J. Appl. Prob. **46,** 372–391 (2009)
Printed in England

BLACKWELL OPTIMALITY FOR CONTROLLED DIFFUSION PROCESSES

HÉCTOR JASSO-FUENTES * ** AND
ONÉSIMO HERNÁNDEZ-LERMA,* *** *CINVESTAV*

Abstract

In this paper we study m-discount optimality ($m \geq -1$) and Blackwell optimality for a general class of controlled (Markov) diffusion processes. To this end, a key step is to express the expected discounted reward function as a Laurent series, and then search certain control policies that lexicographically maximize the mth coefficient of this series for $m = -1, 0, 1, \ldots$. This approach naturally leads to m-discount optimality and it gives Blackwell optimality in the limit as $m \to \infty$.

Keywords: Controlled diffusions; average reward; Laurent series; sensitive discount optimality; Blackwell optimality

2000 Mathematics Subject Classification: Primary 93E20; 60J60

1. Introduction

The most common optimality criteria for infinite-horizon optimal control problems are the *expected discounted reward* criterion and the *long-run expected average reward* criterion, also known as the *ergodic reward* or *gain*. These two criteria have opposite aims: the former considers early periods of the infinite time horizon (since it essentially vanishes for large time intervals—see (2.7)), whereas the latter concerns only the asymptotic behavior; it simply ignores what happens in finite time intervals. To avoid these two extremal situations, we must consider refinements of the average reward criterion such as overtaking optimality, bias optimality, and the so-called sensitive discount criteria, which include m-discount optimality for an integer $m \geq -1$ and Blackwell optimality for $m = +\infty$. They are called 'refinements' because they concern control policies that optimize the average reward and, in addition, they have some other convenient features. In this work we are interested in some of these refinements. Namely, we will give conditions that guarantee m-discount optimality for every integer $m \geq -1$ and also for Blackwell optimality when the controlled system is a Markov diffusion process of the form

$$\mathrm{d}x(t) = b(x(t), u(t))\,\mathrm{d}t + \sigma(x(t))\,\mathrm{d}B(t) \quad \text{for all } t \geq 0 \text{ and } x(0) = x,$$

where $B(\cdot)$ is a d-dimensional Brownian motion, and the coefficients $b(x, u)$ and $\sigma(x)$ satisfy suitable assumptions. (See Section 2 for details.)

Blackwell optimality was introduced by Blackwell [3], whose work deals in fact with a more restrictive concept, known in the literature as *strong Blackwell optimality*, for discrete-time Markov decision processes (MDPs) with finite state space and finite-action spaces. (See

Received 17 January 2008; revision received 6 March 2008.
* Postal address: Department of Mathematics, CINVESTAV-IPN, A. Postal 14-740, Mexico DF 07000, Mexico.
Research partially supported by CONACYT grant 45693-F.
** Email address: hjasso@math.cinvestav.mx
Research also supported by a CONACYT scholarship.
*** Email address: ohernand@math.cinvestav.mx

also [29]). In this work we use a weaker concept, called simply *Blackwell optimality* (see Definition 2.3).

Blackwell optimality for discrete-time MDPs with denumerable state space was studied in [5]. Hordijk and Yushevich [16] presented a fairly complete description of known results on Blackwell optimality for discrete-time MDPs. Special cases of strong m-discount optimality ($m = -1, 0$) for discrete-time models has been studied in [15] and [32] (see also [14]). On the other hand, for continuous-time models, Puterman [27] considered one-dimensional diffusion processes with values in a compact interval, whereas Prieto-Rumeau and Hernández-Lerma [25] and Prieto-Rumeau [23] considered continuous-time controlled Markov chains with a *denumerable* state space. Finally, Jasso-Fuentes and Hernández-Lerma [19] studied special cases of strong m-discount optimality for the cases $m = -1, 0$ for controlled diffusion processes. Our approach in this paper is mainly based on [19] and [23].

The remainder of this paper is organized as follows. In Section 2 we introduce the control system and our main assumptions. In addition, we define the optimality criteria we are concerned with, and we summarize some known results on the Hamilton–Jacobi–Bellman (HJB) equation [2], [4], [8], [10], [17], [18], which is essentially our point of departure to analyze m-discount optimality and Blackwell optimality. In Section 3 we express the expected α-discounted v-reward (see (3.6)) for some function v as a Laurent series (see (3.11)). In Section 4 we define the so-called -1th, 0th, ..., mth Poisson equations and the -1th, 0th, ..., mth average reward HJB equations. In addition, we ensure the existence of solutions to these equations and the existence of policies that maximize the -1th, 0th, ..., mth average reward HJB equations (4.4)–(4.6). Section 5 concerns the existence of m-discount policies *for any* integer $m \geq -1$, which leads to the proof of the existence of Blackwell optimal policies. In Section 6 we present an example that illustrates our results. Finally, we conclude in Section 7 with some remarks.

Throughout the following sections, for vectors and matrices, we use the usual norms

$$|x|^2 := \sum_i x_i^2 \quad \text{and} \quad |A|^2 := \operatorname{tr}(AA^\top) = \sum_{i,j} A_{ij}^2,$$

where $A^\top$ and $\operatorname{tr}(\cdot)$ denote the *transpose* of $A = (A_{ij})$ and the *trace* of a square matrix, respectively.

2. Model definition and basic optimality criteria

The control system we are concerned with is the controlled diffusion process

$$\mathrm{d}x(t) = b(x(t), u(t))\,\mathrm{d}t + \sigma(x(t))\,\mathrm{d}B(t) \qquad x(0) = x,\ t \geq 0, \tag{2.1}$$

where $b(\cdot,\cdot)\colon \mathbb{R}^n \times U \to \mathbb{R}^n$ and $\sigma(\cdot)\colon \mathbb{R}^n \to \mathbb{R}^{n\times d}$ are given functions, and $B(\cdot)$ is a d-dimensional Brownian motion. The set $U \subset \mathbb{R}^m$ is called the *control* (or *action*) *set*, and $u(\cdot)$ is a U-valued stochastic process representing the controller's action at each time $t \geq 0$.

Let $\mathbb{F}$ be the set of all measurable functions $f\colon \mathbb{R}^n \to U$, and let $\mathbb{M} \supset \mathbb{F}$ be the family of all U-valued measurable functions on $[0, \infty) \times \mathbb{R}^n$. A function $f \in \mathbb{M}$ is called a *Markov policy*, whereas $f \in \mathbb{F}$ is called a *stationary Markov policy* or simply a *stationary policy*. Under a policy $f \in \mathbb{M}$ or $f \in \mathbb{F}$, the function $u(\cdot)$ in (2.1) becomes $u(t) := f(t, x(t))$ or, respectively, $u(t) := f(x(t))$.

The following assumption ensures that, for each Markov policy $f \in \mathbb{M}$, (2.1) admits an almost surely unique strong solution $x(\cdot) := \{x(t) \mid t \geq 0\}$, which is a Markov–Feller process. For details, see, for instance, [2, Theorem 2.2.7], [9, Theorem 2.1], or [10, Theorem 3.1].

https://doi.org/10.1239/jap/1245676094 Published online by Cambridge University Press

Assumption 2.1. (a) *The control set U is compact.*

(b) *b is continuous on $\mathbb{R}^n \times U$, and there exists a positive constant K such that, for each x and y in $\mathbb{R}^n$,*

$$\sup_{u \in U} |b(x,u) - b(y,u)| \leq K|x - y|.$$

(c) *There exist positive constants K and γ such that, for each x and y in $\mathbb{R}^n$,*

$$|\sigma(x) - \sigma(y)| \leq K|x - y|,$$

and the matrix $a := \sigma\sigma^\top$ satisfies

$$x^\top a(y)x \geq \gamma|x|^2 \quad \textit{for all } x, y \in \mathbb{R}^n \textit{ (uniform ellipticity).}$$

Let $C^2(\mathbb{R}^n)$ be the space of real-valued continuous functions on $\mathbb{R}^n$ with continuous first and second partial derivatives. For $u \in U$ and $h \in C^2(\mathbb{R}^n)$, let

$$L^u h(x) := h_x(x)b(x,u) + \tfrac{1}{2}\mathrm{tr}(h_{xx}(x)a(x)),$$

where $a(\cdot)$ is as in Assumption 2.1(c), and h_x and h_{xx} represent the gradient vector and the Hessian matrix of h, respectively.

For each $f \in \mathbb{F}$ and $x \in \mathbb{R}^n$, let

$$L^f h(x) := L^{f(x)}h(x). \tag{2.2}$$

For a stationary policy $f \in \mathbb{F}$, the operator L^f in (2.2) coincides with the infinitesimal generator associated to the diffusion $x(\cdot)$ in (2.1).

To emphasize the dependence on $f \in \mathbb{F}$, we sometimes write $x(\cdot)$ as $x^f(\cdot)$. Also, we shall denote by $\mathrm{P}_x^f(t,\cdot)$ the corresponding transition probability, i.e.

$$\mathrm{P}_x^f(t,B) = \mathrm{P}(x^f(t) \in B \mid x^f(0) = x)$$

for every Borel set $B \subset \mathbb{R}^n$. The associated conditional expectation is written as $\mathrm{E}_x^f(\cdot)$.

