A Survey of
Blackwell optimality

Arie Hordijk
University of Leiden, Mathematical Institute,
P.O. Box 9512, 2300 RA Leiden, The Netherlands
E-mail: hordijk@math.leidenuniv.nl

Alexander A. Yushkevich
University of North Carolina at Charlotte,
Department of Mathematics, Charlotte, NC 28223, USA
E-mail: aayushke@email.uncc.edu

1. Finite models

In this introductory section we consider Blackwell optimality in Controlled Markov Processes (CMPs) with finite state and action spaces; for brevity,we call them finite models. We introduce the basic definitions, the Laurent-expansion technique, the lexicographical policy improvement, and the Blackwell optimality equation, which were developed at the early stage of the study of sensitive criteria in CMPs. We also mention some extensions and generalizations obtained afterwards for the case of a finite state space.

1.1 Definition and existence of Blackwell optimal policies. We consider an infinite horizon CMP with a finite state space $\mathbb{X}$, a finite action space $\mathbb{A}$, action sets $\mathbb{A}(x) = A_x$, transition probabilities $p_{xy}(a) = p(y|x,a)$, and reward function $r(x,a)$ $(x \in X, a \in A_x, y \in X)$. Let m be the number of states in $\mathbb{X}$.

We refer to Chapter 1 for definitions of various policies, of probability distributions and expectations corresponding to them, and notations. We also use the notation

$$\mathbb{K} = \{(x,a):\ a \in \mathbb{A}(x),\ x \in \mathbb{X}\}, \tag{1}$$

so that, in particular,

$$P^a f(x) = \sum_{y \in \mathbb{X}} p_{xy}(a) f(y), \qquad (x,a) \in \mathbb{K}. \tag{2}$$

For every discount factor $\beta \in (0,1)$ the expected total reward

$$v(x,\pi,\beta) = v_\beta(x,\pi) := \mathbb{E}_x^\pi \left[\sum_{t=0}^{\infty} \beta^t r(x_t, a_t) \right] \tag{3}$$

converges absolutely and uniformly in the initial state x and policy π, so that the value function

$$V(x,\beta) = V_\beta(x) := \sup_{\pi \in \Pi} v_\beta(x,\pi), \qquad x \in X$$

is well defined and finite. Following Blackwell [3], in this chapter we say that a policy π is β-*optimal* if $v_\beta(x,\pi) = V_\beta(x)$ for all $x \in X$ (not to confuse with ϵ-optimal policies, for which $v_\beta(\pi) \geq V_\beta - \epsilon$; in this chapter we do not use them).

In the case of a stationary policy $\varphi \in \Pi^s$ it is convenient to write (3) in matrix notations. In that case we have an $m \times m$ transition matrix $P(\varphi) = P^\varphi$ with entries $p_{xy}(\varphi(x)) = p^\varphi_{xy}$, and (3) can be written in the form

$$v_\beta(\varphi) = v^\varphi_\beta = \sum_{t=0}^{\infty} (\beta P^\varphi)^t r^\varphi = (I - \beta P^\varphi)^{-1} r^\varphi \tag{4}$$

where r^φ is a vector with entries $r(x, \varphi(x))$, $x \in \mathbb{X}$ (formula (4) has sense also for complex β with $|\beta| < 1$), (in the notation (2) $P^\varphi f(x) = P^{\varphi(x)} f(x)$). For every $\beta \in (0, 1)$ there exists a β-optimal policy $\varphi_\beta \in \Pi^s$; namely, one may set

$$\varphi_\beta(x) = \operatorname*{argmax}_{a \in A_x} \Big[r(x, a) + \beta \sum_{y \in X} P^a v_\beta(x) \Big], \qquad x \in \mathbb{X}.$$

In the important case of undiscounted rewards, when $\beta = 1$, the total expected reward in general diverges, and the simplest performance measure is the average expected reward $w(x, \pi) = w^\pi(x)$ (see Chapter 1). For a stationary policy φ

$$w^\varphi = \lim_{n \to \infty} \frac{1}{n} \sum_{t=0}^{n-1} (P^\varphi)^t r^\varphi = Q^\varphi r^\varphi = \lim_{\beta \uparrow 1} (1 - \beta) v_\beta^\varphi, \tag{5}$$

where $Q^\varphi = Q(\varphi)$ is the *stationary matrix*

$$Q^\varphi = \lim_{N \to \infty} \frac{1}{N+1} \sum_{t=0}^{N} (P^\varphi)^t, \quad \text{and} \quad Q^\varphi P^\varphi = P^\varphi Q^\varphi = Q^\varphi. \tag{6}$$

The last expression for w^φ in (5) follows from (3) and the fact that Cesaro summability of a divergent series implies its Abel summability to the same limit. Howard [26] proved the existence of average optimal policies in finite CMPs with the class Π^s of admissible policies, and developed a policy improvement algorithm to find them, related with his name. Almost at the same time Wagner [41] determined that such policies are average optimal in the class Π too.

However, the average reward criterion is unsensitive, underselective since it is entirely determined by the arbitrarily far tail of the rewards; in accordance with this criterion, two policies providing rewards $100 + 0 + 0 + \cdots$ and $0 + 0 + 0 + \cdots$ are equally good (or bad). Blackwell [3] in his study of finite CMPs introduced a much more sensitive concept of optimality, that bears now his name, and proved the existence of stationary policies optimal in this new sense.

Definition 1. *A policy π is said to be Blackwell optimal, if π is β-optimal for all values of β in an interval $\beta_0 < \beta < 1$.*

A stationary Blackwell optimal policy φ is average optimal. Indeed, there exists a stationary average optimal policy ψ, and by (5)

$$w^\psi = \lim_{\beta \uparrow 1} (1 - \beta) v_\beta^\psi \le \lim_{\beta \uparrow 1} (1 - \beta) v_\beta^\varphi = w^\varphi,$$

so that $w^\varphi = w^\psi$. Since the last limit is the same for all Blackwell optimal policies, stationary or not (as follows from Definition 1), and since by Theorem 2 below there is a stationary Blackwell optimal policy, *every Blackwell optimal policy $\pi \in \Pi$ is average optimal.*

Theorem 2 *In finite CMP there exists a stationary Blackwell optimal policy.*

Proof. Since for every positive $\beta < 1$ there exists a β-optimal policy $\varphi_\beta \in \Pi^s$, and because the set Π^s of stationary policies is finite together with $\mathbb{X}$ and $\mathbb{A}$, there exists a stationary policy φ which is β-optimal for all $\beta = \beta_n$ where $\beta_n \uparrow 1$. We claim that φ is Blackwell optimal.

Suppose the contrary. Then, because $\mathbb{X}$ and Π^s are finite sets, there are a state x_0, a policy $\psi \in \Pi^s$, and a sequence $\gamma_n \uparrow 1$ such that

$$v_\beta^\varphi(x_0) < v_\beta^\psi(x_0) \quad \text{for} \quad \beta = \gamma_n, \quad \gamma_n \uparrow 1.$$

On the other hand, by the selection of φ

$$v_\beta^\varphi(x_0) \geq v_\beta^\psi(x_0) \quad \text{for} \quad \beta = \beta_n \uparrow 1.$$

It follows that the function

$$f(\beta) = v_\beta^\varphi(x_0) - v_\beta^\psi(x_0)$$

defined for all complex β with $|\beta| < 1$ takes on the value 0 at an infinite sequence of different points $z_n \uparrow 1$, and takes on nonzero values at the points $\gamma_n \uparrow 1$.

By using Cramer's rule to compute the inverse matrix, we find that each entry of $(I - \beta P^\varphi)^{-1}$ is a rational function of β, and the same is true with ψ in place of φ. Therefore and by (4), $f(\beta)$ is a rational function of the complex variable β in the circle $|\beta| < 1$ (and hence on the whole complex plane). A rational function cannot have infinitely many different zeros z_n if it is not an identical zero. The obtained contradiction proves that φ is Blackwell optimal.

□

The above proof is a purely existence argument, without any indication how to find a Blackwell optimal policy φ. Blackwell's original proof also did not provide a complete algorithm to obtain φ, but it contained some essential elements in this direction. Blackwell used, besides the limiting matrix Q^φ, the *deviation matrix* D^φ corresponding to $\varphi \in \Pi^s$. If the Markov chain with the transition matrix P^φ is aperiodic, then

$$D^\varphi = \sum_{t=0}^{\infty} [(P^\varphi)^t - Q^\varphi], \tag{7}$$

and the above series converges geometrically fast; in general

$$D^\varphi = \lim_{N\to\infty} \frac{1}{N+1} \sum_{n=0}^{N} \sum_{t=0}^{n} [(P^\varphi)^t - Q^\varphi]. \tag{7'}$$

An important property of this matrix is that D^φ is uniquely determined by the equations

$$D^\varphi Q^\varphi = Q^\varphi D^\varphi = 0, \tag{8}$$

$$D^\varphi (I - P^\varphi) = (I - P^\varphi) D^\varphi = I - Q^\varphi \tag{9}$$

(see, for instance, Kemeny and Snell [27]). Blackwell derived and utilized the expansion

$$v_\beta^\varphi = \frac{h_{-1}^\varphi}{1-\beta} + h_0^\varphi + o(1) \qquad \text{as} \quad \beta \uparrow 1, \tag{10}$$

where

$$h_{-1}^\varphi = Q^\varphi r^\varphi, \qquad h_0^\varphi = D^\varphi r^\varphi, \tag{11}$$

and introduced the notion of a *nearly optimal* policy $\pi \in \Pi$. For such a policy $V_\beta - v_\beta^\pi = o(1)$ as $\beta \uparrow 1$.

The existence of a Blackwell optimal policy $\varphi \in \Pi^s$ implies a similar expansion for the value function

$$V_\beta(x) = \frac{h_{-1}}{1-\beta} + h_0 + o(1) \qquad \text{as} \quad \beta \uparrow 1. \tag{12}$$

It is easy to see using (12), that a policy π is average optimal iff $v_\beta^\pi = h_{-1}/\alpha + o(1/\alpha)$, where $\alpha = 1-\beta$, and that π is nearly optimal iff $v_\beta^\pi = h_{-1}/\alpha + h_0 + o(1)$.

1.2. Laurent series expansions and n-discount optimality. Average optimal and nearly optimal policies, as well as relations (10)-(12), are at the start of a chain of notions and equations developed by Miller and Veinott [30] and Veinott[38], which lead to a deeper insight into Blackwell optimal policies and to an algorithm to find them. We present their main ideas in a slightly modified form.

The approach is based on the Laurent series expansion of the resolvent

$$R_\beta = (I - \beta P)^{-1} = I + \beta P + \beta^2 P^2 + \cdots \qquad (|\beta| < 1) \tag{13}$$

of a Markov chain with the transition kernel P in the neighborhood of the point $\beta = 1$. This expansion is a general fact known in functional analysis (see, for instance, [43]). In the particular case of an aperiodic Markov chain, it follows immediately from the geometric convergence of P^t to the limiting matrix Q. Indeed, the difference

$$R_\beta - \frac{1}{1-\beta} Q = (I - Q) + \beta(P - Q) + \beta^2(P^2 - Q) + \cdots,$$

in which $\|P^n - Q\| \le C\gamma^n$ for some $\gamma < 1$, is an analytic function of the complex variable β in the circle $|\beta| < 1/\gamma$ (we use the norm in the space of $(m \times m)$-matrices generated by the supremum norm in the space of m-vectors). The point $\beta = 1$ is inside this circle, thus R_β has the same singularity at the point $\beta = 1$ as $\frac{1}{1-\beta}Q$, i.e. has a single pole. Therefore in some ring $0 < |\beta - 1| < \alpha_0$ a Laurent expansion

$$R_\beta = \frac{R_{-1}}{\alpha} + R_0 + R_1\alpha + R_2\alpha^2 + \cdots, \qquad \alpha = 1 - \beta \tag{14}$$

holds. If the Markov chain is periodic, consider the least common multiple d of the periods of all its ergodic classes. The chain with a kernel P^d is then aperiodic, so that P^{nd} converges geometrically fast to a stochastic matrix $\widetilde{Q}$ as $n \to \infty$. Similar to the preceding argument, it follows that the infinite sum

$$\widetilde{R}_\beta = I + \beta^d P + \beta^{2d} P^2 + \cdots$$

is analytic in a circle $|\beta|^d < 1/\gamma$ of a radius greater than 1, and thus has a simple pole at $\beta = 1$. Then the same is true for

$$R_\beta = (I + \beta P + \cdots + \beta^{d-1} P^{d-1}) \widetilde{R}_\beta.$$

Instead of the Laurent series (14), one may write a similar series in powers of another small parameter ρ equivalent to α, which has the meaning of an interest rate:

$$\rho = \frac{1-\beta}{\beta} = \frac{\alpha}{1-\alpha}, \qquad \beta = \frac{1}{1+\rho} = 1 - \alpha. \tag{15}$$

Veinott [38] and most of the subsequent authors used series in ρ. Chitashvili [6][7][51] and following him Yushkevich [44]-[50] used series in α. We present both versions.

