Total Expected Discounted Reward MDPs: Existence of Optimal Policies

Eugene A. Feinberg*

Department of Applied Mathematics and Statistics

State University of New York at Stony Brook

Stony Brook, NY 11794-3600

Abstract

This article describes the results on the existence of optimal and nearly optimal policies for Markov Decision Processes (MDPs) with total expected discounted rewards. The problem of optimization of total expected discounted rewards for MDPs is also known under the name of discounted dynamic programming.

1 Introduction

Deterministic optimal policies always exist for discounted dynamic programming problems with finite action sets. Such policies also exist when action sets satisfy certain compactness conditions, and transition probabilities and reward functions satisfy certain continuity conditions. If either compactness or continuity conditions do not hold, deterministic ϵ-optimal policies exist for problems with countable state spaces. For problems with uncountable Borel state spaces, the results similar to the existence of deterministic ϵ-optimal policies hold, but in this more general case either the notion of ϵ-optimality should be replaced with the weaker notion of (p, ϵ)-optimality or a broader definition of a policy is required. Since the theory is simpler when the state space is countable, problems with countable and uncountable state spaces are considered separately in this chapter.

2 Countable state space

2.1 Definitions

Consider a Markov Decision Process (MDP) with the state space X, action space A, sets of actions $A(x)$ available at states $x \in X$, transition probabilities p, and rewards r.

*efeinberg@notes.cc.sunysb.edu

We assume throughout this section that the state space X is countable. In particular, it can be either finite or countably infinite.

We assume that the action set A is a complete separable metric space. For simplicity, A can be imagined as a finite set, countably infinite set, R^n, or its natural subset. All sets $A(x)$ are nonempty Borel subsets of A. In particular, if A is countable then $A(x)$ are arbitrary nonempty subsets of A.

If an action $a \in A(x)$ is selected at the state x then the one-step reward is $r(x, a)$ and the probability that the system will be at the state y at the next step is $p(y|x, a)$. The standard assumption is that the functions $r(x, a)$ and $p(y|x, a)$ are measurable in a for all $x, y \in X$.

Unless otherwise specified, we shall assume that the sum of transition probabilities is 1 and the reward function is bounded above. The former means that $\sum_{y \in X} p(y|x, a) = 1$ for all $x \in X$ and for all $a \in A(x)$. The latter means that there exists a finite constant K such that $r(x, a) \leq K$ for all $x \in X$ and for all $a \in A(x)$.

For the classical dynamic programming problems introduced by Blackwell [3], the reward functions $r(x, a)$ are assumed to be bounded, i.e., $|r(x, a)| \leq K$, $x \in X$ and $a \in A(x)$, for some finite constant K. However, in many operations research applications the reward functions are bounded above, i.e., $r(x, a) \leq K$ when $x \in X$ and $a \in A(x)$. For example, in mathematical models of inventory and queueing systems, the one-step holding costs can tend to ∞ as the inventory levels or number of waiting customers increases to ∞. Therefore, the corresponding reward functions can be unlimited from below. So, we consider the more general case when the reward function is bounded from above.

A general policy may be randomized and history-dependent. Let Π be the set of all policies. A deterministic policy is defined as a function ϕ that always selects action $\phi(x)$ at a state $x \in X$. In other words, a deterministic policy is history-independent and nonrandomized. Let $\mathcal{D}$ denote the set of deterministic policies.

Let the initial state be x and a policy π be chosen. For a positive constant $\alpha < 1$ called a discount factor, the expected total discounted reward is

$$v^\pi(x) = E_x^\pi \sum_{n=0}^{\infty} \alpha^n r(x_n, a_n),$$

where E_x^π is the expectation when the initial state is x and a policy π is chosen, and x_t and a_t are states and selected actions at epochs $t = 0, 1, \ldots$. The value at each state x is defined as $V(x) = \sup_{\pi \in \Pi} v^\pi(x)$.

