On the Reduction of Total-Cost and Average-Cost MDPs to Discounted MDPs

Eugene A. Feinberg [1] Jefferson Huang [2]

[1] *Department of Applied Mathematics and Statistics Stony Brook University, Stony Brook, New York 11794-3600*

[2] *School of Operations Research and Information Engineering Cornell University, Ithaca, New York 14853-3801*

Received 15 May 2015; revised 3 May 2017; accepted 6 May 2017
DOI 10.1002/nav.21743
Published online 00 00 2017 in Wiley Online Library (wileyonlinelibrary.com).

Abstract: This article provides conditions under which total-cost and average-cost Markov decision processes (MDPs) can be reduced to discounted ones. Results are given for transient total-cost MDPs with transition rates whose values may be greater than one, as well as for average-cost MDPs with transition probabilities satisfying the condition that there is a state such that the expected time to reach it is uniformly bounded for all initial states and stationary policies. In particular, these reductions imply sufficient conditions for the validity of optimality equations and the existence of stationary optimal policies for MDPs with undiscounted total cost and average-cost criteria. When the state and action sets are finite, these reductions lead to linear programming formulations and complexity estimates for MDPs under the aforementioned criteria. © 2017 Wiley Periodicals, Inc. Naval Research Logistics 00: 000–000, 2017

Keywords: Markov decision process; reduction; linear program; transient; total cost; average cost

1. INTRODUCTION

This article deals with the reduction of undiscounted total-cost and average-cost Markov decision processes (MDPs) to discounted MDPs. For undiscounted total costs, we consider a weighted-norm version of the *transient* case introduced by Veinott [52] in the context of finite state and action sets and by Pliska [38] in the context of Borel state and action spaces. A feature of such MDPs is that nonnegative transition *rates*, which may not be transition probabilities, are considered. One of the applications of such models is to the control of branching processes; see for example, Rothblum and Veinott [42] and Pliska [38]. Other references for branching processes and other models with transition rates greater than one are given in Section 2.1. *Absorbing* MDPs, which were introduced by Hordijk [32] and studied in the constrained setting by Altman [2] and Feinberg and Rothblum [23], can also be viewed as transient MDPs.

It is well-known that discounted MDPs can be reduced to absorbing or transient MDPs (see e.g., [2, p. 137]). Theorem 6 in this article provides conditions under which the converse is also true. In particular, the reduction comes from a version of the similarity transformation considered by Veinott [52], which is attributed there to Alan Hoffman. This reduction relates the value function and optimality equation of the original transient model with those of the corresponding discounted model. It implies the existence of stationary optimal policies for transient models if certain natural conditions hold. It also implies that the sets of optimal actions for these two models coincide. In the case of finite state and action sets, the reduction shows that complexity estimates for Howard's [34] policy iteration algorithm for discounted MDPs imply corresponding estimates for transient MDPs. Ye [55] proved that Howard's policy iteration algorithm, which corresponds to a block-pivoting simplex method, and the simplex method with Dantzig's rule compute optimal policies for discounted MDPs with a fixed discount factor in strongly polynomial time. The complexity estimates from [55] were improved in Hansen et al. [26] and further improved in Scherrer [46]. Ye [55] and Denardo [9] also obtained complexity estimates for transient MDPs. In Section 3.3, Denardo's [9] estimate for Howard's policy iteration algorithm, which corresponds to a block-pivoting simplex method, is derived from Scherrer's [46] estimate by using the reduction of a transient MDP to a discounted one. We remark that, unlike Howard's policy iteration algorithm, any member of a broad class of modified policy iteration algorithms, which includes value iteration, is not strongly polynomial for discounted MDPs with a fixed discount factor [20, 21].

Correspondence to: Eugene A. Feinberg (eugene.feinberg@sunysb.edu)

© 2017 Wiley Periodicals, Inc.

On the other hand, the discounted-cost criterion plays an important role in the theory of average-cost MDPs. Many results have been proved using the so-called "vanishing discount factor" approach, where discounted total costs with discount factor tending to one are used to obtain a stationary average-cost optimal policy via an optimality inequality or equation; see for example, Sennott [47, Chapter 7], Schäl [45], Hernández-Lerma and Lasserre [27, Chapter 5], and Feinberg et al. [22].

A direct reduction of average-cost MDPs to discounted ones, which yields sufficient conditions for the existence of stationary average-cost optimal policies, was established by Ross [39, 40] for MDPs with Borel state space, finite action sets, bounded costs, and a state to which the process will transition from any state under any action with probability at least $\alpha > 0$. This reduction and Ye's [55] results were used by Feinberg and Huang [19] to obtain iteration bounds for average-cost policy iterations. Gubenko and Štatland [25] showed that a reduction is also possible for MDPs with Borel state space, bounded costs, and compact action sets, if a "minorization" condition, which generalizes Ross's [40] assumption, is satisfied; see also Dynkin and Yushkevich [12, Chapter 7, §10].

More recently, Akian and Gaubert [1] used methods from non-linear Perron–Frobenius theory to reduce a perfect-information zero-sum stochastic game with finite state and action sets, containing a state being recurrent under every pair of stationary strategies, to a discounted game with state-dependent discount factors. In this article, we provide a slightly modified version of their transformation for the case of MDPs with possibly infinite state and action spaces. This reformulation makes the connection between their transformation and the work of Ross [39, 40] and Veinott and Hoffman [52] more apparent. In the context of MDPs with transition probabilities, this transformation yields a reduction of a finite state and action average-cost problem with a state recurrent under every stationary policy to a discounted MDP. The transformation also allows one to write the optimality equation, prove the existence of stationary optimal policies, and, in the case of finite state and action sets, formulate an alternative linear program for such average-cost problems. This program is based on the linear program formulation for the discounted MDP, to which the original problem is reduced. Therefore, an average-cost problem can be solved in strongly polynomial time with complexity estimates similar to those in Scherrer [46]. In addition, Howard's policy iterations for the obtained discounted MDPs coincide with Howard's policy iterations for the initial average-cost unichain MDP. Therefore, Scherrer's [46] results on discounted MDPs imply that Howard's policy iteration algorithm for the average-cost problem computes an optimal policy in strongly polynomial time with the complexity estimates similar to the estimates in [46]. This also implies that, if there exists a state recurrent under all stationary policies, the block-pivoting simplex method for the linear programming problem for average-cost MDPs, is also strongly polynomial with the same complexity estimates.

Previously, Zadorojniy et al. [56] showed that, if every state is recurrent under every stationary policy and an MDP satisfies a coupling property introduced there, then both discounted and average-cost optimal policies can be computed in strongly polynomial time. This is proved in [56] by introducing an algorithm that, as was shown by Even and Zadorojniy [14], is equivalent to applying the Gass–Saaty pivoting rule to the appropriate LP formulation for an MDP. As is shown in [56], the aforementioned coupling property holds for discrete-time versions of M/M/1 queues.

The model and the optimality criteria considered in this article are described in Section 2. In Section 3, we formulate the *Hoffman-Veinott* (*HV*) transformation [52], and give conditions under which it leads to the reduction of the original transient total-cost MDP to a discounted MDP with transition probabilities. Finally, in Section 4 we consider a version of Akian and Gaubert's [1] transformation for average-cost MDPs and the associated reduction to discounted MDPs. Most of the paper deals with countable-state MDPs. Sections 3.3 and 4.3 deal with finite-state problems, while Sections 3.4 and 4.4 study MDPs with Borel state spaces.

2. MODEL DESCRIPTION

Consider a discrete-time MDP with *state space* $\mathbb{X}$ and *action space* $\mathbb{A}$. Most of this article, except Sections 3.4 and 4.4, deals with countable-state MDPs. We start by introducing a countable-state MDP. Let $\mathbb{X}$ be countable and $\mathbb{A}$ be a Borel subset of a complete separable metric space. For each $x \in \mathbb{X}$, the *set of available actions* $A(x)$ is a nonempty Borel subset of $\mathbb{A}$. The *one-step cost* function $c(x, a)$ is (Borel-)measurable in $a \in A(x)$ for each $x \in \mathbb{X}$. The *transition rates* $q(y|x, a) \geq 0$ are measurable in $a \in A(x)$ for each $x, y \in \mathbb{X}$ and satisfy

$$\sup\left\{\sum_{y\in\mathbb{X}} q(y|x,a) : x \in \mathbb{X},\ a \in A(x)\right\} < \infty. \qquad (1)$$

2.1. Remarks on Transition Rates Whose Sum May Be Greater than One

The case where $\sum_{y\in\mathbb{X}} q(y|x, a)$ is possibly greater than one for some state-action pairs has been studied under various names. In Rothblum and Veinott [42] and in Rothblum and Whittle [43], such models are called *branching Markov decision chains*. They have also been referred to as *Markov population decision chains* in [13, 54]. As is explained in Remark 1 below, such models can be viewed as Markov decision processes with transition probabilities and a state-action-dependent discount factor that is possibly greater than one. The case of a constant discount factor, which is possibly greater than one is studied in Hinderer and Waldmann [31].

Such models are applicable in a diverse array of contexts. For example, Markov decision models with transition rates with values possibly greater than one appear in multi-armed bandit problems with risk-seeking utility functions; see Denardo et al. [10, 11]. In addition, their relevance to the control of multitype branching processes, which can be used to model problems in infinite particle systems, marketing, and population genetics, is explained in Pliska [37, 38]. Other relevant application areas are described in Eaves and Veinott [13].

REMARK 1: Equivalently to considering transition rates $q(\cdot|x,a)$, one can consider transition probabilities $p(\cdot|x,a)$ and a discount function $\alpha : \mathbb{X}\times\mathbb{A}\to[0,\infty)$. In particular, given an MDP in the latter form, let $q(\cdot|x,a) := \alpha(x,a)p(\cdot|x,a)$ for $x\in\mathbb{X}$, $a\in A(x)$; conversely, given transition rates $q(\cdot|x,a)$, let $\alpha(x,a) := q(\mathbb{X}|x,a)$ and $p(\cdot|x,a) := q(\cdot|x,a)/q(\mathbb{X}|x,a)$ for $x\in\mathbb{X}$, $a\in A(x)$. Expected total costs for arbitrary policies can be defined in a standard way via the Ionescu Tulcea Theorem (see e.g., [3, pp. 140–141]) by interpreting $\alpha(x,a)$ as a state-action dependent discount factor, and p as a transition probability. The existing literature on total-cost MDPs with transition rates having values possibly greater than one deals only with Markov policies; see for example [13, 37, 38, 43]. This remark overcomes this limitation. However, for transient total-cost models this remark and the reduction to a discounted MDP with transition probabilities and a discount factor less than one (Section 3.2) imply the optimality of stationary policies over all randomized history-dependent ones. Therefore, we mostly consider only stationary policies in this article. We remark that, when (1) holds, it is also possible to transform the original total-cost problem to a discounted one with a constant discount factor possibly greater than one; see [30, Remark 5].

2.2. Optimality Criteria

A *stationary policy* is a mapping $\phi : \mathbb{X}\to\mathbb{A}$ satisfying $\phi(x)\in A(x)$ for each $x\in\mathbb{X}$; let $\mathbb{F}$ denote the set of all such policies. It can be shown that it suffices to consider such policies for the optimality criteria considered in this article; see Remarks 4 and 15. Under $\phi\in\mathbb{F}$, the decision-maker always selects the action $\phi(x)$ when the current state is x. For $\phi\in\mathbb{F}$, consider the matrix of one-step transition rates Q_ϕ with elements $q(y|x,\phi(x))$, $x, y\in\mathbb{X}$. Also, given a *weight function* $W : \mathbb{X}\to[1,\infty)$ and a matrix B with elements $B(x, y)$, $x, y\in\mathbb{X}$, let

$$\|B\|_W := \sup_{x\in\mathbb{X}} W(x)^{-1}\sum_{y\in\mathbb{X}}|B(x,y)|W(y).$$

If $W(x)=1$ for all $x\in\mathbb{X}$, then $\|B\|_W = \|B\| := \sup_{x\in\mathbb{X}}\sum_{y\in\mathbb{X}}|B(x,y)|$. If the function W is bounded from above and below by a finite constant C, then

$$\|B\|_W \le C\|B\|. \qquad (2)$$

In particular, if $\mathbb{X}$ is a finite set, then (2) holds with $C = \max_{x\in\mathbb{X}}W(x)$.

For undiscounted total costs, which are considered in Section 3, the following generalization of the transience condition studied in Veinott [52] and Pliska [38] is assumed to hold.

ASSUMPTION T:

(i) The MDP is transient, that is, there is a weight function $V : \mathbb{X}\to[1,\infty)$ and a constant $K\ge 1$ that satisfy

$$\left\|\sum_{n=0}^{\infty} Q_\phi^n\right\|_V \le K < \infty \quad \text{for all } \phi\in\mathbb{F}. \qquad (3)$$

(ii) There is a constant $\bar{c}<\infty$ satisfying

$$\sup_{a\in A(x)}|c(x,a)| \le \bar{c}V(x) \quad \text{for all } x\in\mathbb{X}. \qquad (4)$$

(iii) For every $x\in\mathbb{X}$ the mapping

$$a\mapsto\sum_{y\in\mathbb{X}} q(y|x,a)V(y)<\infty, \quad a\in A(x),$$

is continuous on $A(x)$.

For $V\equiv 1$, a number of conditions sufficient for or equivalent to (3) are provided in Pliska [38]. If the state and action sets are finite, then Assumption T is equivalent to the assumption that there exists a constant K such that $\|\sum_{n=0}^{\infty}Q_\phi^n\|\le K<\infty$. For finite state and action sets, Assumption T can be checked in strongly polynomial time using the procedure described in [53, proof of Theorem 1], where it is attributed to Eric Denardo; see also [9, Lemma 10].