Theorem 3 *In a finite CMP there exists a number $\beta_0 \in (0, 1)$ such that for every policy $\varphi \in \Pi^s$*

$$v_\beta^\varphi = (1+\rho) \sum_{n=-1}^{\infty} h_n^\varphi \rho^n = \sum_{n=-1}^{\infty} k_n^\varphi \alpha^n, \qquad \beta_0 < \beta < 1 \tag{16}$$

where

$$h_{-1}^\varphi = k_{-1}^\varphi = Q^\varphi r^\varphi = w^\varphi, \qquad h_0^\varphi = k_0^\varphi = D^\varphi r^\varphi \tag{17}$$

(cf. (10) and (11)), and where for $n \geq 1$

$$h_n^\varphi = (-D^\varphi)^n h_0^\varphi, \qquad k_n^\varphi = (I - D^\varphi)^n k_0^\varphi. \tag{18}$$

A similar expansion is valid for the value function

$$V_\beta = (1+\rho) \sum_{n=-1}^{\infty} h_n \rho^n = \sum_{n=-1}^{\infty} k_n \alpha^n, \qquad \beta_0 < \beta < 1. \tag{19}$$

Proof. The existence and convergence of Laurent expansions (16) follow from expansions in powers of ρ or α of βR^φ_β, respectively R^φ_β, and from the formula $v^\varphi_\beta = R^\varphi_\beta r^\varphi$ equivalent to (4). To get the coefficients (17)-(18), observe that by (4) $v^\varphi_\beta = r^\varphi + \beta P^\varphi v^\varphi_\beta$, so that by (15) and (16)

$$(1+\rho)\sum_{-1}^{\infty} h^\varphi_n \rho^n = r^\varphi + P^\varphi \sum_{-1}^{\infty} h^\varphi_n \rho^n.$$

By the uniqueness of the coefficients of power series, this results in equations (to simplify writing, we temporarily skip the superscript φ):

$$h_{-1} = Ph_{-1}, \tag{20}$$

$$h_0 + h_{-1} = r + Ph_0, \tag{21}$$

$$h_n + h_{n-1} = Ph_n \qquad (n \geq 1). \tag{22}$$

From (6) and (20) by iteration and taking a limit, we find $h_{-1} = Qh_{-1}$. For the stationary matrix $Q = QP = PQ$, and a multiplication of (21) by Q gives $Qh_{-1} = Qr$, so that $h_{-1} = Qr$ as in (17). A multiplication of (22) by Q provides $Qh_n = 0$ $(n \geq 0)$. Using this, the relation $h_{-1} = Qh_{-1}$ and (8)-(9), we get after a multiplication of (21) by $D = D^\varphi$, that $D(I-P)h_0 + DQh_{-1} = Dr$, or $(I-Q)h_0 = Dr$, or finally $h_0 = Dr$ as in (17). Multiplying (22) by D, in a similar way we get $D(I-P)h_n + Dh_{n-1} = 0$, or $h_n - Qh_n = -Dh_{n-1}$, or $h_n = -Dh_{n-1}$ $(n \geq 1)$, and this proves that $h_n = (-D)^n h_0$ as in (18). Formulas (17)-(18) for k^φ_n follow absolutely similarly from equations $k_{-1} = Pk_{-1}$, $k_0 + Pk_{-1} = r + Pk_0$ and $k_n + Pk_{n-1} = Pk_n$ instead of (20)-(22).

Since the set Π^s is finite, we have the expansions (16) simultaneously for all $\varphi \in \Pi^s$ in some interval $(\beta_0, 1)$. Formula (19) follows now from Theorem 2. $\square$

Formulas of Theorem 3 are a generalization of (10) and (11). They stimulate a similar generalization of the average optimality and nearly optimality criteria. The following definition is due to Veinott [39].

Definition 4 *For $n \geq 1$, a policy $\pi^* \in \Pi$ is said to be n-discount optimal, if for every $\pi \in \Pi$*

$$\varliminf_{\beta \uparrow 1} \rho^{-n}[v_\beta(\pi^*) - v_\beta(\pi)] \geq 0 \tag{23}$$

(*with α in place of ρ we have an equivalent condition*).

By substituting in (23) a Blackwell optimal policy π, for which $v_\beta(\pi) = V_\beta$ and $v_\beta(\pi^*) - v_\beta(\pi) \leq 0$, one may see that in finite CMPs condition (23) is equivalent to a simpler (and formally stronger) condition

$$\lim_{\beta \uparrow 1} \rho^{-n}[V_\beta - v_\beta(\pi^*)] = 0. \tag{24}$$

However, condition (23) appeared to be more suitable for an extension of sensitive criteria to denumerable and Borelian CMPs. To avoid confusion, mention that in literature 0-discount optimal policies are sometimes called bias-optimal or 1-optimal; the latter name originates from Veinott [38]. Also, as seen from a comparison of (16) and (19), a stationary policy is Blackwell optimal iff it is n-discount optimal for every natural n, or, briefly speaking, is ∞-discount optimal.

A convenient description of n-discount optimal policies can be made in terms of sequences of coefficients of series (16) and (19) and a lexicographical ordering in spaces of them. Define

$$H^\varphi = \{h_{-1}^\varphi, h_0^\varphi, \ldots\}, \quad K^\varphi = \{k_{-1}^\varphi, k_0^\varphi, \ldots\}, \tag{25}$$

let H_n^φ and K_n^φ be the initial segments of H^φ and K^φ up to the n-th term, and let H, K, H_n and K_n have the same meaning for the series (19) (each h_n^φ etc. is an m-vector). For those sequences and segments we introduce a natural lexicographical ordering denoted by symbols $\succ$, $\succeq$, $\preceq$, $\prec$. So, $H^\varphi \prec H^\psi$ means that $H^\varphi \neq H^\psi$, and that there exists a number $N < \infty$ and a state $x_0 \in \mathbb{X}$, such that $H_{N-1}^\varphi = H_{N-1}^\psi$ (if $N \geq 0$), and $h_N^\varphi(x_0) < h_N^\psi(x_0)$ while $h_N^\varphi(x) \leq h_N^\psi(x)$ for all other $x \in \mathbb{X}$. The relation $H^\varphi \preceq H^\psi$ means that either $H^\varphi = H^\psi$ or $H^\varphi \prec H^\psi$. The relations $H^\psi \succ H^\varphi$ and $H^\psi \succeq H^\varphi$ are equivalent to $H^\varphi \prec H^\psi$ and $H^\varphi \preceq H^\psi$.

With this notation we have $H^\varphi \preceq H$ and $K^\varphi \preceq K$ for every $\varphi \in \Pi^s$, and the policy φ is n-discount optimal (or Blackwell optimal) iff $H_n^\varphi = H_n$ or $K_n^\varphi = K_n$ (respectively, if $H^\varphi = H$ or $K^\varphi = K$).

The following theorem due to Veinott [39] shows that in finite CMPs the n-th discount optimality of a stationary policy for large values of n coincides with its Blackwell optimality. Let Φ_n be the subset of Π^s consisting of all stationary n-discount optimal policies ($n \geq -1$), and let Φ_∞ be the set of all Blackwell optimal policies in Π^s. Evidently,

$$\Phi_{-1} \supset \Phi_0 \supset \Phi_1 \supset \cdots, \qquad \Phi_\infty = \bigcap_n \Phi_n.$$

Theorem 5 *In finite CMPs with $m \geq 2$ states*

$$\Phi_{m-1} = \Phi_m = \ldots = \Phi_\infty.$$

Proof. It is sufficient to show that $\Phi_{m-1} = \Phi_\infty$. Consider any policy $\varphi \in \Phi_{m-1}$. We have $H_{m-1}^\varphi = H_{m-1}$, or in more detail

$$h_n^\varphi = h_n, \qquad n = -1, 0, 1, \ldots, m-1. \tag{26}$$

Since $m \geq 2$, both h_0 and h_1 are present in (26). We claim that m column m-vectors $h_0, h_1, \ldots, h_{m-1}$ are linearly dependent. It is sufficient to show that m row vectors of the corresponding square matrix are linearly dependent; these rows are $\{h_0(x),\dots,h_{m-1}(x)\} = \{h_0^\varphi(x),\dots,h_{m-1}^\varphi(x)\}$, $x\in\mathbb{X}$. In fact even the infinite sequences

$$\{h_0^\varphi(x), h_1^\varphi(x),\dots,h_t^\varphi(x),\dots\}, \qquad x\in\mathbb{X} \tag{27}$$

are linearly dependent. Indeed, in the finite Markov chain generated by P^φ there exists a stationary distribution $\{\mu(x),\ x\in\mathbb{X}\}$. The total discounted expected reward corresponding to the initial distribution μ and policy φ is equal to

$$\begin{aligned} v_\beta^\varphi(\mu) &:= \sum_{x\in\mathbb{X}}\mu(x)v_\beta^\varphi(x) = \sum_{x\in\mathbb{X}}\mu(x)\,\mathbb{E}_x^\varphi\sum_{t=0}^\infty \beta^t r(x_t,\varphi(x_t)) = \\ &= \sum_{t=0}^\infty \beta^t \sum_{x\in\mathbb{X}}\mu(x)\,\mathbb{E}_x^\varphi\, r(x_t,\varphi(x_t)) = \sum_{t=0}^\infty \beta^t\,\mathbb{E}_\mu^\varphi\, r(x_t,\varphi(x_t)). \end{aligned} \tag{28}$$

Here the $\mathbb{P}_\mu^\varphi$-distribution of x_t does not depend on t because μ is a stationary distribution, and hence the factor at β^t in (28) is some constant C. Thus

$$v_\beta^\varphi(\mu) = C\sum_{t=0}^\infty \beta^t = \frac{C}{1-\beta} = C\frac{1+\rho}{\rho} = (1+\rho)\left[\frac{C}{\rho} + \sum_{n=0}^\infty 0\cdot\rho^n\right] \tag{29}$$

(cf. (15)). On the other hand, by (28) and (16),

$$v_\beta^\varphi(\mu) = (1+\rho)\sum_{n=-1}^\infty \rho^n \sum_{x\in\mathbb{X}}\mu(x)h_n^\varphi(x).$$

A comparison with (29) together with the uniqueness of the Laurent coefficients show that $\sum\limits_x \mu(x)h_n^\varphi(x) = 0$ for all $n\ge 0$, so that the sequences (27) are linearly dependent.

Now, by (18)

$$h_{n+1}^\varphi = -D^\varphi h_n^\varphi, \qquad h_{n+1} = -D^\psi h_n \qquad (n=0,1,2,\dots) \tag{30}$$

where ψ is a Blackwell optimal policy. Let t be the maximal integer such that the vectors $h_0 = h_0^\varphi,\dots,h_t = h_t^\varphi$ in (26) are linearly independent; such $t\ge 0$ exists if only $h_0\ne 0$, and as just proved, $t<m-1$. If $h_0 = 0$, then by (26) also $h_0^\varphi = 0$, and by (30) $h_n^\varphi = 0 = h_n$ for all $h\ge 0$, so that $H^\varphi = H$ and $\varphi\in\Phi_\infty$. If there is the required t, then $h_{t+1} = h_{t+1}^\varphi$ is a linear combination of $h_0 = h_0^\varphi,\dots,h_t = h_t^\varphi$:

$$h_{t+1} = \sum_{i=0}^t C_i h_i, \qquad h_{t+1}^\varphi = \sum_{i=0}^t C_i h_i^\varphi. \tag{31}$$

Due to (30) multiplying the first identity by $-D^{\psi}$ and the second by $-D^{\varphi}$, we only increase every subscript in (31) by 1, and since $h_i^{\varphi} = h_i$ for $0 \le i \le t+1$, we get $h_{t+2}^{\varphi} = h_{t+2}$. Repeating this, by induction we get $h_n^{\varphi} = h_n$ for every $n \ge 0$, so that $\varphi \in \Phi_{\infty}$. $\square$

The sets Φ_{m-2} and Φ_{m-1} are in general different. The following example, taken from [39], confirms this statement. To make it more visual, we present it for $m = 5$.

Example 6. There are $m = 5$ states $1, 2, \ldots, 5$ with mandatory transitions $2 \to 3 \to 4 \to 5$, the state 5 is absorbing. In the state 1 there is a choice between two actions, which determine two stationary policies φ and ψ (see Fig. 1). Under φ we have a mandatory transition $1 \to 2$, under ψ transitions $1 \to 1$ and $1 \to 2$ are equally likely. The numbers under arrows indicate the rewards $r(x, a)$. Mention that the rewards $2, -6, 6, -2$ are the binomial coefficients of $(A - B)^{m-2} = (A - B)^3$ multiplied by 2.

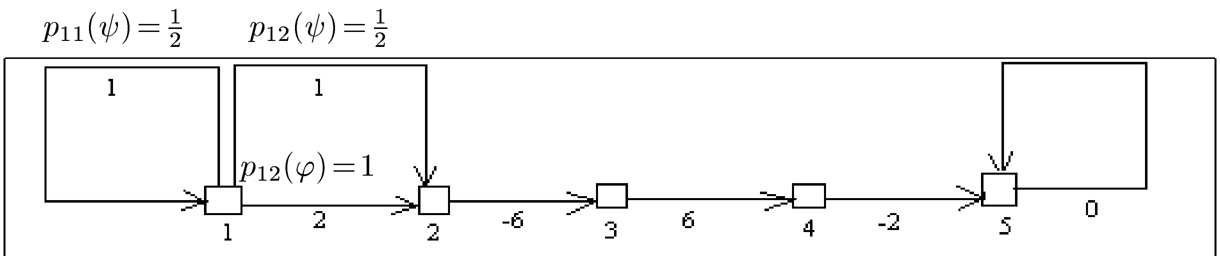

Figure 1.

The expected rewards v_β^{φ} and v_β^{ψ} differ only at the initial state 1. For φ we have

$$v_\beta^{\varphi}(1) = 2 - 6\beta + 6\beta^2 - 2\beta^3 = 2(1-\beta)^3 = 2\alpha^3.$$

For ψ, by the formula $v_\beta^{\psi} = r^{\psi} + \beta P^{\psi} v_\beta^{\psi}$ (cf. (4)) we have the equation

$$v_\beta^{\psi} = 1 + \beta[\tfrac{1}{2} v_\beta^{\psi}(1) + \tfrac{1}{2} v_\beta^{\psi}(2)].$$

Thus

$$(2 - \beta) v_\beta^{\psi}(1) = 2 + \beta v_\beta^{\psi}(2),$$

where

$$\beta v_\beta^{\psi}(2) = \beta(-6 + 6\beta - 2\beta^2) = v_\beta^{\varphi}(1) - 2.$$

Hence

$$v_\beta^{\psi}(1) = \frac{v_\beta^{\varphi}(1)}{2 - \beta} = \frac{2\alpha^3}{1 + \alpha} = 2\alpha^3 - 2\alpha^4 + 2\alpha^5 - \cdots.$$

This means that $V_\beta(1) = 2\alpha^3$, that φ is Blackwell optimal, and that ψ is 3-discount optimal, but not 4-discount optimal. Thus $\Phi_3 \neq \Phi_4 = \Phi_{\infty}$.