2.1. Ergodicity

The following assumption is a standard Lyapunov stability condition for continuous-time (controlled and uncontrolled) Markov processes—see, for instance, [2], [7], [8], [10], [11], [22], [24], [26], and [30].

Assumption 2.2. *There exist a function $w \geq 1$ in $C^2(\mathbb{R}^n)$ and constants $d \geq c > 0$ such that*

(a) $\lim_{|x|\to\infty} w(x) = +\infty$;

(b) $L^u w(x) \leq -cw(x) + d$ *for all $u \in U$ and $x \in \mathbb{R}^n$.*

Assumption 2.2 gives that, for each $f \in \mathbb{F}$, the Markov process $x^f(\cdot)$ is Harris positive recurrent with a unique invariant probability measure μ_f for which

$$\mu_f(w) := \int_{\mathbb{R}^n} w(y)\mu_f(\mathrm{d}y) < \infty. \tag{2.3}$$

(See [2], [8], [11], [12], and [22] for details). Assumption 2.2 also ensures the boundedness of $\mathrm{E}_x^f(w(x(t)))$ in the following sense. (See [18, Lemma 2.10] or [19, Lemma 2.3].)

https://doi.org/10.1239/jap/1245676094 Published online by Cambridge University Press

Lemma 2.1. *Assumption 2.2(b) implies that*

$$\mathrm{E}_x^f(w(x(t))) \le \mathrm{e}^{-ct}w(x) + \frac{d}{c}(1-\mathrm{e}^{-ct}) \tag{2.4}$$

for every $f \in \mathbb{F}$, $x \in \mathbb{R}^n$, *and* $t \ge 0$.

We now introduce the concept of the w-weighted norm, where w is the function in Assumption 2.2.

Definition 2.1. Let $B_w(\mathbb{R}^n)$ denote the Banach space of real-valued measurable functions v on $\mathbb{R}^n$ with finite w-norm, which is defined as

$$\|v\|_w := \sup_{x\in\mathbb{R}^n} \frac{|v(x)|}{w(x)}.$$

Assumption 2.3, below, concerns the w-exponential ergodicity of $x^f(\cdot)$. Sufficient conditions for Assumption 2.3 are given, for instance, in [17, Theorem 1.3.6] or [18, Theorem 2.7].

Assumption 2.3. *For each* $f \in \mathbb{F}$, *the process* $x(\cdot) \equiv x^f(\cdot)$ *is* w*-exponentially ergodic; that is, there exist positive constants* C *and* δ *such that*

$$\sup_{f\in\mathbb{F}} |\mathrm{E}_x^f(v(x(t))) - \mu_f(v)| \le C\mathrm{e}^{-\delta t}\|v\|_w w(x) \tag{2.5}$$

for all $x \in \mathbb{R}^n$, $v \in B_w(\mathbb{R}^n)$, *and* $t \ge 0$, *where* $\mu_f(v) := \int_{\mathbb{R}^n} v(y)\mu_f(\mathrm{d}y)$.

2.2. Optimization problems

Let $r\colon \mathbb{R}^n \times U \to \mathbb{R}$ be a measurable function, called the *reward rate*, which satisfies the following conditions.

Assumption 2.4. (a) $r(x,u)$ *is continuous on* $\mathbb{R}^n \times U$ *and locally Lipschitz in* x, *uniformly in* $u \in U$, *i.e. for each* $R > 0$, *there exists a constant* $K(R)$ *such that*

$$\sup_{u\in U} |r(x,u) - r(y,u)| \le K(R)|x-y| \quad \textit{for all } |x|, |y| \le R.$$

(b) $r(\cdot,u)$ *is in* $B_w(\mathbb{R}^n)$ *uniformly in* u; *that is, there exists an* $M > 0$ *such that, for all* $x \in \mathbb{R}^n$,

$$\sup_{u\in U} |r(x,u)| \le Mw(x).$$

For each Markov policy $f \in \mathbb{M}$, $t \ge 0$, and $x \in \mathbb{R}^n$, we write

$$r(t,x,f) := r(x, f(t,x)), \tag{2.6}$$

which reduces to $r(x,f) := r(x, f(x))$ if $f \in \mathbb{F}$ is stationary.

Definition 2.2. (*Discounted reward criterion.*) Given the 'discount factor' $\alpha > 0$, let

$$V_\alpha(x,f) := \mathrm{E}_x^f\left(\int_0^\infty \mathrm{e}^{-\alpha t} r(t, x(t), f)\,\mathrm{d}t\right) \tag{2.7}$$

be the *expected* α*-discounted reward*, when using the policy $f \in \mathbb{M}$, given the initial state $x \in \mathbb{R}^n$. The corresponding optimal value function is

$$V_\alpha^*(x) = \sup_{f\in\mathbb{M}} V_\alpha(x,f).$$

https://doi.org/10.1239/jap/1245676094 Published online by Cambridge University Press

A stationary policy f_α^* is said to be *α-discount optimal* if

$$V_\alpha^*(x) = V_\alpha(x, f_\alpha^*) \quad \text{for all } x \in \mathbb{R}^n.$$

Assumptions 2.1, 2.2, and 2.4 ensure the existence of α-discount optimal policies in the class $\mathbb{F}$ of stationary policies. (See, for instance, [2], [9], and [17].)

The discounted reward criterion is closely related to Blackwell optimality, which is the main subject in this paper. In fact, as shown in the following pages, Blackwell optimality is related to a whole sequence of optimality criteria.

Definition 2.3. (*Blackwell optimality.*) A policy f^* in $\mathbb{F}$ is called *Blackwell optimal* if, for each policy $f \in \mathbb{F}$ and each state $x \in \mathbb{R}^n$, there exists a discount factor $\alpha^* = \alpha^*(x, f) > 0$ such that

$$V_\alpha(x, f^*) \geq V_\alpha(x, f) \quad \text{for all } 0 < \alpha < \alpha^*.$$

The proof of the existence of Blackwell optimal policies is much more involved than for α-discount optimality, for *fixed* $\alpha > 0$, but the basic approach can be traced back to Blackwell's analysis [3] for a finite-state, finite-action Markov decision process. This approach hinges on the Laurent series (introduced in Section 3), which in turn is based on α-discount optimality (Definition 2.2) and average reward optimality (see Definition 2.4, below). Then the coefficients of the Laurent series are used to define a sequence of 'nested' control problems (see Section 4) that in the limit give Blackwell optimality (see Section 5).

In the reminder of this section we summarize the definition of and some facts about average optimality.

Definition 2.4. (*Average reward criterion.*) For each $f \in \mathbb{M}$, $x \in \mathbb{R}^n$, and $T \geq 0$, let

$$J_T(x, f) := \mathrm{E}_x^f \left(\int_0^T r(t, x(t), f)\, \mathrm{d}t \right).$$

The long-run *average reward*—also known as the *gain*—of f, given the initial state x, is

$$J(x, f) := \liminf_{T \to \infty} \frac{1}{T} J_T(x, f). \tag{2.8}$$

The function

$$J^*(x) := \sup_{f \in \mathbb{M}} J(x, f) \quad \text{for } x \in \mathbb{R}^n \tag{2.9}$$

is referred to as the *optimal gain* or *optimal average reward*. If there is a policy $f^* \in \mathbb{M}$ for which $J(x, f^*) = J^*(x)$ for all $x \in \mathbb{R}^n$, then f^* is called *average optimal*.

Assumptions 2.1, 2.2, 2.3, and 2.4 ensure the existence of average optimal stationary policies. Indeed, under these assumptions, there exists a pair (g, h) consisting of a constant $g \in \mathbb{R}$ and a function $h \in C^2(\mathbb{R}^n) \cap B_w(\mathbb{R}^n)$ satisfying the *average reward HJB* equation

$$g = \max_{u \in U} [r(x, u) + L^u h(x)] \quad \text{for all } x \in \mathbb{R}^n. \tag{2.10}$$

Moreover, there exists a policy $f \in \mathbb{F}$ that attains the maximum in (2.10), i.e.

$$g = r(x, f) + L^f h(x) \quad \text{for all } x \in \mathbb{R}^n. \tag{2.11}$$

https://doi.org/10.1239/jap/1245676094 Published online by Cambridge University Press

A policy $f \in \mathbb{F}$ that satisfies (2.11) is referred to as a *canonical* policy. Denote by $\mathbb{F}_{\text{ca}}$ and $\mathbb{F}_{\text{ao}}$ the sets of canonical and average optimal policies, respectively. It can be shown that, under our standing assumptions, the set of canonical policies coincides with the set of average optimal policies, so $\mathbb{F}_{\text{ca}} = \mathbb{F}_{\text{ao}}$. (See [2], [4], [10], [20], and [21].)