For $\phi\in\mathbb{F}$, let $c_\phi(x) := c(x,\phi(x))$ for $x\in\mathbb{X}$. Under Assumption T, the *total cost* incurred under $\phi\in\mathbb{F}$, when the initial state is $x\in\mathbb{X}$, is

$$v^\phi(x) := \sum_{n=0}^{\infty} Q_\phi^n c_\phi(x).$$

A policy ϕ_* is *total-cost optimal* if $v^{\phi_*}(x) = \inf_{\phi\in\mathbb{F}}v^\phi(x) =: v(x)$ for all $x\in\mathbb{X}$.

The following characterization of Assumption T will be used to define the transformations described in Sections 3.1 and 4.1 for total-cost MDPs.

PROPOSITION 1: Assumption T(i) holds if and only if there is a function $\mu : \mathbb{X} \to [1, \infty)$ such that $V(x) \le \mu(x) \le KV(x)$ for all $x \in \mathbb{X}$ and

$$\mu(x) \ge V(x) + \sum_{y \in \mathbb{X}} q(y|x,a)\mu(y), \quad x \in \mathbb{X},\ a \in A(x). \tag{5}$$

PROOF: Suppose there is a function $\mu : \mathbb{X} \to [1, \infty)$ that satisfies $V(x) \le \mu(x) \le KV(x)$ for all $x \in \mathbb{X}$ and (5). Consider an arbitrary $\phi \in \mathbb{F}$. According to (5),

$$\mu(x) \ge V(x) + \sum_{y \in \mathbb{X}} q(y|x,\phi(x))\mu(y) \quad \text{for all } x \in \mathbb{X},$$

which, since μ is nonnegative and majorized by *KV*, implies that for $N = 1, 2, \ldots$

$$\begin{aligned} KV(x) &\ge \sum_{n=0}^{N-1} Q_\phi^n V(x) + Q_\phi^N \mu(x) \\ &\ge \sum_{n=0}^{N-1} Q_\phi^n V(x) \quad \text{for all } x \in \mathbb{X}. \end{aligned}$$

Hence

$$K \ge V(x)^{-1} \lim_{N\to\infty} \sum_{n=0}^{N-1} Q_\phi^n V(x) \quad \text{for all } x \in \mathbb{X}. \tag{6}$$

Since $\phi \in \mathbb{F}$ is arbitrary, it follows from (6) that Assumption T holds.

Conversely, suppose Assumption T holds and consider the operator $\mathcal{U}$ defined for functions $u : \mathbb{X} \to [0, \infty)$ by

$$\mathcal{U}u(x) := \sup_{A(x)} \left[V(x) + \sum_{y \in \mathbb{X}} q(y|x,a)u(y) \right], \quad x \in \mathbb{X}.$$

Let $u_0 := V$, and for $n = 1, 2, \ldots$ let $u_n = \mathcal{U}u_{n-1}$. Note that the positivity of V implies $V \le u_n \le u_{n+1}$ for all n. Furthermore, letting $\mu := \lim_{n\to\infty} u_n$, Lebesgue's monotone convergence theorem implies that $\mu = \mathcal{U}\mu$. Hence to complete a proof, it suffices to show that $u_n \le KV$ for all n.

Note that $u_0 = V \le KV$ because $K \ge 1$. Next, suppose $u_n \le KV$ for some nonnegative integer n, and consider an arbitrary $\epsilon > 0$. Let ϕ^ϵ be a stationary policy satisfying

$$\begin{aligned} V(x) + \sum_{y \in \mathbb{X}} q(y|x,\phi^\epsilon(x))u_n(y) & \\ \ge \mathcal{U}u_n(x) - \epsilon(KV(x))^{-1}, &\quad x \in \mathbb{X}. \end{aligned}$$

Define $\tilde{u}_0 := u_n$. For $N = 1, 2, \ldots$ let

$$\tilde{u}_N(x) := \sum_{i=0}^{N-1} Q_{\phi^\epsilon}^i V(x) + Q_{\phi^\epsilon}^N u_n(x), \quad x \in \mathbb{X}. \tag{7}$$

Since $0 \le u_n \le KV$, it follows that $0 \le Q_{\phi^\epsilon}^N u_n \le K Q_{\phi^\epsilon}^N V$ for all N, which according to Assumption T implies that $Q_{\phi^\epsilon}^N u_n \to 0$ as $N \to \infty$. Hence it follows from (7) and Assumption T that

$$\lim_{N\to\infty} \tilde{u}_N(x) \le \sum_{i=0}^{\infty} Q_{\phi^\epsilon}^i V(x) \le KV(x) \quad \text{for all } x \in \mathbb{X}. \tag{8}$$

Next, we claim that

$$\tilde{u}_N(x) \ge u_{n+1}(x) - \epsilon(KV(x))^{-1} \sum_{i=0}^{N-1} Q_{\phi^\epsilon}^i V(x) \quad \text{for all } x \in \mathbb{X},\ N \ge 1. \tag{9}$$

Observe that (8) and (9) together with Assumption T imply

$$KV(x) \ge u_{n+1}(x) - \epsilon \quad \text{for all } x \in \mathbb{X}.$$

Since $\epsilon > 0$ is arbitrary, this implies by induction that $u_n \le KV$ for all n, from which the Proposition follows. To verify that (9) holds, first observe that for all $x \in \mathbb{X}$,

$$\begin{aligned} \tilde{u}_1(x) = V(x) + Q_{\phi^\epsilon} u_n(x) &\ge \mathcal{U}u_n(x) - \epsilon(KV(x))^{-1} \\ &= u_{n+1}(x) - \epsilon(KV(x))^{-1}. \end{aligned}$$

Next, suppose $\tilde{u}_N \ge u_{n+1} - \epsilon(KV)^{-1} \sum_{i=0}^{N-1} Q_{\phi^\epsilon}^i V$ for some $N \ge 1$. Then, since $u_{n+1} \ge u_n$, it follows that for $x \in \mathbb{X}$

$$\begin{aligned} \tilde{u}_{N+1}(x) &= V(x) + Q_{\phi^\epsilon} \tilde{u}_N(x) \\ &\ge V(x) + Q_{\phi^\epsilon} u_{n+1}(x) - \epsilon(KV(x))^{-1} \sum_{i=0}^{N-1} Q_{\phi^\epsilon}^{i+1} V(x) \\ &\ge V(x) + Q_{\phi^\epsilon} u_n(x) - \epsilon(KV(x))^{-1} \sum_{i=1}^{N} Q_{\phi^\epsilon}^i V(x) \\ &\ge \mathcal{U}u_n(x) - \epsilon(KV(x))^{-1} - \epsilon(KV(x))^{-1} \sum_{i=1}^{N} Q_{\phi^\epsilon}^i V(x) \\ &= u_{n+1}(x) - \epsilon(KV(x))^{-1} \sum_{i=0}^{(N+1)-1} Q_{\phi^\epsilon}^i V(x). \qquad \square \end{aligned}$$

LEMMA 2: Suppose statements (i) and (iii) of Assumption T hold, and let μ be the function described in the statement of Proposition 1. Further, suppose that for every $x, y \in \mathbb{X}$ the mappings $a \mapsto q(y|x,a)$ and $a \mapsto q(\mathbb{X}|x,a)$ are continuous on $A(x)$. Then for every $x \in \mathbb{X}$ the mapping

$$a \mapsto \sum_{y \in \mathbb{X}} q(y|x,a)\mu(y), \quad a \in A(x),$$

is continuous on $A(x)$.

PROOF: Fix $x \in \mathbb{X}$, and let $\{a_n\}$ be any sequence in $A(x)$ converging to some $a \in A(x)$. Under the hypotheses of the lemma, the sequence of measures $\{q(\cdot|x, a_n)\}$ converges setwise to $q(\cdot|x, a)$; for a definition of setwise convergence of measures, see for example [44, p. 269]. Since $0 \le \mu(x) \le KV(x)$ for all $x \in \mathbb{X}$, and $\sum_{y\in\mathbb{X}} q(y|x,a)V(y) < \infty$, it follows from the dominated convergence theorem for setwise converging measures (see e.g., [44, Proposition 18]) that

$$\lim_{n\to\infty} \sum_{y\in\mathbb{X}} q(y|x,a_n)\mu(y) = \sum_{y\in\mathbb{X}} q(y|x,a)\mu(y). \qquad \square$$

REMARK 2: For $\phi \in \mathbb{F}$ and $x \in \mathbb{X}$, let $\tau^{\phi}(x) := \sum_{n=0}^{\infty} Q_{\phi}^{n} e(x)$ and $\tau(x) := \sup_{\phi\in\mathbb{F}} \tau^{\phi}(x)$. Then it follows from [28, Proposition 9.6.4] that $\tau(x) \le KV(x)$ for all $x \in \mathbb{X}$. When the transition rates q are substochastic, that is, $\sum_{y\in\mathbb{X}} q(y|x,a) \le 1$ for all $x \in \mathbb{X}$ and $a \in A(x)$ (if equality holds for all x and a, then q is called *stochastic*), the quantity $\tau^{\phi}(x)$ can be interpreted as the expected total lifetime of the process under the policy ϕ when x is the initial state.

For average costs, which are dealt with in Section 4, Assumption HT on hitting times formulated below is assumed to hold. To state it, for $z \in \mathbb{X}$ and $\phi \in \mathbb{F}$ consider the matrix ${}_zQ_{\phi}$ with elements

$${}_zQ_{\phi}(x,y) := \begin{cases} q(y|x,\phi(x)), & \text{if } x \in \mathbb{X},\ y \ne z, \\ 0, & \text{if } x \in \mathbb{X},\ y = z. \end{cases}$$

ASSUMPTION HT:

(i) There is a state $\ell \in \mathbb{X}$ and a constant K^* satisfying

$$\left\| \sum_{n=0}^{\infty} {}_{\ell}Q_{\phi}^{n} \right\| \le K^* < \infty \quad \text{for all } \phi \in \mathbb{F}. \qquad (10)$$

(ii) The one-step cost function c is bounded.

REMARK 3: Observe that Assumptions T and HT are related. If an MDP satisfies Assumption HT then, if state ℓ and all transition rates to it are removed, the truncated MDP is transient with $V \equiv 1$. In particular, when the transition rates are substochastic or the sets $\mathbb{X}$ and $A(\ell)$ are finite, Assumption HT for the initial MDP and Assumption T with $V \equiv 1$ for the MDP with the state ℓ removed are equivalent. For the substochastic case, $K^* \le K+1$, where K is the constant from Assumption T for the truncated MDP. This is true because the truncated MDP does not contain the state ℓ, whereas K^* is an upper bound on the mean recurrence time for all the states of the original MDP, including the state ℓ, under any stationary policy.

When q is substochastic, Assumption HT means that when the initial state is x, the expected hitting time to state ℓ under any stationary policy is bounded above by K^*. When the state and action sets are finite, Assumption HT is equivalent to state ℓ being recurrent under all stationary policies. According to Feinberg and Yang [24], Assumption HT can be checked in strongly polynomial time. We remark that any MDP satisfying Assumption HT is unichain, and that in general the problem of checking whether an MDP is unichain is NP-hard [50]. In addition, Assumption HT is related to many other recurrence conditions that have been used to study average-cost MDPs; see for example, the surveys by Federgruen et al. [15], Thomas [49], and Hernández-Lerma et al. [29].

For the initial state $x \in \mathbb{X}$, the *average cost* incurred under $\phi \in \mathbb{F}$ is

$$w^{\phi}(x) := \limsup_{N\to\infty} \frac{1}{N} \sum_{n=0}^{N-1} Q_{\phi}^{n} c_{\phi}(x).$$

A policy ϕ_* is *average-cost optimal* if $w^{\phi_*}(x) = \inf_{\phi\in\mathbb{F}} w^{\phi}(x) =: w(x)$ for all $x \in \mathbb{X}$.

According to Rothblum [41], a stationary policy ϕ is called *normalized* if $\sum_{n=0}^{\infty} \beta^n Q_{\phi}^{n}$ converges for all $\beta \in (0,1)$. If Assumption T holds or the transition rates q are substochastic, then any stationary policy is normalized. Given $\beta \in [0,1)$ and an initial state $x \in \mathbb{X}$, the *β-discounted cost* incurred under a normalized policy $\phi \in \mathbb{F}$ is

$$v_{\beta}^{\phi}(x) := \sum_{n=0}^{\infty} \beta^n Q_{\phi}^{n} c_{\phi}(x).$$

A policy ϕ_* is *β-optimal* if $v_{\beta}^{\phi_*}(x) = \inf_{\phi\in\mathbb{F}} v_{\beta}^{\phi}(x) =: v_{\beta}(x)$ for all $x \in \mathbb{X}$.

In this article, transformations to discounted MDPs with stochastic transition rates are considered. Discounted MDPs with nonstochastic transition rates are mentioned only in Remark 9, where complexity estimates for discounted MDPs with transition rates satisfying Assumption T are provided.

3. UNDISCOUNTED TOTAL COSTS

The transformation of the original transient MDP to a discounted one, which we call the *Hoffman-Veinott (HV)* transformation, is given in Section 3.1. Under the hypotheses of Theorem 6 in Section 3.2, a stationary optimal policy exists for the transformed discounted MDP, and the sets of optimal policies for the transformed and original MDPs coincide. The finite state and action case is considered in Section 3.3. The Borel-state case is treated in Section 3.4.