1.3. Lexicographical policy improvement and Blackwell optimality equation. Policy improvement is both a practical method to approach an optimal policy in CMPs and an important tool in their theory. Its essense is that if φ and ψ are two stationary policies, if $\pi = \psi\varphi^\infty$ is a Markov policy coinciding with ψ at the first step of the control and coinciding with φ afterwards, and if π is better than φ, then ψ is also better than φ. This method, almost trivial for the discounted reward criterion with a fixed $\beta < 1$, was developed by Howard [26] for the average reward criterion. Howard used, besides the avarage reward $w^\varphi (= h^\varphi_{-1})$, a bias function, in fact connected with the term h^φ_0 in the expansions (10) and (16). Blackwell [3] provided a rigorous proof that a slightly different version of Howard's policy improvement method does converge. Miller and Veinott [30] have extended policy improvement to the case of Blackwell optimality, and Veinott [39] refined it using the classes Φ_n. We expose this topic in a modernized form, using an operator approach developed in Dekker and Hordijk [8] in the framework of CMPs with a countable state space $\mathbb{X}$. To avoid additional formulas, we do all calculations in terms of ρ; in terms of α formulas are slightly different.

From the structure of π and (14) we have

$$v^\pi_\beta = r^\psi + \beta P^\psi v^\varphi_\beta = r^\psi + \frac{1}{1+\rho} P^\psi v^\varphi_\beta = r^\psi + P^\psi \sum_{n=-1}^{\infty} h^\varphi_n \rho^n,$$

while

$$v^\varphi_\beta = h^\varphi_{-1}\rho^{-1} + \sum_{n=0}^{\infty} (h^\varphi_n + h^\varphi_{n-1})\rho^n.$$

Subtracting, we get

$$v^\pi_\beta - v^\varphi_\beta = (P^\psi h_{-1} - h_{-1})\rho^{-1} + (r^\psi + P^\psi h_0 - h_0 - h_{-1}) + \sum_{n=1}^{\infty} (P^\psi h_n - h_n - h_{n-1})\rho^n \tag{32}$$

where it is understood that $h_n = h^\varphi_n$. By (18), the supremum norm $\|h^\varphi_n\|$ is growing no more than geometrically fast with n.

It is convenient to introduce the space $\mathfrak{H}$ of all sequences $H = \{h_n, n \geq -1\}$ of m-vectors satisfying this growth condition, and to treat the sequences of Laurent coefficients of the series (16), (32) etc. as elements of $\mathfrak{H}$. In particular $H^\varphi \in \mathfrak{H}$ (see (25)), and in $\mathfrak{H}$ we consider the same lexicographical ordering as we have introduced in connection with H^φ. Also, it is convenient to define the spaces $\mathfrak{H}_n$ of finite collections $H_n = \{h_t, -1 \leq t \leq n\}$ of m-vectors.

The right side of (32) defines an *operator* L^ψ in the spaces $\mathfrak{H}$ and $\mathfrak{H}_n$. Since the matrix P^ψ has entries $p_{xy}(a)$ with $a = \psi(x)$, we express L^ψ through the corresponding operators L^a transforming functions (vectors) on $\mathbb{X}$ into functions of pairs (x, a) on the state-action space $\mathbb{K}$ defined in (1). We have

$$(L^\psi H)(x) = L^{\psi(x)} H(x), \qquad x \in \mathbb{X}, \tag{33}$$

$$L^a H(x) = \{\ell h^a_{-1}(x), \ell h^a_0(x), \ell h^a_1(x), \ldots\}, \qquad (x, a) \in \mathbb{K}, \tag{34}$$

where according to (32)

$$
\begin{aligned}
\ell h^a_{-1}(x) &= P^a h_{-1}(x) - h_{-1}(x),\\
\ell h^a_0(x) &= r(x,a) + P^a h_0(x) - h_0(x) - h_{-1}(x),\\
\ell h^a_n(x) &= P^a h_n(x) - h_n(x) - h_{n-1}(x) \qquad (n \geq 1).
\end{aligned} \tag{35}
$$

The same formulas define L^a and L^ψ, as operators on $\mathfrak{H}_n$.

Lemma 7 *Let $\varphi, \psi \in \Pi^s$. If $(L^\psi H^\varphi)_{n+1} \succeq 0$ for some $n \geq -1$, then $H^\psi_n \succeq H^\varphi_n$. Moreover, if in addition $(L^\psi H^\varphi)_{n+1}(x_0) \succ 0$ at some $x_0 \in \mathbb{X}$, then $H^\psi_n(x_0) \succ H^\varphi_n(x_0)$. The same is true with the reverse inequality signs.*

In particular, if $L^\psi H^\varphi = 0$, then $H^\psi = H^\varphi$.

Proof. The condition $(L^\psi H^\varphi)_n \succeq 0$ means that

$$v^\pi_\beta = v^\varphi_\beta + Q_n(\rho) + O(\rho^{n+1}) \tag{36}$$

where $Q_n(\rho)$ is a vector consisting of polynomials of degree $\leq n$ with lexicographically nonnegative coefficients, and where $O(\rho^{n+1})$ is uniform in $x \in \mathbb{X}$ since $\mathbb{X}$ and $\mathbb{A}$ are finite sets (compare (32) with (33)-(35)). Consider policies $\pi_t = \psi^t \varphi^\infty$, and let $v(t) = v_\beta(\pi_t)$, so that, in particular, $v(0) = v^\varphi_\beta$, $v(1) = v^\pi_\beta$. We have

$$v(t+1) = r^\psi + \beta P^\psi v(t), \qquad t = 0, 1, 2, \ldots \tag{37}$$

and by (36)

$$v(1) = v(0) + Q_n + R, \tag{38}$$

where the remainder R is of order ρ^{n+1}. From (37) and (38) by induction we get

$$v(t) = v(0) + (I + \beta P^\psi + (\beta^2 P^\psi)^2 + \cdots + (\beta P^\psi)^{t-1})(Q_n + R), \qquad t \geq 1.$$

(we use that $r^\psi + \beta P^\psi v(0) = v(0) + Q_n + R$ according to (37) and (38)). Since $\beta < 1$, in the limit $v(t)$ becomes $v(\infty) = v^\psi_\beta$, so that

$$v^\psi_\beta = v^\varphi_\beta + \sum_{t=0}^{\infty} (\beta P^\psi)^t (Q_n(\rho) + R) = v^\varphi_\beta + R^\psi_\beta (Q_n + R).$$

Here $Q_n \geq 0$ for small $\rho > 0$, R is of order $O(\rho^{n+1})$, and the resolvent R^ψ_β is of order $O(1 + \beta + \beta^2 + \cdots) = O(\rho^{-1})$. This proves that $H^\psi_{n-1} \succeq H^\varphi_{n-1}$ if $L^\psi H^\varphi_n \succeq 0$. Other assertions are proved in a similar way. □

To proceed further, we need the lexicographical *Bellman operator* L in the spaces $\mathfrak{H}$ and $\mathfrak{H}_n$:

$$LH(x) = \max_{a \in \mathbb{A}_x} L^a H(x), \qquad H \in \mathfrak{H}, \quad x \in \mathbb{X}, \tag{39}$$

where the maximum is understood in the lexicographical sense $\succeq$; the same formula holds for $H_n \in \mathfrak{H}_n$. This maximum always exists because the sets $\mathbb{A}_x$ are finite. Since one may use all combinations of actions in stationary policies, formula

$$LH = \max_{\psi \in \Pi^s} L^\psi H \tag{40}$$

defines the same operator L.

If in (32) $\psi = \varphi$ then $\pi = \psi\varphi^\infty$ coincides with φ, and the left side of (32) is zero. Hence all the coefficients at the right side vanish, and this means that $L^\varphi H^\varphi = 0$ for every $\varphi \in \Pi^s$. Therefore $LH^\varphi \geq 0$ for every $\varphi \in \Pi^s$. If $LH^\varphi = 0$, we say that φ is *unimprovable*; if $LH_n^\varphi = 0$, then φ is *unimprovable of order* n. The equation

$$LH = 0 \qquad H \in \mathfrak{H} \tag{41}$$

is called the *Blackwell optimality equation* in the honor of Blackwell; the similar equation $LH_n = 0$ for $H_n \in \mathfrak{H}_n$ is the *n-order optimality equation*. Let $H = \{h_n\}$ be the element of $\mathfrak{H}$ corresponding to the value function V_β (see (19)), and let H_n be the initial segments of H. We say that a stationary policy φ is *conserving* (or n-order *conserving*) if $L^\varphi H = 0$ (respectively, $L^\varphi H_n = 0$).

Theorem 8 A. *The Blackwell optimality equation has a unique solution* $H^* = \max_{\varphi \in \Pi^s} H^\varphi$. *A policy* $\varphi \in \Pi^s$ *is Blackwell optimal iff* $H^\varphi = H^*$, *and iff* φ *is a conserving policy.*

B. *For every* $n \geq -1$, H_n^* *is uniquely determined by the equation* $LH_{n+1} = 0$. *A policy* $\varphi \in \Pi^s$ *is* n*-discount optimal iff* $H_n^\varphi = H_n^*$, *and is* n*-discount optimal if* φ *is* $(n+1)$*-order conserving.*

Proof. By Theorem 2 there exists a Blackwell optimal policy $\varphi \in \Pi^s$. Evidently, $H^\varphi \succeq H^\psi$, $\psi \in \Pi^s$ and φ is unimprovable, so that $H^\varphi = H^* := \max_{\psi \in \Pi^s} H^\psi$, and $LH^* = LH^\varphi = 0$. Since $L^\varphi H^\varphi = 0$, also $L^\varphi H^* = 0$, and φ is conserving. In the part A it remains to prove that the solution of (41) is unique, and that a conserving stationary policy is Blackwell optimal. If ψ is conserving, then $L^\psi H^* = 0$, hence $L^\psi H^\varphi = 0$ for a Blackwell optimal φ, therefore by Lemma 7 (applied to every n) $H^\psi = H^\varphi = H^*$, so that ψ is Blackwell optimal too. Finally, suppose that $\widetilde{H}$ is a solution to (41). By taking for each $x \in \mathbb{X}$ a lexicographical maximizer $a \in \mathbb{A}_x$ of $L^a \widetilde{H}(x)$, we obtain a stationary policy ψ for which $L^\psi \widetilde{H} = \widetilde{H}$. One may check (we omit the proof) that Lemma 7 is true for any $H \in \mathfrak{H}$ in place of H^φ, in particular, for $\widetilde{H}$. It follows that $H^\psi = \widetilde{H}$, and since $LH^\psi = L\widetilde{H} = 0$, the policy ψ is unimprovable. Hence ψ is Blackwell optimal, so that $\widetilde{H} = H^\psi = H^*$.

The proof of part B is similar, with a reference to Lemma 7. □

Policy improvement provides an algorithm to compute a Blackwell optimal policy in a finite CMP. Start with some $\varphi \in \Pi^s$ and compute H^φ_m using formulas of Theorem 3 (here m is the number of states in $\mathbb{X}$). Check the values of $\ell^a_{-1}h(x), (x,a) \in \mathbb{K}$. For $a = \varphi(x)$ those values are zeros, and if $\ell^{a^*}h_{-1}(x^*) > 0$ for some pair (x^*, a^*), then the policy

$$\psi(x) = \begin{cases} a^* & \text{if } x = x^*, \\ \varphi(x) & \text{otherwise} \end{cases}$$

improves φ. If there are no such pairs (x^*, a^*), repeat the same procedure with $\ell^a h_0$ and the shrinked sets $\mathbb{A}_0(x) = \{a \in \mathbb{A}_{-1}(x),\ \ell^a h_{-1}(x) = 0\}$, $\mathbb{K}_0 = \{(x,a) :\ a \in \mathbb{A}_0(x), x \in \mathbb{X}\}$ (where $\mathbb{A}_{-1}(x) = \mathbb{A}(x)$). A policy ψ as above with $\ell^{a^*}h_0(x^*) > 0$, $(x^*, a^*) \in \mathbb{K}$ improves φ. If there are no such pairs (x^*, a^*), repeat the procedure with all subscripts increased by 1, etc., until either you get a better policy ψ, or reach the set $\mathbb{K}_m$. In the latter case φ is $(m-1)$-order discount optimal, and therefore Blackwell optimal by Theorem 5. Otherwise, proceed in the same way with the obtained policy ψ. Since the set Π^s is finite, this algorithm leads to a Blackwell optimal policy in a finite number of steps. In practice, one may improve φ simultaneously at several states x^*.

On the other hand, the lexicographical policy improvement approach opens a new way to prove the existence of Blackwell optimal policies via a maximization of H^φ over all stationary policies φ and the related Blackwell optimality equation(41). The latter idea can be used in CMPs with an infinite state space $\mathbb{X}$, in which the proof of Theorem 2, based on the fact that the set Π^s is finite, is inapplicable.

1.4. Extensions and generalizations. Veinott [39] introduced also the notion of n-average optimality in addition to the n-discount optimality. Let

$$v^{(1)}_T(x,\pi) = \mathbb{E}^\pi_x\left[\sum_{t=0}^{T-1} r(x_t, a_t)\right]$$

and define recursively for $n \geq 1$

$$v^{(n+1)}_T(x,\pi) = \sum_{t=1}^{T} v^{(n)}_t(x,\pi).$$

Then π^* is n-average optimal if for every policy π

$$\varliminf_{T\to\infty} \frac{1}{T}\left[v^{(n+2)}_T(\pi^*) - v^{(n+2)}_T(\pi)\right] \geq 0.$$

Veinott[39],[40] and Sladky [34] showed that in a finite CMP a policy is n-discount optimal iff it is n-average optimal.