A policy π is called optimal if $v^\pi(x) = V(x)$ for all $x \in X$. Thus, the optimality is defined with respect to all possible initial states. If an optimal policy is deterministic, it is called a deterministic optimal policy.

A policy π is called ϵ-optimal for a nonnegative constant ϵ if $v^\pi(x) \geq V(x) - \epsilon$ for all initial states x from X. In particular, the notions of 0-optimal and optimal policies coincide.

For $x \in X$, $a \in A(x)$, and for a bounded above real-valued function f on X, define

$$T^a f(x) = r(x, a) + \alpha \sum_{y \in X} p(y|x, a) f(y).$$

This value can be interpreted as the expected reward over two steps starting from the state x, when the action a from $A(x)$ is selected and the reward at the second step is defined by the function f. For a deterministic policy ϕ we can consider the operator T^ϕ defined for functions f bounded above,

$$T^\phi f(x) = T^{\phi(x)} f(x), \qquad x \in X.$$

Then

$$v^\phi(x) = T^\phi v^\phi(x), \qquad x \in X. \tag{1}$$

The optimality operator T is defined as

$$Tf(x) = \sup_{a \in A(x)} T^a f(x), \qquad x \in X.$$

The value function V satisfies the *optimality equation*

$$U(x) = TU(x), \qquad x \in X. \tag{2}$$

In particular, if the reward function r is bounded then T^ϕ and T are contraction mappings defined on the set of bounded functions on X endowed with the metric d induced by supremum norms. This is true for an arbitrary, possibly uncountable, set X. The supremum norm $||v||$ of a bounded function v on X is $||v|| = \sup_{x \in X} |v(x)|$. The metric d is defined as $d(v,u) = ||u-v|| = \sup_{x \in X} |v(x)-u(x)|$. The contraction mapping property in our case means that $d(T^\phi v, T^\phi u) \le \alpha d(v,u)$ and $d(Tv,Tu) \le \alpha d(v,u)$.

Therefore, when the reward function r is bounded, the Banach fixed point theorem (also known as the contraction mapping theorem or contraction mapping principle) implies that the functions v^π and V are the unique bounded solutions of the equations $u = T^\phi u$ and $U = TU$ respectively. The Banach fixed point theorem also provides the convergence of the value iteration algorithm and it provides estimates for its convergence; see Bertsekas [1, Chapter 1] on such estimates and Feinberg [8] on convergence of value iterations for total-reward criteria. We notice that in addition to a unique bounded solution, each of these equations may have unbounded solutions; see Example 6.4 in Feinberg [8].

The analysis of discounted problems with reward functions bounded above can be reduced to the analysis of a negative dynamic programming problem by replacing the reward function r with the nonpositive reward function $\tilde{r} = r - K$. Then the expected total reward $\tilde{v}^\pi$ for the new problem is $\tilde{v}^\pi(x) = v^\pi(x) - K/(1-\alpha)$ for all $x \in X$ and for all $\pi \in \Pi$. Therefore, the existence of optimal and ϵ-optimal policies for discounted problems can be obtained from the corresponding results for negative programming. For example, for negative programming the value function is the largest solution of (2) with $\alpha \in [0,1]$. Therefore, if the reward function r is bounded above for a discounted dynamic programming problem, then the value function V is the largest solution of equation (2) satisfying the inequality $U(x) \le K/(1-\alpha)$ for all $x \in X$. However, stronger results sometimes hold for discounted problems than for negative problems. The optimality equation and its properties are summarized in the following theorem.

Theorem 1 *The value function V is the largest solution of the optimality equation (2) satisfying the inequality $U(x) \le K/(1-\alpha)$ for all $x \in X$. If the reward function r is bounded, then V is the unique bounded solution of the optimality equation (2).*

In the similar way, discounted MDPs with reward functions bounded below can be reduced to positive dynamic programming. Such reward functions are used in some economics applications.

