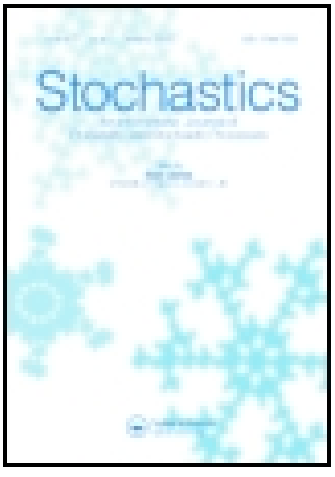

Stochastics An International Journal of Probability and Stochastic Processes: formerly Stochastics and Stochastics Reports

Publication details, including instructions for authors and subscription information:
http://www.tandfonline.com/loi/gssr20

Approximation of average cost Markov decision processes using empirical distributions and concentration inequalities

François Dufour[a] & Tomás Prieto-Rumeau[b]

[a] Team CQFD, Université Bordeaux I, INRIA Bordeaux Sud-Ouest, Bordeaux, France

[b] Statistics Department, UNED, Madrid, Spain

Published online: 07 Nov 2014.

To cite this article: François Dufour & Tomás Prieto-Rumeau (2014): Approximation of average cost Markov decision processes using empirical distributions and concentration inequalities, Stochastics An International Journal of Probability and Stochastic Processes: formerly Stochastics and Stochastics Reports, DOI: 10.1080/17442508.2014.939979

To link to this article: http://dx.doi.org/10.1080/17442508.2014.939979

Stochastics: An International Journal of Probability and Stochastic Processes, 2014
http://dx.doi.org/10.1080/17442508.2014.939979

Approximation of average cost Markov decision processes using empirical distributions and concentration inequalities

François Dufour[a,1] and Tomás Prieto-Rumeau[b]*

[a]*Team CQFD, Université Bordeaux I, INRIA Bordeaux Sud-Ouest, Bordeaux, France;*
[b]*Statistics Department, UNED, Madrid, Spain*

(*Received 6 January 2014; accepted 26 June 2014*)

We consider a discrete-time Markov decision process with Borel state and action spaces, and possibly unbounded cost function. We assume that the Markov transition kernel is absolutely continuous with respect to some probability measure μ. By replacing this probability measure with its empirical distribution μ_n for a sample of size n, we obtain a finite state space control problem, which is used to provide an approximation of the optimal value and an optimal policy of the original control model. We impose Lipschitz continuity properties on the control model and its associated density functions. We measure the accuracy of the approximation of the optimal value and an optimal policy by means of a non-asymptotic concentration inequality based on the 1-Wasserstein distance between μ and μ_n. Obtaining numerically the solution of the approximating control model is discussed and an application to an inventory management problem is presented.

Keywords: Markov decision processes; long-run average cost; approximation of the optimal value and an optimal policy; concentration inequalities; Wasserstein distance

AMS Subject Classification: 90C40; 90C05

1. Introduction

This paper is concerned with numerical methods for Markov decision processes (MDPs). We are interested in approximating numerically the optimal value function and an optimal policy of a discrete-time MDP with Borel state and action spaces, and unbounded cost function under the long-run expected average cost criterion.

MDPs with general (Borel) state and action spaces, and unbounded cost function have been extensively studied from a theoretical point of view; see, e.g. [2,16,18]. The existence of optimal policies and the characterization of the optimal value function have been established using various techniques such as, for instance, dynamic programming and related algorithms (the value iteration and the policy iteration algorithms), and the linear programming (LP) approach. The issue of the approximation of the optimal value and an optimal policy remains, in general, an open issue. This is because, except for some particular cases, the usual approaches to MDPs do not allow to obtain explicitly the optimal value and an optimal policy (not even to approximate them). This is due to the nature itself of the above techniques. As an illustration, the policy and the value iteration algorithms require to perform successive maximizations (over a Borel domain) of functions with Borel domain. Moreover, convergence of the policy iteration algorithm for an average cost control problem requires particularly demanding hypotheses; see [17] and

*Corresponding author. Email: tprieto@ccia.uned.es

the references therein. Also, the LP formulation of an MDP is stated on an infinite-dimensional space of measures over a Borel space, and so it is hardly tractable from a numerical perspective.

On the other hand, a finite MDP (that is, with finite state and actions spaces) can be, in principle, solved numerically. The computational effort, however, grows exponentially with the number of variables involved, thus limiting drastically the practical interest of the above-mentioned techniques. In this context, several approaches have been proposed to solve (or approximate) numerically a finite MDP: reinforcement learning, neuro-dynamic programming, approximate dynamic programs and simulation-based techniques, to name just a few; see the survey [24] and the books [3,8,22,23]. In the particular context of the average cost control problem, there exist several approximation techniques [1,6,7,9,10,13,20,21,25] related to randomized/simulation-based approaches. All these methods are exclusively focused on MDPs with finite or countable state and action spaces, and bounded cost function. We can also mention [19], in which the authors study the convergence of several actor-critic algorithms for MDPs with Borel state and action spaces by minimizing the long-run average cost criterion over a parametrized family of policies; therefore, the optimization problem is not addressed in its full generality.

Summarizing, though there exists an extensive literature on the approximation of discrete and finite MDPs, the challenge of approximating an MDP with general state and action spaces, and unbounded cost function remains open. Our aim is to address this problem for such general MDPs under the long-run expected average cost criterion. We base our approach on the hypothesis that the transition kernel $Q(\mathrm{d}y|x,a)$ defining the dynamics of the original control model $\mathcal{M}$ has a density function $q(y|x,a)$, which satisfies suitable Lipschitz continuity properties, with respect to a reference probability measure $\mu(\mathrm{d}y)$, that is, $Q(\mathrm{d}y|x,a) = q(y|x,a)\mu(\mathrm{d}y)$. The idea is to approximate $\mathcal{M}$ with a control model $\mathcal{M}_{n,\mathfrak{d}}$ defined through the dynamics $q(y|x,a)\mu_n(\mathrm{d}y)$, in which μ is replaced with its empirical distribution μ_n obtained from a sample of size n. Moreover, the action sets $A(x)$ of the original model are replaced in $\mathcal{M}_{n,\mathfrak{d}}$ with *smaller* sets $A_{\mathfrak{d}}(x)$, where the Hausdorff distance between $A(x)$ and $A_{\mathfrak{d}}(x)$ is of (small) order $\mathfrak{d} > 0$.

Concerning our main contributions, let us mention the following points.

- Our framework (an MDP with general state and action spaces with possibly unbounded cost function) is clearly more general than most of the settings studied in the literature. Moreover, the average cost control problem is much more technically demanding in the context of Borel state and action spaces than in the discrete setting. As already mentioned, the policy iteration algorithm, which has been the basis for the development of several numerical procedures to approximate MDPs, is not of great applicability in this context.
- Our approach is based on the construction and the analysis of a simpler model $\mathcal{M}_{n,\mathfrak{d}}$ with finite state space. From this model, we provide an approximation of the value function and we construct an ϵ-optimal control policy for the original control model $\mathcal{M}$.
- The convergence of these approximations is proved and, in addition, the accuracy of the approximations is characterized in terms of a concentration inequality measuring the non-asymptotic deviation between the value function of the original MDP and its approximations. These inequalities are based on the 1-Wasserstein distance between μ and μ_n.