Chitashvili [6],[7],[51] extended results of Theorem 8 to more general models with a finite state space. In [6] he treated CMPs with arbitrary (indeed, compactified) action sets. He considered also what can be called (n, ϵ)-discount optimal policies; in their definition one should replace 0 by $-\epsilon$ in formula (23). In [7] he studied n-discount optimality in finite models with discount factors depending on the state x and action a: $\beta(x, a) = c_1\beta + c_2\beta^2 + \cdots + c_k\beta^k$ where k and c_i are functions of (x, a). In this case the reward functions were of some specific average form. In [51] Theorem 8 is generalized to a finite model with two reward functions $r(x, a)$ and $c(x, a)$. More precisely,

$$v_\beta^\varphi(x) = \mathbb{E}_x^\varphi \left[\sum_{t=0}^{\infty} \beta^t (r(x_t, a_t) + (1-\beta) c(x_t, a_t)) \right]$$

(in [51] Chitashvili considered only stationary policies). This expected discounted reward corresponds to an undiscounted reward

$$\sum_{t=0}^{\infty} r(x_t, a_t) + \lim_{T\to\infty} \frac{1}{T} \sum_{t=0}^{T-1} c(x_t, a_t).$$

In that case all formulas related to Theorem 8 remain valid, with one exception: in equations (35) defining L^a, the term ℓh_1 should be changed to

$$\ell h_1^a(x) = c(x, a) + P^a h_1(x) - h_1(x) - h_0(x)$$

(similar to the term ℓh_0).

As explained in the proof of Theorem 2, in finite CMPs the expected discounted reward $v_\beta^\varphi(x)$ is a rational function of β. Hordijk e.a. [18] introduced a non-Archimedian ordered field of rational functions, used a simplex method in this field, and developed a linear programming method for the computation of β-optimal policies over the entire range (0,1) of the discount factor. In particular, their method allows to compute a Blackwell optimal policy. More precisely, for some m one may find numbers $\beta_0 := 0 < \beta_1 < \cdots < \beta_{m-1} < \beta_m := 1$ and stationary policies $\varphi_1, \varphi_2, \ldots, \varphi_m$ such that φ_j is β-optimal for all $\beta \in [\beta_{j-1}, \beta_j]$, $1 \le j \le m-1$, and φ_m is β-optimal in the interval $[\beta_{m-1}, \beta_m)$ (which means that φ_m is Blackwell optimal).

In CMPs with constraints the controller wants to maximize expected (discounted) rewards while keeping other expected (discounted) costs in some given bounds. For such CMPs Altman e.a. [1] gave a constructive proof for the following (weaker) version of the result obtained in [18]. There exist numbers m and β_j as above such that for every $j = 1, \ldots, m$ either the constrained problem is not feasible in the open interval $(\beta_{i-1}\beta_j)$ or the value function is a rational function of β in the closed interval $[\beta_{j-1}, \beta_j]$, $j \le m-1$

and $[\beta_{m-1}, 1)$. Consequently, if the constrained problem is feasible in the neighborhood of $\beta = 1$, then v_β has a Laurent series expansion at $\beta = 1$.

As shown in the proof of Theorem 2, the limits of β-optimal policies, for β tending to 1, are Blackwell optimal. A counterexample in Hordijk and Spieksma [21] shows that in general this is not true in unichain CMPs with a finite state space and compact action sets. This disproves a conjecture in Cavazos-Cadena and Lasserre [4].

2. Denumerable state models

In this section we consider CMPs for which the state space $\mathbb{X}$ is denumerable. There are many applications of controlled Markov chains for which it is natural to take an infinite number of states. An important class of models is that of open stochastic networks, used for the modelling of controlled communication systems.

Even in the case when the action space $\mathbb{A} = \{0, 1\}$ consists of only two elements, but the state space $\mathbb{X}$ is denumerable, the situation with Blackwell optimal policies is much more complicated than in finite models. It turns out that a Blackwell optimal policy may be not average optimal. A corresponding counterexample, based on the fact that the Cesaro lower limit of a sequence of numbers can be different from the Abel lower limit of the same sequence, was constructed by Flynn [52]. Also, there can be no Blackwell optimal policy not only in the sense of Definition 1, but also in the sense of a weaker Definition 12 given below, in which the interval $(\beta_0, 1)$ may depend on the state x and the nonoptimal policy π. Maitra [53] presented such a counterexample, and in connection with it formulated this weaker definition.

The analysis of sensitive and Blackwell optimality for denumerable state models is mostly done under the following assumption.

Assumption 9 (a) *Action sets $\mathbb{A}(x)$, $x \in \mathbb{X}$, are compact metric sets.*

(b) *Transition probabilities $p_{xy}(a)$ and rewards $r(x, a)$ are continuous functions of $a \in A(x)$ for all $x, y \in \mathbb{X}$.*

2.1. n-discount optimality The following Liapunov function condition introduced in Hordijk [17] implies, together with Assumption 9, the existence of a stationary n-discount optimal policy.

Assumption 10 *There exist a state, say state 0, and nonnegative functions $g_0, g_1, \ldots, g_{n+1}$ on $\mathbb{X}$ such that*

(a) $\max\limits_{a\in\mathbb{A}_x} |r(x, a)| \le g_0(x), \qquad x \in \mathbb{X},$

(b) $\inf\limits_{x\in\mathbb{X}} g_0(x) > 0,$

$g_m(x) + \sum\limits_{y\neq 0} p_{xy}(a) g_{m+1}(y) \le g_{m+1}(x)$ *for all* $a \in \mathbb{A}$, $x \in \mathbb{X}$, $m = 0, 1, \ldots, n,$

(c) $P^a g_{n+1}(x) = \sum\limits_{y\in\mathbb{X}} p_{xy}(a) g_{n+1}(y)$ *is a continuous function of* $a \in \mathbb{A}_x$, $x \in \mathbb{X}$.

It is easily seen that $g_m(x) \le g_{m+1}(x)$, $m = 0, 1, \ldots, n$, $x \in \mathbb{X}$. Hence, by the dominated convergence theorem, it follows from (c) that $P^a g_m(x)$ is a continuous function of $a \in \mathbb{A}_x$, $x \in \mathbb{X}$ also for $m = 0, \ldots, n$.

In the case of a finite model, this assumption requires the accessibility of the state 0 under each stationary policy from each state. In the denumerable models it requires a strong version of recurrence to the state 0. More precisely, it assumes the finiteness under any policy of the n-th absolute moment of the total cost until the state 0 is reached, with immediate "cost" $c(x,a)$ equal to $|r(x,a)| \vee 1$, $(x,a) \in \mathbb{K}$. For simplicity, in [17] Assumption 10 is supposed to hold for all n. However, the proofs there remain true and provide sharper results if this assumption holds for a fixed n.

As shown in [17], Assumptions 9 and 10 imply for every $\varphi \in \Pi^s$ a partial Laurent expansion of the form

$$v_\beta^\varphi = (1+\rho) \sum_{k=-1}^{n} h_k^\varphi \rho^k + O(\rho^n) \tag{42}$$

where $O(\rho^n)$ is uniform in φ. Moreover, the coefficients h_k^φ are continuous in $\varphi \in \Pi^s$ in the following topology. The space Π^s in the case of a denumerable $\mathbb{X}$ is the Cartesian (direct) product $\prod_{x \in \mathbb{X}} \mathbb{A}(x)$, and we take in each $\mathbb{A}(x)$ the Borel topology of a metric space, and in Π^s the product topology.

Using this continuity and the compactness of $A(x)$, we can find for each $x \in \mathbb{X}$ a lexicographically maximal element $H_n^*(x) = \{h_k^*(x),\ -1 \le k \le n\}$ of $H_n^\varphi(x)$ over all $\varphi \in \Pi^s$, and the corresponding maximizer φ_x (see Subsection 1.2 for the lexicographical ordering and notations). Clearly, then

$$\varliminf_{\beta \uparrow 1} \rho^{-(n-1)} [v_\beta(x, \varphi_x) - v_\beta(x, \varphi)] \ge 0, \qquad x \in \mathbb{X}$$

for any other $\varphi \in \Pi^s$, so that φ_x can be called $(n-1)$-discount optimal in the class Π^s for the initial state x. However, in the proper n-dscount optimality the same policy should fit for all states x.

The following theorem derived in Hordijk [17] leads to this goal (in Hordijk and Sladky [20] it can be found with a different proof).

Theorem 11 *Suppose Assumptions 9 and 10. Then there exist functions $u_{-1}, u_0, \ldots, u_n$ from $\mathbb{X}$ to $\mathbb{R}$ (u_{-1} is a constant) statisfying bounds*

$$|u_m(x)| \le c_m g_m(x), \qquad m = 0, 1, \ldots, n$$

for some constants c_m, and nonempty compact (in the product topology) sets

$$\mathcal{P}_{-1} = \Pi^s \supset \mathcal{P}_0 \supset \cdots \supset \mathcal{P}_n$$

of stationary policies of the form $\mathcal{P}_m = \prod_{x \in \mathbb{X}} \mathbb{A}_m(x)$ ($\mathbb{A}_{-1}(x) = \mathbb{A}(x)$) with the following property. Let $U_m = \{u_{-1}, \ldots, u_m\}$. We have $L^\varphi U_m \le 0$ for every $\varphi \in \Pi^s$ and $m = -1, \ldots, n$, and $\varphi \in \mathcal{P}_m$ if and only if $L^\varphi U_m = 0$.

It is shown in [20] that $\varphi \in \mathcal{P}_{m+1}$ $(m = -1, \ldots, n-1)$ *if and only if* φ *is* m*-discount optimal in the class* Π^s (this means that (23) holds for $n = m$ for $\pi^* = \varphi$ and any $\pi \in \Pi^s$). If $\varphi \in \mathcal{P}_{m+1}$, then the coefficients h_k^{φ}, $-1 \le k \le m$ in (42) coincide with u_k, so that in Theorem 11 we have an analogue of the conserving property.

Also, we can conclude from the above results that the value function V_β has a partial Laurent expansion with the coefficients equal to the functions u_m. Hence

$$V_\beta(x) = (1+\rho)\left[u_{-1}\rho^{-1} + u_0(x) + \sum_{k=1}^{m-1} u_k(x)\rho^k\right] + O(\rho^m) \qquad x \in \mathbb{X}, \tag{43}$$

where $u_k = h_k^{\varphi}$, $k = 1, \ldots, m-1$ and $\varphi \in \mathcal{P}_{m+1}$. For $m = 1$ this result is established in Cavazos-Cadena and Lasserre [4] under more restrictive recurrence conditions.

Hordijk and Sladky [20] also proved that m-discount optimality is equivalent to m-average optimality for $m = 0, 1, \ldots, n-1$ in denumerable CMPs satisfying Assumptions 9 and 10.

Let us now assume that Assumption 10 holds for all $n \in \mathbb{N}$. Then we have nonempty compact sets $\mathcal{P}_n$ for every n, and their intersection

$$\mathcal{P}_\infty = \bigcap_{n=1}^{\infty} \mathcal{P}_n$$

is also a nonempty compact set in Π^s. A policy $\varphi \in \mathcal{P}_\infty$ is n-discount optimal (in the class Π^s) for every $n \in \mathbb{N}$, and it is tempting to conjecture that φ is Blackwell optimal. However, in general this is not true. We return to this question in Section 2.2.

2.2. On Blackwell optimality in infinite state models. The original Blackwell definition (Definition 1) is too strong for the denumerable state CMPs, as we will see in the counterexample below. In the following definition it is weakened, and a policy satisfying Blackwell's version is renamed into a *strong Blackwell optimal policy.* It is easy to see that in finite models both definitions coincide. Note that Veinott's Definition 4 of n-discount optimality is stated in weak terms, applicable to general models.

Definition 12. *For any set* $\Pi' \subset \Pi$ *a policy* $\pi^* \in \Pi'$ *is said to be Blackwell optimal within the class* Π'*, if for every* $x \in \mathbb{X}$ *and* $\pi \in \Pi$ *there exists a number* $\beta_0(x,\pi) < 1$ *such that*

$$v_\beta(x,\pi^*) \ge v_\beta(x,\pi) \quad \textit{for all } \beta \in (\beta_0(x,\pi), 1).$$

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

Counterexample 13. The state space is

$$\mathbb{X} = \mathbb{X}_0 \cup \mathbb{X}_1 \cup \mathbb{X}_2 \cup \cdots$$

with

$$\mathbb{X}_0 = \{(0,0)\},$$
$$\mathbb{X}_n = \{(n,0),(n,1,1),\ldots,(n,n,1),(n,1,2),\ldots,(n,n,2)\}.$$

The action sets are

$$A((0,0)) = (0,1,\tfrac{1}{2},\tfrac{1}{3},\ldots\},$$
$$A(n,0)) = \{1,2\}, \quad A((n,i,j)) = \{1\} \quad 1 \le i \le n, \quad n \ge 1 \quad j = 1,2.$$

The transition probabilities are

$$p((1,0)|(0,0),\tfrac{1}{n}) = 1 - p((n,0)|(0,0),\tfrac{1}{n}) = 1 - 2^{-n}, \quad n = 1,2,\ldots$$
$$p((1,0)|(0,0),0) = 1,$$

and for $n = 1,2,\ldots$

$$p((n,1,1)|(n,0),1) = p((n,1,2)|(n,0),2) = 1,$$
$$p((n,i+1,1)|(n,i,1),1) = p((n,i+1,2)|(n,i,2),1) = 1, \quad 1 \le i \le n-1,$$
$$p((n,0)|(n,n,1),1) = p((n,0)|(n,n,2),1) = 1.$$

The immediate rewards are

$$r((0,0),a) = 1 \quad \forall a \in A(0,0),$$

and for $n = 1,2,\ldots$

$$r((n,0),2) = n, \quad r((n,0),1) = 1,$$
$$r((n,i,1),1) = 1, \quad r((n,i,2),1) = 0, \qquad 1 \le i \le n.$$

Note that this CMP satisfies assumption 9.
Define $\varphi_k, \quad k = 1,2,\ldots$ as follows:

$$\varphi_k(0,0) = \tfrac{1}{k}$$
$$\varphi_1(n,0) = 1 \quad \text{and} \quad \varphi_k(n,0) = 2 \quad \text{for} \quad k = 2,3,\ldots,\ n = 1,2,3,\ldots$$

It is easy to calculate that

$$v((n,0),\varphi_1,\beta) = (1-\beta)^{-1},$$

and

$$v((n,0),\varphi_k,\beta) = n(1-\beta^{n+1})^{-1} \qquad \text{for} \quad n,k \ge 2.$$

Hence

$$v((n,0),\varphi_1,\beta) \ge v((n,0),\varphi_k,\beta)$$

if and only if $\beta \ge \beta_n$ with β_n being the unique solution of the equation $1+\beta+\cdots+\beta^n = n$ in the interval $0 \le \beta \le 1$. Since β_n is monotone increasing to 1 as $n \to \infty$, there is no $\beta_0 < 1$ such that for $k \ge 2$,

$$v(x,\varphi_1,\beta) \ge v(x,\varphi_k,\beta) \quad \text{for all} \quad \beta \in [\beta_0, 1) \quad \text{and all} \quad x \in \mathbb{X}.$$

Hence φ_1 is not a *strongly* Blackwell optimal policy. Clearly, for fixed initial state $(n, 0)$, $n \geq 1$ there is an β_n such that φ_1 is discounted optimal for $\beta \in [\beta_n, 1)$.
This is not true for the state $(0, 0)$. Indeed, for $k \geq 2$

$$\begin{aligned}
&v((0,0),\varphi_1,\beta) - v((0,0),\varphi_k,\beta) = \\
&\beta[(1-\beta)^{-1} - (1-2^{-n})(1-\beta)^{-1} - 2^{-n}n(1-\beta^{n+1})^{-1}] = \\
&\beta 2^{-n}[(1-\beta)^{-1} - n(1-\beta^{n+1})^{-1}],
\end{aligned}$$

which is nonnegative if and only if $\beta \geq \beta_k$. Thus φ_1 is (*weakly*) *Blackwell optimal* in the class Π^s.

