A chapter for
MARKOV DECISION PROCESSES
Models, Methods, Directions, and Open Problems
written by

Eugene A. Feinberg
Department of Applied Mathematics and Statistics
SUNY at Stony Brook
Stony Brook, 11794-3600, NY, USA

May 28, 2000

Chapter 1

Total reward criteria

ch:MC

Abstract

This chapter deals with total reward criteria. One-step rewards can be as positive as negative. We describe results on the existence of optimal and nearly optimal strategies and the results on the convergence of value iteration algorithms under the so-called General Convergence Condition. This condition requires that, for any initial state and for any policy, the sum of positive parts of reward converges to a finite number.

1.1 Introduction

s:TC-intro

This chapter deals with the total reward criterion. This criterion is natural for finite horizon problems and for infinite horizon problems in which the number of steps is not fixed but it is finite along most trajectories. The examples include discounted criteria, which can be interpreted as problems with geometrically distributed horizon, as well as other problems such as sequential statistical procedures, stopping, search, and optimal selection problems.

The analysis of finite horizon models is usually based on the analysis of optimality equations and optimality operators. The value function satisfies the optimality equation and a Markov policy at which these equations are achieved at each step is optimal. If optimality cannot be achieved, a Markov policy, which is (ε/N)-close to the optimal value, is ε-optimal for the N-horizon problem.

For infinite horizon problems, we can distinguish two distinct approaches. The first approach deals with the analysis of optimality (also called, dynamic programming) equations and operators. The second approach studies probability distributions on the sets of trajectories. In fact, the most interesting results have been achieved by combining these two approaches.

As was observed in sixties, certain properties of optimality operators, namely contracting and motonicity properties, imply the existence of optimal or nearly optimal strategies within natural classes of strategies. These properties also imply the convergence of algorithms. In general, dynamic programming operators may not possess these properties. However, if the one-step reward function is uniformly bounded and the nonnegative discount factor is less then one, then the contracting property holds. If rewards are nonnegative (nonpositive) then the value iteration algorithm, applied to the zero terminal value, forms a monotone and therefore convergent sequence. The mentioned three models are called discounted, positive, and negative respectively.

The first comprehensive results were obtained for these three models. (Blackwell bl65 [7] Denardo de67 [17], Strauch st66 [53]) studied discounted models, (Blackwell bl67 [8], Ornstein or [46], and Strauch st66 [53]) studied positive models, and Strauch st66 [53] studied negative models. The results differ significantly from one model to another. To illustrate these differences, let us consider the situation when the state space is countable and there are no additional assumptions such as compactness of action sets. For discounted and positive models, for each initial state the supremum of the expected total rewards over the class of all strategies is equal to the corresponding supremum over the class of all stationary policies; Blackwell [7, 8]. However, it is not true for negative models; Strauch [53]. For discounted models, for any positive constant ε, there exist stationary ε-optimal policies. Such policies may not exist for positive models; Blackwell [8]. However, for positive models there exist stationary εV-optimal policies also called multiplicatively ε-optimal; Ornstein [46]. There are other significant differences between positive, negative, and discounted programming. The differences are so significant than for a long period of time it was even was not clear how to formulate unified results. As the result, all textbooks on dynamic programming and Markov decision processes, that deal with infinite-horizon models with total rewards, consider only positive, negative, and discounted models and deal with them separately; see e.g. Ross [49] and Puterman [48].

In addition to positive, negative, and discounted models, problems with arbitrary reward functions and without discounting have been considered in the literature for a long period of time. It turned out that the most comprehensive results can be proved when so-called General Convergence Condition holds. This condition means that the positive part of the reward function satisfies the positive programming assumptions. Positive, negative, and discounted models are particular cases of models satisfying the General Convergence Condition.

Blackwell [6] and Krylov [41] proved the existence of stationary optimal policies for MDPs with finite state and action sets. Dubins and Savage [19] and Hordijk [39] introduced sufficient (and almost necessary) conditions for optimality, so-called conserving and equalizing conditions. Derman and Strauch [18], and Strauch [53] proved that, for a given initial state, any policy can be substituted with an equivalent randomized Markov policy. Krylov [41] and Gikhman and Skorohod [36] showed that nonrandomized strategies are as good as randomized strategies. Dynkin and Yushkevich [20] proved that if someone randomizes between different strategies, the objective function cannot be improved. Feinberg [21, 22] proved that, for a given initial state, (nonrandomized) Markov policies are as good as general ones; earlier van Hee [62] showed that the supremum of the expected total rewards over the class of Markov policies is equal to the supremum over the class of all policies.

Seminal papers by Blackwell [7, 8] and Strauch [53] dealt with models with Borel state and action spaces. Some of the following papers dealt just with countable state spaces. For some results, their extension from countable to uncountable models is a straightforward exercise. For other results, such extensions are either difficult or impossible. In addition, Blackwell and Strauch considered Borel-measurable strategies and discovered that ε-

optimal strategies may not exist but for any initial measure they exist almost everywhere. In order to establish the existence of everywhere ε-optimal policies, one should expand the set of policies to universally measurable or analytically measurable policies. Such extension was introduced by Blackwell, Freedman, and Orkin [10] and it was done in a systematic and comprehensive way in the book by Bersekas and Shreve [4]. In particular, this book expanded in a natural way almost all results on positive, negative and discounted programming in a way that ε-optimality was established instead of almost sure ε-optimality. The major exception was Ornstein's theorem [46]. It was proved by Ornstein [46] for a countable state space; see also Hordijk [39]. It's extension to Borel models in the sense of almost sure multiplicative nearly-optimality was formulated by Blackwell [8] as an open question. Frid [34] solved this problem (Schäl and Sudderth [52] found a correctable gap in Frid's proof). The natural conjecture is that under more general measurability assumptions, in the spirit of Bertsekas and Shreve [4], almost sure nearly-optimality can be replaced with nearly optimality everywhere in Ornstein's theorem. Blackwell and Ramachandran [11] constructed a counter-example to this conjecture.

For the General Convergence Condition, significantly deeper results are available for countable state models than for Borel state problems. First, three important particular results were discovered by van der Wal: (i) the supremum of the expected total rewards over all policies is equal to the supremum over stationary policies if the action sets are finite; [58, Theorem 2.22] (this result was generalized by Schäl [51] to Borel state models with compact action sets); (ii) extension of Ornstein's theorem to the case when some action sets are compact; [60] (this results was generalized by van Dawen and Scäl [57, 56]; (iii) existence of uniformly nearly-optimal Markov policies; [59]. The survey by van der Wal and Wessels [61] describes these and many preceding results.

Feinberg and Sonin [31] generalized Ornstein's [46] theorem to models satisfying the General Convergence Condition. For the long period of time, it had not been clear even how to formulate such results. The first clue is that in the more general formulation the value function of the class of stationary strategies should be considered instead of the value function of the class of all strategies. The second clue is that in the definition of multiplicative ε-optimal policies function V should be replaced with an excessive majorant of a value function of the class of stationary policies. The proofs in Feinberg and Sonin [31] are non-trivial and differ from Ornstein's [32] proofs. Feinberg and Sonin [32] extended these results to non-stationary policies and to more general classes of functions that approximate optimal values.

Feinbeg fe87, fe87a [24, 25] described the structure of uniformly nearly-optimal policies in countable state models satisfying the General Convergence Condition. As mentioned above, stationary policies can be significantly outperformed by nonstationary policies in negative problems; see Example ex6 3.6. It turned out that this example demonstrates the only pathological situation when the value of the class of stationary policies is less than the value of the class of all policies. Considers the set of states for which the value functions equal to zero and there are no conserving actions (an action at which the optimality operator applied to the value function achieves the maximum). Then there are policies which are uniformly nearly optimal and which are stationary outside of this set; see Theorems t:prel 19 and t:main2 20.

Another important feature of the papers by Feinberg and Sonin feso85, fe87, fe87a [32, 24, 25] is that they consider general classes of nonstationary policies and general methods how to deals with nonstationary policies. In particular, the information about the past plays an important role. It is possible to identify two properties of this information: (i) non-repeating and (ii) transitivity conditions. Non-repeating condition implies that randomized policies are as good as non-randomized ones and there exist uniformly nearly optimal policies within any class of policies that satisfies this condition. Transitivity condition implies that the model can be transformed into a new model in a way that the class of strategies satisfying this condition in the old model becomes the class of stationary strategies in the new one.

This paper is a survey of results and methods for models satisfying the General Convergence Condition. We present the theory for countable models and discuss open questions, most of which deal with uncountable state spaces. In order to illustrate major concepts and counter-examples, we start our presentation with classical discounted, positive, and negative problems.

1.2 Definitions of discounted, positive, negative, and general convergent models

s:MC-model

We say that an MDP is *discounted* if function r is bounded if there is a constant $\beta \in [0,1[$, called the discount factor, such that

$$w(x,\pi) = \mathbb{E}_x^\pi \sum_{t=0}^{\infty} \beta^t r(x_t,a_t). \tag{1.1}$$

Dis

An MDP is called unbounded discounted if (Dis 3.1) holds and the function r is bounded above, $r(x,a) \le C < \infty$ for all $x \in \mathbb{X}$, $a \in \mathbb{A}(x)$ and for some C.

Discounted and unbounded discounted MDPs can be reduced to an MDP with a discount factor equal to 1. In order to do it, we add an additional state to the state space X. This state has only one action under which it is absorbent and all rewards are equal to zero in this state. This state sometimes is called a grave. The one-step transition probabilities between states in X become equal to $\beta p_{xy}(a)$ and the transition probability from any state in X to the new absorbent states become $(1-\beta)$. The expected total rewards for all initial distributions on X in the new MDP are equal to the expected total discounted rewards for the same initial distribution in the original system. In the new MDP, the original constant β can be interpreted as the probability that the system remains alive at the next step if it is alive on the current step. This construction is well-known and it is described in details in Altman's book al99 [2, Section 10.1]. Thus we can consider discounted MDPs as a special case of total reward MDPs with the discount factor equal to 1.

An MDP is called *positive* if $r(x,a) \geq 0$ for all $x \in X$ and $a \in \mathbb{A}(x)$ and $w(x,\pi) < \infty$ for all $x \in X$ and for all $\pi \in \Pi^R$. An MDP is called *negative* if $r(x,a) \leq 0$ for all $x \in \mathbb{X}$ and for all $a \in \mathbb{A}(x)$.

We we indicate by **D**, **UD**, **P**, and **N** when we assume that the MDP is respectively discounted, unbounded discounted, positive, and negative. For an arbitrary MDP we define

$$w_+(x,\pi) = \mathbb{E}_x^\pi \sum_{t=0}^{\infty} r^+(x_t,a_t), \tag{1.2} \qquad \boxed{\texttt{PosCrit}}$$

$$w_-(x,\pi) = \mathbb{E}_x^\pi \sum_{t=0}^{\infty} r^-(x_t,a_t). \tag{1.3} \qquad \boxed{\texttt{NegCrit}}$$

Let

$$V_+(x) = \sup_{\phi\in\Pi^R} w_+(x,\phi) \qquad V_-(x) = \sup_{\phi\in\Pi^R} w_-(x,\phi)$$

It is well-known that $V_+(x) = \sup_{\phi\in\Pi^S} w_+(x,\phi)$; see Blackwell bl67 [8]. The following condition holds for **D**, **UD**, **P**, and **N** MDPs.

General Convergence Condition. $V_+(x) < \infty$ for all $x \in \mathbb{X}$.

The General Convergence Condition is equivalent to the condition that $w_+(x,\pi) < \infty$ for all $x \in \mathbb{X}$ and for all $\pi \in \Pi^R$; see van der Wal vdw81 [58, Theorem 2.3]. If the General Convergence Condition holds, $w(x,\pi) = w_+(x,\pi) + w_-(x,\pi)$ for all initial states and for all policies. In this paper we always assume the General Convergence Condition. The value of

$w(x,\pi)$ does not change if we re-group summands $r(x_t,a_t)$ or change the order of summation. We say that an MDP is General Convergent (**GC**) if the General Convergence Condition holds.

Let

$$s(x) = \sup_{\phi\in\Pi^S} w(x,\phi)$$

be the value of the class of stationary policies.

s:SMOF

1.3 Properties of strategic measures and objective functions

t:MC-Markov

Theorem 1 ([53, 18, 39]) *Let $\pi^1,\pi^2,\ldots$ be an arbitrary sequence of policies and $\lambda_1,\lambda_2,\ldots$ a sequence of nonnegative numbers summing to 1. Consider a randomized Markov policy π defined by*

$$\pi_t\left(C\mid y\right) \overset{def}{=} \frac{\sum_{i=1}^{\infty}\lambda_i\,\mathbb{P}_x^{\pi^i}\left(x_t=y,a_t\in C\right)}{\sum_{i=1}^{\infty}\lambda_i\,\mathbb{P}_x^{\pi^i}\left(x_t=y\right)}\ , \quad t\geq 0,\ y\in\mathbb{X}, \tag{1.4}$$

e:eqrandM

whenever the denominator in (3.4) is not equal to 0. Then, for all $t\geq 0$, $y\in\mathbb{X}$ and Measurable subsets C of $\mathbb{A}(y)$,

$$\mathbb{P}_x^{\pi}\left(x_t=y,a_t\in C\right) = \sum_{i=1}^{\infty}\lambda_i\,\mathbb{P}_x^{\pi^i}\left(x_t=y,a_t\in C\right)\ . \tag{1.5}$$

e:MC-MarkovP

Note that the randomized Markov policy defined in Theorem 1 depends on the initial state. Our first observation related to Theorem 1 is that setting $\lambda_1 = 1$ and $\lambda_i = 0$, $i>1$, we have that for any given policy and initial state, we can find a randomized Markov policy that produces the same one-dimensional distributions for the pair (x_t,a_t). Consequently, for any criterion depending only on such distributions Markov policies suffice. In particular, we have the following result.

c:randMv

Corollary 2 *Given an initial state x, for any policy σ consider a randomized Markov policy π defined in Theorem 1 for $\pi^1=\sigma$ and $\lambda_1=1$. Then $v(x,\pi)=v(x,\sigma)$.*

Our second observation is that we can expand the notion of a policy by allowing the decision maker to select policies randomly at epoch 0. For example, we can say that a sequence $\gamma = \{(\lambda_i,\pi^i)\}_{i=1}^{\infty}$, where λ_i are nonnegative numbers with the sum equal to 1 and π^i are policies is a mixed strategy. Then any couple (γ, μ), where γ is a mixed strategy and μ is the initial probability distribution on $\mathbb{X}$, define a strategic measure $\mathbb{P}^{\gamma}_{\mu} = \sum_{i=1}^{\infty} \lambda_i \mathbb{P}^{\pi^i}_{\mu}$. Theorem 1 shows that mixed policies do not outperform randomized Markov policies. Given a mixed policy γ, formula (6) in Section 1.3 in Dynkin and Yushkevich [20] defines a usual randomized policy σ such that $\mathbb{P}^{\sigma}_{x} = \mathbb{P}^{\gamma}_{x}$. This is even a stronger argument that there is no need to deal with mixed policies.

