A PERTURBATION APPROACH TO A CLASS OF DISCOUNTED APPROXIMATE VALUE ITERATION ALGORITHMS WITH BOREL SPACES

ÓSCAR VEGA-AMAYA* AND JOAQUÍN LÓPEZ-BORBÓN

Departmento de Matemáticas
Universidad de Sonora
Rosales s/n, Col. Centro, 83000
Hermosillo, Sonora, México

(Communicated by Bruno Oliveira)

Abstract. The present paper gives computable performance bounds for the approximate value iteration (AVI) algorithm when approximation operators satisfy the following properties: (i) they are positive linear operators; (ii) constant functions are fixed points of such operators; (iii) they have certain continuity property. Such operators define transition probabilities on the state space of the controlled systems. This has two important consequences: (a) one can see the approximating function as the average value of the target function with respect to the induced transition probability; (b) the approximation step in the AVI algorithm can be thought of as a perturbation of the original Markov model. These two facts enable us to give finite-time bounds for the AVI algorithm performance depending on the operators accuracy to approximate the cost function and the transition law of the system. The results are illustrated with numerical approximations for a class of inventory systems.

1. Introduction. Markov decision processes provide a highly flexible framework to analyze many sequential optimization problems in a number of fields. See, for instance, the books by Bertsekas [3], Bertsekas and Tsitsiklis [5], Hernández-Lerma [11], Hernández-Lerma and Lasserre [12, 13], Puterman [19] for the theory, and the survey papers by Stidham and Weber [23], White [27, 28, 29], Yakowitz [30], Yeh [31] for real and potential applications. However, its range of usefulness is seriously limited by the so-called curse of dimensionality since it prevents the numerical computations in most applications, specially in those cases having large or infinite state spaces. It has been proposed a plethora of approximation schemes to break down or to alleviate the curse of dimensionality, which produce suboptimal but hopefully “good” solutions (see, for instance, Arruda et al. [2], Bertsekas [4], Jiang and Powell [14], Powell [16, 17, 18], Rust [20], Sutton [24]). The bulk of the references considers the approach called approximate value iteration (AVI) algorithm and deals mainly with finite models. In the present paper we also study AVI algorithms but considering models with Borel spaces.

2010 Mathematics Subject Classification. Primary: 93E20, 90C59; Secondary: 90C40.

Key words and phrases. Markov decision processes, discounted criterion, approximate value iteration algorithm, perturbed models.

* Corresponding author: Ó. Vega-Amaya.
Roughly speaking, the AVI algorithms combine suitable function approximation schemes with the standard value iteration algorithm. In many cases, the approximation schemes are represented by operators and the quality of the resulting AVI algorithms depend strongly on their properties; for instance, the convergence can not be guaranteed unless the approximation operator has a non-expansive property (de Farias and Van Roy [9], Gordon [10]). Thus, this kind of procedures raises the following three important issues: (I) convergence or stability of the AVI algorithms; (II) once the convergence is ensured, the second issue is to provide computable bounds for the error produced by the approximating functions; (III) the third one, perhaps the most important from the applications point of view, is to provide computable error bounds for the suboptimal policies generated by the algorithms.

Concerning issues (I)-(III) above, a quick glance at the references by Bertsekas [4], Bertsekas and Tsitsiklis [5], Powell [16], [17], [18], Rust [20] –and their extensive references–evidences, on one hand, the lack of general bounds for the performance of AVI algorithms–and for others approximation schemes as well–and for the other hand, that the most part of the papers focuses on finite models. There are of course several exceptions dealing with Borel spaces, among which we can mention the papers by Almudevar [1], Dufour and Prieto-Rumeau [8], Munos [15] and Stachurski [21]. The main differences of the present work with these latter papers are discussed below.

The present work studies a class of AVI algorithms for discounted Markov decision models with Borel spaces and bounded costs, and addresses issues (I)-(III) above for approximation operators with the following properties: (i) they are positive linear operators; (ii) constant functions are fixed points of such operators; (iii) they have a certain continuity property–see Definition 3.1, Section 3, for a precise statement of these properties. Many operators studied in approximation theory satisfy these properties as, for instance, piece-wise constant approximation operators, linear and multilinear interpolators, kernel-based interpolators (Gordon [10], Stachurski [21]), certain aggregation-projection operators (Van Roy [25]), Schoenberg’ splines, Hermite-Fejér and Bernstein operators (Beutel [6], DeVore [7]), among others.

The key point in this approach is that operators satisfying properties (i)-(iii) define transition probabilities on the state space of the controlled system. This has two important consequences: (a) one can see the approximating function as the average value of the target function with respect to the induced transition probability; (b) the approximation step in the AVI algorithm can be thought of as a perturbation of the original Markov model. These two facts allow us to give finite-time bounds for the AVI algorithm performance in terms of the accuracy of the approximations given by such operators for the primitive data model, namely, the one-step reward function and the system transition law. The accuracy of the approximations is measured by means of the supremum norm of bounded functions and the total variation norm of finite signed-measures.

A remarkable and perhaps somewhat surprising fact is that once the approximation step is seen as a perturbation of the original Markov model, the convergence in (I) and the bounds for problems (II)-(III) are directly established with quite elementary proofs. To the best authors’ knowledge, facts (a) and (b) have largely passed unnoticed except for the paper by Gordon [10], who refers to property (a) as an “intriguing property” that allows to see the “averagers as a Markov processes”, whereas property (b) is plainly ignored. Thus, Gordon [10] does not take advantage
of these facts to provide error bounds for the AVI algorithm performance. In fact, Gordon sets aside properties (i)-(iii) and focuses on non-expansive operators. As we shall see in Remark 3.2, properties (i)-(iii) imply the averagers have this latter property. In spite of these differences we follow Gordon’s practice and call averagers to operators with properties (i)-(iii) because of the property (a) above.

The papers by Almudevar [1], Munos [15] and Stachurski [21] also deal with approximations for models with Borel spaces, but they differ with the present work in several ways. For instance, Almudevar [1] takes an abstract approach and first studies general approximate fixed-point iteration schemes, and then he applies the results to several types of Markov decision problems. However, Almudevar [1] only obtains asymptotic bounds under the assumption the algorithm is stable—which means that the errors of the algorithm are uniformly bounded—and the bounds are not tied to the accuracy with which the approximation scheme represents the primitive data model. Additionally, for the continuous state space case Almudevar [1] requires the transition law has a (conditional) density function, which is not required in the present work.

On the other hand, Munos [15] provides two kind of $L_p$ bounds using certain quantities called “transition probabilities concentration coefficients” and “first and second discounted state future distribution concentration coefficients.” To obtain the first kind of bounds, he assumes that the algorithm is stable and bounds the algorithm performance asymptotically, as in the Almudevar’s paper [1]. The second ones are based on the so-called “Bellman residual”, but nothing is said about how it can be bounded with computable quantities. These “concentration coefficients” seem to be quite difficult to compute or to bound excepting some simple cases. Moreover, in the continuous state space case, Munos requires that the transition law has a density.

Stachurski [21] focus on non-expansive operators and first gives performance bounds in terms of quantities that are not directly computable. He removes this drawbacks under the additional assumption that the dynamic programming or Bellman operator preserves monotonicity of functions. This assumption seems suitable for some problems in economics but obviously limits the general usefulness of his results. Jiang and Powell [14] also study the AVI algorithm in problems with “monotone structure” and prove the algorithm convergence for finite models under several technical conditions.

An alternative to AVI algorithms is given by discretization procedures. These procedures require that the control model satisfies very nice structural properties such as Lipschitz continuity in the one-step cost function and the system evolution law, as well as in the multifunction defined by the admissible actions sets (Hernández-Lerma [11], Dufour and Prieto-Rumeau [8]). Under such kind of hypothesis, the latter two references give explicit bounds for the approximation error to the optimal value function; however, Hernández-Lerma [11] only analyzes the performance of the approximating policies in a asymptotical sense as the size of the discretization mesh goes to zero, whereas Dufour and Prieto-Rumeau [8] provide a performance bound depending on a quantity that cannot be controlled by the size of the discretization mesh. Here it is worth mentioning that in the present paper we obtain the performance bounds only assuming that the model satisfies the standard continuity-compactness hypothesis.

