Memorandum COSOR 83 - 03

Note on the existence of uniformly nearly-optimal stationary strategies in negative dynamic programming

by

Jan van der Wal

Eindhoven, the Netherlands

January 1983
NOTE ON THE EXISTENCE OF UNIFORMLY NEARLY-
OPTIMAL STATIONARY STRATEGIES IN NEGATIVE
DYNAMIC PROGRAMMING

by

Jan van der Wal

Abstract.
This note relaxes a condition given by Demko and Hill for the existence of a uniformly nearly-optimal stationary strategy in negative dynamic programming.
1. Introduction

In DEMKO and HILL [1] it has been shown, among other things, that a sufficient condition for the existence of a uniformly nearly-optimal stationary strategy in the negative dynamic programming model is that the rewards do not depend on the action but on the state only and are strictly negative. This note shows that the condition can be relaxed to action-dependent rewards as long as per state they are bounded away from zero.

So let us consider a dynamic system with countable state space $S$, for convenience say $S = \{1, 2, 3, \ldots \}$ and arbitrary action space $A$ endowed with some $\sigma$-field $A$ containing all one-point sets. If in state $i \in S$ an action $a \in A$ is taken, two things happen: a negative reward $r(i, a)$ is earned and a transition is made to state $j$, $j \in S$, with probability $p(i, a, j)$, where $\sum_{j} p(i, a, j) = 1$. The functions $r(i, \cdot)$ and $p(i, \cdot, j)$ are assumed to be $A$-measurable.

Note that Demko and Hill consider a general state space with discrete transition probabilities. The results to be obtained here seem to hold in their setting as well.

Strategies are defined in the usual way. Each strategy, $\pi$ say, and initial state $i$ define a probability measure $P_{i, \pi}$ on $(S \times A)^\infty$ and a stochastic process $\{ (X_n, A_n), n = 0, 1, \ldots \}$, where $X_n$ is the state at time $n$ and $A_n$ the action chosen at time $n$. Expectations with respect to $P_{i, \pi}$ will be denoted by $E_{i, \pi}$.

\footnote{1) We use the usual negative dynamic programming set up maximizing negative payoffs instead of minimizing costs as in [1].}
For each initial state $i \in S$ and strategy $\pi$ the total expected reward is defined by

$$v(i,\pi) := \mathbb{E}_{i,\pi} \sum_{n=0}^{\infty} r(X_n, A_n).$$

The value of the problem is denoted by $v^*$, so

$$v^*(i) = \sup_{\pi} v(i,\pi).$$

The condition used by Demko and Hill is

$$r(i, a) < 0 \quad \text{for all } a \in A \text{ and } i \in S.$$  

We will relax (3) to

\underline{Condition BAFO.} (Bounded away from zero.)

For all $i \in S$ there is a number $r(i)$ such that

$$r(i, a) \leq r(i) < 0 \quad \text{for all } a \in A.$$  

The result to be proved in Section 2 is

\underline{Theorem 1.}

Under condition BAFO there exists for each $\varepsilon > 0$ a stationary strategy, $\pi$ say, satisfying

$$v(i,\pi) \geq v^*(i) - \varepsilon \quad \text{for all } i \in S.$$  

Finally Section 3 contains a minor extension of Theorem 1.
2. The proof of Theorem 1

The proof is essentially the one given by Demko and Hill and consists of a series of lemmas.

The first lemma (cf. Lemma 3.2 in DEMKO and HILL [1]) states that good strategies pay only a limited number of visits to each state.

So, define \( N(i,j,n) \) to be the expected number of visits to state \( j \) if the initial state is \( i \) and strategy \( n \) is used. Then we have

Lemma 1.

Let \( \pi \) be \( \varepsilon \)-optimal for initial state \( i \), i.e. \( v(i,\pi) \geq v^*(i) - \varepsilon \), then

\[
N(i,j,\pi) \leq \frac{v^*(j) - \varepsilon}{r(j)}.
\]

Proof. Let \( \tau \) be the first entry time to state \( j \) with \( \tau = 0 \) if \( j = 1 \). Then

\[
v^*(i) - \varepsilon \leq v(i,\pi) = E_{i,\pi} \left[ \sum_{n=0}^{\tau-1} r(X_n, A_n) + \sum_{n=\tau}^{\infty} r(X_n, A_n) \right]
\]

\[
\leq E_{i,\pi} \left[ \sum_{n=0}^{\tau-1} r(X_n, A_n) + N(i,j,\pi) r(j) \right]
\]

\[
\leq E_{i,\pi} \left[ \sum_{n=0}^{\tau-1} r(X_n, A_n) + v^*(X_{\tau}) \right] + N(i,j,\pi) r(j) - v^*(j)
\]

\[
\leq v^*(i) + N(i,j,\pi) r(j) - v^*(j).
\]

Hence

\[
N(i,j,\pi) \leq \frac{v^*(j) - \varepsilon}{r(j)}.
\]
The next lemma states that a specific sufficiently small perturbation of the rewards has little influence on the value of the problem.

