A Markovian Decision Process

RICHARD BELLMAN

1. Introduction. The purpose of this paper is to discuss the asymptotic behavior of the sequence $\{f_N(i)\}$, $i = 1, 2, \cdots, M$, $N = 1, 2, \cdots$, generated by the non-linear recurrence relations

$$f_N(i) = \operatorname{Max}_{\mathbf{q}} \left[b_i(\mathbf{q}) + \sum_{j=1}^{M} a_{ij}(\mathbf{q}) f_{N-1}(j) \right], \qquad N = 1, 2, \cdots, \tag{1}$$

$$f_0(i) = c_i, \qquad i = 1, 2, \cdots, M.$$

Although these equations are non-linear, they possess certain quasi-linear properties which permit a more thorough discussion than might be imagined upon first glance.

As we shall discuss below, this question arises from the consideration of a dynamic programming process. A related process gave rise to an equation of the above form which was discussed in [1], under a particular set of assumptions concerning the functions $b_i(\mathbf{q})$ and the matrices $A(\mathbf{q}) = (a_{ij}(\mathbf{q}))$. Here we shall impose restrictions of a quite different type. Any complete discussion of relations of the foregoing type is at least as detailed as a corresponding discussion of the linear case, and, as in the linear case, the assumptions made determine the techniques employed to a considerable extent.

We shall discuss elsewhere some interesting quadratically non-linear recurrence relations which arise from specializing the forms of $b_i(\mathbf{q})$ and $a_{ij}(\mathbf{q})$. These are related to the types of differential equations discussed in [3], being essentially a particular type of discrete version.

It will be clear from what follows that similar techniques can be utilized to treat the determination of the asymptotic behavior of the sequences defined by

$$f_N(x) = \operatorname{Max}_{\mathbf{q}} \left[b(x, \mathbf{q}) + \int_0^1 K(x, y, \mathbf{q}) f_{N-1}(y)\, dy \right], \qquad N = 1, 2, \cdots, \tag{2}$$

$$f_0(x) = c(x), \qquad 0 \leqq x \leqq 1,$$

and by the equation

$$u_N = \operatorname{Max}_{\mathbf{q}} \left[b(\mathbf{q}) + \sum_{i=1}^{k} a_i(\mathbf{q}) u_{N-i} \right], \qquad N = k+1, k+2, \cdots, \tag{3}$$

$$u_i = c_i, \qquad i = 0, 1, \cdots, k,$$

under corresponding assumptions.

Journal of Mathematics and Mechanics, Vol. 6, No. 5 (1957).

2. Statement of Results. We shall suppose that the functions $b_i(\mathbf{q})$ and $a_{ij}(\mathbf{q})$ satisfy either of the following sets of conditions:

(1)
A. The functions $b_i(\mathbf{q})$ and $a_{ij}(\mathbf{q})$ are functions of finite dimensional vectors $\mathbf{q}$ whose components assume only a finite set of values, which, in general, depend upon i and j.

B. The functions $b_i(\mathbf{q})$ and $a_{ij}(\mathbf{q})$ are continuous functions of finite dimensional vectors whose components assume values in certain closed, bounded regions in $\mathbf{q}$-space, which, in general, depend upon i and j.

Either of these sets of conditions ensures that the maximum is assumed in the recurrence relations of (1.1).

Our principal result is

Theorem. *Let us assume that either* (1A) *or* (1B) *is satisfied and that*

(2)
$$\text{a. } b_i(\mathbf{q}) \geqq 0 \quad \textit{and} \quad b_i(\mathbf{q}) > 0 \qquad \textit{for some } i \textit{ and all } \mathbf{q},$$
$$\text{b. } a_{ij}(\mathbf{q}) \geqq d > 0, \qquad i, j = 1, 2, \cdots, M, \qquad \textit{for all } \mathbf{q},$$
$$\text{c. } \sum_{i=1}^{M} a_{ij}(\mathbf{q}) = 1, \qquad i = 1, 2, \cdots, M.$$

In other words, $A(\mathbf{q})$ is for each $\mathbf{q}$ the transpose of a positive Markoff matrix.

