f(\sum_{n=0}^{\infty} e^{-\beta T_n} r(Y_n, a_n) \mid Y_0 = i)$, where $\beta > 0$ is a discount factor, and $\Psi_f(i) = \lim_{t\to\infty} E_f(t^{-1} \sum_{n=0}^{N_t} r(Y_n, a_n) \mid Y_0 = i)$. Here the controlled Markov process

Published in Operations Research on June 01, 1979 as DOI: 10.1287/opre.27.3.616.
This article has not been copyedited or formatted. The final version may differ from this version.

$Y = \{Y(t): t \le 0\}$ is described by $Y(t) = Y_n$ if $T_n \le t < T_{n+1}$, where T_n is the time of the nth jump of Y which is to state Y_n, the $a_n = f(Y_n)$, and $N_t = \max\{n: T_n \le t\}$ is the number of jumps that occur in time t. We use the convention that $T_k = \infty$ for $k \ge n+1$ when Y is absorbed in state Y_n. We shall denote this controlled Markov process by $Y = (S, A, r, \lambda, p, \beta)$.

Note that we are considering $r(i, a)$ as a lump reward received at the beginning of a sojourn in state i when action a is taken: it is a composite reward associated with the sojourn and the next jump. In typical applications, the r is of the form $r(i, a) = E_f(\rho_0(i, Y_1) + \int_0^{T_1} e^{-\beta t}\rho_1(i, Y_1)dt + e^{-\beta T_1}\rho_2(i, Y_1) | Y_0 = i) = \sum_j p(i, a, j)[\rho_0(i, j) + (\rho_1(i, j) + \rho_2(i, j)\lambda(i, a))/(\beta + \lambda(i, a))]$ where $a = f(i)$. Here $\rho_0(i, j)$ and $\rho_2(i, j)$ are lump rewards received at the beginning and end of a sojourn in state i, and $\rho_1(i, j)$ is the reward rate received during the sojourn when the next state is j.

We shall also consider a discrete time controlled Markov chain, that moves as follows. Upon arriving at a state $i \in S$, an action $a \in A$ is taken, a reward $r(i, a)$ is received, and then the process jumps to a state $j \in S$ with probability $p(i, a, j)$. This series of events is repeated indefinitely. As above, we let f denote a stationary policy and we assume that the discounted and average rewards under f exist. We denote these by $V_f(i) = E_f(\sum_{n=0}^{\infty} \alpha^n r(X_n, a_n) | X_0 = i)$ where $0 < \alpha < 1$ is a discount factor, and $\Phi_f(i) = \lim_{n\to\infty} E_f(n^{-1}\sum_{k=0}^{n-1} r(X_k, a_k) | X_0 = i)$. Here X_n is the nth state of the process and $a_n = f(X_n)$ is the nth action taken. We shall denote this controlled Markov chain by $X = (S, A, r, p, \alpha)$.

The following result asserts that if a controlled Markov process has uniformly bounded transition rates, then one can construct a controlled Markov chain which is equivalent to it.

THEOREM. *Let $\hat{Y} = (S, A, \hat{r}, \hat{\lambda}, \hat{p}, \beta)$ be a controlled Markov process with $(1 - \hat{p}(i, a, i))\hat{\lambda}(i, a) \le c < \infty$ for all a and i. Let $Y = (S, A, r, \lambda, p, \beta)$ be a controlled Markov process where*

$$r(i, a) = \begin{cases} \hat{r}(i, a)(\beta + \hat{\lambda}(i, a))/(\beta + c), & \text{when considering discounted rewards} \\ \hat{r}(i, a)\hat{\lambda}(i, a)/c, & \text{when considering average rewards,} \end{cases} \quad (1)$$

$$p(i, a, j) = \begin{cases} 1 - \hat{\lambda}(i, a)(1 - \hat{p}(i, a, i))/c, & \text{if } i = j \\ \hat{\lambda}(i, a)\hat{p}(i, a, j)/c, & \text{if } i \ne j, \end{cases}$$

and $\lambda(i, a) = c$ for all a and i. Let $X = (S, A, r, p, c/(\beta + c))$ be a controlled Markov chain, where r and p are defined by (1). *If $\hat{Y}$, Y and X are controlled by a stationary policy f, then $\hat{W}_f = W_f = V_f$ and $\hat{\Psi}_f = \Psi_f = c\Phi_f$.*

Remark. Suppose that each of the decision processes $\hat{Y}$, Y and X has a stationary discounted optimal policy within its respective class of

Published in Operations Research on June 01, 1979 as DOI: 10.1287/opre.27.3.616.
This article has not been copyedited or formatted. The final version may differ from this version.

randomized nonstationary policies. This is true, for example, when A is finite and $\hat{r}$ is bounded. Of course $\hat{Y}$ and Y have a larger class of policies than X, but all three have the same class of (deterministic) stationary policies. From the above result, it follows that if a stationary policy f is discounted optimal for any one of the $\hat{Y}$, Y or X, then it is discounted optimal for the other two. A similar comment applies to average optimal policies, ϵ-discounted or ϵ-average optimal policies, etc. In this sense, the $\hat{Y}$, Y and X are equivalent decision processes.