Using (2.5), it is easy to verify that, for every $f \in \mathbb{F}$, the gain (2.8) becomes

$$J(x, f) = \lim_{T\to\infty} \frac{1}{T} J_T(x, f) = \bar{r}(f) \quad \text{for all } x \in \mathbb{R}^n, \tag{2.12}$$

where $\bar{r}(f) := \int r(y, f)\mu_f(\mathrm{d}y)$. Furthermore, (2.3) and (2.4) yield $\mu_f(w) \leq d/c$, and, therefore, by the Assumption 2.4(b) we have

$$|\bar{r}(f)| \leq \int_{\mathbb{R}^n} |r(y, f)|\mu_f(\mathrm{d}y) \leq M\frac{d}{c} < \infty. \tag{2.13}$$

Finally, as in [2], [4], [10], [20], and [21], it can be seen that the constant

$$r^* := \sup_{f\in\mathbb{F}} \bar{r}(f) < \infty \tag{2.14}$$

coincides with the optimal gain in (2.9), i.e.

$$r^* = J^*(x) \quad \text{for all } x \in \mathbb{R}^n;$$

hence,

$$J^*(x) = \sup_{f\in\mathbb{M}} J(x, f) = \sup_{f\in\mathbb{F}} J(x, f) = r^* \quad \text{for all } x \in \mathbb{R}^n.$$

3. The Laurent series

The main objective of this section is to show that, for each $f \in \mathbb{F}$, the α-discounted reward (2.7) can be expressed as a Laurent series. This result for controlled Markov processes comes, of course, from the Laurent series expansion for the resolvent of Markov semigroups—see, for instance, [28] or [31, p. 212]. First, we will recall some facts from bias optimality.

For each $f \in \mathbb{F}$, we define the *bias* of f as the function

$$h_f(x) := \int_0^\infty (\mathrm{E}_x^f(r(x(t), f)) - \bar{r}(f))\,\mathrm{d}t \quad \text{for } x \in \mathbb{R}^n. \tag{3.1}$$

By (2.5) and Assumption 2.4(b), this function is finite valued; in fact, it belongs to $B_w(\mathbb{R}^n)$.

Definition 3.1. (*Bias optimality.*) The function $\hat{h} \in B_w(\mathbb{R}^n)$ defined as

$$\hat{h}(x) := \sup_{f\in\mathbb{F}_{\text{ao}}} h_f(x) \quad \text{for } x \in \mathbb{R}^n \tag{3.2}$$

is called the *optimal bias function*. Moreover, a stationary average optimal policy $\hat{f}$ is said to be *bias optimal* if it attains the maximum in (3.2), i.e.

$$h_{\hat{f}}(x) = \hat{h}(x).$$

We denote by $\mathbb{F}_{\text{bias}}$ the set of optimal bias policies.

https://doi.org/10.1239/jap/1245676094 Published online by Cambridge University Press

The following result shows that $\mathbb{F}_{\text{bias}}$ is nonempty, and it also gives some characterizations of bias optimal policies. For a proof, see Theorems 5.4, 5.5, and 6.2 of [18].

Proposition 3.1. *Under Assumptions 2.1, 2.2, 2.3, and 2.4, the following holds.*

(a) $\mathbb{F}_{\text{bias}}$ *is nonempty.*

(b) *The pair* $(r^*, \hat{h})$ *consisting of the constant* r^* *in (2.14) and the optimal bias function* $\hat{h}$ *in (3.2) form the unique solution satisfying the bias optimality equations*

$$g = \max_{u \in U}[r(x,u) + L^u h(x)], \tag{3.3}$$

$$h(x) = \max_{u \in U^0(x)} L^u \kappa(x), \tag{3.4}$$

for all $x \in \mathbb{R}^n$, *where* $U^0(x) := \{u \in U \mid g = r(x,u) + L^u h(x)\}$ *and* κ *is some function in* $C^2(\mathbb{R}^n) \cap B_w(\mathbb{R}^n)$.

(c) f *is bias optimal if and only if, for every* $x \in \mathbb{R}^n$, $f(x)$ *attains the maximum in (3.3) and (3.4).*

In this section we will consider the optimality criteria (2.7) and (2.8) with reward rates different from the function $r(x,u)$. These reward rates will be restricted to the class of functions defined as follows.

Definition 3.2. Let w be the function in Assumption 2.2. We denote by $B_w(\mathbb{R}^n \times U)$ the space of real-valued measurable functions v on $\mathbb{R}^n \times U$ such that

$$\sup_{u \in U} |v(x,u)| \le M_v w(x) \quad \text{for all } x \in \mathbb{R}^n, \tag{3.5}$$

where M_v is a positive constant depending of v.

Note that the space $B_w(\mathbb{R}^n)$ is contained in $B_w(\mathbb{R}^n \times U)$, because any function $v \in B_w(\mathbb{R}^n)$ can be written as $v(x) \equiv v(x,u)$ for $u \in U$.

As in (2.6), for $v \in B_w(\mathbb{R}^n \times U)$, $f \in \mathbb{M}$, and $x \in \mathbb{R}^n$, we define $v(t,x,f) := v(x, f(t,x))$. The following definitions generalize the optimality criteria (2.7) and (2.8).

Given $f \in \mathbb{M}$, $x \in \mathbb{R}^n$, $v \in B_w(\mathbb{R}^n \times U)$, and $\alpha > 0$, we define

$$V_\alpha(x,f,v) := \mathrm{E}_x^f\left(\int_0^\infty \mathrm{e}^{-\alpha t} v(t, x(t), f)\,\mathrm{d}t\right), \tag{3.6}$$

the expected α-discounted v-reward when using the policy $f \in \mathbb{M}$, given the initial state $x \in \mathbb{R}^n$. Similarly, the v-gain of f given the initial state x is defined as

$$J(x,f,v) := \liminf_{T \to \infty} \frac{1}{T}\,\mathrm{E}_x^f\left(\int_0^T v(t, x(t), f)\,\mathrm{d}t\right). \tag{3.7}$$

Note that if $v(x,u) = r(x,u)$ then (3.6) and (3.7) reduce to (2.7) and (2.8), respectively.

Given $f \in \mathbb{F} \subset \mathbb{M}$, let

$$\bar{v}(f) = \int_{\mathbb{R}^n} v(y,f)\mu_f(\mathrm{d}y). \tag{3.8}$$

https://doi.org/10.1239/jap/1245676094 Published online by Cambridge University Press

Under Assumptions 2.1, 2.2, and 2.3, we can see, as in (2.12), that the v-gain in (3.7) is the constant (3.8); that is,

$$J(x, f, v) = \lim_{T\to\infty} \frac{1}{T} \mathrm{E}_x^f \left(\int_0^T v(x(t), f)\, \mathrm{d}t \right) = \bar{v}(f).$$

In addition, as in (2.13), $\bar{v}(f)$ is uniformly bounded; in fact,

$$\sup_{f\in\mathbb{F}} |\bar{v}(f)| \le M_v \frac{d}{c}.$$

Now, for each $f \in \mathbb{F}$, consider the *bias operator* G_f on $B_w(\mathbb{R}^n \times U)$ defined, for every $v \in B_w(\mathbb{R}^n \times U)$, by

$$G_f v(x) := \int_0^\infty [\mathrm{E}_x^f (v(x(t), f)) - \bar{v}(f)]\, \mathrm{d}t. \tag{3.9}$$

Observe that (3.9) reduces to (3.1) when v coincides with the reward rate r in Assumption 2.4. Also, note that, from (2.5) and (3.5),

$$|G_f v(x)| \le \delta^{-1} C M_v w(x),$$

that is, $G_f v$ is in $B_w(\mathbb{R}^n)$, and its w-norm is uniformly bounded in $f \in \mathbb{F}$ because

$$\sup_{f\in\mathbb{F}} \|G_f v\|_w \le \delta^{-1} C M_v.$$

Hence, G_f maps $B_w(\mathbb{R}^n \times U)$ into $B_w(\mathbb{R}^n)$. Finally, observe that

$$\mu_f(G_f v) = 0 \quad \text{for all } f \in \mathbb{F} \text{ and } v \in B_w(\mathbb{R}^n \times U). \tag{3.10}$$

The next theorem ensures that the α-discounted v-reward (3.6) can be written as the Laurent series (3.11).

Theorem 3.1. *Let $\delta > 0$ be the constant in Assumption 2.3. Let $f \in \mathbb{F}$ and $v \in B_w(\mathbb{R}^n \times U)$ be arbitrary. Then, for $\alpha \in (0, \delta)$, the α-discounted v-reward (3.6) can be expressed as*

$$V_\alpha(x, f, v) = \frac{1}{\alpha}\bar{v}(f) + \sum_{k=0}^{\infty} (-\alpha)^k G_f^{k+1} v(x) \quad \textit{for all } x \in \mathbb{R}^n, \tag{3.11}$$

where G_f^{k+1} is the $(k+1)$-composition of G_f with itself. Moreover, the series (3.11) converges in the w-norm.

