J. Appl. Prob. **43,** 603–621 (2006)
Printed in Israel

AN ANALYSIS OF TRANSIENT MARKOV DECISION PROCESSES

HUW W. JAMES * AND
E. J. COLLINS,** *University of Bristol*

Abstract

This paper is concerned with the analysis of Markov decision processes in which a natural form of termination ensures that the expected future costs are bounded, at least under some policies. Whereas most previous analyses have restricted attention to the case where the set of states is finite, this paper analyses the case where the set of states is not necessarily finite or even countable. It is shown that all the existence, uniqueness, and convergence results of the finite-state case hold when the set of states is a general Borel space, provided we make the additional assumption that the optimal value function is bounded below. We give a sufficient condition for the optimal value function to be bounded below which holds, in particular, if the set of states is countable.

Keywords: Pursuit problem; first passage problem; stochastic shortest path problem; value iteration; policy iteration

2000 Mathematics Subject Classification: Primary 90C40
Secondary 93E20

1. Introduction

This paper is concerned with the analysis of Markov decision processes in which a natural form of termination ensures that the expected future costs are bounded, at least under some policies. While it is possible to analyse such processes using the existing theory for Markov decision processes with the total cost criterion, much of this theory applies only when the cost function is either nonnegative or nonpositive. This paper analyses the case where neither of these assumptions necessarily holds. Instead, a condition is formulated which ensures that termination can occur, using the notion of a *transient policy*.

Finite-state, finite-action transient Markov decision processes with positive cost functions were first formulated and studied by Eaton and Zadeh [6] as *pursuit problems*. Veinott [19] derived similar results under the assumption that all stationary, deterministic policies are transient. Derman [4, pp. 53–63] extended these results under the title of *first passage problems*. Referring to this type of problem as a *stochastic shortest path problem*, Bertsekas and Tsitsiklis [1, pp. 317–323] generalized the results of both [6] and [19] by relaxing the assumption that all policies are transient, instead introducing the assumption that every stationary, deterministic policy which is not transient has an associated value function that is unbounded above. Bertsekas and Tsitsiklis [2] then strengthened their original results by weakening the assumption that the set of actions available in each state is finite, instead assuming that the set of actions available in each state is compact, the transition kernel is continuous over the set of actions available in

Received 23 September 2005; revision received 27 February 2006.
* Current address: Commerzbank Corporates and Markets, 60 Gracechurch Street, London EC3V 0HR, UK. Email address: huw.james@commerzbank.com
** Postal address: School of Mathematics, University of Bristol, University Walk, Bristol BS8 1TW, UK.

each state, and the cost function is both lower semicontinuous over the set of actions available in each state and bounded. More recently, Hinderer and Waldmann [13], [14] extended the results of [19] in terms of a *critical discount factor*, defined as the smallest number such that, for all discount factors δ smaller than this number, the limit of the k-stage, δ-discounted optimal value function exists and is finite.

Pliska [16] was the first to generalize the problem to Borel state and action spaces. In addition to the standard assumptions of compact action space, continuous transition kernel, and lower-semicontinuous cost function, his key assumptions were that the cost function was bounded and that all policies were transient. Hernández-Lerma *et al.* [10] extended the results of [16] in one direction, retaining the assumption that all policies were transient but relaxing the assumption that the cost function was bounded and instead assuming that it was dominated by some given function.

This paper extends the analysis in a different direction, retaining the assumption that the cost function is bounded but allowing for the existence of policies which are not necessarily transient. Thus, there is no direct overlap with [10], even though both treat similar problems of existence, uniqueness, and convergence. Indeed, it is not difficult to construct examples of processes which are covered by the results presented here but are not covered by the results of [10] (see [2]), since any process for which there exist policies under which one can get trapped in a given subset of the set of states, with zero probability of escaping, constitutes such an example.

Thus, this paper generalizes the results of [16] to the case where there may exist some policies which are not transient and those of [2] to the case where the set of states is not necessarily finite. Owing to the greater generality of the model considered here, an assumption additional to those of [2] is required, namely that the optimal value function is bounded below. In particular, we show that, under this additional assumption, (i) there exists an optimal stationary, deterministic policy; (ii) the optimal value function is the unique solution to the optimality equation; (iii) the value iteration algorithm converges to the optimal value function starting from any bounded function; and (iv) the policy iteration algorithm converges to the optimal value function starting from any transient, stationary, deterministic policy.

To help identify cases satisfying our assumptions, we show that if the sequence of stationary, deterministic policies generated by the policy iteration algorithm has a pointwise convergent subsequence – as is the case when, in particular, the set of states is countable – then the optimal value function is bounded below. We also strengthen the results of Pliska [16], by showing that, under his original assumptions, the dynamic programming operator is a k-stage contraction (for some k) with respect to the usual supremum norm and a 1-stage contraction with respect to a weighted supremum norm.

Finally, we note that our results do not apply to *optimal stopping processes* (see [5] and [7]), which are a special type of transient Markov decision process where a state-dependent cost is incurred only when invoking a stopping action which forces the system to terminate and all costs are 0 prior to stopping. For such processes, the policy under which the stopping action is never taken is not transient but its associated value function is equal to 0 at all states. Thus, they do not satisfy our assumption that that every stationary, deterministic policy which is not transient has an associated value function which is unbounded above.

The remainder of the paper is organized as follows. In Section 2, general definitions are given and some notational conventions are introduced. In Section 3, the main existence, uniqueness, and convergence results set out above are proved under the assumption that the optimal value function is bounded below. In Section 4, it is shown that if the sequence of

stationary, deterministic policies generated by the policy iteration algorithm has a pointwise-convergent subsequence, then the optimal value function is bounded below and, therefore, the results of Section 3 apply. In Section 5, the uniform termination assumption under which Pliska [16] derived his results is introduced and discussed. We show that, under this assumption, the dynamic programming operator has the contraction properties outlined above.

2. Definitions and notation

We define a *Markov decision process* to be a 5-tuple $M = (X, A, \Gamma, q, c)$, where X is the set of *nonterminal states*, A is the set of *actions*, Γ is the set of *feasible state–action pairs*, q is the *transition kernel*, and c is the *cost function*. We assume that X and A are nonempty Borel spaces and that Γ is a Borel-measurable subset of $X \times A$. For $x \in X$, the set

$$A(x) = \{a \in A : (x, a) \in \Gamma\}$$

represents the set of actions available when the system is in state x. Since X and A are Borel spaces, Γ is also a Borel space. The Borel σ-algebras on X, A, and Γ are denoted by $\mathcal{B}(X)$, $\mathcal{B}(A)$, and $\mathcal{B}(\Gamma)$, respectively. The transition kernel q is a function mapping $\mathcal{B}(X) \times \Gamma$ to $[0, 1]$ such that, for all $B \in \mathcal{B}(X)$, $q(B|\cdot)$ is a Borel-measurable function on Γ and, for all $(x, a) \in \Gamma$, $q(\cdot|x, a)$ is a subprobability measure on $\mathcal{B}(X)$ (since there is a 'loss of probability' from the system if action a is chosen in state x and the system terminates with probability $1 - q(X|x, a)$). The cost function c is a Borel-measurable function mapping Γ to the set of real numbers, $\mathbb{R}$.

To model termination, we augment the set of states X with an extra *terminal state*, $x^0 \notin X$, in which there is a single available control action, $a^0 \in A$, under which the system remains in state x^0 forever at no further cost. Let $X^0 = X \cup \{x^0\}$ denote the augmented set of states and let $\Gamma^0 = \Gamma \cup \{(x^0, a^0)\}$ denote the augmented set of feasible state–action pairs. The transition kernel q and cost function c can be extended to $\mathcal{B}(X^0) \times \Gamma^0$ and Γ^0, respectively, by setting

$$q(\{x^0\}|x^0, a^0) = 1, \qquad c(x^0, a^0) = 0. \tag{1}$$

This $q(\cdot|x, a)$ is now a probability measure for each $(x, a) \in \Gamma$.