Thus, Theorem 3.5 implies that

$$V(x) = \sup_{\pi \in \Pi^{RM}} v(x, \pi), \qquad x \in \mathbb{X}. \tag{1.6}$$

Our next step is to show that nonrandomized Markov policies are as good randomized Markov ones. It follows from the fact that, for a given initial state or distribution, any strategic measure for a randomized Markov policies can be presented as a convex combination of strategic measures for nonrandomized Markov policies.

We recall that a Markov policy ϕ selects action $\phi_n(x)$ when the system is at state x on epoch n. Sometimes it is more convenient to write $\phi(x, n)$ instead $\phi_n(x)$. We shall use both these notations. We observe that the set of all Markov policies Π^M is the set of all functions from $\mathbb{X} \times \mathbb{N}$ to $\mathbb{A}$ such that $\phi(x, n) \in \mathbb{A}(x)$ for all x and n. For $B \in \mathcal{A}$ we denote by $\mathcal{A}(B)$ the σ-field on B which elements belong to $\mathcal{A}$.

Now we introduce a measurable structure on Π^M. For each couple (x, n) and for each $C \in \mathcal{A}(\mathbb{A}(x))$ we consider the cylinder $\Pi^{NR}_{x,n}(C) = \{\phi \in \Pi^{NR} : \phi(x, n) \in C\}$. Then we consider a σ-field of all cylinders with the base (x, n), $\mathcal{F}_{x,n} = \{\Pi^{NR}_{x,n}(C) : \ C \in \mathcal{A}(\mathbb{A}(x))\}$. For a set $E \subseteq \mathbb{X} \times \mathbb{N}$ we define the σ-fiels of cylindrical subsets $\mathcal{F}_E$ with the base at E, $\mathcal{F}_E = \cap_{(x,n) \in E} \mathcal{F}_{(x,n)}$. We denote $\mathcal{F} = \mathcal{F}_{\mathbb{X} \times \mathbb{N}}$.

We observe that $\phi \to \mathbb{P}^{\phi}_{x}(C)$ is a measurable function on $(\Pi^M, \mathcal{F})$ for any $C \in \mathcal{H}_{\infty}$ and for any given initial state x. The main idea of the proof is that we interpret the problem in the following way. The decision-maker selects a Markov policy first and then uses this policy. The initial state is always x. So, we add the point ϕ before each trajectory. A trajectory $x, a_0, x_1, a_1 \ldots$ transforms into $\phi, x, a_0, x_1, a_1 \ldots$. The set of trajectories H_{∞} transforms the set H_{∞} into the set $\Pi^M \times H_{\infty}$. Transition probabilities from a_n to x_{n+1}, given the history $\phi, x_0, a_0, \ldots, x_n, a_n$ are defined by $p(dx_{n+1} | x_n, a_n)$, $n = 0, 1, \ldots$. Transition probabilities from x_n to a_n, given the history $\phi, x_0, a_0, \ldots, x_n$ are defined by $\mathbf{I}\{\phi(x_n, n) = a_n\}$, $n = 0, 1, \ldots$. Since the measurability conditions from the Ionesco Tulcea [47, Proposition V.1.1] theorem hold, we have that the mapping $\phi \to \mathbb{P}_x^{\phi}(C)$ is measurable on $(\Pi^M, \mathcal{F})$ for any $C \in \mathcal{H}_\infty$. The arguments in this paragraph also imply that, for any fixed $x \in X$ and $C \in \mathcal{H}_n$, the function $\mathbb{P}_x^{\phi}(C)$ is $\mathcal{F}_{\mathbb{X}\times\{0,1,\dots,n-1\}}$-measurable; see Lemmas 3.1 and 3.2 in Feinberg [24] for details.

Consider a randomized Markov policy π. We also will use sometimes the notation $\pi(\cdot|x,n)$ instead of $\pi_n(\cdot|x)$. Consider the measure m^π on $(\Pi^M, \mathcal{F})$ defined as the product of measures $\pi(\cdot|x,n)$. This means that for any set of $B(x,n) \in \mathcal{A}(\mathbb{A}(x))$, where $x \in \mathbb{X}$ and $n \in \mathbb{N}\}$, we have

$$m^\pi\{\phi \in \Pi^M |\, \phi(x,n) \in B(x,n), (x,n) \in \mathbb{X}\times\mathbb{A}\} = \prod_{(x,n)\in\mathbb{X}\times\mathbb{N}} m^\pi(B(x,n)|x,n).$$

The measure m^π exists and unique for each Markov policy π. This follows from the general construction of product measures on products of measurable spaces; see e.g. Proposition V.1.2 in Neveu [47]. The following theorem is a special case of Theorem 3.1 in Feinberg [47].

t:MC-nonran **Theorem 3** *For any randomized Markov policy π and for any $x \in \mathbb{X}$*

$$\mathbb{P}_x^\pi(C) = \int\limits_{\Pi^M} \mathbb{P}_x^\phi(C) m^\pi(d\phi), \qquad C \in \mathcal{H}_\infty. \tag{1.7}$$ e:intrep

Proof. If (3.7) holds for all $C \in \mathcal{H}_n$, $n \in \mathbb{N}$, it holds for all $C \in \mathcal{H}_\infty$ because measures P_x^σ can be continued from $\cup_{n=0}^\infty \mathcal{H}_n$ to $\mathcal{H}_\infty$. Therefore, it is sufficient to prove (3.7) for $C \in \mathcal{H}_n$, $n \in \mathbb{N}$. We shall do it by induction.

We have that $P_x^\sigma\{x_0 = y\} = \mathbf{I}\{x = y\}$ for any $y \in X$ and for any policy σ. This implies that (3.7) holds for all $C \in \mathcal{H}_0$.

Let (3.7) holds for all $C \in \mathcal{H}_n$ for some $n = 0, 1 \dots$. We take an arbitrary $C \in \mathcal{H}_n$ and an arbitrary $B \in \mathcal{A}$.

We observe that the function $\mathbb{P}_x^\phi(C')$ is $\mathcal{F}_{\mathbb{X}\times\{0,1,\dots,n-1\}}$-measurable for all $C' \in \mathcal{H}_n$. The mapping $\phi(y,n)$ is $\mathcal{F}_{\mathbb{X}\times\{n\}}$-measurable. Since $(\mathbb{X}\times\{0,1,\dots,n-1\}) \cap (\mathbb{X}\times\{n\}) = \emptyset$, the σ-fields $\mathcal{F}_{\mathbb{X}\times\{0,1,\dots,n-1\}}$ and $\mathcal{F}_{\mathbb{X}\times\{n\}}$ are independent with respect to measure m^π. This implies that for any $D = \{x_0, a_0 \in B_0, x_1, \dots, x_{n-1}, a_{n-1} \in B_{n-1}, x_n\}$

$$\begin{gathered}\int\limits_{\Pi^M} \mathbf{I}\{\phi(x_n,n) \in B_n\} m^\pi(d\phi) \int\limits_{\Pi^M} \mathbb{P}_x^\phi(D) m^\pi(d\phi) = \\ \int\limits_{\Pi^M} \mathbf{I}\{\phi(x_n,n) \in B_n\}\, \mathbb{P}_x^\phi(D) m^\pi(d\phi)\end{gathered} \tag{1.8}$$ e:indst1

for all $n = 0, 1, \dots$ and for all $x_t \in \mathbb{X}$, $B_t \in \mathcal{A}_t$, $t = 0, \dots, n$.

We have that

$$\mathbb{P}_x^\pi\{C \times B\} = \int\limits_C \pi(B|x_n, n)\,\mathbb{P}_x^\pi(dh_n) =$$

$$\int\limits_C \int\limits_{\Pi^M} \mathbf{I}\{\phi(x_n, n) \in B\} m^\pi(d\phi) \int\limits_{\Pi^M} \mathbb{P}_x^\phi(dh_n) m^\pi(d\phi) = \qquad (1.9) \quad \boxed{\texttt{e:indst2}}$$

$$\int\limits_C \int\limits_{\Pi^M} \mathbf{I}\{\phi(x_n, n) \in B\}\,\mathbb{P}_x^\phi\{dh_n\} m^\pi(d\phi) = \int\limits_C \mathbb{P}_x^\phi(C \times B) m^\pi(d\phi),$$

where the first and the last equalities in (e:indst2 3.9) follow from the definition of measures P_x^σ, the second equality follows from the definition of the measure m^π and from the induction assumption, and the third equality in (e:indst2 3.9) follows from (e:indst1 3.8).

We have from (e:indst1 3.8) that (e:intrep 3.7) holds for any $C \in \mathcal{H}_n \times \mathcal{A}$. Consider arbitrary $C \in \mathcal{H}_n \times \mathcal{X}$ and $y \in \mathbb{X}$. We have

$$\mathbb{P}_x^\pi\{C \times \{y\}\} = \int\limits_C p(y|x_n, a_n)\,\mathbb{P}_x^\pi(dh_n) = \int\limits_C p(y|x_n, a_n) \int\limits_{\Pi^M} \mathbb{P}_x^\phi(dh_n) m^\pi(d\phi) =$$

$$\int\limits_{\Pi^M} m^\pi(d\phi) \int\limits_C p(y|x_n, a_n)\,\mathbb{P}_x^\phi(dh_n) = \int\limits_{\Pi^M} \mathbb{P}_x^\phi(C \times \{y\}\} m^\pi(d\phi),$$

$$(1.10) \quad \boxed{\texttt{e:indstep2}}$$

where the first and the last equalities in (e:indstep2 3.10) follow from the definition of measures P_x^σ, the second one follows from the validity of (e:intrep 3.7) for $C \in \mathcal{H}_n \times \mathcal{A}$ established in (e:indst2 3.9), and the third one follows from Fubini's theorem. ■

The General Convergence Condition and Theorem t:MC-nonran 2 yield the following result.

c:MC-nonran **Corollary 4** *For any randomized Markov policy π and for any initial state $x \in \mathbb{X}$*

$$v(x, \pi) = \int\limits_{\Pi^M} v(x, \phi) m^\pi(d\phi).$$

Corollaries c:randMv 3.1 and c:MC-nonran 3.2 imply the following statement.

c:nonrMv **Corollary 5** *Given an initial state x, for any policy σ there exists a Markov policy ϕ such that*

$$v(x, \phi) \geq v(x, \sigma).$$

The latter corollary yields the following result.

c:supM **Corollary 6** $v(x) = \sup\{v(x,\phi| \ \phi \in \Pi^M\}$ *for all* $x \in \mathbb{X}$.

A nonrandomized policy ϕ is called *semi-Markov* if $\phi(h_n) = \phi(x_0, x_n)$ for all $h_n = x_0, a_0, \dots, x_n$, $n = 1, 2, \dots$. Since the state space $\mathbb{X}$ is discrete, Corollary c:supM 3.4 implies the following statement.

c:semi-Mar **Corollary 7** *For any positive function* $\varepsilon(x)$ *on* $\mathbb{X}$ *there exists a semi-Markov policy* ϕ *such that* $v(x,\phi) \geq V(x) - \varepsilon(x)$ *for all* $x \in \mathbb{X}$.

t:OE **Theorem 8** (Optimality Equation) $V(x) = TV(x)$ *for all* $x \in \mathbb{X}$.

Proof. For a Markov policy $\phi = \{\phi_0, \phi_1, \dots\}$ we define a shifted policy $\phi^1 = \{\phi_1, \phi_2, \dots\}$. Then $v(x,\phi) = T^{\phi(x)} v(x, \phi^1)$. We fix an arbitrary $x \in \mathbb{X}$. The proof consists of two simple steps.

Step 1 ($V(x) \leq TV(x)$). Consider an arbitrary constant $\varepsilon > 0$. According to Corollary c:supM 3.4 there exists a Markov policy ϕ for which $v(x,\phi) \geq V(x) - \varepsilon$. We use obvious monotonicity properties of optimality operators and get $V(x) - \varepsilon \leq v(x,\phi) = T^{\phi(x)} v(x,\phi^1) \leq Tv(x,\phi^1) \leq TV(x)$. Since $\varepsilon > 0$ is arbitrary, step 1 is proved.

