MANAGEMENT SCIENCE
Vol. 24, No. 5, January 1978
Printed in U.S.A

Notes* II

A NOTE ON DYNAMIC PROGRAMMING WITH UNBOUNDED REWARDS*†

J. A. E. E. VAN NUNEN‡ AND J. WESSELS§

In a recent paper, Lippman presents sufficient conditions for Denardo's N-stage contraction in discounted semi-Markov decision processes with unbounded rewards. In this note it is demonstrated that Lippman's conditions may be replaced by weaker conditions which even imply 1-stage contraction. The verification of the conditions of this note is somewhat easier.

Lippman [4] considers a discounted semi-Markov decision process with general state space S and action space A. Generalizing his approach of an earlier paper [3], he presents sufficient conditions for the existence of a normed Banach space of real valued functions on S in which Denardo's N-stage contraction approach [1] may be used. This approach allows the rewards of the Markov decision process to be unbounded to a certain extent. This is made possible by the specific choice of the norm in the Banach space. Lippman uses weighted supremum norms which were later studied more generally for Markov decision processes by Wessels in [7]. Another method of handling certain types of unbounded rewards has been introduced by Harrison [2]. His idea of a shifted space has been combined with the weighted supremum norm approach by van Nunen in [5].

In Lippman's notation $q(\cdot \mid x, a)$, $r(x, a)$ denote the transition probability and one period reward, respectively, for state $x \in S$ and action $a \in A$; $\alpha > 0$ is the discount factor; $t(\cdot \mid x, a)$ is the probability distribution function of the time until the next transition (given state $x \in S$, action $a \in A$). Also, a policy f maps S into A.

The conditions in [4] are the following:

A function w on S exists with $w(x) \geqslant 1$, an integer $m \geqslant 1$ exists, a number β $(0 \leqslant \beta < 1)$ exists, positive numbers b and M exist, such that for all $x \in S$, $a \in A$:

$$\beta(x, a) \equiv \int_0^\infty e^{-\alpha\tau} t(d\tau \mid x, a) \leqslant \beta,$$

$$|r(x, a)| w(x)^{-m} \leqslant M,$$

$$\int_S w^n(y) q(dy \mid x, a) \leqslant [w(x) + b]^n \quad \text{for } n = 1, \ldots, m.$$

In this note it will be shown that these conditions may be replaced by somewhat weaker and simpler conditions:

THEOREM. *Under Lippman's conditions the following holds: A function v on S exists with $v(x) > 0$, a number β $(0 \leqslant \beta < 1)$ exists, a number ρ $(\beta < \rho < 1)$ exists, a positive*

* All Notes are refereed.

† Accepted by Bennett L. Fox; received September 17, 1975. This paper has been with the authors 3 months, for 1 revision.

‡ Graduate School of Management, Delft.

§ Eindhoven University of Technology.

number M exists, such that for all $x \in S$, $a \in A$:

$$\beta(x, a) \equiv \int_0^\infty e^{-\alpha\tau} t(d\tau \mid x, a) \leqslant \beta,$$

$$|r(x, a)| v(x)^{-1} \leqslant M,$$

$$\beta \int_S v(y) q(dy \mid x, a) \leqslant \rho v(x).$$

Moreover, the existence of a function v and numbers β, ρ, M with the properties mentioned above already guarantees that the operator T_f defined by

$$(T_f u)(x) \equiv r(x, f(x)) + \beta(x, f(x)) \int_S u(y) q(dy \mid x, f(x))$$

is a one-stage contraction on the Banach space of real-valued functions on S with norm

$$\|u\|_v = \sup_{x \in S} |u(x)| v(x)^{-1}.$$

Lippman's Banach space consists of real valued functions u and S with the following weighted supremum norm:

$$\|u\| \equiv \sup_x |u(x)| w(x)^{-m}.$$

In [4] it is proved that under these conditions there exists an integer $J \geqslant 1$, such that for any sequence of policies $f_1, \ldots, f_j$ the operator $T_{f_1} \cdots T_{f_j}$ is a contraction.

The proof of the theorem follows from the two lemmas below.

LEMMA 1. *Under Lippman's conditions the following holds: for any $\rho > \beta$ and any $c \geqslant b[(\rho/\beta)^{1/m} - 1]^{-1}$ the positive function v on S with*

$$v(x) \equiv [w(x) + c]^m$$

satisfies

$$\beta \int_S v(y) q(dy \mid x, a) \leqslant \rho v(x) \quad \textit{for all } x \in S, a \in A.$$

PROOF. Note that c satisfies $b + c \leqslant (\rho/\beta)^{1/m} c$. Hence

$$\begin{aligned}
\int_S v(y) q(dy \mid x, a) &= \int_S [w(y) + c]^m q(dy \mid x, a) \\
&= \sum_{n=0}^m \binom{m}{n} c^{m-n} \int_S w^n(y) q(dy \mid x, a) \\
&\leqslant \sum_{n=0}^m \binom{m}{n} c^{m-n} [w(x) + b]^n = [w(x) + b + c]^m \\
&\leqslant [w(x) + (\rho/\beta)^{1/m} c]^m \leqslant \rho v(x)/\beta.
\end{aligned}$$

LEMMA 2. *Under Lippman's conditions the following holds: for any ρ ($\beta < \rho < 1$) there exists a function v on S with $v(x) > 0$, such that for any policy f*

$$\|T_f u_1 - T_f u_2\|_v \leqslant \rho \|u_1 - u_2\|_v,$$

$$\|r_f\|_v \leqslant M,$$

where $r_f(x) \equiv r(x, f(x))$.