Summing up, the main contribution of the present work is a novel perturbation approach to analyze AVI algorithms defined by a class of approximation operators
we call averagers. This framework allows to give performance bounds depending on
the averagers accuracy to approximate the one-step cost function and the transition
law. The averagers includes many of the operators used in approximation function
theory and approximate dynamic programming itself. For instance, the examples of
nonexpansive operators given in Stachurski[21] and the projection-aggregation op-
erator in Van Roy [25] are in fact averagers. Note that the discretization procedures
can be recast as piece-wise constant approximations, which are also represented by
averagers. Hence, the results of the present paper are applicable to all these cases.

The remainder of the present work is organized as follows. Section 2 is largely
expository since it contains a brief description of the Markov control model and
some well-known results for the discounted optimal control problem with bounded
costs. Section 3 introduces the approximate value iteration algorithm, the kind
of approximation operators we are considering and the perturbed Markov models
associated with these operators. Next, Section 4 provides bounds for the approxi-
mation problem (II) and for the performance of the AVI algorithm (III). In section
5, the results are illustrated with an inventory system with finite capacity, linear
production cost, no set-up cost and no back-orders. The paper ends with some
concluding remarks in Section 6.

2. The discounted cost criterion. Throughout the work we use the following
notation. For a topological space $(S, \tau)$, $\mathcal{B}(S)$ denotes the Borel $\sigma$-algebra generated
by the topology $\tau$ and “measurability” will always mean Borel measurability.
Moreover, $M(S)$ is the class of measurable functions on $S$ whereas $M_b(S)$ is the sub-
space of bounded measurable functions endowed with the supremum norm given as
$||u||_\infty := \sup_{s \in S} |u(s)|$, $u \in M_b(S)$. The subspace of bounded continuous functions
is denoted by $C_b(S)$. For a subset $A \subseteq S$, $I_A$ stands for the indicator function of
$A$, that is, $I_A(s) = 1$ for $s \in S$ and $I_A(s) = 0$ for $s \notin A$. A Borel space $Y$ is a
measurable subset of a complete separable metric space endowed with the inherited
metric.

Let $M = (X, \mathcal{A}, \{A(x) : x \in X\}, R, Q)$ be the standard Markov control model.
This is thought as a model of a controlled stochastic process $\{x_n, a_n\}$, where the
state process $\{x_n\}$ takes values in the Borel space $X$ and the control process $\{a_n\}$
takes values in the Borel space $\mathcal{A}$. The controlled process evolves as follows: at
each time $n \in \mathbb{N}_0 := \{0, 1, \ldots\}$, the controller observes the system in some state
$x_n = x$ and chooses a control $a_n = a$ from the admissible control subset $A(x)$, which
is assumed to be a Borel subset of $\mathcal{A}$. It is also assumed that the admissible pairs
set $\mathcal{K} := \{(x, a) : x \in X, a \in A(x)\}$ belongs to $\mathcal{B}(X \times \mathcal{A})$. Then, the controller
receives a reward $R(x, a)$ where $R$ is a real-valued Borel measurable function defined
on $\mathcal{K} := \{(x, a) : x \in X, a \in A(x)\}$. Moreover, the controlled system moves to a
new state $x_{n+1} = x'$ according to the distribution measure $Q(\cdot|x, a)$, where $Q$ is a
stochastic kernel on $X$ given $\mathcal{K}$, that is, $Q(\cdot|x, a)$ is a probability measure on $X$ for
each pair $(x, a) \in \mathcal{K}$, and $Q(B, \cdot)$ is a Borel measurable function on $\mathcal{K}$ for each
Borel subset $B$ of $X$. Then, the controller chooses a new control $a_{n+1} = a' \in A(x')$
receiving a reward $R(x', a')$ and so on.

Let $H_n = \mathcal{K}^n \times X$ for $n \in \mathbb{N}$ and $H_0 := X$. Observe that a generic element
of $H_n$ has the form $h_n = (x_0, a_0, x_1, a_1, \ldots, x_{n-1}, a_{n-1}, x_n)$ where $(x_k, a_k) \in \mathcal{K}$ for
$k = 0, \ldots, n - 1$ and $x_n \in X$. A control policy is a sequence $\pi = \{\pi_n\}$ where $\pi_n(\cdot|\cdot)$
is a stochastic kernel on $\mathcal{A}$ given $H_n$ satisfying the constraint $\pi_n(A(x_n)|h_n) = 1$ for
all $h_n \in H_n$, $n \in \mathbb{N}_0$. Now let $F$ be the class of all measurable functions $f : X \rightarrow \mathcal{A}$.
such that \( f(x) \in A(x) \) for each \( x \in X \). A control policy \( \pi = \{\pi_n\} \) is said to be (deterministic) stationary if there exists \( f \in F \) such that the measure \( \pi_n(\cdot|x) \) is concentrated at \( f(x) \) for each \( x \in X \) and \( n \in N_0 \). Following a standard convention, the stationary policy \( \pi \) is identified with the selector \( f \). The class of all policies is denoted by \( \Pi \) and the class of all stationary policies is identified with the class \( \pi \).

Let \( \Omega := (X \times A)^{\infty} \) be the canonical sample space and \( F \) the product \( \sigma \)-algebra. For each policy \( \pi = \{\pi_n\} \in \Pi \) and “initial” state \( x_0 = x \in X \) there exists a probability measure \( P^\pi_x \) on the measurable space \((\Omega, F)\) that governs the evolution of the controlled process \((x_n, a_n)\). The discounted reward criterion is given as

\[
V_\pi(x) := E^\pi_x \sum_{k=0}^{\infty} \alpha^k R(x_k, a_k)
\]

where the discount factor \( \alpha \in (0, 1) \) is fixed and \( E^\pi_x \) denotes the expectation operator with respect to the probability measure \( P^\pi_x \). The optimal control problem is to find a control policy \( \pi^* \in \Pi \) such that

\[
V_{\pi^*}(x) = V^*(x) := \sup_{\pi \in \Pi} V_\pi(x) \quad \forall x \in X.
\]

The policy \( \pi^* \) is called discounted optimal policy, while \( V^* \) is called the discounted optimal value function.

Each one of Assumptions 1 and 2 below guarantees the discounted reward criterion is well defined and the existence of stationary optimal policies as well.

**Assumption 1.** (a) The function \( R(\cdot, \cdot) \) is bounded by a constant \( K > 0 \); (b) \( A(x) \) is a non-empty compact subset of \( A \) for each \( x \in X \) and the mapping \( x \to A(x) \) is continuous; (c) \( R(\cdot, \cdot) \) is a continuous function on \( K \); (d) \( Q(\cdot|\cdot, \cdot) \) is weakly continuous on \( K \), that is, the mapping

\[
(x, a) \to \int_X u(y)Q(dy|x, a)
\]

is continuous for each function \( u \in C_b(X) \).

**Assumption 2.** (a) The function \( R(\cdot, \cdot) \) is bounded by a constant \( K > 0 \); moreover, the following holds for each \( x \in X \) : (b) \( A(x) \) is a non-empty compact subset of \( A \); (c) \( R(x, \cdot) \) is a continuous function on \( A(x) \); (d) \( Q(\cdot|x, \cdot) \) is strongly continuous on \( A(x) \), that is, the mapping

\[
a \to \int_X u(y)Q(dy|x, a)
\]

is continuous for each function \( u \in M_b(X) \).

Throughout the paper \( C(X) \) will denote either \( C_b(X) \) or \( M_b(X) \) depending on whether Assumption 1 or 2 is being used, respectively. Then, under either one of Assumption 1 or 2, the dynamic programming operator

\[
Tu(x) := \sup_{a \in A(x)} [R(x, a) + \alpha \int_X u(y)Q(dy|x, a)], \quad x \in X,
\]

is a contraction operator from the Banach space \((C(X), \|\cdot\|_\infty)\) into itself with contraction factor \( \alpha \) (see, for instance, Hernández-Lerma [11], Lemma 2.5, p. 20.)
Moreover, by a selection theorem (Hernández-Lerma [11], D.3, p. 130), for each \( u \in C(X) \) there exists an selector \( f_u \in \mathcal{F} \) such that
\[
Tu(x) = R(x, f_u(x)) + \alpha \int_X u(y)Q(dy|x, f_u(x)) \quad \forall x \in X.
\] (1)

We shall refer to the policy \( f_u \) as \( u \)-greedy policy.