Step 2 ($V(x) \geq TV(x)$). Again, we consider an arbitrary positive constant ε. Corollary c:semi-Mar 3.5 implies that for any $\varepsilon > 0$ there exists an ε-optimal semi-Markov policy ϕ. For any $a \in \mathbb{A}(x)$ we consider a nonrandomized policy $\sigma[a,\phi]$ such that

$$\sigma[a,\phi](x_0, a_0, x_1, \dots, x_n) = \begin{cases} a, & \text{if } n = 0 \text{ and } x_0 = x; \\ \phi_0(x_1), & \text{if } n = 1; \\ \phi_{n-1}(x_1, x_n), & \text{if } n > 1. \end{cases}$$

This policy uses action a at epoch 0 in the initial state x and than it switches to the semi-Markov policy ϕ with the initial state x_1 and starting epoch $t = 1$. We have that

$$V(x) \geq V(x, \sigma[a,\phi]) = T^a v(x,\phi) \geq T^a(V - \varepsilon)(x) = T^a V(x) - \varepsilon.$$

and $V(x) \geq T^a V(x) - \varepsilon$ for all $a \in \mathbb{A}(x)$ and for all $\varepsilon > 0$. Step 2 is proved. ■

We recall that an action $a \in \mathbb{A}(x)$ is called conserving in state x if $V(x) = T^a V(x)$. A policy that uses only conserving actions in all states is called conserving. Obviously, a stationary optimal policy is conserving. The following simple example shows that a conserving stationary policy may not be optimal.

ex:cons **Example 9** Let $\mathbb{X} = \mathbb{A} = \{0,1\}$, $\mathbb{A}(0) = \mathbb{A}$, and $\mathbb{A}(1) = \{0\}$. State 1 is absorbing with one-step reward equal to 0. If action 0 is selected in state 0, the process remains in state 0 and the reward is not collected. Action 1 moves the process to state 1 and brings the one-step reward equal to 1. The formal definitions are $p(x|x,0) = p(1|0,1) = 1$, $r(x,0) = 0$, and $r(0,1) = 1$ where $x \in \mathbb{X}$. In this example there are two stationary policies ϕ_a, $a = 0,1$. Policy ϕ_a selects action a in state 0. Both ϕ_0 and ϕ_1 are conserving. However, $V(0,\phi_0) = 0$ and $V(0,\phi_1) = 1$. ■

A policy π is called equalizing at state x if $\limsup_{n\to\infty} \mathbb{E}_x^\pi V(x_n) \le 0$. A policy is called equalizing if it is equalizing all all states x. It is easy to see that if a policy is conserving and equalizing then it is optimal. An optimal policy π is conserving and the limit of $\mathbb{E}_x^\pi V(x_n)$ exists and equal to 0 for all states x in which $V(x) > -\infty$.

A natural question is when conserving policies exist. The Optimality Equation implies that a stationary conserving policy exists if and only if TV(x) is achieved for each $x \in \mathbb{X}$ at some action $a \in \mathbb{A}(x)$.

Consider the following conditions that are essential for the existence of conserving policies.

- (i) A is a metric space and $\mathcal{A}$ is its Borel σ-field;
- (ii) $\mathbb{A}(x)$ is compact for some $x \in \mathbb{X}$;
- (iii) $p(y|x,a)$ are continuous in a and $r(x,a)$ are upper semi-continuous in a.

l:max **Lemma 10** *If conditions (i-iii) hold for a given $x \in \mathbb{X}$ and f is a bounded above function on $\mathbb{X}$ then $T^a f(x) = Tf(x)$ for some $a \in \mathbb{A}(x)$.*

See Lemma 4.2(i) in fesh96 [29] for the proof. In order to guarantee the existence of conserving strategies, we consider the following assumption.

Compactness Assumption Condition (i) holds and Conditions (ii) and (iii) hold for all $x \in \mathbb{X}$.

Lemma l:max 3.1 and the Optimality Equation imply the following statement.

c:conv **Corollary 11** *If the value function $V(x)$ bounded above then the Compactness Assumption implies the existence of stationary conserving strategies.*

In some applications, action sets are not compacts but can be approximated by compacts in a way that it is expensive to select actions outside of compact subsets. For example, in some inventory models, the size of the order may not be limited but large orders are expensive. To cover this situation, we consider the following condition.

- (iv) For any positive number N there exist a compact subset $B_N(x)$ of $\mathbb{A}(x)$ such that $r(x,a) \le -N$ for $a \in \mathbb{A}(x) \setminus B_N(x)$.

l:maxgen

Lemma 12 *If conditions (i – iv) hold for a given $x \in \mathbb{X}$ and f is a bounded above function on $\mathbb{X}$ then $T^a f(x) = Tf(x)$ for some $a \in \mathbb{A}(x)$.*

Proof. If $Tf(x) = -\infty$ then $T^a f(x) = Tf(x)$ for all $a \in \mathbb{A}(x)$. Let $T^{a^*} f(x) > -\infty$ for some $a^* \in \mathbb{A}(x)$. Let $C \ge f(z)$ for all $z \in \mathbb{X}$. We set $N = C - T^{a^*} f(x) + 1$. Lemma 3.1 applied to $B_N(x)$ implies that $T^{a'} f(x) = \sup\{T^a f(x) : \ a \in B_N(x)\}$ for some $a' \in B_N(x)$. For any $a \in \mathbb{A}(x) \setminus B_N(x)$ we have that $T^a f(x) \le C - N = T^{a^*} f(x) - 1 < Tf(x)$. Thus $T^{a'} f(x) = Tf(x)$. ■

Pre-Compactness Assumption Condition (i) holds and Conditions (iv) and (iii) hold for all $x \in \mathbb{X}$.

The following statements strengthens Corollary 3.6.

c:conv1

Corollary 13 *If the value function $V(x)$ bounded from above then the Pre-Compactness Assumption implies the existence of stationary conserving strategies. In particular, the Pre-Compactness Assumption implies the existence of stationary conserving policies for* **D**, **GD**, *and* **N** *MDPs.*

ex2

Example 14 There are no conserving policies in positive MDPs satisfying the Compactness Assumption.

Let $\mathbb{X} = \{0, 1, \dots\} \cup \{g\}$, $\mathbb{A} = \mathbb{A}(0) = \{1/2, 1/3, \dots, 0\}$, and $\mathbb{A}(x) = \{0\}$ for $x \in \mathbb{X} \setminus \{0\}$. We set $p(g|x,0) = 0$ for $x \in \mathbb{X}$, $r(g,0) = 0$, and $r(x,0) = x-1$ for $x = 1, 2, \dots$. We also set $r(0,a) = 0$ for all $a \in \mathbb{A}(x)$ and $p(x|0,1/x) = 1/x$, $p(g|0,1/x) = (x-1)/x$. We have that $V(g) = 0$, $V(x) = x-1$ for $x \ge 1$, and $T^{1/x}V(0) = 1 - 1/x$. Since $T^0 V(0) = 0$, there is no conserving action at state 0. ■

In view of Corollary 3.3, we consider the question if for any randomized Markov policy π there exists a Markov policy ϕ such that $v(x,\phi) \ge v(x,\pi)$ for all $x \in \mathbb{X}$. This natural question was asked in Dynkin and Yushkevich [20, Section 4.7]. The following example illustrates that the answer to this question is negative even when π is randomized stationary.

ex3 **Example 15** (Feinberg and Sonin feso81 [30]) Let $\mathbb{X} = \{(0,0)\} \cup \{(i,j)\},\ i = 1,2,\dots,\ j = -i+1,\dots,0\}$, $A = \{c,s\}$, where c stands for continue and s stands for stop. We also have that $\mathbb{A}(i,j) = \{c\}$ when $j < 0$, $\mathbb{A}(0,0) = \{s\}$, and $\mathbb{A}(i,0) = \{c,s\}$ when $i > 0$. The state $(0,0)$ is a "grave." This means that $p((0,0)|(0,0),s) = 1$ and $r((0,0),s) = 0$. The system always moves with zero reward from a state (i,j) to $(i,j+1)$ when $j < 0$. In other words, $p((i,j+1)|(i,j),c) = 1$ and $r((i,j),c) = 0$ when $j < 0$. For $i > 0$ we set $p((i+1,0)|(i,0),c) = p((0,0)|(i,0),s) = 1$ and $r((i,0),c) = 2^{-(i+1)}$, $r((i,0),s) = 1 - 2^{-(i+1)}$.

We consider a randomized stationary policy π that selects actions c and c with equal probabilities, $\pi(s|(i,0),c) = \pi(s|(i,0),s) = 0.5$ for $i > 0$. Expected one-step rewards in states $(i,0)$ are equal to 0.5 when $i > 0$. If the initial state is $(i,0)$ with positive i, the income stream forms a geometric progression with the first term and ratio equal to 0.5. The sum of this progression is 1. From each state (i,j) with $j < 0$, the system moves to state $(i,j+1)$ with zero one-step rewards until it reaches state $(i,0)$. Therefore, $v(x,\pi) = 1$ for all $x \in \mathbb{X} \setminus \{(0,0)\}$.

Consider a Markov policy ϕ. If $\phi((i,0),i-1) = c$ for all $i \geq 1$ then $v((1,0),\pi) = 0.5 < v((1,0),\phi)$. Therefore, if $v(x,\phi) \geq v(x,\pi)$ for all $x \in \mathbb{X}$ then $\phi((i,0),i-1) = s$ for some $i \geq 1$. However, in this case $v((i,-i+1),\phi) = 1 - 2^{-(i+1)} < v((i,-i+1),\phi)$. ■

1.4 Non-homogeneous and finite-horizon models

s:NFM

Sometimes it is natural to consider models in which all parameters of the change over time. In this section, we consider models in which action sets, transition probabilities, and rewards depend also on time. For such models, called non-homogeneous, it is natural to expand the state space and consider the standard homogeneous model with the state space $\tilde{\mathbb{X}} = \mathbb{X} \times \mathbb{N}$. By doing this, we can expand all previous notations and results to non-homogeneous models. The major distinction is that in the new model the states are couples (x,n) instead of x. For example, values functions are $V(x,n)$.

A particular important example of non-homogeneous models is a finite-horizon model. For an n-horizon model, the state and action sets, reward and transition functions remain unchanged at epochs $t = 0,1,\dots,n-1$. At epoch n, the one-step reward is $V_0(x)$, where V_0 is a so-called final reward, and the system stops. The final reward is also known under several other names such as terminal reward or salvage value.

For each n-horizon model, it is natural to construct a non-homogeneous model, in which the model remains unchanged at steps $0, 1, \dots, n-1$. At step n the reward function is V_0 and the system goes to a "grave" which is a state $\mathbf{x}$ with one available action under which this state is absorbent, i.e. $p(\mathbf{x}|\mathbf{x}, a) = 1$, and one-step rewards are equal to zero, $r(\mathbf{x}, a) = 0$.

If the original MDP contains nonnegative rewards, V_0 is a nonnegative function, and the expected total rewards are finite for any policy and any initial state then the corresponding homogeneous model is positive. If the original model is negative and V_0 is a nonpositive function then the appropriate homogeneous model for an n-horizon model is negative. If the original model satisfies the General Convergence Condition with V_0 considered as a reward at n-th epoch then the appropriate homogeneous model for an n-horizon model satisfies the General Convergent Condition. In particular, if the original model satisfies the General Convergent Condition and $V_0 = 0$ then the homogeneous model, that corresponds to an n-horizon model, satisfies the General Convergent Condition too.

For n-horizon models, it is natural to use notations $V_i(x, V_0)$ instead of $V(x, n-i)$, $i = 0, \dots, n$. Then Theorem t:OE 3 implies that

$$V_{n+1}(x, V_0) = TV_n(x, V_0), \qquad x \in \mathbb{X}. \tag{1.11}$$

e:dp

Since any policy is equalizing, we obtain a dynamic programming algorithm for N-horizon problems: first solve iteratively (e:dp 3.11) for $n = 0, 1, \dots, N-1$ with $V_0(x, V_0) = V_0(x)$. Then for $\varepsilon > 0$ we define a Markov policy $\phi = (\phi_N, phi_{N-1}, \dots, \phi_1)$ by $\phi_{n+1}(x) = a$ where $n = N-1, N-2, \dots, 0$ $x \in X$, and a is an element of $\mathbb{A}(x)$ such that

$$T^a V_n(x, V_0) \geq TV_n(x, V_0) - \varepsilon/N.$$

Any policy is equalizing in the new homogeneous model. It is easy to see that Markov policy ϕ is ε-optimal. In addition, if this policy can be defined as a conserving policy ($\varepsilon = 0$) then it is optimal. For example, if functions V_0 and r are bounded above and the Pre-Compactness Assumption holds then a conserving policy can be defined; see Corollary c:conv1 3.7.

For the important case V_0 identical to 0, we write $V_n(x)$ instead of $V_n(x, V_0)$. An important question is whether the sequence $V_n(x, V_0)$ converges. We denote $V_\infty(x, V_0) = \lim_{n\to\infty} V_n(x, V_0)$ and $V_\infty(x) = V_\infty(x, 0)$ when these limits, possibly infinite, exist. Sequential computation of V_n is called value iteration or successive approximation. When the limit V_∞ exists, another important question is whether it equals V.

s:SPMO

1.5 Particular models

1.5.1 Discounted MDPs

We write the optimality operator T in the explicit form

$$Tf(x) = \sup\{r(x,a) + \beta P^a f(x) : a \in \mathbb{A}(x)\}.$$

This operator is a contracting mapping of the set of bounded functions endowed with the norm $\|f\| = \sup_x f(x)$ into itself; see Denardo [17] or Blackwell [7]. This implies that the optimality equation has a unique bounded solution and the limit $V_\infty(x, V_0)$ exists and is equal to this solution for any bounded V_0. Theorem 3 implies that this solution is V.

t:disc

Theorem 16 Blackwell [8], Denardo [17]) *Consider a* **D** *MDP.*

(i) For any $\varepsilon > 0$ there exists a ε-stationary optimal policy.

(ii) If a stationary policy is conserving, it is optimal.

(iii) If the Pre-Compactness Assumption holds then there exists a stationary optimal policy.

(iv) The limit V_∞ exists and equals V.

We remark that (i) and (ii) follow from the equalizing property applied either to a stationary policy ϕ with $T^{\phi(x)}V(x) \geq V(x) - (1-\beta)\varepsilon$ for all $x \in \mathbb{X}$ or to a stationary conserving policy; (iii) follows from Corollary 3.7, boundness of $V(x)$, and (iv); and (iv) follows from the fixed point theorem for contracting mappings. We also remark that, as the following simple example shows, the Optimality Equation may have additional unbounded solutions.

ex4

Example 17 $\mathbb{X} = \{0, 1, \dots\}$ and $\mathbb{A}(x) = \{a\}$. All rewards are equal to 0. If the system is in state i, it moves to state $i+1$; $p(i+1|i,a) = 1$. We have that $V(x) = 0$ for all x. However, any function $u(i) = C/\beta^i$ satisfies the Optimality Equation $u = Tu$. ■

1.5.2 Generalized Discounted MDPs

t:gdisc

Theorem 18 *Consider a* **GD** *MDP.*

(i) For any $\varepsilon > 0$ there exists a stationary ε-optimal policy.