Proof. Fix an arbitrary $f \in \mathbb{F}$. Clearly, we can rewrite (3.6) as

$$V_\alpha(x, f, v) = \frac{1}{\alpha}\bar{v}(f) + \int_0^\infty \mathrm{e}^{-\alpha t} [\mathrm{E}_x^f (v(x(t), f)) - \bar{v}(f)]\, \mathrm{d}t. \tag{3.12}$$

To simplify the notation, define

$$Z_t^f v(x) := \mathrm{E}_x^f (v(x(t), f)) - \bar{v}(f).$$

https://doi.org/10.1239/jap/1245676094 Published online by Cambridge University Press

Observe that the integral in (3.12) is finite because, by (2.5),

$$|Z_t^f v(x)| \le C M_v \mathrm{e}^{-\delta t} w(x). \tag{3.13}$$

Furthermore, by the semigroup property of transition probabilities [6], [13], it is easy to verify that the family $\{Z_t^f\}_{t\ge 0}$ satisfies

$$Z_{t+s}^f = Z_t^f Z_s^f \quad \text{for all } s, t \ge 0 \text{ and } f \in \mathbb{F}. \tag{3.14}$$

Expanding the factor $\mathrm{e}^{-\alpha t}$ in (3.12) as a Taylor series and using the dominated convergence theorem (recall (3.13)), we obtain

$$\int_0^\infty \mathrm{e}^{-\alpha t} Z_t^f v(x)\,\mathrm{d}t = \sum_{k=0}^\infty (-\alpha)^k \int_0^\infty \frac{t^k}{k!} Z_t^f v(x)\,\mathrm{d}t.$$

Let

$$Y_k^f v(x) := \int_0^\infty \frac{t^k}{k!} Z_t^f v(x)\,\mathrm{d}t.$$

To obtain (3.11), we will use induction on k to prove that $Y_k^f v(x) = G_f^{k+1} v(x)$.

By (3.9), $Y_0^f = G_f$. Now suppose that $Y_{k-1}^f = G_f^k$ for some $k \ge 1$. Hence, we have

$$\begin{aligned} Y_k^f v(x) &= \int_0^\infty \left(\int_0^t \frac{s^{k-1}}{(k-1)!}\,\mathrm{d}s \right) Z_t^f v(x)\,\mathrm{d}t \\ &= \int_0^\infty \frac{s^{k-1}}{(k-1)!} \left(\int_s^\infty Z_t^f v(x)\,\mathrm{d}t \right) \mathrm{d}s \quad \text{(Fubini's theorem)} \\ &= \int_0^\infty \frac{s^{k-1}}{(k-1)!} \left(\int_0^\infty Z_s^f Z_t^f v(x)\,\mathrm{d}t \right) \mathrm{d}s, \end{aligned} \tag{3.15}$$

where the last equality is due to (3.14).

Observe that

$$\begin{aligned} \int_0^\infty (Z_s^f Z_t^f v)(x)\,\mathrm{d}t &= \int_0^\infty \int_{\mathbb{R}^n} Z_t^f v(y)[\mathrm{P}_x^f(s,\,\mathrm{d}y) - \mu_f(\mathrm{d}y)]\,\mathrm{d}t \\ &= \int_{\mathbb{R}^n} \int_0^\infty Z_t^f v(y)\,\mathrm{d}t[\mathrm{P}_x^f(s,\,\mathrm{d}y) - \mu_f(\mathrm{d}y)] \\ &= \int_{\mathbb{R}^n} G_f v(y)[\mathrm{P}_x^f(s,\,\mathrm{d}y) - \mu_f(\mathrm{d}y)] \quad \text{(by (3.9))} \\ &= \mathrm{E}_x^f((G_f v)(x(s))) - \mu_f(G_f v) \\ &= Z_s^f G_f v(x). \end{aligned}$$

Therefore, (3.15) becomes

$$Y_k^f v = \int_0^\infty \frac{s^{k-1}}{(k-1)!} (Z_s^f G_f v)\,\mathrm{d}s = Y_{k-1}^f G_f v = G_f^{k+1} v,$$

where the last equality follows from the induction hypothesis.

https://doi.org/10.1239/jap/1245676094 Published online by Cambridge University Press

Now we will prove that the series in (3.11) converges in the w-norm. To do this, note that (3.13) implies that the terms of the series in (3.11) are bounded, because

$$|(-\alpha)^k G_f^{k+1} v(x)| = |(-\alpha)^k Y_k^f v(x)| \leq \frac{CM_v}{\delta}\left(\frac{\alpha}{\delta}\right)^k w(x). \tag{3.16}$$

Hence, since α is in $(0, \delta)$, we obtain

$$\left\| \sum_{k=m+1}^{m+p} (-\alpha)^k G_f^{k+1} v(x) \right\|_w \leq \frac{CM_v}{\delta} \sum_{k=m+1}^{m+p} \left(\frac{\alpha}{\delta}\right)^k \to 0 \quad \text{as } m \to \infty.$$

This implies that the series in (3.11) is a Cauchy series in the Banach space $B_w(\mathbb{R}^n)$, and so it converges in $B_w(\mathbb{R}^n)$.

The following proposition establishes that the residual terms in the Laurent series are *bounded* in the w-norm.

Proposition 3.2. *Let $\theta \in \mathbb{R}$ be such that $0 < \theta < \delta$, where δ is the constant in Assumption 2.3. For each $v \in B_w(\mathbb{R}^n \times U)$, $f \in \mathbb{F}$, and $k = 0, 1, \ldots$, define the k-residual of the Laurent series (3.11) as*

$$R_k(f, v, \alpha) := \sum_{j=k}^{\infty} (-\alpha)^j G_f^{j+1} v.$$

Then, for all $|\alpha| \leq \theta$ and $k = 0, 1, \ldots$,

$$\sup_{f \in \mathbb{F}} \|R_k(f, v, \alpha)\|_w \leq \frac{M_v C}{\delta^k(\delta - \theta)} |\alpha|^k. \tag{3.17}$$

Proof. This is a straightforward consequence of inequality (3.16).

For each $v \in B_w(\mathbb{R}^n \times U)$, $f \in \mathbb{F}$, and $k = 0, 1, \ldots$, define $h_f^k v \in B_w(\mathbb{R}^n)$ as

$$h_f^k v(x) := (-1)^k G_f^{k+1} v(x) \quad \text{for all } x \in \mathbb{R}^n.$$

For $v = r$, with r as in Assumption 2.4, we simply write $h_f^k := h_f^k r$. On the other hand, by (3.1), $h_f^0 = h_f$ is in fact the bias of f, and $h_f^1 = G_f(-h_f^0)$ is the bias of f when the reward rate is $-h_f^0 = -h_f$. Also, note that the function $\bar{v}(f)$ is the gain $\bar{r}(f)$ in (2.12). In general, it can be seen by induction that

$$h_f^k = G_f(-h_f^{k-1}), \tag{3.18}$$

and so, h_f^k is the bias when the reward rate is $-h_f^{k-1}$. Therefore, the Laurent series (3.11), with r in lieu of $v \in B_w(\mathbb{R}^n \times U)$, becomes

$$V_\alpha(x, f) = \frac{1}{\alpha}\bar{r}(f) + \sum_{k=0}^{\infty} \alpha^k h_f^k(x) \tag{3.19}$$

for all $f \in \mathbb{F}$, $x \in \mathbb{R}^n$, and α as in Theorem 3.1. Finally, applying (3.10) to (3.18), we obtain

$$\mu_f(h_f^k) = 0 \quad \text{for all } k = 0, 1, 2, \ldots. \tag{3.20}$$

https://doi.org/10.1239/jap/1245676094 Published online by Cambridge University Press

4. The Poisson and the average reward HJB equations

In this section we associate the coefficients of the Laurent series (3.19) with the solutions of the average reward HJB equations, (4.4)–(4.6), below. To this end, given $f \in \mathbb{F}$, let L^f be the operator defined in (2.2). For each $m \geq 0$, consider the following system of equations: for $x \in \mathbb{R}^n$,

$$g = r(x, f) + L^f h^0(x), \tag{4.1}$$

$$h^0(x) = L^f h^1(x), \tag{4.2}$$

$$\vdots$$

$$h^m(x) = L^f h^{m+1}(x), \tag{4.3}$$

where g is a constant and $h^0, h^1, \ldots, h^{m+1}$ are functions in $C^2(\mathbb{R}^n)$. Equations (4.1)–(4.3) are referred to as the -1th, 0th, and mth *Poisson equations* for f, respectively.

The following theorem establishes a relation between the coefficients of the Laurent series (3.19) and the solution to the Poisson equations (4.1)–(4.3).

Theorem 4.1. *Fix $m \geq -1$. The constant $g \in \mathbb{R}$ and the functions $h^0, h^1, \ldots, h^{m+1} \in C^2(\mathbb{R}^n) \cap B_w(\mathbb{R}^n)$ are solutions to the Poisson equations (4.1)–(4.3) if and only if $g = \bar{r}(f)$, $h^k = h_f^k$ for $0 \leq k \leq m$, and $h^{m+1} = h_f^{m+1} + z$ for $z \in \mathbb{R}$.*

Proof. First we shall prove that $\bar{r}(f), h_f^0, \ldots, h_f^{m+1}$ satisfy the Poisson equations (4.1)–(4.3), and also that they are in $C^2(\mathbb{R}^n) \cap B_w(\mathbb{R}^n)$. The proof will be by induction on m. For $m = -1$, by [18, Proposition 4.1], h_f^0 is in $C^2(\mathbb{R}^n) \cap B_w(\mathbb{R}^n)$ and the pair $(\bar{r}(f), h_f^0)$ satisfies the -1th Poisson equation, (4.1).

