Statistics, Probability and Game Theory
IMS Lecture Notes – Monograph Series (1996) Volume 30

BLACKWELL OPTIMAL POLICIES IN COUNTABLE DYNAMIC PROGRAMMING WITHOUT APERIODICITY ASSUMPTIONS

ALEXANDER A. YUSHKEVICH[1]
University of North Carolina at Charlotte

Abstract

The existence of stationary Blackwell optimal policies is proved in denumerable dynamic programming models satisfying compactness-continuity conditions and a uniform Doeblin condition. The latter is given in terms compatible with periodic chains.

1. Introduction and Summary. Dekker and Hordijk (1988), (1992) studied the existence of stationary Blackwell optimal policies in denumerable Markov decision chains with compact action sets and continuous in action transition probabilities and rewards. In the first of the cited works, their most complete result is proved under a uniform geometric convergence condition which excludes periodic chains. Recently Tijms (1994) developed an elegant way to treat the periodicity in connection with the average optimality equation: namely, to substitute the given controlled chain by a perturbed chain with a same geometric sitting time at all states. In this short note we show that Tijms' (1994) idea, combined with the Dekker and Hordijk's (1988) results, can be used in connection with the Blackwell optimality too. In the second of the cited publications, Dekker and Hordijk substitute the uniform geometric convergence condition by a uniform geometric recurrence condition which does not exclude periodic chains. However, the relation between the latter condition and the Tijms' condition we use here is not clear. It is a great pleasure to contribute this paper to a volume in the honor of David Blackwell, whom the author considers as his teacher in dynamic programming.

A dynamic programming model is determined by a state space X, action sets $A(x)$, a transition function $p(x, a, B)$, and a real-valued reward function $r(x, a)$, $a \in A(x)$, $x \in X$, $B \subset X$ (we omit measurability assumptions and other formalities in this preliminary paragraph). The selection of an initial state x and a policy π defines a probability distribution $\mathbf{P}_x^\pi$ in the space of sequences $x_0 a_1 x_1 a_2 \ldots$ of consecutively visited states x_t and actions

[1]Supported by NSF Grant DMS-9404177.

$a_{t+1} \in A(x_t)$ applied at them; the corresponding expectation is $\mathbf{E}_x^\pi$. Denote by Π the set of all (in general, randomized and history dependent) policies, and by Φ the set of all stationary (non-randomized) policies. Elements of Φ are identified with functions $\varphi : x \to a = \varphi(x) \in A(x)$, $x \in X$. Under suitable convergence conditions, for every discount factor $0 < \beta < 1$ the expected total discounted reward

$$v(x,\pi) = \mathbf{E}_x^\pi \sum_{t=0}^{\infty} \beta^t r(x_t, a_{t+1}), \qquad x \in X, \pi \in \Pi, \tag{1}$$

is well-defined, and the following definition makes sense.

DEFINITION 1. Given a class $\Pi' \subset \Pi$ of policies, a policy π^* is said to be *Blackwell optimal within* Π' if for every $x \in X$ and $\pi \in \Pi$ there exists some $\beta_0(x,\pi) < 1$ such that

$$v_\beta(x,\pi^*) \geq v_\beta(x,\pi), \qquad \beta_0(x,\pi) < \beta < 1. \tag{2}$$

In the case $\Pi' = \Pi$, $\pi^\star$ is called *Blackwell optimal.*

This definition, up to mentioning Π', is due to Dekker and Hordijk (1988), (1992), whose works remain the main sources on Blackwell optimality in the denumerable models. If one may choose $\beta_0 < 1$ independent of x and π (as it is in the case of finite sets X and $A(x)$ treated in the basic Blackwell's (1962) work), then (2) reduces to the original Blackwell's definition of optimality:

$$v_\beta(x,\pi^*) = v_\beta(x), \qquad x \in X, \beta_0 < \beta < 1 \tag{3}$$

where $v_\beta(x) = \sup_{\pi\in\Pi} v_\beta(x,\pi)$, $x \in X$ is the value function of the model. In the case of an infinite space X, the definition (3) appears to be too strong to work with. For a further discussion of concepts and more survey we refer the reader to Dekker and Hordijk (1988) and Yushkevich (1994).