(ii) If a stationary policy is conserving, it is optimal.

(iii) If the Compactness Assumption holds then there exists a stationary optimal policy.

(iv) The limit $V_\infty(x)$ exists and $V_\infty(x) \geq V(x)$ for all $x \in \mathbb{X}$.

Proof. The proof of statements (i-iii) coincides with the proof of the corresponding statements in Theorem t:disc 4. (iv) Let $r(x,a) \le C$ for all x and a. If we subtract C from r then $V(x,\pi)$ will be reduced by $C/(1-\beta)$ for all x and π. Therefore, we received an equivalent **N** MDP and the inequality follows from Theorem t:negat 6(iv) . ∎

ex5 **Example 19** ($V_\infty(x) > V(x)$) Let $\mathbb{X} = \{(0,0)\} \cup \{1,2,\dots\} \times \{0,1,\dots\}$, $\mathbb{A} = \mathbb{A}(0,0) = \{a^0, a^1, \dots\}$, and $\mathbb{A}(x) = \{a^0\}$ when $x \in \mathbb{X} \setminus \{(0,0)\}$. From any state (i,j) with $i > 0$, the system moves to $(i+1,j)$. This means that $p((i+1,j)|(i,j),a^0) = 1$ when $i > 0$. If action a^j is selected at state $(0,0)$, the system moves to $(1,j)$. In other words, $p((1,j)|(0,0),a^j)$=1. The rewards are

$$r((i,j),a^0) = \begin{cases} -\beta^{-i}, & \text{if } i = j \\ 0, & \text{otherwise.} \end{cases}$$

We have that $V_n(0,0) = 0$ for $n > 0$ and therefore $V_\infty(0,0) = 0$. We also have that $V(0,0) = -1$. ∎

1.5.3 Negative MDPs

t:negat **Theorem 20** (Strauch st66 [53]) *Consider an* **N** *MDP.*

(i) For any $\varepsilon > 0$ there exists a Markov ε-optimal policy.

(ii) If a stationary policy is conserving, it is optimal.

(iii) If the Pre-Compactness Assumption holds then there exists a stationary optimal policy.

(iv) The sequence $V_n(x)$ is nonincreasing, the limit $V_\infty(x)$ exists, and $V_\infty(x) \ge V(x)$ for all $x \in \mathbb{X}$.

(v) The value function V is the maximum nonpositive solution of the Optimality Equation.

(vi) If either $\mathbb{X}$ is finite or all sets $\mathbb{A}(x)$ are finite then $V_\infty(x) = V(x)$ for all $x \in \mathbb{X}$.

In order to verify (i), one should consider a Markov policy $\phi = \{\phi_1, \phi_2, \dots, \}$ such that $T^{\phi_n(x)} v(x) \ge v(x) - \varepsilon_n$ for all $x \in \mathbb{X}$ and for all $n \in \mathbb{N}$, where $\varepsilon_0 + \varepsilon_1 + \ldots \le \varepsilon$. Then the inequality $v(x) \le 0$ implies ε-optimality of ϕ. (ii) follows from the same reasons where ϕ is a stationary policy such that actions $\phi(x)$ are conserving at x. Since $V(x) \le 0$, the value function V is bounded above and Corollary c:conv1 3.7 implies (iii). (iv) is correct because $v(x,\pi) \le v(x,\pi,n+1) \le v(x,\pi,n)$ for any policy π and therefore

$V(x) \leq V_{n+1}(x) \leq V_n(x)$. Statements (i), (ii), and (iii) hold for general convergent MDPs with $V(x) \leq 0$ for all $x \in X$. The proof of these generalizations is the same as for **N** MDPs. It follows from the fact that any policy is equalizing. Statement (vi) was proved by Strauch [53]. When $\beta = 1$, Example 3.5 demonstrates the possibility of $V_\infty(x) > V(x)$; see also Strauch [53].

The following example demonstrates that stationary ε-optimal policies may not exist.

ex6 **Example 21** $\mathbb{X} = \{1,2\}$, $\mathbb{A} = \mathbb{A}(1) = [0,1)$, and $\mathbb{A}(2) = \{0\}$. State 2 is absorbent with zero one-step rewards, $p(2|2,0) = 1$ and $r(2,0) = 0$. If action a is selected in state 1, $p(1|1,a) = a$, $p(2|1,a) = 1-a$, and $r(1,a) = (a-1)$. Then $w(1,\phi) = -1$ and $V(1) = 0$. ■

1.5.4 Positive MDPs

The notion of an ε-optimal policies has been defined for a constant ε. It is easy to extend it when a function $f(x)$ is considered instead of constant ε. We say that a policy π is f-optimal for a given function f on $\mathbb{X}$ if $w(x,\pi) \geq v(x) - f(x)$ for all $x \in \mathbb{X}$.

t:negat1 **Theorem 22** *Consider a* **P** *MDP.*

(i) (Blackwell [8]) *$s(x) = V(x)$ for all $x \in \mathbb{X}$.*

(ii) (Ornstein [46]) *For any $\varepsilon > 0$ there exists a stationary εV-optimal policy.*

(iii) The sequence $V_n(x)$ is nondecreasing and the limit $V_\infty(x)$ exists and $V_\infty(x) = V(x)$ for all $x \in \mathbb{X}$.

(iv) The value function V is the minimum nonnegative solution of the Optimality Equation.

The first result was established by Blackwell [8]. Its proof consists of the following steps. First, we introduce a discount factor $\beta \in [0,1[$. Then $v(x,\pi,\beta)$ is non-decreasing in β and $v(x,\pi,\beta) \to v(x,\pi)$ as $\beta \to 1$. We fix an arbitrary $\varepsilon > 0$. If ϕ is a stationary policy such that $T^{\phi(x)}v(x,\beta) \geq v(x) - (1-\beta)\varepsilon$ for all $x \in \mathbb{X}$ then $v(x,\pi,\beta) \geq v(x) - \varepsilon$ for all $x \in \mathbb{X}$. This implies that for any positive $\beta < 1$ and for any policy π there exist a stationary policy ϕ such that $v(x,\phi,\beta) \geq v(x,\pi,\beta)$ for all $x \in \mathbb{X}$. Consider an ε-optimal policy π. We fix $x \in \mathbb{X}$ and consider β such that $v(x,\pi,\beta) \geq v(x,\pi) - \varepsilon$. We have that

$$s(x) \geq v(x,\phi,\beta) \geq v(x,\pi,\beta) - \varepsilon \geq v(x,\pi) - 2\varepsilon \geq V(x) - 3\varepsilon.$$

Since $\varepsilon > 0$ is arbitrary, $s(x) \geq V(x)$. However, $s(x) \leq V(x)$. Thus, $s(x) = V(x)$ for all $x \in \mathbb{X}$.

The proof of (ii) is non-trivial and it was explained in greater details in Hordijk [39]. Statements (ii) and (iii) are easy. Example 3.7 shows that conserving strategies may not be optimal. In addition it shows that the Compactness Assumption does not imply the existence of a stationary optimal policy even when the state space is finite.

ex7 **Example 23** (Cavazos-Cadena, Feinberg, Montes-de-Oca [13]). Consider an MDP with state and action spaces given by $\mathbb{X} = \{0, 1\}$ and $\mathbb{A} = [0, 1]$, respectively. The sets of admissible actions are defined by $\mathbb{A}(1) = A = [0, 1]$, and $\mathbb{A}(0) = \{0\}$, whereas the transition law and the reward function are determined by

$$p(0|0,0) = 1, \quad p(0|1,a) = a = 1 - p(1|1,a), \quad a \in [0,1],$$

and

$$r(0,0) = 0, \quad \text{and} \quad r(1,a) = a(1-a), \; a \in [0,1].$$

From these definitions, it is clear that 0 is an absorbing state under every policy and that $V(0) = 0$. Also, stationary policies are naturally indexed by the action they prescribe at state 1: $\phi_a \in \mathbb{F}$ is given by

$$\phi_a(1) = a, \quad \phi_a(0) = 0, \quad a \in [0,1].$$

The following statements are true for this example:

(a) the expected total-reward at state 1 under policy ϕ_a is given by $v(1, \phi_a) = 1 - a$ if $a \in (0, 1]$, and $v(1, \phi_a) = 0$ if $a = 0$;
(b) the optimal expected total-reward at state 1 is $V(1) = 1$;
(c) an optimal policy does not exist.

Indeed, (i) implies that $s(1) = 1$. Theorem 7(i) and (a) imply (c). From (a) and (b) we have that there is no stationary optimal policy. However, if an optimal policy exists in positive dynamic programming then a stationary optimal policy exists (Puterman 1994 , p. 324). Therefore, (a) and (b) yield (c).

Now we verify (a). Let $a \in (0, 1]$ be fixed. Starting at state $X_0 = 1$, under policy ϕ_a the system will arrive to state 0 at a random time T which is geometrically distributed with parameter a, that is, for each positive integer n, $\mathbb{P}_1^{\phi_a}\{T = n\} = a(1-a)^{n-1}$ and a reward $r(1, a) = a(1 - a)$ will be earned at each integer time $t \in \{0, 1, \dots, T-1\}$. Since $r(0,0) = 0$ and state 0 is absorbing, we have

$$V(1, \phi_a) = \mathbb{E}_1^{\phi_a}\left[\sum_{t=0}^{T-1} r(X_t, A_t)\right] = r(1,a)\,\mathbb{E}_1^{\phi_a}[T]$$
$$= a(1-a) \times \mathbb{E}_1^{\phi_a}[T] = a(1-a) \times \frac{1}{a} = (1-a).$$

Consider now $a = 0$. In this case, state 1 is absorbing under $\phi_a = \phi_0$, and a reward $r(1,0) = 0$ is earned forever, so that $V(1, \phi_0) = 0$. ■

Blackwell bl67 [8] constructed an example when there is no stationary ε-optimal policy for a positive countable MDP with a finite state of actions. The following example, which is Example 2 in Feinberg and Sonin feso81 [30] shows that randomized Markov K-optimal policies may not exist for any $K > 0$; see also relevant Example 2.26 in van der Wal vdw81 [58].

ex8 **Example 24** Let $X = \{(0,0)\} \cup \{(i,j) :\ i = 1, 2, \dots, j = -i+1, \dots, 0, 1\}$, and $\mathbb{A} = \{c, s\}$. In states $(0,0)$ and (i,j), where $i = 1, 2, \dots$, and $j < 0$, the action sets, transition probabilities, and rewards are the same as in Example ex3 3.3. We set $\mathbb{A}(i,1) = \{c\}$ and $p((0,0)|(1,1), c) = 1$, $p((i,1)|(i+1,1), c) = 1$, and $r((i,1), c) = 1$, $i = 1, 2, \dots$. We also set $\mathbb{A}(i,0) = \mathbb{A}$, $p((0,0)|(i,0), c) = p((i+1,0)|(i,0), c) = 0.5$, $p((2^i - i^2 + i - 1, 1)|(i,0), s) = 1$, and $r((i,0), a) = 0$, $i = 1, 2, \dots$, $a \in \mathbb{A}$.

We denote $g(i) = i^2 - i + 1$. The sequence $g(i)$ is used in this example because of its three properties: (a) $\frac{g(n+i)}{2^i} \to 0$ as $i \to \infty$, (b) $g(i) > 0$ when $i \geq 1$, and (c) $\sum_{i=1}^{\infty} \frac{1}{g(i)} < \infty$.

Let i be a positive integer. If action s is selected in state $(i,0)$, the total future reward is $2^i - g(i)$. If action c is selected in this state, the system moves with fifty-fifty chances to the "grave" $(0,0)$ or to the state $(i+1, 0)$. Let ϕ be a Markov policy and $m = \min\{n \geq 0 |\ \phi_n(i+n) = s,\ i = 0, 1, \dots\}$. If $m < \infty$, we have that $v((i,0), \phi) = 2^{-n}(2^{i+n} - g(i+n))$. If $m = \infty$ then $v((i,0), \phi) = 0$. Therefore, $v(i,0) = 2^i$. The optimality equation implies that $v(i,j) = 2^i$ when $j < 0$.

Let ψ be a randomized Markov K-optimal policy for some constant $K > 0$. Since $v((i,j), \psi) \geq V(i,j) - K = 2^i - K$, $j \leq 0$, we have that $\pi_j(s|(i,0)) \leq$

$K/g(i)$, $j = 0, 1, \ldots, i-1$. Then

$$v((i,0),\psi) = \sum_{t=0}^{\infty} P\{x_t = (i+t,0), a_t = s\}(2^{i+t} - g(i+t)) =$$
$$\pi_0(s|(i,0))(2^i - g(i)) + \sum_{n=1}^{\infty} \pi_n(s|(i+n,0))(\frac{1}{2})^n \times$$
$$[\prod_{j=0}^{n-1}(1 - \pi_j(s|(s+j,0)))](2^{n+i} - g(n+i)) \leq$$
$$2^i \sum_{n=0}^{\infty} \pi_n(s|(i+n,0)) < 2^i K \sum_{j=i}^{\infty} \frac{1}{g(j)} = 2^i - 2^i(1 - K\sum_{j=i}^{\infty}\frac{1}{g(j)}) < 2^i - K.$$

when i is large enough. The last inequality holds since $\sum_{j=i}^{\infty} \frac{1}{g(j)} \to 0$ as $j \to \infty$. Thus ψ is not K-optimal and for any $K > 0$ there is no randomized Markov K-optimal policy. ■

Puterman [48] describes a generalization of positive models called positive bounded models. For these models the assumption that the reward function r is nonnegative is replaced with a weaker assumption that for each $x \in \mathbb{X}$ there is at least one $a \in \mathbb{A}(x)$ for which $r(x,a) \geq 0$. As explained in Puterman [48], these more general models inherits many properties of positive models. Our remark is that, though $S = V$ for positive bounded models, stationary εV-optimal may not exist for them. The following example shows that randomized Markov εV-optimal policies may not exist in positive bounded MDPs.

ex9 **Example 25** Consider Example 3.8. Let $\tilde{v}$ and $\tilde{V}$ denote functions v and V in that example. If ψ is a randomized Markov policy such that $\tilde{v}((i,j),\psi) > \tilde{V}(i,j) - 1$ for all $i \geq 1$ and $j < 0$ then $\tilde{v}((i,0),\psi) < \tilde{V}(i,0) - 1$ for large i. Using the same arguments as in the previous example but with slightly more complicated calculations, it is possible to show that $\tilde{v}((i,-1),\psi) < \tilde{V}(i,-1) - K$ for some i. We recall that $\tilde{V}(i,j) = 2^i$.