For each \( f \in \mathcal{F} \) and measurable function \( v \) on \( K \) set \( v_f(x) := v(x, f(x)), x \in X \).

Thus, in particular, we have
\[
R_f(x) = R(x, f(x)) \quad \text{and} \quad Q_f(\cdot|x) = Q(\cdot|x, f(x)), x \in X.
\]

Moreover, we shall write
\[
Q_fu(x) := \int_X u(y)Q_f(dy|x), x \in X,
\]
for each measurable function \( u \) on \( X \) for which the integral is well defined.

Now define the operators
\[
T_fu := R_f + \alpha Qfu, \quad f \in \mathcal{F},
\]
and observe that with this notation equation (1) becomes \( Tu = T_{fu}u \).

Assumption 1(a) (or Assumption 2(a)) implies that \( T_f \) is a contraction operator from the Banach space \( (M_b(X), ||\cdot||_\infty) \) into itself with contraction factor \( \alpha \). The Banach fixed-point theorem and standard dynamic programming arguments yield the following well-known result (Hernández-Lerma [11], Theorem 2.2, p. 19).

**Theorem 2.1.** Assume either Assumption 1 or Assumption 2 holds. Then:

(a) the optimal value function \( V^* \) is the only fixed-point in \( C(X) \) of operator \( T \);

(b) a stationary policy \( f \in \mathcal{F} \) is optimal if and only if \( V^* = T_fV^* \);

(c) there exists a stationary policy \( f^* \) such that \( V^* = T_{f^*}V^* \); hence, \( f^* \) is optimal;

(d) \( ||T^n u - V^*||_\infty \to 0 \) at geometric rate for any \( u \in C(X) \).

Theorem 2.1(a)-(c) gives a solution to the optimal control problem; however, the computation of an optimal stationary policy requires the optimal value function is known in advance, which, unfortunately, only occurs in few very simple cases.

Thus, based on Theorem 2.1(d), one can seek approximations of the value function \( V^* \) by means of the value iteration (VI) algorithm given as
\[
V_{k+1} = TV_k, \quad k \in \mathbb{N}_0,
\] (2)

where \( V_0 \in C(X) \) is an arbitrary function. The VI algorithm prescribes the computation of a \( V_k \)-greedy policy \( f \in \mathcal{F} \) once some stopping rule is satisfied, and then approximates the optimal value function \( V^* \) by means of \( V_f \). This section is closed with a result that gives a bound for the approximation error \( ||V^* - V_f||_\infty \).

**Theorem 2.2.** Suppose either Assumption 1 or Assumption 2 holds. Let \( V \in C(X) \) be arbitrary and \( f \) be a \( V \)-greedy policy. Then,
\[
||V^* - V_f||_\infty \leq \frac{2\alpha}{1 - \alpha} ||TV - V||_\infty.
\]

In particular, if \( V = V_k \), then \( ||V^* - V_f||_\infty \leq \frac{2\alpha}{1 - \alpha} ||V_{k+1} - V_k||_\infty \).

The proof of Theorem 2.2 follows elementary arguments and can be found, for instance, in Stachurski [22], Thm. 10.2.1, p. 239.
3. Approximation operators and perturbed models. The computational bur-
den associated with the VI algorithm (2) is prohibitive for systems with large finite
state spaces and is plainly infeasible for the continuous case. The standard way of
tackling this difficulty is to intersperse the application of the dynamic programming
operator with an approximation scheme. In many cases the approximation scheme
can be represented by means of an operator $L$ that maps a function $v$ into a func-
tion $Lv$ belonging to suitable subspace of functions; in this case, $Lv$ represents an
approximation to the function $v$. This yields two slightly different approximation
procedures depending on which operator, either $T$ or $L$, is applied first. Both ap-
proximation procedures are known as approximate value iteration (AVI) algorithm
and they are defined as follows:

$$
\hat{V}_{k+1} = \hat{T}\hat{V}_k \quad \text{and} \quad \tilde{V}_{k+1} = \tilde{T}\tilde{V}_k, \quad k \in \mathbb{N}_0,
$$

where $\hat{V}_0, \tilde{V}_0$ are arbitrary functions, $\hat{T} := LT$ and $\tilde{T} := TL$.

There are three important issues concerning these algorithms:

(I) the first one is the convergence of sequences in (3), say, to some functions $\hat{V}$
and $\tilde{V}$, respectively;

(II) the second one is to bound the approximation errors $\|V^* - \hat{V}\|_\infty$ and
$\|V^* - \tilde{V}\|_\infty$ provided the algorithms converge, or to bound the sequence of Bellman
residuals $\|\hat{V}_k - T\hat{V}_k\|_\infty$ and $\|\tilde{V}_k - T\tilde{V}_k\|_\infty$, $k \in \mathbb{N}$, otherwise;

(III) the third one is to bound the algorithm performance, that is, to bound
the quantity $\|V^* - V_f\|_\infty$ where the policy $f \in F$ is $\hat{V}_k$-greedy or $\tilde{V}_k$-greedy.
These issues are addressed below down for a class of approximation operators we call
averagers.

Definition 3.1. An operator $L : M(X) \rightarrow M(X)$ is said to be an averager iff it
satisfies the following conditions:

(a) $L(1X) = 1x$;

(b) $L$ is a linear operator;

(c) $L$ is a positive operator, that is, $Lu \geq 0$ for each $u \geq 0$ in $M(X)$.

(d) If $v_n \downarrow 0, v_n \in M_b(X)$, then $Lv_n \downarrow 0$.

The class of averagers may seem somewhat restrictive at a first sight, but many
approximation schemes define operators with these properties. Examples of aver-
agers are given by piece-wise constant approximation operators, linear and multilin-
ear interpolators, kernel-based interpolators (Gordon [10], Stachurski [21]), certain
aggregation-projection operators (Van Roy [25]), Schoenberg’ splines, Hermite-Fejér
and Bernstein operators (Beutel [6], DeVore [7]), among others.

The key point is that averagers allow to see the approximation step in the AVI
algorithms (3) as perturbations of the original Markov model. To introduce these
perturbed models we need several simple but important properties of averagers.

Remark 1. Suppose that $L$ is an averager. Then:

(a) $L$ is monotone, that is, $Lu \geq Lv$ whenever $u \geq v$. Moreover, $L$ is non-expansive
with respect to the supremum norm $\| \cdot \|_\infty$, that is, $\|Lu - Lv\|_\infty \leq \|u - v\|_\infty$ for all $u, v \in M_b(X)$. These properties follow directly from properties in Definition
3.1(a)-(c).

(b) Thus, if Assumption 2 holds, then the operator $\hat{T} = LT$ is a contraction from
$M_b(X)$ into itself with modulus $\alpha$. 

(c) Moreover, if the averager \( L \) maps \( C_b(X) \) into itself and Assumption 1 holds, the operators \( \tilde{T} = TL \) and \( \tilde{T} = LT \) are contraction operators from \( C_b(X) \) into itself with modulus \( \alpha \). This is the case, for instance, if the averager interpolates values using continuous functions.

The following lemma plays a key role in our approach; it shows the averagers can also be seen as transition probabilities. Its proof is omitted because it follows from standard arguments.

**Lemma 3.2.** Let \( L \) be an averager and define \( L(D|x) := LI_D(x) \) for \( x \in X \) and \( D \in \mathcal{B}(X) \). Then:

(a) \( L(\cdot|\cdot) \) is a transition probability on \( X \), that is, \( L(\cdot|x) \) is a probability measure on \( X \) for each \( x \in X \), and \( L(D|\cdot) \) is a measurable function for each \( D \in \mathcal{B}(X) \).

(b) Moreover, \( L v(x) = \int_X v(y) L(dy|x) \) for all \( x \in X, v \in M_b \).

(c) If in addition \( L \) maps \( C_b(X) \) into itself, then \( L(\cdot|\cdot) \) is weakly continuous.