Let $\|\cdot\|$ be the supremum norm. We make the following

ASSUMPTION 1. (i) X is a denumerable space, and the transition function is given by transition probabilities $p(x,a,y) \geq 0$ satisfying the condition $\sum_y p(x,a,y) = 1$ $(x \in X,\ a \in A(x),\ y \in X)$;

(ii) $A(x)$ are non-empty compact sets in a Polish space $(x \in X)$;

(iii) $p(x,a,y)$ and $r(x,a)$ are continuous in a $(x \in X,\ y \in X)$;

(iv) $\| r \| < \infty$.

Parts (i) - (iii) are standard in the denumerable dynamic programming, when dealing with sensitive criteria. As to part (iv), Dekker and Hordijk (1988), (1992) consider a more general case of rewards bounded with respect to some positive weight function $\mu(x) > 0$, $x \in X : \| \frac{r}{\mu} \| < \infty$. This leads to additional conditions, such as convergence and continuity in a of the sums of series $\sum_y p(x,a,y)\mu(y)$ etc. We restrict ourselves to the case $\mu = 1$.

Dekker and Hordijk's (1988) work contains a scrupulous analysis of various steps leading to stationary Blackwell optimal policies via the Laurent expansions technique developed for finite models by Veinott (1969). As to the concluding results, the most complete one is the existence of such policies under a *uniform geometric convergence* condition, which we formulate for $\mu = 1$. Given $\varphi \in \Phi$, let $p^{\varphi}(x,y) = p(x,\varphi(x),y)$, $x \in X$, $y \in X$. The matrix $P^{\varphi} = (p^{\varphi}(x,y))$ is the transition matrix of the corresponding Markov chain on X. Consider its t-step transition probabilities $p_t^{\varphi}(x,y)$, $t = 1,2,\ldots$ As in every countable Markov chain, when $t \to \infty$, the matrix $(P^{\varphi})^t = (p_t^{\varphi}(x,y))$ converges, at least in the Cesáro sense, to a limiting (sub)stochastic matrix $\overline{P}^{\varphi} = (\overline{p}^{\varphi}(x,y))$.

CONDITION UGC. There are constants $c > 0$ and $\gamma \in (0,1)$ such that uniform in $x \in X$ and $\varphi \in \Phi$

$$\sum_{y \in X} |p_t^{\varphi}(x,y) - \overline{p}^{\varphi}(x,y)| \leq c\gamma^t, \qquad t = 1,2,\ldots$$

This condition excludes periodic chains (but not a decomposition into several ergodic classes). On the page 409 of their work Dekker and Hordijk (1988) mention that the above condition "can be altered to include periodic chains by introducing an initial distribution but we will not elaborate this here". As far as we know, this approach was not elaborated elsewhere. We make the following assumption, equivalent to Condition C1 in Tijms (1994).

ASSUMPTION 2. There exist a number $\varepsilon > 0$ and an integer $T > 0$ such that for every $\varphi \in \Phi$ one may find a state $y^{\varphi} \in X$ with

$$p_1^{\varphi}(x,y^{\varphi}) + p_2^{\varphi}(x,y^{\varphi}) + \cdots + p_T^{\varphi}(x,y^{\varphi}) \geq \varepsilon, \quad x \in X. \tag{4}$$

This assumption does not exclude periodicity (although it excludes more than one ergodic class). We prove the following

THEOREM 1. *Under Assumptions* 1 *and* 2 *there exists a stationary Blackwell optimal policy.*

The *uniform geometric recurrence* condition introduced in Dekker and Hordijk (1992) requires, in the case $\mu = 1$, the existence of a fixed finite set M of states such that for all initial states $x \in X$ and all policies $\varphi \in \Phi$ the probability not to hit M in t steps decays exponentially fast as $t \to \infty$, uniform in x and φ. Tijms (1994) makes a conjecture about the equivalence of a so called *simultaneous Doeblin condition* (which is essentially the same as

the uniform geometric recurrence condition), and his Condition C1 (equivalent to Assumption 2). As far as we know, this still remains to be only a conjecture.