For $k = 0, 1, \ldots$, a *k-stage trajectory* for M is a $2(k + 1)$-tuple

$$\omega_k = (x_0, a_0, x_1, a_1, \ldots, x_k, a_k),$$

where $(x_i, a_i) \in \Gamma^0$ for $i = 0, 1, \ldots, k$. For $k = 0, 1, \ldots$, let $\Omega_k = \Gamma^0 \times \Gamma^0 \times \cdots \times \Gamma^0$ (with $k + 1$ factors) denote the set of k-stage trajectories for M and let $\mathcal{F}_k$ denote the Borel σ-algebra on Ω_k. A *policy* for M is a sequence $\pi = (\pi_0, \pi_1, \ldots)$ where, for $k = 0, 1, \ldots$, π_k is a function mapping $\mathcal{B}(A) \times \Omega_{k-1} \times X$ to $[0, 1]$ such that, for all $B \in \mathcal{B}(A)$, $\pi_k(B|\cdot)$ is a Borel-measurable function and, for all $\omega_{k-1} \in \Omega_{k-1}$ and $x \in X$, $\pi_k(\cdot|\omega_{k-1}, x)$ is a probability measure concentrated on $A(x)$ (with the convention that $\Omega_{-1} = \varnothing$). Let Π denote the set of policies for M.

The policy $\pi = (\pi_k) \in \Pi$ is said to be a *Markov policy* if, for $k = 1, 2, \ldots$, there exist $\psi_k : \mathcal{B}(A) \times X \to [0, 1]$ such that, for all $\omega_{k-1} \in \Omega_{k-1}$ and $x \in X$,

$$\pi_k(\cdot|\omega_{k-1}, x) = \psi_k(\cdot|x);$$

this is written as $\pi = (\psi_k)$. The Markov policy $\pi = (\psi_k) \in \Pi$ is said to be *stationary* if $\psi_0 = \psi_1 = \cdots$. The stationary policy $(\psi, \psi, \ldots) \in \Pi$ said to be *deterministic* if, for all $x \in X$, $\psi(\cdot|x)$ assigns unit mass to some $a \in A(x)$.

https://doi.org/10.1239/jap/1158784933 Published online by Cambridge University Press

Let Ψ denote the set of functions $\psi: \mathcal{B}(A) \times X \to [0, 1]$ with the properties that, for all $B \in \mathcal{B}(A)$, $\psi(B|\cdot)$ is a Borel-measurable function and, for all $x \in X$, $\psi(\cdot|x)$ is a probability measure concentrated on $A(x)$, and let F denote the set of Borel measureable functions $f: X \to A$ satisfying $f(x) \in A(x)$ for all $x \in X$. By a standard abuse of notation, if ψ is an element of Ψ then the same symbol, 'ψ', is used to denote the associated stationary policy, and if f is an element of F then the same symbol, 'f', is used to denote the associated stationary, deterministic policy. Thus, under this convention, $F \subset \Psi \subset \Pi$.

A *trajectory* for M is a sequence

$$\omega = (x_0, a_0, x_1, a_1, \dots),$$

where $(x_k, a_k) \in \Gamma^0$ for $k = 0, 1, \dots$. Let $\Omega = \Gamma^0 \times \Gamma^0 \times \cdots$ denote the set of trajectories for M and let $\mathcal{F}$ denote the Borel σ-algebra on Ω. It follows from [15] that, for all $\pi \in \Pi$ and all probability measures μ on $(X, \mathcal{B}(X))$, we can define a unique probability measure $\mathrm{P}(\pi, \mu)$ on $\mathcal{F}$ in a canonical way. For $\pi \in \Pi$ and $x \in X$, if μ is the probability measure concentrated on $\{x\}$ then let $\mathrm{P}(\pi, x) = \mathrm{P}(\pi, \mu)$ and let $\mathrm{E}(\pi, x)$ denote the expectation operator associated with $\mathrm{P}(\pi, x)$.

The optimality criterion of interest in this paper is the so-called *total cost criterion* (originally studied in detail by Strauch [18] and Blackwell [3]), where the *value function* associated with the policy $\pi \in \Pi$ is the function $v(\pi): X \to \bar{\mathbb{R}}$ defined by

$$v(\pi)(x) = \liminf_{k \to \infty} \mathrm{E}(\pi, x)\left[\sum_{i=0}^{k} c(x_i, a_i)\right]. \tag{2}$$

Here $\bar{\mathbb{R}}$ denotes the affinely extended set of real numbers: $\bar{\mathbb{R}} = \mathbb{R} \cup \{-\infty, \infty\}$. Note that the function $v(\pi)$ is well defined for all $\pi \in \Pi$ if the cost function c is *bounded*. The *optimal value function* is the function $v^*: X \to \bar{\mathbb{R}}$ defined by

$$v^*(x) = \inf_{\pi \in \Pi} v(\pi)(x).$$

The policy $\pi \in \Pi$ is defined to be *optimal* if $v(\pi) = v^*$.

Let V denote the Banach space of real-valued, bounded, Borel-measurable functions on X with the supremum norm $\|\cdot\|$. If L is a linear operator on V then define the supremum norm of L by

$$\|L\| = \sup_{v \in V}\{\|Lv\|: \|v\| \le 1\}.$$

The linear operator L is said to be *nonnegative* if, for all $v \in V$,

$$v \ge 0 \quad \Longrightarrow \quad Lv \ge 0,$$

where $v \ge 0$ means that $v(x) \ge 0$ for all $x \in X$. If L is a nonnegative linear operator on V then it is clear that L is *monotone* in the sense that, for all $v, \bar{v} \in V$,

$$v \le \bar{v} \quad \Longrightarrow \quad Lv \le L\bar{v}.$$

It follows that if L is a nonnegative linear operator on V then $\|L\| = \|L\chi\|$, where $\chi(x) = 1$ for all $x \in X$.

https://doi.org/10.1239/jap/1158784933 Published online by Cambridge University Press

If $\psi \in \Psi$ is a stationary policy then define the function $c(\psi)\colon X \to \mathbb{R}$ by

$$c(\psi)(x) = \int_A c(x,a)\psi(\mathrm{d}a|x),$$

and define the operators $Q(\psi)$ and $T(\psi)$ on V by

$$Q(\psi)v(x) = \int_A \int_X v(y)q(\mathrm{d}y|x,a)\psi(\mathrm{d}a|x),$$
$$T(\psi)v(x) = \int_A \left[c(x,a) + \int_X v(y)q(\mathrm{d}y|x,a)\right]\psi(\mathrm{d}a|x),$$

respectively, implying that $T(\psi)v = c(\psi) + Q(\psi)v$ for all $v \in V$. A straightforward calculation shows that, for all $\psi \in \Psi$, the operator $T(\psi)$ is *monotone* in the sense that, for all $v, \bar{v} \in V$,

$$v \le \bar{v} \quad \Longrightarrow \quad T(\psi)v \le T(\psi)\bar{v}.$$

If $\pi = (\psi_k) \in \Pi$ is a Markov policy then let $Q(\pi)^0 = T(\pi)^0 = I$ (the identity operator) and, for $k = 1, 2, \ldots$, define the operators $Q(\pi)^k$ and $T(\pi)^k$ on V by

$$Q(\pi)^k = Q(\psi_0)Q(\psi_1)\cdots Q(\psi_{k-1}), \qquad T(\pi)^k = T(\psi_0)T(\psi_1)\cdots T(\psi_{k-1}), \tag{3}$$

respectively. A straightforward calculation shows that, for all Markov policies $\pi = (\psi_k) \in \Pi$, $k = 1, 2, \ldots$, and all $v \in V$,

$$T(\pi)^k v = \sum_{i=0}^{k-1} Q(\pi)^i c(\psi_i) + Q(\pi)^k v. \tag{4}$$

Furthermore, it follows from the definition of $\mathrm{P}(\pi, x)$ and (2) that, for all Markov policies $\pi = (\psi_k) \in \Pi$,

$$v(\pi) = \liminf_{k\to\infty} \sum_{i=0}^{k} Q(\pi)^i c(\psi_i) = \liminf_{k\to\infty} T(\pi)^k 0. \tag{5}$$

The following definition dates back to Veinott [19] and is key in what follows.