We next show that φ_1 is Blackwell optimal in the class of randomized stationary policies, it is also optimal in the class of all policies. Since the set of states $\{(n,0),(n,1,1),\ldots,$ $(n,n,1),(n,1,2),\ldots,(n,n,2)\}$ is a closed set under any policy, it follows from the results for finite models that φ_1 is Blackwell optimal on this set in the class of all policies. This holds for all $n \geq 1$. Hence it is sufficient to consider a policy φ which only randomizes in state $(0, 0)$, say with probability p_k it takes action $\frac{1}{k}$. If φ takes action 1 in $(n, 0)$ then $v((n,0),\varphi,\beta) = (1-\beta)^{-1} = v((n,0),\varphi_1,\beta)$, and for computing the difference between $v((0,0),\varphi_1,\beta) - v((0,0),\varphi,\beta)$ we may as well set $p_n = 0$ in this case. So without loss of generality suppose $\varphi(n, 0) = 2$ if $p_n > 0$. Then

$$\begin{aligned}
v((0,0),\varphi,\beta) &= 1 + \beta \sum_{k=2}^{\infty} p_k \left(2^{-k} \cdot \frac{k}{1-\beta^{k+1}} + (1-2^{-k})\frac{1}{1-\beta}\right) \\
&= 1 + \frac{\beta}{1-\beta} \sum_{k=2}^{\infty} p_k \left(2^{-k} \frac{k}{1+\beta+\cdots+\beta^k} + (1-2^{-k})\right).
\end{aligned}$$

On the other hand

$$v((0,0),\varphi_1,\beta) = 1 + \frac{\beta}{1-\beta}.$$

Hence,

$$\begin{aligned}
f(\beta) :=& \frac{1-\beta}{\beta}(v((0,0),\varphi_1,\beta) - v((0,0),\varphi,\beta)) = \\
& \sum_{k=2}^{\infty} p_k \left(2^{-k}\left(1 - \frac{k}{1+\beta+\cdots+\beta^k}\right)\right).
\end{aligned}$$

By dominated convergence

$$\lim_{\beta \uparrow 1} f(\beta) = \sum_{k=2}^{\infty} p_k 2^{-k} \cdot \frac{1}{k+1}.$$

Consequently,

$$\lim_{\beta \uparrow 1}(v((0,0),\varphi_1,\beta) - v((0,0),\varphi,\beta)) = \begin{cases} 0 & \text{if } p_k = 0,\ k \geq 2 \\ \infty & \text{otherwise,} \end{cases}$$

and φ_1 dominates φ for β sufficiently close to 1. Hence φ_1 is Blackwell optimal in the class of randomized stationary policies. With similar arguments it can be shown that it is Blackwell optimal in the class of all policies. However, φ_1 is not a strong Blackwell optimal policy, since there does not exist a $\beta_0 < 1$ such that $v((n,0),\varphi_1,\beta) = v((n,0),\beta)$ for all $n \geq 1$ and $\beta_0 < \beta < 1$.

We now return to the question whether a policy $\varphi \subset \mathcal{P}_\infty$ is Blackwell optimal (if Assumption 10 holds for all n). Take any policy $\psi \in \Pi^s$ and initial state $x \in \mathbb{X}$ and consider the infinite sequences $H^\varphi(x) = \{h_k^\varphi(x)\} = \{u_k(x)\}$ and $H^\psi(x) = \{h_k^\psi(x)\}$ (see (41) and Theorem 11). By this theorem, $H_m^\varphi(x) \succeq H_m^\psi(x)$ for every $m \in \mathbb{N}$, hence the same is true for the infinite sequences: $H^\varphi(x) \succeq H^\psi(x)$. Therefore either $H^\varphi(x) = H^\psi(x)$, or there is an integer m such that

$$\begin{aligned} &h_k^\varphi(x) = h_k^\psi(x) \quad \text{for} \quad -1 \leq k \leq m-1, \\ &h_m^\varphi(x) > h_m^\psi(x). \end{aligned}$$

In the second case it follows from expansions (42) for φ and ψ with $n = m+1$, that $v_\beta^\varphi(x) > v_\beta^\psi(x)$ for all β in some interval $(\beta_0(x,\psi),1)$. The difficulty arises in the first case: there is no guarantee that v_β^φ and v_β^ψ have complete Laurent expansions of the form $v_\beta^\varphi = (1+\rho)\sum_{-1}^{\infty} h_k\rho^k$. Indeed, they may not, as one may see from the following example.

Consider the one-server queue with a controllable Poisson arrival process as studied in Hordijk [17, Section 2.2]. It is shown there that Assumptions 9 and 10 are satisfied for a given n, if $\mathbb{E}\, S^{n+1} < \infty$, where S is the service time of one customer. Hence if $\mathbb{E}\, S^k < \infty$ for all $k \in \mathbb{N}$ and the rewards $r(x,a)$ are bounded by a polynomial in x, the set $\mathcal{P}_\infty$ is nonempty. However, if the Laplace-Stieltjes transform of the service time

$$f(z) = \int_0^\infty e^{-zx}\, dF_S(x)$$

is not analytic at $z = 0$, then the complete Laurent expansion of discounted rewards does not exist for $r(x,a) = \mathbf{1}\{x = 0\}$ and the Poisson arrival process with a positive parameter λ. Indeed, in this case we have a homogeneous random walk as described in Hordijk et al. [19, Section 2].

2.3. Operator theoretical approach to Blackwell optimality. A satisfactory theory of denumerable state model should contain as a special case the finite model. Assumption 10 is not suitable for this purpose. It presupposes a unichain CMP, while the theory of finite CMPs covers the multichain case too. So it is too restrictive. On the other hand, it does not guarantee the existence of complete Laurent expansions of discounted rewards under stationary policies, and henceforth is too weak to obtain the Blackwell optimality.

In Dekker and Hordijk [8] a theory of denumerable CMPs has been developed free of these inadequacies. The operator theoretical approach to Blackwell optimality is introduced there. In this approach a *bounding function* μ is used satisfying the condition $\mu(x) \geq 1$, $x \in \mathbb{X}$. We relate with μ the Banach space V_μ of all real-valued functions f on $\mathbb{X}$ with the finite μ-norm

$$\|f\|_\mu = \sup_{x \in \mathbb{X}} \frac{|f(x)|}{\mu(x)}.$$

The associated operator norm is

$$\|T\|_\mu = \sup_{f : \|f\|_\mu \leq 1} \|Tf\|_\mu$$

for any operator $T : V_\mu \to V_\mu$. In the following "bounding assumption" the notations (1), (2), and also the notation

$$\hat{f}(x) = \sup_{a \in \mathbb{A}(x)} |f(x,a)|, \qquad x \in \mathbb{X} \tag{44}$$

for any function f on K, are used.

Assumption 14 *For some constant $C > 0$*
(a) $\|\hat{r}\|_\mu \leq C$,
(b) $P^a\mu(x) \leq C\mu(x), \qquad (x,a) \in \mathbb{K}$.

In [8] the part (b) of the compactness-continuity Assumption 9 concerning the transition probabilities is strengthened to the following form.

Assumption 15 *$P^a f(x)$ is continuous in $a \in \mathbb{A}(x)$ for every $f \in V_\mu$ and $x \in \mathbb{X}$.*

Besides the bounding and compactness-continuity assumptions, we need also a condition to guarantee the Laurent series expansion for v_β^φ, $\varphi \in \Pi^s$ (it should be pointed out that Assumption 10 implies Assumptions 14 and 15 with $\mu = \text{const} \cdot g_n$, but not the complete Laurent expansions). Dekker and Hordijk [8] introduced an ergodicity condition, renamed in Hordijk and Spieksma [22] into μ-geometric ergodicity in the case of a Markov chain, and into uniform μ-geometric ergodicity in the case of a CMP. Note that a CMP can be seen as a compact product set of Markov chains (see Hordijk [16]), and that therefore any ergodicity property of a CMP becomes a corresponding property of a Markov chain if CMP consists of a single chain. Since its introduction in [8], the μ-geometric ergodicity became also a new notion in the Markov processes literature, and has been intensively studied (see Meyn and Tweedie [29]). The uniform μ-geometric condition is

Assumption 16 *For every $\varphi \in \Pi^s$, the t-th convolution $P^t(\varphi)$ of the operator $P(\varphi)$ converges to a limiting stochastic operator $Q(\varphi)$ geometrically fast in the μ-norm, i.e. for some constants $C < \infty$ and $\gamma < 1$*

$$\|P^t(\varphi) - Q(\varphi\|_\mu \leq C\gamma^t, \qquad \varphi \in \Pi^s, t \in \mathbb{N}. \tag{45}$$

Note that (45) requires the aperiodicity of all Markov chains with kernels P^φ. As shown in Hordijk and Yushkevich [24], the following weaker version of (45) suffices for the study of Blackwell optimality, which covers the case of periodic chains.

Assumption 16' *For some constants T, $C < \infty$ and $\gamma < 1$*

$$\left\| \frac{1}{T} \sum_{k=1}^{T} P^{k+t}(\varphi) - Q(\varphi) \right\|_\mu \leq C\gamma^t, \quad \varphi \in \Pi^s, \quad t \in N. \tag{46}$$

Note that condition (46) (even with T independent of φ) is fulfilled in any finite CMP (because it is true for a finite Markov chain, and the number of Markov chains corresponding to stationary policies is finite). In a denumerable Markov chain, this condition implies the existence of the deviation operator analoguous to the deviation matrix considered in Section 1.1. The following lemma is indeed a general fact on operators in Banach spaces with a convergent resolvent. We state it in the context of the μ-norm and operators $P(\varphi)$.

Lemma 17 *Assumption* 16' *implies the existence of the stationary operator $Q(\varphi) = Q^\varphi$ and the deviation operator $D(\varphi) = D^\varphi$ as defined in* (6) *and* (7') (*the limits are understood in the μ-norm*). *Moreover, the equations in* (6) *for Q^φ and in* (8)-(9) *for D^φ are satisfied.*

We have seen in Section 1.2 that in the case of a finite state space the resolvent

$$R_\beta(P) = R(\rho, P) = \sum_{t=0}^{\infty} \left(\frac{P}{1+\rho} \right)^t = \sum_{t=0}^{\infty} (\beta P)^t = (I - \beta P)^{-1} \tag{47}$$

of the transition operator P has a Laurent series expansion in the neighborhood of $\rho = 0$, reflected in formulas (13) and (16)-(18) together with $v_\beta^\varphi = R_\beta(P^\varphi) r^\varphi$, implied by equations (6) and (8)-(9). The same expansion remains true in the case of a Markov chain on a general state space $\mathbb{X}$, if there is a geometric convergence (46).

Lemma 18 *If Assumption* 16' *holds, then there exists a number $\rho_0 > 0$ such that for all complex values of ρ in the ring $0 < |\rho| < \rho_0$*

$$R(\rho, P^\varphi) = (1+\rho) \left[\frac{Q^\varphi}{\rho} + \sum_{n=0}^{\infty} (-\rho)^n (D^\varphi)^{n+1} \right], \tag{48}$$

where Q^φ and D^φ are the same bounded (*in the μ-norm*) *operators as in Lemma* 17.

