Óscar Vega-Amaya; Joaquín López-Borbón

A perturbation approach to approximate value iteration for average cost Markov decision processes with Borel spaces and bounded costs

Kybernetika, Vol. 55 (2019), No. 1, 81–113

Persistent URL: `http://dml.cz/dmlcz/147707`

KYBERNETIKA — VOLUME *55* (2019), NUMBER 1, PAGES 81–113

A PERTURBATION APPROACH TO APPROXIMATE VALUE ITERATION FOR AVERAGE COST MARKOV DECISION PROCESSES WITH BOREL SPACES AND BOUNDED COSTS

Óscar Vega-Amaya and Joaquín López-Borbón

The present paper studies the *approximate value iteration* (AVI) algorithm for the average cost criterion with bounded costs and Borel spaces. It is shown the convergence of the algorithm and provided a performance bound assuming that the model satisfies a standard continuity-compactness assumption and a uniform ergodicity condition. This is done for the class of approximation procedures that can be represented by linear positive operators which give exact representation of constant functions and also satisfy certain continuity property. The main point is that these operators define transition probabilities on the state space of the controlled system. This has the following important consequences: (a) the approximating function is the average value of the target function with respect to the induced transition probability; (b) the approximation step in the AVI algorithm can be seen as a perturbation of the original Markov model; (c) the perturbed model inherits the ergodicity properties imposed on the original Markov model. These facts allow to bound the AVI algorithm performance in terms of the accuracy of the approximations given by this kind of operators for the primitive data model, namely, the one-step reward function and the system transition law. The bounds are given in terms of the supremum norm of bounded functions and the total variation norm of finite-signed measures. The results are illustrated with numerical approximations for a class of single item inventory systems with linear order cost, no set-up cost and no back-orders.

Keywords: Markov decision processes, average cost criterion, approximate value iteration algorithm, contraction and non-expansive operators, perturbed Markov decision models

Classification: 90C40, 90C59, 93E20

1. INTRODUCTION

The *approximate value iteration* (AVI) algorithms is a class of approximating procedures aiming to cope the numerical computation of solutions to the optimal control problem in Markov decision processes. The research on approximation procedures is by now a very active subdiscipline of Markov decision processes with a diversity of problems, and the literature dealing with them has experienced an explosive growth in the recent years;

DOI: 10.14736/kyb-2019-1-0081

see, for instance, the books [7, 10, 38, 48] and the survey papers [8, 31, 39, 40, 41, 43, 44]. In this concern, Powell [40] says that this situation "can sometimes seem like a jungle of algorithmic strategies." Nonetheless, the major part of the works are concentrated either on the discrete space case (mainly, on the finite case) or on the discounted cost criterion. Bertsekas [8] and Powell and Ma [41] give succinct but comprehensive reviews on the approximate policy iteration for the discounted cost criterion; the former deals with finite models, whereas the latter one reviews the continuous spaces case. For further results on the discounted criterion with uncountable state and control spaces see the references [3, 14, 15, 34, 42, 45, 47, 53].

The present work studies a class of approximate value iteration (AVI) algorithms for the *average cost criterion* with *Borel spaces* and *bounded costs*. The average cost optimal control problem, being by far more difficult that the discounted one, has been much less studied from the numerical viewpoint. However, there are a number of important contributions among which we can mention the references [1, 9, 11, 16, 18, 19, 21, 30, 36, 46]. The approaches, assumptions and results of the these references differ with those of the present work in several respects. Here we follow the perturbation approach developed and applied to discounted problems in [42, 53]. We will comment briefly on the former references at the end of this section.

The main approach to solve the optimal control problem, either with respect to the average cost criterion or the discounted cost criterion, is to find solutions to the corresponding optimality equations. The *average cost optimality equation* has the form

$$\rho^* + h^* = Th^*, \tag{1}$$

where ρ^* is a constant, h^* is a measurable function and T is the dynamic programming operator (see (9) below). If such a pair (ρ^*, h^*) exists and h^* is a bounded function, for instance, it is proved using standard arguments that ρ^* is the *optimal average cost* and also that any stationary policy f^* that attains the minimum at the right hand-side of the optimality equation is optimal. For general results on the average cost criterion see, for instance, [4, 23, 24, 27, 28, 50, 51, 52].

However, equation (1) can be rarely solved analytically, so its solutions have to be approximated. The major schemes to approximate a solution to the optimality equation (1) are the *value iteration algorithm*, the *policy iteration algorithm,* and the *linear programming approach* [6, 22, 23, 24, 37]. The present work concerns with the value iteration algorithm, which, roughly speaking, aims to get a solution (ρ^*, h^*) as limit of the sequences

$$\rho_n := J_n - J_{n-1}, \qquad h_n := J_n - J_n(z), \tag{2}$$

for $n \in \mathbb{N} := \{1, 2, \ldots\}$, where J_n is the *n-step optimal cost function* (with $J_0 \equiv 0$), and $z \in \mathbf{X}$ is an arbitrary but fixed state.

Using the dynamic programming operator these sequences are related by the equations

$$\rho_{n+1}(z) + h_{n+1} = Th_n \tag{3}$$

with $h_0 \equiv 0$. Note that $\rho_{n+1}(z) = Th_n(z)$ because $h_{n+1}(z) = 0$. Thus, we have that

$$h_{n+1} = T_z h_n \quad \forall n \in \mathbb{N}, \tag{4}$$

where the operator T_z is defined as

$$T_z u := Tu - Tu(z) \tag{5}$$

for functions u belonging to an appropriate space of functions. Observe that (ρ^*, h^*) satisfies (1) with $h^*(z) = 0$ if and only if h^* is a *fixed point* of T_z.

Thus, it is said that the *value iteration algorithm* (3)–or (4)–converges if the sequence $(\rho_n(z), h_n), n \in \mathbb{N}$, converges to a solution (ρ^*, h^*) of the average cost optimality equation (1). If this is the case and the algorithm (4) is stopped at stage $n \in \mathbb{N}$, then it is computed a *greedy* stationary policy f_n with respect to the function h_n–that is, a policy that attains the minimum at Th_n– and the optimal value ρ^* is approximated by the average cost $J(f_n)$ induced by policy f_n.

Unfortunately, the procedure (4)-or (2) or (3)-is numerically infeasible for systems with large or continuous spaces if it is pursued an exact representation of the sequence $\{(\rho_n, h_n)\}$. This obstacle is circumvented by considering an *approximate* or *fitted value iteration algorithm* which interleaves an approximation step between two consecutive applications of the dynamic programming operator T. In many cases the approximation step is represented by an operator L, so Lv is the approximation of function v.

There are two slight different approximate procedures depending on which operator, either T or L, acts first. These procedures are given by the composite operators $\widehat{T} = TL$ and $\widetilde{T} = LT$. Thus, the standard value iteration algorithm (4) is substituted by

$$\widehat{h}_{n+1} = \widehat{T}_z \widehat{h}_n := \widehat{T\widehat{h}}_n - \widehat{T\widehat{h}}_n(z), \tag{6}$$

or

$$\widetilde{h}_{n+1} = \widetilde{T}_z \widetilde{h}_n := \widetilde{T\widetilde{h}}_n - \widetilde{T}_z \widetilde{h}_n(z), \tag{7}$$

for $n \in \mathbb{N}_0 := \mathbb{N} \cup \{0\}$.

In several approaches, Lv is the projection of function v into some class of approximating functions $\mathcal{F}$ that depends on finitely many parameters. The family $\mathcal{F}$ is usually taken as the spanned space by linear combinations of a finite "basis function" $\phi_0, \phi_1, \ldots, \phi_N$. The class $\mathcal{F}$ is called an *architecture* for the approximating problem and Lv is called *scoring function*. The main steps in the projection method is the choice of a good architecture $\mathcal{F}$, which is problem dependent, and the computation of the projections. Usually the projections are very difficult to find analytically, thus they are approximated by a variety of simulation methods (see, e. g., [3, 7, 34, 38]). Clearly, this introduces a second kind of approximation errors, which in many cases turn out to be quite difficult to measure or to control.

Other kind of approximation operators is given by the so-called self-approximating operators, that is, operators that do not need auxiliary methods to produce the approximation Lv for any function v. Examples of such operators are piecewise constant approximations, linear and multilinear approximations, splines, Chebyshev polynomials, kernel-based approximations, etc. (see, e. g., [43, 47]).

Then, as in the exact VI algorithm (3) or (4), if the algorithm (7) is stopped at stage n, it computes an $\widetilde{h}_n$-greedy policy f_n–i. e., a stationary policy that attains the minimum at $T\widetilde{h}_n$–and approximates the optimal average cost ρ^* by means of the average cost $J(f_n)$ incurred by policy f_n. Thus, this approximate value iteration algorithm raises several important problems:

P₁ The first one is the convergence of the approximate sequence $\{(\widetilde{\rho}_n, \widetilde{h}_n)\}$;

P₂ Once the convergence is ensured, say to $(\widetilde{\rho}, \widetilde{h})$, the second one concerns with establishing computable bounds for the approximation errors $\widetilde{\rho} - \rho^*$ and $\widetilde{h} - h^*$.

P₃ The third one–perhaps, the most important from the practical viewpoint–consists in providing performance bounds for the algorithm, that is, bounds for the approximation error $J(f_n) - \rho^*$.

Obviously, the same problems are raised for the algorithm (6) replacing the sequence $\{(\widetilde{\rho}_n, \widetilde{h}_n)\}$ and $(\widetilde{\rho}, \widetilde{h})$ by $\{(\widehat{\rho}_n, \widehat{h}_n)\}$ and $(\widehat{\rho}, \widehat{h})$, respectively.

Concerning problem $\mathbf{P}_1$, it was noted in previous works that the approximate value iterations need not to converge even for the discounted cost criterion, at least the approximator L has a non-expansive property [17, 20, 43]. For instance, it is well known that the β-discounted dynamic programming operator T_β is a contraction operator on the space of bounded measurable functions whenever the one-step cost is a bounded function; thus, if L is non-expansive, the operators $\widehat{T}_\beta := T_\beta L$ and $\widetilde{T}_\beta := LT_\beta$ are contractions operator too. Hence, the Banach fixed-point theorem guarantees the convergence of the approximate sequences. On the other hand, the projection method, which is the most used approximating procedure, usually leads to expansive operators. (One exception to this rule is the aggregation-projection scheme used in [49].) Thus, the convergence of the approximating iterates (6) and (7) may fail and the hope in this case is that the sequences of n-step approximation errors or *Bellman residuals* $||L\widehat{h}_n - TL\widehat{h}_n||$ and $||\widetilde{h}_n - T\widetilde{h}_n||, n \in \mathbb{N}$, remain bounded with respect to some suitable norm $|| \cdot ||$. If the convergence fails, the problem $\mathbf{P}_3$ still make sense provided the Bellman residuals remain bounded, but it becomes in a pretty difficult task; in fact, the authors are unaware of general computable finite-time bounds either for the discounted or the average cost criterion.

We address the problems $\mathbf{P}_1 - \mathbf{P}_3$ assuming that the Markov decision model satisfies standard compactness and continuity conditions (Assumptions 2.2 and 2.3, Section 2), a uniform ergodicity condition (Assumption 2.4) and considering positive linear operators– as suggested in [20] – that give exact representation of the constant functions and additionally have the following continuity property: $Lv_n \downarrow 0$ whenever $v_n \downarrow 0$ (Definition 4.1). Following Gordon [20], we call *averagers* to these operators. The averagers is a rich class of approximators that includes many of the operators studied in approximation theory as piecewise constant approximation operators, linear and multilinear interpolators, kernel-based interpolators [20, 47], certain aggregation-projection operators [49], Schoenberg' splines, Hermite-Fejér and Bernstein operators [5, 12], among others.

The key point in our approach is that the averagers define transition probabilities on the state space (see Lemma 4.2). This fact allows to view the approximating action of an averager as a perturbation of the original Markov decision model and then to provide the bounds asked in $\mathbf{P}_2$ and $\mathbf{P}_3$. The convergence in $\mathbf{P}_1$ comes from the fact that the perturbations induced by averagers preserve the continuity and ergodic properties imposed on the original model (see Assumption 2.4, Section 2). This implies that operators $\widehat{T}_z$ and $\widetilde{T}_z$ are contraction mappings with respect to the span semi-norm (see Lemma 4.5), establishing the convergence in problem $\mathbf{P}_1$. To the best of the authors' knowledge, the aforementioned link between approximation schemes using averagers and perturbed models has passed unnoticed in the literature excepting the paper by Gordon [20], who does not take advantage of it. In fact, Gordon referred to such link as an "intriguing property" that allows to view the "averagers as a Markov processes" [20, Section 4]). A remarkable fact is that, once the approximation step is seen as a perturbation, the analysis of problems $\mathbf{P}_1 - \mathbf{P}_3$ is direct and the proofs follow by somewhat elementary arguments.

As mentioned previously, references [1, 9, 11, 16, 18, 19, 21, 30, 36, 46] study numerical methods for solving the average cost optimal control problem; among these, only [16, 30, 46] deal with Borel spaces. Next we briefly comment on them just for a rough comparison with the perturbation approach described above.