2. Proofs. In the first two lemmas we summarize some results following more or less directly from Dekker and Hordijk (1988).

LEMMA 1. *Assumption* 1 *and Condition* UGC *imply that:*
(i) *for every* $\varphi \in \Phi$ *a Laurent series expansion*

$$v_\beta(x,\varphi) = (1+\rho)\sum_{n=-1}^{\infty} h_n^\varphi(x)\rho^n, \qquad x \in X,\ 0 < \rho = \frac{1-\beta}{\beta} < \rho_0 \tag{5}$$

holds with $|h_n^\varphi| \le C^{n+2}(n \ge -1)$, *where* $\rho_0 > 0$ *and* $C < \infty$ *are the same for all* x *and* φ;

(ii) *if* $\varphi_k \in \Phi\,(k \ge 1)$ *and the limit* $\varphi(x) = \lim_{k\to\infty} \varphi_k(x)$ *exists for every* $x \in X$, *then* $\varphi \in \Phi$ *and*

$$h_n^\varphi(x) = \lim_{k\to\infty} h_n^{\varphi_k}(x), \qquad x \in X,\ n \ge -1;$$

(iii) *there exists a policy* φ^* *Blackwell optimal within* Φ.

PROOF. Part (i) follows from *ibid.*, Theorems 4.8 to 4.1, with bounds for h_n^φ obtained as in Theorem 4.4 (the continuity of $P^\varphi\mu$ in φ and other conditions besides UGC present in those theorems become trivial in the case $\mu = 1$). Part (ii) follows from the continuity in φ of the deviation matrix D^φ (*ibid.*, Theorem 4.8) and of the function

$$r^\varphi(x) = r(x,\varphi(x)), \qquad x \in X,$$

from the resulting continuity in φ of $\overline{P}^\varphi r^\varphi$ and $(D^\varphi)^n r^\varphi$ (*ibid.*, Lemmas 4.5 and 4.6), and from formulas relating h_n^φ with $\overline{P}^\varphi, D^\varphi$ and r^φ (*ibid.*, (4.5)). To get part (iii), follow along *ibid.*, Theorems 4.8 to 4.7 to 3.1 to 3.2. □

LEMMA 2. *Assumption* 1 *and parts* (i), (ii) *of Lemma* 1 *imply that a policy* φ^* *Blackwell optimal within* Φ *is also Blackwell optimal within* Π.

Proof. See *ibid*, Theorem 5.4, with two adjustments. Firstly, instead of the continuity of D^φ we assume parts (i)-(ii) of Lemma 1. One may see that the above continuity is used only to obtain what is stated in Lemma 1 (i, ii). Secondly, Theorem 5.4 states only that

$$\liminf_{\rho\downarrow 0} \frac{v_\beta(x,\varphi^*) - v_\beta(x,\pi)}{\rho^n} \ge 0, \qquad n \ge -1,\ x \in X,\ \pi \in \Pi,$$

i.e. that φ^* is n-discount optimal for each $n = -1, 0, 1, 2, \ldots$ in the terminology of Veinott (1969). This is weaker than the Blackwell optimality

(2), because the Laurent expansion (5) is not known for $v_\beta(x,\pi)$. In fact, as shown for another model in Yushkevich (1994), in the case of bounds for h_n^φ given in Lemma 1(i), essentially the same reasoning gives for φ^* the relation (2) too. □