The above convergence results are derived in the general context where the sets $A_{\mathfrak{d}}(x)$ satisfy some weak technical hypotheses. However, the construction of model $\mathcal{M}_{n,\mathfrak{d}}$ and the

associated numerical approximation method depends on the choice of the family of sets $A_{\mathfrak{d}}(x)$ and, in particular, on their topological structure. From a practical point of view, there exist two natural choices for the family $A_{\mathfrak{d}}(x)$:

- The first possibility corresponds to a finite $A_{\mathfrak{d}}(x)$, leading to a finite model $\mathcal{M}_{n,\mathfrak{d}}$. It is then shown that the numerical approximations of the value function and the ϵ-optimal policy can be obtained by solving *two* finite LP problems, for which we can use the powerful solvers available nowadays that can handle really large LPs. In this case, in order to solve $\mathcal{M}_{n,\mathfrak{d}}$, our method could be also combined with the techniques that deal with finite MDPs with 'large' state and actions spaces; see [1,6,7,9,10,13,20,21,25].
- The second natural choice is $A_{\mathfrak{d}}(x) = A(x)$. Then, by using similar arguments as in section 5.3 in [12], it can be shown that the finite-state, possibly infinite-action (hence, still infinite-dimensional) LP problem associated with $\mathcal{M}_{n,\mathfrak{d}}$ reduces to a *finite-dimensional nonlinear optimization problem* that can be solved numerically by using nonlinear optimization tools such as, e.g. the simulated annealing technique.

Finally, let us mention Ref. [14], which follows an approach related to ours. Starting from the recursive equation formulation of an MDP $x_{t+1} = F(x_t, a_t, \xi_t)$ and by replacing the disturbances distribution of $\{\xi_t\}$ with its empirical distribution, the authors measure the accuracy of an optimal policy for the perturbed model in terms of the 1-Wasserstein (or Kantorovich) distance between the original and the empirical disturbances. The nature of the approximation method in [14] is fundamentally different to ours. Indeed, in [14] the idea consists in perturbing the disturbances process, while in our work we approximate (perturb) a probability measure underlying the Markov transition kernel. These two different approaches yield approximating models having different properties. Among them, let us emphasize that the perturbed model in [14] does not have a finite state space, neither considers a modification of the action spaces. Consequently, these points preclude any tractable numerical approach, as opposed to our case. Let us also mention that, in our paper, our hypotheses are only imposed on the original control model, while in [14] the assumptions concern both the original and the perturbed control models; e.g. Assumption 1 in [14]. In connection with the approach in [14], see also Remark 4.2 below.

The rest of the paper is organized as follows. In Section 2 we define the control model $\mathcal{M}$, state our main assumptions and provide some useful basic results on $\mathcal{M}$. The approximating control models $\mathcal{M}_{n,\mathfrak{d}}$ are defined in Section 3 and some of their basic properties are studied. Section 4 addresses the issue of the approximation of the optimal value of $\mathcal{M}$, while Section 5 is concerned with the approximation of an optimal policy. In Section 6 we discuss how to solve numerically the approximating control model $\mathcal{M}_{n,\mathfrak{d}}$. Finally, Section 7 shows an application of our technique to an inventory management system.

Notation. The following notation will be used throughout the paper. Given x and y in the Euclidean space $\mathbb{R}^n$, let $\langle x, y\rangle$ be the usual inner product of x and y. By $|x| = \langle x, x\rangle^{1/2}$ we will denote the norm of $x \in \mathbb{R}^n$. Let $\mathbf{0}$ and $\mathbf{1}$ be the elements of $\mathbb{R}^n$ with all components equal to 0 and 1, respectively. If θ_1 and θ_2 are in $\mathbb{R}^n$, we shall write $\theta_1 \geq \theta_2$ (respectively, $\theta_1 > \theta_2$) when all the components of θ_1 are greater than or equal to (respectively, strictly greater than) the corresponding components of θ_2.

We recall that E is said to be a Borel space if it is a Borel subset of a complete and separable metric space. Its Borel σ-algebra will be denoted by $\mathcal{B}(E)$. If γ is a measure on $(E, \mathcal{B}(E))$ and $v : E \to \mathbb{R}^q$ is a measurable mapping, then the (component-wise) integral of

v with respect to γ will be denoted by $\gamma(v) := \int_E v \, d\gamma \in \mathbb{R}^q$, provided that it is well defined and finite. The Dirac probability measure concentrated at $x \in E$ will be denoted by δ_x; that is, for $B \in \mathcal{B}(E)$ we have $\delta_x(B) = \mathbf{1}_B(x)$, where $\mathbf{1}_B$ denotes the indicator function.

If E is a Borel space and $w : E \to [1, +\infty)$ is a measurable function, the class of measurable functions $f : E \to \mathbb{R}$ such that $\|f\|_w := \sup_{x \in E}\{|f(x)|/w(x)\} < \infty$ is a Banach space denoted by $\mathbb{B}_w(E)$, and $\|\cdot\|_w$ is the associated norm, called the w-norm. The family of real-valued measurable functions on E with finite w-norm which, in addition, are continuous (respectively, Lipschitz continuous), is denoted by $\mathbb{C}_w(E)$ (respectively, $\mathbb{L}_w(E)$). We denote by $\mathcal{M}_w(E)$ the family of measures ν on E such that $\nu(w)$ is finite. For continuous w, the w-weak topology on $\mathcal{M}_w(E)$ is the coarsest topology for which all the real-valued mappings defined on $\mathcal{M}_w(E)$ by $\eta \mapsto \eta(f)$ for $f \in \mathbb{C}_w(E)$ are continuous.

Let $\mathcal{M}^1(E)$ be the family of probability measures λ on E with finite first moment, that is, $\int_E \rho(x, x_0)\lambda(dx) < \infty$ for some $x_0 \in E$, with ρ the metric in E. The 1-Wasserstein metric on $\mathcal{M}^1(E)$ ([4], p. 234) is defined as

$$W_1(\lambda, \lambda') = \sup_{f \in \mathbb{L}} \left| \int_X f \, d\lambda - \int_X f \, d\lambda' \right| \quad \text{for } \lambda, \lambda' \in \mathcal{M}^1(E), \tag{1.1}$$

where $\mathbb{L}$ is the family of 1-Lipschitz continuous functions $f : E \to \mathbb{R}$.

Finally, we let $\mathbb{R}_+ = [0, \infty)$ and $\mathbb{N} = \{0, 1, 2, \dots\}$.

2. The control model

We will deal with the Markov control model $\mathcal{M} := (X, A, \{A(x) : x \in X\}, Q, c)$, where

- X is the state space, assumed to be a Borel space (i.e. a measurable subset of a complete and separable metric space), with metric ρ_X.
- A is the action space, assumed to also be a Borel space, with metric ρ_A.
- The set of feasible controls in state $x \in X$ is $A(x)$, which is a non-empty measurable subset of A. We suppose that $\mathbb{K} := \{(x, a) \in X \times A : a \in A(x)\}$ is a measurable subset of $X \times A$ and that it contains the graph of a measurable function from X to A.
- The stochastic kernel Q on X given $\mathbb{K}$ is the transition probability function.
- The measurable function $c : \mathbb{K} \to \mathbb{R}$ is the cost-per-stage function.

In the family of closed subsets of A the Hausdorff metric is defined as

$$\rho_H(C_1, C_2) := \sup_{a \in C_1} \inf_{a' \in C_2} \{\rho_A(a, a')\} \vee \sup_{a' \in C_2} \inf_{a \in C_1} \{\rho_A(a, a')\}.$$

It is a well-known result that ρ_H is indeed a metric, except that it might not be finite. The following notation will be used throughout. Given a measurable function $v : X \to \mathbb{R}$ we define $Qv : \mathbb{K} \to \mathbb{R}$ as $Qv(x, a) := \int_X v(y)P(dy|x, a)$, provided that the corresponding integrals are well defined and finite.