2.2 Existence of optimal policies

A deterministic policy ϕ is called conserving if $T^{\phi}V(x) = V(x)$ for all $x \in X$.

Theorem 2 *A deterministic policy ϕ is optimal if and only if it is conserving.*

Therefore, the existence of a deterministic optimal policy and the existence of a conserving policy are equivalent statements. Finding a deterministic optimal policy is equivalent to finding a conserving policy. Consider the sets of conserving actions

$$A_c(x) = \{a \in A(x) | T^a V(x) = V(x)\}, \qquad x \in X.$$

Theorem 2 implies that a deterministic optimal policy exists if and only if $A_c(x) \neq \emptyset$. In addition, the optimality equation implies that $v^{\pi}(x) < V(x)$ for any policy π if $A_c(x) = \emptyset$. Therefore, Theorem 2 implies the following corollary.

Corollary 3 *An optimal policy exists if and only if for each $x \in X$ the set of conserving actions $A_c(x)$ is not empty. In addition, if an optimal policy exists then there exists a deterministic optimal policy.*

Thus, it is sufficient to verify $A_c(x) \neq \emptyset$ for all $x \in X$ to prove the existence of optimal deterministic policies. In particular, since $V = TV$, optimal policies exist if all the sets $A(x)$ are finite. However, they also exist under more general assumptions.

Definition 4 A real-valued function f defined on a topological space Z is called sup-compact if the set $Z^{\lambda} = \{z \in Z | \, f(z) \geq \lambda\}$ is compact for any real number λ.

The following assumption is sufficient for $A_c(x) \neq \emptyset$ for all $x \in X$.

Assumption 5 (a) For each $x, y \in X$ the function $p(y|x,a)$ is continuous in $a \in A(x)$.

(b) For each $x \in X$ the function $r(x,a)$ is sup-compact in $a \in A(x)$. This means that for any $x \in X$ and for any number λ the set $A^{\lambda}(x) = \{a \in A(x) | \, r(x,a) \geq \lambda\}$ is compact.

Assumption 5 implies that all the sets of conserving actions $A_c(x)$ are nonempty and therefore the following statement holds.

Corollary 6 *If an MDP satisfies Assumption 5 then there exists a deterministic optimal policy.*

The following assumption is stronger than Assumption 5.

Assumption 7 (a) Assumption 5(a) holds.

(b) The set $A(x)$ is compact for each $x \in X$.

(c) For each $x \in X$ the reward function $r(x,a)$ is upper semi-continuous in $a \in A(x)$.

Corollary 8 *If an MDP satisfies Assumption 7 then there exists a deterministic optimal policy.*

In particular, Assumption 7 holds when all the sets $A(x)$ are finite. Therefore, the mentioned above fact on the existence of optimal policies for finite action sets can be formulated as a particular case of Corollary 8.

Corollary 9 *If for each state $x \in X$ the set of available actions $A(x)$ is finite then there exists a deterministic optimal policy.*

According to the above statements, the existence of optimal policies requires certain continuity and compactness conditions to ensure the existence of conserving actions for all states. The existence of deterministic ϵ-optimal policies, where ϵ is an arbitrary fixed positive number, does not require such conditions.

Let ϵ be a positive constant. Consider a deterministic policy ϕ such that $T^{\phi}(x)V(x) \geq TV(x) - (1-\alpha)\epsilon$ for all $x \in X$. Then the optimality equation $V = TV$ and straightforward calculations imply that $v^{\pi}(x) \geq V(x) - \epsilon$ for all $x \in X$. Therefore, the following result holds.

Theorem 10 *For any $\epsilon > 0$ there exists a deterministic ϵ-optimal policy.*

3 Borel state problems

A topological space $(E, \mathcal{T})$ is called Polish if it is separable and completely metrizable. We recall that a topological space E is called completely metrizable if it admits a compatible metric d such that (X, d) is a complete metric space. A Borel σ-field on a topological space $(E, \mathcal{T})$ is defined as the minimal σ-field on E containing all the open subsets from E. A subset of B is called Borel if it belongs to the Borel σ-field on B.