3.1. HV Transformation

Let Assumption T hold. By Proposition 1, there is a non-negative function μ on $\mathbb{X}$ that satisfies $V \le \mu \le KV$ and (5). Objects associated with the discounted MDP will be indicated by a tilde. The state space is $\tilde{\mathbb{X}} := \mathbb{X} \cup \{\tilde{x}\}$, where $\tilde{x} \notin \mathbb{X}$ is a cost-free absorbing state. Letting $\tilde{a}$ denote the only action available at state $\tilde{x}$, the action space is $\tilde{\mathbb{A}} := \mathbb{A} \cup \{\tilde{a}\}$ and for $x \in \tilde{\mathbb{X}}$ the set of available actions is unchanged if $x \in \mathbb{X}$, namely

$$\tilde{A}(x) := \begin{cases} A(x), & \text{if } x \in \mathbb{X}, \\ \{\tilde{a}\}, & \text{if } x = \tilde{x}. \end{cases}$$

Define the one-step costs $\tilde{c}$ as

$$\tilde{c}(x,a) := \begin{cases} \mu(x)^{-1} c(x,a), & \text{if } x \in \mathbb{X},\ a \in A(x), \\ 0, & \text{if } (x,a) = (\tilde{x}, \tilde{a}). \end{cases}$$

To complete the definition of the discounted MDP, choose a discount factor

$$\tilde{\beta} \in \left[\frac{K-1}{K}, 1\right),$$

and let

$$\tilde{p}(y|x,a) := \begin{cases} \frac{1}{\tilde{\beta}\mu(x)} q(y|x,a)\mu(y), & \text{if } x, y \in \mathbb{X},\ a \in A(x), \\ 1 - \frac{1}{\tilde{\beta}\mu(x)} \sum_{y \in \mathbb{X}} q(y|x,a)\mu(y), & \\ \quad \text{if } y = \tilde{x},\ x \in \mathbb{X},\ a \in A(x), & \\ 1 \quad \text{if } y = x = \tilde{x},\ a = \tilde{a}. & \end{cases} \tag{11}$$

Note that $\tilde{p}(\cdot|x,a)$ is a probability distribution on $\tilde{\mathbb{X}}$ for each $x \in \tilde{\mathbb{X}}$ and $a \in \tilde{A}(x)$. Also, since $\tilde{A}(\tilde{x})$ is a singleton, the sets of policies for these two models coincide. Let $\tilde{v}^{\phi}_{\tilde{\beta}}(x)$ denote the $\tilde{\beta}$-discounted cost incurred under the policy ϕ when the initial state of this MDP is $x \in \tilde{\mathbb{X}}$, and let $\tilde{v}_{\tilde{\beta}}(x) = \inf_{\phi \in \mathbb{F}} \tilde{v}^{\phi}_{\tilde{\beta}}(x)$ for $x \in \tilde{\mathbb{X}}$.

3.1.1. *Relation to Veinott's positive similarity transformation*

Veinott's [52] positive similarity transformation is defined for transient MDPs with finite state and action sets as follows. Given a diagonal matrix B with positive diagonal entries, let

$$\tilde{c}_{\phi} := Bc_{\phi} \quad \text{and} \quad \tilde{P}_{\phi} := BQ_{\phi}B^{-1}, \quad \phi \in \mathbb{F}.$$

According to Veinott [52], properties that are invariant under this transformation include the transience of a policy, the optimality of a policy, and the geometric convergence of value iteration to the unique fixed point of the optimality operator. Further, letting μ be the unique vector satisfying

$$\mu(x) = \max_{\phi \in \mathbb{F}} [1 + Q_{\phi}\mu(x)], \quad x \in \mathbb{X}, \tag{12}$$

and letting $\mu(x)^{-1}$ be the nonzero entry on the x-th row of B, it follows from [52, Lemma 3] that if the spectral radii of the matrices Q_{ϕ} are all less than one, then the row sums of the matrices $\tilde{P}_{\phi}$ are all less than one; Veinott attributes this result to Alan Hoffman. The first line of (11) is an implementation of Veinott's similarity transformation that is applicable to all policies. Transformations of the form $\mu(x)^{-1} q(y|x,a)\mu(y)$ have also been used in the literature to reduce MDPs with unbounded one-step costs to MDPs with bounded one-step costs; see for example [51, p. 101].

3.2. Results

Given $\phi \in \mathbb{F}$, the following proposition relates the total costs incurred in the original undiscounted MDP with those incurred in the discounted MDP defined by the HV transformation.

PROPOSITION 3: Suppose statements (i) and (ii) of Assumption T hold. Then the one-step cost function $\tilde{c}$ is bounded and $v^{\phi}(x) = \mu(x)\tilde{v}^{\phi}_{\tilde{\beta}}(x)$ for each $\phi \in \mathbb{F}$ and $x \in \mathbb{X}$.

PROOF: Consider the matrix $\tilde{P}_{\phi}$ with elements $\tilde{P}_{\phi}(x,y) := \tilde{p}(y|x,\phi(x)),\ x, y \in \mathbb{X}$. Then

$$\tilde{v}^{\phi}_{\tilde{\beta}}(x) = \sum_{n=0}^{\infty} \tilde{\beta}^n \tilde{P}^n_{\phi} \tilde{c}_{\phi}(x), \quad x \in \tilde{\mathbb{X}}. \tag{13}$$

Since the state $\tilde{x}$ is cost-free and absorbing, it follows from the definitions of $\tilde{P}_{\phi}$ and $\tilde{c}_{\phi}$ that

$$\tilde{\beta}^n \tilde{P}^n_{\phi} \tilde{c}_{\phi}(x) = \mu(x)^{-1} Q^n_{\phi} c_{\phi}(x) \quad \text{for all } x \in \mathbb{X},\ n \ge 0. \tag{14}$$

Observe that, since μ majorizes V, according to (4) the mapping $x \mapsto \mu(x)^{-1} c_{\phi}(x)$ is bounded. Hence, combining (13) and (14), for $x \in \mathbb{X}$

$$\tilde{v}^{\phi}_{\tilde{\beta}}(x) = \mu(x)^{-1} \sum_{n=0}^{\infty} Q^n_{\phi} c_{\phi}(x) = \mu(x)^{-1} v^{\phi}(x).$$

Proposition 1 and Assumption T(ii) imply that $|\tilde{c}(x,a)| \le \bar{c}$ for all $x \in \mathbb{X}$ and $a \in A(x)$. □

The optimality results in this section and Section 4.2 rely on the following compactness-continuity conditions.

Compactness Conditions (cf. [17, p. 181]).

(i) *$A(x)$ is compact for each $x \in \mathbb{X}$;*
(ii) *$c(x, a)$ is lower semicontinuous in $a \in A(x)$ for each $x \in \mathbb{X}$;*
(iii) *the transition rates $q(y|x,a)$ are continuous in $a \in A(x)$ for each $x, y \in \mathbb{X}$;*
(iv) *the transition rates $q(\mathbb{X}|x,a) := \sum_{y\in\mathbb{X}} q(y|x,a)$ are continuous in $a \in A(x)$ for each $x \in \mathbb{X}$*

Observe that, if the state set is finite, then assumption (iii) of the Compactness Conditions implies assumption (iv). Also, if the transition rates are stochastic, that is, $q(\mathbb{X}|x,a) = 1$ for all $x \in \mathbb{X}$ and $a \in A(x)$, then assumption (iv) of the Compactness Conditions always holds.

LEMMA 4: Suppose Assumption T and the Compactness Conditions hold. Then the discounted MDP defined by the HV transformation also satisfies the Compactness Conditions.

PROOF: Assumptions (i–ii) of the Compactness Conditions imply that the sets $\tilde{A}(x)$ are compact and $\tilde{c}$ is bounded and is lower semicontinuous in a. In addition, assumption (iii) of the Compactness Conditions and Lemma 2 imply that $\tilde{p}(y|x,a)$ is continuous in $a \in A(x)$ for all $x, y \in \mathbb{X}$, and assumption (iv) implies that $\tilde{p}(\tilde{x}|x,a)$ is continuous in $a \in A(x)$ for all $x \in \mathbb{X}$. Since $\tilde{p}$ is also stochastic, it follows that the Compactness Conditions hold for the transformed MDP. □

The main result (Theorem 6) of this section relies on the following proposition. To state it, for $\beta \in [0,1)$ define

$$A^*_\beta(x) := \left\{ a \in A(x) \,\middle|\, v_\beta(x) = c(x,a) + \beta \sum_{y\in\mathbb{X}} q(y|x,a) v_\beta(y) \right\}, \quad x \in \mathbb{X}. \tag{15}$$

PROPOSITION 5 (cf. [17, pp. 181, 184]): If an MDP with transition probabilities q and bounded one-step costs c satisfies the Compactness Conditions, then for any discount factor $\beta \in [0,1)$:

(i) the value function v_β is the unique bounded function satisfying the optimality equation

$$v_\beta(x) = \min_{A(x)} \left[c(x,a) + \beta \sum_{y\in\mathbb{X}} q(y|x,a) v_\beta(y) \right], \quad x \in \mathbb{X}; \tag{16}$$

(ii) there is a stationary β-optimal policy;
(iii) a policy $\phi \in \mathbb{F}$ is β-optimal if and only if $\phi(x) \in A^*_\beta(x)$ for all $x \in \mathbb{X}$.

PROOF: The Compactness Conditions imply that, if $\mathbb{X}$ is endowed with the discrete topology, then the transition probabilities q are weakly continuous in (x, a) where $x \in \mathbb{X}$ and $a \in A(x)$. This implies that the MDP satisfies Assumption (W *) in [22]. The validity of (16) and statements (ii), (iii) follows from [22, Theorem 2]. The uniqueness claim in (i) follows from the contraction mapping principle; see Denardo [8] for details. □

To state Theorem 6, let

$$A^*(x) := \left\{ a \in A(x) \,\middle|\, v(x) = c(x,a) + \sum_{y\in\mathbb{X}} q(y|x,a) v(y) \right\}, \quad x \in \mathbb{X}, \tag{17}$$

where v is the value function of the original undiscounted total cost MDP.

THEOREM 6: Suppose the original undiscounted total-cost MDP satisfies Assumption T and the Compactness Conditions. Then:

(i) the value function $v = \mu \tilde{v}_{\tilde{\beta}}$ is the unique function satisfying the optimality equation

$$v(x) = \min_{A(x)} \left[c(x,a) + \sum_{y\in\mathbb{X}} q(y|x,a) v(y) \right], \quad x \in \mathbb{X}, \tag{18}$$

and such that

$$\|v\|_V := \sup_{x\in\mathbb{X}} V(x)^{-1} |v(x)| < \infty; \tag{19}$$

(ii) there is a stationary total-cost optimal policy;
(iii) a policy $\phi \in \mathbb{F}$ is total-cost optimal if and only if $\phi(x) \in A^*(x)$ for all $x \in \mathbb{X}$, and

$$A^*(x) = \left\{ a \in A(x) \mid \tilde{v}_{\tilde{\beta}}(x) = \tilde{c}(x,a) + \tilde{\beta} \sum_{y\in\tilde{\mathbb{X}}} \tilde{p}(y|x,a) \tilde{v}_{\tilde{\beta}}(y) \right\}, \quad x \in \mathbb{X}; \tag{20}$$

in other words, the sets of optimal actions for the original transient MDP and for the transformed discounted MDP with transition probabilities $\tilde{p}$ coincide.

PROOF: By Lemma 4, the transformed discounted MDP satisfies the Compactness Conditions. Hence statements (i–iii) of Proposition 5 hold for the transformed MDP.

Straightforward calculations show that the function $v = \mu\tilde{v}_{\tilde{\beta}}$ satisfies the optimality equation (18) if and only if the function $v_\beta := \tilde{v}_{\tilde{\beta}}$ satisfies the optimality equation (19) for the $\tilde{\beta}$-discounted MDP defined by the HV transformation. In view of Proposition 1, $\|v\|_V < \infty$ if and only if the function $\tilde{v}_{\tilde{\beta}}$ is bounded. Lemma 4 and Propositions 3, 5 imply statement (i).

According to Proposition 5(i), there is a $\phi_* \in \mathbb{F}$ that is $\tilde{\beta}$-optimal for the transformed MDP. By Proposition 3, $v^{\phi_*} = \mu\tilde{v}_{\tilde{\beta}}^{\phi_*} = \mu\tilde{v}_{\tilde{\beta}} = v$, so ϕ_* is total-cost optimal for the original MDP. Therefore (ii) holds.

It follows from the definitions of $\tilde{\mathbb{X}}$, $\tilde{A}$, $\tilde{c}$, $\tilde{\beta}$, and $\tilde{p}$ that (20) holds. Suppose $\phi \in \mathbb{F}$ is total-cost optimal for the original MDP. Then $v^\phi = v$, so since $v^\phi = c_\phi + Q_\phi v^\phi$ it follows that $\phi(x) \in A^*(x)$ for all $x \in \mathbb{X}$. Conversely, if $\phi(x) \in A^*(x)$ for all $x \in \mathbb{X}$, then according to Proposition 5(iii) and (37) the policy ϕ is $\tilde{\beta}$-optimal for the transformed MDP. By Proposition 3, this means ϕ is total-cost optimal for the original MDP. Hence (iii) holds. □

COROLLARY 7: Suppose Assumption T and the Compactness Conditions hold. If an algorithm computes an optimal policy for the discounted MDP defined by the HV transformation, then this policy is optimal for the original undiscounted total-cost MDP.