Let $\mathbb{F}$ be the family of measurable functions $f : X \to A$ such that $f(x) \in A(x)$ for all $x \in X$. By hypothesis, $\mathbb{F}$ is non-empty. Let $H_0 := X$ and $H_t := \mathbb{K} \times H_{t-1}$ for $t \geq 1$ be the history of the MDP up to time t.

Definition 2.1. A control policy is a sequence $\pi = \{\pi_t\}_{t \in \mathbb{N}}$ of stochastic kernels π_t on A given H_t satisfying $\pi_t(A(x_t)|h_t) = 1$ for all $h_t \in H_t$ and $t \in \mathbb{N}$, where $h_t := (x_0, a_0, \dots, x_{t-1}, a_{t-1}, x_t)$. Let Π be the class of all policies.

Let Φ be the family of all stochastic kernels φ on A given X such that $\varphi(A(x)|x) = 1$ for all $x \in X$. A policy π is said to be randomized stationary if there exists $\varphi \in \Phi$ such that $\pi_t(\cdot|h_t) = \varphi(\cdot|x_t)$ for all $t \in \mathbb{N}$ and $h \in H_t$. The class of randomized stationary policies is identified with Φ.

Each $f \in \mathbb{F}$ is identified the policy $\pi \in \Pi$ such that $\pi_t(\cdot|h_t)$ is the Dirac measure concentrated at $f(x_t)$, for all $x_t \in X$ and $t \in \mathbb{N}$. We say that $f \in \mathbb{F}$ is a deterministic stationary policy. Clearly, we have $\mathbb{F} \subseteq \Phi \subseteq \Pi$.

Let $((X \times A)^\infty, \mathcal{B}((X \times A)^\infty))$ be the canonical space consisting of the set of sample paths $(X \times A)^\infty = \{(x_t, a_t)\}_{t \in \mathbb{N}}$ and the associated product σ-algebra. Therefore, $\{x_t\}_{t \in \mathbb{N}}$ stands for the state process and $\{a_t\}_{t \in \mathbb{N}}$ is the action process. For notational convenience, we define $\mathcal{F}_t$ as the σ-algebra generated by (x_s, a_s) for $0 \leq s \leq t$. From ([16], Chapter 2) we know that for every policy $\pi \in \Pi$ and any initial state $x \in X$ there exists a unique probability measure $P^{\pi,x}$ on $((X \times A)^\infty, \mathcal{B}((X \times A)^\infty))$ such that, for any $B \in \mathcal{B}(X)$, $C \in \mathcal{B}(A)$ and $h_t \in H_t$ with $t \in \mathbb{N}$,

$$P^{\pi,x}(x_0 \in B) = \mathbf{I}_B(x), \quad P^{\pi,x}(a_t \in C|h_t) = \pi_t(C|h_t), \quad \text{and}$$
$$P^{\pi,x}(x_{t+1} \in B|h_t, a_t) = Q(B|x_t, a_t).$$

The expectation operator associated with $P^{\pi,x}$ is denoted by $E^{\pi,x}$.

Let $0 < \alpha < 1$ be a given discount factor. The total expected α-discounted cost of the policy $\pi \in \Pi$ for the initial state $x \in X$ is defined as

$$V_\alpha(x, \pi) := E^{\pi,x}\left[\sum_{t=0}^{\infty} \alpha^t c(x_t, a_t)\right],$$

and the long-run expected average cost of the policy $\pi \in \Pi$ for the initial state $x \in X$ is given by

$$J(x, \pi) := \overline{\lim_{t \to \infty}} \frac{1}{t} E^{\pi,x}\left[\sum_{k=0}^{t-1} c(x_k, a_k)\right].$$

The value function of the α-discounted control problem is

$$V_\alpha^*(x) := \inf_{\pi \in \Pi} V_\alpha(x, \pi) \quad \text{for } x \in X$$

and a policy $\pi^* \in \Pi$ is said to be α-discount optimal if $V_\alpha(x, \pi^*) = V_\alpha^*(x)$ for every $x \in X$. Similarly, the value function of the average cost control problem is

$$J^*(x) := \inf_{\pi \in \Pi} J(x, \pi) \quad \text{for } x \in X$$

and a policy $\pi^* \in \Pi$ is said to be AC-optimal if $J(x, \pi^*) = J^*(x)$ for every $x \in X$.

Next we state our assumptions on the elements of the control model $\mathcal{M}$.

Assumption A.

(A1) The action set $A(x)$ is compact for each $x \in X$. The multifunction Ψ from X to A defined by $\Psi(x) := A(x)$ is L_Ψ-Lipschitz continuous with respect to the Hausdorff metric.

(A2) There exists an L_w-Lipschitz continuous function $w : X \to [1, \infty)$ such that there are constants $d \in (0, 1)$ and $b \in \mathbb{R}_+$ with

$$Qw(x, a) \leq dw(x) + b \quad \text{for all } (x, a) \in \mathbb{K}. \tag{2.1}$$

(A3) The cost function c is in $\mathbb{L}_w(\mathbb{K})$, i.e. c is L_c-Lipschitz continuous on $\mathbb{K}$ and there is a positive constant $\bar{c}$ such that, for all $(x, a) \in \mathbb{K}$,

$$|c(x, a)| \leq \bar{c}w(x).$$

The Lipschitz condition in Assumption (A1) means that there exists a positive L_Ψ such that $\rho_H(A(x), A(y)) \leq L_\Psi \rho_X(x, y)$ for every $x, y \in X$. As a consequence of ([11], Lemma 2.6), the multifunction $\Psi : X \to A$ is continuous.

Assumption B.

(B1) There exists a probability measure μ on $(X, \mathcal{B}(X))$ and a measurable function $q : X \times \mathbb{K} \to \mathbb{R}_+$ such that $\mu(w) < +\infty$ and

$$Q(B|x, a) = \int_B q(y|x, a)\mu(\mathrm{d}y) \quad \text{for all } B \in \mathcal{B}(X) \text{ and } (x, a) \in \mathbb{K}.$$

(B2) There exist positive constants $\overline{q}$, L_q and L_{wq} such that function q satisfies the following properties:

$$q(y|x, a) \leq \bar{q}w(x), \tag{2.2}$$

$$|q(y|x, a) - q(z|x, a)| \leq L_q \rho_X(y, z), \tag{2.3}$$

$$|q(y|x, a) - q(y|x', a')| \leq L_q[\rho_X(x, x') + \rho_A(a, a')], \tag{2.4}$$

$$|w(y)q(y|x, a) - w(z)q(z|x, a)| \leq L_{wq} w(x) \rho_X(y, z), \tag{2.5}$$

for every (x, a) and (x', a') in $\mathbb{K}$, and $y, z \in X$.

(B3) There exists some $x_0 \in X$ and $\mathfrak{a} > 0$ such that

$$\int_X \exp\{\mathfrak{a}\rho_X(x, x_0)\}\mu(\mathrm{d}x) < \infty$$

and, in particular, $\mu \in \mathcal{M}^1(X)$.

Remark 2.2. From Assumptions (B1) and (2.4) in Assumption (B2), it follows easily that for any $v \in \mathbb{B}_w(X)$, (x, a) and (x', a') in $\mathbb{K}$

$$|Qv(x, a) - Qv(x', a')| \leq L_q \|v\|_w \mu(w)[\rho_X(x, x') + \rho_A(a, a')], \tag{2.6}$$

and so Qv is Lipschitz continuous on $\mathbb{K}$. In particular, Q is strongly continuous (that is, Qv is continuous on $\mathbb{K}$ for any bounded measurable function v on X) and Qw is continuous on $\mathbb{K}$.