Definition 1. The Markov policy $\pi \in \Pi$ is said to be *transient* if

$$\left\| \sum_{k=0}^{\infty} Q(\pi)^k \right\| < \infty.$$

It follows from [16] that, for a given Markov policy $\pi \in \Pi$, the following are equivalent:

(i) π is transient;

(ii) there exist $\alpha > 0$ and $\beta \in (0, 1)$ such that $\|Q(\pi)^k\| \le \alpha\beta^k$ for $k = 0, 1, \ldots$;

(iii) $\|Q(\pi)^k\| \to 0$;

(iv) $\sum \|Q(\pi)^k\| < \infty$;

(v) $\rho(Q(\pi)) < 1$.

https://doi.org/10.1239/jap/1158784933 Published online by Cambridge University Press

Here $\rho(L)$ denotes the spectral radius of the linear operator L, defined by

$$\rho(L) = \sup_{z \in Z(L)} |z|,$$

where $Z(L)$ denotes the spectrum of L. Note that if the cost function c is bounded and the Markov policy $\pi = (\psi_k) \in \Pi$ is transient, then it follows from (5) that the function $v(\pi)$ is bounded and that we can legitimately write

$$v(\pi) = \sum_{k=0}^{\infty} Q(\pi)^k c(\psi_k).$$

Let $\pi \in \Pi$ be a Markov policy. For $k = 0, 1, \ldots$, the operator $Q(\pi)^k$ is nonnegative and, therefore, $\| Q(\pi)^k \| = \| Q(\pi)^k \chi \|$, so it follows from (i) and (iii) above that π is transient if and only if $\| Q(\pi)^k \chi \| \to 0$. It follows from the definition of $\mathrm{P}(\pi, x)$ and (3) that, for $k = 0, 1, \ldots$ and all $x \in X$,

$$Q(\pi)^k \chi(x) = \mathrm{P}(\pi, x)(x_k \neq x^0). \tag{6}$$

Thus, a Markov policy is transient if and only if under this policy the probability that termination has occurred by stage k converges uniformly to 1 as $k \to \infty$.

For another interpretation of a transient policy, let ℓ denote the length of ω, defined by $\ell = \min\{k : x_k = x^0\}$. Then, for $k = 0, 1, \ldots$, although $\{\ell > k\} \subset \{x_k \neq x^0\}$, it follows from (1) and the definition of $\mathrm{P}(\pi, x)$ that, for all $\pi \in \Pi$ and $x \in X$,

$$\mathrm{P}(\pi, x)(\ell > k) = \mathrm{P}(\pi, x)(x_k \neq x^0). \tag{7}$$

It follows from this and (6) that, for all Markov policies $\pi \in \Pi$ and states $x \in X$,

$$\mathrm{E}(\pi, x)[\ell] = \sum_{k=0}^{\infty} k \, \mathrm{P}(\pi, x)(\ell = k) = \sum_{k=0}^{\infty} \mathrm{P}(\pi, x)(\ell > k) = \sum_{k=0}^{\infty} Q(\pi)^k \chi(x).$$

Thus, from the definition, a Markov policy is transient if and only if under this policy there is a uniform bound over the set of states on the expected number of stages until termination.

If $f \in F$ is transient then, using (5), $v(f)$ can be shown to satisfy the equation

$$v = T(f)v = c(f) + Q(f)v. \tag{8}$$

In fact, if f is transient then $v(f)$ is the unique bounded solution to (8), and if (8) has a unique bounded solution v, then f is transient and $v = v(f)$. To see this, note that (8) has a unique bounded solution if and only if the operator $I - Q(f)$ is invertible, and this solution is (uniquely) given by

$$v = [I - Q(f)]^{-1} c(f). \tag{9}$$

However, from [16], f is transient if and only if $\rho(Q(f)) < 1$. It follows that $f \in F$ is transient if and only if $1 \notin Z(Q(f))$, which is true if and only if (8) has a unique bounded solution. Note that this solution can in theory be calculated using (9).

Define the *dynamic programming operator* T on V by

$$Tv(x) = \inf_{a \in A(x)} \left\{ c(x, a) + \int_X v(y) q(\mathrm{d}y | x, a) \right\}.$$

https://doi.org/10.1239/jap/1158784933 Published online by Cambridge University Press

Note that $T(\psi)v \geq Tv$ for all $\psi \in \Psi$ and $v \in V$. Furthermore, a straightforward calculation shows that the operator T is *monotone* in the sense that, for all $v, \bar{v} \in V$,

$$v \leq \bar{v} \quad \Longrightarrow \quad Tv \leq T\bar{v}.$$

Value iteration is the algorithm defined by transforming the so-called *optimality equation* $v = Tv$ into a recursive formula, as follows:

$$v_{k+1} = Tv_k$$

or, equivalently,

$$v_{k+1}(x) = \inf_{a \in A(x)} \left\{ c(x,a) + \int_X v_k(y) q(\mathrm{d}y|x,a) \right\},$$

where $v_0 \in V$. *Policy improvement* is the process of calculating a stationary, deterministic policy $f \in F$ satisfying

$$T(f)v = Tv \tag{10}$$

or, equivalently,

$$f(x) \in \arg\min_{a \in A(x)} \left\{ c(x,a) + \int_X v(y) q(\mathrm{d}y|x,a) \right\},$$

for a given function $v \in V$. *Policy iteration* is the process of calculating the functions $v_0, v_1, \ldots \in V$ and the stationary, deterministic policies $f_0, f_1, \ldots \in F$ defined by transforming (9) and (10) into recursive formulae, as follows:

$$v_k = [I - Q(f_k)]^{-1} c(f_k), \qquad T(f_{k+1})v_k = Tv_k,$$

or, in less abstract notation,

$$v_k(x) = [I - Q(f_k)]^{-1} c(f_k)(x),$$
$$f_{k+1}(x) \in \arg\min_{a \in A(x)} \left\{ c(x,a) + \int_X v_k(y) q(\mathrm{d}y|x,a) \right\},$$

with the convention that $f_{k+1}(x) = f_k(x)$ if possible, where $f_0 \in F$.

3. General results

3.1. General assumptions

All the results of this paper are derived under the following assumptions.