Now assume that the stated result holds for some $m \geq -1$. We will show that h_f^{m+1} and h_f^{m+2} verify the $(m+1)$th Poisson equation, and that h_f^{m+2} is in $C^2(\mathbb{R}^n) \cap B_w(\mathbb{R}^n)$.

Consider the reward rate $-h_f^{m+1}$. By (3.20), the expected average reward is $\mu_f(-h_f^{m+1}) = 0$, and by (3.18), the bias of f when the reward rate is $-h_f^{m+1}$ becomes h_f^{m+2}. Thus, by [18, Proposition 4.1] we conclude that h_f^{m+2} is in $C^2(\mathbb{R}^n) \cap B_w(\mathbb{R}^n)$, and its corresponding Poisson equation is precisely the $(m+1)$th equation; that is,

$$h_f^{m+1}(x) = L^f h_f^{m+2}(x) \quad \text{for all } x \in \mathbb{R}^n.$$

Hence, h_f^{m+1} and $h_f^{m+2} + z$ for any $z \in \mathbb{R}$ satisfy the $(m+1)$th Poisson equation.

Conversely, suppose that $g \in \mathbb{R}$ and $h^1, h^2, \ldots, h^{m+1} \in C^2(\mathbb{R}^n)$ are solutions to (4.1)–(4.3). For $m = -1$, [18, Proposition 4.1] ensures that $(\bar{r}(f), h_f^0)$ is the unique solution to the -1th Poisson equation, (4.1); hence, $g = \bar{r}(f)$ and $h^0 = h_f^0$.

Now suppose that the solutions to the mth Poisson equation are h_f^m and $h_f^{m+1} + z$ for $z \in \mathbb{R}$. By the induction hypothesis, $h^{m+1} = h_f^{m+1} + z'$ for $z' \in \mathbb{R}$. Since μ_f is an invariant probability measure, we have

$$\int_{\mathbb{R}^n} L^f h(y) \mu_f(\mathrm{d}y) = 0 \quad \text{for all } h \in C^2(\mathbb{R}^n);$$

see, for instance, [2, Lemma 4.2.5]. Therefore, applying μ_f to the $(m+1)$th equation, we deduce that $\mu_f(h^{m+1}) = 0$. We also find, by (3.20), that $\mu_f(h_f^{m+1}) = 0$ and so $z' = 0$, that is, $h^{m+1} = h_f^{m+1}$. Finally, interpreting the $(m+1)$th Poisson equation as the -1th Poisson equation with reward rate $-h_f^{m+1}$, we obtain, as in the $m = -1$ case, $h^{m+2} = h_f^{m+2} + z$ for $z \in \mathbb{R}$.

https://doi.org/10.1239/jap/1245676094 Published online by Cambridge University Press

In addition to the Poisson equations (4.1)–(4.3), we now consider the following system of equations for $g \in \mathbb{R}$ and $h^0, h^1, \ldots, h^{m+1} \in C^2(\mathbb{R}^n)$, which will be referred to as the -1th, 0th, $\ldots$, mth average reward HJB equations, respectively: for $x \in \mathbb{R}^n$,

$$g = \max_{u \in U}[r(x,u) + L^u h^0(x)], \tag{4.4}$$

$$h^0(x) = \max_{u \in U^0(x)}[L^u h^1(x)], \tag{4.5}$$

$$\vdots$$

$$h^m(x) = \max_{u \in U^m(x)}[L^u h^{m+1}(x)], \tag{4.6}$$

where, letting $U^{-1}(x) := U$ for all $x \in \mathbb{R}^n$, the set $U^k(x)$ for $0 \le k \le m$ consists of the elements (controls) $u \in U^{k-1}(x)$ attaining the maximum in the $(k-1)$th average reward HJB equation. In particular, for $m = 0$, (4.4)–(4.5) coincide with the bias optimality equations (3.3)–(3.4).

Remark 4.1. (i) By Assumptions 2.1 and 2.4, the maps $u \mapsto r(x,u)$ and $u \mapsto L^u h^k(x)$ for $k = 0, \ldots, m+1$ are continuous on U for each $x \in \mathbb{R}^n$. Also, it is easily seen by induction that the sets $U^m(x)$, $m \ge 0$, form a nonincreasing sequence of nonempty compact sets for each $x \in \mathbb{R}^n$, and, furthermore, the set-valued mappings $x \mapsto U^m(x)$ are upper semicontinuous; see [18, Lemma 5.2].

(ii) Since $\{U^m(x)\}_{m \ge 0}$ is a nonincreasing sequence of nonempty compact sets, the set

$$U^\infty(x) := \bigcap_{m \ge -1} U^m(x)$$

is nonempty and compact.

The following definition concerns policies $f \in \mathbb{F}$ that attain the maximum in the average reward HJB equations (4.4)–(4.6).

Definition 4.1. Given an integer $m \ge -1$, let $\mathbb{F}_m$ be the set of all policies $f \in \mathbb{F}$ such that $f(x) \in U^{m+1}(x)$ for each $x \in \mathbb{R}^n$; that is, f is in $\mathbb{F}_m$ if it attains the maximum in the -1th, 0th, $\ldots$, mth average reward HJB equations.

By Remark 4.1 and Theorem 4.2, below, $\{\mathbb{F}_m\}_{m \ge -1}$ is a nonincreasing sequence ($\mathbb{F}_m \supseteq \mathbb{F}_{m+1}$) of nonempty sets, and it converges to the set

$$\mathbb{F}_\infty := \bigcap_{m=-1}^{\infty} \mathbb{F}_m. \tag{4.7}$$

By Remark 4.1, the set $\mathbb{F}_\infty$ is also nonempty.

Theorem 4.2 gives the existence and uniqueness of solutions to the average reward HJB equations (4.4)–(4.6). In addition, it guarantees the existence of policies that attain the maximum in these equations.

Theorem 4.2. *The average reward HJB equations (4.4)–(4.6) admit a unique solution $g \in \mathbb{R}$, $h^0, \ldots, h^{m+1} \in C^2(\mathbb{R}^n) \cap B_w(\mathbb{R}^n)$, where $g, h^0, \ldots, h^m$ are unique and h^{m+1} is unique up to an additive constant. Moreover, $\mathbb{F}_m$ is nonempty.*

Proof. First, we will prove the existence of solutions. For $m = 0$, the claim follows from Proposition 3.1(b). Suppose now that the result holds for some $m \ge 0$, and consider a new

https://doi.org/10.1239/jap/1245676094 Published online by Cambridge University Press

control model, which will be referred to as the 'm-bias problem', and which has the following components:

- the dynamical system (2.1);
- the action set $U^m(x)$ for each state $x \in \mathbb{R}^n$; and
- the reward rate $-h^m$.

It is evident that this model satisfies Assumptions 2.1, 2.2, 2.3, and 2.4. Hence, Proposition 3.1(b) gives the existence of the functions h^{m+1} and h^{m+2} in the class $C^2(\mathbb{R}^n) \cap B_w(\mathbb{R}^n)$ that satisfy the mth and the $(m+1)$th average reward HJB equations, and such that h^{m+1} is unique and h^{m+2} is unique up to additive constants. This proves the first statement of the theorem.

To prove that $\mathbb{F}_m$ is nonempty, we again proceed by induction on m. For $m = 0$, we use Proposition 3.1(a) to guarantee the existence of a bias optimal policy $f \in \mathbb{F}$. Then, from Proposition 3.1(c) we deduce that f maximizes the bias optimality equations, which coincide with the -1th and 0th average reward HJB equations. This implies that f is in $\mathbb{F}_0$. Now suppose that $f \in \mathbb{F}_m$ for some $m \geq -1$; that is, $f(x)$ is in $U_{m+1}(x)$ for all $x \in \mathbb{R}^n$. Consider again the m-bias problem defined above and observe that the set of average optimal policies associated to this problem coincides with the set $\mathbb{F}_m$. Then, since the m-bias problem satisfies the hypotheses of Proposition 3.1, we use this proposition to ensure the existence of a bias optimal policy $f \in \mathbb{F}_m$ associated to the m-bias problem. Hence, using the characterization of f in Proposition 3.1(c), we deduce that $f \in \mathbb{F}_m$ maximizes the mth and $(m+1)$th optimality equations; that is, f is in $\mathbb{F}_{m+1}$. This completes the proof.

Note that if f is in $\mathbb{F}_m$ then, by Theorem 4.1, the solution to the average reward HJB equations consists of $g = \bar{r}(f)$, $h^0 = h^0_f, \ldots, h^m = h^m_f$, and $h^{m+1} = h^{m+1}_f + z$ for some constant z.

As a consequence of the Remark 4.1(ii) and Theorem 4.2, we obtain the following result.