The following facts on discount optimality are well known. For a proof, the reader is referred to ([15], Lemma 3.2). For every discount factor $0 < \alpha < 1$, the discounted optimal value function V_α^* verifies

$$\left|V_\alpha^*(x)\right| \leq \frac{\bar{c}}{1-d}\left(w(x) + \frac{b}{1-\alpha}\right) \quad \text{for all } x \in X \tag{2.7}$$

and, in particular, $V_\alpha^* \in \mathbb{B}_w(X)$. Moreover, V_α^* is a solution of the α-discounted cost optimality Equation (α-DCOE)

$$V_\alpha^*(x) = \min_{a \in A(x)} \{c(x,a) + \alpha \int_X V_\alpha^*(y) Q(\mathrm{d}y|x,a)\} \quad \text{for } x \in X. \tag{2.8}$$

When dealing with average optimality, we impose another condition.

Assumption C. There exists $x_0 \in X$ such that function h_α defined by $h_\alpha(x) = V_\alpha^*(x) - V_\alpha^*(x_0)$ satisfies $\bar{h} = \sup_{\alpha \in (0,1)} \|h_\alpha\|_w < \infty$.

Let us introduce $g_\alpha = (1-\alpha)V_\alpha^*(x_0)$. A standard calculation shows that the α-DCOE can be re-written as

$$g_\alpha + h_\alpha(x) = \min_{a \in A(x)} \{c(x,a) + \alpha \int_X h_\alpha(y) Q(\mathrm{d}y|x,a)\}.$$

Lemma 2.3. For any $0 < \alpha < 1$, function h_α is in $\mathbb{L}_w(X)$ and its Lipschitz constant is

$$\left[L_c + L_q \mu(w)\bar{h}\right](1 + L_\Psi).$$

Proof. For notational convenience, let us introduce

$$\Delta_\alpha(x,a,x',a') := |c(x,a) - c(x',a')| + \alpha |Qh_\alpha(x,a) - Qh_\alpha(x',a')|$$

for (x,a) and (x',a') in $\mathbb{K}$. Then we have

$$c(x,a) + \alpha Qh_\alpha(x,a) \leq c(x',a') + \alpha Qh_\alpha(x',a') + \Delta_\alpha(x,a,x',a')$$

for any (x,a) and (x',a') in $\mathbb{K}$, implying that

$$\begin{aligned} h_\alpha(x) &\leq -g_\alpha + c(x',a') + \alpha Qh_\alpha(x',a') + \inf_{a \in A(x)} \{\Delta_\alpha(x,a,x',a')\} \\ &\leq h_\alpha(x') + \sup_{a' \in A(x')} \inf_{a \in A(x)} \{\Delta_\alpha(x,a,x',a')\}. \end{aligned}$$

By symmetry, we obtain that

$$h_\alpha(x') \leq h_\alpha(x) + \sup_{a \in A(x)} \inf_{a' \in A(x')} \{\Delta_\alpha(x,a,x',a')\}.$$

Therefore, for all x and x' in X

$$|h_\alpha(x) - h_\alpha(x')| \leq \sup_{a\in A(x)} \inf_{a'\in A(x')} \{\Delta_\alpha(x,a,x',a')\} \vee \sup_{a'\in A(x')} \inf_{a\in A(x)} \{\Delta_\alpha(x,a,x',a')\}.$$

On the other hand, from Assumptions (A3), C and Remark 2.2

$$\Delta_\alpha(x,a,x',a') \leq \big[L_c + \alpha L_q \mu(w)\bar{h}\big](\rho_X(x,x') + \rho_A(a,a')).$$

Now using Assumption (A1), the result follows. □

We say that the pair $(g,h) \in \mathbb{R}\times\mathcal{B}_w(X)$ is a solution to the average cost optimality Equation (ACOE) for $\mathcal{M}$ if

$$g + h(x) = \min_{a\in A(x)} \left\{ c(x,a) + \int_X h(y) Q(\mathrm{d}y|x,a) \right\} \quad \text{for any } x \in X. \tag{2.9}$$

It is important to note that, in general, there might not exist solutions to the ACOE; rather, there are solutions to the average cost optimality inequality (or inequalities); see [15] or ([18], Theorem 10.3.1). In our next proposition we prove that, under our hypotheses on the control model, there exists a solution (g,h) to the ACOE for which, besides, h is Lipschitz continuous.

Proposition 2.4. Under Assumptions A, B and C there exists a solution $(g^*,h) \in \mathbb{R}\times\mathbb{L}_w(X)$ to the ACOE for the control model $\mathcal{M}$. Moreover, $g^* = J^*(x)$ for every $x \in X$.

Proof. From (2.7), we have that g_α is bounded for $0 < \alpha < 1$. In addition, from Lemma 2.3, the family of functions $\{h_\alpha : 0 < \alpha < 1\}$ is equicontinuous. Therefore, from Assumption C and Ascoli's theorem, it follows that there exist $(g^*,h) \in \mathbb{R}\times\mathbb{B}_w(X)$, with $\|h\|_w \leq \bar{h}$, and a sequence $\{\alpha_k\}$ such that

$$\alpha_k \to 1, \quad g_{\alpha_k} \to g^* \quad \text{and} \quad h_{\alpha_k}(x) \to h(x) \quad \text{for any } x \in X$$

as $k \to \infty$. In addition, the fact that the Lipschitz constants of h_α do not depend on α (recall Lemma 2.3) implies that $h \in \mathbb{L}_w(X)$, with Lipschitz constant $[L_c + L_q\mu(w)\bar{h}](1 + L_\Psi)$. Now, by using standard arguments, see, for instance, Theorem 4.1 in [15], we obtain the result. □

We conclude this section with a technical result that will be useful in the sequel.

Lemma 2.5. Suppose that Assumptions A and B hold. Given $v \in \mathbb{L}_w(X)$ and $(x,a) \in \mathbb{K}$, the function $y \mapsto v(y)q(y|x,a)$ is Lipschitz continuous on X with Lipschitz constant given by $\mathcal{K}_v w(x)$, with

$$\mathcal{K}_v := \|v\|_w(L_{wq} + \bar{q}L_w) + \bar{q}L_v.$$

Proof. For every $y,z \in X$ we have, from (2.2) in Assumption (B2),

$$\begin{aligned} |v(y)q(y|x,a) - v(z)q(z|x,a)| &\leq |v(y)||q(y|x,a) - q(z|x,a)| + q(z|x,a)|v(y) - v(z)| \\ &\leq \|v\|_w w(y)|q(y|x,a) - q(z|x,a)| + L_v\bar{q}w(x)\rho_X(y,z). \end{aligned}$$

Function w being Lipschitz continuous – Assumption (A2) – and from (2.2) and (2.5) in Assumption (B2),

$$w(y)|q(y|x,a) - q(z|x,a)| \le |w(y)q(y|x,a) - w(z)q(z|x,a)| + |w(y) - w(z)|q(z|x,a)$$
$$\le (L_{wq} + \bar{q}L_w)w(x)\rho_X(y,z).$$

The stated result follows. □

3. The approximating control models

Throughout this section, we suppose that Assumptions A and B are satisfied.