Under these conditions, we have the asymptotic result

$$f_N(i) \sim Nr, \qquad N \to \infty, \qquad i = 1, 2, \cdots, M, \tag{3}$$

where the scalar quantity r is obtained as follows:

$$r\mathbf{1} = \operatorname*{Max}_{\mathbf{q}} \lim_{N\to\infty} \left[\frac{\mathbf{b}(\mathbf{q}) + A(\mathbf{q})\mathbf{b}(\mathbf{q}) + \cdots + A(\mathbf{q})^{N-1}\mathbf{b}(\mathbf{q})}{N} \right]. \tag{4}$$

Here

$$\mathbf{1} = \begin{bmatrix} 1 \\ 1 \\ \vdots \\ 1 \end{bmatrix}, \qquad \mathbf{b}(\mathbf{q}) = \begin{bmatrix} b_1(\mathbf{q}) \\ b_2(\mathbf{q}) \\ \vdots \\ b_N(\mathbf{q}) \end{bmatrix}, \qquad A(\mathbf{q}) = (a_{ij}(\mathbf{q})), \tag{5}$$

and maximization over $\mathbf{q}$ means maximization over all the vectors $\mathbf{q}$ appearing in all the relations of (1.1).

3. Preliminaries on Markoff Matrices. Before proceeding to the proof of this result, let us note for future reference some known results concerning the asymptotic behavior of the iterates of the transpose of a positive Markoff matrix.

If $\mathbf{y}$ is a non-negative non-trivial vector, we have

$$A^n\mathbf{y} \sim r\mathbf{1}, \tag{1}$$

where r is a scalar quantity dependent upon $\mathbf{y}$, and $\mathbf{1}$, as above, denotes the vector defined in (2.5). Furthermore, the limit

$$\lim_{n\to\infty} \sum_{k=1}^{n} A^k\mathbf{y} - nr\mathbf{1} = \mathbf{x} \tag{2}$$

exists and yields a vector $\mathbf{x}$ which satisfies the system of linear equations

$$r\mathbf{1} + \mathbf{x} = \mathbf{y} + A\mathbf{x}. \tag{3}$$

These results are well known from the theory of Markoff chains, or else may be viewed as simple consequences of PERRON's theorem asserting the existence of a positive characteristic root of largest absolute value of a positive matrix. The associated characteristic vector may be taken to be $\mathbf{1}$.

4. Proof of Theorem. We begin the proof of the theorem with a discussion of the linear system of (3.3). Let $\bar{\mathbf{q}}$ denote a value of $\mathbf{q}$ which maximizes the limit in (2.4). It is easy to show that the assumptions we have made concerning the range of values of $\mathbf{q}$ ensure the existence of a maximizing $\bar{\mathbf{q}}$. Let $r = r(\bar{\mathbf{q}})$ denote the maximum value of r determined by $\bar{\mathbf{q}}$, and let $\mathbf{x} = \mathbf{x}(\bar{\mathbf{q}})$ denote the vector determined by (3.2).

Then we have the system of equations

$$r + x_i = b_i(\bar{\mathbf{q}}) + \sum_{j=1}^{M} a_{ij}(\bar{\mathbf{q}})x_j\,, \qquad i = 1, 2, \cdots, M. \tag{1}$$

Actually, each $\bar{\mathbf{q}}$ above should be $\bar{\mathbf{q}}_i$, but we feel no confusion will result if we omit this subscript and use a generic $\bar{\mathbf{q}}$.

Our first task is to show that this linear system is equivalent to the non-linear system

$$r + x_i = \operatorname*{Max}_{\mathbf{q}} \left[b_i(\mathbf{q}) + \sum_{j=1}^{M} a_{ij}(\mathbf{q})x_j \right], \qquad i = 1, 2, \cdots, M, \tag{2}$$

in the sense that the set of x_i satisfying (1) also satisfies (2).

It is clear, to begin with, that

$$r + x_i \leqq \operatorname*{Max}_{\mathbf{q}} \left[b_i(\mathbf{q}) + \sum_{j=1}^{M} a_{ij}(\mathbf{q})x_j \right], \qquad i = 1, 2, \cdots, M. \tag{3}$$

If the x_i do not satisfy (2), there will be strict inequality in at least one of the relations in (3). Without loss of generality, let the strict inequality occur in the first relation. Finally, let $\mathbf{q}'$ be a value of $\mathbf{q}$ yielding the maximum on the right side of (2). Again we drop the subscripts in $\mathbf{q}'$ and use a generic symbol to simplify the notation.