A measurable space $(B, \mathcal{B})$ is called Borel (or standard Borel) if there is a one-to-one measurable mapping $f : B \to [0, 1]$ such that $f(B)$ is a Borel subset of the interval $[0, 1]$. Any Polish space B is a Borel space with $\mathcal{B}$ being the Borel σ-field on B. In addition, any Borel space and, therefore, any Polish space is either countable or continuum [7, Appendix 1] or [2, Corollary 7.16.1]. Countable sets include finite sets. If a Borel space B is continuum, there is a measurable one-to-one mapping of B on the interval $[0, 1]$. If two Borel spaces B_1 and B_2 have the same cardinality, there is a measurable one-to-one mapping of B_1 on B_2 such that the mapping f^{-1} is measurable too.

Any countable or continuum set is Polish in the appropriate topology. If a set is countable, it is Polish if the discrete topology on it is chosen. If a set B is continuum, we can choose any one-to-one correspondence $f : B \to [0, 1]$ of B on the interval $[0, 1]$, and define the topology on B whose basis consists of the pro-images $f^{-1}(C)$ of all open intervals $C = (y, z) \subset [0, 1]$. In particular, if B is a Borel space, then f can be selected measurable. Therefore, any Borel space is a Polish space in the described topology.

Let $(B, \mathcal{B})$ be a Borel space and $\mathcal{P}(B)$ be the set of all probability measures on $(B, \mathcal{B})$). For any probability measure p on $(B, \mathcal{B})$, consider the p-completion $\mathcal{B}_p$ of $\mathcal{B}$, that is, $\mathcal{B}_p$ is the minimal σ-field containing $\mathcal{B}$ and any subset Y of B such that $Y \subset C$

for some $C \in \mathcal{B}$ with $p(C) = 0$. The σ-field $\mathcal{U} = \cap_{p \in \mathcal{P}(B)} \mathcal{B}_p$ is called the universal σ-field. A subset C of B is called universally measurable if $C \in \mathcal{U}$. A function $f : B \to Y$, where B and Y are Borel spaces, is called universally measurable, if $f^{-1}(C)$ is universally measurable for any Borel set $C \subseteq Y$. If f is a universally measurable function and p is a probability measure on $(B, \mathcal{B})$ then there exists a Borel function f_p such that $P(\{z : f(z) \neq f_p(z)\}) = 0$. Thus, universally measurable functions $f : B \to R$ can be integrated with respect any probability measure on $(B, \mathcal{B})$ in the same way as Borel measurable functions.

A Borel state MDP is defined by the same objects as the MDP with a countable state space. The differences are that: (i) the state space X is a Borel space, (ii) the transition probability $p(E|x, a)$ is a probability measure on the Borel σ-field of X and it satisfies the condition that for any Borel subset Y of X the function $p(Y|x, a)$ is measurable in $(x, a) \in X \times A$.

A general policy is defined by transition probabilities π_n, $n = 0, 1, \ldots$ such that for any $h_n = x_0, a_0, x_1, a_1, \ldots, a_n$, $\pi_n(da_n|h_n)$ is a probability measure on A satisfying the following two conditions: (a) $\pi_n(A(x_n)|h_n) = 1$ and (b) for each Borel subset B of A the function $\pi_n(B|h_n)$ is Borel-measurable on the set of all histories H_n up to time n, $H_n = X \times (A \times X)^n$. Thus, a general policy can be randomized and history-dependent.

The following assumption is standard for Borel state MDPs and is always assumed in this article: the graph

$$\mathrm{Gr}(A) = \{(x, a) |\, x \in X, a \in A(x)\}$$

is a Borel subset of $X \times A$. A deterministic policy is a measurable mapping of X to A such that $\phi(x) \in A(x)$ for all $x \in X$. In other words, $(x, \phi(x) \in \mathrm{Gr}(A)$ for all $x \in X$. In general, for some $D \subset X \times A$, a mapping $f : X \to A$ is called a selector if $(x, f(x)) \in D$ for all $x \in X$. So, a deterministic policy is a measurable selector from X to A with respect to $D = \mathrm{Gr}(A)$. So, a deterministic policy is sometimes called a measurable selector.