We suppose that there is a probability space $(\Omega, \mathcal{F}, \mathbb{P})$ and a family $\{Y_n\}_{n\ge1}$ of i.i.d. random variables taking values in X with distribution μ. For each $n \ge 1$, the $\mathcal{M}^1(X)$-valued mapping μ_n defined on Ω by

$$\mu_n(\mathrm{d}y) = \frac{1}{n}\sum_{k=1}^{n} \delta_{Y_k}(\mathrm{d}y) \tag{3.1}$$

is called the empirical probability measure. It is a random variable since the $\mathcal{M}^1(X)$-valued mapping defined on X^n by $(x_1, \ldots, x_n) \mapsto 1/n\sum_{i=1}^n \delta_{x_i}$ is continuous. As a consequence, the 1-Wasserstein distance $W_1(\mu_n, \mu)$ is a real-valued random variable. We will denote by $\mathbb{P}^*$ the outer measure associated with the probability measure $\mathbb{P}$, which is defined on the class of all subsets of Ω.

Let us recall a known result. For a proof, the reader is referred to Corollary 2.5 and Theorem A.5 in [5].

THEOREM 3.1. Suppose that the probability measure μ satisfies Assumption (B3). There exists some γ_0 such that, given any $0 < \gamma \le \gamma_0$, there are constants $\mathcal{S}, \mathcal{T} > 0$, depending on γ, with

$$\mathbb{P}\{W_1(\mu_n, \mu) > \gamma\} \le \mathcal{S}\exp\{-\mathcal{T}n\} \quad \text{for all } n \ge 1.$$

The following notation will be useful in the forthcoming. Given $z > 0$, define for $n \ge 1$

$$F_n(z) = \{\omega \in \Omega : W_1(\mu_n(\omega), \mu) \le z\} \in \mathcal{F}.$$

DEFINITION 3.2. The constants $\mathfrak{c}_i$ are defined as

$$\mathfrak{c}_1 = \frac{1}{2L_q}, \quad \mathfrak{c}_2 = \frac{1-d}{4(L_q + L_{wq})} \quad \text{and} \quad \mathfrak{c}_3 = \frac{1-d}{4(L_{wq} + L_q(1 + 4(d+b)))}.$$

They verify $\mathfrak{c}_3 \le \mathfrak{c}_2 \le \mathfrak{c}_1$, and so, for $n \ge 1$ we have $F_n(\mathfrak{c}_3) \subseteq F_n(\mathfrak{c}_2) \subseteq F_n(\mathfrak{c}_1)$.

The meaning of constants $\mathfrak{c}_i$ will become clear later. We make another hypothesis.

Assumption D. For all $\mathfrak{d} > 0$ there exists a family $A_{\mathfrak{d}}(x)$, for $x \in X$, of subsets of A satisfying the following hypotheses.

(D1) For every $x \in X$, $A_{\mathfrak{d}}(x)$ is a non-empty closed subset of $A(x)$. We suppose that

$$\mathbb{K}_{\mathfrak{d}} = \{(x,a) \in X \times A : a \in A_{\mathfrak{d}}(x)\}$$

is a measurable subset of $X \times A$, containing the graph of a measurable function from X to A.

(D2) For every $x \in X$,

$$\rho_H(A(x), A_{\mathfrak{d}}(x)) \leq \mathfrak{d} w(x). \tag{3.2}$$

(D3) The multifunction $\Psi_{\mathfrak{d}}$ from X to A defined by $\Psi_{\mathfrak{d}}(x) := A_{\mathfrak{d}}(x)$ is $L_{\Psi,\mathfrak{d}}$-Lipschitz continuous with respect to the Hausdorff metric. We suppose that $\sup_{\mathfrak{d}>0} L_{\Psi,\mathfrak{d}} = L_{\Psi}^{*} < \infty$.

We define, for $n \geq 1$ and $(x,a) \in \mathbb{K}_{\mathfrak{d}}$,

$$\beta_n(x,a) = \int_X q(y|x,a)\mu_n(\mathrm{d}y). \tag{3.3}$$

Observe that

$$\beta_n(x,a) - 1 = \int_X q(y|x,a)\mu_n(\mathrm{d}y) - \int_X q(y|x,a)\mu(\mathrm{d}y)$$

and so, from Lipschitz continuity of $y \mapsto q(y|x,a)$ for fixed (x,a) – see (2.3) in Assumption (B2)

$$|\beta_n(x,a) - 1| \leq L_q W_1(\mu, \mu_n) \quad \text{for all } (x,a) \in \mathbb{K}_{\mathfrak{d}}. \tag{3.4}$$

We introduce a family of random kernels Q_n for $n \geq 1$ depending on $\omega \in \Omega$ through the empirical measure μ_n associated with μ.

Definition 3.3. Given $n \geq 1$ and for $\omega \in F_n(\mathfrak{c}_1)$ consider the kernel Q_n on X given $\mathbb{K}_{\mathfrak{d}}$

$$Q_n(B|x,a) = \frac{1}{\beta_n(x,a)} \int_B q(y|x,a)\mu_n(\mathrm{d}y) = \frac{\sum_{\{k:Y_k \in B\}} q(Y_k|x,a)}{\sum_{k=1}^{n} q(Y_k|x,a)}$$

for $B \in \mathcal{B}(X)$ and $(x,a) \in \mathbb{K}_{\mathfrak{d}}$.

Note that for every fixed $\omega \in \Omega$, the kernel Q_n is supported on the finite set $\{Y_k(\omega)\}_{1 \leq k \leq n}$. The condition $\omega \in F_n(c_1)$ ensures that $\beta_n(x,a) \geq 1/2$ for all $(x,a) \in \mathbb{K}_{\mathfrak{d}}$; therefore, $Q_n(\cdot|x,a)$ is well defined on $F_n(\mathfrak{c}_1)$. We also have that Q_n is a stochastic kernel on $F_n(\mathfrak{c}_1)$, meaning that $Q_n(X|x,a) = 1$ for $(x,a) \in \mathbb{K}_{\mathfrak{d}}$.

Definition 3.4. Given $n \geq 1$, $\omega \in F_n(\mathfrak{c}_1)$ and $\mathfrak{d} > 0$, the control model $\mathcal{M}_{n,\mathfrak{d}}$ is defined by the following elements

$$\big(X, A, \{A_{\mathfrak{d}}(x) : x \in X\}, Q_n, c\big)$$

(cf. the definition of the control model $\mathcal{M}$).

Associated with the control model $\mathcal{M}_{n,\mathfrak{d}}$, the set $\Pi_{\mathfrak{d}}$ denotes the family of all control policies. Given $\pi \in \Pi_{\mathfrak{d}}$ and any initial state $x \in X$, let $P_{n,\mathfrak{d}}^{\pi,x}$ be the underlying probability measure on $((X \times A)^{\infty}, \mathcal{B}((X \times A)^{\infty}))$ such that, for any $B \in \mathcal{B}(X)$, $C \in \mathcal{B}(A)$, and $h_t \in H_t$ with $t \in \mathbb{N}$,

$$P_{n,\mathfrak{d}}^{\pi,x}(x_0 \in B) = \mathbf{I}_B(x), \quad P_{n,\mathfrak{d}}^{\pi,x}(a_t \in C|h_t) = \pi_t(C|h_t), \quad \text{and}$$
$$P_{n,\mathfrak{d}}^{\pi,x}(x_{t+1} \in B|h_t, a_t) = Q_n(B|x_t, a_t).$$

For notational convenience, $E_{n,\mathfrak{d}}^{\pi,x}$ will denote the associated expectation operator.