Proof. Essentially, it is the same algebra based on equations (8)-(9) as in the proof of Theorem 3. However, there we knew beforehand that the resolvent has a simple pole at $\rho = 0$. To avoid this, one may check by direct algebra that the series (48) (which converges in some ring $0 < |\rho| < \rho_0$, because the operator D^φ is bounded) defines an operator inverse to $I - \beta P$ (see (47)). Indeed, we have according to (8)-(9) and (6) (and omitting the index φ)

$$
\begin{aligned}
(1+\rho)\left[\frac{Q}{\rho} + \sum_0^\infty (-\rho^n)(D)^{n+1}\right]\left(I - \frac{P}{1+\rho}\right) &= \\
= \left[\frac{Q}{\rho} + \sum_0^\infty (-\rho)^n D^{n+1}\right][(I-P)+\rho I] &= \\
= Q + \sum_0^\infty (-\rho)^n D^{n+1}(I-P) + \sum_0^\infty (-1)^n (\rho D)^{n+1} &= \\
= Q + \sum_0^\infty (-\rho)^n D^n (I-Q) - \sum_1^\infty (-\rho)^n D^n &= \\
= Q + I - \sum_0^\infty (-\rho)^n D^n Q = Q + I - Q = I&,
\end{aligned}
$$

and the same holds if we multiply in the reverse order. □

To proceed further, we need to extend the space $\mathfrak{H}$ and operators in it, introduced in Section 1.3, to the case of a countable state space $\mathbb{X}$ and the Banach space V_μ of functions on $\mathbb{X}$. Clearly, instead of sequences of Laurent coefficents we may consider the Laurent series themselves.

Definition 19 (a) *The linear space $\mathfrak{H}_\mu$ consists of all Laurent series of the form*

$$h := h(x) = h(x,\rho) = \sum_{n=-1}^\infty h_n(x)\rho^n, \qquad h_n \in V_\mu, \quad x \in \mathbb{X}$$

in the complex variable ρ with coefficients satisfying the geometric growth condition

$$\overline{\lim_{n\to\infty}} \|h_n\|_\mu^{\frac{1}{n}} < \infty.$$

(b) *For every $\psi \in \Pi^s$, operators L^ψ and U^ψ in the space $\mathfrak{H}_\mu$ are given by the formulas*

$$
\begin{aligned}
L^\psi h &= r^\psi + P^\psi h - (1+\rho)h \qquad h \in \mathfrak{H}_\mu, \\
U^\psi &= \frac{\rho}{1+\rho} R(\rho, P^\psi).
\end{aligned}
\tag{49}
$$

We use in $\mathfrak{H}_\mu$ the lexicographical ordering defined in Section 1.2 for the sequences of their coefficients $H = \{h_{-1}, h_{0,\dots}\}$. Clearly, if $H' \preceq H''$, or equivalently $h' \preceq h''$, then $h'(x,\rho) \preceq h''(x,\rho)$ for all positive sufficiently small ρ, etc. Note that the definition (49) of L^ψ is consistent with formulas (33)-(35).

Lemma 20 *Suppose Assumption* 16'. *Then for every* $\varphi \in \Pi^S$ *the formulas* (16)-(18) *of Theorem* 3 *are valid, with coefficients* h_n^φ *(or* k_n^φ*)* $\in V_\mu$. *If Assumption* 16 *holds* (*or if* T *in Assumption* 16' *is the same for all* φ), *then*

$$\overline{\lim_{n\to\infty}} \left[\sup_{\varphi \in \Pi^s} \|h_n^\varphi\| \right]^{\frac{1}{n}} < \infty.$$

This lemma is a direct consequence of the formula $v_\beta^\varphi = R(\rho, P^\varphi) r^\varphi$ and Lemma 18. The corresponding element of $\mathfrak{H}_\mu$ we denote $h^\varphi = \sum_{-1}^{\infty} h_n^\varphi \rho^n$, so that

$$v_\beta^\varphi = (1+\rho) h^\varphi.$$

The following comparison lemma is an extension to denumerable models of a result derived by Veinott [39] for finite models.

Lemma 21 (Comparison lemma) *Suppose Assumptions* 9,14,15 *and* $16'$. *Then for every* $h \in \mathfrak{H}_\mu$ *and* $\psi \in \Pi^s$ *there exists* $\rho_0 > 0$ *such that*

$$h^\psi - h = \frac{1}{\rho} U^\psi L^\psi h, \qquad 0 < |\rho| < \rho_0.$$

Central in the analysis of [8] is the following lemma which is the key lemma in the operator theoretical approach.

Lemma 22 (Key lemma) *Under assumptions of Lemma* 21, *for every* $\psi \in \Pi^s$ *the operator* U^ψ *is a positive operator: if* $h \in \mathfrak{H}_\mu$, $h \succeq 0$ *then* $U^\psi h \succeq 0$ (*if* $h \preceq 0$ *then* $U^\psi h \preceq 0$). *Moreover, if* $h \succeq 0$ *and* $h(x_0) \succ 0$ *for some* $x_0 \in \mathbb{X}$, *then* $Uh(x_0) \succ 0$.

The key lemma together with the comparison lemma yield the lexicographical policy improvement approach for the denumerable models (cf. Lemma 7 for finite models). For the average criterion in finite models policy improvement was developed by Howard [26] and Blackwell [3] and for the sensitive criteria in those models by Veinott [39]. In their honor we call it Howard-Blackwell-Veinott policy improvement.

In finite models policy improvement is a constructive way to find a Blackwell optimal policy, as sketched in Section 1.3. In infinite state space models this algorithm is not sufficient, since the improvements may not terminate in a finite number of steps. Lemmas 21-22 provide also for finite models an approach to Blackwell optimality different from that of Blackwell and Veinott.

The following theorem gives several equivalent formulations of Blackwell optimality in the class Π^s of stationary policies. Its proof is rather a direct consequence of the comparison and key lemmas. We refer to formulas (39)-(41) and Theorem 8 for notations, terminology and a comparison with the case of a finite model.

Theorem 23 *Suppose Assumptions* 9,14,15 *and* 16'. *Then the following statements concerning a policy* $\varphi \in \Pi^s$ *are equivalent:*
(a) φ *is Blackwell optimal within the class* Π^s;
(b) $h^\varphi \succeq h^\psi$ *for every* $\psi \in \Pi^s$;
(c) h^φ *is a solution of the Blackwell optimality equation* $Lh = 0$;
(d) $L^\varphi h = 0$ *for a solution* h *of the equation* $Lh = 0$ (*i.e.* φ *is a conserving policy*).
Moreover, the solution of the equation $Lh = 0$ (*if any*) *is unique.*

It is shown in [8] that under continuity and uniform μ-geometric ergodicity assumptions, $P^\varphi\mu$, $Q^\varphi\mu$ and $D^\varphi\mu$ are continuous functions of φ on the compact Π^s (in the product topology). This, together with the bounding assumption, implies the continuity in φ of the Laurent coefficients $h_n^\varphi(x)$ given by formulas (17) and (18). This, together with a diagonal process on the countable set X, allows to get a maximizer $h^\varphi = \max_\psi h^\psi$, $\psi \in \Pi^s$ as in part (b) of Theorem 23. By this theorem, φ is Blackwell optimal in the class Π^s, and the Blackwell optimality equation has a solution. A technical proof given in [8] shows that φ is Blackwell optimal also in the class Π of all policies (versions of this proof for Borel models can be found in [44], [50] and [25]). Thus, the following result holds.

Theorem 24 *In a denumerable state space model satisfying Assumptions* 9 *and* 14-16 *there exists a Blackwell optimal policy.*

A related question is whether a limit φ of β-optimal policies φ_β (if the limit exists as $\beta \uparrow 1$) is Blackwell optimal. Under weaker assumptions than above, it is shown in Hordijk [17], that such φ is 0-discount optimal. Under another set of assumptions, which can be shown to be more restrictive than Hordijk's conditions, the same result is obtained by Cavazos-Cadena and Lasserre [4]. On the other hand, Hordijk and Spieksma [21] have constructed an example in which the limiting policy φ is not 1-discount optimal, so a fortiori not Blackwell optimal.

Lasserre [28], starting from the ideas developed in [8], obtained the existence of a policy Blackwell optimal within Π^s (and hence, according to [8], in the class Π too) without the policy improvement, making use of more results in the spectral theory of bounded linear operators.

Yushkevich [47] has shown how one may get the existence of Blackwell optimal policies in denumerable models with periodic chains by perturbing them into aperiodic models. In this work, besides Assumption 9, the boundedness of the reward function r was assumed, as well as the following condition taken from Tijms [37]: there are a number $\epsilon > 0$ and an integer T such that for every $\varphi \in \Pi^s$ there exists a state $y_T = y(\varphi)$ such that

$$\sum_{t=1}^{T} \mathbb{P}_x^\varphi\{x_t = y\} \geq \epsilon, \qquad x \in X.$$

In fact, we get the most general results if we use Assumptions 16 (and 16') in a nonuniform way; this is true for the above results except the continuity of $Q(\varphi)$ and $D(\varphi)$. So, for Theorem 23 the following weak versions of Assumptions 16 and 16' are sufficient.

Assumption 16(w) *For every $\varphi \in \Pi^s$, there exist constants $C(\varphi) < \infty$ and $\gamma(\varphi) < 1$ such that*

$$\|P^t(\varphi) - Q(\varphi)\|_\mu \leq C(\varphi)\gamma^t(\varphi).$$

Assumption 16'(w) *For every $\varphi \in \Pi^s$, there exist $T(\varphi)$, $C(\varphi)$ and $\gamma(\varphi) < 1$ such that*

$$\left\| \frac{1}{T(\varphi)} \sum_{k=1}^{T(\varphi} P^{k+t}(\varphi) - Q(\varphi) \right\|_\mu \leq C(\varphi)\gamma^t(\varphi).$$

For continuity of $Q(\varphi)$ and $D(\varphi)$ in φ and therefore for Theorem 24 we need the uniform Assumption 16 (or 16').

2.4. Recurrence conditions for Blackwell optimality. In [10] recurrence conditions are introduced which imply the existence of complete Laurent series and of Blackwell optimal policies. Starting in Ross [31], recurrence conditions have been extensively used and studied in undiscounted nonfinite CMPs.

The first analysis based on the notion of 'simultaneous Doeblin condition', can be found in Hordijk [16]. Many equivalent formulations of this condition have been derived. We refer to [16], Federgruen, Hordijk and Tijms [13], [14], Thomas [36], Hernández-Lerma, Montes-de-Oca, Cavazos-Cadena [15], Hordijk and Spieksma [22], Dekker, Hordijk and Spieksma [11]. We present here some of the results which appeared in the last of these papers.

The taboo transition matrix ${}_MP$ with taboo set $M \subset X$ is defined by

$${}_Mp_{xy} = \begin{cases} p_{xy} & y \notin M \\ 0 & y \in M, \end{cases}$$

with the convention that ${}_MP^t$ is the t-fold matrix-product of ${}_MP$, and ${}_MP^0 = I$, with I the identity-matrix. The uniform μ-geometric recurrence condition (in the weak form) is

Assumption 25 *There is a finite set M and constants $c < \infty$ and $\gamma < 1$ such that*

$$\| {}_MP^t(\varphi)\|_\mu < c\gamma^t, \qquad t = 0, 1, \ldots, \quad \varphi \in \Pi^S.$$

For the special case $\mu = e$, where e the function with $e(x) = 1$ for all $x \in X$, uniform μ-geometric recurrence is equivalent to the simultaneous Doeblin condition (see Hordijk [17], Theorem 11.3 and especially relation (11.3.2)). The generalization from e to a general (mostly unbounded) bounding function μ is important. It gives not only results for unbounded reward functions (see the bounding Assumption 14), but also covers the class of CMP satisfying the uniform μ-geometric recurrence condition for a suitable chosen μ which is essentially larger than that of satisfying the simultaneous Doeblin condition. Indeed, let τ be the recurrence time to the set M, then

$$\{\tau > t\} = \{x_k \notin M, \quad 1 \le k \le t\}.$$

Hence,

$$\mathbb{P}_x^{\varphi}\{\tau > t\} = \left({}_M P^t(\varphi) e\right)(x)$$

and under Assumption 25,

$$\mathbb{E}_x^{\varphi}\, \tau = \sum_{t=0}^{\infty} \mathbb{P}_x\{\tau > t\} = \sum_{t=0}^{\infty} \left({}_M P^t(\varphi) e\right)(x) \le \frac{c}{1-\gamma} \cdot \mu(x).$$

Consequently, if μ is bounded then the expected recurrence time is bounded in the starting state. Clearly, this does not hold for most queueing models. See Spieksma [35] for CMPs, especially controlled queues, which do satisfy the uniform μ-geometric recurrence condition.

Let $\nu(\varphi)$ denote the number of closed classes in the Markov chain with transition probabilities $P(\varphi)$. A set $B(\varphi) \subset X$ is called a set of 'reference states' if it contains precisely one state from each closed class and no other states.

An apparently stronger version of Assumption 25 is

Assumption 25' *There is a finite set M and constants $c < \infty$ and $\gamma < 1$ such that for every $\varphi \in \Pi^s$ there exists a reference set $B(\varphi) \subset M$, and moreover*

$$\| {}_{B(\varphi)} P^t(\varphi) \|_{\mu} < c\gamma^t, \quad t = 0, 1, \ldots.$$

In [11] it is shown that Assumption 25, together with the continuity of $v(\varphi)$ as function of φ, is equivalent to Assumption 25'. Dekker and Hordijk [10] analyzed and proved the existence of Laurent expansions and Blackwell optimality under Assumption 25'.

It was Hordijk's conjecture that (uniform) μ-geometric ergodicity is equivalent to (uniform) μ-geometric recurrence. This was proved in Hordijk and Spieksma [22] for one Markov chain and has been generalized for the unichain case to general Borel state space by Meyn and Tweedie [29]. For CMPs the equivalence is more complicated , it can be found in Dekker, Hordijk and Spieksma [11].

Note that for the finite model $\nu(\varphi)$ is automatically continuous since Π^S is a finite set. Moreover, in Assumption 25 we may take $M = \mathbb{X}$; then ${}_M P$ is the zero matrix. Therefore Assumption 25, and hence also Assumption 25', are always fulfilled in finite models, also in the multichain case.

One might ask whether Assumption 25' can be weakened. Using the existence and continuity of the Laurent expansion of the discounted rewards, one may show that for the operator theoretical approach of Dekker and Hordijk the Assumption 25' is also necessary (see Lasserre [28], Spieksma [35]).

Let us conclude this section with pointing out the relation between the Assumptions 10 and 25.