Now we slightly modify Example 3.8 by setting $\mathbb{A}(i,-1) = \{c, s\}$ with $p((i,0)|(i,-1),c) = 1$, $r((i,-1),c) = 1 - 2^i$ and $p((0,0)|(i,-1),s) = 1$, $r((i,-1),s) = 0$. In other words, the decision maker can either pay $2^i - 1$ and move to the state $(i,0)$ or collect zero return and stop the process. Since $v(i,0) = 2^i$, the optimality equation implies that $V(i,j) = 1$ when $j < 0$. Fix $\varepsilon > 0$. Let ϕ be a randomized Markov policy and $v((i,j),\phi) \geq 1 - \varepsilon$

when $j < 0$. Consider the model from the Example 3.8. For that model consider a randomized Markov policy ψ which coincides with ϕ at all states except $(i, -1)$, $i = 1, 2, \ldots$. Since in the MDP from Example 3.8 we have that $\mathbb{A}(i, -1) = \{c\}$, policy ϕ always selects action c at states $(i, -1)$. We observe that $v((i, j), \phi) \leq \tilde{v}((i, j), \psi) + 2^i - 1$. Therefore, if $v((i, j), \phi) \geq V(i, j) - \varepsilon V(i, j)$ for all states (i, j) with $j < 0$ then $\tilde{v}((i, j), \psi) \geq 2^i - \varepsilon$ for all states with $j < 0$. As explained in the previous paragraph, this is impossible. ∎

1.6 The first main theorem: uniformly nearly-optimal stationary policies

s:SP

Comparison of positive and negative MDPs creates a strong impression that these models are totally different; see the table on page 324 in Puterman [48]. For example, for **P** MDPs $s = V$ but it is not true for **N** MDPs. In addition, it is even not obvious how to define a unified notion of ε-optimality. The notion of εV-optimality, which is natural for **P** MDPs, is not applicable to **N** MDPs because policies cannot better than optimal. The notion of ε-optimality, which is natural for **N** MDPs, is not applicable to **P** models; see Example 3.8.

As the result, almost all books treat these models separately. The major exceptions are Hinderer [38], Dynkin and Yushkevich [20], and van der Wal [58]. The first two books were written when very little was known about **GC** MDPs. Van der Wal [58] introduced several new results including several counterexamples, the proof that $s = V$ when all actions $\mathbb{A}(x)$ are finite, and interesting results on value iteration. The following result has been known for a long period of time.

t:FIN

Theorem 26 (Blackwell [6], Krylov [41], Dynkin and Yushkevich [20], Kallenberg [40]) *If $\mathbb{X}$ and A are finite then there exists a stationary optimal policy.*

As Examples 3.6 and 3.8 demonstrate, if either $\mathbb{X}$ or one of the sets $\mathbb{A}(x)$ is infinite then optimal policies may not exist. Van der Wal [58] showed that if all states $\mathbb{A}(x)$ are finite then $s(x) = V(x)$ for all x. Schäl [51] extended this result to Borel state problems with compact action sets.

For a long period time, there were no results available on the existence of stationary uniformly nearly-optimal policies in **GC** MDPs. Probably the first result was van der Wal's [60] theorem that states that stationary εV_+-optimal policies exist if for each $x \in \mathbb{X}_{\leq} = \{x \in \mathbb{X} | \ V(x) \leq 0\}$ there exists a conserving action. This result generalizes Ornsten's theorem (Theorem t:negat1 7(ii)). The weak point of this statement is that the use of function V_+ in the definition of ε-optimality does not look natural when reward functions can be negative. For example, if $V(x)$ is always nonpositive and there is a conserving action in each state, such actions form a stationary optimal policy; this fact follows from the same arguments as Theorem t:negat 6(ii). However, it is possible that $V_+(x)$ is unbounded from above and van der Wal's ydw84 [60] result implies the existence of stationary policies which are far from optimal. Van der Wal and Wessels ydwwe84 [61] asked whether V_+ could be substituted with a better function.

Now we describe the result from Feinberg and Sonin feso83 [31] on the existence of uniformly nearly-optimal policies within the class of stationary policies for **GC** MDPs. This result implies Ornstein's theorem and many other specific results. First, we define the sets

$$X_S = \{x \in X|\ v(x,\phi) = s(x) \text{ for some } \phi \in S\};$$

$$X_\Pi = \{x \in X|\ v(x,\pi) = V(x) \text{ for some } \pi \in \Pi\}.$$

For any x from X_Π there is a policy π which is optimal for this state. Similarly, for any x from X_S there is a stationary policy which is the best stationary policy for this state.

Let Φ be the class of numerical functions on $\mathbb{X}$ such that $Pf^+ < \infty$. We observe that functions V, S, V_+ belong to Φ. Constants also belong to Φ. For $Z \subseteq \mathbb{X}$ and for any function $f \in \Phi$, we denote by $L(f,Z)$ the set of nonnegative functions ℓ on $\mathbb{X}$ such that $\ell(x) > 0$ and $\ell(x) \geq \max\{f, P\ell\}$ when $x \in \mathbb{X} \setminus Z$. We also define $L(f) = L(f,\emptyset)$.

We remark that $L(f,Z)$ is the set of nonnegative functions which are positive excessive (or super-harmonic) majorants of f on $\mathbb{X} \setminus Z$. The sets $L(f,Z)$ possess the following two properties: (i) if $f(x) \geq g(x)$ for all $x \in \mathbb{X} \setminus Z$ then $L(g,Z) \supseteq L(f,Z)$, and (ii) if $Z \subset Y$ then $L(f,Z) \supseteq L(f,Y)$. We also observe that if $\ell \in L(f,Z)$ then $\ell_0 \in L(f,Z)$ where $l_0(x) = l(x)$ for $x \in \mathbb{X} \setminus Z$ and $l_0(x) = 0$ for $x \in Z$.

We observe that $V_+ + 1 \in L(V)$. Therefore, the set $L(V,Z) \neq \emptyset$ for any Z. We also observe that $V \in L(V, X_S)$ in **P** MDPs and $V_+ \in L(V, X_S)$ if for each $x \in \mathbb{X}_\leq$ there exists a conserving action. Indeed, consider the set $X^* = \{x \in X|V_+(x) = 0\}$. If an initial state belongs to X^* the system will never leave X^*. Therefore, we get a negative MDP if we restrict X to X^*. Since there are conserving actions in all sets $\mathbb{A}(x)$ when $x \in X_\leq$ and $X^* \subseteq X_\leq$, there is a stationary optimal policy in the negative MDP with the state space X^*. therefore, there is an optimal stationary policy for any initial point $x \in X^*$. Thus, $X^* \subseteq X_S$. And we have that $V_+ \in L(V_+, X^*) \subseteq L(V, X_S)$.

t:main1 **Theorem 27** (First main theorem: the existence of uniformly optimal stationary policies; Feinberg and Sonin [31], Theorem 2.1.) *For any $\varepsilon > 0$ and for any $\ell \in L(s, X_S)$ there exists a stationary policy ϕ such that $v(x, \phi) \geq s(x) - \varepsilon\ell$ for all $x \in \mathbb{X}$.*

The proof of Theorem 9 is not trivial and we do not consider it here. This theorem provides a unified way to prove the existence of uniformly $\varepsilon\ell$-optimal policies: it is sufficient to prove that $s(x) = V(x)$ for any given x and that $\ell \in L(s, X_S.)$ As we saw in Theorem 7, the proof that $s(x) = V(x)$ is significantly easier that the direct proof that there are uniformly nearly-optimal policies. For example, for **P** MDPs, we have $s(x) = V(x)$ (Theorem 7(i)) and $V \in L(V, X_S)$. Therefore, Theorem 9 implies Ornstein's theorem. If there is a conserving action for any $x \in X_{\leq}$ then $s(x) = V(x)$; this and more general results follow from our main statement, Theorem 20. Therefore, Theorem 9 implies van der Wal's [60] described before its formulation. We also observe that if V is bounded above by a constant K then $K \in L(V)$. Therefore, if $S = V$ and V is bounded above then there exist stationary ε-optimal policies. For example, Theorem 9 implies the existence of stationary ε-optimal policy for positive bounded models from Puterman [48]. Indeed, it is easy to see that in this model $V(x) \geq 0$ for all x and if $V(x) = 0$ then there is a conserving action at x. Thus, $s(x) = V(x)$ for all x in positive bounded models.

An important corollary from Theorem 9 is that the value function s is the solution of the optimality equation.

t:oes **Theorem 28** (Feinberg and Sonin [31, Theorem 2.2]) $s(x) = Ts(x)$ *for all* $x \in \mathbb{X}$.

We remark that we do not know how to prove $s = Ts$ in **GC** MDPs without using Theorem 9. It is easy to show that $Ts(x) \geq s(x)$ for all $x \in X$. In order to prove $s(x) \geq Ts(x)$ we need to use the existence of a stationary policy ϕ such that $v(z, \phi) \geq s(z) - \varepsilon(z)$ for all $z \in X$ and for some function $\varepsilon(z)$. In the proof of $V \geq TV$, we used Corollary 3.5 for the similar result. After Theorem 9 is established, the proof of $s = Ts$ is similar to the proof of $V = TV$; see Feinberg and Sonin [31] for details.

An important question is how to expand the class of functions $L(s, X_S)$. One possible approach is to replace s in $L(s, X_S)$ with a function $d \leq S$. Our particular interest is to consider d defined in a way that it is possible that s is not bounded above but d is bounded above. Van Dawen and Schäl [57], van Dawen [56], and Schäl and Sudderth [52] considered functions d of this type for particular models. These functions were related to the limiting behavior of $s(x_n)$. For countable MDPs, the broadest known class of such functions was introduced in Feinberg and Sonin [32]. Let $Q^b = \cup_{n=1}^{\infty}\{\tau < n\}$ be the set of all uniformly bounded stopping times. Let

$$d_S(x) = \sup_{\phi S}\inf_{\tau\in Q^b} \mathbb{E}_x^{\phi}\, s(x_\tau).$$

The following theorem generalizes Theorem 9 in the sense that it describes a larger class of functions ℓ with respect to which ε-optimal policies exist.

t:main1m **Theorem 29** (Feinberg and Sonin [32], Theorem 1.) *For any $\varepsilon > 0$ and for any $\ell \in L(d, X_S)$ there exists a stationary policy ϕ such that $v(x,\phi) \geq s(x) - \varepsilon\ell$ for all $x \in \mathbb{X}$.*

We say that an MDP is deterministic if all transition probabilities $p(y|x,a)$ are equal to 0 or 1. Bertsekas and Shreve [5] proved the existence of stationary ε-optimal policies in **P** MDPs.

t:main1d **Theorem 30** (Feinberg and Sonin [31, Section 5]) *Consider a deterministic MDP. Consider an arbitrary nonnegative function ℓ on $\mathbb{X}$ such that $\ell(x) \geq \ell(y)$ for $x, y \in \mathbb{X}\setminus X_S$ when there is an action $a \in \mathbb{A}(x)$ such that $p(y|x,a) = 1$. Then for any $\varepsilon > 0$ there exists a stationary policy ϕ such that $v(x,\phi) \geq s(x) - \varepsilon\ell(x)$ for all $x \in \mathbb{X}$. In particular, in a* **P** *MDP for any $\varepsilon > 0$ there exists an $\varepsilon \min 1, V$-optimal stationary policy.*

We see that stationary $\varepsilon\ell$-optimal policies exist in deterministic models for a broader class of functions ℓ. A natural direction of research is to expand the class $L(d, X_s)$ in Theorem 11. An open question is to get results that unify deterministic and stochastic models; see Feinberg [23] for additional details. Function d is related to value functions in gambling problems; see Dubins and Savage [19] and Maitra and Sudderth [45]. However, the meaning of this relationship currently is not clear.

In conclusion, we illustrate how Theorems 9 11 can be used to prove the existence of uniformly nearly-optimal policies in particular models.

t:co **Theorem 31** (cp. Cavazos-Cadena and Montes-de-Oca [14]) *Consider a negative MDP, all rewards r are nonpositive. Assume that if $V(x) > -\infty$ for some $x \in \mathbb{X}$ then*

$$\lim_{n\to\infty}\inf_{\phi\in S} \mathbb{E}_x^{\phi}\, v(x_n,\phi) = 0. \qquad (1.12) \quad \text{e:co}$$

If this assumption holds then for any $\varepsilon > 0$ there exists a stationary ε-optimal policy.

Proof. In view of Theorem t:main1 9, it is sufficient to prove that $s(x) = V(x)$ for all $x \in \mathbb{X}$. Since $s(x) \le V(x)$ we have that $s(x) = V(x)$ when $V(x) = -\infty$. We observe that it is sufficiently to prove the theorem for the situation when $V(x) > -\infty$ for all $x \in \mathbb{X}$. Indeed, let $Y = \{x \in \mathbb{X} |\ V(x) = -\infty\}$. If $Y = \mathbb{X}$ then the problem is trivial. So, we consider the case, $Y \ne \mathbb{X}$. In this case, we exclude the subset Y from $\mathbb{X}$ and remove all actions a such that $p(Y|x, a) > 0$ from sets $\mathbb{A}(x)$ when $x \in \mathbb{X} \setminus Y$. We remark that $\mathbb{A}(x)$, $X \in \mathbb{X} \setminus Y$ are still nonempty sets after this procedure because otherwise we would have $V(x) = -\infty$. So, we have received a new model in which condition (e:co 3.12) holds for all $x \in \mathbb{X}$.

