8 BLACKWELL OPTIMALITY

Arie Hordijk
Alexander A. Yushkevich

8.1 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. In Chapter 2 the algorithmic approach to Blackwell optimality for finite models is given. We refer to that chapter for computational methods. Especially for the linear programming method, which we do not introduce.

8.1.1 *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 $p_{xy}(\varphi(x)) = p_{xy}^\varphi$, and (3) can be written in the form

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

where r^φ is a vector with entries $r(x,\varphi(x))$, $x \in \mathbb{X}$ (formula (4) makes 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* (or *limiting*) *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 [27] 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 [46] determined that such policies are average optimal in the class Π too.

However, the average reward criterion is insensitive, under selective 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 8.1 *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 8.1), and since by Theorem 8.1 below there is a stationary Blackwell optimal policy, *every Blackwell optimal policy $\pi \in \Pi$ is average optimal.*

Theorem 8.1 *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

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

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 [29]). 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 $v_\beta^\pi = h_{-1}/\alpha + h_0 + o(1)$.

8.1.2 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 [33] and Veinott [42], 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, [48]). 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 $\|P^n - Q\| \leq C\gamma^n$ 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 [42] and most of the subsequent authors used series in ρ. Chitashvili [6, 7, 56] and following him Yushkevich [49]–[55] used series in α. We present both versions.

Theorem 8.2 *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 8.1. ∎

Formulas of Theorem 8.2 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 [43].

Definition 8.2 *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 [42]. 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. For bias optimality in models with finite state and action spaces see Chapter 3.

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

$$H^\varphi = \{h^\varphi_{-1}, h^\varphi_0, \dots\}, \quad K^\varphi = \{k^\varphi_{-1}, k^\varphi_0, \dots\}, \tag{25}$$

let H^φ_n and K^φ_n 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^φ_n 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 $H^\varphi_{N-1} = H^\psi_{N-1}$ (if $N \geq 0$), and $h^\varphi_N(x_0) < h^\psi_N(x_0)$ while $h^\varphi_N(x) \leq h^\psi_N(x)$ 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 $H^\varphi_n = H_n$ or $K^\varphi_n = K_n$ (respectively, if $H^\varphi = H$ or $K^\varphi = K$).

The following theorem due to Veinott [43] 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 8.3 *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 $H^\varphi_{m-1} = H_{m-1}$, or in more detail

$$h^\varphi_n = h_n, \qquad n = -1, 0, 1, \dots, m-1. \tag{26}$$

Since $m \geq 2$, both h_0 and h_1 are present in (26). We claim that m column m-vectors $h_0, h_1, \dots, h_{m-1}$ are linearly dependent. It is sufficient to show that m row vectors of the corresponding square matrix are linearly dependent; these rows are $\{h_0(x), \dots, h_{m-1}(x)\} = \{h^\varphi_0(x), \dots, h^\varphi_{m-1}(x)\}$, $x \in \mathbb{X}$. In fact even the infinite sequences

$$\{h^\varphi_0(x), h^\varphi_1(x), \dots, h^\varphi_t(x), \dots\}, \qquad x \in \mathbb{X} \tag{27}$$

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

$$v_\beta^\varphi(\mu) := \sum_{x\in\mathbb{X}} \mu(x) v_\beta^\varphi(x) = \sum_{x\in\mathbb{X}} \mu(x)\, \mathbb{E}_x^\varphi \sum_{t=0}^{\infty} \beta^t r(x_t, \varphi(x_t)) =$$

$$= \sum_{t=0}^{\infty} \beta^t \sum_{x\in\mathbb{X}} \mu(x)\, \mathbb{E}_x^\varphi\, r(x_t, \varphi(x_t)) = \sum_{t=0}^{\infty} \beta^t\, \mathbb{E}_\mu^\varphi\, r(x_t, \varphi(x_t)). \quad (28)$$

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

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

(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 $\sum_x \mu(x) h_n^\varphi(x) = 0$ for all $n \geq 0$, so that the sequences (27) are linearly dependent.