Let $\mathbb{F}_{\mathfrak{d}}$ be the family of measurable functions $f : X \to A$ such that $f(x) \in A_{\mathfrak{d}}(x)$ for all $x \in X$. Clearly, $\mathbb{F}_{\mathfrak{d}} \subseteq \mathbb{F}$. Also, let $\Phi_{\mathfrak{d}} \subseteq \Phi$ be the family of stochastic kernels on A given X such that $\varphi(A_{\mathfrak{d}}(x)|x) = 1$ for all $x \in X$. Sets $\mathbb{F}_{\mathfrak{d}}$ and $\Phi_{\mathfrak{d}}$ are identified with the class of stationary deterministic and randomized policies for $\mathcal{M}_{n,\mathfrak{d}}$, respectively. We will also use the notation $Q_n(\cdot|x, f)$ and $Q_n(\cdot|x, \varphi)$ to denote the Markov kernels associated with the stationary policies $f \in \mathbb{F}_{\mathfrak{d}}$ and $\varphi \in \Phi_{\mathfrak{d}}$, i.e.

$$Q_n(B|x, f) = Q_n(B|x, f(x)) \quad \text{and} \quad Q_n(B|x, \varphi) = \int_{A_{\mathfrak{d}}(x)} Q_n(B|x, a)\varphi(\mathrm{d}a|x)$$

for $B \in \mathcal{B}(X)$ and $x \in X$.

It is worth noting that the Markov chain with transition function $Q_n(\cdot|x, \varphi)$, for $x \in X$ and $\varphi \in \Phi_{\mathfrak{d}}$, is essentially a *finite state* Markov chain. Indeed, for whatever initial state $x_0 \in X$, the subsequent states x_t for $t \geq 1$ lie in the finite set $\Gamma_n = \{Y_k(\omega)\}_{1 \leq k \leq n}$. This finite state space, however, varies with $\omega \in \Omega$ and $n \geq 1$.

Propositions 3.5 and 3.6 below explore some properties of the control models $\mathcal{M}_{n,\mathfrak{d}}$. They suppose that Assumptions A, B and D hold.

PROPOSITION 3.5. Fix $n \geq 1$, $\omega \in F_n(\mathfrak{c}_1)$ and $\mathfrak{d} > 0$. For any $v \in \mathbb{B}_w(X)$,

$$|Q_n v(x, a) - Q_n v(x', a')| \leq 2\|v\|_w L_q \left[\mu(w) + \frac{L_w}{2L_q}\right] \left[1 + 2\bar{q}w(x)\right] \left[\rho_X(x, x') + \rho_A(a, a')\right],$$

for any (x, a) and (x', a') in $\mathbb{K}_{\mathfrak{d}}$. In particular, the mapping $(x, a) \mapsto Q_n v(x, a)$ is locally Lipschitz continuous on $\mathbb{K}_{\mathfrak{d}}$.

Proof. First observe that

$$|\beta_n(x, a) - \beta_n(x', a')| \leq L_q[\rho_X(x, x') + \rho_A(a, a')], \tag{3.5}$$

by definition of β_n (see Equation (3.3)) and

$$\mu_n(w) \leq \mu(w) + \frac{L_w}{2L_q}, \tag{3.6}$$

by using the fact that w is L_w-Lipschitz continuous and $\omega \in F_n(\mathfrak{c}_1)$. Now, $Q_n v(x, a) - Q_n v(x', a')$ equals

$$\frac{1}{\beta_n(x, a)} \int_X v(y)q(y|x, a)\mu_n(\mathrm{d}y) - \frac{1}{\beta_n(x', a')} \int_X v(y)q(y|x', a')\mu_n(\mathrm{d}y)$$

and so

$$\begin{aligned}|Q_n v(x,a) - Q_n v(x',a')| &\le \frac{1}{\beta_n(x,a)}\|v\|_w \int_X w(y)|q(y|x,a) - q(y|x',a')|\mu_n(\mathrm{d}y) \\ &\quad + \frac{|\beta_n(x,a) - \beta_n(x',a')|}{\beta_n(x,a)\beta_n(x',a')}\|v\|_w \int_X w(y)q(y|x,a)\mu_n(\mathrm{d}y) \\ &\le 2\|v\|_w L_q \mu_n(w)\big[1 + 2\bar{q}w(x)\big]\big[\rho_X(x,x') + \rho_A(a,a')\big],\end{aligned}$$

by using Equation (3.5) and the fact that $\beta_n(x,a) \ge 1/2$. Finally, combining the previous equation and (3.6) the result follows. □

As a direct consequence of this lemma, we obtain that the kernel Q_n is strongly continuous, and also that $Q_n w$ is continuous on $\mathbb{K}_\mathfrak{d}$. It is worth noting that, for $v \in \mathbb{B}_w(X)$, the mapping Qv is Lipschitz continuous (see Remark 2.2). For the kernel Q_n this might not hold, and we have that $Q_n v$ is locally Lipschitz continuous.

PROPOSITION 3.6. Given $n \ge 1$ and $\mathfrak{d} > 0$, on $F_n(\mathfrak{c}_2)$ we have

$$Q_n w(x,a) \le \frac{1+d}{2} w(x) + 2b \quad \text{for any } (x,a) \in \mathbb{K}_\mathfrak{d}.$$

As a consequence,

$$E^{\pi,x}_{n,\mathfrak{d}}[w(x_t)] \le \left(\frac{1+d}{2}\right)^t w(x) + \frac{4b}{1-d} \tag{3.7}$$

for any $x \in X$, $\pi \in \Pi_\mathfrak{d}$ and $t \in \mathbb{N}$.

Proof. Note that

$$Q_n w(x,a) \le \frac{|1 - \beta_n(x,a)|}{\beta_n(x,a)} \int_X w(y)q(y|x,a)\mu_n(\mathrm{d}y) + \int_X w(y)q(y|x,a)\mu_n(\mathrm{d}y).$$

By Lipschitz continuity of wq and Assumption (A2) we have

$$\begin{aligned}\int_X w(y)q(y|x,a)\mu_n(\mathrm{d}y) &\le L_{wq} w(x) W_1(\mu,\mu_n) + Qw(x,a) \\ &\le L_{wq} w(x) W_1(\mu,\mu_n) + dw(x) + b.\end{aligned}$$

Therefore, from Equation (3.4)

$$Q_n w(x,a) \le \big[2L_q W_1(\mu,\mu_n) + 1\big]\big[L_{wq} w(x) W_1(\mu,\mu_n) + dw(x) + b\big]$$

and so,

$$Q_n w(x,a) \le d_n w(x) + 2b \quad \text{for all } (x,a) \in \mathbb{K}_\mathfrak{d},$$

with $d_n = d + 2W_1(\mu,\mu_n)(L_q + L_{wq})$, and the first statement easily follows. The second statement of this proposition is now a direct consequence. □

Given $n \geq 1$, $\mathfrak{d} > 0$ and $\omega \in F_n(\mathfrak{c}_2)$ we can introduce, for the control model $\mathcal{M}_{n,\mathfrak{d}}$, the total expected α-discounted cost of the policy $\pi \in \Pi$ for the initial state $x \in X$ defined as

$$V_{\alpha,n,\mathfrak{d}}(x,\pi) := E^{\pi,x}_{n,\mathfrak{d}}\left[\sum_{t=0}^{\infty} \alpha^t c(x_t, a_t)\right],$$

and the long-run expected average cost of the policy $\pi \in \Pi_{\mathfrak{d}}$ for the initial state $x \in X$ is given by

$$J_{n,\mathfrak{d}}(x,\pi) := \overline{\lim_{t\to\infty}} E^{\pi,x}_{n,\mathfrak{d}}\left[\frac{1}{t}\sum_{k=0}^{t-1} c(x_k, a_k)\right].$$

The value function of the α-discounted control problem is

$$V^*_{\alpha,n,\mathfrak{d}}(x) := \inf_{\pi\in\Pi_{\mathfrak{d}}} V_{\alpha,n,\mathfrak{d}}(x,\pi) \quad \text{for } x \in X$$

and a policy $\pi^* \in \Pi_{\mathfrak{d}}$ is said to be α-discount optimal if $V_{\alpha,n,\mathfrak{d}}(x,\pi^*) = V^*_{\alpha,n,\mathfrak{d}}(x)$ for every $x \in X$. Similarly, the value function of the average cost control problem is

$$J^*_{n,\mathfrak{d}}(x) := \inf_{\pi\in\Pi_{\mathfrak{d}}} J_{n,\mathfrak{d}}(x,\pi) \quad \text{for } x \in X$$

and a policy $\pi^* \in \Pi_{\mathfrak{d}}$ is said to be AC-optimal if $J_{n,\mathfrak{d}}(x,\pi^*) = J^*_{n,\mathfrak{d}}(x)$ for every $x \in X$.