The perturbed model \( \hat{M} \). For this model, the state and control spaces, the admissible decision sets and the one-step reward function are the same as the original Markov model \( M \). The perturbed transition law is defined by

\[
\hat{Q}(B|x,a) := \int_X L(B|y) Q(dy|x,a), \quad (x,a) \in \mathbb{K}, B \in \mathcal{B}(X).
\]

If the averager \( L \) maps \( C_b(X) \) into itself, then \( \hat{Q}(\cdot|\cdot,\cdot) \) is clearly a weakly continuous stochastic kernel on \( X \) given \( \mathbb{K} \) because it is the composition of weakly continuous stochastic kernels (see Lemma 3.2(a) and (c)).

For a policy \( \pi \in \Pi \) and initial state \( x \in X \), let \( \{(\hat{x}_k, \hat{a}_k)\} \) be the resulting controlled process and \( \hat{P}^\pi_x \) the corresponding probability measure defined on the measurable space \((\Omega, \mathcal{F})\). Denote by \( \hat{E}^\pi_x \) the expectation operator with respect to such measure. The discounted reward criterion and the discounted optimal value for the perturbed model \( \hat{M} \) are given as

\[
\hat{V}_\pi(x) := \hat{E}^\pi_x \sum_{k=0}^{\infty} \alpha^k R(\hat{x}_k, \hat{a}_k) \quad \text{and} \quad \hat{V}_\pi(x) := \sup_{\pi \in \Pi} \hat{V}_\pi(x), \quad x \in X, \pi \in \Pi,
\]

respectively. A policy \( \pi^* \) is said to be optimal for model \( \hat{M} \) iff \( \hat{V}_\pi(\cdot) = \hat{V}_{\pi^*}(\cdot) \).

The perturbed model \( \hat{M} \). For this model the state and control spaces are the same as the original Markov model \( M \). The perturbed one-step rewards and transition laws are only defined for the class of stationary policies as follows:

\[
\hat{R}_f(\cdot) := LR_f(\cdot) \quad \text{and} \quad \hat{Q}_f(B|\cdot) := LQ_f(B|\cdot)
\]

for each \( f \in \mathbb{F} \) and \( B \in \mathcal{B}(X) \). Again by Lemma 3.2(a), we have that \( \hat{Q}_f(\cdot|\cdot) \), \( f \in \mathbb{F} \), is a transition probability on \( X \) because

\[
\hat{Q}_f(B|x) = \int_X Q_f(B|y) L(dy|x) \quad \forall x \in X, B \in \mathcal{B}(X).
\]

Thus, for each stationary policy \( f \in \mathbb{F} \) and initial state \( x \in X \) there exists a Markov chain \( \{\hat{x}_n\} \) and probability measure \( \hat{P}^f_x \) defined on the measurable space \((\Omega, \mathcal{F})\) such that \( \hat{Q}_f(\cdot|\cdot) \) is the one-step transition probability of \( \{\hat{x}_n\} \). The expectation operator with respect to \( \hat{P}^f_x \) is denoted by \( \hat{E}^f_x \). The discounted reward criterion and
optimal value function are given as
\[
\tilde{V}_f(x) := E_x \sum_{k=0}^{\infty} \alpha^k \tilde{R}_f(\tilde{x}_k) \quad \text{and} \quad \tilde{V}_\ast(x) := \sup_{f \in \mathcal{F}} \tilde{V}_f(x), \quad x \in \mathbf{X}, f \in \mathcal{F}.
\]

A policy \( f^* \in \mathcal{F} \) is said to be optimal for the perturbed model \( \tilde{M} \) iff \( \tilde{V}_* = \tilde{V}_{f^*} \).

Lemmas 3.3 and 3.4, given below, contain the analog results for the perturbed models of those given in Theorems 2.1 and 2.2 for the original model \( M \). In fact, Lemma 3.3 is a direct corollary of Theorems 2.1 and 2.2.

**Lemma 3.3.** Suppose Assumption 1 holds and also the averager \( L \) maps \( C_b(\mathbf{X}) \) into itself. Then, the perturbed Markov model \( \tilde{M} \) satisfies Assumptions 2.1. Therefore, all the conclusions in Theorems 2.1 and 2.2 hold with \( C(\mathbf{X}) = C_b(\mathbf{X}) \).

**Lemma 3.4.** Suppose that either one of the following set of conditions holds:
(a) Assumption 1 holds and \( L \) is an averager mapping \( C_b(\mathbf{X}) \) into itself;
(b) Assumption 2 holds and \( L \) is an averager.

Then, there exists a \( \tilde{V}_* \)-greedy policy \( \tilde{f} \in \mathcal{F} \) and the optimal value function \( \tilde{V}_* \) is the unique solution in \( C(\mathbf{X}) \) of the optimality equation
\[
\tilde{V}_* = \sup_{f \in \mathcal{F}} [\tilde{R}_f + \alpha \tilde{Q}_f \tilde{V}_*] = \tilde{R}_{\tilde{f}} + \alpha \tilde{Q}_{\tilde{f}} \tilde{V}_*.
\]

Hence, \( \tilde{f} \) is optimal for model \( \tilde{M} \). Moreover, the remainder conclusions in Theorem 2.1(b)-(d) and Theorem 2.2 all hold true.

**Proof of Lemma 3.4.** From Remark 1(b)-(c) and (1), there exists a unique function \( \tilde{W} \) in the space \( C(\mathbf{X}) \) and a \( \tilde{W} \)-greedy policy \( \tilde{f} \in \mathcal{F} \) such that \( \tilde{W} = \tilde{T} \tilde{W} = \tilde{L} \tilde{T} \tilde{f} \tilde{W} \). Then, the linearity and monotonicity of \( L \) together with Lemma 3.2(b) imply, for one hand, that \( \tilde{W} = L(\tilde{R}_{\tilde{f}} + \alpha \tilde{Q}_{\tilde{f}} \tilde{W}) = \tilde{R}_{\tilde{f}} + \alpha \tilde{Q}_{\tilde{f}} \tilde{W} \) and also that \( \tilde{W} \geq \tilde{R}_{\tilde{f}} + \alpha \tilde{Q}_{\tilde{f}} \tilde{W} \) for each \( f \in \mathcal{F} \). Then,
\[
\tilde{W} = \sup_{f \in \mathcal{F}} [\tilde{R}_f + \alpha \tilde{Q}_f \tilde{W}] = \tilde{R}_{\tilde{f}} + \alpha \tilde{Q}_{\tilde{f}} \tilde{W}.
\]

Now, standard dynamic programming arguments yield \( \tilde{W} = \tilde{V}_* = \tilde{V}_{\tilde{f}} \), which proves the first two statement of the lemma. The proof of the remainder results also follow standard dynamic programming arguments since the operators
\[
\tilde{T}_f u = \tilde{R}_f + \alpha \tilde{Q}_f u, \quad f \in \mathcal{F}, u \in M_b(\mathbf{X}),
\]
are contraction from \( M_b(\mathbf{X}) \) into itself; so it is omitted. \( \square \)

The AVI algorithms defined in (3) only differ in the order the operators \( T \) and \( L \) are applied. Propositions 1 and 2 show how this difference is passed down to the corresponding optimal control problems.

**Proposition 1.** Suppose Assumption 1 holds and that the averager \( L \) maps \( C_b(\mathbf{X}) \) into itself. Then:
(a) \( \tilde{V}_* = L \tilde{V}_* \) and \( \tilde{V}_\ast = T \tilde{V}_\ast \);
(b) \( f \) is optimal for model \( \tilde{M} \) if and only if it is \( \tilde{V}_* \)-greedy, that is, \( T \tilde{V}_* = T_f \tilde{V}_* \);
(c) if \( f \) is optimal for the model \( \tilde{M} \), then \( f \) is optimal for model \( M \).