References [1] and [21] show the almost surely convergence of Q-learning algorithms for finite models; the Q-learning variant used in [21] is based on the policy iteration algorithm and it applies to both Markov and semi-Markov models. Reference [11] studies finite models and also follows a simulation-based approach; it is proposed a policy iteration algorithm whose evaluation step estimates the exact solution of the Poisson equations by "direct" simulations and the improvement step is realized with respect to such estimates; moreover, it is provided a set of verifiable conditions that guarantee that the proposed simulation-based policy iteration algorithm eventually reaches, and never leaves, the set of optimal policies almost surely; it is also given three simulation estimators for the evaluation step that satisfy such conditions. In reference [30] is shown the convergence of a class of actor-critic algorithms for models with both finite and Borel spaces under several technical requirements; the actor-critic algorithms are simulation-based methods that works in a parameterized policy space: the critic estimates the policy value using temporal difference learning and the actor makes the improvement in the parameterized policy space in an approximate gradient direction. Reference [36] proposes an state aggregation method based on the notion of similarity among states defined in terms of certain pseudometrics; its main result establishes an upper bound for the loss of optimality in the aggregated model. Reference [9] studies an approximate version of a receding or rolling horizon approach for models with countable state spaces and finite action sets; it is analyzed the performance of an approximate receding horizon policy assuming that such a policy is computed using good approximations for the finite horizon costs.

On the other hand, [18] develops a two-phase approximate linear program for finite models; the first phase is to approximate the optimal average cost, whereas the second phase is to control the accuracy of the approximations of the (differential cost) function that solves the optimality equation. Reference [19] also develops an approximate linear programming approach but considering models with countable state spaces and finite action sets in discrete and continuous time; this last paper exploits the fact that the average cost problem is the limit of discounted problems when the discount factor vanishes, and the discounted optimal problem is seen as a perturbed version of the average cost problem. In [18] and [19], the approximate linear program is based on a linear architecture.

Reference [16] considers models with Borel spaces and unbounded cost but assumes–among other technical conditions–that the transition law is absolutely continuous with respect to some reference probability measure. This probability measure is approximated by means of empirical measures to obtain a finite-state models, which in turn are used to compute approximations to an optimal policy as well as to the optimal average cost; the accuracy of the approximations is measured by means of concentration inequalities based on the 1-Wassersttein distance of probability measures.

On the other hand, reference [46] also study finite approximations for models with Borel spaces but using a nearest neighborhood quantizer method to produce discretizations of the state space. (As can be seen in Remark 5.5 below, this kind of methods defines a particular class of averagers.) It is shown the asymptotic optimality of the optimal stationary policies corresponding to the discretized models. This is done in two settings. In the first one, it is assumed that the state and the control spaces are both compact subsets, the one-step cost is a bounded continuous function, and the transition law is continuous with respect to the total variation norm and also that it satisfies an ergodicity condition. In the second one, it is assumed that the state space is locally compact and the control space is compact, the transition law satisfies a Lyapunov condition (as in [50, 51, 52]) and that it is continuous with respect to a weighted norm for finite signed measures. It is also assumed that the cost function is continuous and unbounded but having a growth dominated by a weighting function. In both cases, the weighting function is that appearing in the Lyapunov condition.

The remainder of the present work is organized as follows. Section 2 introduces the Markov decision model, the average cost criterion and the assumptions imposed on the model. Section 3 contains some well-known results on the value iteration algorithm, which are our departing point. The core of the work are Sections 4 and 5. Section 4 introduces the averagers and the perturbed models, while Section 5 addresses problems $\mathbf{P}_1-\mathbf{P}_3$; specifically, Theorem 5.2 proves the convergence of the AVI algorithm, among other properties, and Theorems 5.3 and 5.5 give the performance bounds. Section 6 illustrates our perturbation approach with numerical results for a single item inventory system with finite capacity, no backlog and no set-up production cost. Section 7, Appendix, collects the proof of the results.

2. THE AVERAGE COST CRITERION

We use the following concepts and notation throughout the paper. For a topological space (S,τ), let $\mathcal{B}(S)$ denote the Borel σ-algebra generated by the topology τ; any statement about "measurability" will always mean Borel measurability. A Borel space S is a measurable subset of a complete separable metric space endowed with the inherited metric. For each subset $A\subset S$, $\mathbb{I}_A$ is the indicator function of A, that is, $\mathbb{I}_A(s)=1$ for $s\in A$ and $\mathbb{I}_A(s)=0$ otherwise. Let $M(S)$ be the class of all measurable functions on S and $M_b(S)$ be the subspace of bounded measurable functions; $C_b(S)$ stands for the subspace of bounded continuous functions. The two latter spaces are endowed with the supremum norm $||v||_\infty:=\sup_{s\in S}|u(s)|$.

We consider a Markov decision model $M = (\mathbf{X}, \mathbf{A}, \{A(x) : x \in \mathbf{X}\}, Q, C)$ with the usual meaning: the *state space* $\mathbf{X}$ and the *action* or *decision space* $\mathbf{A}$ are Borel spaces; $A(x)$ is a nonempty measurable subset of $\mathbf{A}$, which stands for the *admissible action* or *decision set* for state $x \in \mathbf{X}$. As usually, it is assumed that the admissible state-action pairs set $\mathbb{K} := \{(x, a) \in \mathbf{X} \times \mathbf{A} : a \in A(x), x \in \mathbf{X}\}$ is a measurable subset of the product space $\mathbf{X} \times \mathbf{A}$. The *transition law* $Q(\cdot|\cdot, \cdot)$ is a stochastic kernel on $\mathbf{X}$ given $\mathbb{K}$, that is, $Q(\cdot|x, a)$ is a probability measure on $\mathbf{X}$ for each pair $(x, a) \in \mathbb{K}$, and $Q(B|\cdot, \cdot)$ is a measurable function on $\mathbb{K}$ for each measurable subset B of $\mathbf{X}$. Finally, the one-step cost or running cost function $C(\cdot, \cdot)$ is a measurable function on $\mathbb{K}$.

A Markov decision model is a model of a stochastic dynamical system that evolves as follows: at time $n = 0$ the decision-maker observes the system in some state $x_0 = x \in \mathbf{X}$ and chooses a decision or action $a_0 = a \in A(x)$ incurring in a cost $C(x, a)$. Then, the system moves to a new state $x_1 = x' \in \mathbf{X}$ according to the probability measure $Q(\cdot|x, a)$ and the decision-maker chooses a new admissible decision $a_1 = a\prime \in A(x')$ with a cost $C(x', a')$ and so on. We will refer to the processes $\{x_t\}$ and $\{a_t\}$ as the state and decision processes, respectively.

At each decision time $n \in \mathbb{N}_0$, the decision-maker chooses the decision variable a_n according to a deterministic or stochastic rule π_n, which may depend on the whole previous history $h_n = (x_0, a_0, \ldots, x_{n-1}, a_{n-1}, x_n)$ of the system. Naturally, the rules have to choose admissible actions with probability one, that is,

$$\pi_n(a_n \in A(x_n)|h_n) = 1 \quad \forall h_n \in \mathbb{H}_n,$$

where $\mathbb{H}_n := \mathbb{K}^n \times \mathbf{X}$ for $n \in \mathbb{N}$ and $\mathbb{H}_0 := \mathbf{X}$. Note that $\mathbb{H}_n$ is the set of all *admissible histories* up to time $n \in \mathbb{N}_0$. We will refer to the sequence $\pi = \{\pi_n\}$ as (admissible) *decision or control policy.* The class of all policies is denoted by Π.

Denote by $\mathbb{F}$ the class of measurable selectors, that is, the class of measurable functions $f : \mathbf{X} \to \mathbf{A}$ such that $f(x) \in A(x)$ for all $x \in \mathbf{X}$. A policy $\pi = \{\pi_n\}$ is called (deterministic) *stationary policy* if there exists $f \in \mathbb{F}$ such that the measure $\pi_n(\cdot|h_n)$ is concentrated at $f(x_n)$ for all history $h_n \in \mathbb{H}_n$ and $n \in \mathbb{N}_0$; thus, the policy $\pi = \{\pi_n)$ is identified with the selector f and the class of all stationary policies with the family of selectors $\mathbb{F}$.

Let $\Omega := (\mathbf{X} \times \mathbf{A})^\infty$ be the canonical sample space and $\mathcal{F}$ the product σ-algebra. It is well-known that for each decision policy $\pi = \{\pi_n\}$ and initial state $x_0 = x \in \mathbf{X}$ there exists a probability measure P_x^π on the measurable space $(\Omega, \mathcal{F})$ that governs the evolution of the *controlled process* $\{(x_n, a_n)\}$ induced by the policy $\pi = \{\pi_n\}$. The expectation operator with respect to the probability measure P_x^π is denoted as E_x^π.

The expected cost incurred in n steps when the policy $\pi = \{\pi_n\} \in \Pi$ is used and $x_0 = x$ is given as

$$J_n(\pi, x) := E_x^\pi \sum_{k=0}^{n-1} C(x_k, a_k).$$

Set $J_0^* = 0$ and let $J_n^*, n \in \mathbb{N}$, be the *n-stage optimal cost function*, that is,

$$J_n^*(x) := \inf_{\pi \in \Pi} J_n(\pi, x), \quad x \in \mathbf{X}. \tag{8}$$

The (expected) *average cost* is given by

$$J(\pi, x) := \limsup_{n\to\infty} \frac{1}{n} J_n(\pi, x), \quad x \in \mathbf{X}, \pi \in \Pi,$$

and the *optimal average control problem* is to find a policy $\pi^* = \{\pi_n^*\}$ such that

$$J(\pi^* x) = J^*(x) := \inf_{\pi\in\Pi} J(\pi, x) \quad \forall x \in \mathbf{X}.$$

If such a policy $\pi^* = \{\pi_n^*\}$ exists it is called (expected) *average optimal*, while J^* is called the (expected) *average cost optimal value function.*

We will use the following notation. For every measurable function v on $\mathbb{K}$ and stationary policy $f \in \mathbb{F}$, set

$$v_f(x) := v(x, f(x)), \quad x \in \mathbf{X}.$$

In particular, for the cost function and the transition law we write

$$C_f(x) := C(x, f(x)) \quad \text{and} \quad Q_f(\cdot|x) := Q(\cdot|x, f(x)), \quad x \in \mathbf{X}.$$

Using this notation, $Q_f^n(\cdot|\cdot)$ stands for the n-step transition probability of the Markov chain induced by the stationary policy $f \in \mathbb{F}$.

The main results of the present work are shown assuming that either one of two standard sets of continuity-compactness conditions holds. These set of conditions are given below in Assumption 2.2 and Assumption 2.3. The first set includes the continuity of the correspondence (or set-valued mapping) $x \to A(x)$. This concept is briefly presented below down.

Let Z and W be topological spaces. A correspondence Φ from Z to W is a (set-valued) mapping $\Phi : Z \to \mathcal{P}_0(W)$ where $\mathcal{P}_0(W)$ stands for the collection of all nonempty subsets of W. Let Φ be a correspondence from Z to W and define

$$\Phi^{-1}[D] := \{z \in Z : \Phi(z) \cap D \neq \emptyset\}$$

for each subset D of W. If $\Phi^{-1}[D]$ is closed (open, resp.) for each closed (open, resp.) subset D of W, it is said that Φ is *upper semicontinuous* (*lower semicontinuous*, resp.). If Φ is both upper and lower semicontinuous, it is said that Φ is a *continuous* correspondence.

The next remark provides a characterization using sequences of lower and upper semicontinuity of correspondences between metric spaces.

Remark 2.1. Suppose that Z and W are metric spaces and also that Φ be a correspondence from Z to W such that $\Phi(z)$ is a compact subset for each $w \in W$. Then:

(a) Φ is upper semicontinuous if and only if for each $z \in Z$ and all sequences $z_n \to z$ and $w_n \in \Phi(z_n), n \in \mathbb{N}$, there exist $w \in \Phi(z)$ and a subsequence $\{w_{n_k}\}$ of $\{w_n\}$ such that $w_{n_k} \to w$;

(b) Φ is lower semicontinuous if and only if for each $z \in Z$ and all sequences $z_n \to z$ and $w \in \Phi(z)$, there exist a subsequence $\{z_{n_k}\}$ of $\{z_n\}$ and $w_k \in \Phi(z_{n_k}), k \in \mathbb{N}$, such that $w_k \to w$.

For detailed discussions of semicontinuity of correspondences and a proof of Remark 2.1 see, for instance, [2, 29, 33].

Assumption 2.2.

(a) $C(\cdot,\cdot)$ is bounded by a constant $K > 0$;

(b) $A(x)$ is a compact subset of $\mathbf{A}$ for each $x \in \mathbf{X}$ and the correspondence $x \to A(x)$ is continuous;

(c) $C(\cdot,\cdot)$ is a continuous function on $\mathbb{K}$;

(d) $Q(\cdot|\cdot,\cdot)$ is weakly continuous on $\mathbb{K}$, that is, the mapping

$$(x,a) \to \int_{\mathbf{X}} u(y)Q(\mathrm{d}y|x,a)$$

is continuous for each function $u \in C_b(\mathbf{X})$.

Assumption 2.3.

(a) $C(\cdot,\cdot)$ is bounded by a constant $K > 0$;

moreover, the following holds for each $x \in \mathbf{X}$:

(b) $A(x)$ is a compact subset of $\mathbf{A}$;

(c) $C(x,\cdot)$ is a continuous function on $A(x)$;

(d) $Q(\cdot|x,\cdot)$ is strongly continuous on $A(x)$, that is, the mapping

$$a \to \int_{\mathbf{X}} u(y)Q(\mathrm{d}y|x,a)$$

is continuous for each function $u \in M_b(\mathbf{X})$.