REMARK 4: The HV transformation also applies to arbitrary policies if the total costs are defined using the equivalent formulation in terms of transition probabilities and state-dependent discount factors; see Remark 1. Since stationary policies are optimal within the class of all policies for discounted MDPs with transition probabilities satisfying the Compactness Conditions [17, p. 184], the stationary total-cost optimal policies referred to in Theorem 6 are optimal over nonstationary policies as well.

3.3. Finite State and Action Sets

In this section, we assume that both $\mathbb{X}$ and $\mathbb{A}$ are finite. Recall from the paragraph after the statement of Assumption T that, when the state and action sets are finite, Assumption T is equivalent to the existence of a constant K such that

$$\sum_{n=0}^{\infty} Q_\phi^n e(x) \le K \quad \text{for all } \phi \in \mathbb{F},\ x \in \mathbb{X}, \tag{21}$$

where e denotes the function on $\mathbb{X}$ that is identically equal to one. Therefore, in this section we assume without loss of generality that (21) holds.

Corollary 7 implies that an optimal policy for the original transient MDP can be computed by solving the following linear program (LP):

$$\begin{aligned}
\text{minimize}\quad & \sum_{x\in\tilde{\mathbb{X}}}\sum_{a\in\tilde{A}(x)} \tilde{c}(x,a)\tilde{z}_{x,a}\\
\text{such that}\quad & \sum_{a\in\tilde{A}(x)} \tilde{z}_{x,a} - \tilde{\beta}\sum_{y\in\tilde{\mathbb{X}}}\sum_{a\in\tilde{A}(y)} \tilde{p}(x|y,a)\tilde{z}_{y,a} = 1,\\
& \quad x \in \tilde{\mathbb{X}},\\
& \tilde{z}_{x,a} \ge 0, \quad x \in \tilde{\mathbb{X}},\ a \in \tilde{A}(x).
\end{aligned} \tag{22}$$

According to Scherrer [46, Theorem 3], the LP (22) can be solved using

$$(m-n)\left\lceil \frac{1}{1-\tilde{\beta}}\log\frac{1}{1-\tilde{\beta}} \right\rceil = O((m-n)K\log K) \tag{23}$$

iterations of the block-pivoting simplex method corresponding to Howard's policy iteration algorithm. Alternatively, if the simplex method with Dantzig's rule is applied to the LP (22), then according to [46, Theorem 4] at most

$$n(m-n)\left(1 + \frac{2}{1-\tilde{\beta}}\log\frac{1}{1-\tilde{\beta}}\right) = O(n(m-n)K\log K) \tag{24}$$

iterations are needed to compute an optimal solution.

Let $z_{x,a} := \tilde{z}_{x,a}/\mu(x)$ for $x \in \mathbb{X}$ and $a \in A(x)$. The LP (22) for this discounted MDP can be written as

$$\begin{aligned}
\text{minimize}\quad & \sum_{x\in\mathbb{X}}\sum_{a\in A(x)} c(x,a)z_{x,a}\\
\text{such that}\quad & \sum_{a\in A(x)} z_{x,a} - \sum_{y\in\mathbb{X}}\sum_{a\in A(y)} q(x|y,a)z_{y,a} = \mu(x)^{-1},\\
& \quad x \in \mathbb{X},\\
& z_{x,a} \ge 0, \quad x \in \mathbb{X},\ a \in A(x).
\end{aligned} \tag{25}$$

This is true because of the following arguments that hold all $x \in \mathbb{X}$ and for all $a \in A(x)$: (i) the objective functions for the LPs (22) and (25) are equal because $\tilde{A}(x) = A(x)$,$c(x,a) = \mu(x)\tilde{c}(x,a)$,$\tilde{A}(\tilde{x}) = \{\tilde{x}\}$, and $\tilde{c}(\tilde{x},\tilde{a}) = 0$, (ii) for x, the equality constraints are equivalent in these LPs because $\tilde{p}(x|\tilde{x},\tilde{a}) = 0$,$q(y|x,a) = \tilde{\beta}\tilde{p}(y|x,a)$, where $y \in \mathbb{X}$, and the inequality constraints are equivalent because $\mu(x) > 0$, (iii) the equality and inequality constraints for $\tilde{x}$ can be excluded from the LP (22) because the former implies that $\tilde{z}_{\tilde{x},\tilde{a}} = (1-\tilde{\beta})^{-1}(1 + \tilde{\beta}\sum_{y\in\mathbb{X}}\sum_{a\in A(y)} \tilde{p}(\tilde{x}|y,a)\tilde{z}_{y,a}) > 0$ and, in view of (i) and (ii), the variable $\tilde{z}_{\tilde{x},\tilde{a}}$ does not appear anywhere else in the LP (22).

Since a policy for this discounted MDP is optimal if and only if it is optimal for the original discounted MDP defined by the HV transformation, Corollary 7 implies that an optimal policy for the original transient MDP can be computed by solving the LP (25). Since any optimal policy derived from the LP (25) is still optimal if the right-hand sides of the equality constraints are replaced by arbitrary positive numbers (see e.g., [33, Corollary 3, Remark 6]), it follows that an optimal policy for the original transient MDP can be computed by solving the LP

$$\begin{aligned} \text{minimize} \quad & \sum_{x\in\mathbb{X}}\sum_{a\in A(x)} c(x,a)z_{x,a} \\ \text{such that} \quad & \sum_{a\in A(x)} z_{x,a} - \sum_{y\in\mathbb{X}}\sum_{a\in A(y)} q(x|y,a)z_{y,a} = 1, \\ & \quad x \in \mathbb{X}, \\ & z_{x,a} \geq 0, \quad x\in\mathbb{X},\ a\in A(x). \end{aligned} \tag{26}$$

This provides an alternative derivation of the LP (26) for transient MDPs provided in Denardo [9], where it is shown that the LP (26) can be solved using at most $(m-n)k^*$ iterations of the block-pivoting simplex method corresponding to Howard's policy iteration algorithm, where k^* is the smallest integer k that satisfies $1 > K(1-(1/K))^k$ [9, Theorem 2]. This implies that the required number of iterations is $O((m-n)K\log K)$, which matches the estimate (23) for the LP (22) obtained using [46, Theorem 3]. If $K=1$, then $\tilde{\beta} = 0$ and the problem can be solved by simply selecting, for each $x \in \mathbb{X}$, an action minimizing $c(x, a)$ over $a \in A(x)$. Denardo [9] also showed that the LP (26) can be solved using at most $(m-n)j^*$ iterations of the simplex method with Dantzig's rule, where j^* is the smallest integer j that satisfies $1 > (e\tau)(1-(1/e\tau))^j$ and τ is the function defined in Remark 2. This implies that the simplex method with Dantzig's rule requires at most

$$O((m-n)(e\tau)\log(e\tau)) \tag{27}$$

iterations to solve the LP (26).

REMARK 5: Applying the simplex method with Dantzig's rule to (22) can be viewed as applying a certain pivoting rule to the LP (26). In particular, given a non-optimal basic feasible solution to (26) corresponding to the non-optimal stationary policy ϕ, the variable $z_{x,a}$ that enters the basis under this pivoting rule is the one minimizing

$$\tilde{c}(x,a) + \tilde{\beta}\sum_{y\in\mathbb{X}}\tilde{p}(y|x,a)\tilde{v}^{\phi}_{\tilde{\beta}}(y) - v^{\phi}_{\tilde{\beta}}(x)$$
$$= \frac{1}{\mu(x)}\left[c(x,a) + \sum_{y\in\mathbb{X}} q(y|x,a)v^{\phi}(y) - v^{\phi}(x)\right], \tag{28}$$

where the expression in the square brackets on the right-hand side of (28) is precisely the reduced cost, for the variable $z_{x,a}$, associated with the basis corresponding to ϕ. It follows from (24) that this pivoting rule for the LP (26) considered in Denardo [9] is strongly polynomial when K is fixed. This algorithm is not the same as applying Dantzig's rule to the LP (26), however; see Remark 8 below.

REMARK 6: To compare the estimates (24) and (27) for the simplex method with Dantzig's rule for LPs (22) and (26) respectively, consider the functions $f(n,K) := nK\log K$ and $g(e\tau) := e\tau\log(e\tau)$. Using these notations, the estimate (24) is

$$O((m-n)f(n,K)), \tag{29}$$

and the estimate (27) is

$$O((m-n)g(e\tau)). \tag{30}$$

If K is fixed, then the estimate (29) is better than the estimate (30). This is because, when K fixed, $f(n,K) = O(n)$ while $nK \geq e\tau \geq n-1+K$ implies that $g(e\tau) = O(n\log n)$. In addition, if $\tau \equiv K$, then the estimate (29) is also better than (30) because $g(e\tau) = nK\log nK \geq nK\log K = f(n,K)$. On the other hand, for some particular values of n, K, and τ it is possible that $f(n,K) > g(e\tau)$. For example, consider the MDP with $n=10$ states and 1 action per state, where for states 1 through 9 the process stops after one transition, and for state 10 the process stops with probability 1/5 and continues with probability 4/5. Then $K=5$ and $e\tau = 9+5 = 14$, which implies $f(n,K) \geq 10\cdot 5\cdot\log(5) > 14\cdot\log(14) = g(e\tau)$.

REMARK 7: Consider Howard's policy iteration algorithm for the discounted MDP defined by the HV transformation, which according to [35, p. 68] is equivalent to a block-pivoting simplex method for the LP (22). Given $\phi \in \mathbb{F}$ and recalling that $\tilde{x}$ is a cost-free absorbing state, an improved policy $\phi^+ \in \mathbb{F}$ is constructed (when possible) as follows. For each $x \in \mathbb{X}$, $\phi^+(x)$ is taken to be any action belonging to

$$\operatorname*{arg\,min}_{a\in A(x)}\left[\tilde{c}(x,a) + \tilde{\beta}\sum_{y\in\mathbb{X}}\tilde{p}(y|x,a)\tilde{v}^{\phi}_{\tilde{\beta}}(y)\right]. \tag{31}$$

It follows from the definitions of $\tilde{c}$ and $\tilde{p}$ and Proposition 3 that for each $x \in \mathbb{X}$, the set (31) is equal to

$$\operatorname*{arg\,min}_{a\in A(x)}\left[c(x,a) + \sum_{y\in\mathbb{X}} q(y|x,a)v^{\phi}(y)\right]. \tag{32}$$

Under Howard's policy iteration algorithm for the original transient MDP, which according to the arguments in [35,

p. 68] and [33, pp. 55–56] is equivalent to a block-pivoting simplex method for the LP (26), given $\phi \in \mathbb{F}$ an improved policy ϕ^+ is constructed (when possible) by taking, for each $x \in \mathbb{X}$, $\phi^+(x)$ to be any action belonging to (32). Since for each $x \in \mathbb{X}$ the sets (31) and (32) are equal, it follows that there is a one-to-one correspondence between sequences of policies generated by Howard's policy iteration algorithm for the discounted MDP defined by the HV transformation, and sequences of policies generated by Howard's policy iteration algorithm for the original transient MDP. Using Scherrer's [46, Theorem 3] $O(mK \log K)$ iteration bound for Howard's policy iteration algorithm for discounted MDPs, we therefore obtain the bound derived by Denardo [9, Theorem 2] for the original transient MDP.

REMARK 8: According to Remark 7, starting with the same basic variables, the sequences of basic variables for implementations of block-pivoting simplex methods for the LPs (22) and (26) coincide. This is not true for the simplex method with Dantzig's rule, however. To confirm this, consider the following transient MDP. The set of states is $\mathbb{X} = \{1, 2\}$, and the sets of available actions are $A(1) = A(2) = \{a, b\}$. The transition rates are $q(1|1, a) = 2/3$, $q(2|1, a) = 1/6$, $q(1|1, b) = q(2|1, b) = 1/3$, $q(1|2, a) = 2/3$, $q(2|2, a) = 1/6$, $q(1|2, b) = 1/12$, and $q(2|2, b) = 5/6$. The one-step costs are $c(1, a) = -0.91$, $c(1, b) = -0.56$, $c(2, a) = -0.19$, and $c(2, b) = -0.8$. One can verify that the function μ defined by $\mu(1) = 8$ and $\mu(2) = 10$ satisfies (5) with $V \equiv 1$. The total-cost LP given by (26) is

$$\begin{aligned} \text{minimize} \quad & -0.91z_{1,a} - 0.56z_{1,b} - 0.19z_{2,a} - 0.8z_{2,b} \\ \text{such that} \quad & \frac{1}{3}z_{1,a} + \frac{2}{3}z_{1,b} - \frac{2}{3}z_{2,a} - \frac{1}{12}z_{2,b} = 1 \\ & -\frac{1}{6}z_{1,a} - \frac{1}{3}z_{1,b} + \frac{5}{6}z_{2,a} + \frac{1}{6}z_{2,b} = 1 \\ & z_{1,a}, z_{1,b}, z_{2,a}, z_{2,b} \geq 0, \end{aligned} \tag{33}$$

and, letting $K = 10$, the LP (22) for the discounted MDP defined by the HV transformation with discount factor $\tilde{\beta} = (K - 1)/K = 9/10$ is

$$\begin{aligned} \text{minimize} \quad & -0.11375\tilde{z}_{1,a} - 0.07\tilde{z}_{1,b} - 0.019\tilde{z}_{2,a} - 0.08\tilde{z}_{2,b} \\ \text{such that} \quad & \frac{1}{3}\tilde{z}_{1,a} + \frac{2}{3}\tilde{z}_{1,b} - \frac{8}{15}\tilde{z}_{2,a} - \frac{1}{15}\tilde{z}_{2,b} = 1 \\ & -\frac{5}{24}\tilde{z}_{1,a} - \frac{5}{12}\tilde{z}_{1,b} + \frac{5}{6}\tilde{z}_{2,a} + \frac{1}{6}\tilde{z}_{2,b} = 1 \\ & \tilde{z}_{1,a}, \tilde{z}_{1,b}, \tilde{z}_{2,a}, \tilde{z}_{2,b} \geq 0, \end{aligned} \tag{34}$$

where, according to the remarks following (25), the variable $\tilde{z}_{\bar{x},\bar{a}}$ has been removed. For both LPs, suppose the initial basic feasible solution for the simplex method with Dantzig's rule is the one defined by the stationary policy ϕ with $\phi(1) = b$ and $\phi(2) = a$; namely, for both LPs the basic variables are those corresponding to the state-action pairs $(1, b)$ and $(2, a)$. Consider the first iterations of the simplex method with Dantzig's rule for the LPs (33) and (34). For the LP (33), the basic variable $z_{2,a}$ is the unique variable to leave the basis, while for the LP (34) the basic variable $\tilde{z}_{1,b}$ is the unique variable to leave the basis.