First, Assumption 10 is more restrictive, since it assumes the existence of one state which is accessible from all other states under each policy, i.e. the unichain case, whereas Assumption 25 allows a finite number of closed sets. Let us assume the unichain case (note that in the unichain case $\nu(\varphi) \equiv 1$ and so it is continuous), then Assumption 25 implies Assumption 25' and with $B(\varphi) = \{0\}$ we have

$$\| {}_0P^t(\varphi)\|_\mu < c\gamma^t, \qquad t = 0, 1, \dots$$

with $c < \infty$ and $\gamma < 1$.
Define

$$\widetilde{\mu} = \sup_\varphi \sum_{t=0}^\infty {}_0P^t(\varphi)\mu.$$

Then,

$$\mu +_0 P(\varphi)\widetilde{\mu} \le \widetilde{\mu} \qquad \forall\varphi \in \Pi^S$$

and

$${}_0P(\varphi)\widetilde{\mu} \le \widetilde{\mu} - \mu \le \left\{1 - \frac{1-\gamma}{c}\right\}\widetilde{\mu}$$

and

$$\widetilde{\mu} +_0 P(\varphi)\,\widetilde{\widetilde{\mu}} \le \widetilde{\widetilde{\mu}}. \tag{50}$$

with $\widetilde{\widetilde{\mu}} = \frac{c}{1-\gamma}\widetilde{\mu}$. Since $\mu \le \widetilde{\mu} \le \frac{c}{1-\gamma}\mu$, Assumption 14 for μ implies the same assumption for $\widetilde{\mu}$. By using (50) recursively, it is easily seen that Assumption 10 is satisfied for

$$g_n = \left(\frac{c}{1-\gamma}\right)^{n+1}\widetilde{\mu}.$$

Hence for the unichain case Assumption 25 implies Assumption 10.

With a slightly more involved argument one may show that Assumption 25 is equivalent to

$$s(\varphi) := \sum_{t=0}^\infty {}_MP^t(\varphi)\mu \le c_1\mu, \tag{51}$$

for some constant c_1 and all $\varphi \in \Pi^S$. Indeed (cf. [10]), (51) implies that

$${}_MP(\varphi)s(\varphi) = s(\varphi) - \mu \le \left(1 - \frac{1}{c_1}\right)s(\varphi).$$

Choose $\gamma_1 < 1$ and let t_0 be such that $\left(1 - \frac{1}{c_1}\right)^{t_0} c_1 < \gamma_1$; then

$$ {}_MP^{t_0}(\varphi)\mu \leq {}_MP^{t_0}(\varphi)s(\varphi) \leq \left(1 - \frac{1}{c_1}\right)^{t_0} s(\varphi) \leq \gamma_1 \mu. $$

Let $c_1 = \sup_{\varphi} {}_MP(\varphi)\mu$, $c = (c_1 \vee 1)^{t_0}\gamma_1^{-1}$ and $\gamma = \gamma_1^{1/t_0}$. Then $\gamma < 1$, and for $kt_0 \leq t < (k+1)t_0$, $k \geq 0$ we have

$$ \begin{aligned} \| {}_MP^t(\varphi)\|_\mu &\leq \| {}_MP^{kt_0}(\varphi)\|_\mu \| {}_MP^{t-kt_0}(\varphi)\|_\mu \\ &\leq \gamma_1^k c_1^{t-kt_0} \leq c\gamma^t. \end{aligned} $$

Hence Assumption 25 is satisfied with $c < \infty$ and $\gamma < 1$.

Remark. The Laurent series expansion of the discounted rewards and the existence of strong Blackwell optimal policies for *semi*-Markov decision chains with a finite number of states and actions has been established in Denardo [12]. Similar results under related recurrence conditions have been obtained for the denumerable state model in Dekker and Hordijk [9].

3. Borel state models

In this section we consider Blackwell optimality in CMPs with a Borel state space. We formulate the existence results, describe distinctive features of the approach to Borel models, state recurrence conditions which imply less verifiable uniform ergodicity and integrability assumptions.

3.1. Existence of Blackwell optimal policies. The study of Blackwell optimality in Borelian models was started by Yushkevich [44,48] and continued by Hordijk and Yushkevich [24], [25]. An extended summary of results obtained in [48] can be found in [50]. A related paper is Yushkevich [49], where the compactness of the policy space is treated.

An advance in the direction of Borel models appeared possible in the case when the transition probabilities are given by transition densities. This is a common case in models with a continous state space. We also need the corresponding versions of the compactness-continuity Assumption 9 and either of the uniform geometric ergodicity Assumption 16 or of recurrence conditions implying Assumption 16. Models with a bounded reward function and a strong minorant or simultaneous Doeblin-Doob condition were treated in [48]; the particular case of finite action sets $\mathbb{A}(x)$ was studied before that in [44]. In models with unbounded rewards considered in [24][25], one needs a stronger version of the bounding Assumption 14, and the ergodicity or recurrence conditions should be stated in the terms of μ-norms; also a technical uniform integrability condition is needed in the absence of recurrence conditions.

To avoid repetitions, we first state the more general results obtained in [24][25]. Before stating the whole set of conditions, we introduce notations related to transition densities and randomized stationary policies; the formulas will become meaningful under subsequent assumptions. There is a reference *measure* $m(dx)$ on the space $\mathbb{X}$, and we often write dx instead of $m(dx)$. The transition probabilities $p(Y \mid x, a)$ are determined by *transition densities* $p(x, a, y)$ so that (for measurable $Y \subset \mathbb{X}$)

$$p(Y \mid x, a) = \int\limits_{Y} p(x, a, y) m(dy), \qquad (x, a) \in \mathbb{K}.$$

Similar to (44), we denote (because the maximum exists)

$$\hat{p}(x, y) = \max_{a \in \mathbb{A}(x)} p(x, a, y), \qquad x, y \in \mathbb{X}.$$

Formula (2) takes on the form

$$P^a f(x) = \int\limits_{\mathbb{X}} p(x, a, y) f(y) m(dy).$$

For uniformity with other notations of this section, we denote by $\sigma(x, da)$ the probability measure $\sigma(\cdot \mid x)$ on $\mathbb{A}(x) \subset \mathbb{A}$ defined by a *randomized stationary policy* $\sigma \in \Pi^{RS}$. The transition density corresponding to $\sigma \in \Pi^{RS}$ is

$$p^\sigma(x, y) = \int\limits_{\mathbb{A}} p(x, a, y) \sigma(x, da), \qquad x, y \in \mathbb{X},$$

the corresponding transition operator is P^σ:

$$P^\sigma f(x) = \int\limits_{\mathbb{X}} p^\sigma(x, y) f(y) m(dy).$$

Finally, we need multistep transition densities corresponding to a *randomized Markov policy* $\pi = \{\sigma_1, \sigma_2, \ldots\} \in \Pi^{RM}$ where $\sigma_t \in \Pi^{RS}$. They are defined recursively by the formulas

$$p_1^\pi(x, y) = p^{\sigma_1}(x, y), \qquad p_{t+1}^\pi(x, y) = \int\limits_{\mathbb{X}} p_t^\pi(x, z) p^{\sigma_{t+1}}(z, y) m(dz).$$

We also have a bounding function μ on X and the corresponding μ-norms (see Section 2.3). In the definition of the space V_μ it is understood that $f \in V_\mu$ is measurable (throughout this section measurability means Borel measurability).

Assumption 26(a) $\mathbb{X}$ *is a standard Borel space with a σ-finite measure m in it, $\mathbb{A}$ is a Borel set in a Polish (= complete separable metric) space, the set $\mathbb{K}$ (see (1)) is measureable in $\mathbb{X} \times \mathbb{A}$, transition densities $p(x, a, y) \geq 0$ and rewards $r(x, a)$ are measurable functions on $\mathbb{K} \times \mathbb{X}$ and $\mathbb{K}$ respectively, $\mu(x) \geq 1$ is a measurable function on $\mathbb{X}$.*

(b) *$\mathbb{A}(x)$, $x \in \mathbb{X}$ are nonempty compact sets, functions $p(x, a, y)$ and $r(x, a)$ are continuous in $a \in \mathbb{A}(x)$ for every $x, y \in \mathbb{X}$.*

(c) $\|\hat{r}\|_\mu < \infty$ (cf. (44)), *and for some constant $C > 0$*

$$\int_{\mathbb{X}} \hat{p}(x, y)\mu(y) \leq C\mu(x), \qquad x \in \mathbb{X}.$$

(d) *Operators $(P^\sigma)^t, \sigma \in \Pi^{RS}$ converge in the μ-norm to limiting operators Q^σ geometrically fast and uniformly in σ as $t \to \infty$: there exist positive constants $C < \infty$ and $\gamma < 1$ such that*

$$\|(P^\sigma)^t - Q^\sigma\|_\mu \leq C\gamma^t, \qquad \sigma \in \Pi^{RS}, t = 0, 1, 2, \ldots$$

The following result is proved in [24].

Theorem 27 *If Assumption 26 holds, then there exists a stationary policy $\varphi \in \Pi^s$ Blackwell optimal in the class Π^{RS} of randomized stationary policies. Also, all assertions of Theorem 23 hold (for policies $\varphi, \psi \in \Pi^S$ or Π^{RS}).*

Some partial results are true under milder assumptions. For example, Laurent series expansion of $v_\beta(\sigma)$ for $\sigma \in \Pi^{RS}$ and the analogue of Theorem 23 are valid under Assumption 16'(w) or 16(w), and also the Laurent series expansion of $v_\beta(\sigma)$ is valid under an analogue of Assumption 16' in place of Assumption 26(d). For Blackwell optimality in the class Π of all policies in general we need the following uniform integrability assumption.

Assumption 28 *For every $x \in \mathbb{X}$, randomized Markov policy $\pi \in \Pi^{RM}$, and $\epsilon > 0$, there exist a set $Y \subset \mathbb{X}$ with $m(Y) < \infty$ and a constant $L > 0$ such that*

$$\int_{\mathbb{X}\setminus Y} p_t^\pi(x, y)\mu(y)m(dy) < \epsilon, \qquad t = 1, 2, 3, \ldots,$$

$$p_t^\pi(x, y)\mu(y) \leq L, \qquad y \in Y, \quad t = 1, 2, 3, \ldots.$$

The following result is proved in [25].

Theorem 29 *Under Assumptions 26 and 28, every policy $\varphi \in \Pi^s$ Blackwell optimal in the class Π^{RS}, is Blackwell optimal in the class Π as well.*

In the earlier work [48], results of Theorems 27 and 29 were obtained in the case of *bounded transition densities* $p(x, a, y)$, *bounded rewards* $r(x, a)$ (so that one may take

$\mu \equiv 1$), a *finite measure* m, and the following *minorant condition*: there exist a set Y with $m(Y) > 0$ and a number $\delta > 0$ such that

$$p(x, a, y) \geq \delta, \qquad (x, a) \in \mathbb{K}, \quad y \in Y.$$

In that case Assumptions 26(c) and 28 hold trivially, while the geometric convergence as in 26(d) is shown to be true even for the densities $p_t^\sigma(x, y)$. Of course, one has to suppose Assumptions 26(a,b) (with $\mu = 1$).

In related papers [45], [46] Yushkevich proved a partial expansion

$$V_\beta(x) = (1+\rho)\left(\frac{h_{-1}}{\rho} + h_0\right) + o(1)$$

(cf.(43)) for Borel models satisfying assumptions of the preceding paragraph.

3.2. Specific features of Borel models. In the study of Blackwell optimality in Borel state models there are several features which make it different from that of finite and denumerable CMPs. They are: (i) utilization of the class Π^{RS} instead of Π^S in the Laurent series expansions and related topics; (ii) introduction of the weak-strong topology in the space Π^{RS} based on Carathéodory functions, (iii) lexicographical maximization of expected discounted rewards not pointwise at every state but for some absolutely continuous initial distribution; (iv) utilization of the policy improvement to get Blackwell optimal policy $\varphi \in \Pi^s$ from a maximizing policy $\sigma \in \Pi^{RS}$.

We have to work with the class Π^{RS} instead of Π^S because the latter is not a compact space in a reasonable sense. It should be clear from the following simple example. Let $\mathbb{X} = [0, 1)$ and $\mathbb{A} = \mathbb{A}(x) = \{1, 2\}$. For every $m = 1, 2, \ldots$ let the stationary policy φ_m be defined by the rule: if $x \in [(k-1)2^{-m}, k2^{-m})$ then $\varphi_m(x) = 1$ for odd values of k and $\varphi_m(x) = 2$ for even values of k. Every $\varphi_m \in \Pi^s$, but the only reasonable limit of the sequence $\varphi_1, \varphi_2, \ldots$ is the randomized policy $\sigma \in \Pi^{RS}$ with the distribution $\sigma(1 \mid x) = \sigma(2 \mid x) = 1/2$.

Under assumptions of Section 3.1, Laurent series expansions for $v_\beta(\sigma)$ with $h^\sigma \in \mathfrak{H}_\mu$ as in Lemma 20 are valid for $\sigma \in \Pi^{RS}$. Along the same way as in denumerable models, with only technical differences, one justifies the lexicographical policy improvement, and this leads to an analogue of Theorem 23, (a) to (d), but for policies σ in the whole space $\Pi^{RS} \supset \Pi^S$.

In denumerable models we used the product topology in the space Π^S defined in Section 2.1. The appropriate topology in Π^{RS} is the so-called weak-strong or ws-topology. In this topology $\sigma_m \to \sigma$ iff

$$\lim_{m\to\infty} \int\limits_{\mathbb{K}} f(x, a)\sigma_m(x, da)m(dx) = \int\limits_{\mathbb{K}} f(x, a)\sigma(x, da)m(dx)$$

for all Carathéodory functions f (i.e. functions continuous in a, measurable in x) satisfying some bounding condition in terms of $\hat{f}(x)$ and measure m. In this topology, Π^{RS} is a compact space with all needed properties. In the deterministic control theory essentially the same fact was used by Warga [42] in connection with relaxed controls. In the nonstationary stochastic dynamic programming the related compactness of the set of all measures corresponding to a given initial distribution in the ws-topology was proved by Schäl [32],[33] and Balder [2]. Another proof, especially for the space Π^{RS}, is given in Yushkevich [49]. Compactness proved in the above references covers the case of a finite reference measure m. For the σ-finite measure m it is proved in [24].