We will assume that either Assumption 2.2 or Assumption 2.3 holds. Thus, $\mathcal{C}(\mathbf{X})$ will denote either the space $C_b(\mathbf{X})$ or $M_b(\mathbf{X})$ depending on whether Assumption 2.2 or Assumption 2.3 is being used, respectively. Thus, using this convention, it follows from [22, Proposition D.3, p. 130] that the *dynamic programming operator*

$$Tu(x) := \inf_{a\in A(x)}[C(x,a) + \int_{\mathbf{X}} u(y)Q(\mathrm{d}y|x,a)], \quad x \in \mathbf{X}, \tag{9}$$

maps the space $\mathcal{C}(X)$ into itself and that for each $u \in \mathcal{C}(\mathbf{X})$ there exists a selector f_u $\in \mathbb{F}$ such that

$$Tu(x) = C_{f_u}(x) + \int_{\mathbf{X}} u(y)Q_{f_u}(\mathrm{d}y|x) \quad \forall x \in \mathbf{X}.$$

We will refer to the policy f_u as *u-greedy policy.*

For each $f \in \mathbb{F}$ define the operator

$$T_f u(x) := C_f(x) + Q_f u(x), \quad x \in \mathbf{X},$$

and observe that it maps $M_b(\mathbf{X})$ into itself whenever the one-step cost function $C(\cdot,\cdot)$ is bounded.

A solution of the *average cost optimality equation* is a pair formed by a constant ρ^* and a measurable function h^* on $\mathbf{X}$ that satisfy the equation

$$\rho^* + h^* = Th^*. \tag{10}$$

If such solution exists and $h^* \in \mathcal{C}(\mathbf{X})$, then there is an h^*-greedy policy $f^* \in \mathbb{F}$, that is,

$$\rho^* + h^* = Th^* = T_{f^*}h^*. \tag{11}$$

The triplet (ρ^*, h^*, f^*) is called *canonical triplet.* Then, standard dynamic programming arguments show that the stationary policy f^* is an optimal policy and that the constant ρ^* is the optimal cost, that is, $J^* = J(f^*,\cdot) = \rho^*$. Moreover, if the pair (ρ, h) also enters in a canonical triplet, then $\rho = \rho^*$ and $h = h^* + k$ for a constant k.

The existence of a solution to the average cost optimality equation and the convergence of the value iteration algorithm are guaranteed under the ergodicity condition given in Assumption 2.4 below. This condition is given in term of the *total variation norm* for finite-signed measures λ on $\mathbf{X}$, which is defined as

$$||\lambda||_{TV} := \sup\left\{\Big|\int_{\mathbf{X}} v(y)\lambda(\mathrm{d}y)\Big|/||v||_\infty : v \in M_b(\mathbf{X}), v \neq 0\right\}. \tag{12}$$

It can be shown for any two probabilities measures P_1 and P_2 on $\mathbf{X}$ that

$$||P_1 - P_2||_{TV} = 2\sup\{|P_1(B) - P_2(B)| : B \in \mathcal{B}(\mathbf{X})\}. \tag{13}$$

Assumption 2.4. There exists a positive number $\alpha < 1$ such that

$$||Q(\cdot|x,a) - Q(\cdot|x',a')||_{TV} \leq 2\alpha \quad \forall (x,a),(x',a') \in \mathbb{K}. \tag{14}$$

Remark 2.5. (c.f. Hernández-Lerma [22, Lemma 3.3, p. 57], Meyn and Tweedie [32, Thm. 16.0.2, p. 384]) Assumption 2.4 implies that

$$\sup_{x\in\mathbf{X}} ||Q_f^n(\cdot|x) - \mu_f||_{TV} \leq 2\alpha^n \quad \forall f \in \mathbb{F}, n \in \mathbb{N}, \tag{15}$$

where μ_f is the (unique) *invariant* probability measure for the transition probability $Q_f(\cdot|\cdot)$, which means that

$$\mu_f(B) = \int_{\mathbf{X}} Q_f(B|x)\mu_f(\mathrm{d}x) \quad \forall B \in \mathcal{B}(\mathbf{X}).$$

Property (15) implies that

$$\sup_{x\in\mathbf{X}} |E_x^f u(x_n) - \mu_f(u)| \leq 2\alpha^n ||u||_\infty \tag{16}$$

for all $u \in M_b(\mathbf{X})$ and $f \in \mathbb{F}$, where

$$\mu_f(v) := \int_{\mathbf{X}} v(y)\mu_f(\mathrm{d}y).$$

Moreover, property (16) in turn leads to the equality

$$\lim_{n\to\infty} \frac{1}{n} E_x^f \sum_{k=0}^{n-1} u(x_k) = \mu_f(u) \tag{17}$$

for all $x \in \mathbf{X}$ and $u \in M_b(\mathbf{X})$. In particular,

$$J(f) := J(f,x) = \mu_f(C_f) \quad \forall x \in \mathbf{X}, f \in \mathbb{F}. \tag{18}$$

For further discussion on Assumption 2.4 and a proof of (15) the reader is referred to [22, Ch. 3.], [32, Ch. 16], [13, 25, 26].

The proof of our main results (Theorem 5.3, Section 5) uses the the following result borrowed from [35].

Remark 2.6. Let $S(\cdot|\cdot)$ and $R(\cdot|\cdot)$ be transition probabilities on $\mathbf{X}$. Define

$$\theta := \sup_{x\in\mathbf{X}} ||S(\cdot|x) - R(\cdot|x)||_{TV} \quad \text{and } \gamma := \frac{1}{2} \sup_{x,y\in\mathbf{X}} ||S(\cdot|x) - S(\cdot|y)||_{TV}.$$

If the transition probability $R(\cdot|\cdot)$ is uniformly ergodic and $\gamma < 1$, then

$$||\pi_S - \pi_R||_{TV} \leq \frac{\theta}{1-\gamma}$$

where π_S and π_R are the invariant probability measures for $S(\cdot|\cdot)$ and $R(\cdot|\cdot)$, respectively.

3. VALUE ITERATION ALGORITHM

Let $z \in \mathbf{X}$ be an arbitrary but fixed state and $M_b^0(\mathbf{X})$ be the subspace of functions $u \in M_b(\mathbf{X})$ satisfying the condition $u(z) = 0$. Similarly, the subspace $C_b^0(\mathbf{X})$ is the class of function $u \in C_b(\mathbf{X})$ with $u(z) = 0$. Usually, the spaces $M_b(\mathbf{X})$ and $C_b(\mathbf{X})$ are endowed with the supremum norm $||u||_\infty = \sup_x |u(x)|$, but here we also consider the span semi-norm defined as

$$||u||_{sp} := \sup_{x\in\mathbf{X}} u(x) - \inf_{x\in\mathbf{X}} u(x), \quad u \in M_b(\mathbf{X}).$$

Note that $||\cdot||_{sp}$ is a semi-norm on $M_b(\mathbf{X})$, but it becomes a norm when it is restricted to the subspace $M_b^0(\mathbf{X})$ or $C_b^0(\mathbf{X})$. Moreover, it holds that

$$||u||_\infty \leq ||u||_{sp} \quad \forall u \in M_b^0(\mathbf{X}), \tag{19}$$

so the subspaces $(M_b^0(\mathbf{X}), ||\cdot||_{sp})$ and $(C_b^0(\mathbf{X}), ||\cdot||_{sp})$ are Banach spaces. It also holds that

$$||u||_{sp} \leq 2||u||_\infty \quad \forall u \in M_b(\mathbf{X}).$$

Let (ρ^*, h^*) be a solution of the optimality equation (10) with $h^*(z) = 0$. Then, $\rho^* = Th^*(z)$ and the optimality equation can be rewritten as

$$h^*(x) = Th^*(x) - Th^*(z) \quad \forall x \in \mathbf{X}.$$

Thus, define the operator

$$T_z u(x) := Tu(x) - Tu(z), \quad x \in \mathbf{X}.$$

Let $\mathcal{C}_0(\mathbf{X})$ denote either $C_b^0(\mathbf{X})$ or $M_b^0(\mathbf{X})$ depending on whether Assumption 2.2 or Assumption 2.3 is being used. Under either one of these assumptions, T_z maps the subspace $\mathcal{C}_0(\mathbf{X})$ into itself and a function $h^* \in \mathcal{C}_0(\mathbf{X})$ satisfies the optimality equation if and only if it is a fixed point of T_z, that is,

$$h^*(x) = T_z h^*(x) = Th^*(x) - Th^*(z) \quad \forall x \in \mathbf{X}.$$

On the other hand, the n-stage optimal cost functions (8) satisfy the recursive equation (see, e. g., [22])

$$J^*_{n+1} = TJ^*_n \quad \forall n \in \mathbb{N}_0,$$

which leads to the equation

$$\rho_{n+1}(z) + h_{n+1} = Th_n \quad \forall n \in \mathbb{N}_0, \tag{20}$$

where h_n and $\rho_n, n \in \mathbb{N}$, are the functions introduced in (2), that is,

$$h_n = J^*_n - J^*_n(z) \quad \text{and} \quad \rho_n = J^*_n - J^*_{n-1}, \; n \in \mathbb{N}.$$

Notice that (20) can be rewritten as

$$h_{n+1} = T_z h_n = T_z^n h_0 \quad \forall n \in \mathbb{N}_0,$$

with $h_0 = 0$.

The *value iteration* (VI) *algorithm* is said to converge if the sequence $(\rho_n(z), h_n), n \in \mathbb{N}$, converges to a solution (ρ^*, h^*) of the average cost optimality equation. The convergence of the VI algorithm is established in Remark 3.1 and Theorem 3.2 below.

Remark 3.1. Suppose that Assumption 2.4 and either one of Assumptions 2.2 or 2.3 holds. Then:

(a) The operator T_z is a contraction from the Banach space $(\mathcal{C}_0(\mathbf{X}), || \cdot ||_{sp})$ into itself with modulus α (see [22, Lemma 3.5, p. 59.]). Thus, by the Banach fixed point theorem and (11), there exists a canonical triplet (ρ^*, h^*, f^*) with h^* belonging to $\mathcal{C}_0(\mathbf{X})$ and $\rho^* = Th^*(z)$.

(b) Moreover,

$$||T_z^n u - h^*||_\infty \leq ||T_z^n u - h^*||_{sp} \leq \alpha^n ||h^*||_{sp} \to 0$$

for all $u \in \mathcal{C}(\mathbf{X})$. In particular, taking $u \equiv 0$, we have

$$||h_n - h^*||_\infty \leq ||h_n - h^*||_{sp} \leq ||h^*||_{sp} \, \alpha^n \to 0.$$

To complete the proof of the convergence of the VI algorithm it only remains to establish the convergence of sequence $\rho_n(z), n \in \mathbb{N}$, to ρ^*. This is shown, among other useful facts, in the next theorem which comes from [22, Thm. 4.8, p. 64].

Theorem 3.2. Suppose that either one of Assumptions 2.2 or 2.3 holds, and also that Assumption 2.4 holds. Let h^*be as in Remark 3.1 and define the sequences

$$s_n := \inf_{x \in \mathbf{X}} \rho_n(x) \quad \text{and} \quad S_n := \sup_{x \in \mathbf{X}} \rho_n(x), n \in \mathbb{N}.$$

Then:

(a) $\{s_n\}$ is nondecreasing and $\{S_n\}$ is nonincreasing; moreover,

$$-\alpha^{n-1}||h^*||_{sp} \leq s_n - \rho^* \leq S_n - \rho^* \leq \alpha^{n-1}||h^*||_{sp} \quad \forall n \in \mathbb{N};$$

hence, in particular, $\rho_n(z) \to \rho^*$.

(b) If f is an h_n-greedy policy, then

$$0 \leq J(f) - \rho^* \leq ||\rho_n||_{sp} \leq \alpha^{n-1}||h^*||_{sp} \quad \forall n \in \mathbb{N}.$$

4. AVERAGERS AND PERTURBED MODELS

As mentioned above, the main concern of the present work are problems $\mathbf{P}_1 - \mathbf{P}_3$–stated in the Introduction–for the approximate value iteration algorithms (6) and (7). These problems are studied following the approach introduced in [53] to study the discounted cost criterion. This latter paper focus on a class of approximating operators called *averagers*, which are introduced next.

Definition 4.1. The operator $L : M_b(\mathbf{X}) \to M_b(\mathbf{X})$ is said to be an *averager* if it satisfies the following properties:

(a) $L\mathbb{I}_{\mathbf{X}} = \mathbb{I}_{\mathbf{X}}$;

(b) L is a *linear* operator;

(c) L is a *positive* operator, that is, $Lu \geq 0$ for each $u \geq 0$ in $M_b(\mathbf{X})$.

(e) L satisfies the following *continuity* property:

$$v_n \downarrow 0, v_n \in M_b(\mathbf{X}) \implies Lv_n \downarrow 0.$$

If in addition L maps $C_b(\mathbf{X})$ into itself it is called *continuous averager.*

The averagers is a rich class of approximation operators. As mentioned previously, it includes many of the approximation operators studied in approximation theory as piecewise constant approximation operators, linear and multilinear interpolators, kernel-based interpolators; see [20, 36, 46, 47]; certain aggregation-projection operators [49]; Schoenberg' splines, Hermite-Fejér and Bernstein operators [5, 12], among others.

The key point is that the approximating step in the AVI algorithms (6) and (7)–when L is an averager–can be seen as a perturbation of the original Markov model. To introduce the perturbed models we need the following simple but important result; in fact, the term averager comes from property (21).