In order to define at least one policy and avoid the possibility that $\Pi = \emptyset$, we need to assume that there exists at least one deterministic policy. We shall avoid this assumption by using the convention that $\sup\{\emptyset\} = -\infty$. Then $V(x) = -\infty$ for all $x \in X$, if $\Pi = \emptyset$. In many cases, for example, under Assumptions S, Su, W, and Wu and for universally measurable policies described below, $\Pi \neq \emptyset$, because the existence of a deterministic policy follows from so-called measurable selection theorems. In some cases it is possible to avoid this assumption by setting $V = -\infty$ if $\Pi = \emptyset$.

A special feature of the above model is that the value function V may not be Borel measurable. However, it belongs to a broader class of universally measurable functions, for which integration is possible. In the same way we had for the countable state set, the value function V satisfies the optimality equation. In the case of a bounded reward function r, the value function is a unique bounded universally measurable function. This is true if either the existence of at least one deterministic policy is assumed or policies are allowed to be selected from a broader class of universally measurable policies.

For Borel-state models, the statements similar to Theorems 1 and 2 hold.

Theorem 11 *(i) The value function V is universally measurable and it is a solution of the optimality equation (2).*

(ii) The value function V is the largest universally measurable solution U of the optimal equation (2) such that $U(x) \le K/(1-\alpha)$ for all $x \in X$.

(iii) If the reward function r is bounded, then V is the unique bounded solution of the optimality equation (2).

Theorem 12 *A deterministic policy is optimal if and only if it is conserving.*

In addition, the following statement holds.

Theorem 13 *If there exists an optimal policy then there exists a deterministic optimal policy.*

To ensure the existence of conserving policies, some continuity and compactness properties are required. Before we describe particular cases, we formulate a general statement.

For a policy π let v_N^π denote the N-horizon expected discounted total rewards. We have $v_0^\pi(x) = 0$, $x \in X$, and for $N = 1, 2, \dots$

$$v_N^\pi(x) = E_x^\pi \sum_{n=0}^{N-1} \alpha^n r(x_n, a_n), \qquad x \in X.$$

Let $V_N(x) = \sup_{\pi \in \Pi} v_N^\pi(x)$. According to these definitions, $V_0(x) = 0$ for all $x \in X$. In the trivial case, when $\Pi = \emptyset$ we have $V_N(x) = -\infty$, for all $x \in X$ and for all $N = 1, 2, \dots$. The values $V_N(x)$ is the supremum of the expected total reward over the finite horizon N with the 0 terminal value V_0. As was shown by Blackwell [3], the functions V_N, $N = 1, 2 \dots$, and V belong to the class of universally measurable functions and therefore they can be integrated in the same way as Borel functions. In addition, the function V_N satisfies the optimality equation $V_{N+1} = TV_N$, $N = 0, 1, \dots$. This is the well-known optimality equation for finite-horizon dynamic programming models. It can be proved by inductions or as a corollary from Theorem 11(i), because a finite-horizon model can be reduced to an infinite-horizon model; see [8, Section 4.6] The following theorem provides a sufficient condition for the existence of optimal deterministic policies.

Theorem 14 *Suppose that there exists a nonnegative integer k such that for any $x \in X$, for any finite number λ, and for any $N \ge k$ the set*

$$A_N^\lambda(x) = \{a \in A(x) |\, r(x,a) + \alpha \int V_N(y) p(dy|x,a) \ge \lambda\}$$

is a compact subset of A. Then $V(x) = \lim_{N\to\infty} V_N(x)$ for all in$X$, and there exists a deterministic optimal policy.