Assumptions of Section 3.1 imply the continuity in $\sigma \in \Pi^{RS}$ of the operators P^σ and Q^σ, and after that, through formulas for h_n^σ as in Theorem 3 and the power series in P^σ for $(D^\sigma)^n$ obtained from (7), the continuity of the coefficients $h_n^\sigma(x)$. This implies the continuity of $h_n^\sigma(\ell) = \int\limits_{\mathbb{X}} h_n^\sigma(x)\ell(x)m(dx)$ for any initial density ℓ. Taking a strictly positive density ℓ on $\mathbb{X}$, we lexicographically maximize $h^\sigma(\ell) = \{h_n^\sigma(\ell), n \geq -1\}$ over Π^{RS}, and get a "best" policy σ^* for the initial distribution ℓ.

Lexicographical policy improvement applied to σ^* at all states X where it is possible to improve, provides a policy $\varphi \in \Pi^s$. With the help of the already proven part of Theorem 23, it is now not difficult to show that φ is Blackwell optimal in the class Π^{RS}, and that the Blackwell optimality equation has a unique solution in $\mathfrak{H}_\mu$.

The proof that a policy Blackwell optimal in Π^{RS} is Blackwell optimal in the whole space Π is even more technical than in the denumerable case. It utilizes the main idea of the proof in [8], Assumption 28, and an additional property of the ws-topology; see [25], or for the special case of bounded rewards, [44] or [48].

3.3. Recurence conditions for Blackwell optimality. The uniform geometric μ-ergodicity conditon and the uniform integrability condition (Assumptions 26(d) and 28) are difficult for a verification in CMPs with a noncompact state space and an unbounded reward function. In Hordijk and Yushkevich [25] simpler recurrence and drift conditions are given, which imply those assumptions. This approach is based on ideas developed in Hordijk and Spieksma [22] and Hordijk et al. [23], with an additional use of the weak-strong topology. Consider the following set of conditions.

Assumption 30 (a) (Uniform minorant condition) *There exist sets $D, Y \subset \mathbb{X}$ with $m(D) > 0$, $m(Y) > 0$ and a number $\delta > 0$ such that*

$$p(x,a,y) \geq \delta, \qquad x \in D, \quad a \in \mathbb{A}(x), \quad y \in Y.$$

(b) (Uniform drift condition) *There exist a set $D \subset \mathbb{X}$ with $m(D) > 0$ and numbers $b > 0$, $0 < \gamma < 1$ such that*

$$\sup_{x \in D} \mu(x) < \infty$$

and

$$\int_{\mathbb{X}} p(x,a,y)\mu(y)m(dy) \le \gamma\mu(x) + b\mathbf{1}_D(x), \qquad (x,a) \in \mathbb{K}.$$

(c) (Uniform accessibility condition) *There exists a set $D \subset \mathbb{X}$, and for every sublevel set*

$$M_c = \{x : \mu(x) \le c\}$$

there exist a number $\eta > 0$ and an integer N such that

$$\int_D P_N^\sigma(x,y)m(dy) \ge \eta, \qquad x \in M_c, \quad \sigma \in \Pi^{RS}.$$

(d) (Dominance integrability condition) *There exist a set $D \subset \mathbb{X}$ with $m(D) > 0$ and a measurable function $\ell \ge 0$ and $\mathbb{X}$ such that*

$$\int_{\mathbb{X}} \ell(x)\mu(x)m(dx) < \infty \quad \text{and} \quad \hat{p}(x,y) \le \ell(y), \quad x \in D, \quad y \in \mathbb{X},$$

and also $m(M_c) < \infty$ for every sublevel set M_c with $c \ge 1$.

Omitting some details, we summarize those relations between conditions which provide the existence of Blackwell optimal policies. It follows from [23], that Assumptions 30 (a,b,c) with the same set D, together with 26(a) and the condition $P^a\mu(x) \le C\mu(x)$ imply the uniform integrability Assumption 26(d). Also, if the density $p(x,a,y)$ is bounded, Assumptions 30(b,d) together with 26(a,b,c) imply Assumption 28 (with a possible change of the function μ, which does not affect the made assumptions). The proof of the last result essentially follows the proof of a similar result for denumerable models in Dekker et al. [11], with the use of ws-topology. As a consequence, we have the following theorem.

Theorem 31 *In CMP satisfying Assumptions* 26(a,b,c) *and* 30(a,b,c), *there exists a stationary policy $\varphi \in \Pi^s$ Blackwell optimal in the class Π^{RS}, and all assertions of Theorem* 23 *hold.*

If in addition the transition density $p(x,a,y)$ is bounded and Assumption 30(d) *holds, then φ is Blackwell optimal in the class Π as well.*

Bibliography

[1] E. Altman, A. Hordijk and L.C.M. Kallenberg, "On the value function in constrained control of Markov chains", *Mathematical Methods of Operations Research* **44**, 387-399, 1996.

[2] E.I. Balder, "On compactness of the space of policies in stochastic dynamic programming", *Stochastic Processes and Applications* **32**, 141-150, 1989.

[3] D. Blackwell, "Discrete dynamic programming", *Annals of Mathematical Statistics* **33**, 719-726, 1962.

[4] R. Cavazos-Cadena and J.B. Lasserre, "Strong 1-optimal stationary policies in denumerable Markov decision processes", *Systems and Control Letters* **11**, 65-71, 1988.

[5] R. Cavazos-Cadena and J.B. Lasserre, "A direct approach to Blackwell optimality", 1-16, Preprint, 1993.

[6] R.Ya. Chitashvili, "A controlled finite Markov chain with an arbitrary set of decisions", *Theory of Probability and Its Applications* **20**, 839-846, 1975.

[7] R.Ya. Chitashvili, "A finite controlled Markov chain with small termination probability", *Theory of Probability and Its Applications* **21**, 158-163, 1976.

[8] R. Dekker and A. Hordijk, "Average, sensitive and Blackwell optimal policies in denumerable Markov decision chains with unbounded rewards", *Mathematics of Operations Research* **13**, 395-421, 1988.

[9] R. Dekker and A. Hordijk. "Denumerable semi-Markov decision chains with small interest rates", *Annals of Operations Research* **28**, 185-212, 1991.

[10] R. Dekker and A. Hordijk, "Recurrence conditions for average and Blackwell optimality in denumerable state Markov decision chains", *Mathematics of Operations Research* **17**, 271-289, 1992.

[11] R. Dekker, A. Hordijk and F.M. Spieksma, "On the relation between recurrence and ergodicity properties in denumerable Markov decision chains", *Mathematics of Operations Research* **19**, 539-559, 1994.

[12] E.V. Denardo, "Markov renewal programming with small interest rates", *Annals of Mathematical Statistics* **42**, 477-496, 1971.

[13] A. Federgruen, A. Hordijk and H.C. Tijms, "A note on simultaneous recurrence conditions on a set of denumerable stochastic matrices", *Journal of Applied Probability* **15**, 842-847, 1978.

[14] A. Federgruen, A. Hordijk and H.C. Tijms, "Recurrence conditions in denumerable state Markov decision processes", in *Dynamic Programming and Its Applications*, edited by M.L. Puterman, 3-22, Academic Press, 1978.

[15] O. Hernández-Lerma, R. Montes-de-Oca and R. Cavazos-Cadena, "Recurrence conditions for Markov decision processes with Borel state space: a survey", *Annals of Operations Research* **28**, 29-46, 1991.

[16] A. Hordijk, *Dynamic Programming and Markov Potential Theory*, Mathematical Centre Tract **51**, Mathematisch Centrum, 1974.

[17] A. Hordijk, "Regenerative Markov decision models", in *Mathematical Programming Study*, **6**, edited by R.J.B. Wets, North Holland, 1976, 49-72.

[18] A. Hordijk, R. Dekker and L.C.M. Kallenberg, "Sensitivity-analysis in discounted Markovian decision problems", *Operations Research Spektrum* **7**, 143-151, 1985.

[19] A. Hordijk, O. Passchier and F.M. Spieksma, "On the existence of the Puisseux expansion of the discounted rewards: a counterexample", *Probability in the Engineering and Informational Sciences* **13**, 229-235, 1999.

[20] A. Hordijk and K. Sladký, "Sensitive optimality criteria in countable state dynamic programming", *Mathematics of Operations Research* **2**, 1-14, 1977.

[21] A. Hordijk and F.M. Spieksma, "Are limits of α-discounted optimal policies Blackwell optimal? A counterexample", *Systems and Control Letters*, **13**, 31-41, 1989.

[22] A. Hordijk and F.M. Spieksma, "On ergodicity and recurrence properties of a Markov chain with an application to an open Jackson network", *Advances in Applied Probability* **24**, 343-376, 1992.

[23] A. Hordijk, F.M. Spieksma and R.L. Tweedie, "Uniform stability conditions for general space Markov decision processes", Technical report, Leiden University and Colorado State University, 1995.

[24] A. Hordijk and A.A. Yushkevich, "Blackwell optimality in the class of stationary policies in Markov decision chains with a Borel state space and unbounded rewards", *Mathematical Methods of Operations Research* **49**, 1-39, 1999.

[25] A. Hordijk and A.A. Yushkevich, "Blackwell optimality in the class of all policies in Markov decision chains with a Borel state space and unbounded rewards", *Mathematical Methods of Operations Research* (to appear).

[26] R.A. Howard, *Dynamic Programming and Markov Processes*, Wiley, 1960.

[27] J.G. Kemeny and J.L. Snell, *Finite Markov chains*, Van Nostrand-Reinhold, 1960.

[28] J.B. Lasserre, "Conditions for existence of average and Blackwell optimal stationary policies in denumerable Markov decision processes", *Journal of Mathematical Analysis and Applications* **136**, 479-490, 1988.

[29] S.P. Meyn and R.L. Tweedie, *Markov Chains and Stochastic Stability*, Springer, 1993.

[30] B.L. Miller and A.F. Veinott, "Discrete dynamic programming with a small interest rate", *Annals of Mathematical Statistics* **40**, 366-370, 1969.

[31] S.M. Ross, "Non-discounted denumerable Markovian decision models", *Annals of Mathematical Statistics* **39**, 412-423, 1968.

[32] M. Schäl, "On dynamic programming: Compactness of the space of policies", *Stochastic Processes and Applications* **3**, 345-354, 1975.

[33] M. Schäl, “On dynamic programming and statistical decision theory”, *Annals of Statistics* **7**, 432-445, 1979.

[34] K. Sladký, “On the set of optimal controls for Markov chains with rewards”, *Kybernetika* (Prague) **10**, 350-367, 1974.

[35] F.M. Spieksma, “Geometrically ergodic Markov chains and the optimal control of queues”, Ph.D. Thesis, University of Leiden, 1990.

[36] L.C. Thomas, “Connectedness conditions for denumerable state Markov decision processes”, in *Recent Developments in Markov Decision Processes*, edited by R. Hartley, L.C. Thomas, D.J. White, Academic Press, 1980, 181-204.

[37] H.C. Tijms , “Average reward optimality equation in Markov decision processes with a general state space”, in *Probability, Statistics and Optimization: a Tribute to Peter Whittle*, edited by F.P. Kelly, Wiley, 1994, 485-495.

[38] A.F. Veinott Jr., “On finding optimal policies in discrete dynamic programming with no discounting”, *Annals of Mathematical Statistics* **37**, 1284-1294.

[39] A.F. Veinott Jr., “Discrete dynamic programming with sensitive optimality criteria”, *Annals of Mathematical Statistics* **40**, 1635-1660.

[40] A.F. Veinott Jr., *Dynamic Programming and Stochastic Control*, Unpublished class notes.

[41] H.M. Wagner, “On optimality of pure strategies”, *Management Science* **6**, 268-269, 1960.

[42] J. Warga, *Optimal Control of Differential and Functional Equations*, Academic Press, 1972.

[43] K. Yosida, *Functional Analysis*, Springer, 1980.

[44] A.A. Yushkevich, “Blackwell optimal policies in a Markov decision process with a Borel state space”, *Mathematical Methods of Operations Research* **40**, 253-288, 1994.

[45] A.A. Yushkevich, “Strong 0-discount optimal policies in a Markov decision process with a Borel state space”, *Mathematical Methods of Operations Research* **42**, 93-108, 1995.

[46] A.A. Yushkevich, “A note on asymptotics of discounted value function and strong 0-discount optimality”, *Mathematical Methods of Operations Research* **44**, 223-231, 1996.

[47] A.A. Yushkevich, “Blackwell optimal policies in countable dynamic programming without aperiodicity assumptions”, in *Statistics, Probability and Game Theory: Papers in Honor of David Blackwell*, edited by T.S. Ferguson, L.S. Shapley and J.B. MacQueen, Institute of Mathematical Statistics, 1996, 401-407.

[48] A.A. Yushkevich, “Blackwell optimality in Markov decision processes with a Borel state space”, *Proceedings of 36th IEEE Conference on Decision and Control* **3**, 2827-2830, 1997.

[49] A.A. Yushkevich, "The compactness of a policy space in dynamic programming via an extension theorem for Carathéodory functions", *Mathematics of Operations Research* **22**, 458-467, 1997.

[50] A.A. Yushkevich, "Blackwell optimality in Borelian continuous-in-action Markov decision processes", *SIAM Journal on Control and Optimization* **35**, 2157-2182, 1997.

[51] A.A. Yushkevich and R.Ya. Chitashvili, "Controlled random sequences and Markov chains", *Russian Mathematical Surveys* **37**, 239-274, 1982.

[52] J. Flynn, "Averaging vs. discounting in dynamic programming: a counterexample", *Annals of Statistics* **2**, 411-413, 1974.

[53] A. Maitra, "Dynamic programming for countable state systems", *Sankhya Ser. A* **27**, 241-248, 1965.