REMARK 9: If (21) holds, it holds with the same upper bound K if the transition rates q are replaced with the transition rates βq, where $\beta \in (0, 1]$. Hence the results in Denardo [9, Theorems 1, 2] and the estimates (23), (24) above imply that the number of arithmetic operations needed to compute an optimal policy for a discounted MDP satisfying Assumption T can be bounded by a polynomial in m that does not depend on the discount factor $\beta \in (0, 1]$. In particular, these bounds hold for all discount factors $\beta \in (0, 1]$. If $\beta = 0$, the discounted problem becomes a one-step problem, which is equivalent to a problem with $K = 1$; this case was discussed in the paragraph preceding Remark 6.

REMARK 10: For $x \in \mathbb{X}$, let $\tau(x) := \sup_{\phi \in \mathbb{F}} \sum_{n=0}^{\infty} Q_\phi^n e(x)$. Then $K_\tau := \max_{x \in \mathbb{X}} \tau(x)$ is the smallest constant K satisfying (21). The natural question is how to compute K_τ. One method to compute K_τ consists in the following. First, compute an optimal policy ϕ_* for a transient MDP that is identical to the original MDP except that all one-step costs are equal to –1. Then, compute the value function v^{ϕ_*} of this optimal policy, and set $K_\tau = \max_{x \in \mathbb{X}} v^{\phi_*}(x)$. As discussed in the paragraph following (26), the policy ϕ_* can be computed using $O((m - n)K_\tau \log K_\tau)$ iterations of Howard's policy iteration algorithm. Further, the function v^{ϕ_*} can be computed by solving a system of n linear equations using Gaussian elimination in $O(n^3)$ arithmetic operations; for other methods see for example [5, 48].

3.4. Extension to Uncountable State Spaces

In this section, we assume that the state space $\mathbb{X}$ is a Borel subset of a complete separable metric space, and that the transition rates are defined by a Borel-measurable transition kernel q on $\mathbb{X}$ given $\mathrm{Gr}(A) := \{(x, a) : x \in \mathbb{X},\ a \in A(x)\}$, which we assume to be a Borel subset of $\mathbb{X} \times \mathbb{A}$. That is, $q(\cdot|x, a)$ is a finite measure for every $(x, a) \in \mathrm{Gr}(A)$, and $q(B|\cdot)$ is a Borel-measurable function on $\mathrm{Gr}(A)$ for every Borel subset B of $\mathbb{X}$. In addition, the one-step cost function $c : \mathrm{Gr}(A) \to \mathbb{R}$ is Borel-measurable.

The set of stationary policies $\mathbb{F}$ is identified with the set of all Borel-measurable functions $\phi : \mathbb{X} \to \mathbb{A}$ satisfying $\phi(x) \in A(x)$ for all $x \in \mathbb{X}$. To formulate a version of Assumption T in this setting, for $\phi \in \mathbb{F}$ define the operator

Q_ϕ for Borel-measurable functions $u : \mathbb{X} \to \mathbb{R}$ by

$$Q_\phi u(x) := \int_{\mathbb{X}} u(y) q(dy|x, \phi(x)), \quad x \in \mathbb{X}, \tag{35}$$

and given a Borel-measurable weight function $W : \mathbb{X} \to \mathbb{R}$ and a Borel-measurable transition kernel $B(\cdot|\cdot)$ on $\mathbb{X}$ given $\mathbb{X}$, let

$$\|B\|_W := \sup_{x \in \mathbb{X}} W(x)^{-1} \int_{\mathbb{X}} W(y) B(dy|x).$$

ASSUMPTION T′:

(i) There is a Borel-measurable weight function $V : \mathbb{X} \to [1, \infty)$ and a constant $K \geq 1$ that satisfy

$$\left\| \sum_{n=0}^{\infty} Q_\phi^n \right\|_V \leq K < \infty \quad \text{for all } \phi \in \mathbb{F}. \tag{36}$$

(ii) Moreover, there is a constant $\bar{c} < \infty$ satisfying

$$\sup_{a \in A(x)} |c(x,a)| \leq \bar{c} V(x) \quad \text{for all } x \in \mathbb{X},$$

and for every $x \in \mathbb{X}$ the mapping

$$a \mapsto \int_{y \in \mathbb{X}} V(y) q(y|x,a) < \infty, \quad a \in A(x),$$

is continuous on $A(x)$.

To obtain a reduction to a discounted MDP, we consider the following setwise-continuity and compactness conditions:

ASSUMPTION S:

(a) Statements (i) and (ii) of the Compactness Conditions hold.
(b) For every $x \in \mathbb{X}$, if the sequence $\{a_n\}$ in $A(x)$ converges to $a \in A(x)$, then for every Borel subset B *of* $\mathbb{X}$ the sequence $\{q(B|x,a_n)\}$ converges to $q(B|x,a)$

PROPOSITION 8: Suppose Assumption S holds. Then Assumption T ′ (i) holds if and only if there is a Borel-measurable function $\mu : \mathbb{X} \to [1, \infty)$ satisfying $V(x) \leq \mu(x) \leq K V(x)$ and

$$\mu(x) \geq V(x) + \int_{\mathbb{X}} \mu(y) q(dy|x,a), \quad (x,a) \in \mathrm{Gr}(A).$$

PROOF: This follows from the proof of Proposition 1, with all sums replaced with integrals, and by applying the Brown and Purves [4, Corollary 1] theorem on Borel-measurable selection. □

In this setting, the analogue of Lemma 2 holds as well.

LEMMA 9: Suppose Assumption S and statements (i) and (iii) of T ′ hold, and let μ be the Borel-measurable function described in the statement of Proposition 8. Then for every $x \in \mathbb{X}$ the mapping

$$a \mapsto \int_{\mathbb{X}} \mu(y) q(dy|x,a), \quad a \in A(x),$$

is continuous on $A(x)$.

PROOF: This follows from the proof of Lemma 2, where all sums are replaced with integrals. □

3.4.1. HV Transformation

Let $\mathcal{B}(\mathbb{X})$ denote the Borel σ-algebra of $\mathbb{X}$. The definition of the HV transformation in the setting of a possibly uncountable state space is identical to the definition presented in Section 3.1, except that the cost-free absorbing state $\tilde{x}$ is taken to be isolated from the original state space $\mathbb{X}$, and the transition probability kernel $\tilde{p}$ is defined by

$$\tilde{p}(B|x,a) := \begin{cases} \frac{1}{\tilde{\beta}\mu(x)} \int_B \mu(y) q(dy|x,a), \\ \quad \text{if } B \in \mathcal{B}(\mathbb{X}),\ (x,a) \in \mathrm{Gr}(A), \\ 1 - \frac{1}{\tilde{\beta}\mu(x)} \int_{\mathbb{X}} \mu(y) q(dy|x,a), \\ \quad \text{if } B = \{\tilde{x}\},\ (x,a) \in \mathrm{Gr}(A), \\ 1, \quad \text{if } B = \{\tilde{x}\},\ (x,a) = (\tilde{x}, \tilde{a}). \end{cases}$$

3.4.2. Results

PROPOSITION 10: Suppose Assumptions S and T ′ hold. Then $v^\phi(x) = \mu(x) \tilde{v}^\phi_{\tilde{\beta}}(x)$ for each $\phi \in \mathbb{F}$ and $x \in \mathbb{X}$.

PROOF: This follows from the proof of Proposition 3 by defining for $\phi \in \mathbb{F}$ the operator $\tilde{P}_\phi$ applied to integrable Borel-measurable functions $u : \mathbb{X} \to \mathbb{R}$,

$$\tilde{P}_\phi u(x) := \int_{\tilde{\mathbb{X}}} u(y) \tilde{p}(dy|x, \phi(x)), \quad x \in \mathbb{X}. \qquad \square$$

LEMMA 11: Suppose Assumptions S and T′ hold. Then the discounted MDP defined by the HV transformation also satisfies Assumption S.

PROOF: This follows from the proof of Lemma 4, Lemma 9, and the fact that the added cost-free absorbing state $\tilde{x}$ is isolated from $\mathbb{X}$. □

The special case of Theorem 12 below for $V \equiv 1$ was proved by Pliska [38, Theorem 1.3]. To state Theorem 12, for $\beta \in (0,1)$ and $x \in \mathbb{X}$, define the sets $A^*_\beta(x)$ and $A^*(x)$ by replacing the sums in (15) and (17), respectively, with integrals.

THEOREM 12: Suppose the original undiscounted total-cost MDP satisfies Assumptions S and T′. Then:

(i) the value function $v = \mu\tilde{v}_{\tilde{\beta}}$ is the unique Borel-measurable function satisfying the optimality equation

$$v(x) = \min_{A(x)}\left[c(x,a) + \int_{\mathbb{X}} v(y)q(dy|x,a)\right],\quad x \in \mathbb{X},$$

and such that

$$\sup_{x\in\mathbb{X}} V(x)^{-1}|v(x)| < \infty;$$

(ii) there is a stationary total-cost optimal policy;

(iii) a policy $\phi \in \mathbb{F}$ is total-cost optimal if and only if $\phi(x) \in A^*(x)$ for all $x \in \mathbb{X}$, and

$$A^*(x) = \left\{a \in A(x) \,\middle|\, \tilde{v}_{\tilde{\beta}}(x) = \tilde{c}(x,a) + \tilde{\beta}\int_{\mathbb{X}} \tilde{v}_{\tilde{\beta}}(y)\tilde{p}(dy|x,a)\right\},\quad x \in \mathbb{X}; \tag{37}$$

in other words, the sets of optimal actions for the original transient MDP and for the transformed discounted MDP with transition probabilities $\tilde{p}$ coincide.

PROOF: This follows from the proof of Theorem 6, where instead of [22] one can use [45, Proposition 2.1]. □

4. AVERAGE COSTS PER UNIT TIME

In Section 4.1, we provide a slight modification of the transformation introduced by Akian and Gaubert [1]. Since it can be viewed as an extension of the HV transformation described in Section 3.1, we refer to the transformation given in Section 4.1 as the *HV-AG* transformation. Like the HV transformation, the HV-AG transformation produces a discounted MDP with transition probabilities. According to Theorem 16 in Section 4.2, for an average-cost MDP with transition probabilities q satisfying Assumption HT and the Compactness Conditions given in Section 3.2, the HV-AG transformation reduces the original problem to a discounted one. The finite state and action case is considered in Section 4.3. The Borel-state case is treated in Section 4.4.

4.1. HV-AG Transformation

Suppose Assumption HT holds. According to Proposition 1, there is a function $\mu : \mathbb{X} \to [1,\infty)$ that satisfies $\mu \le K^*$ and

$$\mu(x) \ge 1 + \sum_{y\in\mathbb{X}\setminus\{\ell\}} q(y|x,a)\mu(y),\quad x \in \mathbb{X},\ a \in A(x). \tag{38}$$

Objects associated with the discounted MDP will be indicated by a horizontal bar. The state space is $\overline{\mathbb{X}} := \mathbb{X} \cup \{\bar{x}\}$, where $\bar{x} \notin \mathbb{X}$ is a cost-free absorbing state. Letting $\bar{a}$ denote the only action available at state $\bar{x}$, the action space is $\overline{\mathbb{A}} := \mathbb{A} \cup \{\bar{a}\}$ and for $x \in \overline{\mathbb{X}}$ the set of available actions is unchanged if $x \in \mathbb{X}$, namely

$$\bar{A}(x) := \begin{cases} A(x), & \text{if } x \in \mathbb{X},\\ \{\bar{a}\}, & \text{if } x = \bar{x}.\end{cases}$$

Define the one-step costs $\bar{c}$ by

$$\bar{c}(x,a) := \begin{cases} \mu(x)^{-1}c(x,a), & \text{if } x \in \mathbb{X},\ a \in A(x),\\ 0, & \text{if } (x,a) = (\bar{x},\bar{a}).\end{cases}$$

To complete the definition of the discounted MDP, choose a discount factor

$$\overline{\beta} \in \left[\frac{K^*-1}{K^*}, 1\right),$$

and let

$$\bar{p}(y|x,a) := \begin{cases} \frac{1}{\overline{\beta}\mu(x)}q(y|x,a)\mu(y),\\ \quad y \in \mathbb{X}\setminus\{\ell\},\ x \in \mathbb{X},\ a \in A(x),\\ \frac{1}{\overline{\beta}\mu(x)}[\mu(x) - 1 - \sum_{y\in\mathbb{X}\setminus\{\ell\}} q(y|x,a)\mu(y)],\\ \quad y = \ell,\ x \in \mathbb{X},\ a \in A(x)\\ 1 - \frac{1}{\overline{\beta}\mu(x)}[\mu(x) - 1],\\ \quad y = \bar{x},\ x \in \mathbb{X},\ a \in A(x)\\ 1,\quad y = \bar{x},\ (x,a) = (\bar{x},\bar{a}).\end{cases}$$

Since μ satisfies (5), $\bar{p}(\cdot|x,a)$ is a probability distribution on $\mathbb{X}$ for each $x \in \overline{\mathbb{X}}$ and $a \in \bar{A}(x)$. In addition, the definition of $\bar{A}$ implies that the sets of policies for the transformed MDP and the original MDP coincide. Let $\bar{v}^{\phi}_{\overline{\beta}}(x)$ denote the $\overline{\beta}$-discounted cost incurred when the initial state of the transformed MDP is $x \in \overline{\mathbb{X}}$ and the policy ϕ is used, and let $\bar{v}_{\overline{\beta}}(x) := \inf_{\phi\in\mathbb{F}}\bar{v}^{\phi}_{\overline{\beta}}(x)$ for $x \in \overline{\mathbb{X}}$.