Observe that $J^*_{n,\mathfrak{d}}(x)$ is defined on $F_n(\mathfrak{c}_2)$ as the optimal value function of the control model $\mathcal{M}_{n,\mathfrak{d}}$. For completeness, we propose the following definition of $J^*_{n,\mathfrak{d}}(x)$. If the control model $\mathcal{M}_{n,\mathfrak{d}}$ is well defined and its optimal average cost value function exists, let $J^*_{n,\mathfrak{d}}(x)$ be the corresponding value function (in particular, this holds when $\omega \in F_n(\mathfrak{c}_2)$). Otherwise, define $J^*_{n,\mathfrak{d}}(x) \in \mathbb{R}$ arbitrarily.

4. Approximation of the optimal average cost

Our next results compare the difference between Qv and $Q_n v$ when v is a Lipschitz continuous function.

PROPOSITION 4.1. Given $n \geq 1$, $v \in \mathbb{L}_w(X)$ and $(x,a) \in \mathbb{K}_{\mathfrak{d}}$, on $F_n(\mathfrak{c}_1)$ we have

$$|Qv(x,a) - Q_n v(x,a)| \leq \mathcal{C}_v w(x) W_1(\mu,\mu_n), \tag{4.1}$$

with $\mathcal{C}_v = 2\mathcal{K}_v + 2L_q\|v\|_w(d+b)$, where constant $\mathcal{K}_v$ comes from Lemma 2.5.

Proof. First of all, observe that $Q_n v(x,a)$ is finite for any $(x,a) \in \mathbb{K}_{\mathfrak{d}}$ by using Proposition 3.6. From Definition 3.3 we have

$$Qv(x,a) - Q_n v(x,a) = \int_X v(y)q(y|x,a)\mu(\mathrm{d}y) - \frac{1}{\beta_n(x,a)}\int_X v(y)q(y|x,a)\mu_n(\mathrm{d}y)$$

and so,

$$\begin{aligned}|Qv(x,a) - Q_n v(x,a)| &\leq \frac{1}{\beta_n(x,a)}\left|\int_X v(y)q(y|x,a)\mu(\mathrm{d}y) - \int_X v(y)q(y|x,a)\mu_n(\mathrm{d}y)\right| \\ &\quad + \frac{|\beta_n(x,a)-1|}{\beta_n(x,a)}\int_X |v(y)|q(y|x,a) \\ &\leq 2\mathcal{K}_v w(x) W_1(\mu,\mu_n) + 2L_q W_1(\mu,\mu_n)\cdot\|v\|_w(dw(x)+b),\end{aligned}$$

where we have used Lemma 2.5, Equation (3.4) and Assumption (A2). The stated result follows. □

Remark 4.2. When comparing our hypotheses to those in [14], note that Assumption 2(d) in [14] implies, using our notation, that $|Qv(x,a) - Q_n v(x,a)|$ is bounded above by a constant multiplied by the distance between the original and the perturbed disturbance distributions. Under our hypotheses, we obtain a weaker inequality in Proposition 4.1 because we have, in addition, the multiplicative term $w(x)$.

Next we state our main results in this section.

THEOREM 4.3. Suppose that Assumptions A, B, C and D are satisfied. There exist positive constants $\mathbf{G}_1$ and $\mathbf{G}_2$ such that for every $n \geq 1$, $\mathfrak{d} > 0$ and $\omega \in F_n(\mathfrak{c}_2)$ we have

$$\left|g^* - J^*_{n,\mathfrak{d}}(x)\right| \leq \mathbf{G}_1 W_1(\mu,\mu_n) + \mathbf{G}_2\mathfrak{d} \quad \text{for } x \in X.$$

Proof. Given $\pi \in \Pi_{\mathfrak{d}}$ and $x \in X$, observe that $E^{\pi,x}_{n,\mathfrak{d}}[h(x_t)]$ is finite by using Proposition 3.6. On the other hand, for any $(x,a) \in \mathbb{K}_{\mathfrak{d}}$ we have from Proposition 4.1

$$g^* + h(x) \leq c(x,a) + Q_n h(x,a) + \mathcal{C}_h w(x) W_1(\mu,\mu_n).$$

Consequently, taking the expectation

$$g^* + E^{\pi,x}_{n,\mathfrak{d}}[h(x_k)] \leq E^{\pi,x}_{n,\mathfrak{d}}[c(x_k,a_k)] + E^{\pi,x}_{n,\mathfrak{d}}[h(x_{k+1})] + \mathcal{C}_h W_1(\mu,\mu_n) E^{\pi,x}_{n,\mathfrak{d}}[w(x_k)]$$

for all $k \geq 0$. Summing over $k = 0, \ldots, t-1$ yields

$$tg^* + h(x) \leq \sum_{k=0}^{t-1} E^{\pi,x}_{n,\mathfrak{d}}[c(x_k,a_k)] + E^{\pi,x}_{n,\mathfrak{d}}[h(x_t)] + \mathcal{C}_h W_1(\mu,\mu_n)\sum_{k=0}^{t-1} E^{\pi,x}_{n,\mathfrak{d}}[w(x_k)]. \tag{4.2}$$

From Proposition 3.6 we have

$$\sum_{k=0}^{t-1} E^{\pi,x}_{n,\mathfrak{d}}[w(x_t)] \leq \frac{2}{1-d}w(x) + \frac{4bt}{1-d} \tag{4.3}$$

and

$$E^{\pi,x}_{n,\mathfrak{d}}|h(x_t)| \leq \bar{h}\left[w(x) + \frac{4b}{1-d}\right], \tag{4.4}$$

with $\overline{h}$ as in Assumption C (recall the proof of Proposition 2.4). Dividing by t in (4.2), combining Equations (4.3) and (4.4), and taking the superior limit as t tends to infinity,

$$g^* \leq \varlimsup_{t\to\infty} \frac{1}{t}\sum_{k=0}^{t-1} E^{\pi,x}_{n,\mathfrak{d}}[c(x_k,a_k)] + \mathcal{C}_h W_1(\mu,\mu_n)\frac{4b}{1-d}.$$

Since π is arbitrary, we obtain

$$g^* \leq J^*_{n,\mathfrak{d}}(x) + \mathcal{C}_h W_1(\mu,\mu_n)\frac{4b}{1-d}. \tag{4.5}$$

Recalling Remark 2.2, Assumptions (A1) and (A3), it follows by using Proposition D.5 in [16] that there is a measurable selector $f^* \in \mathbb{F}$ such that

$$g^* + h(x) = c(x,f^*(x)) + Qh(x,f^*(x)). \tag{4.6}$$

Now observe that the function defined on $\mathbb{K}_{\mathfrak{d}}$ by $(x,a) \to \rho_A(a,f^*(x))$ satisfies the hypothesis of Proposition D.5 in [16]. Consequently, by using Assumption D there exists $\tilde{f} \in \mathbb{F}_{\mathfrak{d}}$ satisfying

$$\min_{a\in A^l_{\mathfrak{d}}(x)} \rho_A(a,f^*(x)) = \rho_A(\tilde{f}(x),f^*(x)) \leq \mathfrak{d}w(x), \tag{4.7}$$

for any $x \in X$.