Now, by (18)

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

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 $t \geq 0$ exists if only $h_0 \neq 0$, 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 $h \geq 0$, 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 $h_0 = h_0^\varphi, \dots, h_t = h_t^\varphi$:

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

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 $0 \leq i \leq t+1$, we get $h_{t+2}^\varphi = h_{t+2}$. Repeating this, by induction we get $h_n^\varphi = h_n$ for every $n \geq 0$, so that $\varphi \in \Phi_\infty$. ∎

The sets Φ_{m-2} and Φ_{m-1} are in general different. The following example, taken from [43], confirms this statement. To make it more visual, we present it for $m = 5$.

Example 8.1 *There are $m = 5$ states $1, 2, \dots, 5$ with mandatory transitions $2 \to 3 \to 4 \to 5$, the state 5 is absorbing. In state 1 there is a choice between*

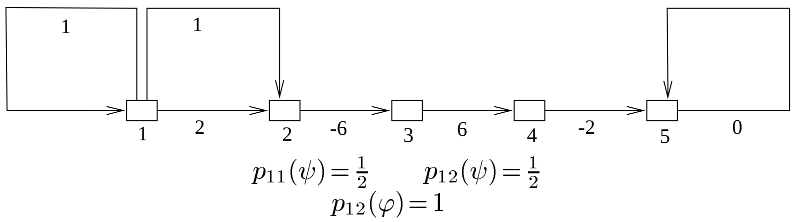

Figure 8.1

two actions, which determine two stationary policies φ and ψ (see Figure 8.1). 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.

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

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

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 $\Phi_3 \neq \Phi_4 = \Phi_\infty$.

8.1.3 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 essence 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 [27] for the average reward criterion. Howard used, besides the average reward $w^\varphi(= h_{-1}^\varphi)$, a second function, in fact equal to the term h_0^φ in the expansions (10) and (16) up to a constant term on each recurrence class of the Markov chain generated by φ. Blackwell [3] provided a rigorous proof that a slightly different version of Howard's policy improvement method does converge. Miller and Veinott [33] have extended policy improvement to the case of Blackwell optimality, and Veinott [43] 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 (13) we have

$$v_\beta^\pi = r^\psi + \beta P^\psi v_\beta^\varphi = r^\psi + \frac{1}{1+\rho} P^\psi v_\beta^\varphi = r^\psi + P^\psi \sum_{n=-1}^{\infty} h_n^\varphi \rho^n,$$

while

$$v_\beta^\varphi = h_{-1}^\varphi \rho^{-1} + \sum_{n=0}^{\infty} (h_n^\varphi + h_{n-1}^\varphi)\rho^n.$$

Subtracting, we get

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

where it is understood that $h_n = h_n^\varphi$. By (18), the supremum norm $\|h_n^\varphi\|$ 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

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

$$L^a H(x) = \{\ell h_{-1}^a(x), \ell h_0^a(x), \ell h_1^a(x), \dots\}, \qquad (x,a) \in \mathbb{K}, \tag{34}$$

where according to (32)

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

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

Lemma 8.1 *Let $\varphi, \psi \in \Pi^s$. If $(L^\psi H^\varphi)_{n+1} \succeq 0$ for some $n \geq -1$, then $H_n^\psi \succeq H_n^\varphi$. Moreover, if in addition $(L^\psi H^\varphi)_{n+1}(x_0) \succ 0$ at some $x_0 \in \mathbb{X}$, then $H_n^\psi(x_0) \succ H_n^\varphi(x_0)$. 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

$$v_\beta^\pi = v_\beta^\varphi + Q_n(\rho) + O(\rho^{n+1}) \tag{36}$$

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, $v(0) = v_\beta^\varphi$, $v(1) = v_\beta^\pi$. 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 $v(\infty) = v_\beta^\psi$, so that

$$v_\beta^\psi = v_\beta^\varphi + \sum_{t=0}^{\infty} (\beta P^\psi)^t (Q_n(\rho) + R) = v_\beta^\varphi + R_\beta^\psi (Q_n + R).$$