Proof. Under the policy f the $\hat{Y}$ is a Markov process with exponential sojourn parameters $\hat{\lambda}_i = \hat{\lambda}(i, f(i))$, and transition probabilities $\hat{Q}(i, j) = \hat{p}(i, f(i), j)$. Similarly, Y is a Markov process with parameters $\lambda_i = c$ and transition probabilities

$$Q(i, j) = p(i, f(i), j) = \begin{cases} 1 - \hat{\lambda}_i(1 - \hat{Q}(i, i))/c & \text{if } i = j \\ \hat{\lambda}_i \hat{Q}(i, j)/c & \text{if } i \neq j. \end{cases} \tag{2}$$

Clearly $\hat{Y}$ and Y have the same infinitesimal generator, and so they are equal in distribution (when they have the same initial distribution). Then in order to prove $\hat{W}_f = W_f$, it suffices to show that the expected discounted rewards associated with $\hat{Y}$ and Y for an "actual" sojourn in a state i are equal for each i.

Consider an actual sojourn in a state i by the process Y. The Y may take a bogus jump from i back to itself with probability $Q(i, i)$. Consequently, the number ν of bogus jumps from i to itself, before a new state is reached, is a geometric random variable with $P(\nu = n) = Q(i, i)^n(1 - Q(i, i))$ for $n \geq 0$. Here $\nu = 0$ when $Q(i, i) = 0$, and $\nu = \infty$ when $Q(i, i) = 1$. Then the expected reward received by Y for an actual sojourn in state i is

$$E\sum_{n=0}^{\nu} e^{-\beta\tau_n} r(i, f(i)) = r(i, f(i))E \sum_{n=0}^{\nu} \gamma^n = r(i, f(i))/(1 - \gamma Q(i, i)). \tag{3}$$

Here τ_n denotes the time of the nth bogus jump in state i: it has a gamma distribution with parameters $\gamma = c/(\beta + c)$ and n, and it is independent of ν.

By a similar argument, it follows that the expected reward received by $\hat{Y}$ for an actual sojourn in i is

$$E \sum_{n=0}^{\hat{\nu}} e^{-\beta\hat{\tau}_n} \hat{r}(i, f(i)) = \hat{r}(i, f(i))/(1 - \hat{\gamma}_i \hat{Q}(i, i)) \tag{4}$$

where $\hat{\tau}_n$ has a gamma distribution with parameters $\hat{\gamma}_i = \hat{\lambda}_i/(\beta + \hat{\lambda}_i)$ and n. Using (1) and (2) in the last term in (3), one can see that the rewards given in (3) and (4) are equal. Thus we have $\hat{W}_f = W_f$. Furthermore, by an obvious use of conditional expectations in W_f, it follows that $W_f = V_f$.

Now consider the average rewards $\hat{\Psi}_f$, Ψ_f and Φ_f under f for $\hat{Y}$, Y and X, respectively. By a well known Abelian theorem [12, p. 182], $\hat{\Psi}_f(i) = \lim_{\beta\to 0} \beta\hat{W}_f(i) = \lim_{\beta\to 0} \beta W_f(i) = \Psi_f(i)$ for all i. Furthermore, using

Abelian theorems for sums and integrals, $\Psi_f(i) = \lim_{\beta\to 0} \beta W_f(i) = \lim_{\beta\to 0} \beta E_f(\sum_{n=0}^{\infty} (c/(\beta + c))^n r(X_n, f(X_n))|X_0 = i) = \lim_{\alpha\to 1} c\alpha^{-1}(1-\alpha)\, E_f(\sum_{n=0}^{\infty} \alpha^n r(X_n, f(X_n))|X_0 = i) = c\Phi_f(i)$.

ACKNOWLEDGMENT

I thank Professors S. Deshmukh, S. Lippman and A. Veinott for their comments on this paper. This research was sponsored in part by the Air Force Office of Scientific Research under Grant AFOSR-74-2627, and by the National Science Foundation under grant ENG-75-13653.

REFERENCES

1. E. Cinlar, *Introduction to Stochastic Processes*, Prentice-Hall, Englewood Cliffs, N.J., 1975.
2. I. Gihman and A. Skorohod, *The Theory of Stochastic Processes II*, Springer-Verlag, New York, 1975.
3. K. Hinderer, *Foundations of Non-Stationary Dynamic Programming with Discrete Time Parameter*, Lecture Notes in Operations Research and Mathematical Economics, Vol. 33, Springer-Verlag, New York, 1970.
4. R. Howard, *Dynamic Programming and Markov Processes*, John Wiley & Sons, New York, 1960.
5. P. Kakumanu, "Continuously Discounted Markov Decision Model with Countable State and Action Space," *Ann. Math. Statist.* **42,** 919–926 (1971).
6. S. Lippman, "Applying a New Device in the Optimization of Exponential Queueing Systems," *Opns. Res.* **23,** 687–710 (1975).
7. B. L. Miller, "Finite State Continuous Time Markov Decision Processes with a Finite Planning Horizon," *SIAM J. Control* **6,** 266–280 (1968).
8. E. Porteus, "Bounds and Transformations for Discounted Finite Markov Decision Chains," *Opns. Res.* **23,** 761–784 (1975).
9. P. Schweitzer, "Iterative Solution of the Functional Equations of Undiscounted Markov Renewal Programming," *J. Math. Anal. Appl.* **34,** 495–501 (1971).
10. R. Serfozo, "Control of Random Walks and Birth and Death Processes," *Adv. Appl. Prob.* to appear.
11. A. Veinott, "Discrete Dynamic Programming with Sensitive Discount Optimality Criteria," *Ann. Math. Statist.* **40,** 1635–1660 (1969).
12. D. Widder, *The Laplace Transform*, Princeton University Press, Princeton, N.J., 1941.