Lemma 4.2. Suppose that L is an averager. Then, the mapping

$$L(B|x) := L\mathbb{I}_B(x), \quad x \in \mathbf{X}, B \in \mathcal{B}(\mathbf{X}),$$

is a transition probability on $\mathbf{X}$ and

$$\int_{\mathbf{X}} v(y)L(\mathrm{d}y|\cdot) = Lv \quad \forall v \in M_b(\mathbf{X}). \tag{21}$$

The proof of Lemma 4.2 is omitted because it follows standard arguments.

Remark 4.3. If L is an averager, then it is monotone and non-expansive with respect to $||\cdot||_\infty$ and $||\cdot||_{sp}$, that is, for all u, v belonging to $M_b(\mathbf{X})$, it holds that

$$||Lu - Lv||_\infty \leq ||u - v||_\infty \quad \text{and} \quad ||Lu - Lv||_{sp} \leq ||u - v||_{sp}.$$

The monotonicity property directly follows from property in Definition 4.1(c). Concerning the second statement, note that it suffices to check such property for an arbitrary function v due to the linearity of the operator L. Then, the monotonicity and the normalizing condition in Definition 4.1(a) implies that $-||v||_\infty \leq Lv \leq ||v||_\infty$ and also that $\inf_{x\in\mathbf{X}} v(x) \leq Lv \leq \sup_{x\in\mathbf{X}} v(x)$, which in turn lead to $||Lv||_\infty \leq ||v||_\infty$ and $||Lv||_{sp} \leq ||v||_{sp}$.

We next introduce two perturbed models $\widehat{M} = (\mathbf{X}, \mathbf{A}, \{A(x) : x \in \mathbf{X}\}, R, \widehat{Q})$ and $\widetilde{M} = (\mathbf{X}, \mathbf{A}, \{\widetilde{R}_f, \widetilde{Q}_f : f \in \mathbb{F}\})$, which are denoted by $\widehat{M}$ and $\widetilde{M}$ for short. Recall that M stands for the original Markov model. We also present in Remarks 4.6 and 4.7 a third perturbed model $\overline{M}$ for models with $A(x) = \mathbf{A}$ for all $x \in \mathbf{X}$.

The perturbed model $\widehat{M}$. In this model, the state and control spaces, the admissible decision sets, and the one-step cost function are as in the original model M. The transition law is defined as

$$\widehat{Q}(B|x, a) := \int_{\mathbf{X}} L(B|y)Q(\mathrm{d}y|x, a), \quad (x, a) \in \mathbb{K}, B \in \mathcal{B}(\mathbf{X}),$$

which is clearly a stochastic kernel on $\mathbf{X}$ given $\mathbb{K}$ because it is the composition of stochastic kernels.

Thus, given a policy $\pi \in \Pi$ and initial state $\widehat{x}_0 = x \in \mathbf{X}$, let $\{(\widehat{x}_k, \widehat{a}_k)\}$ be the resulting controlled process and $\widehat{P}_x^\pi$ the corresponding probability measure, which are defined on the measurable space $(\Omega, \mathcal{F})$. Let $\widehat{E}_x^\pi$ be the expectation operator with respect to such probability measure. The (expected) *average cost* and the *average optimal value* in the perturbed model $\widehat{M}$ are given as

$$\widehat{J}(\pi, x) := \limsup_{n\to\infty} \frac{1}{n} \widehat{E}_x^\pi \sum_{k=0}^{n-1} C(\widehat{x}_k, \widehat{a}_k), \quad x \in \mathbf{X}, \pi \in \Pi,$$

$$\widehat{J}_*(x) := \sup_{\pi\in\Pi} \widehat{J}(\pi, x), \quad x \in \mathbf{X},$$

respectively. A policy π^* is said to be *optimal* in the model $\widehat{M}$ if $\widehat{J}_* = \widehat{J}(\pi^*, \cdot)$.

The dynamic programming operator $\widehat{T}$ in the perturbed model $\widehat{M}$ is given as

$$\begin{aligned}\widehat{T}u(x) &:= \inf_{a \in A(x)} [C(x,a) + \int_{\mathbf{X}} u(y)\widehat{Q}(\mathrm{d}y|x,a)], \quad x \in \mathbf{X},\\ &= TLu(x).\end{aligned}$$

Moreover, define the operator

$$\widehat{T}_z u := \widehat{T}u - \widehat{T}u(z)$$

and the value iteration functions

$$\widehat{J}_n := \widehat{T}\widehat{J}_{n-1}, \quad \text{and} \quad \widehat{\rho}_n = \widehat{J}_n - \widehat{J}_{n-1}, \quad n \in \mathbb{N},$$

with $\widehat{J}_0 = 0$. Observe that the functions defined in (6) can also be expressed as

$$\widehat{h}_n = \widehat{J}_n - \widehat{J}_n(z) \quad \text{and} \quad \widehat{\rho}_n(z) = \widehat{T}\widehat{h}_n(z), \quad n \in \mathbb{N}.$$

Remark 4.4. **(a)** Suppose Assumption 2.2 holds and also that L is a continuous averager. Then, $\widehat{Q}(\cdot|\cdot,\cdot)$ is weakly continuous, $\widehat{T}(C_b(\mathbf{X})) \subset C_b(\mathbf{X})$ and $\widehat{T}_z(C_b(\mathbf{X})) \subset C_b^0(\mathbf{X})$. Moreover, for each $u \in C_b(\mathbf{X})$ there exists a policy $f \in \mathbb{F}$ such that $\widehat{T}u = \widehat{T}_f u$, where

$$\widehat{T}_f u(x) := C_f(x) + \int_{\mathbf{X}} u(y)\widehat{Q}_f(\mathrm{d}y|x) \quad x \in \mathbf{X}.$$

(b) Similarly, if L is an averager and Assumption 2.3 holds, then $\widehat{T}(M_b(\mathbf{X})) \subset M_b(\mathbf{X})$ and $\widehat{T}_z(M_b(\mathbf{X})) \subset M_b^0(\mathbf{X})$. Moreover, for each $u \in M_b(\mathbf{X})$ there exists a policy $f \in \mathbb{F}$ such that $\widehat{T}u = \widehat{T}_f u$.

The perturbed model $\widetilde{M}$. In this model, the state and control spaces and the admissible decision sets are the same of the orginal Markov model M. The perturbed one-step costs and transition laws are only defined for the class of stationary policies as follows:

$$\widetilde{C}_f := LC_f \quad \text{and} \quad \widetilde{Q}_f(B|\cdot) := LQ_f(B|\cdot)$$

for each $f \in \mathbb{F}$ and $B \in \mathcal{B}(\mathbf{X})$. By Lemma 4.2, $\widetilde{Q}_f(\cdot|\cdot), f \in \mathbb{F}$, is a transition probability on $\mathbf{X}$ because it is the composition of two of them:

$$\widetilde{Q}_f(B|x) = \int_{\mathbf{X}} Q_f(B|y)L(\mathrm{d}y|x), \quad \forall x \in \mathbf{X}, B \in \mathcal{B}(\mathbf{X}).$$

Thus, for each stationary policy $f \in \mathbb{F}$ and initial state $\widetilde{x}_0 = x \in \mathbf{X}$ there exists a Markov chain $\{\widetilde{x}_n\}$ and probability measure $\widetilde{P}_x^f$ defined both on the measurable space $(\Omega, \mathcal{F})$ such that $\widetilde{Q}_f(\cdot|\cdot)$ is the one-step transition probability of $\{\widetilde{x}_n\}$. The expectation operator with respect to $\widetilde{P}_x^f$ is denoted by $\widetilde{E}_x^f$. The corresponding *average cost criterion* and *optimal value* function are given as

$$\widetilde{J}(f,x) := \limsup_{n\to\infty} \frac{1}{n} \widetilde{E}_x^f \sum_{k=0}^{n-1} \widetilde{C}_f(\widetilde{x}_k), \quad x \in \mathbf{X}, f \in \mathbb{F},$$

$$\widetilde{J}_*(x) := \sup_{f\in\mathbb{F}} \widetilde{J}(f,x), \quad x \in \mathbf{X}.$$

A policy $f^* \in \mathbb{F}$ is said to be optimal in the perturbed model $\widetilde{M}$ if $\widetilde{J}_* = \widetilde{J}(f^*,\cdot)$.

The dynamic programming operator $\widetilde{T}$ in the model $\widetilde{M}$ is defined as

$$\widetilde{T}u(x) := \inf_{f\in\mathbb{F}} [\widetilde{C}_f + \int_{\mathbf{X}} u(y)\widetilde{Q}_f(\mathrm{d}y|x)], \quad x \in \mathbf{X},$$

for any $u \in M_b(\mathbf{X})$. Moreover, define

$$\widetilde{T}_z u := \widetilde{T}u(x) - \widetilde{T}u(z), \quad u \in M_b(\mathbf{X}),$$

and the value iteration functions

$$\widetilde{J}_n = \widetilde{T}\widetilde{J}_{n-1}, \ \widetilde{J}_0 = 0 \quad \text{and} \quad \widetilde{\rho}_n = \widetilde{J}_n - \widetilde{J}_{n-1}, n \in \mathbb{N}.$$

Functions in (7) can also be expressed as

$$\widetilde{h}_n = \widetilde{J}_n - \widetilde{J}_n(z) \quad \text{and} \quad \widetilde{\rho}_n(z) = \widetilde{T}h_n(z), \quad n \in \mathbb{N}.$$

Remark 4.5. (a) Suppose L is a continuous averager and that Assumption 2.2 holds. Then, the dynamic programming operator $\widetilde{T}$ maps $C_b(\mathbf{X})$ into itself and $\widetilde{T} = LT$. To verify the last fact, let $u \in C_b(\mathbf{X})$ be a fixed function and note that

$$Tu(\cdot) \le C_f(\cdot) + \int_{\mathbf{X}} u(y)Q_f(\mathrm{d}y|\cdot) \ \ \forall f \in \mathbb{F}.$$

Thus, using the monotonicity and linearity of L, Lemma 4.2 implies that

$$LTu(\cdot) \le \widetilde{C}_f(\cdot) + \int_{\mathbf{X}} u(y)\widetilde{Q}_f(\mathrm{d}y|\cdot).$$

On the other hand, recall that for each $u \in C_b(\mathbf{X})$ there exists $f_u \in \mathbb{F}$ such that

$$Tu(\cdot) = C_{f_u}(\cdot) + \int_{\mathbf{X}} u(y)Q_{f_u}(\mathrm{d}y|\cdot).$$

Then, using the linearity of L and Lemma 4.2 again, we see that

$$LTu(\cdot) = \widetilde{C}_{f_u}(\cdot) + \int_{\mathbf{X}} u(y)\widetilde{Q}_{f_u}(\mathrm{d}y|\cdot),$$

which yields that $\widetilde{T}u = LTu$. Hence, $\widetilde{T} = LT$.

(b) If L is an averager and Assumption 2.3 holds, then part (a) holds replacing $C_b(\mathbf{X})$ by $M_b(\mathbf{X})$.

Remark 4.6 If $A(x) = \mathbf{A}$ for all $x \in \mathbf{X}$, one can consider a third perturbed model $\overline{M} = (\mathbf{X}, \mathbf{A}, \overline{Q}, \overline{C})$ where

$$\overline{C}(x,a) := \int_X C(y,a)L(y|x) \quad \text{and} \quad \overline{Q}(B|x,a) := \int_X Q(B|y,a)L(y|x)$$

for $(x,a) \in \mathbf{X} \times \mathbf{A}$. Thus, the dynamic programming operator is

$$\overline{T}u(x) := \inf_{a \in \mathbf{A}} [\widetilde{C}(x,a) + \int_{\mathbf{X}} u(y)\overline{Q}(\mathrm{d}y|x,a)], \quad x \in \mathbf{X}.$$

Under the continuity assumptions in Remark 4.5, it can be easily seen that $\overline{T} = \widetilde{T}$. Hence, the value iteration functions in $\overline{M}$ and $\widetilde{M}$ coincide.

Remark 4.7 Consider again a model with $A(x) = \mathbf{A}$ for all $x \in \mathbf{X}$, and the averager given by a set-wise constant approximator. Thus, let $\{Z_i\}_{i=1}^N$ be a partition of the state space $\mathbf{X}$ formed by Borel measurable subsets and consider points $z_i \in Z_i$ for $i = 1, \ldots, N$. The partition can be defined following a nearest neighborhood quantization method–as in [46]–or any other procedure whenever it yields a partition with Borel measurable subsets.

Next, fix a reference probability measure ν on $\mathbf{X}$ such that $\nu(Z_i) > 0$ for each $i = 1, \ldots, N$ and define the probability measures

$$\nu_i(B) := \frac{\nu(B \cap Z_i)}{\nu(Z_i)}, \quad B \in \mathcal{B}(X), i = 1, \ldots, N.$$

Next consider the "setwise" constant interpolator

$$Lu := \sum_{i=1}^N \nu_i(u)\mathbb{I}_{Z_i},$$

where

$$\nu_i(u) := \int_X u(y)\nu_i(\mathrm{d}y) = \frac{1}{\nu(Z_i)} \int_{Z_i} u(y)\nu(\mathrm{d}y)$$

for bounded measurable functions $u : \mathbf{X} \to \mathbb{R}$. Clearly the constant interpolator L is an averager; thus, the mapping

$$L(B|x) := \sum_{i=1}^N \nu_i(B)\mathbb{I}_{Z_i}(x) \quad \forall B \in \mathcal{B}(\mathbf{X}), x \in \mathbf{X}$$

defines a transition kernel on $\mathbf{X}$. Moreover, observe that

$$\overline{C}(x,a) = \frac{1}{\nu(Z_i)} \int_{Z_i} C(y,a)\nu(\mathrm{d}y),$$

$$\overline{Q}(B|x,a) = \frac{1}{\nu(Z_i)} \int_{Z_i} Q(B|y,a)\nu(\mathrm{d}y),$$

for all $x \in Z_i, a \in \mathbf{A}, i = 1, \ldots, N$.