Corollary 4.1. *There exists a policy $f \in \mathbb{F}$ that maximizes the mth average optimality equations* for every $m = -1, 0, \ldots$. *In other words, f is in $\mathbb{F}_\infty$.*

5. Blackwell optimality

We are finally arriving at the main problem we are concerned with, namely, the characterization and existence of Blackwell optimal policies. Indeed, we will ensure the existence of m-discount optimal policies ($m \geq -1$), based on a lexicographic maximization of the mth coefficient of the Laurent series (3.19). This fact is then used to prove the existence of Blackwell optimal policies. The concept of m-discount optimality is defined as follows.

Definition 5.1. (*m-discount optimality.*) Let $m \geq -1$ be an integer. A stationary policy $f^* \in \mathbb{F}$ is called *m-discount optimal* if

$$\liminf_{\alpha \downarrow 0} \alpha^{-m}[V_\alpha(x, f^*) - V_\alpha(x, f)] \geq 0 \quad \text{for all } f \in \mathbb{F} \text{ and } x \in \mathbb{R}^n.$$

Clearly, if $f^* \in \mathbb{F}$ is m-discount optimal (for $m \geq 0$) then f^* is $(m-1)$-discount optimal. Hence, if $\mathbb{F}^{\mathrm{d}}_m \subset \mathbb{F}$ denotes the set of m-discount optimal stationary policies then the sequence $\{\mathbb{F}^{\mathrm{d}}_m\}_{m \geq -1}$ is nonincreasing.

To prove the existence of Blackwell optimal policies, we first recall the definition of a *lexicographic order*.

https://doi.org/10.1239/jap/1245676094 Published online by Cambridge University Press

Definition 5.2. Given a pair of vectors x and y in $\mathbb{R}^d$, we say that x is lexicographically greater than or equal to y, denoted by $x \succeq y$, if the first nonzero component of $x - y$ is positive. Furthermore, we write $x \succ y$ if $x \succeq y$ and $x \neq y$.

The following proposition relates the sets $\mathbb{F}_m$ in Definition 4.1 with the coefficients of the Laurent series (3.19).

Proposition 5.1. *For every integer $m \geq 0$, a stationary policy $f \in \mathbb{F}$ belongs to $\mathbb{F}_m$ if and only if it lexicographically maximizes the terms $\bar{r}(f), h_f^0, \ldots, h_f^m$ of the Laurent series (3.19) in the class $\mathbb{F}$.*

Proof. We will use induction on m.

Consider the $m = 0$ case. Suppose that f belongs to $\mathbb{F}_0$, that is, f maximizes the -1th and the 0th average HJB equations (4.4)–(4.5); or, equivalently, f maximizes the bias optimality equations (3.3)–(3.4). This implies, from Proposition 3.1(c), that f is bias optimal. Now let $f' \in \mathbb{F}$. If f' is not average optimal, i.e. $f' \notin \mathbb{F}_{\text{ao}}$, then $\bar{r}(f) > \bar{r}(f')$. Hence,

$$(\bar{r}(f), h_f) \succeq (\bar{r}(f'), h_{f'}). \tag{5.1}$$

Otherwise, if $f' \in \mathbb{F}_{\text{ao}}$ then $\bar{r}(f) = \bar{r}(f')$ and $h_f \geq h_{f'}$, because f is bias optimal, and, hence, (5.1) holds.

To prove the converse, suppose that f lexicographically maximizes the terms $\bar{r}(f)$ and h_f. Then, $\bar{r}(f) > \bar{r}(f')$ or $\bar{r}(f) = \bar{r}(f')$ and $h_f \geq h_{f'}$ for all $f' \in \mathbb{F}$. This implies that f is bias optimal. Hence, by Proposition 3.1(c), f satisfies the bias optimality equations (3.3)–(3.4); in other words, it satisfies the -1th and the 0th average HJB equations (4.4)–(4.5). This proves that f is in $\mathbb{F}_0$.

Now suppose that the result holds for some $m \geq 0$. If f is in $\mathbb{F}_{m+1}$ then the solutions to the -1th, 0th, $\ldots$, $(m+1)$th average reward HJB equations become $g = \bar{r}(f), h^0 = h_f^0, \ldots, h^m = h_f^m$, and $h^{m+1} = h_f^{m+1}$; therefore, f is the optimal bias for the m-bias problem. Let $f' \in \mathbb{F}$. If $f' \notin \mathbb{F}_m$ then, by the induction hypothesis,

$$(\bar{r}(f), h_f^0, \ldots, h_f^{m+1}) \succeq (\bar{r}(f'), h_{f'}^0, \ldots, h_{f'}^{m+1}). \tag{5.2}$$

Otherwise, if $f' \in \mathbb{F}_m$,

$$(\bar{r}(f), h_f^0, \ldots, h_f^m) = (\bar{r}(f'), h_{f'}^0, \ldots, h_{f'}^m)$$

and f' is average optimal for the m-bias problem. Also, $h_f \geq h_{f'}$ for the m-bias problem, and, by (3.18),

$$h_f^{m+1} = G_f(-h_f^m) \geq G_{f'}(-h_f^m) = h_{f'}^{m+1}, \tag{5.3}$$

so (5.2) follows.

Conversely, suppose that f lexicographically maximizes $\bar{r}(f), h_f^0, \ldots, h_f^{m+1}$. By the induction hypothesis, f is in $\mathbb{F}_m$. Now let $f' \in \mathbb{F}_m$. Then $h^m = h_f^m = h_{f'}^m$, and (5.3) holds. Therefore, f is bias optimal for the m-bias problem and it verifies its corresponding bias optimality equations; that is, the mth and $(m+1)$th optimality equations. Hence, f is in $\mathbb{F}_{m+1}$.

The following theorem relates policies $f \in \mathbb{F}_m$ with the concept of m-discount optimality in Definition 5.1. In addition, it establishes the equivalence between Blackwell optimal policies with policies in $\mathbb{F}_\infty$.

https://doi.org/10.1239/jap/1245676094 Published online by Cambridge University Press

Theorem 5.1. (i) *Given an integer $m \geq -1$, a policy $f \in \mathbb{F}$ is m-discount optimal if and only if f is in $\mathbb{F}_m$; that is, $\mathbb{F}_m^{\mathrm{d}} = \mathbb{F}_m$.*

(ii) *$f \in \mathbb{F}$ is Blackwell optimal if and only if f is in $\mathbb{F}_\infty$.*

Proof. (i) We will use induction on m. To begin, consider $m = -1$. From (3.17) and (3.19), we have

$$\lim_{\alpha \downarrow 0} \alpha[V_\alpha(x, f^*) - V_\alpha(x, f)] = \bar{r}(f^*) - \bar{r}(f) \quad \text{for all } f^*, f \in \mathbb{F}. \tag{5.4}$$

Now suppose that f^* is in $\mathbb{F}_{-1}$, or, equivalently, that f^* is in $\mathbb{F}_{\mathrm{ca}} (= \mathbb{F}_{\mathrm{ao}})$ (recall the comments after (2.10)). Hence, f^* is -1-discount optimal, because $\bar{r}(f^*) - \bar{r}(f) \geq 0$ for every $f \in \mathbb{F}$. Conversely, suppose that f^* is -1-discount optimal; that is,

$$\liminf_{\alpha \downarrow 0} \alpha[V_\alpha(x, f^*) - V_\alpha(x, f)] \geq 0 \quad \text{for all } f \in \mathbb{F}.$$

From this expression and (5.4), we deduce that $\bar{r}(f^*) \geq \bar{r}(f)$ for every $f \in \mathbb{F}$; hence, f^* is average optimal.

Now suppose that (i) holds for some $m \geq 0$, and let $f \in \mathbb{F}_{m+1}$. We want to prove that $f \in \mathbb{F}_{m+1}^{\mathrm{d}}$. To this end, note that, by (3.19),

$$\begin{aligned}
&\frac{1}{\alpha^{m+1}}[V_\alpha(x, f^*) - V_\alpha(x, f)] \\
&\quad = \frac{1}{\alpha}\left[\frac{1}{\alpha^{m+1}}(\bar{r}(f^*) - \bar{r}(f)) + \frac{1}{\alpha^m}(h_{f^*}^0(x) - h_f^0(x)) + \cdots + (h_{f^*}^m(x) - h_f^m(x))\right] \\
&\qquad + (h_{f^*}^{m+1}(x) - h_f^{m+1}(x)) + \frac{1}{\alpha^{m+1}} \sum_{k=m+2}^{\infty} \alpha^k (h_{f^*}^k(x) - h_f^k(x)).
\end{aligned} \tag{5.5}$$

Taking $\alpha \downarrow 0$ we deduce that, by the induction hypothesis, the first term in (5.5) is nonnegative, and, by (3.17), the third term in (5.5) goes to 0. Moreover, since f^* is in $\mathbb{F}_{m+1}$, it is optimal for the $(m+1)$-bias problem, which consists of

- the dynamical system (2.1);
- the action set $U^{m+1}(x)$ for each state $x \in \mathbb{R}^n$; and
- the reward rate $-h_f^{m+1}$.