We then have the inequalities

$$
\begin{aligned}
r + x_1 &< b_1(\mathbf{q}') + \sum_{j=1}^{M} a_{1j}(\mathbf{q}')x_j , \\
r + x_i &\leqq b_i(\mathbf{q}') + \sum_{j=1}^{M} a_{ij}(\mathbf{q}')x_j , \qquad i = 2, \cdots, M.
\end{aligned} \tag{4}
$$

The first inequality can be strengthened to read

$$
(1 + a)r + x_1 \leqq b_1(\mathbf{q}') + \sum_{j=1}^{M} a_{1j}(\mathbf{q}')x_j , \tag{5}
$$

where a is a positive quantity.

Let us now iterate these inequalities. We obtain

$$
\begin{aligned}
r + x_i \leqq b_i(\mathbf{q}') + \sum_{j=1}^{N} a_{ij}(\mathbf{q}')b_j(\mathbf{q}') + \sum_{j=1}^{N} a_{ij}^{(2)}(\mathbf{q}')x_j \\
- r - ara_{i1}(\mathbf{q}'), \qquad i = 1, 2, \cdots, M,
\end{aligned} \tag{6}
$$

where $(a_{ij}^{(2)}(\mathbf{q}')) = A(\mathbf{q}')^2$.

Since, by hypothesis, $a_{i1}(\mathbf{q}') \geqq d > 0$, we obtain, upon reverting to matrix notation,

$$
\mathbf{x} \leqq \mathbf{b}(\mathbf{q}') + A(\mathbf{q}')\mathbf{b}(\mathbf{q}') + A(\mathbf{q}')^2\mathbf{x} - r(2 + ad)\mathbf{1}. \tag{7}
$$

Let us now iterate this inequality N times. The result is

$$
\mathbf{x} \leqq \mathbf{b}(\mathbf{q}') + A(\mathbf{q}')\mathbf{b}(\mathbf{q}') + \cdots + A(\mathbf{q}')^{2N-1}\mathbf{b}(\mathbf{q}') + A(\mathbf{q}')^{2N}\mathbf{x} - N(2 + ad)\mathbf{1}, \tag{8}
$$

upon recalling that $A(\mathbf{q}')^2\mathbf{1} = \mathbf{1}$.

The maximal property of $r = r(\bar{\mathbf{q}})$ asserts that the vector

$$
\mathbf{b}(\mathbf{q}') + A(\mathbf{q}')\mathbf{b}(\mathbf{q}') + \cdots + A(\mathbf{q}')^{2N-1}\mathbf{b}(\mathbf{q}') - N(2 + ad)\mathbf{1} \tag{9}
$$

becomes an arbitrarily large negative vector as N increases. This contradicts (8) for sufficiently large N.

Hence (2) is equivalent to (1).

5. Proof of Theorem (Continued). It is now easy to complete the proof of the theorem. Let r and x_i be the quantities defined above. We wish to show that $f_N(i)$ satisfies the inequality

$$
Nr + x_i - k \leqq f_N(i) \leqq Nr + x_i + k \tag{1}
$$

for $i = 1, 2, \cdots, M$ and $N = 1, 2, \cdots$ with a suitable choice of k.

The proof is inductive. Choose k so that the inequalities hold for $N = 0$. Suppose that they are valid for $n = 0, 1, \cdots, N$. Then (1.1) yields

$$
\begin{aligned}
f_{N+1}(i) &\leqq \operatorname{Max}\left[b_i(\mathbf{q}) + \sum_{j=1}^{M} a_{ij}(\mathbf{q})[Nr + x_j + k]\right] \\
&\leqq Nr + k + \operatorname{Max}_{\mathbf{q}}\left[b_i(\mathbf{q}) + \sum_{j=1}^{M} a_{ij}(\mathbf{q})x_j\right] \qquad (2) \\
&\leqq Nr + k + r + x_i = (N+1)r + x_i + k.
\end{aligned}
$$

The lower bound is established in the same manner.

The inequalities in (1) above yield the desired asymptotic behavior, and even a more precise result.