Notice that model $\overline{M}$ is essentially a finite-state model. In fact, Saldi, Yuksel and Linder [46] consider the model $\overline{M}_d := (\mathbf{X}_d, \mathbf{A}, p, c)$ where

$$\mathbf{X}_d := \{z_{1,} \ldots, z_N\},$$

$$c(z_j, a) := \overline{C}(z_j, a),$$

$$p(z_j | z_i, a) := \overline{Q}(Z_j | z_i, a),$$

for $a \in \mathbf{A}, i, j = 1, \ldots, N$.

We end the description of the perturbed models emphasizing that the approximate value iteration algorithms given in (6) and (7) are exactly the same that the standard value iteration algorithms in the perturbed models $\widehat{M}$ and $\widetilde{M}$, respectively.

5. CONVERGENCE AND PERFORMANCE BOUNDS

This section addresses problems $\mathbf{P}_1 - \mathbf{P}_3$ posed in the Introduction. The main point here is that models $\widehat{M}$ and $\widetilde{M}$ retain practically all the properties of the original model M. For instance, it is shown in Lemma 5.1 below that models $\widehat{M}$ and $\widetilde{M}$ satisfy the ergodicity property in Assumption 2.4 provided the original model M does. The proofs of all results stated in this section are given in Appendix, Section 7.

Lemma 5.1. Suppose that model M satisfies Assumption 2.4. Then,

$$||\widehat{Q}(\cdot|x, a) - \widehat{Q}(\cdot|x', a')||_{TV} \leq 2\alpha \quad \forall (x, a), (x', a') \in \mathbb{K},$$

$$||\widetilde{Q}_f(\cdot|x) - \widetilde{Q}_g(\cdot|x')||_{TV} \leq 2\alpha \quad \forall x, x' \in \mathbf{X}, f, g \in \mathbb{F}.$$

Lemma 5.1 implies that properties (15)-(18) hold in both models $\widehat{M}$ and $\widetilde{M}$, where $\widehat{\mu}_f$ and $\widetilde{\mu}_f$ are the invariant probabilities measures for $\widehat{Q}_f(\cdot|\cdot)$ and $\widetilde{Q}_f(\cdot|\cdot)$, respectively, for each $f \in \mathbb{F}$. In particular,

$$\widehat{J}(f) := \widehat{J}(f, \cdot) = \int_{\mathbf{X}} C_f(y) \widehat{\mu}_f(\mathrm{d}y),$$
$$\widetilde{J}(f) := \widetilde{J}(f, \cdot) = \int_{\mathbf{X}} \widetilde{C}_f(y) \widetilde{\mu}_f(\mathrm{d}y).$$

Lemma 5.1 combined with the continuity and compactness conditions (either Assumption 2.2 or Assumption 2.3) implies that operators $\widehat{T}_z$ and $\widetilde{T}_z$ are contractions from the space $\mathcal{C}_0(\mathbf{X})$ into itself. This later fact implies the existence of canonical triplets in the perturbed models and the convergence of the corresponding AVI algorithms as well, establishing thus the convergence asked in $\mathbf{P}_1$. These results are stated in the next theorem.

Theorem 5.2. Suppose Assumption 2.4 holds together either one of the following set of conditions:

(i) L is a continuous averager and Assumption 2.2 holds;

(ii) L is an averager and Assumption 2.3 holds.

Then:

(a) There exist canonical triplets $(\widehat{\rho}, \widehat{h}, \widehat{f})$ and $(\widetilde{\rho}, \widetilde{h}, \widetilde{f})$ for models $\widehat{M}$ and $\widetilde{M}$, respectively, where $\widehat{h}$ and $\widetilde{h}$ are functions in $\mathcal{C}_0(\mathbf{X})$;

(b) $\widehat{\rho} = \widetilde{\rho}$, $\widehat{h} - T\widetilde{h} = k_1$ and $L\widehat{h} - \widetilde{h} = k_2$, where k_1 and k_2 are constants;

(c) the AVI algorithms (6) and (7) converge to $(\widehat{\rho}, \widehat{h})$ and $(\widetilde{\rho}, \widetilde{h})$, respectively.

Now, to address problems $\mathbf{P}_2 - \mathbf{P}_3$, first it is needed to specify how the accuracy of averagers is measured. Recall that the goal is to provided performance bounds for the AVI algorithms in terms of the accuracy of the approximations provided by L for the primitive data of the original Markov model, namely, the one-step cost function $C(\cdot, \cdot)$ and the transition law $Q(\cdot|\cdot, \cdot)$.

To this end, let $\widehat{\mathbb{F}}_0$ be a subset of stationary policies that contains the subclasses $\mathbb{F}_* := \{f \in \mathbb{F} : Th^* = T_f h^*\}$, $\widehat{\mathbb{F}}_1 := \{f \in \mathbb{F} : TL\widehat{h} = T_f L\widehat{h}\}$, $\widehat{\mathbb{F}}_2 := \{f \in \mathbb{F} : TL\widehat{h}_n = T_f L\widehat{h}_n$ for some $n \in \mathbb{N}\}$. Similarly, let $\widetilde{\mathbb{F}}_0$ be a subset of stationary policies that contains $\mathbb{F}_*$ and the subclasses $\widetilde{\mathbb{F}}_1 := \{f \in \mathbb{F} : T\widetilde{h} = T_f\widetilde{h}\}$ and $\widetilde{\mathbb{F}}_2 := \{f \in \mathbb{F} : T\widetilde{h}_n = T_f\widetilde{h}_n$ for some $n \in \mathbb{N}\}$.

The accuracy of the approximations given by an operator L is measured by the constant

$$\delta_Q(\widehat{\mathbb{F}}_0) := \sup_{x \in \mathbf{X}, f \in \widehat{\mathbb{F}}_0} ||Q_f(\cdot|x) - \widehat{Q}_f(\cdot|x)||_{TV}, \tag{22}$$

in the model $\widehat{M}$, and by the constants

$$\delta_C(\widetilde{\mathbb{F}}_0) := \sup_{f \in \widetilde{\mathbb{F}}_0} ||C_f - \widetilde{C}_f||_\infty, \quad \delta_Q(\widetilde{\mathbb{F}}_0) := \sup_{x \in \mathbf{X}, f \in \widehat{\mathbb{F}}_0} ||Q_f(\cdot|x) - \widetilde{Q}_f(\cdot|x)||_{TV},$$

in the model $\widetilde{M}$.

With this notation we can give the bounds as asked in $\mathbf{P}_2$.

Theorem 5.3. Suppose that assumptions in Theorem 5.2 hold and let $\sigma^* := \widehat{\rho} = \widetilde{\rho}$. Then:

(a) $|\rho^* - \sigma^*| \leq \dfrac{K}{1-\alpha}\delta_Q(\widehat{\mathbb{F}}_0)$;

(b) $||h - \widehat{h}||_\infty \leq \dfrac{2K}{(1-\alpha)^2}\delta_Q(\widehat{\mathbb{F}}_0)$;

(c) $|\rho^* - \sigma^*| \leq 2[\delta_C(\widetilde{\mathbb{F}}_0) + \dfrac{K}{1-\alpha}\delta_Q(\widetilde{\mathbb{F}}_0)]$;

(d) $||h - \widetilde{h}||_\infty \leq \dfrac{2K}{(1-\alpha)^2}[\delta_Q(\widetilde{\mathbb{F}}_0) + \delta_Q(\widetilde{\mathbb{F}}_0)]$.

The next proposition provides bounds for the performance in the model M of policies that are canonical in models $\widehat{M}$ and $\widetilde{M}$. The proof is omitted because it is pratically the same of Theorem 5.5 below.

Proposition 5.4. Suppose that assumptions in Theorem 5.2 hold. Then:

(a) If $f \in \mathbb{F}$ is $L\widehat{h}$-greedy, that is, $TL\widehat{h} = T_f L\widehat{h}$, then

$$0 \leq J(f) - \rho^* \leq \frac{2K}{1-\alpha}\delta_Q(\widehat{\mathbb{F}}_0);$$

(b) If $g \in \mathbb{F}$ is $\widetilde{h}$-greedy, that is, $T\widetilde{h} = T_g\widetilde{h}$, then

$$0 \leq J(g) - \rho^* \leq 2[\delta_C(\widetilde{\mathbb{F}}_0) + \frac{K}{1-\alpha}\delta_Q(\widetilde{\mathbb{F}}_0)].$$

Finally, the next theorem gives the performance bounds asked in $\mathbf{P}_3$.

Theorem 5.5. Suppose that assumptions in Theorem 5.2. Then:

(a) If $f \in \mathbb{F}$ is $L\widehat{h}_n$-greedy, that is, $TL\widehat{h}_n = T_f L\widehat{h}_n$, then

$$0 \leq J(f) - \rho^* \leq ||\widehat{\rho}_n||_{sp} + \frac{2K}{1-\alpha}\delta_Q(\widehat{\mathbb{F}}_0);$$

(b) If $g \in \mathbb{F}$ is $\widetilde{h}_n$-greedy, that is, $T\widetilde{h}_n = T_g\widetilde{h}_n$, then

$$0 \leq J(g) - \rho^* \leq ||\widetilde{\rho}_n||_{sp} + 2[\delta_C(\widetilde{\mathbb{F}}_0) + \frac{K}{1-\alpha}\delta_Q(\widetilde{\mathbb{F}}_0)].$$

Remark 5.6. In a first comparison of parts (a) and (c)–or (b) and (d)–in Theorem 5.3 it would seem that perturbed model $\widetilde{M}$ gives better bounds than model $\widehat{M}$, but it may not be the case. For instance, suppose that $Q(B|x,a) = 0$ for each discrete subset $B \subset \mathbf{X} := [a,b]$ and $(x,a) \in \mathbb{K}$, and also that L is an interpolator on the grid $s_0 = a < s_1 < \cdots < s_N = b$, so $v(s_i) = Lv(s_i)$ for $i = 0,\ldots,N$, for each function v on $\mathbf{X}$. Then, $L\mathbb{I}_{B_1} = 0$ for $B_1 := \mathbf{X}\backslash\{s_0, s_1, \ldots, s_N\}$, and so $\widehat{Q}(B_1|x,a) = 0$. Thus, $\delta_Q(\widehat{\mathbb{F}}_0) = 2$ independently how fine the grid is.

The above fact is an example of a well-known "anomaly"of the total variation norm: it is too strong to measure the closeness among a discrete measure and a continuous one. One can go around this obstacle considering the 1-Wassertein distance but paying the cost of imposing strong Lipschitz-continuity conditions on the control model; see, for instance, references [16, 46]. The problem can also be dodged using the approximating model $\widetilde{M}$ in lieu of the model $\widehat{M}$, since the former one preserves the discrete or continuous nature of measure Q. In fact, in the next section, the model $\widetilde{M}$ is used to compute numerical approximations for an inventory system.

6. AN EXAMPLE FROM INVENTORY SYSTEMS

The goal of this section is to illustrate the approach developed previously with some numerical results. It was chosen an inventory system for which is known an analytical solution with the end of contrasting the numerical solutions and the performance bounds of the approximate algorithm with the exact solution.

Then, consider an single item inventory system with no backlog, no set-up cost and finite capacity θ. Let x_n and a_n be the stock of the item and the amount of it ordered to the production unit at the beginning of the nth-stage, and w_n the product's demand during that period. It is assumed the quantity a_n is immediately supplied at beginning of nth-stage. Since excess demand is not backlogged, the stock evolves according to

$$x_{n+1} = (x_n + a_n - w_n)^+, n \in \mathbb{N}_0,$$

where $x_0 = x$ and $v^+ := \max(v, 0)$ for each real number v. The demand process $\{w_n\}$ is formed by independent and identically distributed nonnegative random variables with distribution function F. The state and control spaces are $\mathbf{X} = \mathbf{A} = [0, \theta]$ and the admissible control set for state $x \in \mathbf{X}$ is $A(x) = [0, \theta - x]$.

The one-step cost function is

$$C(x, a) = pE_{w_0}(w_0 - x - a)^+ + h(x + a) + ca, \quad (x, a) \in \mathbb{K},$$

where E_{w_0} denotes the expectation operator with respect to the distribution F of the random variable w_0 and the constants p, h and c stand for the unit penalty cost for unmet demand, the unit holding cost for the stock at hand and the unit production cost, respectively.

The transition law of the system is

$$Q(B|x, a) = \Pr[(x + a - w_0)^+ \in B], \quad B \in \mathcal{B}(\mathbf{X}), (x, a) \in \mathbb{K}.$$

In order to guarantee Assumptions 2.2 and 2.4 hold, we suppose that the inventory system satisfies the following conditions.

Assumption 6.1.

(a) The random variable w_0 has finite expectation;

(b) $F(\theta) < 1$;

(c) the distribution function F has a density ρ which is bounded and Lipschitz continuous on $[0, \theta]$ with module l, that is,

$$|\rho(x) - \rho(y)| \leq l|x - y| \quad \forall x, y \in [0, \theta];$$

(d) $p > c > 0$ and $h > 0$.