PROOF. Choose c and v as in lemma 1. Then

$$|(T_f u_1 - T_f u_2)(x)| \leqslant \beta \int_S |u_1(y) - u_2(y)| q(dy \mid x, f(x))$$

$$\leqslant \beta \|u_1 - u_2\|_v \int_S v(y) q(dy \mid x, f(x))$$

$$\leqslant \rho \|u_1 - u_2\|_v v(x).$$

Furthermore: $|r(x, a)| v(x)^{-1} \leqslant |r(x, a)| w(x)^{-m} \leqslant M$.

Lemma 2 proves in fact our theorem, namely, if our conditions are satisfied T_f is a ρ-contraction with respect to the norm $\|\cdot\|_v$ and $\|r_f\|_v \leqslant M$. Note that this ρ-contraction holds with respect to a norm which differs slightly from Lippman's. In fact, we replace his weight function $w(x)^m$ by $[w(x) + c]^m$. If Lippman's conditions hold with $b = 0$, then we may choose $c = 0$ and both norms are identical. The contraction factor ρ can be chosen arbitrarily close to β by choosing c sufficiently large. For computational purposes (especially finding upper and lower bounds for the value function) it is more favorable to have this one-stage ρ-contraction, than to have a J-stage contraction with "average" contraction per stage $\beta(1 + Jb)^{m/J}$ for a J with $\beta^J(1 + Jb)^m < 1$ (see [4]).

REMARKS. (1) As demonstrated in [7], the discounting requirement is not essential in our analysis: if we replace $\beta(x, a)q(\cdot \mid x, a)$ by $p(\cdot \mid x, a)$ then our conditions become:

$$|r(x, a)| v(x)^{-1} \leqslant M < \infty,$$

$$\int_S v(y) p(dy \mid x, a) \leqslant \rho v(x) \quad \text{with } \rho < 1.$$

These conditions allow the situation $\alpha = 0$ in certain cases and give some weakening for $\alpha > 0$.

(2) In [4] it is essential that $w(x) \geqslant \delta$ for some positive δ. This implies for $v(x)$ in lemma 1: $v(x) \geqslant (\delta + c)^m$. However, in order to make T_f contracting with respect to $\|\cdot\|_v$ such a condition is not required for v. It suffices if $v(x) > 0$ for all x. For treating unbounded rewards, the possibility of having v-values approaching zero is not essential. However, for finding v-values satisfying the second requirement in Remark 1, it may be of help to have this extra possibility.

(3) In this paper we showed for a special situation in Markov decision processes that J-stage contraction with respect to some weighted supremum norm for the relevant operator implies one-stage contraction with respect to some other weighted supremum norm. In fact this can be proved more generally (see e.g. [5]), although the general form of the new weight function is less simple. In Markov decision processes the most important operator is $Uu \equiv \sup_f T_f u$ and the contraction properties of T_f are used for proving contraction of U. However, any operator U that is J-stage contracting with respect to some norm ρ in a metric space is one-stage contracting with respect to another norm in a larger metric space (both norms need not be weighted supremum norms and even if ρ is a weighted supremum norm, the new one can be of a different type). This fact has been proved by Walter in [6] by choosing for the new metric

$$\sigma(u_1, u_2) \equiv \rho(u_1, u_2) + \alpha^{-1}\rho(Uu_1, Uu_2) + \cdots + \alpha^{-(J-1)}\rho(U^{J-1}u_1, U^{J-1}u_2),$$

if U is J-stage contracting with contraction factor α^J.

In a personal communication B. L. Fox has suggested another choice for a new

metric:

$$\lambda(u_1, u_2) \equiv \max\left[\rho(u_1, u_2), \alpha^{-1}\rho(Uu_1, Uu_2), \ldots, \alpha^{-(J-1)}\rho(U^{J-1}u_1, U^{J-1}u_2)\right],$$

which guarantees 1-stage contraction if U is J-stage contracting. The idea of Fox can be used for a still more general result:

LEMMA 3. *Suppose U is an operator in some metric space with metric ρ. Suppose that for some $\alpha > 0$ and all u_1, u_2*

$$\sup_{n=0, 1, \ldots} \alpha^{-n}\rho(U^n u_1, U^n u_2) < \infty.$$

Then U is a one-stage contraction with contraction factor α with respect to the metric

$$\nu(u_1, u_2) \equiv \sup_{n=0, 1, 2, \ldots} \alpha^{-n}\rho(U^n u_1, U^n u_2).$$

(4) Our third condition involves only one inequality instead of m inequalities as Lippman requires. On the other hand, it seems that our third condition lacks the extra degree of freedom that Lippman's constant b might provide. However, it is easily verified that with each weighting function $v(x)$ also $v(x) + d$ (with d any positive constant) satisfies the inequality.