Assumption 1. *The following conditions hold:*

(a) *$A(x)$ is compact for all $x \in X$;*

(b) *the function*

$$a \mapsto \int_X v(y) q(\mathrm{d}y|x,a)$$

is continuous for all $v \in V$ and $x \in X$;

(c) *c is bounded and the function $a \mapsto c(x,a)$ is lower semicontinuous for all $x \in X$.*

https://doi.org/10.1239/jap/1158784933 Published online by Cambridge University Press

Assumption 2. *The following conditions hold:*

(a) *there exists a transient, stationary, deterministic policy;*

(b) *every stationary, deterministic policy which is not transient has an associated value function that is unbounded above.*

Recall that the function $g\colon A \to \mathbb{R}$ is said to be lower semicontinuous if, for all $r \in \mathbb{R}$, the set $\{a \in A\colon g(a) > r\}$ is open. Since, by assumption, A is a subset of a metric space, g is lower semicontinuous if and only if, for all sequences (a_k) of elements of A converging to $a \in A$,

$$\liminf_{k\to\infty} g(a_k) \geq g(a).$$

It follows from [11] that if Assumption 1 holds, then for all $v \in V$ there exists a policy $f \in F$ such that $T(f)v = Tv$ and, therefore, the operator T maps V to itself; indeed, this is the main motivation for the introduction of this assumption. Note that Assumption 1 holds in particular if $A(x)$ is finite for all $x \in X$.

A simple condition which implies Assumption 2(b) is that there exists an $\alpha > 0$ such that $c(x, a) \geq \alpha$ for all $(x, a) \in \Gamma$. To see this, note that if this condition holds, then, for all $f \in F$,

$$\|v(f)\| = \left\|\sum_{k=0}^{\infty} Q(f)^k c(f)\right\| \geq \alpha \left\|\sum_{k=0}^{\infty} Q(f)^k\right\|$$

and, therefore, if f is not transient, then $v(f)$ must be unbounded above. Another case where Assumption 2(b) is satisfied is when all stationary, deterministic policies are transient, or, more restrictively, when there is a uniform bound on the expected number of stages until termination which holds under all stationary, deterministic policies. The latter case is considered in Section 5.

As motivation for Assumption 2(b), we note that when it and Assumption 2(a) both hold, there cannot exist an optimal stationary, deterministic policy which is not transient. Thus, intuitively speaking, we would expect to be able to restrict attention to transient, stationary, deterministic policies, although it is feasible that there may exist a policy which is not transient but which, for some initial states, has smaller expected total cost than all transient stationary, deterministic policies. In the following section it will be shown that if the optimal value function is bounded below, then this is not the case and we can indeed restrict attention to transient, stationary, deterministic policies.

3.2. Preliminary lemmas

The proofs of the main results of this section are based on the following three lemmas. The first of these gives an essential characterization of the value function associated with a transient stationary policy and generalizes Lemma 1(a) of [2] to the case where the set of states may be infinite. Although this result will only be used in the context of stationary, deterministic policies, for completeness it is stated in full generality.

Lemma 1. *Let $\psi \in \Psi$ be a transient, stationary policy. Then, for all $v \in V$,*

$$\lim_{k\to\infty} T(\psi)^k v = v(\psi).$$

Proof. It follows from (4) that, for $k = 0, 1, \ldots$ and $v, \bar{v} \in V$,

$$\|T(\psi)^k v - T(\psi)^k \bar{v}\| = \|Q(\psi)^k (v - \bar{v})\| \leq \|Q(\psi)^k\| \|v - \bar{v}\|. \tag{11}$$

https://doi.org/10.1239/jap/1158784933 Published online by Cambridge University Press

Since ψ is transient, there exists a positive integer k such that $\|Q(\psi)^k\| < 1$. It follows from this and (11) that the operator $T(\psi)^k$ is a contraction mapping on V with respect to $\|\cdot\|$. The result therefore follows from the Banach fixed-point theorem.

The following lemma is a simple consequence of Assumption 2 and generalizes Lemma 1(b) of [2] to the case where the set of states may be infinite.

Lemma 2. *Let Assumption 2 hold and let $f \in F$ be a stationary, deterministic policy. If there exists a $v \in V$ such that $v \geq T(f)v$, then f is transient and $v \geq v(f)$.*

Proof. Let $v \in V$ be such that $v \geq T(f)v$. Then, by the monotonicity of $T(f)$ and (4),

$$v \geq T(f)^k v = \sum_{i=0}^{k-1} Q(f)^i c(f) + Q(f)^k v.$$

Taking the limit inferior as $k \to \infty$ of both sides of the above inequality and using (5) yields $v \geq v(f) + \alpha\chi$, where α is a lower bound on v. If f were not transient then $v(f)$ would be unbounded above, by Assumption 2, which is a contradiction. Thus, f is transient and, therefore, $Q(f)^k \to 0$, yielding $v \geq v(f)$.

The following lemma gives a useful 'monotone convergence' property of the dynamic programming operator T.

Lemma 3. *Let $v_0, v_1, \ldots, v \in V$.*

(i) *If $v_k \downarrow v$ then $Tv_k \downarrow Tv$.*

(ii) *If Assumption 1 holds and $v_k \uparrow v$, then $Tv_k \uparrow Tv$.*

Proof. For $k = 0, 1, \ldots$, define the function $u_k : \Gamma \to \mathbb{R}$ by

$$u_k(x, a) = c(x, a) + \int_X v_k(y) q(\mathrm{d}y|x, a),$$

and define the function $u : \Gamma \to \mathbb{R}$ by

$$u(x, a) = c(x, a) + \int_X v(y) q(\mathrm{d}y|x, a).$$

If $v_k \downarrow v$ then $u_k \downarrow u$ by the monotone convergence theorem, and it follows from [12, p. 18] that, for all $x \in X$,

$$\inf_{a \in A(x)} u_k(x, a) \downarrow \inf_{a \in A(x)} u(x, a),$$

from which (i) follows. If $v_k \uparrow v$ then $u_k \uparrow u$ by the monotone convergence theorem, and it follows from [9] that if Assumption 1 holds then, for all $x \in X$,

$$\inf_{a \in A(x)} u_k(x, a) \uparrow \inf_{a \in A(x)} u(x, a),$$

from which (ii) follows.

https://doi.org/10.1239/jap/1158784933 Published online by Cambridge University Press

3.3. The optimality equation

The following lemma gives initial results concerning the optimality equation and the convergence of the policy iteration algorithm.

Lemma 4. *Let Assumptions 1 and 2 hold and suppose that the optimal value function is bounded below. Then there exists a unique bounded function $v \in V$ satisfying the optimality equation $v = Tv$. Furthermore, if (v_k) is generated by the policy iteration algorithm starting from any transient, stationary, deterministic policy, then $v_k \downarrow v$.*

Proof. First it will be shown that if the optimality equation has a bounded solution, then it is unique. Let $v, \bar{v} \in V$ satisfy $v = Tv$ and $\bar{v} = T\bar{v}$, respectively. Then, since Assumption 1 holds, we can choose stationary, deterministic policies $f, \bar{f} \in F$ such that $v = T(f)v$ and $\bar{v} = T(\bar{f})\bar{v}$. It follows from Lemma 2 that f and $\bar{f}$ are transient, and it follows from Lemma 1 that $v = v(f)$ and $\bar{v} = v(\bar{f})$. Since, for $k = 0, 1, \ldots,$

$$v = T^k v \leq T(\bar{f})^k v,$$

it follows from Lemma 1 that

$$v \leq \lim_{k\to\infty} T(\bar{f})^k v = \bar{v}.$$

Similarly $\bar{v} \leq v$, so $v = \bar{v}$.