Theorem 14 is a useful tool to establish the existence of deterministic optimal policies under an assumption similar to Assumption 5. When X is continuum, there are two groups of assumptions in the literature corresponding to Assumption 5. These

assumptions correspond to two different types of convergence of probability measures: setwise convergence and weak convergence. Such assumptions, the so-called S and W were introduced by Schäl [17, 18] for the case of compact action spaces. Since we start with possibly non-compact action sets, we shall use abbreviations Su and Wu, where the symbol "u" indicate that the sets $A(x)$ can be unbounded and therefore non-compact. In many applications, noncompact action sets are unbounded.

Assumption Su (i) For each $x \in X$ the transition probability $p(dy|x,a)$ is setwise continuous in $a \in A(x)$. This means that $p(Y|x,a_n) \to p(Y|x,a)$ for every $x \in X$ and for any sequence $a_n, n = 1,2,\dots,$ of elements of $A(x)$ converging to a, where Y is an arbitrary measurable subset of X.

(ii) For each $x \in X$, the reward function $r(x,a)$ is sup-compact in a, i.e., the set $A^\lambda(x) = \{a \in A(x) |\, r(x,a) \ge \lambda\}$ is compact for any real number λ.

Assumption Wu (i) The transition probability $p(dy|x,a)$ is weakly continuous in $(x,a) \in X \times A$, i.e., for any bounded continuous function f on X

$$\int f(y)p(dy|x_i,a_i) \to \int f(y)p(dy|x,a).$$

for any $x \in X$ and any $a \in A(x)$, if $(x_i,a_i) \to (x,a)$ and $a_i \in A(x_i)$.

(ii) The reward function $r(x,a)$ is sup-compact on $\mathrm{Gr}(A)$, i.e., the set $\{(x,a) \in \mathrm{Gr}(A) |\, r(x,a) \ge \lambda\}$ is compact for any finite number λ.

Theorem 15 *Either Assumption Su or Assumption Wu implies the existence of a deterministic optimal policy. In addition, the value function V is Borel measurable under Assumption Su and sup-compact under Assumption Wu.*

The following assumptions are sufficient for the existence of deterministic optimal policies when all the sets of available actions $A(x)$ are compact.

Assumption S. (i) The sets $A(x)$ are compact for all $x \in X$.

(ii) The transition probability $p(\cdot|x,a)$ is setwise continuous in $a \in A(x)$, i.e., Assumption Su(i) holds.

(iii) The reward function $r(x,a)$ is upper semi-continuous in $a \in A(x)$ for all $x \in X$.

Assumption W. (i) The sets $A(x)$ are compact for all $x \in X$.

(ii) The set-valued mapping $A(x)$ is upper semi-continuous; i.e. for any open subset G of A, the set $\{x \in X |\, A(x) \subseteq G\}$ is open in X.

(iii) The transition probability $p(\cdot|x,a)$ is weakly continuous on $(X \times A)$, i.e., Assumption Wu(i) holds.

(iv) The reward function $r(x,a)$ is upper semi-continuous on $\mathrm{Gr}(A)$.

Theorem 16 *Under each Assumption S or W there exists a deterministic optimal policy. In addition, the value function V is Borel measurable under Assumption S and upper semi-continuous under Assumption W.*

The following Corollary follows from Theorem 16 under Assumption S.

Corollary 17 *If each set $A(x)$, $x \in X$, is finite then there exists a deterministic optimal policy.*

Now we discuss the existence of ϵ-optimal policies. For expected discounted reward MDPs with countable state spaces, deterministic ϵ-optimal policies always exist; see Theorem 10. For a Borel state space, this is true if all the sets of available actions $A(x)$ are countable. In this case, there always exists a deterministic policy, $V(x)$ is a Borel measurable function, and the following theorem holds.