(5) In order to show how the weighting function $v(x)$ can be determined in a realistic problem we consider as in Lippman [4] the $M/G/1$ queue with removable server in which the system is controlled by turning the server on or off. The arriving rate is $\lambda > 0$ and the service times are nonnegative random variables with distribution function G, and mean μ, where $0 < \mu < 1/\lambda$. The cost structure includes four types of costs: a holding cost $h(n)$ depending on the number of customers in the system, a running cost r per unit of time the server is on and a fixed cost R_1 [R_2] for turning the server on [off]. As in Lippman [4] the states of the system are $\langle n, i\rangle$, where n is the number of customers ($n = 0, 1, 2, \ldots$) and i is 0 or 1 when the server is off or on respectively. The law of motion is given by the mass function q where

$$q(\langle n + 1, 0\rangle \mid \langle n, i\rangle, 0) = 1$$

and

$$q(\langle n + j - \delta_n, 1\rangle \mid \langle n, i\rangle, 1) = \int_0^\infty \frac{e^{-\lambda\xi}(\lambda\xi)^j}{j!}\, dG(\xi),$$

with $\delta_0 = 0$ and $\delta_n = 1$ for $n \geqslant 1$.

For the case of linear holding costs we choose $v(\langle n, i\rangle) \equiv n + c$, where c should be chosen such that

$$(n + 1 + c)q(\langle n + 1, 0\rangle \mid \langle n, i\rangle, 0) \leqslant \rho(n + c)/\beta \quad \text{for } n = 0, 1, \ldots$$

and

$$\sum_{j=0}^{\infty} [n + j - \delta_n + c]q(\langle n + j - \delta_n, 1\rangle \mid \langle n, i\rangle, 1) \leqslant \rho(n + c)/\beta \quad \text{for } n = 0, 1, \ldots .$$

Noting that the greatest lower bound for c occurs when $n = 0$ in the first inequality above, we have $c \geqslant (\rho/\beta - 1)^{-1}$.

For the case of quadratic holding costs we choose $v(\langle n, i\rangle) \equiv n^2 + c$, where c should be chosen such that

$$((n + 1)^2 + c)q(\langle n + 1, 0\rangle \mid \langle n, i\rangle, 0) \leqslant \rho(n^2 + c)/\beta \quad \text{for } n = 0, 1, \ldots$$

and

$$\sum_{j=0}^{\infty} \left[(n+j-\delta_n)^2 + c\right] q(\langle n+j-\delta_n, 1\rangle \mid \langle n, i\rangle, 1)$$

$$\leqslant \rho(n^2 + c)/\beta \quad \text{for } n = 0, 1, \ldots$$

These requirements lead easily to the condition

$$c \geqslant \max\left\{ \frac{2-\rho/\beta}{(\rho/\beta - 1)^2}, \frac{\lambda^2\mu_2 + \lambda\mu}{\rho/\beta - 1} \right\},$$

where μ_2 is the second moment of G. In both inequalities $n = 0$ yields the critical value.

In an inventory problem with backlogging our conditions give a nice condition for the tails of the demand distribution and for the backlogging costs. It appears that exponential weight functions can be used for a large class of demand distributions and for strongly increasing costs (see [5] for details).[1]

[1] The authors wish to thank an anonymous associate editor and a referee for some helpful remarks. They are grateful to the area editor Professor B. L. Fox for several comments, especially for calling Walter's paper to their attention and for the suggestion of λ in remark 3.

References

1. DENARDO, E. V., "Contraction Mappings in the Theory Underlying Dynamic Programming," *SIAM Review*, Vol. 9 (1967), pp. 165–177.
2. HARRISON, J. M., "Discrete Dynamic Programming with Unbounded Rewards," *Ann. Math. Statist.*, Vol. 43 (1972), pp. 636–644.
3. LIPPMAN, S. A., "Semi-Markov Decision Processes with Unbounded Rewards," *Management Science*, Vol. 19 (1973), pp. 717–731.
4. ———, "On Dynamic Programming with Unbounded Rewards," *Management Science*, Vol. 21 (1975), pp. 1225–1233.
5. VAN NUNEN, J. A. E. E., "Contracting Markov Decision Processes," *Mathematical Centre Tract* 71, Amsterdam 1976.
6. WALTER, W., "A Note on Contraction," *SIAM Review*, Vol. 18 (1976), pp. 107–111.
7. WESSELS, J., "Markov Programming by Successive Approximations with Respect to Weighted Supremum Norms," *J. Math. Anal. Appl.*, Vol. 58 (1977), pp. 326–335.