**Proof of Proposition 1.** (a) First recall, from Lemma 3.3, that \( \tilde{V}_* \) is the unique fixed-point of \( \tilde{T} = TL \) in \( C_b(\mathbf{X}) \). Then, \( w := L \tilde{V}_* = LT(L \tilde{V}_*) = \tilde{T}w \). Observe that \( w \) belongs to \( C_b(\mathbf{X}) \); thus, there exists a \( w \)-greedy policy \( g \in \mathcal{F} \), that is, \( T w = T_g w \).
This latter fact implies that \( w = \tilde{R}_g + \alpha \tilde{Q}_g w \geq \tilde{R}_f + \alpha \tilde{Q}_f w \) for all \( f \in \mathbb{F} \), which in turn implies that \( w = \tilde{V}_* \) and also that \( \tilde{V}_* = \tilde{T} \tilde{V}_* \), proving thus the first equality. Now, observe that \( TV_* = TL \tilde{V} = \tilde{V}_* \), which proves the second one.

(b) This is a direct consequence of part (a), Lemma 3.3, and Theorem 2.1(b).

(c) Suppose \( f \) is optimal for model \( \tilde{M} \). Then, \( TV_* = T_f \tilde{V} = R_f + \alpha \tilde{Q}_f \tilde{V}_* \), which in turn implies that inequality in part (a) holds. On the other hand, part (b) follows from (4) and part (a).

The Proposition 2 below gives bounds for the approximation errors \( ||V^* - \tilde{V}_*||_\infty \) and \( ||V^* - \tilde{V}_*||_\infty \). They are unsatisfactory because they depend on the (unknown) optimal value function \( V^* \): however, they evidence that such errors can be controlled by chosen “enough accuracy” approximating operators. This is the case, for instance, if the function \( V^* \) is continuous, \( X \) is a compact subset of the real numbers set \( \mathbb{R} \), and \( L \) is given by piece-wise constant approximations or linear interpolation schemes.

**Proposition 2.** Suppose assumptions in Proposition 1 hold. Then:

(a) \( ||V^* - \tilde{V}_*||_\infty \leq \frac{1}{1-\alpha} ||V^* - LV^*||_\infty \);

(b) \( ||V^* - \tilde{V}_*||_\infty \leq \frac{\alpha}{1-\alpha} ||V^* - LV^*||_\infty \).

**Proof of Proposition 2.** The contraction property of \( T \), Lemma 3.3 and Proposition 1(a) imply that

\[
||V^* - \tilde{V}_*||_\infty = ||TV^* - TL \tilde{V}_*||_\infty \\
\leq \alpha ||V^* - L \tilde{V}_*||_\infty \\
\leq \alpha ||V^* - \tilde{V}_*||_\infty.
\]

(4)

Now, from Remark 1, Proposition 1(a) and (4), it follows that

\[
||V^* - \tilde{V}_*||_\infty \leq ||V^* - LV^*||_\infty + ||LV^* - \tilde{V}_*|| \\
\leq ||V^* - LV^*||_\infty + ||LV^* - L \tilde{V}_*|| \\
\leq ||V^* - LV^*||_\infty + ||V^* - \tilde{V}_*|| \\
\leq ||V^* - LV^*||_\infty + \alpha ||V^* - \tilde{V}_*||_\infty,
\]

which in turn implies that inequality in part (a) holds. On the other hand, part (b) follows from (4) and part (a). \( \square \)

4. **Bounds for the approximate value iteration algorithm.** The accuracy of the approximations provided for the averagers are expressed in terms of the supremum norm \( || \cdot ||_\infty \) for bounded functions and the total variation norm \( || \cdot ||_{TV} \) for finite signed-measures. The latter one is defined as

\[
||\lambda||_{TV} := \sup \left\{ \left| \int_X v(y) \lambda(dy) \right| : v \in M_b(X), ||v||_\infty \leq 1 \right\}
\]

where \( \lambda \) is a finite signed-measure on \( X \). From the definition it follows that

\[
\int_X v(y) \lambda(dy) \leq ||\lambda||_{TV} ||v||_\infty \quad \forall v \in M_b(X).
\]

Moreover, one can prove that if \( P_1 \) and \( P_2 \) are probabilities measures, then

\[
||P_1 - P_2||_{TV} = 2 \sup_{B \in B(X)} |P_1(B) - P_2(B)|.
\]

(5)
Let \( \widehat{V}_0 \) be a subclass of stationary policies that contains the stationary optimal policies for the original model \( M \) and the perturbed model \( \widehat{M} \), and the policies \( V_n \)-greedy for each \( n \in \mathbb{N} \) as well. Similarly, \( \overline{F}_0 \) is a subclass of \( F \) that contains the optimal policies for \( M \) and \( \widehat{M} \), and the \( V_n \)-greedy policies. Next, define the quantities
\[
\delta_Q(\widehat{F}_0) := \sup\{|Q_f(\cdot|x) - \widehat{Q}_f(\cdot|x)|_{TV}: x \in X, f \in \overline{F}_0\},
\]
\[
\delta_R(\overline{F}_0) := \sup\{|R_f - \widetilde{R}_f|_{\infty}: f \in \overline{F}_0\},
\]
\[
\delta_Q(\overline{F}_0) := \sup\{|Q_f(\cdot|x) - \widetilde{Q}_f(\cdot|x)|_{TV}: x \in X, f \in \overline{F}_0\},
\]
and observe they measure the averager accuracy for approximating the one-step reward function and the transition law.

Now we are ready to state the main results of the present work.

**Theorem 4.1.** Suppose Assumption 1 and that the averager \( L \) maps \( C_b(X) \) into itself. Then:
(a) for each \( f \in F \) it holds that
\[
\|V_f - \widehat{V}_f\|_{\infty} \leq \frac{\alpha K}{(1 - \alpha)^2} \sup_{x \in X} |Q_f(\cdot|x) - \widetilde{Q}_f(\cdot|x)|_{TV};
\]
(b) \( \|V^* - \widehat{V}_n\|_{\infty} \leq \alpha K(1 - \alpha)^{-2}\delta_Q(\overline{F}_0) \);
(c) if \( f \in F \) is \( V_n \)-greedy, then
\[
\|V^* - V_f\|_{\infty} \leq \frac{2\alpha}{1 - \alpha} \|\widehat{V}_{n+1} - \widehat{V}_n\|_{\infty} + \frac{2\alpha K}{(1 - \alpha)^2}\delta_Q(\overline{F}_0).
\]

The next result provides the analog bounds for the perturbed model \( \widehat{M} \).

**Theorem 4.2.** Suppose that either one of the following set of conditions holds:
(i) Assumption 1 holds and \( L \) is an averager mapping \( C_b(X) \) into itself;
(ii) Assumption 2.2 holds and \( L \) is an averager.
Then:
(a) for each \( f \in F \) it holds that
\[
\|V_f - \overline{V}_f\|_{\infty} \leq \frac{1}{1 - \alpha} \|R_f - \tilde{R}_f\|_{\infty} + \frac{\alpha K}{(1 - \alpha)^2} \sup_{x \in X} |Q_f(\cdot|x) - \widetilde{Q}_f(\cdot|x)|_{TV};
\]
(b) moreover,
\[
\|V^* - \overline{V}_n\|_{\infty} \leq \frac{1}{1 - \alpha} \delta_R(\overline{F}_0) + \frac{\alpha K}{(1 - \alpha)^2}\delta_Q(\overline{F}_0);
\]
(c) if \( g \in F \) is a \( \overline{V}_n \)-greedy policy, then
\[
\|V^* - V_g\|_{\infty} \leq \frac{1}{1 - \alpha} \{2\alpha \|\overline{V}_n - \overline{V}_{n-1}\|_{\infty} + \delta_R(\overline{F}_0)\} + \frac{\alpha K}{(1 - \alpha)^2}\delta_Q(\overline{F}_0)
\]

**Remark 2.** The constants \( \delta_Q(\overline{F}_0) \) and \( \delta_Q(\overline{F}_0) \) are less or equal than 2 (see (5)); however, in general, it may be quite hard to get sharper bounds, at least some additional conditions are imposed on the transition law. For instance, Almudevar [1] and Munos [15] study the AVI algorithm for systems with continuous state.
spaces assuming the transition law is absolutely continuous with respect to some \( \sigma \)-finite measure \( m \); this latter means that there exists a conditional density function \( q(y|x,a) \) such that

\[
Q(B|x,a) = \int_B q(z|x,a) dm(z) \quad \forall B \in \mathcal{B}(X), (x,a) \in K.
\]

(6)

If this is the case, the constant \( \delta_Q(\mathcal{F}_0) \) has a more concrete and less prohibitive form as it is shown in the next proposition. Using analog arguments one can verify a similar result holds if the state space \( X \) is a finite or denumerable set.