Assumption 6.1(a) implies that one-step cost function $C(\cdot, \cdot)$ is finite. It is also continuous and bounded because of the bounded convergence theorem and the compactness of $\mathbb{K}$. Using Remark 2.1 it can be easily proved that the correspondence $x \to A(x)$ is continuous too. Moreover, note that the equality

$$\int_{\mathbf{X}} v(y) Q(\mathrm{d}y|x,a) = E_{w_0} v((x+a-w_0)^+)$$

holds for all $(x,a) \in \mathbb{K}$ and $v \in M_b(\mathbf{X})$. This equality and the bounded convergence theorem imply that the mapping

$$(x,a) \to \int_{\mathbf{X}} v(y) Q(\mathrm{d}y|x,a)$$

is continuous for each $v \in C_b(\mathbf{X})$. Therefore, the inventory system satisfies Assumption 2.2.

On the other hand, Assumption 2.4 follows from Assumption 6.1(b) and the inequality

$$\begin{aligned} Q(\{0\}|x,a) &= \Pr[w_0 \geq x+a] \\ &\geq 1 - F(\theta) > 0 \quad \forall (x,a) \in \mathbb{K}. \end{aligned}$$

In fact, if $0 \in B$ then $Q(B|x,a) \geq 1 - F(\theta)$ for all $(x,a) \in \mathbb{K}$; if $0 \notin B$, then $Q(B|x,a) \leq Q((0,\theta]|x,a) \leq F(\theta)$. Hence,

$$|Q(B|x,a) - Q(B|x',a')| \leq F(\theta) \quad \forall (x,a) \in \mathbb{K}, B \in \mathcal{B}(\mathbf{X}),$$

which in turn, from (13), implies that Assumption 2.4 holds with $\alpha = F(\theta)$.

Now, consider the linear interpolation scheme with evenly spaced nodes $s_0 = 0 < s_1 < \ldots < s_N = \theta$. Thus, the approximating operator L is given as

$$Lv(x) = b_i(x) v(s_i) + \bar{b}_i(x) v(s_{i+1}), \; x \in [s_i, s_{i+1}],$$

for each function $v \in M_b(\mathbf{X})$, where

$$b_i(x) := \frac{s_{i+1} - x}{s_{i+1} - s_i} \quad \text{and} \quad \bar{b}_i(x) := 1 - b_i(x), \quad x \in [s_i, s_{i+1}],$$

for $i = 0, \ldots, N-1$. The operator L is clearly a continuous averager, that is, it is an averager that maps $C_b(\mathbf{X})$ into itself.

In the above inventory model and its perturbation as well, there exist base-stock average cost optimal policies. This can be shown following, for instance, the arguments given in [54] combined with the geometric ergodicity property (15). Recall, that a stationary policy f is a base-stock policy if $f(x) = S - x$ for $x \in [0, S]$, and $f(x) = 0$ otherwise, where the constant $S \geq 0$ is the so-called re-order point. In fact, the optimal re-order point S^* in the original inventory model satisfies the equation

$$F(S) = \frac{p - h - c}{p - c}$$

if $p > h + c$, and $S^* = 0$ otherwise. Moreover, the optimal average cost is

$$\rho^* = p E_{w_0} (w_0 - S^*)^+ + h S^* + c E_{w_0} \min(S^*, w_0).$$

In view of Remark 5.5, we only consider the perturbed model $\widetilde{M}$ and take $\widetilde{\mathbb{F}}_0$ as the class of the base-stock policies. Then, Assumption 6.1(c) implies that

$$\delta_C(\widetilde{\mathbb{F}}_0) = \sup_{f \in \widetilde{\mathbb{F}}_0} ||C_f - \widetilde{C}_f||_\infty \leq (p + h + c)\triangle s,$$

$$\delta_Q(\widetilde{\mathbb{F}}_0) = \sup_{f \in \widetilde{\mathbb{F}}_0} ||Q_f - \widetilde{Q}_f||_{TV} \leq (2\theta l + 4M')\triangle s,$$

where M' is a bound of the density ρ and $\triangle s := \theta/N$. These bounds follow by elementary but cumbersome computations, so their derivation is omitted.

tol. $\varepsilon = 10^{-4}$	$N = 10^2$	$N = 10^3$	$N = 5 \times 10^3$	$N = 10^4$
n_*	7	7	7	7
S_{n_*}	22	21.975	21.975	21.9725
$\delta_C(\widetilde{\mathbb{F}}_0)$	1.25	0.125	0.025	0.0125
$\delta_Q(\widetilde{\mathbb{F}}_0)$	0.08125	0.008125	0.001625	0.0008125
A_E	40.60904	4.060904	0.8121807	0.4060904
T_E	40.60906	4.060929	0.8122065	0.4061161

Tab. 1. Approximated optimal policies and performance bounds for the AVI algorithm with a linear interpolation scheme with $N+1$ evenly spaced nodes for the inventory system with exponentially distributed demand with parameter $\lambda = 0.05$, and parameters $c = 1.5, h = 0.5, p = 3, \theta = 25$.

The numerical results displayed in Tables 1 and 2, and Figure 1 correspond to the parameter values $c = 1.5, h = 0.5, p = 3, \theta = 25$. The product's demand has an exponential density ρ with parameter $\lambda = 0.05$. Note that in this case, the density ρ is bounded by $M' = \lambda = 0.05$ and also that it has Lipschitz module $l = \lambda^2 = 0.0025$. With these parameter values, the above bounds become in

$$\delta_C(\widetilde{\mathbb{F}}_0) \leq 5\triangle s \quad \text{and} \quad \delta_Q(\widetilde{\mathbb{F}}_0) \leq 0.325\triangle s.$$

Moreover, the optimal re-order point and the optimal average cost are $S^* = 21.97225$ and $\rho^* = (c+h)\lambda^{-1} + hS^* = 50.98612$, respectively.

The approximate value iteration algorithm was stopped once the stopping error $S_E := ||\widetilde{\rho}_n||_{sp}$ falls below the tolerance $\varepsilon = 10^{-4}$. Let n_* be the first time S_E is less than ε and $S(n_*)$ the re-order point of the policy $\widetilde{h}_{n_*}$-greedy. Table 1 and Figure 1 show that the algorithm converges very fast and practically gets the true optimal policy. The quantities $A_E := 2[\delta_C(\widetilde{\mathbb{F}}_0) + K(1-\alpha)^{-1}\delta_Q(\widetilde{\mathbb{F}}_0)]$ and $T_E := A_E + S_E$ are bounds for the approximation error and the total error, respectively.

The optimal average cost ρ^* is approximated by the sequences $\widetilde{s}_n := \inf_{x \in \mathbf{X}} \widetilde{\rho}_n(x)$ and $\widetilde{S}_n := \sup_{x \in \mathbf{X}} \widetilde{\rho}_n(x), n \in \mathbb{N}$. Table 2 display the values of these sequences for the grid with $N = 10^3$.

n	2	3	4	5	6	7
$\widetilde{S}_n$	51.79333	51.03147	50.98637	50.98612	50.98612	50.98612
$\widetilde{s}_n$	42.80326	49.35696	50.80725	50.97339	50.98547	50.98610

Tab. 2. Approximated optimal re-order points with a linear interpolation scheme with $10^3 + 1$ evenly spaced nodes for the inventory system with exponentially distributed demand with parameter $\lambda = 0.05$, and parameters $c = 1.5, h = 0.5, p = 3, \theta = 25$.

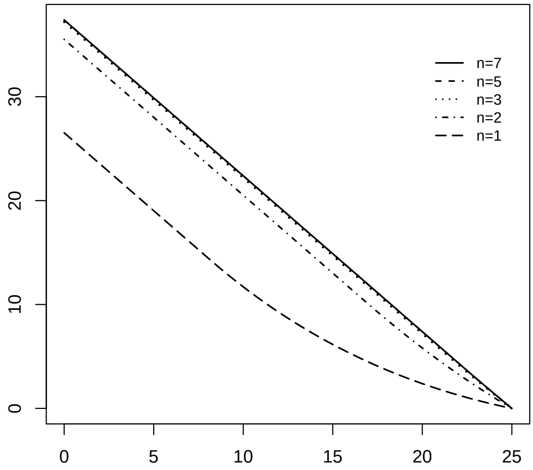

Fig. 1. Functions $\widetilde{h}_n$ with $N = 10^3$ and $\theta = 25$.

Figures 2 and 3 show functions $\widetilde{h}_n$ and sequences $\widetilde{s}_n$ and $\widetilde{S}_n$, respectively, for $\theta = 120$ and $N = 10^3$.

7. APPENDIX

Proof of Lemma 5.1. Assumption 2.4 can be equivalenty rewritten as

$$\left|\int_{\mathbf{X}} v(y)Q(\mathrm{d}y|x,a) - \int_{\mathbf{X}} v(y)Q(\mathrm{d}y\right| x',a')| \le 2\alpha||v||_\infty$$

for all $(x,a) \in \mathbb{K}, v \in M_b(\mathbf{X})$. Now, taking $v = Lu$ with $u \in M_b(\mathbf{X})$, the above inequality yields

$$\left|\int_{\mathbf{X}} Lu(y)Q(dy|x,a) - \int_{\mathbf{X}} Lu(y)Q(\mathrm{d}y\right| x',a')| \le 2\alpha||Lu||_\infty.$$

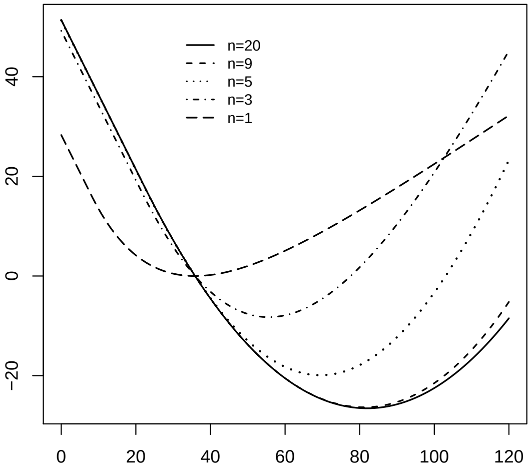

Fig. 2. Functions $\widetilde{h}_n$ with $N = 10^3$ and $\theta = 120$.

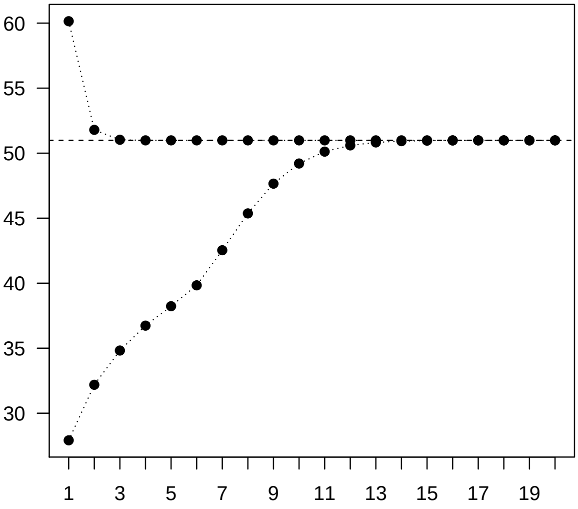

Fig. 3. Sequences $\widetilde{S}_n, \widetilde{s}_n$ with $N = 10^3$ and $\theta = 120$.

From definition of $\widehat{Q}(\cdot|\cdot,\cdot)$, (21) and the non-expansiveness property of L, it follows that

$$\left|\int_{\mathbf{X}} u(y)\widehat{Q}(\mathrm{d}y|x,a) - \int_{\mathbf{X}} u(y)\widehat{Q}(\mathrm{d}y|x',a')\right| \leq 2\alpha||u||_\infty,$$

which leads to

$$||\widehat{Q}(\cdot|x,a) - \widehat{Q}(\cdot|x',a')||_{TV} \leq 2\alpha \ \ \forall (x,a),(x',a') \in \mathbb{K}.$$

To prove the second inequality fix policies $f,g \in \mathbb{F}$ and $B \in \mathcal{B}(\mathbf{X})$. Assumption 2.4 implies that

$$-\alpha + Q_g(B|y) \leq Q_f(B|x) \leq Q_g(B|y) + \alpha \quad \forall x,y \in \mathbf{X}.$$

Fixing $y \in \mathbf{X}$, properties in Definition 4.1(a)-(c) imply

$$-\alpha + Q_g(B|y) \leq LQ_f(B|x) \leq Q_g(B|y) + \alpha \ \ \forall x \in \mathbf{X}.$$

Now fixing $x \in \mathbf{X}$, the latter inequality implies

$$-\alpha + LQ_g(B|y) \leq LQ_f(B|x) \leq LQ_g(B|y) + \alpha \ \ \forall y \in \mathbf{X}.$$

Then, from definition of $\widetilde{Q}_f, \widetilde{Q}_g$ and (21), it follows that

$$|\widetilde{Q}_f(B|x) - \widetilde{Q}_g(B|y)| \leq \alpha \quad \forall x,y \in \mathbf{X}, B \in \mathcal{B}(\mathbf{X}).$$

Hence, from (13),

$$||\widetilde{Q}_f(\cdot|x) - \widetilde{Q}_g(\cdot|y)||_{TV} \leq 2\alpha \quad \forall x,y,f,g \in \mathbb{F}.$$