Hence, $h_{f^*}^{m+1} \geq h_f^{m+1}$. Therefore, f^* is in $\mathbb{F}_{m+1}^{\mathrm{d}}$, i.e. $\mathbb{F}_{m+1} \subset \mathbb{F}_{m+1}^{\mathrm{d}}$.

Conversely, assume that f^* is in $\mathbb{F}_{m+1}^{\mathrm{d}}$, but that it is *not* in $\mathbb{F}_{m+1}$. Since $\mathbb{F}_{m+1}^{\mathrm{d}} \subset \mathbb{F}_m^{\mathrm{d}}$, it follows that, by the induction hypothesis, f^* is in $\mathbb{F}_m$. Take an arbitrary $f \in \mathbb{F}_{m+1}$ (which is possible by Theorem 4.2). By Proposition 5.1, $h_{f^*}^{m+1} < h_f^{m+1}$. Then

$$\begin{aligned}
&\frac{1}{\alpha^{m+1}}[V_\alpha(x, f^*) - V_\alpha(x, f)] \\
&\quad = (h_{f^*}^{m+1}(x) - h_f^{m+1}(x)) + \frac{1}{m+1} \sum_{m+2}^{\infty} \alpha^k (h_{f^*}^k(x) - h_f^k(x)).
\end{aligned} \tag{5.6}$$

https://doi.org/10.1239/jap/1245676094 Published online by Cambridge University Press

Letting $\alpha \downarrow 0$, we see that the last term in (5.6) converges to 0. Thus,

$$\liminf_{\alpha \downarrow 0} \frac{1}{\alpha^{m+1}} [V_\alpha(x, f^*) - V_\alpha(x, f)] < 0,$$

which contradicts the fact that f^* is in $\mathbb{F}^{\mathrm{d}}_{m+1}$. This completes the proof of part (i).

(ii) First, suppose that f^* is in $\mathbb{F}_\infty$. Choose an arbitrary $f \in \mathbb{F}$ and $x \in \mathbb{R}^n$. Then, by (3.19),

$$V_\alpha(x, f^*) - V_\alpha(x, f) = \alpha^{-1}[\bar{r}(f^*) - \bar{r}(f)] + \sum_{k=0}^{\infty} \alpha^k [h^k_{f^*}(x) - h^k_f(x)]. \tag{5.7}$$

By (4.7) and Proposition 5.1, the right-hand side of (5.7) is nonnegative for every $\alpha > 0$, and yields Blackwell optimality (see Definition 2.3).

Conversely, suppose that f^* is Blackwell optimal. Pick an arbitrary $f \in \mathbb{F}$ and $x \in \mathbb{R}^n$, and let $\alpha^* = \alpha^*(x, f) > 0$ be as in Definition 2.3. It follows immediately that f^* is in $\mathbb{F}^d_m$ for $m = 0$ and $m = -1$ (see (5.4)); equivalently, by part (i), f^* is in $\mathbb{F}_m$ for $m = 0$ and -1. In fact, it is evident that f^* is in $\mathbb{F}^d_m = \mathbb{F}_m$ for all $m \geq -1$, and, therefore, f^* is in $\mathbb{F}_\infty = \mathbb{F}^d_\infty$, where $\mathbb{F}^d_\infty := \bigcap_{m=-1}^{\infty} \mathbb{F}^d_m$.

Corollary 5.1. *Under Assumptions 2.1, 2.2, 2.3, and 2.4,*

(i) *for each $m \geq -1$, the set $\mathbb{F}_m \subset \mathbb{F}$ of m-discount optimal policies is nonempty;*

(ii) *there exists a Blackwell optimal policy in $\mathbb{F}$.*

Proof. This follows by combining Theorem 4.2 and Corollary 4.1 with Theorem 5.1.

6. An example

We now give an example to illustrate our results. Our example is motivated by the manufacturing system studied in [1], which is also used as an application in [9] and [10].

Consider the one-dimensional linear system

$$\mathrm{d}x(t) = [\gamma x(t) + \beta u(t)]\,\mathrm{d}t + \sigma\,\mathrm{d}B_t, \qquad x(0) = x_0,\ t \geq 0, \tag{6.1}$$

where γ, β, and σ are given constants, with $\beta > 0$. The control $u(t)$ takes values in the compact set $U := [0, a]$, $a > 0$.

Now let $r : \mathbb{R} \mapsto \mathbb{R}$ be the reward rate, which is supposed to be concave and locally Lipschitz. This choice of r, which depends on the state variable but not on the control, is motivated by some applications to inventory systems (see, for instance, [1]).

We will suppose the existence of a function $w \geq 1$ that satisfies our Assumptions 2.2 and 2.4(b). (This function depends, of course, on r and the coefficients of (6.1). For instance, assuming that $\gamma < 0$ in (6.1), if $r(x) := -x^2$ then a quadratic function such as $w(x) := x^2 + 1$ usually works for our present purposes.)

Our aim is to find m-discount and Blackwell optimal policies in the class of stationary policies $u(t) = f(x(t))$. To this end, for each $m \geq -1$, we will prove the existence of policies $f_m \in \mathbb{F}$ that maximize the -1th, 0th, ..., mth average reward HJB equations (4.4)–(4.6) associated to certain optimal control problems. Hence, according to Theorem 5.1, such policies are m-discount optimal. This will be crucial to find Blackwell optimal policies.

https://doi.org/10.1239/jap/1245676094 Published online by Cambridge University Press

Consider the -1th average reward HJB equation associated to the problem of maximizing the expected average reward

$$\lim_{T\to\infty}\frac{1}{T}\,\mathrm{E}_x^f\left(\int_0^T r(x(t))\,\mathrm{d}t\right)$$

subject to (6.1). By (6.1) and (4.4),

$$-g + r(x) + \gamma x h_0'(x) + \frac{\sigma^2}{2}h_0''(x) + \max_{u\in[0,a]}\{\beta u h_0'(x)\} = 0, \tag{6.2}$$

where $h_0'(x)$ and $h_0''(x)$ denote the first and the second derivatives of h_0, with h_0 in $C^2(\mathbb{R}^n)\cap B_w(\mathbb{R}^n)$, and g is a constant. Since $r(\cdot)$ is concave, the same arguments as in [1, p. 117], show that $h_0(\cdot)$ is also concave. Hence, we have two trivial cases: if h_0 is strictly increasing ($h_0' > 0$) or decreasing ($h_0' < 0$), then the control policy $f_a(x) \equiv a$ or, respectively $f_0(x) \equiv 0$ is the unique policy that attains the maximum in (6.2); in other words, $\mathbb{F}_{-1}$ is the singleton $\{f_a\}$ or $\{f_0\}$. This set coincides with the set of average optimal policies, according to Theorem 3.3 of [18] (see also [2], [4], and [10]).

Now suppose that h_0 attains a maximum, say x_0. Hence,

$$h_0'(x)\begin{cases} > 0 & \text{if } x < x_0,\\ = 0 & \text{if } x = x_0,\\ < 0 & \text{if } x > x_0.\end{cases}$$

Thus, any control of the form

$$f_b(x) = \begin{cases} a & \text{if } x < x_0,\\ b & \text{if } x = x_0 \text{ and } 0 \le b \le a,\\ 0 & \text{if } x > x_0,\end{cases} \tag{6.3}$$

maximizes (6.2), and these controls constitute the set $\mathbb{F}_{-1}$ (the set of average optimal policies).

We are now interested in finding bias optimal policies. To this end, let us consider a new control problem consisting of the following components.

- The dynamic system (6.1).
- The control set $U^0(x) := \{u \in [0, a] \mid u \text{ attains the maximum in (6.2)}\}$.
- The reward rate $-h_0$.

It is easy to verify that this new problem satisfies all of our assumptions. Then, by Proposition 3.1, there exists an h_1 in $C^2(\mathbb{R}^n)\cap B_w(\mathbb{R}^n)$ that satisfies the 0th average reward HJB equation

$$-h_0(x) + \gamma x h_1'(x) + \frac{\sigma^2}{2}h_1''(x) + \max_{u\in U^0(x)}\{\beta u h_1'(x)\} = 0. \tag{6.4}$$

Note that, by Proposition 3.1(c), f maximizes both equations (6.2) and (6.4) if and only if it is bias optimal.

https://doi.org/10.1239/jap/1245676094 Published online by Cambridge University Press

If h_0 is strictly increasing or decreasing, by definition of bias optimality and the uniqueness of the average optimal policies, f_a or f_0 is bias optimal, respectively, depending on the sign of h_0'. Hence, these policies maximize the -1th and the 0th average reward HJB equations. Otherwise, suppose that h_0 attains its maximum in x_0. As we already noted above, h_0 is concave, and so $-h_0$ is convex. Therefore, again using the arguments of [1, p. 117], h_1 is convex. This means that we have again two trivial cases: if h_1 is strictly increasing or decreasing, we see that f_a or, respectively, f_0 maximizes both equations (6.2) and (6.4); that is, $\mathbb{F}_0 = \{f_a\}$ or, respectively, $\mathbb{F}_0 = \{f_0\}$. Otherwise, if h_1 attains a minimum, say x_1, then

$$h_1'(x) \begin{cases} < 0 & \text{if } x < x_1, \\ = 0 & \text{if } x = x_1, \\ > 0 & \text{if } x > x_1. \end{cases}$$

This gives the following.