**Proposition 3.** Suppose the transition law \( Q(\cdot|x) \) satisfies (6). Then, the perturbed transition probability \( Q_f(\cdot|x) \) is absolutely continuous with respect to the measure \( m \) with conditional density function

\[
\tilde{q}_f(\cdot|x) := Lq_f(\cdot|x) = \int_X q_f(\cdot|y)L(dy|x), \quad x \in X, \ f \in \mathcal{F},
\]

and

\[
\delta_Q(\mathcal{F}_0) = \sup_{x \in X, f \in \mathcal{F}} \int_X |q_f(z|x) - \tilde{q}_f(z|x)| dm(z).
\]

If in addition \( q(z|x,a) \leq K' \) for all \( z \in X, (x,a) \in K \) and \( m \) is a finite measure, then

\[
\delta_Q(\mathcal{F}_0) \leq \sup_{x \in X, f \in \mathcal{F}} ||q_f(\cdot|x) - \tilde{q}_f(\cdot|x)||_\infty m(X).
\]

Now we proceed to prove Theorems 4.1 and 4.2, and Proposition 3.

**Proof of Theorem 4.1.** (a) First note that \( V_f = R_f + \alpha Q_f V_f \) and \( \hat{V}_f = R_f + \alpha \hat{Q}_f \hat{V}_f \) for each stationary policy \( f \in \mathcal{F} \). Then,

\[
||V_f - \hat{V}_f||_\infty = \alpha ||Q_f V_f - \hat{Q}_f \hat{V}_f||_\infty
\]

\[
\leq \alpha ||Q_f V_f - Q_f \hat{V}_f||_\infty + \alpha ||Q_f \hat{V}_f - \hat{Q}_f \hat{V}_f||_\infty
\]

\[
\leq \alpha ||V_f - \hat{V}_f||_\infty + \alpha \sup_{x \in X} ||Q_f(\cdot|x) - \hat{Q}_f(\cdot|x)||_{TV} ||\hat{V}_f||_\infty.
\]

Since \( ||\hat{V}_f||_\infty \leq K(1-\alpha)^{-1} \), the latter inequality yields

\[
||V_f - \hat{V}_f||_\infty \leq \frac{\alpha K}{(1-\alpha)^2} \sup_{x \in X} ||Q_f(\cdot|x) - \hat{Q}_f(\cdot|x)||_{TV},
\]

which proves the desired result.

(b) It follows from (a) that \( -K' + \hat{V}_f \leq V_f \leq \hat{V}_f + K' \) for each stationary policy \( f \in \hat{\mathcal{F}}_0 \), where

\[
K' = \frac{\alpha K}{(1-\alpha)^2} \delta_Q(\mathcal{F}_0).
\]

Since \( \hat{\mathcal{F}}_0 \) contains the optimal stationary policies for models \( \hat{M} \) and \( \hat{M} \), we get the inequality \( -K' + \hat{V}_* \leq V_* \leq \hat{V}_* + K' \) after taking supremum over this class policies.

(c) This part follows directly from the inequality

\[
||V_* - V_f||_\infty \leq ||V_* - \hat{V}_*||_\infty + ||\hat{V}_* - \hat{V}_f||_\infty + ||\hat{V}_f - V_f||_\infty
\]

combined with parts (b)-(c), Lemma 3.3 and Theorem 2.2. \( \square \)
Proof of Theorem 4.2. To prove part (a) note that $\tilde{V}_f = \tilde{R}_f + \alpha \tilde{Q}_f \tilde{V}$ for each policy $f \in \mathbb{F}$; thus we see that

$$||V_f - \tilde{V}_f||_\infty \leq ||R_f - \tilde{R}_f||_\infty + \alpha ||Q_f V_f - \tilde{Q}_f \tilde{V}||_\infty$$

$$\leq ||R_f - \tilde{R}_f||_\infty + \alpha ||V_f - \tilde{V}_f||_\infty$$

$$+ \alpha \sup_{x \in \mathbb{X}} ||Q_f(\cdot|x) - \tilde{Q}_f(\cdot|x)||_{TV} ||\tilde{V}_f||_\infty.$$ 

Then,

$$||V_f - \tilde{V}_f||_\infty \leq \frac{1}{1-\alpha} ||R_f - \tilde{R}_f||_\infty + \frac{\alpha K}{(1-\alpha)^2} \sup_{x \in \mathbb{X}} ||Q_f(\cdot|x) - \tilde{Q}_f(\cdot|x)||_{TV},$$

because $||\tilde{V}_f||_\infty \leq K(1-\alpha)^{-1}$ for all $f \in \mathbb{F}$.

The proofs of part (b) and (c) follow similar arguments to those given in the proof of Theorem 4.1. Thus, they are omitted. \(\Box\)

Proof of Proposition 3. To prove (7), let $f \in \mathbb{F}$ and note that

$$\tilde{Q}_f(B|x) = \int_\mathbb{X} \left[ \int_B q_f(z|y) dm(z) \right] L(dy|x) = \int_B \tilde{q}_f(z|x) dm(z).$$

Moreover, from the total variation norm property in Hernandez-Lerma [11], Appendix B.3, p. 125, we see that

$$||Q_f(\cdot|x) - \tilde{Q}_f(\cdot|x)||_{TV} = \int_\mathbb{X} |q_f(y|x) - \tilde{q}_f(y|x)| dm(y).$$

This equality implies that

$$\delta_Q(\mathbb{F}_0) = \sup_{x \in \mathbb{X}, f \in \mathbb{F}_0} \int_\mathbb{X} |q_f(y|x) - \tilde{q}_f(y|x)| dm(y),$$

which proves the desire result.

The last statement of Proposition 3 follows directly from the latter equality. \(\Box\)

5. An inventory system with finite capacity. This section provides numerical results to illustrate the approach developed in the previous sections. We chose a simple inventory control problem for which the optimal stationary policy is already known in order to allow the comparison among the analytical and the numerical solutions. However, a similar analysis can be carried out for other models as inventory systems with positive set-up cost, fishery management, optimal replacement problems, optimal growth models, etc.

Then, consider a single item inventory system with finite capacity $\theta > 0$, no set-up cost and no backorders. Let $x_n$ be the item stock level and $a_n$ the quantity ordered at the beginning of the $n$th-decision epoch, and $w_n$ the quantity demanded during the same epoch. Assuming that the quantity ordered is immediately supplied, the inventory system evolves according to the difference equation

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

where $x_0 = x$ is the initial stock and $v^+ := \max(0, v)$. Thus, $\mathbb{X} = A = [0, \theta]$ and $A(x) = [0, \theta - x]$ for each $x \in \mathbb{X}$. The mapping $x \rightarrow A(x) = [0, \theta - x]$ is clearly continuous. Instead of a reward function $R$ we consider a one-step cost function $C$, so the optimal control problem is to find a policy with minimal cost.
We assume the demand process \( \{w_n\} \) is a sequence of independent identically distributed nonnegative random variables with continuous distribution function \( F \).

Then, the inventory dynamic can also be expressed by means of the stochastic kernel

\[
Q(B|x,a) = E_{w_0} \mathbb{1}_B((x+a-w_0)^+) \quad (x,a) \in \mathbb{K}, B \in \mathcal{B}(X).
\]

Moreover, observe that

\[
\int_X v(y)Q(dy|x,a) = E_{w_0}(v((x+a-w_0)^+)) \quad \forall (x,a) \in \mathbb{K}, v \in M_0(X).
\]

Here \( E_{w_0} \) stands for the expectation with respect to distribution \( F \) of the random variable \( w_0 \).

The latter equality implies that \( Q(\cdot,\cdot) \) is weakly continuous on \( \mathbb{K} \). Hence, this inventory model satisfies Assumption 1.

Now, consider the approximating operator \( L \) defined by the linear interpolation scheme with nodes \( s_0 = 0 < s_1 < \ldots < s_N = \theta \). Thus, for each bounded measurable function \( v \) on \( X \), the operator \( L \) is defined as

\[
Lv(x) = \frac{s_{i+1} - x}{s_{i+1} - s_i} v(s_i) + \frac{x - s_i}{s_{i+1} - s_i} v(s_{i+1}), \quad x \in [s_i, s_{i+1}], i = 0, \ldots, N.
\]

The operator \( L \) is clearly an averager that maps \( C_b(X) \) into itself.