Now it will be shown that the policy iteration algorithm converges to v starting from any transient, stationary, deterministic policy. Given a transient, stationary, deterministic policy $f \in F$, since Assumption 1 holds we can choose an $\bar{f} \in F$ such that

$$T(\bar{f})v(f) = Tv(f).$$

It follows from Lemma 2 that $\bar{f}$ is transient, and by the monotonicity of $T(\bar{f})$ and Lemma 1 that

$$v(f) = T(f)v(f) \geq Tv(f) = T(\bar{f})v(f) \geq \lim_{k\to\infty} T(\bar{f})^k v(f) = v(\bar{f}).$$

Continuing in this manner, we can construct a sequence (f_k) of stationary, deterministic policies such that, for $k = 0, 1, \ldots,$ f_k is transient and

$$v(f_k) \geq Tv(f_k) \geq v(f_{k+1}). \tag{12}$$

For $k = 0, 1, \ldots,$ let $v_k = v(f_k)$. Since the functions $v_0, v_1, \ldots$ are nonincreasing, there exists a function $v \geq v^*$ such that $v_k \downarrow v$. Clearly v is bounded, and it follows from Lemma 3 that $Tv_k \downarrow Tv$. Thus, taking the limit as $k \to \infty$ in (12) shows that $v = Tv$ and, hence, that v is the unique bounded fixed point of T.

3.4. Convergence of policy iteration

The following lemma gives a convergence result for the policy iteration algorithm which is stronger than that given by Lemma 4. The proof of the lemma follows along similar lines to the proof of the analogous result given in [2] for the case where the set of states is finite.

Lemma 5. *Let Assumptions 1 and 2 hold and suppose that the optimal value function is bounded below. If (v_k) is generated by the policy iteration algorithm starting from any transient, stationary, deterministic policy, then there exists an $f \in F$ such that $v_k \downarrow v(f)$.*

https://doi.org/10.1239/jap/1158784933 Published online by Cambridge University Press

Proof. Let (v_k) be generated by the policy iteration algorithm and, for $k = 0, 1, \ldots$, define the function $u_k : \Gamma \to \mathbb{R}$ by

$$u_k(x, a) = c(x, a) + \int_X v_k(y) q(\mathrm{d}y|x, a). \tag{13}$$

Then, for $k = 0, 1, \ldots$ and $x \in X$,

$$v_k(x) = u_k(x, f_k(x)). \tag{14}$$

Also, since $v_0 \geq v_1 \geq \cdots$ by Lemma 4, it follows from (13) that $u_0 \geq u_1 \geq \cdots$ and, therefore, that $u_k \downarrow u$ by the monotone convergence theorem, where

$$u(x, a) = c(x, a) + \int_X v(y) q(\mathrm{d}y|x, a),$$

and where v is as defined in Lemma 4. It follows from [17] that there exists a stationary, deterministic policy $f \in F$ such that, for all $x \in X$, $f(x)$ is an accumulation point of $(f_k(x))$. Fix an $x \in X$. Then there exists a sequence of positive integers (k_i) such that $f_{k_i}(x) \to f(x)$. It follows from [8], Assumption 1, (13), and (14) that

$$v(x) = \lim_{k\to\infty} u_k(x, f_k(x)) = \lim_{i\to\infty} u_{k_i}(x, f_{k_i}(x)) \geq u(x, f(x)) = T(f)v(x).$$

Thus, f is transient and $v \geq v(f)$ by Lemma 2. From Lemma 4, for $k = 0, 1, \ldots$,

$$T(f)v(f_k) \geq Tv(f_k) \geq v.$$

Taking the limit as $k \to \infty$ in the above and using the monotone convergence theorem gives $T(f)v \geq v$. From Lemma 1, this implies that

$$v(f) = \lim_{k\to\infty} T(f)^k v \geq v,$$

whence $v = v(f)$.

3.5. Convergence of value iteration

The following lemma gives a convergence result for the value iteration algorithm. Again, the proof of the lemma follows along the same lines as the proof of the analogous result given in [2] for the case where the set of states is finite.

Lemma 6. *Let Assumptions 1 and 2 hold and suppose that the optimal value function is bounded below. Let $f \in F$ be such that the policy iteration algorithm converges to $v(f)$ starting from any stationary, deterministic policy (such a policy exists, according to Lemma 5). Then the value iteration algorithm also converges to $v(f)$ starting from any bounded function.*

Proof. Fix an $\alpha > 0$ and let v_α be the unique element of V satisfying

$$T(f)v_\alpha = v_\alpha - \alpha\chi.$$

To see that there is a unique such element of V, note that since f is transient, it must also be transient for a transformed problem in which the cost function is increased by α, and, therefore,

https://doi.org/10.1239/jap/1158784933 Published online by Cambridge University Press

the equation $v = c(f) + \alpha\chi + Q(f)v$ has a unique solution in V. Moreover, it is clear that $v_\alpha \geq v(f)$. It follows from this and the monotonicity of T that

$$v(f) = Tv(f) \leq Tv_\alpha \leq T(f)v_\alpha = v_\alpha - \alpha\chi \leq v_\alpha.$$

In turn, it follows from this and the monotonicity of T that, for $k = 0, 1, \ldots,$

$$v(f) = T^{k+1}v(f) \leq T^{k+1}v_\alpha \leq T^k v_\alpha \leq v_\alpha. \tag{15}$$

From (15), the sequence $(T^k v_\alpha)$ is nonincreasing and bounded below by $v(f)$, and therefore converges to some bounded function. Furthermore, it follows from Lemma 3 that this function satisfies the optimality equation and, therefore, that $T^k v_\alpha \downarrow v(f)$ since, as shown earlier, $v(f)$ is the unique fixed point of T in V. A straightforward calculation shows that, for all $v \in V$,

$$Tv + \alpha\chi \geq T(v + \alpha\chi), \qquad Tv - \alpha\chi \leq T(v - \alpha\chi);$$

this is often referred to as the *cost shifting* property of the operator T. It follows from this and the monotonicity of the operator T that

$$v(f) - \alpha\chi = Tv(f) - \alpha\chi \leq T(v(f) - \alpha\chi) \leq Tv(f) = v(f). \tag{16}$$

From (16), the sequence $(T^k(v_\alpha - \alpha\chi))$ is nondecreasing and bounded above by $v(f)$. Therefore, as above, it follows that $T^k(v_\alpha - \alpha\chi) \uparrow v(f)$. Note that, for all $\alpha > 0$, since $v_\alpha \geq v(f)$ and f is transient, we have

$$v_\alpha = T(f)v_\alpha + \alpha\chi \geq T(f)v(f) + \alpha\chi = v(f) + \alpha\chi.$$

Thus, for all $v \in V$, we can find an $\alpha > 0$ such that $v(f) - \alpha\chi \leq v \leq v_\alpha$. By the monotonicity of T, for $k = 0, 1, \ldots$ we have

$$T^k(v(f) - \alpha\chi) \leq T^k v \leq T^k v_\alpha,$$

and since $T^k(v(f) - \alpha\chi) \to v(f)$ and $T^k v_\alpha \to v(f)$, it follows that $T^k v \to v(f)$.

3.6. Optimality results

The main results of this section are given by the following theorem. These results are in fact simple consequences of Lemmas 4, 5, and 6, and a result of [18] which states that the optimal value function v^* is equal to the infimum over the set of *Markov* policies $\pi \in \Pi$ of the functions $v(\pi)$. Thus, to show that the stationary, deterministic policy $f \in F$ is optimal, it suffices to show that $v(\pi) \geq v(f)$ for all Markov policies $\pi \in \Pi$.