(i) If $x_0 > x_1$ then f_a maximizes both (6.2) and (6.4). Hence, $\mathbb{F}_0 = \{f_a\}$.

(ii) If $x_0 = x_1$ then f_b in (6.3) maximizes (6.2) and (6.4), which implies that $\mathbb{F}_0 = \{f_b \mid 0 \le b \le a\}$.

(iii) If $x_0 < x_1$ then f_0 maximizes (6.2) and (6.4), and so $\mathbb{F}_0 = \{f_0\}$.

In general, suppose that $m \ge -1$, and, for each $i = 0, \dots, m$, let x_i be the point where h_{i+1} attains either the maximum or minimum. Then, by induction, it can be seen that, if m is even, the set $\mathbb{F}_m$ is given as

$$\mathbb{F}_m = \begin{cases} \{f_a\} & \text{if } x_0 = x_1 = \cdots = x_{m-1} > x_m, \\ \{f_0\} & \text{if } x_0 = x_1 = \cdots = x_{m-1} < x_m, \\ \{f_b \mid 0 \le b \le a\} & \text{if } x_0 = x_1 = \cdots = x_{m-1} = x_m, \end{cases}$$

or, if m is odd, the set $\mathbb{F}_m$ is given as

$$\mathbb{F}_m = \begin{cases} \{f_0\} & \text{if } x_0 = x_1 = \cdots = x_{m-1} > x_m, \\ \{f_a\} & \text{if } x_0 = x_1 = \cdots = x_{m-1} < x_m, \\ \{f_b \mid 0 \le b \le a\} & \text{if } x_0 = x_1 = \cdots = x_{m-1} = x_m. \end{cases}$$

Hence, by Theorem 5.1, $\mathbb{F}_m$ coincides with the set of m-discount optimal policies. Finally, the set $\mathbb{F}_\infty$, which is composed of the Blackwell optimal policies, is

$$\mathbb{F}_\infty = \begin{cases} \{f_a, f_0\} & \text{if } x_i \ne x_{i+1} \text{ for some } i = -1, 0, \dots, \\ \{f_b \mid 0 \le b \le a\} & \text{if } x_i = x_{i+1} \text{ for all } i = -1, 0, \dots. \end{cases}$$

7. Concluding remarks

In this paper we analyzed m-discount optimality for every integer $m \ge -1$, which essentially gives Blackwell optimality in the 'limit' as $m \to \infty$. A key step to obtain these results was to express the expected discounted reward (2.7) as the Laurent series (3.19). Similar results have been obtained previously for discrete-time and continuous-time controlled Markov chains (see [3], [5], [16], [25], and [29]). To the best of the authors' knowledge, however, the only

https://doi.org/10.1239/jap/1245676094 Published online by Cambridge University Press

related work dealing with controlled diffusion processes is Puterman's paper [27], where he considered a one-dimensional diffusion process with values in a compact interval. We should also mention the work by Taylor [28], who obtained a Laurent series for a general, uncontrolled, continuous-time Markov process with values in a compact metric space.

References

[1] Akella, R. and Kumar, P. R. (1986). Optimal control of production rate in a failure prone manufacturing system. *IEEE Trans. Automatic Control* **31,** 116–126.

[2] Arapostathis, A., Ghosh, M. K. and Borkar, V. S. (2009). *Ergodic Control of Diffusion Processes*. To appear.

[3] Blackwell, D. (1962). Discrete dynamic programming. *Ann. Math. Statist.* **33,** 719–726.

[4] Borkar, V. S. and Ghosh, M. K. (1990). Ergodic control of multidimensional diffusions. II. Adaptive control. *Appl. Math. Optimization* **21,** 191–220.

[5] Dekker, R. and Hordijk, A. (1992). Recurrence conditions for average and Blackwell optimality in denumerable state Markov decision chains. *Math. Operat. Res.* **17,** 271–289.

[6] Dynkin, E. B. (1965). *Markov Processes*, Vol. 1. Springer, Berlin.

[7] Fort, G. and Roberts, G. O. (2005). Subgeometric ergodicity of strong Markov processes. *Ann. Appl. Prob.* **15,** 1565–1589.

[8] Ghosh, M. K. and Marcus, S. I. (1991). Infinite horizon controlled diffusion problems with nonstandard criteria. *J. Math. Systems Estim. Control* **1,** 45–69.

[9] Ghosh, M. K., Arapostathis, A. and Marcus, S. I. (1993). Optimal control of switching diffusions with application to flexible manufacturing systems. *SIAM J. Control Optimization* **31,** 1183–1204.

[10] Ghosh, M. K., Arapostathis, A. and Marcus, S. I. (1997). Ergodic control of switching diffusions. *SIAM J. Control Optimization* **35,** 1952–1988.

[11] Glynn, P. W. and Meyn, S. P. (1996). A Liapounov bound for solutions of the Poisson equation. *Ann. Prob.* **24,** 916–931.

[12] Has'minskiĭ, R. Z. (1980). *Stochastic Stability of Differential Equations*. Sijthoff and Noordhoff, Germantown, Md.

[13] Hernández-Lerma, O. (1994). *Lectures on Continuous-Time Markov Control Processes*. Sociedad Matemática Mexicana, Mexico.

[14] Hernández-Lerma, O. and Lasserre, J. B. (1999). *Further Topics on Discrete-Time Markov Control Processes* (Appl. Math. **42**). Springer, New York.

[15] Hilgert, N. and Hernández-Lerma, O. (2003). Bias optimality versus strong 0-discount optimality in Markov control processes with unbounded costs. *Acta Appl. Math.* **77,** 215–235.

[16] Hordijk, A. and Yushkevich, A. A. (2002). Blackwell optimality. In *Handbook of Markov Decision Processes* (Internat. Ser. Operat. Res. Manag. Sci. **40**), eds E. A. Feinberg and A. Shwartz, Kluwer, Boston, MA, pp. 231–267.

[17] Jasso-Fuentes, H. (2007). Infinite-horizon optimal control problems for Markov diffusion processes. Doctoral Thesis, Mathematics Department, CINVESTAV-IPN.

[18] Jasso-Fuentes, H. and Hernández-Lerma, O. (2008). Characterizations of overtaking optimality for controlled diffusion processes. *Appl. Math. Optimization* **57,** 349–369.

[19] Jasso-Fuentes, H. and Hernández-Lerma, O. (2009). Ergodic control, bias, and sensitive discount optimality for Markov diffusion processes. *Stoch. Anal. Appl.* **27,** 363–385.

[20] Leizarowitz, A. (1988). Controlled diffusion processes on infinite horizon with the overtaking criterion. *Appl. Math. Optimization* **17,** 61–78.

[21] Leizarowitz, A. (1990). Optimal controls for diffusion in $\mathbb{R}^d$—min-max max-min formula for the minimal cost growth rate. *J. Math. Anal. Appl.* **149,** 180–209.

[22] Meyn, S. P. and Tweedie, R. L. (1993). Stability of Markovian processes. III. Foster–Lyapunov criteria for continuous-time precesses. *Adv. Appl. Prob.* **25,** 518–548.

[23] Prieto-Rumeau, T. (2006). Blackwell optimality in the class of Markov policies for continuous-time controlled Markov chains. *Acta Appl. Math.* **92,** 77–96.

[24] Prieto-Rumeau, T. and Hernandez-Lerma, O. (2005). Bias and overtaking equilibria for zero-sum continuous-time Markov games. *Math. Meth. Operat. Res.* **61,** 437–454.

[25] Prieto-Rumeau, T. and Hernandez-Lerma, O. (2005). The Laurent series, sensitive discount and Blackwell optimality for continuous-time controlled Markov chains. *Math. Meth. Operat. Res.* **61,** 123–145.

[26] Prieto-Rumeau, T. and Hernández-Lerma, O. (2006). Bias optimality for continuous-time controlled Markov chains. *SIAM J. Control Optimization* **45,** 51–73.

[27] Puterman, M. L. (1974). Sensitive discount optimality in controlled one-dimensional diffusions. *Ann. Prob.* **2,** 408–419.

[28] Taylor, H. M. (1976). A Laurent series for the resolvent of a strongly continuous stochastic semi-group. *Math. Program.* **6,** 258–263.
[29] Veinott, A. F. Jr. (1969). Discrete dynamic programming with sensitive discount optimality criteria. *Ann. Math. Statist.* **40,** 1635–1660.
[30] Veretennikov, A. Y. and Klokov, S. A. (2005). On the subexponential rate of mixing for Markov processes. *Theory Prob. Appl.* **49,** 110–122.
[31] Yosida, K. (1995). *Functional Analysis* (Reprint). Springer, Berlin.
[32] Zhu, Q. and Guo, X. (2005). Another set of conditions for strong n ($n = -1, 0$) discount optimality in Markov decision processes. *Stoch. Anal. Appl.* **23,** 953–974.