**Bounds for the perturbed model \( \tilde{M} \).** For this model we consider an arbitrary continuous one-step cost function \( C \) and take \( \tilde{F}_0 = F \). The perturbed transition law \( \tilde{Q}(\cdot,\cdot) \) is given as

\[
\tilde{Q}(B|x,a) = \int_X L\mathbb{1}_B(y)Q(dy|x,a), \quad (x,a) \in \mathbb{K}, B \in \mathcal{B}(X).
\]

To compute \( \delta_Q(\tilde{F}_0) \) take \( B_1 := X \setminus \{s_0, s_1, \ldots, s_N\} \); next observe that \( \tilde{Q}(B_1|x,a) = 0 \) and \( Q(B_1|x,a) = F(x+a) \) for all \( (x,a) \in \mathbb{K} \). Then, \( ||Q(\cdot|x,a) - \tilde{Q}(\cdot|x,a)||_{TV} = 2F(x+a) \) for all \( (x,a) \in \mathbb{K} \). Hence, \( \delta_Q(\tilde{F}_0) = 2F(\theta) \). Therefore, from Theorem 4.1(b), if \( f \) is a \( \hat{\mathbb{V}}_n \)-greedy policy then

\[
||V^\star - V_f||_{\infty} \leq \frac{2\alpha}{1 - \alpha} ||\hat{\mathbb{V}}_n - \hat{\mathbb{V}}_{n-1}||_{\infty} + \frac{4\alpha KF(\theta)}{(1 - \alpha)^2}.
\]

Note that the latter bound is given in terms of known or at least computable quantities. Moreover, it is quite general because it holds under very weak assumption, namely, the continuity of the distribution function \( F \) and the one-step cost function \( C \). However, it does not depend on the grid \( \{s_0, s_1, \ldots, s_N\} \) and thus it is insensitive to refinements, which obviously is a disadvantage. This fault is not shared by the other AVI algorithm given in (3), as we show below.

**Bounds for the perturbed model \( \tilde{M} \).** Consider the one-step cost function

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

where \( p, h \) and \( c \) are positive constants representing the unit penalty cost for unmet demand, the unit holding cost for inventory at hand, and the unit production cost, respectively. Clearly, \( C \) is continuous on \( \mathbb{K} \). Moreover, assume the random variable \( w_0 \) has positive finite expectation \( \mathbb{E} \) and also that the distribution function \( F \) has a bounded continuous density \( \rho \). The latter assumption implies that \( Q(\cdot|x,\cdot) \) is strongly continuous for each \( x \in \mathbb{K} \). Hence, Assumption 2 holds.

Numerical experiments evidence that the optimal policies for the models \( M, \tilde{M} \) and \( M \) are base-stock policies. Recall that a stationary policy \( f \) is a base-stock policy if \( f(x) = S - x \) for \( x \in [0,K] \), and \( f(x) = 0 \) otherwise, where the constant
$S \geq 0$ is the so-called re-order point. In fact, it is shown in Vega-Amaya and Montes-de-Oca [26] that a base-stock policy is optimal for the inventory system (8) with a more general one-step cost function than (11), and also that the optimal re-order point $S^*$ for the cost (11) satisfies the equation $F(S) = (p - h - c)/(p - \alpha c)$ if $p > h + c$, and $S^* = 0$ otherwise. Thus, take $\bar{F}_0$ as the class of all base-stock policies. In order to estimate $\delta_C(\bar{F}_0)$ and $\delta_Q(\bar{F}_0)$, introduce the following auxiliary functions:

\[
pv(x) := \int_0^x v(z)\rho(x-z)dz, \quad P\nu(x) := v(0)(1-F(x)) + \pv(x), \quad I(x) := \int_0^x F(y)dy,
\]

for $x \in \mathcal{X}$ and $v \in M_b(\mathcal{X})$. Moreover, denote by $\bar{F}$, $\bar{P}v$, $\bar{P}v$ and $\bar{I}$ the linear interpolation functions of $F$, $P\nu$, $P\nu$ and $I$, respectively, with nodes at $s_0 = 0, s_1, \ldots, s_N = \theta$.

Let $f_0$ be a base-stock policy and suppose that $s_i \leq S < s_{i+1}$. Then, after some elementary computation, we obtain that

\[
Q_{f_0}v(x) = \begin{cases} P\nu(S), & 0 \leq x \leq S, \\ P\nu(x), & S < x \leq \theta, \end{cases}
\]

and

\[
\tilde{Q}_{f_0}\nu(x) = \begin{cases} P\nu(S), & 0 \leq x \leq S, \\ \tilde{P}\nu(x), & S < x \leq \theta. \end{cases}
\]

Similarly, we have that

\[
C_{f_0}(x) = \begin{cases} -cx + (c + h - p)S + p\overline{m} + pI(S), & 0 \leq x \leq S, \\ (h - p)x + p\overline{m} + pI(x), & S < x \leq \theta, \end{cases}
\]

\[
\tilde{C}_{f_0}(x) = \begin{cases} C_{f_0}(x), & 0 \leq x \leq s_i, \\ C_{f_0}(s_i)\tilde{b}_i(x) + C_{f_0}(s_{i+1})\tilde{b}_i(x), & s_i < x \leq s_{i+1}, \\ (h - p)x + p\overline{m} + p\tilde{I}(x), & s_{i+1} < x \leq \theta. \end{cases}
\]

One can verify, after some elementary but cumbersome computations, that

\[
\delta_C(\bar{F}_0) = \sup_{f \in \bar{F}_0} ||C_f - \tilde{C}_f||_\infty \leq (4p - 2h + c)\Delta s,
\]

where $\Delta s := \max_i(s_{i+1} - s_i)$.

Now, in order to get an estimate for $\delta_Q(\bar{F}_0)$ we impose a last condition on the density $\rho$. We assume that the density function $\rho$ is Lipschitz continuous with module $\lambda > 0$ on $[0, \theta]$, that is, $|\rho(x) - \rho(y)| \leq \lambda|x - y|$ for all $x, y \in [0, \theta]$. This implies that $||\pv - \bar{P}v||_\infty \leq 2\theta l + K'\Delta s$ and $||F - \bar{F}||_\infty \leq 2K'\Delta s$ for all $v \in M_b(\mathcal{X})$ with $||v||_\infty \leq 1$, where $K'$ is a bound for the density function $\rho$. Thus, we have

\[
\delta_Q(\bar{F}_0) = \sup_{f \in \bar{F}_0} ||Q_f - \tilde{Q}_f||_{TV} \leq (2\theta l + 4K')\Delta s.
\]

The bounds (12) and (13) may be quite conservative, but they show that the constants $\delta_C(\bar{F}_0)$ and $\delta_Q(\bar{F}_0)$ can be made arbitrarily small by taking fine enough grids. Moreover, they can be sharpened in specific cases as it is done in the numerical results given below.

**Numerical results.** Take $c = 1.5, h = 0.5, p = 3, \theta = 20$ and $\alpha = 0.6$ and consider a grid with $N+1$ evenly spaced nodes. Moreover, assume that the product demand has an exponential density function $\rho$ with parameter $\lambda = 0.1$. Note that $\rho$ is bounded by $K' = \lambda = 0.1$ and also that it is Lipschitz with module $l = \lambda^2 = 0.01$. For this case, the bound given in (12) is improved to $\delta_C(\bar{F}_0) \leq (p + h + c)\Delta s = 5\Delta s$.