Theorem 1. *Let Assumptions 1 and 2 hold and suppose that the optimal value function is bounded below. Then*

(i) *there exists an optimal stationary, deterministic policy;*

(ii) *the optimal value function is the unique solution to the optimality equation;*

(iii) *the value iteration algorithm converges to the optimal value function starting from any bounded function; and*

(iv) *the policy iteration algorithm converges to the optimal value function starting from any transient, stationary, deterministic policy.*

https://doi.org/10.1239/jap/1158784933 Published online by Cambridge University Press

Proof. It follows from Lemmas 4, 5, and 6 that there exists a stationary, deterministic policy $f \in F$ such that the function $v(f)$ is the unique solution to the optimality equation, the value iteration algorithm converges to the function $v(f)$ starting from any bounded function, and the policy iteration algorithm converges to the function $v(f)$ starting from any transient, stationary, deterministic policy. All that remains to be shown is that f is optimal, that is, that $v(f) = v^*$. To do this, let $\pi \in \Pi$ be an arbitrary Markov policy. Then, for $k = 0, 1, \ldots,$

$$T(\pi)^k 0 \geq T^k 0.$$

Taking the limit inferior as $k \to \infty$ of both sides of the above inequality and using (5) yields $v(\pi) \geq v(f)$, so $v(f) = v^*$ since π was arbitrary.

This section concludes with the following theorem, which gives necessary and sufficient conditions for a given stationary, deterministic policy to be optimal under Assumptions 1 and 2. The theorem essentially says that if we know the optimal value function v^*, and v^* is bounded below, then we can, in theory, obtain an optimal policy.

Theorem 2. *Let Assumptions 1 and 2 hold. Then the stationary, deterministic policy $f \in F$ is optimal if and only if v^* is bounded below and $v^* = T(f)v^*$.*

Proof. If f is optimal then f is transient and $v(f) = v^*$, so v^* is bounded below and satisfies $v^* = T(f)v^*$ by Lemma 1. Conversely, if v^* is bounded below and $v^* = T(f)v^*$, then f is transient by Lemma 2, so $v^* = v(f)$ by Lemma 1 and f is optimal.

The condition that $v^* = T(f)v^*$ or, in less abstract notation, that, for all $x \in X$,

$$v^*(x) = c(x, f(x)) + \int_X v^*(y) q(\mathrm{d}y|x, f(x)),$$

is referred to in the literature as the *conserving* property of the policy $f \in F$. Theorem 2 is similar to Theorem 4.12 of [10], which was derived under assumptions different to Assumptions 1 and 2.

4. Boundedness of the optimal value function

In Section 3, it was shown that a sufficient condition for parts (i)–(iv) of Theorem 1 to hold under Assumptions 1 and 2 is that the optimal value function be bounded below. In fact, this is also a necessary condition, since if there exists an optimal stationary, deterministic policy, then this policy must be transient by Assumption 2, and the optimal value function is therefore bounded below. However, this condition may be difficult to directly verify in practice.

In this section an alternative sufficient condition is therefore given for parts (i)–(iv) of Theorem 1 to hold – or, alternatively, for the optimal value function to be bounded below – under Assumptions 1 and 2. We start with the following lemma, which generalizes Lemma 3 of [2] to the case where the set of states may be infinite.

Lemma 7. *Let Assumptions 1 and 2 hold and let (f_k) be a sequence of transient, stationary, deterministic policies which converges to $f \in F$.*

(i) *If f is transient then*

$$\liminf_{k \to \infty} v(f_k) \geq v(f).$$

(ii) *If f is not transient then the sequence $(v(f_k))$ is unbounded above.*

https://doi.org/10.1239/jap/1158784933 Published online by Cambridge University Press

Proof. Since $f_k \to f$, it follows from Assumption 1 and [8] that

$$\liminf_{k\to\infty} Q(f_k)c(f_k) \geq \lim_{k\to\infty} Q(f_k)\Big[\liminf_{k\to\infty} c(f_k)\Big] \geq Q(f)c(f).$$

Suppose that, for some nonnegative integer i,

$$\liminf_{k\to\infty} Q(f_k)^i c(f_k) \geq Q(f_k)^i c(f). \tag{17}$$

It then follows from [8] that

$$\begin{aligned}\liminf_{k\to\infty} Q(f_k)^{i+1}c(f_k) &= \liminf_{k\to\infty} Q(f_k)Q(f_k)^i c(f_k)\\ &\geq \lim_{k\to\infty} Q(f_k)\Big[\liminf_{k\to\infty} Q(f_k)^i c(f_k)\Big]\\ &\geq Q(f)^{i+1}c(f)\end{aligned}$$

and, hence, by induction, (17) holds for all nonnegative integers i.

To prove (i), suppose that f is transient. Then, since, from (17),

$$\sum_{i=0}^{\infty}\Big[\liminf_{k\to\infty} Q(f_k)^i c(f_k)\Big]^- \leq \sum_{i=0}^{\infty}[Q(f)^i c(f)]^-,$$

we can legitimately sum over i in (17), to obtain

$$\sum_{i=0}^{\infty}\liminf_{k\to\infty} Q(f_k)^i c(f_k) \geq \sum_{i=0}^{\infty} Q(f)^i c(f).$$

By Fatou's lemma,

$$\liminf_{k\to\infty}\sum_{i=0}^{\infty} Q(f_k)^i c(f_k) \geq \sum_{i=0}^{\infty}\liminf_{k\to\infty} Q(f_k)^i c(f_k) \geq \sum_{i=0}^{\infty} Q(f)^i c(f),$$

and the result follows.

To prove (ii), suppose that f is not transient and fix an $r \in \mathbb{R}$. Then, by Assumption 2 there exists an $x \in X$ such that $v(f)(x) > r$, so, by (5),

$$\sum_{i=0}^{m} Q(f)^i c(f)(x) > r$$

for all sufficiently large m. By Fatou's lemma and (17), for $m = 0, 1, \ldots,$

$$\liminf_{k\to\infty}\sum_{i=0}^{m} Q(f_k)c(f_k)(x) \geq \sum_{i=0}^{m}\liminf_{k\to\infty} Q(f_k)c(f_k)(x) \geq \sum_{i=0}^{m} Q(f)^i c(f)(x),$$

from which it follows that

$$\sum_{i=0}^{m} Q(f_k)^i c(f_k)(x) > r$$

for all sufficiently large k and m. In taking the limit as $m \to \infty$ in the above inequality, it is clear that $v(f_k)(x) > r$ for all sufficiently large k. However, $r \in \mathbb{R}$ was arbitrary, so the result is proved.

https://doi.org/10.1239/jap/1158784933 Published online by Cambridge University Press

The main result of this section is given by the following theorem.

Theorem 3. *Let Assumptions 1 and 2 hold and suppose that the sequence of transient, stationary, deterministic policies (f_k) generated by the policy iteration algorithm has a subsequence which converges to a stationary, deterministic policy $f \in F$. Then parts (i)–(iv) of Theorem 1 hold and, furthermore, f is optimal.*

Proof. Suppose that the subsequence (f_{k_i}) converges to the stationary, deterministic policy $f \in F$. As in the proof of Lemma 4, the sequence $(v(f_k))$ converges and is bounded above. Therefore, by Lemma 7, f is transient and

$$\lim_{k\to\infty} v(f_k) = \liminf_{i\to\infty} v(f_{k_i}) \geq v(f).$$

In fact, by continuing as in the proof of Lemma 5 we can show that

$$\lim_{k\to\infty} v(f_k) = v(f).$$

By continuing as in the proof of Theorem 1 we can now verify that parts (i)–(iv) of Theorem 1 hold and that f is optimal.

If the set of states is countable then the set of stationary, deterministic policies is a compact subset of a metric space and, therefore, any sequence (f_k) of stationary, deterministic policies has a convergent subsequence. Thus, in this case, Theorem 3 applies and the optimal value function is bounded below.