□

Proof of Theorem 5.2. Under Assumption of Theorem 5.2, model $\widehat{M}$ satisfies all conditions in Remark 3.1 and Theorem 3.2 (see Remark 4.4). Thus, $\widehat{T}_z$ is contraction from $\mathcal{C}_0(\mathbf{X})$ into itself with contraction modulus α and there exists a canonical triplet $(\widehat{\rho},\widehat{h},\widehat{f})$ in the model $\widehat{M}$ with $\widehat{h}$ belonging to $\mathcal{C}_0(\mathbf{X})$ and $\widehat{\rho} = \widehat{T}\widehat{h}(z)$; moreover, all the conclusions in Theorem 3.2 hold replacing $\rho^*, h^*, s_n, S_n, h_n$ and $J(f)$ with $\widehat{\rho}, \widehat{h}, \widehat{s}_n, \widehat{S}_n, \widehat{h}_n$ and $\widehat{J}(f)$, respectively, where

$$\widehat{s}_n := \inf_{x\in\mathbf{X}} \widehat{\rho}_n \quad \text{and} \quad \widehat{S}_n := \sup_{x\in\mathbf{X}} \widehat{\rho}_n, \quad n \in \mathbb{N}.$$

Regarding to model $\widetilde{M}$, the contractiveness property of operator $\widetilde{T}_z$ on $\mathcal{C}_0(\mathbf{X})$ and the existence of a canonical triplet $(\widetilde{\rho},\widetilde{h},\widetilde{f})$ with $\widetilde{h}$ in $\mathcal{C}_0(\mathbf{X})$ and $\widetilde{\rho} = \widetilde{T}\widetilde{h}(z)$ follow from the same arguments given in [22, Lemma 3.5, p. 59] for the proof of Remark 3.1, while the convergence of the value iteration algorithm in the model $\widetilde{M}$ follows from the arguments given in [22, Thm. 4.8, p. 64] for Theorem 3.2 replacing $\rho^*, h^*, s_n, S_n, h_n$ and $J(f)$ with $\widetilde{\rho}, \widetilde{h}, \widetilde{s}_n, \widetilde{S}_n, \widetilde{h}_n$ and $\widetilde{J}(f)$, respectively, where

$$\widetilde{s}_n := \inf_{x\in\mathbf{X}} \widetilde{\rho}_n(x) \quad \text{and} \quad \widetilde{S}_n := \sup_{x\in\mathbf{X}} \widetilde{\rho}_n(x), \quad n \in \mathbb{N}.$$

Part (c) follows directly since the AVI algorithms (6) and (7) coincide with the standard value iteration algorithm in the model $\widehat{M}$ and $\widetilde{M}$, respectively.

Thus, it only remains to prove part (b). To do this, first note that $\widehat{\rho}+\widehat{h} = \widehat{T}\widehat{h} = TL\widehat{h}$; then, $\widehat{\rho}+L\widehat{h} = LT(L\widehat{h})$, which in turn implies that function $v = L\widehat{h}$ satisfies the equation

$$\widehat{\rho} + v = \widetilde{T}v.$$

Hence, $\widehat{\rho} = \widetilde{\rho}$ and functions $v = L\widehat{h}$ and $\widetilde{h}$ differ by a constant.

On the other hand, the pair $(\widetilde{\rho}, \widetilde{h})$ satisfies the equalities $\widetilde{\rho} + \widetilde{h} = \widetilde{T}\widetilde{h} = LT\widetilde{h}$. Now, observe that $\widetilde{\rho} + T\widetilde{h} = TL(T\widetilde{h})$, which yields that the function $u := T\widetilde{h}$ satisfies the equation

$$\widetilde{\rho} + u = \widehat{T}u,$$

which implies that $u = T\widetilde{h}$ and $\widehat{h}$ differ by a constant because $\widetilde{\rho} = \widehat{\rho}$. □

Lemma 7.1. Let $f \in \mathbb{F}$ be an arbitrary stationary policy and K the bound of $C(\cdot,\cdot)$. Then:

(a) $\sigma_f := \widehat{J}(f) = \widetilde{J}(f)$;

(b) $|J(f) - \sigma_f| \leq \frac{K}{1-\alpha} \sup_{x\in\mathbf{X}} ||Q_f(\cdot|x) - \widehat{Q}_f(\cdot|x)||_{TV}$;

(c) $|J(f) - \sigma_f| \leq ||C_f - \widetilde{C}_f||_\infty + \frac{K}{1-\alpha} \sup_{x\in\mathbf{X}} ||Q_f(\cdot|x) - \widetilde{Q}_f(\cdot|x)||_{TV}$.

Proof of Lemma 7.1. To prove part (a), let $f \in \mathbb{F}$ be a fixed policy and define the operators $\widehat{T}_f := T_f L$ and $\widetilde{T}_f := LT_f$. From Lemma 5.1, it follows that operators

$$\begin{aligned}\widehat{T}_{z,f}v &:= \widehat{T}_f v - \widehat{T}_f v(z),\\ \widetilde{T}_{z,f}v &:= \widetilde{T}_f v - \widetilde{T}_f v(z)\end{aligned}$$

are contractions operators from $M_b^0(\mathbf{X})$ into itself with contraction modulus α. Hence, there exist functions $\widehat{h}_f$ and $\widetilde{h}_f$ in $M_b^0(\mathbf{X})$ that satisfies the Poisson equations

$$\begin{aligned}\widehat{\rho}_f + \widehat{h}_f &= C_f + \widehat{Q}_f\widehat{h}_f\\ \widetilde{\rho}_f + \widetilde{h}_f &= \widetilde{C}_f + \widetilde{Q}_f\widetilde{h}_f,\end{aligned} \tag{23}$$

where $\widehat{\rho}_f := \widehat{T}_f\widehat{h}_f(z)$ and $\widetilde{\rho}_f := \widetilde{T}_f\widetilde{h}_f(z)$. Clearly, it holds that $\widehat{J}(f) = \widehat{\rho}_f$ and $\widetilde{J}(f) = \widetilde{\rho}_f$. Equation (23) implies that

$$\begin{aligned}\widehat{\rho}_f + L\widehat{h}_f &= LC_f + L\widehat{Q}_f\widehat{h}_f\\ &= LC_f + LQ(L\widehat{h}_f)\\ &= \widetilde{C}_f + \widetilde{Q}(L\widehat{h}_f).\end{aligned}$$

Hence, $\widehat{\rho}_f = \widetilde{\rho}_f$, which proves part (a). Additionally, note that functions $L\widehat{h}_f$ and $\widehat{h}_f$ differ only by a constant.

To prove part (b), note that property (18) and Lemma 5.1 imply

$$|J(f)-\widehat{J}(f)| = |\mu_f(C_f)-\widehat{\mu}_f(C_f)| \le K||\mu_f-\widehat{\mu}_f||_{TV}.$$

Now taking $S(\cdot|\cdot)=Q_f(\cdot|\cdot)$, $R(\cdot|\cdot)=\widehat{Q}_f(\cdot|\cdot)$, $\theta=\sup_{x\in\mathbf{X}}||Q_f(\cdot|x)-\widetilde{Q}_f(\cdot|x)||_{TV}$ and $\gamma=\frac{1}{2}\sup_{x,y\in\mathbf{X}}||Q_f(\cdot|x)-Q_f(\cdot|y)||_{TV}$, Remark 2.6 yields

$$|J(f)-\widehat{J}(f)| \le \frac{K}{1-\gamma}\sup_{x\in\mathbf{X}}||Q_f(\cdot|x)-\widetilde{Q}_f(\cdot|x)||_{TV},$$

which implies the first inequality because $\gamma\le\alpha$.

The inequality (b) follows in a similar manner. In fact, note that

$$\begin{aligned}|J(f)-\widetilde{J}(f)| &\le |\mu_f(C_f)-\mu_f(\widetilde{C}_f)|+|\mu_f(\widetilde{C}_f)-\widetilde{\mu}_f(\widetilde{C}_f)|\\ &\le ||C_f-\widetilde{C}_f||_\infty + K\ ||\mu_f-\widetilde{\mu}_f||_{TV}.\end{aligned}$$

Now use Remark 2.6 again but now taking $R(\cdot|\cdot)=\widetilde{Q}_f(\cdot|\cdot)$ to obtain the desire result.□

Proof of Theorem 5.3. Parts (a) and (c) follow directly from Lemma 7.1 after noting that

$$\begin{aligned}|\rho^*-\sigma^*| &= |\inf_{f\in\widehat{\mathbb{F}}_0} J(f)-\inf_{f\in\widehat{\mathbb{F}}_0}\widehat{J}(f)| \le \sup_{f\in\widehat{\mathbb{F}}_0}|J(f)-\widehat{J}(f)|,\\ |\rho^*-\sigma^*| &= |\inf_{f\in\widetilde{\mathbb{F}}_0} J(f)-\inf_{f\in\widetilde{\mathbb{F}}_0}\widetilde{J}(f)| \le \sup_{f\in\widetilde{\mathbb{F}}_0}|J(f)-\widetilde{J}(f)|.\end{aligned}$$

To prove part (b) recall that h and $\widehat{h}$ are the unique fixed points of T_z and $\widehat{T}_z$ in the space $(\mathcal{C}_0(\mathbf{X}),||\cdot||_{sp})$, respectively. Then,

$$\begin{aligned}||h-\widehat{h}||_{sp} &= ||T_zh-\widehat{T}_z\widehat{h}||_{sp}\\ &\le ||T_zh-T_z\widehat{h}||_{sp}+||T_z\widehat{h}-\widehat{T}_z\widehat{h}||_{sp}\\ &\le \alpha||h-\widehat{h}||_{sp}+||T_z\widehat{h}-\widehat{T}_z\widehat{h}||_{sp}\\ &\le \alpha||h-\widehat{h}||_{sp}+||T\widehat{h}-\widehat{T\widehat{h}}||_{sp}.\end{aligned}$$

Since $||u||_\infty\le||u||_{sp}$ for all $u\in\mathcal{C}_0(\mathbf{X})$ and $||u||_{sp}\le 2||u||_\infty$ for all $u\in M_b(\mathbf{X})$, the last inequality implies that

$$||h-\widehat{h}||_\infty \le \frac{2}{1-\alpha}||T\widehat{h}-\widehat{T\widehat{h}}||_\infty. \tag{24}$$

Now put $\widehat{h}_c:=\widehat{h}-c$, where c is an arbitrary constant, and notice that

$$\begin{aligned}|T\widehat{h}(x)-\widehat{T\widehat{h}}(x)| &= |T\widehat{h}_c(x)-\widehat{T\widehat{h}}_c(x)|\\ &= |\inf_{f\in\widehat{\mathbb{F}}_0} T_f\widehat{h}_c(x)-\inf_{f\in\widehat{\mathbb{F}}_0}\widehat{T\widehat{h}}_c(x)|\\ &\le \sup_{f\in\widehat{\mathbb{F}}_0}|T_f\widehat{h}_c(x)-\widehat{T\widehat{h}}_c(x)|\\ &\le \sup_{f\in\widehat{\mathbb{F}}_0}|Q_f\widehat{h}_c(x)-\widehat{Q}_f\widehat{h}_c(x)|\\ &\le ||\widehat{h}_c||_\infty\ \delta_Q(\widehat{\mathbb{F}}_0).\end{aligned}$$

Therefore,

$$||T\widehat{h}-\widehat{T}\widehat{h}||_{\infty}\leq||\widehat{h}_c||_{\infty}\ \delta_Q(\widehat{\mathbb{F}}_0). \tag{25}$$

Next consider the canonical policy $\widehat{f}$; thus,

$$\widehat{h}(x)=C_{\widehat{f}}(x)-\widehat{\rho}+\int_{\mathbf{X}}\widehat{h}(y)\widehat{Q}_{\widehat{f}}(\mathrm{d}y|x)\ \ \forall x\in\mathbf{X}.$$

This implies that

$$\widehat{h}(x)-\widehat{\mu}_{\widehat{f}}(\widehat{h})=\sum_{n=0}^{\infty}[E_x^{\widehat{f}}C_{\widehat{f}}(x_k)-\widehat{\rho}],$$

which in turn yields

$$||\widehat{h}-\widehat{\mu}_{\widehat{f}}(\widehat{h})||_{\infty}\leq\frac{1}{1-\alpha}||C_{\widehat{f}}||_{\infty}\leq\frac{K}{1-\alpha}.$$

The last inequality combined with (24) and (25) with $c:=\widehat{\mu}_{\widehat{f}}(\widehat{h})$, implies that

$$||h-\widehat{h}||_{\infty}\leq\frac{K}{(1-\alpha)^2}\delta_Q(\widehat{\mathbb{F}}_0),$$

which is the desired result.

The proof of (d) follows the same arguments of part (b) after noting that

$$||T\widetilde{h}-\widetilde{T}\widetilde{h}||_{\infty}\leq\delta_C(\widetilde{\mathbb{F}}_0)+\delta_Q(\widetilde{\mathbb{F}}_0).$$

□

Proof of Theorem 5.5. To prove this theorem recall, as discussed in the proof of Theorem 5.2, that all the conclusions in Theorem 3.2 remain valid for models $\widehat{M}$ and $\widetilde{M}$. Thus, let $f\in\mathbb{F}$ be a $L\widehat{h}_n$-greedy policy, that is,

$$TL\widehat{h}_n=T_fL\widehat{h}_n,$$

which also means that f is $\widehat{h}_n$-greedy in the model $\widehat{M}$. Then, from Theorem 3.2(b),

$$0\leq\widehat{J}(f)-\widehat{\rho}\leq||\widehat{\rho}_n||_{sp}. \tag{26}$$

On the other hand,

$$0\leq J(f)-\rho^*\leq|J(f)-\widehat{J}(f)|+|\widehat{J}(f)-\widehat{\rho}|+|\widehat{\rho}-\rho^*|.$$

Thus, Theorem 5.3(a), Lemma 7.1(b) and inequality (26) imply

$$0\leq J(f)-\rho^*\leq||\widehat{\rho}_n||_{sp}+\frac{2K}{1-\alpha}\delta_{\widehat{\mathbb{F}}_0}(Q),$$

which proves (a).

