DENUMERABLE STATE MARKOVIAN DECISION PROCESSES—AVERAGE COST CRITERION[1]

By Cyrus Derman

Columbia University and Stanford University

1. Introduction. We are concerned with the optimal control of certain types of dynamic systems. We assume such a system is observed periodically at times $t = 0, 1, 2, \cdots$. After each observation the system is classified into one of a possible number of states. Let I denote the space of possible states. We assume I to be denumerable. After each classification one of a possible number of decisions is made. Let K_i denote the number of possible decisions when the system is in state i, $i \varepsilon I$. The decisions interact with the chance environment in the evolution of the system.

Let $\{Y_t\}$ and $\{\Delta_t\}$, $t = 0, 1, \cdots$, denote the sequences of states and decisions. A basic assumption concerning the type of systems under consideration is that

$$P\{Y_{t+1} = j \mid Y_0, \Delta_0, \cdots, Y_t = i, \Delta_t = k\} = q_{ij}(k),$$

for every i, j, k and t; i.e., the transition probabilities from one state to another are functions only of the last observed state and the subsequently made decision. It is assumed that the $q_{ij}(k)$'s are known.

A *rule* or *policy* R for controlling the system is a set of functions $\{D_k(Y_0, \Delta_0, \cdots, Y_t)\}$ satisfying $0 \leqq D_k(Y_0, \Delta_0, \cdots, Y_t) \leqq 1$, for every k, and $\sum_{k=1}^{K_i} D_k(Y_0, \Delta_0, \cdots, Y_t = i) = 1$, for every history $Y_0, \Delta_0, \cdots, Y_t$ $(t = 0, 1, \cdots)$. As part of a controlling rule, $D_k(Y_0, \Delta_0, \cdots, Y_t)$ is the instruction at time t to make decision k with probability $D_k(Y_0, \Delta_0, \cdots, Y_t)$ if the particular history $Y_0, \Delta_0, \cdots, Y_t$ has occurred. We remark that although we have assumed a kind of Markovian property regarding the behavior of the system, the process $\{Y_t\}$, or even the joint process $\{Y_t, \Delta_t\}$, is not necessarily a Markov process; for a rule may or may not depend upon the complete history of the system.

We further assume that there is a known cost (or expected cost) w_{ik} incurred each time the system is in state i and decision k is made. Thus, we can define a sequence of random variables $\{W_t\}$, $t = 0, 1, 2, \cdots$ by $W_t = w_{ik}$ if $Y_t = i$, $\Delta_t = k$, $t = 0, 1, \cdots$. For a given $Y_0 = i$ and rule R we can talk about $E_R W_t$, provided it exists. Let

$$Q_{T,R}(i) = (T + 1)^{-1} \sum_{t=0}^{T} E_R W_t, \quad \text{when} \quad Y_0 = i;$$

thus, $Q_{T,R}(i)$ is the expected average cost per unit time up to time period T. Let $Q_R(i) = \lim_{T\to\infty} Q_{T,R}(i)$, if the limit exists; otherwise, let $Q_R(i) = \limsup_{T\to\infty} Q_{T,R}(i)$.

Received 3 March 1966.

[1] This work was supported in part by an Office of Naval Research contract (Nonr-225(53)-(NR-042-002)) at Stanford University.