Next, following Tijms (1994), we select a number $0 < b < 1$ and consider a *perturbed model* with the same X, $A(x)$ and r but with the transition probabilities changed to

$$q(x,a,y) = bp(x,a,y) + (1-b)\delta(x,y), \quad x \in X,\ a \in A(x),\ y \in X \tag{6}$$

(δ is the Kronecker symbol). Let $q^\varphi(x,y)$, $q_t^\varphi(x,y)$, $\overline{q}^\varphi(x,y)$, Q^φ and $w_\beta(x,\pi)$ have the same meaning in the perturbed model as $p^\varphi(x,y)$, $p_t^\varphi(x,y)$, $\overline{p}^\varphi(x,y)$, P^φ and $v_\beta(x,\pi)$ do in the original model. Then

$$Q^\varphi = bP^\varphi + (1-b)I, \qquad \varphi \in \Phi \tag{7}$$

where I is the identity matrix. The following lemma is a key one.

LEMMA 3. *For linear operators P with $\| P \| \leq 1$ (acting in some Banach space), let the resolvent $R_\rho(P)$ be defined by*

$$R_\rho(P) = (I - \frac{P}{1+\rho})^{-1} = \sum_{t=0}^{\infty} \left(\frac{P}{1+\rho}\right)^t, \qquad \rho > 0 \tag{8}$$

(I is the identity operator). Then

$$R_\rho(P) = \frac{b+b\rho}{1+b\rho} R_{b\rho}(bP + (1-b)I), \qquad \rho > 0,\ 0 < b < 1. \tag{9}$$

PROOF. Using (8) and (9), verify that

$$\left(I - \frac{bP + (1-b)I}{1+b\rho}\right) R_{b\rho}(bP + (1-b)I) = I. \quad \square$$

LEMMA 4. *Assumptions* 1 *and* 2 *for the original model imply Assumption* 1 *and Condition* UGC *for the perturbed model.*

PROOF. Assumption 1 is evident. As to Condition UGC, by (6) we have $q^\varphi(x,y) \geq bp^\varphi(x,y)$ for every x, y and φ, and hence, by an induction in t,

$$q_t^\varphi(x,y) \geq b^t p_t^\varphi(x,y), \qquad x \in X,\ y \in X,\ t = 1,2,\ldots,\ \varphi \in \Phi.$$

In a similar way $q^\varphi(x,x) \geq 1-b$, and hence

$$q_t^\varphi(x,x) \geq (1-b)^t, \qquad x \in X,\ t = 1,2,\ldots,\ \varphi \in \Phi.$$

Therefore for every $t = 1, 2, \dots, T$

$$q_T^\varphi(x,y) \geq q_t^\varphi(x,y) q_{T-t}^\varphi(y,y) \geq b^t(1-b)^{T-t} p_t^\varphi(x,y) \geq b^T(1-b)^T p_t^\varphi(x,y).$$

Now a summation over t and (4) yield the following *uniform Doeblin condition* for the perturbed model:

$$q_T^\varphi(x, y^\varphi) \geq \alpha = \frac{\varepsilon}{T} b^T(1-b)^T, \qquad x \in X,\ \varphi \in \Phi. \tag{10}$$

Condition (10) implies the existence of the limits $\overline{q}^\varphi(y) = \lim_{t\to\infty} q_t^\varphi(x,y)$ $(x, y \in X)$ and uniform in x, φ and $B \subset X$ geometric bounds

$$\left| \sum_{y\in B} [q_t^\varphi(x,y) - \overline{q}^\varphi(y)] \right| \leq (1-\alpha)^{\frac{t}{T}-1}, \qquad t = 1, 2, \dots \tag{11}$$

(see, for example, Doob (1953), page 197, case (b) (take Doob's measure φ on X equal to 1 at the state y^φ)). Another reference is Meyn and Tweedy (1993), p. 384, Theorem 16.02, part (v), formula (16.11). Condition UGC for the perturbed model follows from (11), with $\gamma = \sqrt[T]{1-\alpha}$. □