Here $Q_n \geq 0$ for small $\rho > 0$, R is of order $O(\rho^{n+1})$, and the resolvent R_β^ψ is of order $O(1 + \beta + \beta^2 + \cdots) = O(\rho^{-1})$. This proves that $H_{n-1}^\psi \succeq H_{n-1}^\varphi$ if $L^\psi H_n^\varphi \succeq 0$. 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 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 8.4 *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 8.1 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 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 8.1 (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 8.1 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 8.1. ∎

Policy improvement is a basis for an algorithm to compute a Blackwell optimal policy in a finite CMP. Theoretically, one may proceed in the following way. Start with some $\varphi \in \Pi^s$ and compute H_m^φ using formulas of Theorem 8.2 (here m is the number of states in $\mathbb{X}$). Check the values of $\ell_{-1}^a h(x), (x,a) \in \mathbb{K}$. 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 8.3. 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^*; see Policy Iteration in chapter 10 of [34].

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 8.1, based on the fact that the set Π^s is finite, is inapplicable.

8.1.4 Extensions and generalizations

In [45] Veinott simplified and updated results of [42, 43]. In particular, he refined Theorem 8.3 as follows: $\Phi_\infty = \Phi_{m-r}$ where r is the number of recurrent classes in X under a Blackwell optimal stationary policy.

Veinott [43] introduced also the notion of n-average optimality in addition to the n-discount optimality. Let

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

and define recursively for $n \geq 1$

$$v_T^{(n+1)}(x,\pi) = \sum_{t=1}^{T} v_t^{(n)}(x,\pi).$$

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

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

Veinott [43, 44] and Sladky [38] showed that in a finite CMP a policy is n-discount optimal iff it is n-average optimal.

Chitashvili [6, 7, 56] extended results of Theorem 8.4 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 [56] Theorem 8.4 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 [56] 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 8.4 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 8.1, in finite CMPs the expected discounted reward $v_\beta^\varphi(x)$ is a rational function of β. Hordijk e.a. [19] 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 $\beta \in [\beta_{j-1}, \beta_j]$, $1 \leq j \leq m-1$, and φ_m is β-optimal in the interval $[\beta_{m-1}, \beta_m)$ (which means that φ_m is Blackwell optimal). In section 5.4 of Chapter 2 this algorithm is studied. In Chapter 4 an asymptotic simplex method based on Laurent series expansions for the computation of a Blackwell optimal policy, is used.

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 [19]. 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 $[\beta_{j-1}, \beta_j]$, $j \leq m-1$ 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 8.1, the limits of β-optimal policies, for β tending to 1, are Blackwell optimal. A counterexample in Hordijk and Spieksma [22] 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].

More recently, Huang and Veinott [28] extended many results concerning Blackwell and n-discount optimality in finite models to the case when (i) the reward $r(x, a, \rho)$ at the state x under the action a depends also on the interest rate ρ, namely is an analytic function of ρ in a neighborhood of $\rho = 0$, (ii) the nonnegative transition coefficients $p_{xy}(a)$ in general do not sum to 1, only for every Markov policy the t-step transition matrices are of order $O(1)$ as $t \to \infty$. They proved the existence of stationary Blackwell optimal policies, extended to their case the lexicographical policy improvement, showed that if $r(x, a, \rho)$ is a polynomial or a rational function of ρ then Blackwell optimality of a stationary policy is equivalent to its n-discount optimality for some n depending on the degrees of involved polynomials. In the latter case the policy iteration algorithm provides a Blackwell optimal policy in a finite number of steps. Another constructive rule is given for finding n-discount optimal policies using a linear programming approach.

8.2 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