6. Discussion. As in the theory of Markoff processes, the condition $a_{ij}(\mathbf{q}) \geqq d > 0$ can be considerably relaxed at the expense of more detailed discussion. However, as the study of Markoff processes show, it cannot be relaxed to mere non-negativity. The essential restriction is that the equations describe one interlinked system, rather than two independent systems arbitrarily considered as one system. The condition $a_{ij}(\mathbf{q}) \geqq d > 0$ is one way of ensuring this, but clearly there are many others. The simplest, perhaps, are those obtained from powers of the matrix, *i.e.* $a_{ij}^{(k)}(\mathbf{q}) \geqq d > 0$, where $A(\mathbf{q})^k = (a_{ij}^{(k)}(\mathbf{q}))$.

7. A Dynamic Programming Process. Let us now briefly describe a dynamic programming process, [2], which gives rise to recurrence relations of the type considered above.

Consider a machine which is used repeatedly to produce a certain type of item. At each stage there is a probability that the machine produces a perfect item, a probability that it produces a defective item, and a probability that the machine breaks down and requires repair. These probabilities depend upon the age of the machine.

Examining the matter in more detail, let us suppose that there are k different sources of failure within the machine, leading to either defective items or breakdown of the machine or both. Let us define the following probabilities:

(1) $p_i(n)$ = probability that a machine breaks down due to failure at the i^{th} source after it has successfully produced n items,

$q_i(n)$ = probability that a defective item will be produced due to hidden failure at the i^{th} source after n items have been successfully produced.

At any particular stage, we face the problem of deciding whether to examine the machine for possible failure at one of the sources of trouble or to wait until a defective item is produced. In addition, if a defective item is produced, there is the question of whether we should repair the machine insofar as the immediate source of failure is concerned, whether we should in addition examine other potential sources of failure, or whether we should automatically provide new parts at various sources of failure, without preliminary inspection.

The decisions, of course, will be dependent upon the costs incurred in carrying

out these operations, the costs due to defective parts, and the costs due to breakdown of the machine.

The state of the system at any time can be characterized by the set of numbers $(n_1, n_2, \cdots, n_k)$ specifying the number of items n_i produced since the i^{th} source of trouble was examined.

The problem is then to determine the inspection and replacement policy which minimizes the expected unit cost of production.

To treat this problem, we begin with the problem of determining the policy which minimizes the expected cost of producing N items. Define the sequence of functions

$f_N(n_1, n_2, \cdots, n_k)$ = expected cost of producing N items using an optimal inspection and replacement policy starting in state $(n_1, n_2, \cdots, n_k)$.

In view of the above discussion, it follows that the effect of any decision, to produce without inspection or repair, to inspect with possible replacement, or to replace, is to transform the system from its present state into another state. Assume that only a finite number of states are permissible, and enumerate them in some order, $i = 1, 2, \cdots, M$. Any particular decision, designated by $\mathbf{q}$, leads to a recurrence relation

$$f_N(i) = b_i(\mathbf{q}) + \sum_{j=1}^{M} a_{ij}(\mathbf{q}) f_{N-1}(j). \tag{2}$$

The principle of optimality, *cf.* [2], asserts that $\mathbf{q}$ is chosen so as to yield the equation

$$f_N(i) = \operatorname*{Min}_{\mathbf{q}} \left[b_i(\mathbf{q}) + \sum_{j=1}^{M} a_{ij}(\mathbf{q}) f_{N-1}(j) \right]. \tag{3}$$

The theorem proved above in §2 states that there is a steady state optimal policy to which we converge as $N \to \infty$, provided that the $a_{ij}(\mathbf{q})$ satisfy certain restrictions.

As we have discussed above, the natural condition is that the system be interlinked, *i.e.* not separable into two distinct systems.

Particular examples of processes of the above general type are discussed in [4] and [5].

References

[1] Bellman, R., *On a quasi-linear equation*, Can. Jour. Math. **8** (1956) pp. 198–202.

[2] ———, *Dynamic Programming*, Princeton University Press, 1957.

[3] ———, *Functional equations in the theory of dynamic programming*, II. *Nonlinear differential equations*, Proc. Nat. Acad. Sci. **41** (1955) pp. 482–5.

[4] ———, *Equipment replacement*, Jour. Soc. Ind. Appl. Math. **3** (September, 1955) No. 3.

[5] Dreyfus, S., *A Note on an Industrial Replacement Process*, The Rand Corporation Paper P–1045, March, 1957.

The Rand Corporation
Santa Monica, California