PROOF OF THEOREM 1. In the case of a stationary policy φ, formula (1) for $v_\beta(x,\varphi)$ reduces, in the vector notations, to

$$v_\beta(\varphi) = \sum_{t=0}^{\infty} (\beta P^\varphi)^t r^\varphi, \qquad 0 < \beta < 1,\ \varphi \in \Phi,$$

or, in the resolvent notations (8), to

$$v_\beta(\varphi) = R_\rho(P^\varphi) r^\varphi, \qquad \beta = \frac{1}{1+\rho},\ \rho > 0,\ \varphi \in \Phi.$$

According to (7), in the perturbed model

$$w_\beta(\varphi) = R_\rho(bP^\varphi + (1-b)I) r^\varphi.$$

Therefore by Lemma 3

$$\frac{v_\beta(\varphi)}{1+\rho} = b\frac{w_\delta(\varphi)}{1+b\rho}, \qquad \beta = \frac{1}{1+\rho},\ \delta = \frac{1}{1+b\rho},\ \rho > 0,\ \varphi \in \Phi. \tag{12}$$

It follows from (12) and Definition 1 that a policy Blackwell optimal within Φ in the perturbed model is also Blackwell optimal within Φ in the original model. By Lemmas 4 and 1, there is such a policy φ^*, and moreover, parts (i) - (ii) of Lemma 1 also hold in the perturbed model. In particular, for every $\varphi \in \Phi$

$$w_\delta(\varphi) = (1+b\rho) \sum_{n=-1}^{\infty} k_n^\varphi(x) (\rho b)^n, \qquad 0 < b\rho < \rho_0',\ \varphi \in \Phi \tag{13}$$

with $\| k_n^\varphi \| \leq (C')^{n+2}$ etc. A comparison of (5), (12) and (13) shows that in the original model the relation (5) also holds, with

$$h_n^\varphi(x) = b^{n+1} k_n^\varphi(x), \qquad x \in X,\ \varphi \in \Phi,\ n \geq -1,$$

and with $\rho_0 = \frac{\rho_0'}{b}$, $C = \frac{C'}{b}$. Since h_n^φ differ from k_n^φ by constant coefficients only, part (ii) of Lemma 1 extends from the perturbed to the original model. By Lemma 2, φ^* is Blackwell optimal within Π in the original model. □

REFERENCES

BLACKWELL, DAVID (1962). Discrete dynamic programming. *Ann. Math. Stat.* **33,** 19-726.

DEKKER, R. AND HORDIJK, A. (1988). Average, sensitive and Blackwell optimal policies in denumerable Markov decision chains with unbounded rewards. *Math. Oper. Res.* **13** 395-420.

DEKKER, R. AND HORDIJK, A. (1992). Recurrence conditions for average and Blackwell optimality in denumerable state Markov decision chains. *Math.Oper.Res.* **17**, 271-289.

DOOB, J.L. (1953). *Stochastic Processes.*Wiley, New York.

MEYN, S. P. AND TWEEDY, R. L. (1993). *Markov Chains and Stochastic Stability.* Springer, New York.

TIJMS, H. C. (1994). Average reward optimality equation in Markov decision processes with a general state space. In *Probability, Statistics and Optimization: A Tribute to Peter Whittle* (ed. Kelley, F. P.), Wiley, New York, 485-495.

VEINOTT, A. F. JR. (1969). Discrete dynamic programming with sensitive optimality criteria. *Ann. Math. Stat.* **40**, 1635-1660.

YUSHKEVICH, A. A. (1994). Blackwell optimal policies in a Markov decision process with a Borel state space. *Z. Oper. Res.* **40**, 253-288.

DEPARTMENT OF MATHEMATICS
UNIV. OF NORTH CAROLINA AT CHARLOTTE
CHARLOTTE, NC 28223
fma00aay@unccvm.uncc.edu