Theorem 18 *If all the action sets $A(x)$, $x \in X$, are countable then for any $\epsilon > 0$ there exists a deterministic ϵ-optimal policy.*

In a general situation, the value function $V(x)$ may not be Borel measurable, but it is universally measurable; see [2, 3, 7] for detail. Since for any fixed policy π the function $v^\pi(x)$ is Borel, ϵ-optimal policies may not exist; Blackwell [3, Example 2]. However, such policies exist if either the definition of ϵ-optimal is changed to the definition of so-called (p, ϵ)-optimal policies or the set of policies is expanded by allowing the policies to be universally measurable.

Let p be a probability measure on X. A policy π is called (p, ϵ)-optimal if there exists a measurable subset Y of X such that $p(Y) = 1$ and $v^\pi(x) \geq V(x)$ for all $x \in Y$.

Theorem 19 *For any probability measure p on X and for any $\epsilon > 0$ there exists a deterministic (p, ϵ)-optimal policy.*

However, for any $\epsilon > 0$ there exists a deterministic ϵ-optimal policy, if the definition of a policy is expanded by allowing transition probabilities $\pi_n(B|h_n)$ to be universally measurable functions on X for all Borel measurable subsets B of A. A deterministic universally measurable policy ϕ is a universally measurable mapping from X to A such that $\phi(x) \in A(x)$ for all $x \in X$. The Jankov-von Neumann selection theorem implies that there exists at least one universally measurable deterministic policy; see Bertsekas and Shreve [2], Dynkin and Yushkevich [7], or Kechris [15] for detail. We also remark that the value function V remains unchanged after the set of Π is extended by allowing universally measurable policies, except the trivial case when there is no Borel measurable selector from X to A.

Theorem 20 *If universally measurable policies are allowed then for any $\epsilon > 0$ there exists a deterministic universally measurable ϵ-optimal policy.*

4 Bibliographic Notes

The literature on discounted MDPs is vast and we mention only a few references. Shapley [19] studied stochastic games with discounted payoffs. Dubins and Savage [6] and then Hordijk [14] studied the relations between conserving and optimal policies. In particular, a deterministic policy for an MDP with the expected total criterion is optimal if and only if it is conserving and equalizing. However, any policy is equalizing for a discounted MDP with a reward function $r(x, a)$ bounded above. So, Theorem 2 can be traced to Dubins and Savage [6]. Theorem 12 is a straightforward extension of Theorem 2 to problems with Borel state spaces.

Blackwell [3] studied a discounted Borel-state MDP with bounded rewards. Theorems 13, 18, and 19 are from [3]. Strauch [20] proved the universal optimality of the value function and optimality equation; see Theorem 11. Blackwell [3] and Strauch [20] considered a model when $A(x) = A$ for all $x \in X$. Dynkin and Yushkevich [6] extended the theory in several directions including the state-dependent action sets $A(x)$. Denardo [5] studied contraction properties of dynamic programming operators. In particular, these properties lead to Theorem 10. This theorem can be viewed as a particular case of Theorem 6.21 from Feinberg [8] which describes the structure of nearly-optimal policies in countable MDPs with expected total rewards; see also [9].

Theorem 14 is Proposition 9.17 from Bertsekas and Shreve [2]. Schäl [17, 18] introduced Assumptions S and W, where S stands for setwise continuity of transition probabilities and W stands for weak continuity. Theorem 16 is from Schäl [18]. The monograph [13] by Hernández-Lerma and Lasserre primarily covers MDPs with setwise continuous transition probabilities. Assumption Su is from Hernández-Lerma [12] and Assumption Wu is from Feinberg and Lewis [10]. Theorem 15 presents the results from [12, 10]. The answer to the question on which type of continuity is more appropriate depends on a particular application. For example, in [10] it was observed that Assumption Wu is applicable to inventory control problems with continuous demand distributions, while Assumption Su is applicable to inventory control problems with general demand distributions.

Blackwell, Freedman and Orkin [4] and Freedman [11] investigated discounted MDPs with analytically measurable policies. Bertsekas and Shreve [2] developed the theory of dynamic programming on Borel state spaces and universally measurable policies. Theorem 20 is from [2].