So, it is sufficient to show that $s(x) = V(x)$ for any $x \in \mathbb{X}$ if (e:co 3.12) holds for all $x \in \mathbb{X}$. In order to do it, we fix an arbitrary $\varepsilon > 0$ and an arbitrary $x \in \mathbb{X}$. Then we select $n \ge 1$ such that

$$\inf_{\phi \in S} \mathbb{E}_x^{\phi}\, v(x_n, \phi) \ge -\varepsilon. \tag{1.13}$$ e:coone

We also select a stationary policy ϕ such that $T^{\phi(z)}V(z) \ge V(z) - \varepsilon/n$ for all $z \in \mathbb{X}$. We have that

$$(T^{\phi})^n V(x) \ge V(x) - \varepsilon. \tag{1.14}$$ e:cotwo

We also have

$$s(x) \ge v(x, \phi) = T^{\phi} v(x, \phi) = (T^{\phi})^n v(x, \phi) = (T^{\phi})^n (v(x, \phi) - V(x) + V(x)) =$$

$$(T^{\phi})^n V(x) + \mathbb{E}_x^{\phi}\, v(x_n, \phi) - \mathbb{E}_x^{\phi}\, V(x_n) \ge V(x) - 2\varepsilon.$$

The last inequality follows from (e:coone e:cotwo 3.13,3.14) and $V(x) \le 0$. Since $\varepsilon > 0$ is arbitrary, $s(x) \ge V(x)$. Thus, $s(x) = V(x)$. ■

We remark that $s = V$ if $s_- = V_-$ where the minus means that the function is related to the model in which r is replaced with r^-; see Theorem t:main2 20. So, if the model, in which r is replaced with r^- satisfies conditions of Theorem t:co 13 then for any $\ell \in L(d_s, X_s)$ and for any $\varepsilon > 0$ there exists a stationary $\varepsilon\ell$-optimal policy. We also observe that $X_\Pi = X_s$ if $V = s$.

t:fest **Theorem 32** (Corollary 2 in Feinberg fe86 [23]) *If for any $x \in \{x \in \mathbb{X} |\ V(x) > -\infty\}$ and for any $\pi \in \Pi^M$*

$$\limsup_{n\to\infty} \mathbb{E}_x^{\pi}\, s(x_n) \ge 0$$

then for any $\ell \in L(d_s, \mathbb{X}_\Pi)$ and for any $\varepsilon > 0$ there exists a stationary $\varepsilon\ell$-optimal policy.

In conclusion, we remark that $s(x) = s_R(x)$ for all $x \in X$ where $s_R(x) = \sup_{\pi \in \Pi^{RS}} v(x, \pi)$ for all $x \in X$. This result was proved for the countable state MDPs by Feinberg and Sonin [32] and for Borel MDPs by Feinberg [27].

1.7 (f, I)-generated policies

s:NSP

Van der Wal [59] proved that for any $\varepsilon > 0$ there exists a Markov $\varepsilon\ell$-optimal policy with $\ell = V_+ + 1$. Theorems 9 and 11 imply similar results for broader sets of functions ℓ. Indeed, let us replace the state space $\mathbb{X}$ with the state space of couples (x, n) where $x \in \mathbb{X}$ and $n = 0, 1, \ldots$. The process moves from states (x, n) to $(y, n+1)$ with transition probabilities $p(y|x, a)$. There is one-to-one correspondence between stationary policies in the new model and Markov policies in the original one. We also observe that if there exists a policy π such that $v(x, \pi) = V(x)$ for some $x \in \mathbb{X}$ then there is a Markov policy ϕ such that $v(x, \pi) = V(x)$; see Corollary 3.3. These observations and Theorems 9 imply the following result.

t:mare

Theorem 33 (Feinberg and Sonin [32]) *For any $\varepsilon > 0$ and for any $\ell \in L(V, X_\Pi)$ there exists a Markov $\varepsilon\ell$-optimal policy.*

We remark that if one uses Theorem 11 instead Theorem 9 then a slightly more general class of function ℓ can be considered. We shall formulate it as a part of a more general statement. The natural question is which classes of strategies contain uniformly mearly-optimal policies. In order to answer this question, we consider the following construction.

Let $I = \{0, 1, \ldots\}$ and $f : H \to I$. A policy π is called randomized (f, I)-generated if $\pi_n(\cdot|h_n) = \sigma(\cdot|x_n, f(h_n))$ for some conditional distribution σ. We also consider nonrandomized (f, I)-generated policies. In the latter case, $\pi_n(h_n) = \sigma(x_n, f(h_n))$ where σ is a mapping from $\mathbb{X} \times I$ to $\mathbb{A}$. If we say that a policy is (f, I)-generated we mean that it is nonrandomized and (f, I)-generated.

Markov policies are an example of (f, I)-generated policies. In this case $f(h_n) = n$. Stationary policies form another example. In this case, $f(h_n) = 0$ for all h_n. Tracking strategies introduced by Hill [37] is the third example. For tracking policies, decision depends on the current state and the number of visits to it, i.e. $f(h_n) = \sum_{i=0}^{n} \mathbf{I}\{x_i = x_n\}$. So, I is the information available about the past and f is a memory function which indicates what information about the past is remembered. We consider the following condition.

Transitivity Condition (Feinberg and Sonin [32]) If $f(h_n) = f(h'_m)$ and $x_n = x'_m$, where $h_n = x_0 a_0 x_1 \dots x_n$ and $h_m = x'_0 a'_0 x'_1 \dots x'_m$, then $f(h_n a z) = f(h'_m a z)$ for all $a \in \mathbb{A}(x_n)$ and for all $z \in \mathbb{X}$.

In other words, the current state x_n, the current information $i_n = f(h_n)$, the next action a, and the next state z completely define $i_{n+1} = f(h_n a z)$. We observe that stationary and Markov policies satisfy the Transitivity Condition and tracking policies do not satisfy this condition. We denote by I^f the set of (f, I)-generated policies. We also denote by RI^f the set of randomized (f, I)-generated policies. We define

$$V^f = \sup\{v(x,\pi) : \ \pi \in I^f\}, \qquad V^f_R = \sup\{v(x,\pi) : \ \pi \in RI^f\}.$$

Let

$$d^f(x) = \sup_{\phi \in I^f} \inf_{\tau \in Q^b} \mathbb{E}^\phi_x V^f(x_\tau).$$

Let also

$$\mathbb{X}^f = \{x \in \mathbb{X} | v(x,\phi) = V^f(x) \text{ for some } \phi \in I^f\}.$$

t:trans **Theorem 34** (Feinberg and Sonin [32]) *Let function f satisfies the transitivity condition. Then $V^f(x) = V^f_R(x)$ for all $x \in \mathbb{X}$ and for any $\ell \in L(d^f, \mathbb{X}^f)$ and for any $\varepsilon > 0$ there exists an (f, I)-generated policy ϕ such that $V(x,\phi) \geq V^f(x) - \varepsilon \ell(x)$ for all $x \in \mathbb{X}$.*

The proof of Theorem 16 is based on the following observation. Consider the MDP with the state space $\mathbb{X} \times \mathbb{N}$. If the transitivity condition holds then there is a one-to-one correspondence between (f, I)-policies and stationary policies in the new model where states are in fact couples $(x_n, f(h_n))$. Then Theorem 16 follows from $s(x) = s_R(x)$ and from Theorem 11.

Markov and stationary policies are two examples of classes of policies satisfying the transitivity conditions. Another important example is so-called Y-policies introduced in Feinberg [25]. For Y-policies, defined by a subset Y of $\mathbb{X}$, decisions depend on the following factors: (i) the current state, (ii) the number of visits to Y, and (iii) the time passed after the last visit to Y. For any history $h_n \in H_n$, $n = 0, 1, \dots$, we define the number of visits to Y

$$m(h_n, Y) = \sum_{i=0}^{n} \mathbf{I}\{x_i \in Y\}.$$

We also agree by that definition $x_{-1} \in Y$ and let $\xi(h_n, Y)$ is the epoch when the system visited Y for the last time,

$$\xi(h_n, Y) = \max\{i = -1, 0, 1, \dots , n | \; x_n \in Y\}$$

and $\theta(h_n, Y) = n - \xi(h_n, Y)$ be the time passed after the last visit.

For Y-embedded policies we define $I = \{0, 1, \dots\} \times \{0, 1, \dots\}$ and $f(h_n) = (m(h_n, Y), \theta(h_n, Y))$. We observe that if $f(h_n) = (m, i)$ then $f(h_n, a, z) = (m, i+1)\mathbf{I}\{z \notin Y\} + (m+1, 0)\mathbf{I}\{z \in Y\}$ where $\mathbf{I}$ is an indicator function. Therefore, the Transitivity Condition holds and Theorem t:trans 16 can be applied to Y-embedded policies. We also remark than in two extreme cases, $Y = \mathbb{X}$ and $Y = \emptyset$, the set of Y-embedded policies coincides with the set of Markov policies.

Let τ^1 be the first epoch when the system hits Y, $\tau^1(h) = \min\{n \geq 0 | \; x_n \in Y\}$ and let τ^m be the $(m+1)$-hitting epoch for $Y : \tau^m = \min\{n > \tau^m | \; x_n \in Y\}$. We observe that $\{\tau^m(h) = n\} = \{(m(h_n, Y) = m, \theta(h_n, Y) = 0$. The following theorem is a direct generalization of Theorem t:MC-Markov 1 to Y-embedded policies. It was proved by induction in Feinberg fe87 [24] when $\lambda_1 = 1$ and $\lambda_i = 0$ for $i > 1$. However, in the case of arbitrary λ_i, the proof remains the same.

t:embed **Theorem 35** (Theorem 4.1 in fe87 [24]) *Let $\pi^1, \pi^2, \dots$ be an arbitrary sequence of policies and $\lambda_1, \lambda_2, \dots$ a sequence of nonnegative numbers summing to 1. For an arbitrary $Y \subseteq X$ consider a randomized Y-embedded policy π defined by*

$$\pi\left(C \mid y, m, k\right) \overset{def}{=} \frac{\sum_{i=1}^{\infty} \lambda_i \, \mathbb{P}_x^{\pi^i} \left(x_{\tau^m+k} = y, \tau^{m+1} > \tau^m + k, a_{\tau^m+k} \in C\right)}{\sum_{i=1}^{\infty} \lambda_i \, \mathbb{P}_x^{\pi^i} \left(\tau^m + k = y, \tau^{m+1} > \tau^m + k\right)} \, , \tag{1.15}$$

e:eqrandE

whenever the denominator in e:eqrandE (3.15) is not equal to 0, where $C \in \mathbb{A}$, $y \in \mathbb{X}$, $m = 1, 2, \dots$, and $k = 0, 1, \dots$. Then, for all $m, k = 0, 1, \dots$, $y \in \mathbb{X}$ and measurable subsets C of $\mathbb{A}(y)$,

$$\begin{aligned} &\mathbb{P}_x^{\pi} \left(x_{\tau^m+k} = y, \tau^{m+1} > \tau^m + k, a_{\tau^m+k} \in C\right) = \\ &\sum_{i=1}^{\infty} \lambda_i \, \mathbb{P}_x^{\pi^i} \left(x_{\tau^m+k} = x, \tau^{m+1} > \tau^m + k, a_{\tau^m+k} \in C\right) \, , \end{aligned} \tag{1.16}$$

e:MC-embed

and therefore

$$v(x, \pi) = \sum_{i=1}^{\infty} \lambda_i v(x, \pi^i). \tag{1.17}$$

Theorem 17 provides an important result in the following direction. Let Δ be some class of randomized policies. Then there is an explicit formula such that indicates a policy $\pi \in \Delta$ for a given initial state x and for an arbitrary policy π such that $v(x,\sigma) = v(x,\pi)$.

The results in this direction were provided for stationary policies by Krylov [42, 43] for discounted controlled diffusion processes and by Borkar [12] for discounted MDPs. Altman [1] proved the same result for MDPs with uniformly bounded life times. Discounted MDPs are a particular case of such models. In this case, the occupation measure for the original policy is equal to the occupation measure for the corresponding randomized stationary policy. For a more general case when the expected number to each state is bounded above, Altman [1] proved a weaker result that the occupation measure for the corresponding randomized stationary policy majorizes the original occupation measure. Feinberg and Sonin [33] constructed an example when the strong inequality takes place.

We remark that our Theorem 17 has no limitation on the life time of the system. In is a direct generalization of Theorem 1 which follows from Theorem 17 when $Y = \mathbb{X}$ aor $Y = \emptyset$. Unfortunately, its relationship of Theorem 17 with the results on the existence of equivalent randomized stationary policies is unknown; see Borkar's paper in this volume.

In addition to the Transitivity Condition, the so-called Non-Repeating Condition for (f,I)-policies plays an important role.

Non-Repeating Condition (Feinberg [24]) If history $h_m = x_0a_0\dots x_m$ is a continuation of history $h_n = x_0a_0\dots x_n$, i.e. $h_m = h_na_n\dots x_m$, then $(x_n, f(h_n)) \neq (x_m, f(h_m))$.

In other words, if the Non-Repeating Condition means that if $m > n$, $h_m = h_na_nx_{n+1}\dots x_m$, and $x_n = x_m$ then $f(h_m) \neq f(h_n)$. Markov, tracking, and Y-embedded policies are examples of (f,I)-generated policies that satisfy the Non-Repeating Condition. Stationary policies do not satisfy this condition.