If the set of states is uncountable then it is not clear whether the sequence of stationary, deterministic policies (f_k) generated by the policy iteration algorithm has a convergent subsequence. It is certainly true that if the set of states is uncountable then a general sequence of stationary, deterministic policies (f_k) does not necessarily have a convergent subsequence. This can be seen by letting $X = [0, 1]$, $A = \{0, 1\}$, and $\Gamma = X \times A$, so $A(x) = A$ for all $x \in X$, and, for $k = 0, 1, \ldots$, letting $f_k(x)$ be equal to the $(k + 1)$th digit in the binary expansion of x. The functions $f_0, f_1, \ldots$ are Borel measurable, and are therefore stationary, deterministic policies as defined in Section 2. However, given any subsequence (f_{k_i}) of (f_k), we can find an $x \in X$ such that $f_{k_i}(x) = 0$ if i is even and $f_{k_i}(x) = 1$ if i is odd, so the sequence (f_k) does not have a convergent subsequence.

We cannot therefore use Theorem 3 directly to show that the optimal value function is bounded below in the general case. It would be nice to know whether parts (i)–(iv) of Theorem 1 hold when the set of states is uncountable without the additional assumption that the optimal value function is bounded below, but the authors have been unable to either prove that this is the case or find a counterexample.

5. Uniform termination results

This section extends the results of Pliska [16] by showing that under his assumptions, which represent a strengthening of those considered so far, the dynamic programming operator is a k-stage contraction with respect to the usual supremum norm, for some k, and a 1-stage contraction with respect to a weighted supremum norm. Specifically, Pliska [16] derived results under Assumption 1 and the following assumption, which is a strengthening of Assumption 2.

Assumption 3. *There exists a $\theta < \infty$ such that, for all $f \in F$,*

$$\left\| \sum_{k=0}^{\infty} Q(f)^k \right\| \leq \theta.$$

https://doi.org/10.1239/jap/1158784933 Published online by Cambridge University Press

Clearly, if Assumption 3 holds then all stationary, deterministic policies are transient. The converse is obviously true when the sets of states and actions available in each state are finite, since in this case the set of stationary, deterministic policies is also finite. Pliska [16] showed that the following are equivalent:

(i) Assumption 3 holds;

(ii) there exist $\alpha > 0$ and $\beta \in (0, 1)$ such that $\|Q(f)^k\| \leq \alpha\beta^k$ for $k = 0, 1, \ldots$ and all $f \in F$;

(iii) for all $\varepsilon > 0$, there exists a k such that $\|Q(f)^k\| \leq \varepsilon$ for all $f \in F$;

(iv) there exists a $\kappa > 0$ such that, for all $f \in F$,

$$\sum_{k=0}^{\infty} \|Q(f)^k\| \leq \kappa;$$

(v) there exists a $\zeta < 1$ such that $\rho(Q(f)) \leq \zeta$ for all $f \in F$.

Hernández-Lerma *et al.* [10] showed that if Assumption 3 holds then

$$\left\| \sum_{k=0}^{\infty} Q(\pi)^k \right\| \leq \theta$$

for all Markov policies $\pi \in \Pi$. This means that if Assumption 3 holds then the optimal value function is bounded below by $-\theta\|c\|$. We can conclude from this that if Assumptions 1 and 3 hold then parts (i)–(iv) of Theorem 1 hold. It fact, it is not actually necessary to use the result of [10] to reach this conclusion, since the proof of Theorem 1 holds directly under Assumptions 1 and 3.

The k-stage contraction property of the dynamic programming operator T under Assumptions 1 and 3 stems from the following lemma, which was essentially proved by Pliska [16]. The proof of the lemma is restated here, partly for convenience and partly because the result as stated here is more general than that proved by Pliska [16], who only proved it in the case $\delta = \frac{1}{2}$, although the extension to different values of δ is trivial.

Lemma 8. *Let Assumptions 1 and 3 hold. Then* $\|Q(\pi)^k\| \leq \delta$ *for all* $\delta \in (0, 1)$, *all* $k > \theta\delta^{-1}$, *and all Markov policies* $\pi \in \Pi$.

Proof. Fix a $\delta \in (0, 1)$ and a $k > \theta\delta^{-1}$ and let $\pi \in \Pi$ be a Markov policy. Suppose that there exists an $x \in X$ such that $Q(\pi)^k\chi(x) > \delta$. Then $Q(\pi)^i\chi(x) > \delta$ for $i = 0, 1, \ldots, k$ and, therefore,

$$\sum_{i=0}^{k} Q(\pi)^i\chi(x) > k\delta > \theta,$$

which contradicts Assumption 3.

The main result of this section is given by the following theorem, which generalizes Theorem 5.1 of [4] to the case where the set of states may be infinite.

Theorem 4. *Let Assumptions 1 and 3 hold. Then, for all* $\delta \in (0, 1)$, $k > \theta\delta^{-1}$, *and* $v, \bar{v} \in V$,

$$\|T^k v - T^k \bar{v}\| \leq \delta\|v - \bar{v}\|.$$

https://doi.org/10.1239/jap/1158784933 Published online by Cambridge University Press

Proof. Let $v, \bar{v} \in V$ and let $f, \bar{f} \in F$ be stationary, deterministic policies satisfying $T(f)v = Tv$ and $T(\bar{f})\bar{v} = T\bar{v}$, respectively. Then

$$Tv - T\bar{v} \leq Q(\bar{f})|v - \bar{v}|, \qquad T\bar{v} - Tv \leq Q(f)|v - \bar{v}|.$$

Define the stationary, deterministic policy $f_0 \in F$ by

$$f_0(x) = \begin{cases} f(x), & Q(f)|v - \bar{v}|(x) \geq Q(\bar{f})|v - \bar{v}|(x), \\ \bar{f}(x), & Q(f)|v - \bar{v}|(x) < Q(\bar{f})|v - \bar{v}|(x). \end{cases}$$

Then it is clear that

$$|Tv - T\bar{v}| \leq Q(f_0)|v - \bar{v}|.$$

By repeating the argument above, we can show that, for $k = 0, 1, \ldots$, there exists a stationary, deterministic policy $f_k \in F$ such that

$$|T^{k+1}v - T^{k+1}\bar{v}| \leq Q(f_k)|T^k v - T^k \bar{v}|.$$

It follows that, for $k = 0, 1, \ldots$,

$$|T^k v - T^k \bar{v}| \leq Q(\pi)^k |v - \bar{v}|, \tag{18}$$

where $\pi = (f_k) \in \Pi$. Fix a $\delta \in (0, 1)$. It follows from Lemma 8 that $\|Q(\pi)^k\| \leq \delta$ for all $k > \theta\delta^{-1}$. It therefore follows from (18) that, for all $k > \theta\delta^{-1}$,

$$\|T^k v - T^k \bar{v}\| \leq \|Q(\pi)^k\| \, \|v - \bar{v}\| \leq \delta \|v - \bar{v}\|,$$

which proves the result.