The theory of Borel spaces and measurable selection theorems plays an important role in studying dynamic programming problems with Borel state spaces. In addition to excellent chapters and appendices in Bertsekas and Shreve [2], Dynkin and Yushkevich [7], and Hernández-Lerma and Lasserre [13], the monograph by Kechris [15] contains important facts on these topics.

Acknowledgement. This author acknowledges support by NSF grant DMI-0600538. The author thanks Peng Dai, Jun Fei, Mark E. Lewis, and Xiaoxuan Zhang for reading a preliminary version of this article and providing their comments.

References

[1] D.P. Bertsekas, *Dynamic Programming and Optimal Control*, Volume 2, Second edition, Athena Scientific, Belmont, 2001.

[2] D.P. Bertsekas and S.E. Shreve, *Stochastic Optimal Control: The Discrete-Time Case*, Athena Scientific, Belmont, MA, 1996 (reprinted from the 1978 Acedemic Press edition).

[3] D. Blackwell, Discounted dynamic programming, *Ann. Math. Statist.,* **36**(1965), 226-235.

[4] D. Blackwell, D. Freedman, and M. Orkin, The optimal reward operator in dynamic programming, *Ann. Probability,* **2**(1974), 926-941.

[5] E.V. Denardo, Contraction mappings in the theory of underlying dynamic programming, *SIAM Rev.*, **9**(1967), 165-177.

[6] L. Dubins and L. Savage, *Inequalities for Stochastic Processes (How to Gamble if You Must*, Dover, New York, 1976 (reprinted from the 1965 McGraw-Hill edition).

[7] E.B. Dynkin and A.A. Yushkevich, *Controlled Markov Processes*, Springer-Verlag, New York, 1979 (translated from the Russian 1975 edition).

[8] E.A. Feinberg, Total reward criteria, in: E.A. Feinberg and A. Shwartz (Eds.), Handbook of Markov Decision Processes, Kluwer, Boston, 2002, pp. 173–207.

[9] E.A. Feinberg, Sufficient classes of strategies in discrete dynamic programming II: Locally Stationary Strategies. *SIAM Theory Prob. Appl.* **32**(1987), 478-493.

[10] E.A. Feinberg and M.E. Lewis, Optimality inequalities for average cost Markov decision processes and the stochastic cash balance problem, *Mathematics of Operations Research,* **32**(2007), 769-783.

[11] D. Freedman, The optimal reward operator in special classes of in dynamic programming problems, *Ann. Probability,* **2**(1974), 942-949.

[12] O. Hernández-Lerma, Average optimality in dynamic programming on Borel spaces, *Systems and Control Letters,* **17**(1991), 237–242.

[13] O. Hernández-Lerma and J.B. Lasserre. *Discrete-Time Markov Control Processes. Basic Optimality Criteria.* Springer, New York, 1996.

[14] A. Hordijk, *Dynamic Programming and Markov Potential Theory.* Second edition. Mathematical Centre Tracts **51**, Mathematisch Centrum, Amsterdam, The Netherlands, 1977.

[15] A.S. Kechris, *Classical Descriptive Set Theory*, Springer-Verlag, New York, 1994.

[16] M.L. Puterman, Markov Decision Processes, Wiley, New York, 1994.

[17] M. Schäl, On dynamic programming: compactness of the space of policies. *Stochastic Processes and their Appliations* **3**(1975), 345-364.

[18] M. Schäl, Conditions for optimality in dynamic programming and for the limit of n-stage optimal policies to be optimal, *Z. Wahrsch. verve. Gebiete* **32**(1975), 179-196.

[19] L.S. Shapley, Stochastic games, *Proc. Natl. Acad. Sci. USA* **39**(1953), 1095-1100.

[20] R. Strauch, Negative dynamic programming, *Ann. Math. Statist.*, **37**(1966), 871-880.