Any randomized (f,I)-generated policy π is defined by transition probabilities $\pi(\cdot|x,i)$. Any nonrandomized (f,I)-generated policy σ is defined by a function $\sigma(x,i)$. Let I^f be the set of all functions ϕ on $\mathbb{X}\times I$ with values on $\mathbb{A}$ and such $\phi(x,i) \in \mathbb{A}(x)$ for all $x \in X$. Any randomized (f,I)-generated policy π defines a measure m^π on the set I^f defined as the product of the countable set of independent measures $\pi(\cdot|x,i)$ over the set $\mathbb{X}\times I$; see Neveu [47, Proposition VI.2] for the existence and uniqueness of such products.

t:nonrep **Theorem 36** (Feinberg [24], Theorem 3.1) *If the function f satisfies the Non-Repeating Condition then for any $C \in \mathcal{F}_\infty$ and for any randomized (f, I)-generated policy π*

$$\mathbb{P}_x^\pi(C) = \int\limits_{I^f} \mathbb{P}_x^\phi(C) m^\pi(d\phi) \qquad (1.18)$$ e:nonrep

and therefore $v(x, \pi) = \int\limits_{I^f} v(x, \phi) m(d\phi)$.

c:nonrep **Corollary 37** *If π is a randomized (f, I)-generated policy, f satisfies the Non-Repeating Condition, and x is any fixed element of $\mathbb{X}$ then $v(x, \phi) \geq v(x, \pi)$ for some (f, I)-generated policy ϕ.*

Thus, we have three groups of results that can help us to prove the existence of uniformly $\varepsilon\ell$-optimal policies in some class of nonrandomized policies. First, Theorems 1, 17, and relevant results for stationary policies related to occupation measures (see Krylov [42, 43], Borkar [12], Altman [1], and Feinberg and Sonin [33]) provide the methods to prove that for any initial state and policy there is an equivalent randomized policy in a given class of policies. Theorem 18, Theorem 16, or equality $s(x) = s_R(x)$ (see [32, 27]) imply that a policy can be selected in a nonrandomized form. Theorems 9, 11, and 16 imply the existence uniformly $\varepsilon\ell$-optimal policies within certain classes of policies.

1.8 The second main theorem: uniformly nearly optimal locally stationary policies

s:tsmt

Ornstein's theorem (Theorem 7(ii)) implies the existence of stationary uniformly nearly optimal policies in **P** MDPs. Example 3.6 implies that such policies do not exist in **N** MDPs. Demko and Hill [16] proved that if $r(x, a) = r(x) < 0$ all $x \in \mathbb{X}$ then for any $\varepsilon > 0$ there exists a stationary ε-optimal policy. This result can be interpreted in the following way: if rewards are negative in a strong sense then there exist stationary uniformly nearly optimal policies.

Since stationary nearly optimal policies may not exist, it is natural to consider subsets of the state space on which such policies exist. We say that a nonrandomized policy ϕ is stationary on the set $Y \subseteq \mathbb{X}$ if $\phi_n(x_0 a_0 \ldots x_n) = \phi(x_n)$ for some function ϕ when $x_n \in Y$. We denote by $\Pi^{S,Y}$ the set of policies stationary on Y. We remark that $\Pi^S = \Pi^{S,\mathbb{X}}$. We also denote

$$s_Y(x) = \sup_{\phi\in\Pi^{S,Y}} w(x,\phi).$$

We consider the sets

$$\mathbb{X}^+ = \{x \in \mathbb{X}|\ V(x) > 0\}, \qquad \mathbb{X}^- = \{x \in \mathbb{X}|\ V(x) < 0\},$$
$$\mathbb{X}^c = \{x \in \mathbb{X}|\ A^c(x) \neq 0\}, \qquad \mathbb{X}^* = X^+ \cup \mathbb{X}^- \cup \mathbb{X}^*.$$

Thus, $\mathbb{X}^+$ is the subset of states where the value function is positive, $\mathbb{X}^-$ is the subset where the value function is positive, and $\mathbb{X}^c$ is the set of states where exist conserving actions. We have that the set $\mathbb{X}^*$ contains all elements of $\mathbb{X}$ except those where the value function is equal to 0 and there is no conserving actions. Chitashvili yuch [64] showed that if $\mathbb{X}$ is finite then $s_{\mathbb{X}^*}(x) = V(x)$ for all $x \in \mathbb{X}$. Feinberg fe87a [25] proved the existence of uniformly nearly optimal policies in $\Pi^{S,\mathbb{X}^*}$ for the countable state space.

Let

$$d(x) = \sup_{\phi\in\Pi^{S,X^*}} \inf_{\tau\in Q^b} \mathbb{E}_x^\phi V(x_\tau). \tag{1.19}$$ e:defd

The following theorem is a particular case of Theorem 6.2 in Feinberg fe87a [25].

t:prel **Theorem 38** *Consider a set (f,I)-generating policies I^f with the function f satisfying the Non-Repeating Condition. Then for any $\varepsilon > 0$ and for any $\ell \in L(d,\mathbb{X}_\Pi)$ there exists a policy ϕ with the following properties: (i) ϕ is stationary on $\mathbb{X}^*$, (ii) $\phi \in I^f$, and (iii) is uniformly $\varepsilon\ell$-optimal, i.e.*

$$v(x,\phi) \geq V(x) - \varepsilon\ell(x), \qquad x \in \mathbb{X}.$$

We observe that if $Y \subseteq Z$ then $s_Y(x) \geq s_Z(x)$ for all $x \in \mathbb{X}$. In view of Theorem t:prel 19 it is natural to find the biggest set Y for which $s_Y(x) = V(x)$ for all $x \in \mathbb{X}$. Unfortunately, even when $\mathbb{X}$ is finite, it is possible that there are several maximum sets with this property (a subset Y is called a maximum subset with a given property if there is no set Z other than Y such that Z satisfies this property and contains Y). The following example is similar to one provided by Chitashvili ch76 [15]. It shows that it is possible that $s_Y = V$ and $s_Z = V$ but $s_{X\cup Y}(x) < V(x)$ for some x.

ex10 **Example 39** Let $\mathbb{X} = \{0,1,g\}$ and $\mathbb{A} = \{0\} \cup \{b^1,b^2,\dots\} \cup \{c^1,c^2,\dots\}$. The state g is absorbing, $\mathbb{A}(g) = \{0\}$, $r(g,0) = 0$, and $p(g|g,0) = 1$. We also have that $\mathbb{A}(x) = \{b^1,b^2,\dots\} \cup \{c^1,c^2,\dots\}$ for $x = 1,2$. We also have $p(g|x,b^i)=1/i$, $p(x|x,b^i)=1-1/i$, $p(1|0,c^i)=p(0|1,c^i)=1$, and $r(x,b^i)=r(x,c^i)=-1/i$ for $i=1,2,\dots$ and $x=0,1$. It is easy to see that $V(g)=s(g)=0$, $V(0)=V(1)=0$, and $s(0)=s(1)=-1$. Let $Y=\{0,g\}$ and $Z=\{1,g\}$. Then $\mathbb{X}=Y\cup Z$. It is easy to see that $s_Y(x)=s_Z(x)=V(x)$. ■

According to Example 6.1 in Feinberg fe87a [25], the following situation is possible. There are finite sets Z_n, $n=1,2,\dots$ such that $Z_n\subseteq Z_{n+1}$, $\mathbb{X}=\cup_{n=1}^{\infty}Z_n$, $s_{Z_n}=V$ but $s\neq V$. Since $s=s_{\mathbb{X}}$, this example shows that if $\mathbb{X}$ is countable, maximal subsets, for which there exist good policies stationary on them, may not exist. The following natural way to expand the set $\mathbb{X}^*$ was described in Feinberg fe87a [25].

Let $r(x,a)=r^1(x,a)+r^2(x,a)$ where r^1 and r^2 are measurable in a. We assume that r^2 is a nonnegative function and consider two MDPs with the same state space, the same action spaces, and the same transition probabilities as the original MDP but with the reward function r^1 and r^2 respectively. Let V^1 and V^2 are the value functions for these MDPs. We assume that $V^2(x)<\infty$ for all $x\in\mathbb{X}$. We also set $Z=\{x\in\mathbb{X}|\ V^1(x)<0\}$. Of course, the set Z depends on the selection of r^2 which defines r^1 and V^1. For example, we can select $r^2(x,a)=0$ for all x and a. In this case, $r=r+0$ and $Z=\mathbb{X}^-$. We can select $r^2(x,a)=r^+(x,a)$. Then $r^1=r^-$ and $Z=\{x\in\mathbb{X}|\ V_-(x)<0.\}$ It is obvious that this set contains $\mathbb{X}^-$. We also can select $r^2=kr^+$ where k is a constant or nonnegative bounded function of x and a.

We also remark that if a function r^2 is replaced with a function, that is greater or equal to $r^2(x,a)$ for all $x\in\mathbb{X}$ and for all $a\in\mathbb{A}(x)$, then the corresponding set Z expands. However, the function r_2 cannot be arbitrary large because of the condition $V^2(x)<\infty$ for all $x\in\mathbb{X}$.

t:main2 **Theorem 40** (The second main theorem; Feinberg fe87a [25]) *Consider a set (f,I)-generating policies I^f with the function f satisfying the Non-Repeating Condition. Let Z be a subset of $\mathbb{X}$ defined above by some nonnegative function r^2 with $V^2(x)<\infty$ for all $x\in\mathbb{X}$. Then for any $\varepsilon>0$ and for any $\ell\in L(d,\mathbb{X}_\Pi)$ there exists a policy ϕ with the following properties: (i) ϕ is stationary on $\mathbb{X}^*\cup Z$, (ii) $\phi\in I^f$, and (iii) is uniformly $\varepsilon\ell$-optimal, i.e.*

$$v(x,\phi)\geq V(x)-\varepsilon\ell(x),\qquad x\in\mathbb{X}.$$

We remark that Theorem t:main2 20 implies almost all known results on the existence of uniformly nearly optimal policies. For example, for **P** MDPs,

$\mathbb{X} = \mathbb{X}^*$ and Theorem 20 implies Ornstein's theorem. For $I^f = \Pi^M$, it implies van der Wal's [59] theorem on the existence of uniformly nearly optimal Markov policies. If I^f is selected to be the set of tracking policies, Theorem 20 gives the positive answer to the question asked by van der Wal and Wessels [61] on the existence of uniformly nearly optimal policies. If $X^0 \subseteq X^c$ Theorem 20 implies the existence of stationary $\varepsilon\ell$-optimal policies and this result strengthens van der Wal's theorem on stationary strategies [60]; see also Theorem 2.22 in van der Wal [58]. In addition, if $\mathbb{X}$ and $\mathbb{A}$ are finite, this result implies the existence of stationary optimal policies because the set of stationary policies if finite and the stationary $\varepsilon\ell$-optimal policy is optimal for small $\varepsilon > 0$.. In addition, Theorem 9 contains the statement that if there exists an optimal policy then there exists a stationary optimal policy.

There are two important open questions related to Theorem 9: (i) how to provide the broadest natural description of the set where stationary policies are sufficient; (ii) how to decrease the function d? With respect to the second question, it would be nice to select d in a form that it equals to 0 for deterministic MDPs. Some results and discussion related to these two important questions can be found in Feinberg [23, 25]. In particular, Theorem 6.2 in Feinberg [25] contains function d which is less than or equal to function d defined in (3.19).

1.9 Uncountable MDPs

Most of the described above results hold for Borel MDPs. In particular, Dynkin and Yushkevich [20] studied the sets of strategic measures $\mathbb{P}^\pi_\mu$. They showed that this is a convex Borel space. Furthermore, this set is convex in a strong sense when strategic measures are integrated with respect to any probability measure; see [20, Section 3.5]. The observation that this set is convex is important. Imagine that we expand the notion of a strategy in the following way: the decision-maker selects randomly a policy at epoch 0 and then follows it. This initial randomization procedure is defined by a probability measure on the set of strategic measures with a given initial distribution or state. We call such strategies mixed. The convexity of the set of strategic measures implies that for each mixed policy there is an equivalent randomized policy. Therefore, in-fact mixed policies do not expand the set of policies. In view of this fact, policy π in Theorem 1 which was originally established by Strauch [53] for Borel MDPs and $\lambda_1 = 1$, can be constructed in two steps for a fixed initial state x: for any sequence of nonnegative numbers λ_i with the sum equal to 1 and for any sequence of policies π^i,

$i = 1, 2, \dots$, there exists a policy σ with $\mathbb{P}_x^\sigma = \sum\limits_{i=1}^{\infty} \mathbb{P}_x^{\sigma^i},$; and (ii) for any policy σ there exists a Markov policy π such that $\mathbb{P}_x^\pi\{x_n \in Y, a_n \in B\} = \mathbb{P}_x^\sigma\{x_n \in Y, a_n \in B\}$ for any Borel subsets $Y \subseteq \mathbb{X}$ and $B \subseteq \mathbb{A}$. In both cases (i) and (ii) there are simple formulae for σ and π.

Another general result for strategic measures is that strategic measures for randomized Markov strategies can be presented as convex integral combinations of strategic measures for nonrandomized Markov policies; Feinberg [28]. This fact implies that for any policy and for any given initial distribution there exists a Markov policy with equal or better performance; Feinberg [21, 22]. Schäl [50], Balder [3], and Yushkevich [63] studied topological properties of the sets on strategic measures and used them to establish sufficient conditions for the existence of optimal policies.

The optimality equation holds for Borel MDPs; see Strauch [53], Dynkin and Yushkevich [20], and Bertsekas and Shreve [4]. The major technical difficulty when a Borel state space is considered instead of a countable state space is related to so-called selection theorems. All results from Section 3.5, except Ornstein's theorem (Theorem 7(ii)) were originally proved for classical Borel models and the existence of (p, ε)-optimal policies was proved.

Blackwell [8] recognized that the validity of Ornstein's theorem for Borel positive MDPs was a difficult question and posted it as an open problem. Frid [34] proved that Ornstein's theorem is valid for Borel MDPs if $(p, \varepsilon V)$-optimality is considered instead of εV-optimality. Schäl and Sudderth [52] found a correctable mistake in Frid's proof: one of the sets where Frid switched policies was universally measurable but not Borel measurable.

Here I would like to mention the name of gifted mathematician Efim Frid who died at a young age as a result of an accident: he was hit by a vehicle when he was crossing a street in Odessa. Efim was a student of Nicolai Krylov. In addition to the proof of Blackwell's conjecture, Frid also wrote one of the first papers on MDPs with multiple criteria [35] and several interesting papers on stochastic games, in particular, on the sequence of nonzero sum two-person games. Though during many years we both have been living in Moscow at the same time, unfortunately I have never met Efim Frid.