Pliska [16] showed that if Assumptions 1 and 3 hold then there exists a unique function $w\colon X \to [1, \theta]$ satisfying

$$w(x) = \max_{a \in A(x)} \left\{ 1 + \int_X w(y) q(\mathrm{d}y|x, a) \right\} \tag{19}$$

for all $x \in X$. In fact, by considering a modified process in which $c(x, a) = -1$ for all $(x, a) \in \Gamma$, it can be seen that the existence and uniqueness of the function w under Assumptions 1 and 3 follows from Theorem 1. Define the norm $\|\cdot\|_w$ on V by

$$\|v\|_w = \left\| \frac{v}{w} \right\| = \sup_{x \in X} \frac{|v(x)|}{w(x)}.$$

The norms $\|\cdot\|$ and $\|\cdot\|_w$ are equivalent, since, for all $v \in V$,

$$\|v\|_w \leq \|v\| \leq \|w\| \, \|v\|_w \leq \theta \|v\|_w$$

(using the fact that $1 \leq w(x) \leq \theta$ for all $x \in X$). Thus, convergence in the norm $\|\cdot\|$ implies convergence in the norm $\|\cdot\|_w$, and vice versa. If L is a linear operator on V then let

$$\|L\|_w = \sup_{v \in V} \{\|Lv\|_w \colon \|v\|_w \leq 1\}.$$

https://doi.org/10.1239/jap/1158784933 Published online by Cambridge University Press

If L is a nonnegative linear operator on V then $\|L\|_w = \|Lw\|_w$. Thus, if $\psi \in \Psi$ is a stationary policy then

$$\|Q(\psi)\|_w = \sup_{x\in X} \frac{1}{w(x)} \int_A \int_X w(y) q(\mathrm{d}y|x,a)\psi(\mathrm{d}a|x). \tag{20}$$

The following theorem says that if Assumption 3 holds then the dynamic programming operator is a contraction mapping on V with respect to $\|\cdot\|_w$. The theorem generalizes Lemma 3 of [19] to the case where the set of states may be infinite.

Theorem 5. *Let Assumptions 1 and 3 hold. Then, for all $v, \bar{v} \in V$,*

$$\|Tv - T\bar{v}\|_w \le \delta \|v - \bar{v}\|_w,$$

where $\delta = 1 - \theta^{-1} < 1$.

Proof. Fix an $x \in X$. Multiplying (19) by $1/w(x)$ and then subtracting $1/w(x)$ yields

$$1 - \frac{1}{w(x)} = \max_{a\in A(x)} \left\{ \frac{1}{w(x)} \int_X w(y) q(\mathrm{d}y|x,a) \right\}.$$

It follows from this and (20) that, for all stationary, deterministic policies $f \in F$,

$$\|Q(f)\|_w \le 1 - \frac{1}{\|w\|} \le 1 - \frac{1}{\theta} < 1. \tag{21}$$

As in the proof of Theorem 4, for all $v, \bar{v} \in V$ there exists a stationary, deterministic policy $f_0 \in F$ such that

$$|Tv - T\bar{v}| \le Q(f_0)|v - \bar{v}|.$$

It follows from this and (21) that

$$\|Tv - T\bar{v}\|_w \le \|Q(f_0)\|_w \|v - \bar{v}\|_w \le \delta \|v - \bar{v}\|_w,$$

which proves the result.

With the aid of either Theorem 4 or Theorem 5 we can prove the following.

Theorem 6. *Let Assumptions 1 and 3 hold and let (v_k) be generated by the value iteration or policy iteration algorithm. Then $\|v_k - v^*\| \to 0$.*

Proof. It follows from Theorem 1 that if (v_k) is generated by the value iteration algorithm then $v_k \to v^*$ and, therefore, $\|v_k - v^*\| \to 0$ from Theorem 4 or Theorem 5 and the Banach fixed-point theorem. Now let $((f_k, v_k))$ be generated by the policy iteration algorithm and let (u_k) be generated by the value iteration algorithm starting from v_0. It will be shown by induction that $u_k \ge v_k$ for all k. The result is true when $k = 0$, by definition. Suppose that the result is true for arbitrary k. Then, by the monotonicity of the operator T,

$$u_{k+1} = Tu_k \ge Tv_k. \tag{22}$$

Also, by the monotonicity of $T(f_k)$ and Lemma 1,

$$v_k = T(f_k)v_k \ge Tv_k = T(f_{k+1})v_k \ge \lim_{i\to\infty} T(f_{k+1})^i v_k = v_{k+1}. \tag{23}$$

Combining (22) and (23) yields $u_{k+1} \ge v_{k+1}$, so the induction hypothesis holds. Since $v_k \downarrow v^*$ by the proof of Theorem 1, $u_k \ge v_k$ for all k, and $\|u_k - v^*\| \to 0$, it follows that $\|v_k - v^*\| \to 0$.

https://doi.org/10.1239/jap/1158784933 Published online by Cambridge University Press

Acknowledgements

The first author was supported by a research studentship from QinetiQ Ltd. The authors are grateful to Simon Maskell (QinetiQ Ltd) for a number of useful discussions during the preparation of this article, and to an anonymous referee for helpful comments on an earlier draft.

References

[1] Bertsekas, D. P. and Tsitsiklis, J. N. (1989). *Parallel and Distributed Computation: Numerical Methods*. Prentice-Hall, Englewood Cliffs, NJ.
[2] Bertsekas, D. P. and Tsitsiklis, J. N. (1991). An analysis of stochastic shortest path problems. *Math. Operat. Res.* **16,** 580–595.
[3] Blackwell, D. (1967). Positive dynamic programming. In *Proc. 5th Berkeley Symp. Math. Statist. Prob.* (Berkeley, CA, 1965/66), Vol. 1, University of California Press, Berkeley, pp. 415–418.
[4] Derman, C. (1970). *Finite State Markovian Decision Processes*. Academic Press, New York.
[5] Dynkin, E. B. (1963). The optimum choice of the instant for stopping a Markov process. *Soviet Math. Dokl.* **150,** 238–240.
[6] Eaton, J. H. and Zadeh, L. A. (1962). Optimal pursuit strategies in discrete-state probabilistic systems. *Trans. ASME Ser. D J. Basic Eng.* **84,** 23–29.
[7] Grigelionis, R. I. and Shiryaev, A. N. (1966). On Stefan's problem and optimal stopping rules for Markov processes. *Theory Prob. Appl.* **11,** 541–558.
[8] Hernández-Lerma, O. and Lasserre, J. B. (2000). Fatou's lemma and Lebesgue's convergence theorem for measures. *J. Appl. Math. Stoch. Anal.* **13,** 137–146.
[9] Hernández-Lerma, O. and Muñoz de Ozak, M. (1992). Discrete-time MDPs with discounted unbounded costs: optimality criteria. *Kybernetika* **528,** 191–212.
[10] Hernández-Lerma, O., Carrasco, G. and Pérez-Hernández, R. (1999). Markov control processes with the expected total cost criterion: optimality, stability, and transient models. *Acta Appl. Math.* **59,** 229–269.
[11] Himmelberg, C. J., Parthasarathy, T. and Van Vleck, F. S. (1976). Optimal plans for dynamic programming problems. *Math. Operat. Res.* **1,** 390–394.
[12] Hinderer, K. (1970). *Foundations of Non-Stationary Dynamic Programming with Discrete Time Parameter*. Springer, New York.
[13] Hinderer, K. and Waldmann, K. H. (2003). The critical discount factor for finite Markovian decision processes with an absorbing set. *Math. Meth. Operat. Res.* **57,** 1–19.
[14] Hinderer, K. and Waldmann, K. H. (2005). Algorithms for countable state Markov decision models with an absorbing set. *SIAM J. Control Optimization* **43,** 2109–2131.
[15] Ionescu Tulcea, C. T. (1949). Measures dans les espaces produits. *Atti Accad. Naz. Lincei. Rende. Cl. Sci. Fis. Mat. Nat. (8)* **7,** 208–211.
[16] Pliska, S. R. (1978). On the transient case for Markov decision chains with general state spaces. In *Dynamic Programming and Its Applications*, ed. M. L. Puterman, Academic Press, New York, pp. 335–349.
[17] Schäl, M. (1974). A selection theorem for optimization problems. *Arch. Math. (Basel)* **25,** 219–224.
[18] Strauch, R. E. (1966). Negative dynamic programming. *Ann. Math. Statist.* **37,** 871–890.
[19] Veinott, A. F. (1969). Discrete dynamic programming with sensitive discount optimality criteria. *Ann. Math. Statist.* **40,** 1635–1660.