REMARK 11: While the HV-AG transformation applies to transition rates in general, the major results in Section 4.2 pertain to the case when these rates are probabilities.

REMARK 12: Akian and Gaubert [1] prove their results by transforming a perfect-information mean-payoff stochastic game into a discounted game with state-dependent discount factors. The version of their transformation presented above uses techniques from [18] to directly obtain a problem with a single discount factor.

REMARK 13: Ross [39, 40] considered MDPs with transition probabilities q satisfying the special case of Assumption HT where there is a constant α such that

$$q(\ell|x,a) \geq \alpha > 0 \quad \text{for all } x \in \mathbb{X},\ a \in A(x),$$

and introduced a transformation of the transition probabilities that can be used to reduce the average-cost MDP to a discounted one. In fact, Ross's [39, 40] transformation can be viewed as a special case of the HV-AG transformation. Namely, taking $\mu \equiv K = 1/\alpha$, the resulting transition probabilities are the same in both cases and the one-step costs differ by a factor of α.

REMARK 14: The HV-AG transformation does not apply to the version of Assumption HT with the norm $\|\cdot\|$ being replaced with $\|\cdot\|_V$, when V is unbounded. In particular, $\bar{p}(\bar{x}|x,a) \geq 0$ implies that $\mu(x) \leq (1-\overline{\beta})^{-1}$.

4.2. Results

The proofs of Proposition 14 and Theorem 16 below rely on the following lemma.

LEMMA 13: If a bounded function $f : \overline{\mathbb{X}} \to \mathbb{R}$ satisfies $f(\bar{x}) = 0$, then for all $x \in \mathbb{X}$ and $a \in A(x)$

$$\bar{c}(x,a) + \overline{\beta} \sum_{y \in \overline{\mathbb{X}}} \bar{p}(y|x,a) f(y) = \frac{1}{\mu(x)} \left[c(x,a) + \sum_{y \in \mathbb{X}} q(y|x,a)\mu(y)[f(y) - f(\ell)] + [\mu(x) - 1] f(\ell) \right]. \tag{39}$$

PROOF: According to the definition of $\bar{c}$, $\overline{\beta}$, and $\bar{p}$ in Section 4.1, for $x \in \mathbb{X}$ and $a \in A(x)$

$$\begin{aligned}
\bar{c}(x,a) &+ \overline{\beta} \sum_{y \in \overline{\mathbb{X}}} \bar{p}(y|x,a) f(y) \\
&= \frac{c(x,a)}{\mu(x)} + \frac{1}{\mu(x)} \sum_{y \in \mathbb{X} \setminus \{\ell\}} q(y|x,a)\mu(y) f(y) \\
&\quad + \frac{1}{\mu(x)} \left[\mu(x) - 1 - \sum_{\mathbb{X} \setminus \{\ell\}} q(y|x,a)\mu(y) \right] f(\ell) \\
&= \frac{1}{\mu(x)} \left[c(x,a) + \sum_{y \in \mathbb{X}} q(y|x,a)\mu(y)[f(y) - f(\ell)] + [\mu(x) - 1] f(\ell) \right].
\end{aligned}$$

□

Given $\phi \in \mathbb{F}$, the following proposition relates the average costs incurred in the original MDP with the discounted costs incurred in the MDP constructed using the HV-AG transformation. Recall that q is *stochastic* if $\sum_{y \in \mathbb{X}} q(y|x,a) = 1$ for all $x \in \mathbb{X}$ and $a \in A(x)$.

PROPOSITION 14: Suppose Assumption HT holds. Let $\phi \in \mathbb{F}$ be a stationary policy and $h^\phi(x) := \mu(x)[\bar{v}^{\phi}_{\overline{\beta}}(x) - \bar{v}^{\phi}_{\overline{\beta}}(\ell)]$ for $x \in \mathbb{X}$. Then

$$\bar{v}^{\phi}_{\overline{\beta}}(\ell) + h^\phi(x) = c(x,\phi(x)) + \sum_{y \in \mathbb{X}} q(y|x,\phi(x)) h^\phi(y), \quad x \in \mathbb{X}. \tag{40}$$

In addition, if the transition rates q are stochastic, then $w^\phi \equiv \bar{v}^{\phi}_{\overline{\beta}}(\ell)$.

PROOF: Since the state $\bar{x}$ in the discounted MDP defined by the HV-AG transformation is cost-free and absorbing, (40) follows from the fact that

$$\bar{v}^{\phi}_{\overline{\beta}}(x) = \bar{c}(x,\phi(x)) + \overline{\beta} \sum_{y \in \overline{\mathbb{X}}} \bar{p}(y|x,\phi(x)) \bar{v}^{\phi}_{\overline{\beta}}(y), \quad x \in \mathbb{X},$$

and Lemma 13. Iterating (40) gives

$$N \bar{v}^{\phi}_{\overline{\beta}}(\ell) + h^\phi(x) = \sum_{n=0}^{N-1} Q^n_\phi c_\phi(x) + Q^N_\phi h^\phi(x), \quad x \in \mathbb{X},\ N = 1, 2, \ldots. \tag{41}$$

Since c is bounded, the function h^ϕ is bounded as well. The equality $w^\phi \equiv \bar{v}^{\phi}_{\overline{\beta}}(\ell)$ then follows by dividing both sides of (41) by N and letting $N \to \infty$. □

LEMMA 15: Suppose Assumption HT and the Compactness Conditions hold. Then the discounted MDP defined by the HV-AG transformation also satisfies the Compactness Conditions.

PROOF: Assumptions (i–ii) of the Compactness Conditions imply that the sets $\bar{A}(x)$ are compact and $\bar{c}$ is bounded and is lower semicontinuous in a. Assumption (iii) of the Compactness Conditions and Lemma 2 imply that $\bar{p}(y|x,a)$ is continuous in $a \in A(x)$ for all $x \in \mathbb{X}$ and $y \in \mathbb{X} \setminus \{\ell\}$. Assumption (iii), for state ℓ, and assumption (iv) of the Compactness Conditions imply that $\bar{p}(\ell|x,a)$ is continuous in $a \in A(x)$ for all $x \in \mathbb{X}$. □

For $x \in \mathbb{X}$, and a constant w and function $h : \mathbb{X} \to \mathbb{R}$ satisfying the average-cost optimality equation (43) given in the statement of Theorem 16 below, consider the sets of actions

$$A^*_{\text{av}}(x) := \Big\{ a \in A(x) \mid w + h(x) = c(x,a) + \sum_{y \in \mathbb{X}} q(y|x,a)h(y) \Big\}, \quad x \in \mathbb{X}. \tag{42}$$

Theorem 16 also follows from Federgruen and Tijms [16, Theorems 2.1, 2.2], where other recurrence conditions are considered as well.

THEOREM 16: Suppose the original MDP with transition probabilities q satisfies Assumption HT and the Compactness Conditions. Then:

(i) the constant $w = \bar{v}_{\overline{\beta}}(\ell)$ and the function $h(x) = \mu(x)[\bar{v}_{\overline{\beta}}(x) - \bar{v}_{\overline{\beta}}(\ell)]$, $x \in \mathbb{X}$, satisfy the optimality equation

$$w + h(x) = \min_{A(x)} \left[c(x,a) + \sum_{y \in \mathbb{X}} q(y|x,a)h(y) \right], \quad x \in \mathbb{X}, \tag{43}$$

and $\bar{v}_{\overline{\beta}}(\ell)$ is the optimal average cost for each initial state.

(ii) there is a $\phi \in \mathbb{F}$ satisfying $\phi(x) \in A^*_{\text{av}}(x)$ for all $x \in \mathbb{X}$, where

$$A^*_{\text{av}}(x) = \Big\{ a \in A(x) \mid \bar{v}_{\overline{\beta}}(x) = \bar{c}(x,a) + \overline{\beta} \sum_{y \in \overline{\mathbb{X}}} \bar{p}(y|x,a)\bar{v}_{\overline{\beta}}(y) \Big\}, \quad x \in \mathbb{X}, \tag{44}$$

and any such policy is average-cost optimal.

PROOF: Lemma 15 implies that statements (i–iii) of Proposition 5 hold for the transformed MDP. In particular, there is a stationary $\overline{\beta}$-optimal policy ϕ for the transformed MDP, which satisfies $\phi(x) \in A^*_{\overline{\beta}}(x)$ for all $x \in \mathbb{X}$.

The validity of (43) follows from applying Lemma 13 to the optimality equation for the $\overline{\beta}$-discounted MDP defined by the HV-AG transformation. Further, Proposition 14 implies that the optimal average cost for each state is $\bar{v}_{\overline{\beta}}(\ell)$, so (i) holds.

Lemma 13 implies that (44) holds, from which the existence of a $\phi \in \mathbb{F}$ satisfying $\phi(x) \in A^*_{\text{av}}(x)$ for all $x \in \mathbb{X}$ follows. Moreover, since the function h is bounded,

$$\lim_{N \to \infty} \frac{1}{N} \mathbb{E}^{\phi}_x h(x_N) = 0 \quad \text{for all } x \in \mathbb{X}.$$

It therefore follows from for example [27, Theorem 5.2.4] that any $\phi \in \mathbb{F}$ satisfying $\phi(x) \in A^*_{\text{av}}(x)$ for all $x \in \mathbb{X}$ is average-cost optimal. □

COROLLARY 17: Suppose Assumption HT and the Compactness Conditions hold. If an algorithm computes an optimal policy for the discounted MDP defined by the HV-AG transformation, then this policy is optimal for the original average-cost MDP.

REMARK 15: The average-cost optimal policy referred to in Theorem 16 is in fact optimal over all randomized history-dependent policies; see for example, Hernández-Lerma and Lasserre [27, Theorem 5.2.4].

REMARK 16: Stationary average-cost optimal policies exist under much more general conditions than the ones considered in Theorem 16. In particular, the Compactness Conditions and Assumption HT imply Conditions (S) and (B) in Schäl [45], as well as Assumptions (W*) and (B) in Feinberg et al. [22].

REMARK 17: Under the hypotheses of Theorem 16, the average-cost optimality equation (43) has a unique bounded solution up to an additive constant; see [6, Lemma 3.3]. This is because Assumption HT is a special case of the more general weighted geometric ergodicity condition considered in [6]; see [7] for relationships between this condition and various other ergodicity and recurrence assumptions.

4.3. Finite State and Action Sets

In this section, we assume that both $\mathbb{X}$ and $\mathbb{A}$ are finite. Recall from the paragraph after Remark 3 that, when the state and action sets are finite, Assumption HT is equivalent to the existence of a constant K^* such that

$$\sum_{n=0}^{\infty} {}_{\ell}Q^n_{\phi} e(x) \le K^* \quad \text{for all } \phi \in \mathbb{F},\ x \in \mathbb{X}, \tag{45}$$

where e denotes the function on $\mathbb{X}$ that is identically equal to one. Therefore, in this section we assume without loss of generality that (45) holds.

For a finite state and action MDP with transition probabilities q that satisfy Assumption HT, Corollary 17 implies that a stationary average-cost optimal policy can be computed by solving the LP

$$\begin{aligned} \text{minimize} \quad & \sum_{x \in \overline{\mathbb{X}}} \sum_{a \in \bar{A}(x)} \bar{c}(x,a)\bar{z}_{x,a} \\ \text{such that} \quad & \sum_{a \in \bar{A}(x)} \bar{z}_{x,a} - \overline{\beta} \sum_{y \in \overline{\mathbb{X}}} \sum_{a \in \bar{A}(y)} \bar{p}(x|y,a)\bar{z}_{y,a} = 1, \\ & \quad x \in \overline{\mathbb{X}}, \\ & \bar{z}_{x,a} \ge 0, \quad x \in \overline{\mathbb{X}},\ a \in \bar{A}(x). \end{aligned} \tag{46}$$

Recall that $m = \sum_{x\in\mathbb{X}} |A(x)|$ and $n = |\mathbb{X}|$. If $K^* > 1$, it follows from Scherrer [46, Theorem 3] that the LP (46) can be solved using

$$(m-n)\left\lceil \frac{1}{1-\overline{\beta}} \log \frac{1}{1-\overline{\beta}} \right\rceil = O((m-n)K^* \log K^*)$$

iterations of the block-pivoting simplex method corresponding to Howard's policy iteration algorithm. In addition, it follows from Scherrer [46, Theorem 4] that the LP (46) can alternatively be solved using

$$n(m-n)\left(1 + \frac{2}{1-\overline{\beta}} \log \frac{1}{1-\overline{\beta}}\right) \\ = O(n(m-n)K^* \log K^*) \qquad (47)$$

iterations of the simplex method with Dantzig's rule. Observe that $K^* = 1$ means that the state ℓ is absorbing under each stationary policy, and a stationary policy ϕ is average-cost optimal if and only if $c(\ell, \phi(\ell)) = \min\{c(\ell, a) : a \in A(\ell)\}$.