Bertsekas and Shreve [4] studied MDPs with universally measurable policies. Except Ornstein's theorem they proved that for all results on the existence of ε-optimal policies from Section 3.5, there are similar ε-optimal policies for Borel MDPs with universally nearly optimal policies. Blackwell and Ramakrishnan [11] constructed an example of a **P** Borel MDP with a bounded function V for which there is no stationary universally measurable policy which is uniformly ε-optimal. This example implies that Ornstein's theorem as well as more general statements, Theorems 9, 11, 16, 19, and 20 cannot be expanded to Borel MDPs with universally measurable polices. The questions whether any of these statements or Theorem 15 hold for Borel MDPs when the notion of p-a.s. $\varepsilon\ell$ policies is considered are completely open. The key issue here is Theorem 9. Its proof in Feinberg and Sonin [31] used explicitly that the state space is not bigger than countable. Except Frid [34] and Blackwell and Ramakrishnan [11], the only research that studied possibilities to expand Ornstein's theorem to uncountable state spaces and other relevant issues, I am familiar with, is the paper by Schäl and Sudderth [52].

As was observed in Feinberg [26], it is not clear whether the Non-Repeating Condition implies the result similar to Theorem 18 for Borel MDPs. It was shown there that a so-called Strong Non-Repeating Condition, which holds for Markov policies, implies such result.

We also mention that Feinberg [27] proved that $s_R(x) = s(x)$ for all $x \in \mathbb{X}$. Schäl [51] proved that $s(x) = V(x)$, $x \in X$, for models with compact action sets. Since Theorem 9 is an open question for Borel MDPs when $(p, \varepsilon\ell)$-optimality is considered, we do not know if $s = Ts$ for **GC** Borel MDPS. The non-trivial part is to prove that $s(x) \geq Ts(x)$ for all $x \in \mathbb{X}$. The measurability properties of function s were established in Feinberg [28] by using the results by Sudderth [54] and Blackwell [9].

Acknowledgement. This research was partially supported by NSF Grant DMI-9908258.

Bibliography

al96 [1] [al96] E. Altman, “Constrained Markov decision processes with total cost criteria: occupation measures and primal LP,” *Math. Methods Oper. Res.* **43** pp. 45–72, 1996.

al99 [2] [al99] E. Altman, *Constrained Markov Decision Processes,* Chapman & Hall/CRC, Boca Raton, 1999.

ba89 [3] [ba89] E.J. Balder, “On compactness of the space of policies in stochastic dynamic programming,” *Stoch. Proc. Appl.* **32** pp. 141–151, 1989.

besh78 [4] [besh78] D.P. Bertsekas and S. Shreve, *Stochastic Optimal Control: the Discrete Time Case,* Academic Press, New York, 1978 (reprinted by Athena Scientific, Belmont, 1996).

besh79 [5] [besh79] D.P. Bertsekas and S. Shreve, “Existence of stationary optimal policies in deterministic optimal control,” *J. Math. Anal. Appl.* **69**, pp. 607–620, 1979.

bl62 [6] [bl62] D. Blackwell, “Discrete dynamic programming,” *Ann. Math. Stat.* **33** pp. 719–726, 1962.

bl65 [7] [bl65] D. Blackwell, “Discounted dynamic programming,” *Ann. Math. Stat.* **36** pp. 226–235, 1965.

bl67 [8] [bl67] D. Blackwell, “Positive dynamic programming,” in *Proceedings of the 5th Berkeley Symposium on Mathematical Statistics and Probability* **1** pp. 415–418, University of California Press, Berkeley, 1967.

bl76 [9] [bl76] D. Blackwell, “The stochastic processes of Borel Gambling and dynamic programming,” *Ann. Statist.* **4** pp. 370–374, 1976.

blfror [10] [blfror] D. Blackwell, D. Freedman and M. Orkin, “The optimal reward operator in dynamic programming,” *Ann. Probab.* **2** pp. 926–941, 1974.

blra [11] [blra] D. Blackwell and S. Ramakrishnan, “Stationary plans need not be uniformly adequate for leavable, Borel gambling problems,” *Proc. Am. Math. Soc.* **102** pp. 1024–1027, 1988.

bo88 [12] [bo88] V.S. Borkar, “A convex analytic approach to Markov decision processes,” *Prob. Theor. Relat. Fields* **78** pp. 583-602, 1988.

ccfemo [13] [ccfemo] R. Cavazos-Cadena, E.A. Feinberg and R. Montes-de-Oca, “A note on the existence of optimal policies in total reward dynamic programs with compact action sets,” *Math. Oper. Res.,* to appear.

ccmo [14] [ccmo] R. Cavazos-Cadena and R. Montes-de-Oca, “Nearly optimal stationary policies in negative dynamic programming,” *Math. Methods Oper. Res.* **49** pp. 441–456, 1999.

ch76 [15] [ch76] R.Ya. Chitashvili, “On the existence of ε-optimal stationary policies for a controlled Markov chain,” *Soobshch. Acad. Nauk. Gruzin. SSR* **83** pp. 549–552, 1976.

dehi [16] [dehi] S. Demko and T.P.Hill, “Decision processes with total cost criteria,” *Ann. Prob.* **9** pp. 293-301, 1981.

de67 [17] [de67] E.V. Denardo, “Contracting mappings in the theory underlying dynamic programming,” *SIAM Rev.* **9** pp. 165–177, 1967.

dest [18] [dest] C. Derman and R. Strauch, “A note on memoryless rules for controlling sequential control processes,” *Ann. Math. Stat.* **37** pp. 276–278, 1966.

dusa [19] [dusa] L.E. Dubins and L.J. Savage, *How to Gamble If You Must: Inequalities for Stochastic Processes.* McGraw-Hill, New York; second edition: Dover, New York, 1976.

dyyu [20] [dyyu] E.B. Dynkin and A.A. Yushkevich, *Controlled Markov Processes.* Springer-Verlag, New York, 1979.

fe82 [21] [fe82] E.A. Feinberg, ”Nonrandomized Markov and semi-Markov strategies in dynamic programming,” *SIAM Theory Prob. Appl.* **27** pp. 116–126, 1982.

fe82a [22] [fe82a] E.A. Feinberg, ”Controlled Markov processes with arbitrary numerical criteria,” *SIAM Theory Prob. Appl.* **27** pp. 486–503, 1982.

fe86 [23] [fe86] E.A. Feinberg, ”The structure of persistently nearly optimal strategies in stochastic dynamic programming,” *Lecture Notes in Control and Inform. Sci.* **81** pp. 22–31, 1986.

fe87 [24] [fe87] E.A. Feinberg, ”Sufficient classes of strategies in discrete dynamic programming. I: decomposition of randomized strategies and imbedded models,” *SIAM Theory Prob. Appl.* **31** pp. 658–668, 1987.

fe87a [25] [fe87a] E.A. Feinberg, ”Sufficient classes of strategies in discrete dynamic programming. I: locally stationary strategies,” *SIAM Theory Prob Appl.* **32** pp. 478–493, 1987.

fe91 [26] [fe91] E.A. Feinberg, “Nonrandomized strategies in stochastic decision processes,” *Ann. Oper. Res.* **29** pp. 315–332, 1991.

fe92 [27] [fe92] E.A. Feinberg, “On stationary strategies in borel dynamic programming,” *Math. Oper. Res.* **17** pp. 393–397, 1992.

fe96 [28] [fe96] E.A. Feinberg, “On measurability and representation of strategic measures in Markov decision processes”, in *Statistics, Probability and Game Theory Papers in Honor of David Blackwell* (eds. T.S. Ferguson et al.), IMS Lecture Notes - Monograph Series, **30**, pp. 29–43, 1996.

fesh96 [29] [fesh96] E.A. Feinberg and A. Shwartz, “Constrained discounted dynamic programming,” *Math. of Operations Research* **21** pp. 922–945, 1996.

feso81 [30] [feso81] E.A. Feinberg and I.M. Sonin, “Markov policies in infinite horizon dynamic programming problems with bounded value functions,” in *Abstracts of 4-th USSR - Japan Symposium on Probability Theory and Mathematical Statistics* **1** pp. 209–210, Tbilisi, 1982.

feso83 [31] [feso83] E.A. Feinberg and I.M. Sonin, “Stationary and Markov policies in countable state dynamic programming,” *Lecture Notes in Math.,* **1021** pp. 111–129, 1983.

feso85 [32] [feso85] E.A. Feinberg and I.M. Sonin, “Persistently nearly optimal strategies in stochastic dynamic programming,” in *Statistics and Control of Stochastic Processes, Steklov Seminar* pp. 69–101, Optimization Software, New York, 1985

feso96 [33] [feso96] E.A. Feinberg and I.M. Sonin, “Notes on equivalent stationary policies in Markov decision processes with total rewards,” *ZOR - Math. Methods of Oper. Res.* **44** pp. 205–221, 1996.

fr70 [34] [fr70] E.B. Frid, “On a problem of D. Blackwell from the theory of dynamic programming,” *SIAM Theory Prob Appl.* **15** pp. 719–722, 1970.

fr72 [35] [fr72] E.B. Frid, “On optimal strategies in cintrol problems with constraints,” *SIAM Theory Prob Appl.* **17** pp. 188–192, 1972.

gisk79 [36] [gisk79] I.I. Gikhman and A.V. Skorokhod, *Controlled Random Processes,* Springer, New York, 1979.

hi79 [37] [hi79] T. Hill, “On the existence of good Markov strategies,” *Trans. Amer. Math. Soc.* **247** pp. 157–176, 1979.

hi [38] [hi] K. Hinderer, *Foundations of Non Stationary Dynamic Programming with Discrete Time Parameter*, Lecture Notes in Operations Research **33**, Springer-Verlag, NY, 1970.

ho [39] [ho] A. Hordijk, *Dynamic Programming and Markov Potential Theory*, Math. Centre Tracts **51**, Math. Centrum, Amsterdam, 1974.

ka [40] [ka] L.C.M. Kallenberg, *Linear Programming and Finite Markovian Problem*, Math. Centre Tracts **148**, Math. Centrum, Amsterdam, 1983.

kr65 [41] [kr65] N.V. Krylov, “Construction of an optimal strategy for a finite controlled chain,” *SIAM Theory Prob. Appl.* **10** pp. 45–54, 1965.

kr85 [42] [kr85] N.V. Krylov, “Once more about the connection between elliptic operators and Itô’s stochastic equations,” in *Statistics and Control of Stochastic Processes, Steklov Seminar*(N.V. Krylov, R.Sh. Liptser, and A.A. Novikov, eds.) pp. 69–101, Optimization Software, New York, 1985

kr87 [43] [kr87] N.V. Krylov, “An approach in the theory of controlled diffusion processes,” *SIAM Theory Prob. Appl.* **31** pp. 604–626, 1987.

masu92 [44] [masu92] A. Maitra and W.D. Sudderth “The optimal return operator in negative dynamic programming,” *Math. Oper. res.* **17** pp. 921–931, 1992.

masu96 [45] [masu96] A. Maitra and W.D. Sudderth, *Discrete Gambling and Stochastic Games,* Springer-Verlag, New York, 1996.

or [46] [or] D. Ornstein, “On the existence of stationary optimal strategies,” *Proc. Am. Math. Soc.* **20** pp. 563–569, 1969.

ne [47] [ne] J. Neveu, “Mathematical Foundations of the Calculus of Probability,” Holden-Day, San Francisco, 1965.

pu [48] [pu] M.L. Puterman, *Markov Decision Processes*, Wiley, New York, 1994.

ro83 [49] S.M. Ross, *Introduction to Stochastic Dynamic Programming,* Academic Press, Orlando, 1983.

sc75 [50] [sc75] M. Schäl, “On dynamic programming: compactness of the space of policies,” *Stoch. Processes Appl.* **3** pp. 345–364, 1975.

sc83 [51] [sc83] M. Schäl, “Stationary policies in dynamic programming models under compactness assumptions,” *Math. Oper. Res.* **8** pp. 366–372, 1983.

scsu [52] [scsu] M. Schäl and W.D. Sudderth, “Statioanary policies and Markov policies in Borel dynamic programming,” *Probab. Th. Rel Fields* **74** pp. 91–111, 1987.

st66 [53] [st66] R. Strauch, “Negative dynamic programming,” *Ann. Math. Stat.* **37** pp. 871–890, 1966.

su [54] [su] W.D. Sudderth “On the existence of good stationary strategies,” *Trans. Amer. Math. Soc.* **126** pp. 399–414, 1969.

vd85 [55] [vd85] R. van Dawen, “Negative dynamic programming,” Preprint no. 458, University of Bonn, 1985.

vd86 [56] [vd86] R. van Dawen, “Pointwise and uniformly good stationary strategies in dynamic programming models,” sl Math. Oper. Res. **9** pp. 521-535 (1986).

vdsh [57] [vdsh] R. van Dawen and M. Schäl, “On the existence of Stationary Optimal Policies in Markov Decision Processes,” *Z. Angew. Math. Mech.* **63**, no 5, pp. T403-T404, 1983.

vdw81 [58] [vdw81] J. van der Wal, *Stochastic Dynamic Programming*, Math. Centre Tracts **139**, Math. Centrum, Amsterdam, 1981.

vdw83 [59] [vdw83] J. van der Wal, “On uniformly nearly-optimal Markov strategies,” in *Operations Research Proceedings* 1982, pp. 461–467. Springer-Verlag, Berlin, 1983.

vdw84 [60] [vdw84] J. van der Wal, “On stationary strategies in countable state total reward Markov decision processes,” *Math. Oper. Res.* **9** pp. 290–300, 1984.

vdwwe84 [61] [vdwwe84] J. van der Wal and J. Wessels, “On the use of information in Markov decision processes,” *Statistics & Decisions* **2**, pp. 1–21, 1984.

vh [62] [vh] K.M. van Hee, “Markov strategies in dynamic programming,” *Math. Oper. Res.* **3** pp. 37–41, 1978.

yu97 [63] [yu97] A.A. Yushkevich “The compactness of a policy space in dynamic programming via an extension theorem for Caratheodory functions,” *Math. Oper. Res.* **22**, pp. 458–467, 1997.

yuch [64] [yuch] A.A. Yushkevich and R.Ya. Chitashvili, “Controlled random sequences and Markov chains,” *Russian Math. Surveys* **37**, no 6, pp. 239-274, 1982.