**Lemma 2** (cf. Proposition 3.1 in DEMKO and HILL [1]).

Choose \( \varepsilon > 0 \) and define

\[
\tilde{r}(i,a) := r(i,a) - d(i) \quad \text{for all } i \text{ and } a,
\]

with

\[
d(i) := \frac{r(i)}{v^*(i) - 1} \varepsilon 2^{-i}, \quad i \in S.
\]

Then

\[
v^*(i) - \varepsilon \leq \tilde{v}^*(i) \leq v^*(i) \quad \text{for all } i \in S.
\]

(All objects concerning the perturbed problem are labeled by a tilde.)

**Proof.** Clearly \( \tilde{v}^*(i) \leq v^*(i) \), so let us consider the other inequality.

Let \( \pi \) be some \( \delta \)-optimal strategy for initial state \( i \) in the original problem with \( \delta \leq 1 \), then by Lemma 1

\[
\tilde{v}(i,\pi) = v(i,\pi) - \sum_j N(i,j,\pi)d(j)
\]

\[
\geq v(i,\pi) - \sum_j v^*(j) - \delta \cdot \frac{r(j)}{v^*(j) - 1} \varepsilon 2^{-j}
\]

\[
\geq v(i,\pi) - \sum_j \varepsilon 2^{-j} = v(i,\pi) - \varepsilon.
\]

So

\[
\tilde{v}(i,\pi) \geq v^*(i) - \varepsilon - \delta.
\]

Since \( \delta \) can be chosen arbitrarily small also

\[
\tilde{v}^*(i) \geq v^*(i) - \varepsilon.
\]

\[ \Box \]
Now we can prove the following result:

**Lemma 3.**

Let \( f \) be a policy satisfying for all \( i \in S \)

\[
(4) \quad r(i,f(i)) + \sum_j p(i,f(i),j) \tilde{v}^*(j) \geq \tilde{v}^*(i),
\]

with \( \tilde{v}^* \) the value of the perturbed problem of Lemma 2.

(That such a policy \( f \) exists is immediate from

\[
\sup_a \left\{ r(i,a) + \sum_j p(i,a,j) \tilde{v}^*(j) \right\} = \tilde{v}^*(i).
\]

Then

\[
v(i,f) \geq v^*(i) - \varepsilon \quad \text{for all } i \in S.
\]

**Proof.** From (4) we have \( v(i,f) \geq \tilde{v}^*(i) \) for all \( i \in S \), since \( v(\cdot,f) \) is the largest nonpositive solution \( v \) of

\[
r(i,f(i)) + \sum_j p(i,f(i),j) v(j) \geq v(i),
\]

a well-known result in negative dynamic programming (see e.g. Van der Wal [2, Theorem 2.18]).

Hence, with Lemma 2

\[
v(i,f) \geq v^*(i) - \varepsilon \quad \text{for all } i \in S.
\]

This establishes Theorem 1, thus condition BAFO is indeed sufficient for the existence of uniformly nearly-optimal stationary strategies.
3. An extension

The requirement that condition BAFO should hold in all states can be relaxed a little.

Theorem 2.
If in each state either BAFO holds or a conserving action \(^1\) exists then a uniformly nearly-optimal stationary strategy exists.

Proof. In those states where BAFO does not hold we restrict the action set to a singleton containing one conserving action. Let \(v'\) denote the value of the problem with the restricted action sets, \(A(i)\) say, \((A(i)\) is either a singleton of \(A)\) then clearly \(v' \leq v^*\). But \(v^*\) still solves

\[
\sup_{a \in A(i)} \left\{ r(i,a) + \sum_j p(i,a,j) v(j) \right\} = v(i) \quad \text{for all } i, \tag{5}
\]

so, since \(v'\) is the largest nonpositive solution of (5), also \(v' \geq v^*\). Hence \(v' = v^*\). Thus the restricted problem is essentially equivalent to the original one. Now let us embed the restricted problem on the set of states where BAFO does hold. This does not change the problem either. For the embedded problem, however, by Theorem 1 a uniformly nearly-optimal stationary strategy exists. Combining this stationary strategy with the conserving actions fixed before gives us a uniformly nearly-optimal stationary strategy for the original problem. \(\square\)

---

\(^1\) An action \(a\) in state \(i\) is called conserving if

\[
r(i,a) + \sum_j p(i,a,j) v^*(j) = v^*(i).\]
References