The AVI algorithm is stopped once $E_n := ||\bar{V}_n - \bar{V}_{n-1}||_\infty$ falls below of a given tolerance error $\varepsilon > 0$. Let $n_\varepsilon$ be the first iteration when this happens and $f_{n_\varepsilon}$ be the corresponding $\bar{V}_{n_\varepsilon}$-greedy policy. As was already mentioned above, the
Table 1. Approximate optimal policies and performance bounds for the AVI algorithm with a linear interpolation scheme with \(N + 1\) nodes for an inventory systems with exponentially distributed demand with parameter \(\lambda = 0.1\).

| \(\varepsilon = 0.01\) | \(N = 1000\) | \(N = 2000\) | \(N = 5000\) | \(N = 10000\) |
|------------------------|--------------|--------------|--------------|-------------|
| \(n_s\)               | 17           | 17           | 17           | 17          |
| \(E_{n_s}\)           | 0.007454679  | 0.007456157  | 0.007487855  | 0.007487864 |
| \(\delta_C(\bar{F}_0)\) | 0.1          | 0.05         | 0.02         | 0.01        |
| \(\delta_Q(\bar{F}_0)\) | 0.016        | 0.008        | 0.0032       | 0.0016      |
| \(S_E\)               | 0.02236404   | 0.02236847   | 0.02246356   | 0.02246359  |
| \(A_E\)               | 33.54504     | 16.77252     | 6.709009     | 3.354504    |
| \(T_E\)               | 33.56741     | 16.79489     | 6.731472     | 3.37698     |
| \(S_m^*\)             | 6.5          | 6.47         | 6.468        | 6.467       |

Greedy policies for the AVI algorithm are base-stock policies; thus let \(\tilde{S}_n\) be the re-order point for the \(\tilde{V}_n\)-greedy policy. The numerical results are shown in Table 1 with the following notation: \(A_E := 2(1 - \alpha)^{-1}\delta_Q(\bar{F}_0) + 2\alpha K(1 - \alpha)^{-2}\delta_Q(\bar{F}_0)\), \(S_E := 2\alpha(1 - \alpha)^{-1}E_n\) and \(T_E := S_E + A_E\).

The Figure 1 displays the graphs of functions \(\tilde{V}_n\) for \(n = 1, 5, 10, 17\). Note that functions \(\tilde{V}_{10}\) and \(\tilde{V}_{17}\) are virtually indistinguishable. The numerical results in Table 1 and Figure 1 show that the approximate value iteration algorithm practically identifies the optimal re-orden point \(S^* = 6.466453\) and also that the performance bounds are pretty good; in fact, the bounds can be done arbitrarily small by chosen fine enough grids.

Figure 1. AVI functions \(\tilde{V}_n\) with \(n = 1, 5, 10, 17\) and \(N = 2000\).

6. Concluding remarks. In this paper we propose a perturbation approach to analyze the AVI algorithm when are used a class of approximation operators we call averagers. The key point is the dual nature of the averagers, which allows to see
the AVI algorithm as the standard value iteration algorithm in certain perturbed Markov decision model. This fact offers a new framework for analyzing the major problems raised by the AVI algorithm (I)-(III) described in the Introduction. We think this perturbation approach can also be combined with the policy iteration algorithm and the linear programming approach as well, and used to approximate the average cost optimal control problem. Moreover, albeit the focus of this paper is the case of “continuous” spaces, it should be noted that the perturbation approach also works in the discrete space case. Thus, it could also be interesting to compare it with the existing numerical procedures for such models using some test problems.

Acknowledgments. We thank the reviewer for the useful comments and also for bringing our attention to references [2] and [14].

REFERENCES

[1] A. Almudevar, Approximate fixed point iteration with an application to infinite horizon Markov decision processes, SIAM Journal on Control and Optimization, 47 (2008), 2303–2347.
[2] E. F. Arruda, M. D. Fragoso and J. B. R. do Val, Approximate dynamic programming via direct search in the space of value function approximations, European Journal of Operational Research, 211 (2011), 343–351.
[3] D. P. Bertsekas, Dynamic Programming: Deterministic and Stochastic Models, Prentice-Hall, Englewood Cliffs, NJ, 1987.
[4] D. P. Bertsekas, Approximate policy iteration: A survey and some new methods, Journal of Control Theory and Applications, 9 (2011), 310–335.
[5] D. P. Bertsekas and J. N. Tsitsiklis, Neuro-Dynamic Programming, Athena Scientific, Belmont, MA, 1996.
[6] L. Beutel, H. Gonska and D. Kacsó, On variation-diminishing Schoenberg operators: new quantitative statements, in Multivariate approximation and interpolations with applications (ed. M. Gasca), Monografías de la Academia de Ciencias de Zaragoza, 20 (2002), 9–58.
[7] R. A. DeVore, The Approximation of Continuous Functions by Positive Linear Operators, Lectures Notes in Mathematics 293, Springer-Verlag, Berlin Heidelberg, 1972.
[8] F. Dufour and T. Prieto-Rumeau, Approximation of infinite horizon discounted cost Markov decision processes, in Optimization, Control, and Applications of Stochastic Systems, In Honor of Onésimo Hernández-Lerma (eds. D. Hernández-Hernández and J. A. Minjárez-Sosa), Birkhäuser, (2012), 59–76.
[9] D. P. de Farias and B. Van Roy, On the existence of fixed points for approximate value iteration and temporal difference learning, Journal of Optimization Theory and Applications, 105 (2000), 589–608.
[10] G. J. Gordon, Stable function approximation in dynamic programming, Proceedings of the Twelfth International Conference on Machine Learning, Tahoe City, CA, 1995, 261–268.
[11] O. Hernández-Lerma, Adaptive Markov Control Processes, Springer-Verlag, NY, 1989.
[12] O. Hernández-Lerma and J. B. Lasserre, Discrete-Time Markov Control Processes. Basic Optimality Criteria, Springer-Verlag, NY, 1996.
[13] O. Hernández-Lerma and J. B. Lasserre, Further Topics on Discrete-Time Markov Control Processes, Springer-Verlag, NY, 1999.
[14] D. R. Jiang and W. B. Powell, An approximate dynamic programming algorithm for monotone value functions, Operations Research, 63 (2015), 1489–1511.
[15] R. Munos, Performance bounds in $L_p$ -norm for approximate value iteration, SIAM Journal on Control and Optimization, 46 (2007), 541–561.
[16] W. P. Powell, Approximate Dynamic Programming. Solving the Curse of Dimensionality, John Wiley & Sons Inc., 2007.
[17] 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, Journal of Control Theory and Applications, 9 (2011), 336–352.
[18] W. P. Powell, Perspectives of approximate dynamic programming, Annals of Operations Research, 241 (2016), 319–356.
[19] M. L. Puterman, *Markov Decision Processes: Discrete Stochastic Dynamic Programming*, Wiley, NY, 1994.

[20] J. Rust, Numerical dynamic programming in economics, in *Handbook of Computational Economics*, (eds. H. Amman, D. Kendrick and J. Rust), Elsevier, 1 (1996), 619–729.

[21] J. Stachurski, Continuous state dynamic programming via nonexpansive approximation, *Computational Economics*, 31 (2008), 141–160.

[22] J. Stachurski, *Dynamic Economic: Theory and Computation*, MIT Press, Cambridge, MA, 2009.

[23] S. Stidham Jr. and R. Weber, A survey of Markov decision models for control of networks of queues, *Queueing Systems*, 13 (1993), 291–314.

[24] R. S. Sutton and A. G. Barto, *Reinforcement Learning: An Introduction*, MIT Press, Cambridge, MA, 1998.

[25] B. Van Roy, Performance loss bounds for approximate value iteration with state space aggregation, *Mathematics of Operations Research*, 31 (2006), 234–244.

[26] O. Vega-Amaya and R. Montes-de-Oca, Application of average dynamic programming to inventory systems, *Mathematical Methods of Operations Research*, 47 (1998), 451–471.

[27] D. J. White, Real applications of Markov decision processes, *Interfaces*, 15 (1985), 73–83.

[28] D. J. White, Further real applications of Markov decision processes, *Interfaces*, 18 (1988), 55–61.

[29] D. J. White, A survey of applications of Markov decision processes, *The Journal of the Operational Research Society*, 44 (1993), 1073–1096.

[30] S. Yakowitz, Dynamic programming applications in water resources, *Water Resource Research*, 18 (1982), 673–696.

[31] W. W.-G. Yeh, Reservoir management and operation models: A state-of-the-art review, *Water Resource Research*, 21 (1985), 1797–1818.

Received December 2015; revised July 2016.

E-mail address: ovega@mat.uson.mx
E-mail address: jlopez@mat.uson.mx