REMARK 18: According to [1, Proposition 12], there is a one-to-one correspondence between sequences of policies generated by Howard's policy iteration algorithm for the discounted MDP defined by the HV-AG transformation, and sequences of policies generated by Howard's policy iteration algorithm for the original unichain average-cost MDP. In particular, under Howard's policy iteration algorithm for the discounted MDP, an improved policy ϕ^+ is constructed (when possible) by taking, for each $x \in \mathbb{X}$, $\phi^+(x)$ to be any action belonging to

$$\operatorname*{argmin}_{a\in A(x)} \left[\bar{c}(x,a) + \overline{\beta} \sum_{y\in\mathbb{X}} \bar{p}(y|x,a) \bar{v}^{\phi}_{\overline{\beta}}(y) \right]. \qquad (48)$$

Under Howard's policy iteration algorithm for unichain average-cost MDPs, given $\phi \in \mathbb{F}$ an improved policy ϕ^+ is constructed by first obtaining a constant g and a function h that satisfy the system of equations

$$g + h(x) = c(x, \phi(x)) + \sum_{y\in\mathbb{X}} q(y|x,\phi(x))h(y), \quad x \in \mathbb{X}, \qquad (49)$$

and then, for every $x \in \mathbb{X}$, taking $\phi^+(x)$ to be any action belonging to

$$\operatorname*{argmin}_{a\in A(x)} \left[c(x,a) + \sum_{y\in\mathbb{X}} q(y|x,a)h(y) \right]. \qquad (50)$$

Let $h^\phi(x) := \mu(x)\left[\bar{v}^{\phi}_{\overline{\beta}}(x) - \bar{v}^{\phi}_{\overline{\beta}}(\ell)\right]$ for $x \in \mathbb{X}$. According to Proposition 14, the constant $\bar{v}^{\phi}_{\overline{\beta}}(\ell)$ and the function h^ϕ satisfy (49). Further, the definitions of $\bar{c}$ and $\bar{p}$ and Lemma 13 imply that for each $x \in \mathbb{X}$ the set (50) is equal to the set (48). This implies that Howard's policy iteration algorithm for the discounted MDP defined in Section 4.1 is equivalent to a particular version of Howard's policy iteration algorithm for the original unichain average-cost MDP. Since both of these policy iteration algorithms correspond to block-pivoting simplex methods (see [35, pp. 68, 122], it follows from Scherrer [46, Theorem 3] that, when there is a state that is recurrent under all stationary policies, the well-known LP for unichain average-cost MDPs, see for example [[35], LP 4.6.7],

$$\begin{aligned} \text{minimize} \quad & \sum_{x\in\mathbb{X}} \sum_{a\in A(x)} c(x,a) z_{x,a} \\ & \sum_{a\in A(x)} z_{x,a} - \sum_{y\in\mathbb{X}} \sum_{a\in A(y)} q(x|y,a) z_{y,a} = 0, \\ & \quad x \in \mathbb{X}, \\ & \sum_{y\in\mathbb{X}} \sum_{a\in A(y)} z_{y,a} = 1, \quad x \in \mathbb{X}, \\ & z_{x,a} \geq 0, \quad x \in \mathbb{X},\ a \in A(x), \end{aligned} \qquad (51)$$

can be solved using $O((m-n)K^* \log K^*)$ iterations of a block-pivoting simplex method.

REMARK 19: For $x \in \mathbb{X}$, let $\tau_\ell(x) := \sup_{\phi\in\mathbb{F}} \sum_{n=0}^{\infty} {}_{\ell}Q^n_\phi e(x)$. Then $K_\ell := \max_{x\in\mathbb{X}} \tau_\ell(x)$ is the smallest constant K^* satisfying (45). The iteration estimate for Howard's policy iteration algorithm for average-cost MDPs satisfying (45) that follows from Akian and Gaubert [1, Corollary 15] is $O((m-n)K_\ell \log K_\ell)$. One method to compute K_ℓ consists of the following. First, compute an optimal policy ϕ_* for a transient MDP that is identical to the original MDP, except that state ℓ is removed and all one-step costs are equal to -1. Then, compute the value function v^{ϕ_*} of this optimal policy, set

$$v^{\phi_*}(\ell) := \max_{a\in A(x)} \left[1 + \sum_{y\neq\ell} q(y|\ell,a) v^{\phi_*}(y) \right],$$

and set $K_\ell = \max_{x\in\mathbb{X}} v^{\phi_*}(x)$. According to Denardo [9, Theorem 2], the policy ϕ_* can be computed using $O((m-n)K_\ell \log K_\ell)$ iterations of Howard's policy iteration algorithm. Further, the function v^{ϕ_*} can be computed by solving a system of $n-1$ linear equations, using Gaussian elimination in $O(n^3)$ arithmetic operations; for other methods see for example [5, 48].

REMARK 20: Applying the simplex method with Dantzig's rule to the LP (46) can be viewed as applying a certain pivoting rule to the LP (51). In particular, for $\phi \in \mathbb{F}$ let $h^\phi(x) := \mu(x)[\bar{v}^{\phi}_{\overline{\beta}}(x) - \bar{v}^{\phi}_{\overline{\beta}}(\ell)]$ for $x \in \mathbb{X}$. Given a non-optimal basic feasible solution to (51) corresponding to the non-optimal stationary policy ϕ, it follows from Lemma 13 and Proposition 14 that the variable $z_{x,a}$ that enters the basis under this pivoting rule is the one minimizing

$$\bar{c}(x,a) + \overline{\beta}\sum_{y\in\mathbb{X}} \bar{p}(y|x,a)\bar{v}^{\phi}_{\overline{\beta}}(y) - \bar{v}^{\phi}_{\overline{\beta}}(x)$$
$$= \frac{1}{\mu(x)}\left[c(x,a) + \sum_{y\in\mathbb{X}} q(y|x,a)h^{\phi}(y) - w^{\phi} - h^{\phi}(x)\right], \tag{52}$$

and the variable that leaves the basis is $z_{x\phi(x)}$. According to (47), this pivoting rule for the LP (51), that is typically used to solve unichain average-cost MDPs, is strongly polynomial when K^* is fixed. This algorithm is not the same as applying Dantzig's rule to the LP (51), however; see Remark 21.

REMARK 21: Since an MDP satisfying Assumption HT is unichain, an optimal policy under the average-cost criterion can be computed by solving the LP (51); see for example [35, LP 4.6.7]. As follows from Remark 18, under Assumption HT, starting with the same basic variables, the sequences of basic variables for implementations of block-pivoting simplex methods for the LPs (46) and (51) coincide. However, this is not true for the simplex method with Dantzig's rule. To confirm this, let us consider the following example. The set of states is $\mathbb{X} = \{1, 2\}$ and the sets of available actions are $A(1) = A(2) = \{a, b\}$. The transition probabilities form stochastic vectors given by $p(1|1, a) = 1/2,\ p(1|1, b) = 0$, $p(1|2, a) = 1/3$, and $p(1|2, b) = 1/2$. The one-step costs are $c(1, a) = c(1, b) = 1$ and $c(2, a) = c(2, b) = 2$. Letting $\ell = 1$, one can verify that the function μ defined by $\mu(1) = 10$ and $\mu(2) = 3$ satisfies (38) with $V \equiv 1$. The average-cost LP given by the LP (51) is

$$\begin{aligned} \text{minimize}\quad & z_{1,a} + z_{1,b} + 2z_{2,a} + 2z_{2,b} \\ \text{such that}\quad & \frac{1}{2}z_{1,a} + z_{1,b} - \frac{1}{3}z_{2,a} - \frac{1}{2}z_{2,b} = 0 \\ & -\frac{1}{2}z_{1,a} - z_{1,b} + \frac{1}{3}z_{2,a} + \frac{1}{2}z_{2,b} = 0 \\ & z_{1,a} + z_{1,b} + z_{2,a} + z_{2,b} = 1 \\ & z_{1,a}, z_{1,b}, z_{2,a}, z_{2,b} \geq 0, \end{aligned} \tag{53}$$

and the LP (46) for the discounted MDP defined by the HV-AG transformation is

$$\begin{aligned} \text{minimize}\quad & \frac{1}{10}z_{1,a} + \frac{1}{10}z_{1,b} + \frac{1}{3}z_{2,a} + \frac{1}{3}z_{2,b} \\ \text{such that}\quad & \frac{1}{4}z_{1,a} + \frac{2}{5}z_{1,b} - \frac{1}{6}z_{2,b} = 1 \\ & -\frac{3}{20}z_{1,a} - \frac{3}{10}z_{1,b} + \frac{1}{3}z_{2,a} + \frac{1}{2}z_{2,b} = 1 \\ & z_{1,a}, z_{1,b}, z_{2,a}, z_{2,b} \geq 0. \end{aligned} \tag{54}$$

For both LPs, suppose the initial basic feasible solution for the simplex method with Dantzig's rule is the one defined by the stationary policy ϕ where $\phi(1) = b$ and $\phi(2) = a$; namely, the basic variables are $z_{1,b}$ and $z_{2,a}$. Consider the first iteration of this simplex method. For the LP (53), the basic variable $z_{1,b}$ is the unique variable to leave the basis, while for the LP (54) the basic variable $z_{2,a}$ is the unique variable to leave the basis.

REMARK 22: Consider an LP with n constraints and m variables, where the positive elements of every basic feasible solution are bounded below by δ and bounded above by γ. By generalizing the analysis in Ye [55] for discounted MDPs, it is proved in Kitahara and Mizuno [36, Theorem 3] that the simplex method with Dantzig's rule requires at most

$$O\left(nm\frac{\gamma}{\delta}\log\frac{\gamma}{\delta}\right)$$

iterations to return an optimal solution. For the LP (46), $\delta = 1$ and $\gamma = (1-\tilde{\beta})^{-1} = K^*$ satisfy the hypotheses of this result. Therefore, it follows from [36, Theorem 3] that an average-cost optimal policy can be computed in strongly polynomial time when K^* is fixed, by applying the simplex method with Dantzig's rule to the LP (46). However, [36, Theorem 3] does not imply an analogous statement for the LP (51) for unichain average-cost MDPs. This is because, for such MDPs, every basic feasible solution of (51) is the vector of state-action frequencies under some stationary policy [35, Remark 4.7.4]. Even for MDPs satisfying Assumption HT with a fixed K^*, these frequencies can decrease exponentially with the number of states. To verify this, for $n = 2, 3, \ldots$ consider an MDP with state set $\mathbb{X} := \{1, \ldots, n\}$, a single action 0 available at every state, transition probabilities $p(1|1, 0) = p(n|i, 0) = p(i|i + 1, 0) := 1/2$ for $i = 1, \ldots, n - 1$, and arbitrary real-valued one-step costs. Observe that for $n = 1, 2, \ldots$, this MDP satisfies Assumption HT with $\ell = n$ and $K^* = 2$. In addition, the unique feasible solution to (51) for this MDP is

$$z_{1,0} = \left(\frac{1}{2}\right)^{n-1}, \quad z_{i,0} = \left(\frac{1}{2}\right)^{n-i+1}, \quad \text{for } i = 2, \ldots, n.$$

Thus, there is no $\delta > 0$ such that $z_{1,0} \geq \delta$ for all $n = 2, 3, \ldots$.

4.4. Extension to Uncountable State Spaces

For $\phi \in \mathbb{F}$, let ${}_{\ell}Q_{\phi}$ be defined for an integrable Borel-measurable $u : \mathbb{X} \to \mathbb{R}$ as

$${}_{\ell}Q_{\phi}u(x) := \int_{\mathbb{X}\setminus\{\ell\}} u(y)q(dy|x, \phi(x)), \quad x \in \mathbb{X}.$$

The version of Assumption HT that we consider when the state space is possibly uncountable is as follows:

ASSUMPTION HT′:

(i) There is a state $\ell \in \mathbb{X}$ and a constant K^* satisfying

$$\left\| \sum_{n=0}^{\infty} {}_\ell Q_\phi^n \right\| \le K^* < \infty \quad \text{for all } \phi \in \mathbb{F}. \tag{55}$$

(ii) The one-step cost function c is bounded.