The proof of (b) is analogous. First note that

$$0 \leq J(g) - \rho^* \leq |J(g) - \widetilde{J}(g)|| + |\widetilde{J}(g) - \widetilde{\rho}| + |\widetilde{\rho} - \rho^*|. \tag{27}$$

Now, since $g \in \mathbb{F}$ is $\widetilde{h}_n$-greedy, then

$$T\widetilde{h}_n = T_f \widetilde{h}_n,$$

which means that g is $\widetilde{h}_n$-greedy in model $\widetilde{M}$. Thus, from Theorem 3.2(b),

$$0 \leq \widetilde{J}(g) - \widetilde{\rho} \leq ||\widetilde{\rho}_n||.$$

Thus, the desire result follows from (27), part (c) and Lemma 7.1(c). □

ACKNOWLEDGMENTS.

The authors thank to the referees their helpful comments and for providing the references [36, 46].

The first author (OVA) was partially supported by project SEP-CA Red en Control Estocástico, and second the author (JLV) was partially supported by Consejo Nacional de Ciencia y Tecnología (Scholarship 61702) and by Universidad de Sonora.

(Received May 4, 2017)

REFERENCES

[1] J. Abounadi, D. Bertsekas, and V. S. Borkar: Learning algorithms for Markov decision processes with average cost. SIAM J. Control Optim. *40* (2001), 681–698. DOI:10.1137/s0363012999361974 ,

[2] C. D. Aliprantis and K. C. Border: Infinite Dimensional Analysis. Third edition. Springer-Verlag, Berlin 2006.

[3] A. Almudevar: Approximate fixed point iteration with an application to infinite horizon Markov decision processes. SIAM J. Control Optim. *46* (2008), 541–561. DOI:10.1137/040614384

[4] A. Araposthatis, V. S. Borkar, E. Fernández-Guacherand, M. K. Gosh, and S. I. Marcus: Discrete-time controlled Markov processes with average cost criterion: a survey. SIAM J. Control Optim. *31* (1993) 282–344. DOI:10.1137/0331018

[5] L. Beutel, H. Gonska, and D. Kacsó: On variation-diminishing Shoenberg operators: new quantitative statements. In: Multivariate Approximation and Interpolations with Applications (M. Gasca, ed.), Monografías de la Academia de Ciencias de Zaragoza No. 20 2002, pp. 9–58.

[6] D. P. Bertsekas: Dynamic Programming: Deterministic and Stochastic Models. Prentice-Hall, Englewood Cliffs NJ 1987.

[7] D. P. Bertsekas and J. N. Tsitsiklis: Neuro-Dynamic Programming. Athena Scientific, Belmont 1996.

[8] D. P. Bertsekas: Approximate policy iteration: a survey and some new methods. J. Control Theory Appl. *9* (2011), 310–335. DOI:10.1007/s11768-011-1005-3

[9] H.,§. Chang and S. I. Marcus: Approximate receding horizon approach for Markov decision processes: average reward case. J. Math. Anal. Appl. *286* (2003), 636–651. DOI:10.1016/s0022-247x(03)00506-7

[10] H. S. Chang, J. Hu, M. C. Fu, and S. I. Marcus: Simulation-Based Algorithms for Markov Decision Processes. Second edition. Springer-Verlag, London 2013. DOI:10.1007/978-1-4471-5022-0

[11] W. L. Cooper, S. G. Henderson, and M. E. Lewis: Convergence of simulation-based policy iteration. Prob. Eng. Inform. Sci. *17* (2003), 213–234. DOI:10.1017/s0269964803172051

[12] R. A. DeVore: The Approximation of Continuous Functions by Positive Linear Operators. Lectures Notes in Mathematics *293*. Springer-Verlag, Berlin, Heidelberg 1972. DOI:10.1007/bfb0059493

[13] C. C. Y. Dorea and A. G. C. Pereira: A note on a variations of Doeblin's condition for uniform ergodicity of Markov chains. Acta Math. Hungar. *110*, Issue 4, (2006), 287–292. DOI:10.1007/s10474-006-0023-y

[14] F. Dufour adn T. Prieto-Rumeau: Approximation of Markov decision processes with general state space. J. Math. Anal. Appl. *388* (2012), 1254–1267. DOI:10.1016/j.jmaa.2011.11.015

[15] F. Dufour and T. Prieto-Rumeau: Stochastic approximations of constrained discounted Markov decision processes. J. Math. Anal. Appl. *413* (2014), 856–879. DOI:10.1016/j.jmaa.2013.12.016

[16] F. Dufour and T. Prieto-Rumeau: Approximation of average cost Markov decision processes using empirical distributions and concentration inequalities. Stochastics *87* (2015), 273–307. DOI:10.1080/17442508.2014.939979

[17] D. P. de Farias and B. van Roy: On the existence of fixed points for approximate value iteration and temporal difference learning. J. Optim. Theory Appl. *105* (2000), 589–608. DOI:10.1023/a:1004641123405

[18] D. P. de Farias and B. van Roy: Approximate linear programming for average-cots dynamic programming. In: Advances in Neural Information Processing Systems 15 (S. Becker, S. Thrun and K. Obermayer, eds.), MIT Press, Cambridge MA 2002, pp. 1587–1594.

[19] D. P. de Farias and B. Van Roy: A cost-shaping linear program for average-cost approximate dynamic programming with performance guarantees. Math. Oper. Res. *31* (2006), 597–620. DOI:10.1287/moor.1060.0208

[20] G. J. Gordon: Stable function approximation dynamic programming. In: Proc. Twelfth International Conference on Machine Learning (A. Prieditis and S. J. Russell, eds.), Tahoe City CA 1995, pp. 261–268. DOI:10.1016/b978-1-55860-377-6.50040-2

[21] A. Gosavi: A reinforcement learning algorithm based on policy iteration for average reward: empirical results with yield management and convergence analysis. Machine Learning *55* (2004), 5–29. DOI:10.1023/b:mach.0000019802.64038.6c

[22] O. Hernández-Lerma: Adaptive Markov Control Processes. Springer-Verlag, NY 1989. DOI:10.1007/978-1-4419-8714-3

[23] O. Hernández-Lerma and J. B. Lasserre: Discrete-Time Markov Control Processes. Basic Optimality Criteria. Springer-Verlag, NY 1996. DOI:10.1007/978-1-4612-0729-0

[24] O. Hernández-Lerma and J. B. Lasserre: Further Topics on Discrete-Time Markov Control Processes. Springer-Verlag, NY 1999. DOI:10.1007/978-1-4612-0561-6

[25] O. Hernández-Lerma and J. B. Lasserre: Markov Chains and Invariant Probabilities. Birkhauser Verlag, Basel 2003. DOI:10.1007/978-3-0348-8024-4

[26] O. Hernández-Lerma, R. Montes-de-Oca, and R. Cavazos-Cadena: Recurrence conditions for Markov decision processes with Borel spaces: a survey. Ann. Oper. Res. *29* (1991), 29–46. DOI:10.1007/bf02055573

[27] O. Hernández-Lerma, O. Vega-Amaya, and G. Carrasco: Sample-path optimality and variance-minimization of average cost Markov control processes. SIAM J. Control Optim. *38* (1999), 79–93. DOI:10.1137/s0363012998340673

[28] A. Jaskiewicz and A. S. Nowak: On the optimality equation for average cost Markov control processes with Feller transitions probabilities. J. Math. Anal. Appl. *316* (2006), 495–509. DOI:10.1016/j.jmaa.2005.04.065

[29] E. Klein and A. C. Thompson: Theory of Correspondences. Wiley, New York 1984.

[30] V. R. Konda and J. N. Tsitsiklis: Actor-critic algorithms. SIAM J. Control Optim. *42* (2003), 1143–1166. DOI:10.1137/s0363012901385691

[31] J. M. Lee and J. H. Lee: Approximate dynamic programming strategies and their applicability for process control: a review and future direction. Int. J. Control Automat. Systems *2* (2004), 263–278.

[32] S. P. Meyn and R. L. Tweedie: Markov Chain and Stochastic Stability. Springer-Verlag, London 1993. DOI:10.1007/978-1-4471-3267-7

[33] R. Montes-de-Oca and E. Lemus-Rodríguez: An unbounded Berge's minimum theorem with applications to discounted Markov decision processes. Kybernetika *48* (2012), 268–286.

[34] R. Munos: Performance bounds in L_p-norm for approximate value iteration. SIAM J. Control Optim. *47* (2007), 2303–2347. DOI:10.1137/s0363012904441520

[35] A. S. Nowak: A generalization of Ueno's inequality for n-step transition probabilities. Appl. Math. *25* (1998), 295–299. DOI:10.4064/am-25-3-295-299

[36] R. Ortner: Pseudometrics for state aggregation in average reward Markov decision processes. In: Algorithmic Learning Theory LNAI 4754 (M. Hutter, R. A. Serveido and E. Takimoto, eds.), Springer, Berlin, Heidelberg 2007, pp. 373–387. DOI:10.1007/978-3-540-75225-7_30

[37] M. L. Puterman: Markov Decision Processes: Discrete Stochastic Dynamic Programming. Wiley, NY 1994. DOI:10.1002/9780470316887

[38] W. P. Powell: Approximate Dynamic Programming. Solving the Curse of Dimensionality. John Wiley and Sons Inc., 2007. DOI:10.1002/9780470182963

[39] W. P. Powell: What you should know about approximate dynamic programming. Naval Res. Logist. *56* (2009), 239–249. DOI:10.1002/nav.20347

[40] W. P. Powell: Perspectives of approximate dynamic programming. Ann. Oper. Res. *241* (2012), 319–356. DOI:10.1007/s10479-012-1077-6

[41] W. P. Powell and J. Ma: A review of stochastic algorithms with continuous value function approximation and some new approximate policy iteration algorithms for multidimensional continuous applications. J. Control Theory Appl. *9* (2011), 336–352. DOI:10.1007/s11768-011-0313-y

[42] M. T. Robles-Alcaraz, O. Vega-Amaya, and J. Adolfo Minjárez-Sosa: Estimate and approximate policy iteration algorithm for discounted Markov decision models with bounded costs and Borel spaces. Risk Decision Anal. *6* (2017), 79–95. DOI:10.3233/rda-160116

[43] J. Rust: Numerical dynamic programming in economics. In: Handbook of Computational Economics, Vol. 13 (H. Amman, D. Kendrick and J. Rust, eds.), North-Holland, Amsterdam 1996, pp. 619–728. DOI:10.1016/s1574-0021(96)01016-7

[44] M. S. Santos: Analysis of a numerical dynamic programming algorithm applied to economic models. Econometrica *66* (1998), 409–426. DOI:10.2307/2998564

[45] M. S. Santos and J. Rust: Convergence properties of policy iteration. SIAM J. Control Optim. *42* (2004) 2094–2115. DOI:10.1137/s0363012902399824

[46] N. Saldi, S. Yuksel, and T. Linder: Asymptotic optimality of finite approximations to Markov decision processes with Borel spaces. Math. Oper. Res. *42* (2017), 945–978. DOI:10.1287/moor.2016.0832

[47] J. Stachurski: Continuous state dynamic programming via nonexpansive approximation. Comput. Economics *31* (2008), 141–160. DOI:10.1007/s10614-007-9111-5

[48] R. S. Sutton and A. G. Barto: Reinforcement Learning: An Introduction. MIT Press, Cambridge MA 1998. DOI:10.1108/k.1998.27.9.1093.3

[49] B. Van Roy: Performance loss bounds for approximate value iteration with state aggregation. Math. Oper. Res. *31* (2006), 234–244. DOI:10.1287/moor.1060.0188

[50] O. Vega-Amaya: The average cost optimality equation: a fixed-point approach. Bol. Soc. Mat. Mexicana *9* (2003), 185–195.

[51] O. Vega-Amaya: Zero-sum average semi-Markov games: fixed point solutions of the Shapley equation. SIAM J. Control Optim. *42* (2003), 1876–1894. DOI:10.1137/s0363012902408423

[52] O. Vega-Amaya: Solutions of the average cost optimality equation for Markov decision processes with weakly continuous kernel: The fixed-point approach revisited. J. Math. Anal. Appl. *464* (2018), 152–163. DOI:10.1016/j.jmaa.2018.03.077

[53] O. Vega-Amaya and J. López-Borbón: A Perturbation approach to a class of discounted approximate value iteration algorithms with Borel spaces. J. Dyn. Games *3* (2016), 261–278. DOI:10.3934/jdg.2016014

[54] O. Vega-Amaya amd R. Montes-de-Oca: Application of average dynamic programming to inventory systems. Math. Methods Oper. Res. *47* (1998) 451–471. DOI:10.1007/bf01198405

Óscar Vega-Amaya, Departamento de Matemáticas, Universidad de Sonora, Luis Encinas y Rosales s/n, C.P. 83180, Hermosillo, Sonora. México.
e-mail: ovega@mat.uson.mx

Joaquín López-Borbón, Departamento de Matemáticas, Universidad de Sonora, Luis Encinas y Rosales s/n, C.P. 83180, Hermosillo, Sonora. México.
e-mail: jlopez@gauss.mat.uson.mx