4.4.1. *HV-AG Transformation*

Suppose Assumption HT′ holds. According to Proposition 8, there is a Borel-measurable function $\mu : \mathbb{X} \to [1, \infty)$ that satisfies $\mu \le K^*$ and

$$\mu(x) \ge 1 + \int_{\mathbb{X}\setminus\{\ell\}} \mu(y) q(dy|x,a), \quad (x,a) \in \mathrm{Gr}(A). \tag{56}$$

Here the HV-AG transformation is defined exactly as described in Section 4.1, except that the cost-free absorbing state $\bar{x}$ is taken to be isolated from $\mathbb{X}$, and the transition probabilities $\bar{p}$ are defined by

$$\bar{p}(B|x,a) := \begin{cases} \frac{1}{\overline{\beta}\mu(x)} \int_B \mu(y) q(dy|x,a), \\ \quad B \in \mathcal{B}(\mathbb{X} \setminus \{\ell\}),\ (x,a) \in \mathrm{Gr}(A), \\ \frac{1}{\overline{\beta}\mu(x)} [\mu(x) - 1 - \int_{\mathbb{X}\setminus\{\ell\}} \mu(y) q(dy|x,a)], \\ \quad B = \{\ell\},\ (x,a) \in \mathrm{Gr}(A), \\ 1 - \frac{1}{\overline{\beta}\mu(x)} [\mu(x) - 1], \\ \quad B = \{\bar{x}\},\ (x,a) \in \mathrm{Gr}(A), \\ 1, \quad B = \{\bar{x}\},\ (x,a) = (\bar{x}, \bar{a}). \end{cases}$$

4.4.2. *Results*

LEMMA 18: If a bounded Borel function $f : \overline{\mathbb{X}} \to \mathbb{R}$ satisfies $f(\bar{x}) = 0$, then for any $x \in \mathbb{X}$ and $a \in A(x)$

$$\begin{aligned} &\bar{c}(x,a) + \overline{\beta} \int_{\overline{\mathbb{X}}} f(y) \bar{p}(dy|x,a) \\ &\quad = \frac{1}{\mu(x)} \Big[c(x,a) + \int_{\mathbb{X}} \mu(y)[f(y) - f(\ell)] q(dy|x,a) \\ &\qquad + [\mu(x) - 1] f(\ell) \Big]. \end{aligned} \tag{57}$$

PROOF: This follows from the proof of Lemma 13, with all sums replaced with integrals. □

PROPOSITION 19: Suppose Assumption HT′ holds. Let $\phi \in \mathbb{F}$ be a stationary policy and $h^\phi(x) := \mu(x)[\bar{v}^\phi_{\overline{\beta}}(x) - \bar{v}^\phi_{\overline{\beta}}(\ell)]$ for $x \in \mathbb{X}$. Then

$$\bar{v}^\phi_{\overline{\beta}}(\ell) + h^\phi(x) = c(x, \phi(x)) + \int_{\mathbb{X}} h^\phi(y) q(dy|x, \phi(x)), \quad x \in \mathbb{X}. \tag{58}$$

In addition, if the transition rates q are stochastic, then $w^\phi \equiv \bar{v}^\phi_{\overline{\beta}}(\ell)$.

PROOF: This follows from the proof of Proposition 14, where sums are replaced with integrals in the appropriate places. □

LEMMA 20: Suppose Assumptions S and HT′ hold. Then the discounted MDP defined by the HV-AG transformation also satisfies Assumption S.

PROOF: This follows from Lemma 9 and the proof of Lemma 15. □

To state the main result in this section, for $x \in \mathbb{X}$ define $A^*_{\mathrm{av}}(x)$ by replacing the sum in (42) with an integral.

THEOREM 21: Suppose the original MDP with transition probabilities q satisfies Assumptions S and HT′. Then:

(i) the constant $w = \bar{v}_{\overline{\beta}}(\ell)$ and the function $h(x) = \mu(x)[\bar{v}_{\overline{\beta}}(x) - \bar{v}_{\overline{\beta}}(\ell)]$, $x \in \mathbb{X}$, satisfy the optimality equation

$$w + h(x) = \min_{A(x)} \left[c(x,a) + \int_{\mathbb{X}} h(y) q(dy|x,a) \right], \quad x \in \mathbb{X}, \tag{59}$$

and $\bar{v}_{\overline{\beta}}(\ell)$ is the optimal average cost for each initial state.

(ii) there is a $\phi \in \mathbb{F}$ satisfying $\phi(x) \in A^*_{\mathrm{av}}(x)$ for all $x \in \mathbb{X}$, where

$$A^*_{\mathrm{av}}(x) = \left\{ a \in A(x) \mid \bar{v}_{\overline{\beta}}(x) = \bar{c}(x,a) + \overline{\beta} \int_{\overline{\mathbb{X}}} \bar{v}_{\overline{\beta}}(y) \bar{p}(dy|x,a) \right\}, \quad x \in \mathbb{X}, \tag{60}$$

and any such policy is average-cost optimal.

PROOF: This follows from the proof of Theorem 16, where sums are replaced with integrals in the appropriate places. □

ACKNOWLEDGMENTS

The authors thank the associate editor, two anonymous referees, Rolando Cavazos-Cadena, Eric Denardo, and Pavlo Kasyanov for useful comments and suggestions. This research was partially supported by NSF grants CMMI-1335296 and and CMMI-1636193.

REFERENCES

[1] M. Akian and S. Gaubert, Policy iteration for perfect information stochastic mean payoff games with bounded first return times is strongly polynomial, Preprint (2013), http://arxiv.org/abs/1310.4953v1.

[2] E. Altman, Constrained Markov decision processes, Chapman and Hall/CRC, Boca Raton, FL, 1999.

[3] D.P. Bertsekas and S.E. Shreve, Stochastic optimal control: The discrete-time case, Athena Scientific, Belmont, MA, 1996.

[4] L.D. Brown and R. Purves, Measurable selections of extrema, Ann Stat 1 (1973), 902–912.

[5] D.S. Coppersmith and S. Winograd, Matrix multiplication via arithmetic progressions, J Symbolic Comput 9 (1990), 251–280.

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

[7] R. Dekker, A. Hordijk, and F.M. Spieksma, On the relation between recurrence and ergodicity properties in denumerable Markov decision chains, Math Oper Res 19 (1994), 539–559.

[8] E.V. Denardo, Contraction mappings in the theory underlying dynamic programming, SIAM Rev 9 (1967), 165–177.

[9] E.V. Denardo, Nearly strongly polynomial algorithms for transient dynamic programs, Preprint, February 1, 2016.

[10] E.V. Denardo, E.A. Feinberg, and U.G. Rothblum, The multi-armed bandit, with constraints, Ann Oper Res 208 (2013), 37–62.

[11] E.V. Denardo, H. Park, and U.G. Rothblum, Risk-sensitive and risk-neutral multiarmed bandits, Math Oper Res 32 (2007), 374–394.

[12] E.B. Dynkin and A.A. Yushkevich, Controlled Markov processes, Springer-Verlag, New York, NY, 1979.

[13] B.C. Eaves and A.F. Veinott, Maximum-stopping-value policies in finite Markov population decision chains, Math Oper Res 39 (2014), 597–606.

[14] G. Even and A. Zadorojniy, Strong polynomiality of the Gass-Saaty shadow-vertex pivoting rule for controlled random walks, Ann Oper Res 201 (2012), 159–167.

[15] A. Federgruen, A. Hordijk, and H.C. Tijms, "Recurrence conditions in denumerable state Markov decision processes," in: M.L. Puterman (Editor), Dynamic programming and its applications, Academic Press, New York, NY, 1978, pp. 3–22.

[16] A. Federgruen and H.C. Tijms, The optimality equation in average cost denumerable state semi-Markov decision problems – recurrency conditions and algorithms, J Appl Probab 15 (1978), 356–373.

[17] E.A. Feinberg, "Total reward criteria," in: E.A. Feinberg and A. Shwartz (Editors), Handbook of Markov decision processes, Kluwer Academic Publishers, Norwell, MA, 2002, pp. 173–207.

[18] E.A. Feinberg, "Constrained discounted semi-Markov decision processes," in: Z. Hou, J.A. Filar, and A. Chen (Editors), Markov processes and controlled Markov chains, Academic Publishers, Dordrecht, 2002, pp. 231–242.

[19] E.A. Feinberg and J. Huang, Strong polynomiality of policy iterations for average-cost MDPs modeling replacement and maintenance problems, Oper Res Lett 41 (2013), 249–251.

[20] E.A. Feinberg and J. Huang, The value iteration algorithm is not strongly polynomial for discounted dynamic programming, Oper Res Lett 42 (2014), 130–131.

[21] E.A. Feinberg, J. Huang, and B. Scherrer, Modified policy iteration algorithms are not strongly polynomial for discounted dynamic programming, Oper Res Lett 42 (2014), 429–431.

[22] E.A. Feinberg, P.O. Kasyanov, and N. V. Zadoianchuk, Average cost Markov decision processes with weakly continuous transition probabilities, Math Oper Res 37 (2012), 591–607.

[23] E.A. Feinberg and U. G. Rothblum, Splitting randomized stationary policies in total-reward Markov decision processes, Math Oper Res 37 (2012), 129–153.

[24] E.A. Feinberg, and F. Yang, On polynomial cases of the unichain classification problem for Markov decision processes, Oper Res Lett 36 (2008), 527–530.

[25] L.G. Gubenko and È.S. Štatland, On controlled discrete-time Markov decision processes, Theor Probab Math Stat 7 (1975), 47–61.

[26] T.D. Hansen, P.B. Miltersen, and U. Zwick, Strategy iteration is strongly polynomial for 2-player turn-based stochastic games with a constant discount factor, J ACM 60 (2013), 1–16.

[27] O. Hernández-Lerma and J.B. Lasserre, Discrete-time markov control processes: Basic optimality criteria, Springer-Verlag, New York, NY, 1996.

[28] O. Hernández-Lerma and J.B. Lasserre, Further topics on discrete-time markov control processes, Springer-Verlag, New York, NY, 1999.

[29] O. Hernández-Lerma, R. Montes-de-Oca, and R. Cavazos-Cadena, Recurrence conditions for Markov decision processes with Borel state space: A survey, Ann Oper Res 28 (1991), 29–46.

[30] K. Hinderer and K.H. Waldmann, The critical discount factor for finite Markovian decision processes with an absorbing set, Math Meth Oper Res 57 (2003), 1–19.

[31] K. Hinderer and K.H. Waldmann, Algorithms for countable state Markov decision models with an absorbing set, SIAM J Control Optim 43 (2005), 2109–2131.

[32] A. Hordijk, Dynamic programming and Markov potential theory, Mathematisch Centrum, Amsterdam, 1974.

[33] A. Hordijk and L.C.M. Kallenberg, Transient policies in discrete dynamic programming: Linear programming including suboptimality tests and additional constraints, Math Program 30 (1984), 46–70.

[34] R.A. Howard, Dynamic programming and Markov processes, The MIT Press, Cambridge, MA, 1960.

[35] L.C.M. Kallenberg, Linear programming and finite markovian control problems, Mathematisch Centrum, Amsterdam, 1983.

[36] T. Kitahara and S. Mizuno, A bound for the number of different basic solutions generated by the simplex method, Math Program Ser A 137 (2013), 579–586.

[37] S.R. Pliska, Optimization of multitype branching processes, Manage Sci 23 (1976), 117–124.

[38] S.R. Pliska, "On the transient case for Markov decision chains with general state spaces," in: M.L. Puterman (Editor), Dynamic programming and its applications, Academic Press, New York, NY, 1978, pp. 335–349.

[39] S.M. Ross, Non-discounted denumerable Markovian decision models, Ann Math Stat 39 (1968), 412–423.

[40] S.M. Ross, Arbitrary state Markovian decision processes, Ann Math Stat 39 (1968), 2118–2122.

[41] U.G. Rothblum, Normalized Markov decision chains I; sensitive discount optimality, Oper Res 23 (1975), 785–795.

[42] U.G. Rothblum and A.F. Veinott, Markov branching decision chains: Immigration-induced optimality, Technical Report No. 45, Department of Operations Research, Stanford University, 1992.

[43] U.G. Rothblum and P. Whittle, Growth optimality for branching Markov decision chains, Math Oper Res 7 (1982), 582–601.

[44] H.L. Royden, Real analysis, 3rd ed., Prentice-Hall, Upper Saddle River, NJ, 1988.

[45] M. Schäl, Average optimality in dynamic programming with general state space, Math Oper Res 18 (1993), 163–172.

[46] B. Scherrer, Improved and generalized upper bounds on the complexity of policy iteration, Math Oper Res 41 (2016), 758–774.

[47] L.I. Sennott, Stochastic dynamic programming and the control of queueing systems, John Wiley and Sons, New York, NY, 1999.

[48] V. Strassen, Gaussian elimination is not optimal, Numer Math 13 (1969), 354–356.

[49] L.C. Thomas, "Connectedness conditions for denumerable state Markov decision processes," in: R. Hartley, L.C. Thomas, and D.J. White (Editors), Recent developments in Markov decision processes, Academic Press, New York, NY, 1980, pp. 181–204.

[50] J.N. Tsitsiklis, NP-hardness of checking the unichain condition in average cost MDPs, Oper Res Lett 35 (2007), 319–323.

[51] J. van der Wal, Stochastic dynamic programming, Mathematisch Centrum, Amsterdam, 1981.

[52] A.F. Veinott, Discrete dynamic programming with sensitive discount optimality criteria, Ann Math Stat 40 (1969), 1635–1660.

[53] A.F. Veinott, "Markov decision chains," in: G.B. Dantzig and B.C. Eaves (Editors) Studies in optimization, MAA Studies in Mathematics Vol. 10, Mathematical Association of America, Washington DC, 1974, pp. 124–159.

[54] A.F. Veinott, Lectures in Dynamic Programming and Stochastic Control, Course Notes, Stanford University, 2008.

[55] Y. Ye, The simplex and policy-iteration methods are strongly polynomial for the Markov decision problem with a fixed discount rate, Math Oper Res 36 (2011), 593–603.

[56] A. Zadorojniy, G. Even, and A. Shwartz, A strongly polynomial algorithm for controlled queues, Math Oper Res 34 (2009), 992–1007.