From Remark 2.2, Assumption (A3) and (4.7), we obtain

$$\begin{aligned} g^* + h(x) &\geq -|c(x,f^*(x)) - c(x,\tilde{f}(x))| - |Qh(x,f^*(x)) - Qh(x,\tilde{f}(x))| \\ &\quad + c(x,\tilde{f}(x)) + Qh(x,\tilde{f}(x)) \\ &\geq -\big(L_c + L_q\mu(w)\bar{h}\big)\mathfrak{d}w(x) + c(x,\tilde{f}(x)) + Qh(x,\tilde{f}(x)), \end{aligned}$$

and so,

$$g^* + h(x) \geq -\big[(L_c + L_q\mu(w)\bar{h})\mathfrak{d} + \mathcal{C}_h W_1(\mu,\mu_n)\big]w(x) + c(x,\tilde{f}(x)) + Q_n h(x,\tilde{f}(x)),$$

by Proposition 4.1. Finally by using the same arguments as before for the stationary policy $\tilde{\pi}$ generated by $\tilde{f}$, we obtain that

$$g^* \geq J_{n,\mathfrak{d}}(x,\tilde{\pi}) - \big[(L_c + L_q\mu(w)\bar{h})\mathfrak{d} + \mathcal{C}_h W_1(\mu,\mu_n)\big]\frac{4b}{1-d}. \tag{4.8}$$

Combining Equations (4.5) and (4.8) and letting

$$\mathbf{G}_1 = \mathcal{C}_h\frac{4b}{1-d} \quad \text{and} \quad \mathbf{G}_2 = (L_c + L_q\mu(w)\bar{h})\frac{4b}{1-d}$$

we obtain the result. □

We note that we have bounds on the norm of h (recall Assumption C) and its Lipschitz constant (as a consequence of Lemma 2.3). Therefore, constants $\mathbf{G}_1$ and $\mathbf{G}_2$ defined above *do not depend on* $n \geq 1$ nor on $\mathfrak{d}$: they depend on constants related to the control model $\mathcal{M}$ that have been introduced in Assumptions A, B and C.

THEOREM 4.4. Suppose that Assumptions A, B, C and D are satisfied. There exists $\varepsilon_0 > 0$ such that for any $0 < \varepsilon \leq \varepsilon_0$ there exist $\mathfrak{d} > 0$ and constants $\mathcal{S}', \mathcal{T}' > 0$ such that

$$\mathbb{P}^*\left\{\left|J^*_{n,\mathfrak{d}}(x) - g^*\right| > \varepsilon\right\} \leq \mathcal{S}' \exp\{-\mathcal{T}' n\}$$

for all $n \geq 1$ and $x \in X$.

Proof. Define $\varepsilon_0 = 2\mathbf{G}_1\gamma_0$, where the constants $\mathbf{G}_1$ and γ_0 are taken from Theorems 4.3 and 3.1, respectively. Fix ε such that $0 < \varepsilon \leq \varepsilon_0$. We define

$$\mathfrak{c} = \mathfrak{c}_2 \wedge \frac{\varepsilon}{2\mathbf{G}_1} \quad \text{and} \quad \mathfrak{d} = \frac{\varepsilon}{2\mathbf{G}_2}.$$

Given arbitrary $n \geq 1$ and $\omega \in F_n(\mathfrak{c})$, $J^*_{n,\mathfrak{d}}(x)$ is the optimal average cost of the control model $\mathcal{M}_{n,\mathfrak{d}}$ for the initial state $x \in X$.

Since $\mathfrak{c} \leq \gamma_0$, we obtain from Theorem 3.1 that there are positive constants $\mathcal{S}'$ and $\mathcal{T}'$ such that

$$\mathbb{P}\{W_1(\mu, \mu_n) > \mathfrak{c}\} \leq \mathcal{S}' \exp\{-\mathcal{T}' n\}.$$

On the other hand, on the set $F_n(\mathfrak{c})$ we have, by Theorem 4.3,

$$\left|J^*_{n,\mathfrak{d}}(x) - g^*\right| \leq \mathbf{G}_1 W_1(\mu, \mu_n) + \mathbf{G}_2\mathfrak{d} \leq \varepsilon.$$

Therefore, $\{|J^*_{n,\mathfrak{d}}(x) - g^*| > \varepsilon\} \subseteq \{W_1(\mu, \mu_n) > \mathfrak{c}\}$. The stated result follows. □

Note that we do not take probability $\mathbb{P}$ of the set $\{|J^*_{n,\mathfrak{d}}(x) - g^*| > \varepsilon\}$ but, rather, its outer probability $\mathbb{P}^*$. This is because the issue of the measurability of $J^*_{n,\mathfrak{d}}(x)$ has not been addressed. We also note that Assumption (D3) has not been used yet. We will need it in our next section.

5. Approximation of an average cost optimal policy

In this section, we introduce another assumption on the control model.

Assumption E. Let $w : X \to [1, \infty)$ be the L_w-Lipschitz continuous function introduced in Assumption A. There exists $(x^*, a^*) \in X$ such that $w^* = Qw(x^*, a^*)$ is finite and, in addition, there is some $0 < d < 1$ such that

$$\int_X w(y)|Q(\mathrm{d}y|x, a) - Q(\mathrm{d}y|x', a')| \leq \mathrm{d}(w(x) + w(x')) \quad \text{for all } (x, a) \text{ and } (x', a') \text{ in } \mathbb{K},$$

where $|Q(\cdot|x, a) - Q(\cdot|x', a')|$ is the total variation of the signed kernel $Q(\cdot|x, a) - Q(\cdot|x', a')$.

We note that Assumption E implies Assumption A2. Indeed, given $(x,a) \in \mathbb{K}$ we have

$$Qw(x,a) \leq \int_X w(y)|Q(\mathrm{d}y|x,a) - Q(\mathrm{d}y|x^*,a^*)| + Qw(x^*,a^*) \leq dw(x) + (w^* + dw(x^*)).$$

Therefore, Assumption (A2) holds by letting $b := w^* + dw(x^*)$. That is why we have used the same constant $0 < d < 1$ in Assumptions (A2) and E. As we shall see, Assumption E is also a sufficient condition for Assumption C.

Our next lemma is a direct consequence of ([18], Theorem 7.3.14).

LEMMA 5.1. Under Assumption E, the control model $\mathcal{M}$ is uniformly w-geometrically ergodic on $\mathbb{F}$. This means that, for each $f \in \mathbb{F}$, the Markov chain on X with transition kernel given by $Q(\cdot|x,f(x))$ has a unique invariant probability measure μ_f on X with $\mu_f(w) < \infty$, and that

$$\sup_{f\in\mathbb{F}} \left|E^{f,x}[u(x_t)] - \mu_f(u)\right| \leq R\mathrm{d}^t \cdot \|u\|_w w(x) \quad \text{for all } t \in \mathbb{N},\ u \in \mathbb{B}_w(X), \text{ and } x \in X,$$

with $R = 1 + (b/(1-d))$.

The important feature of this result is that constants R and b are the same for *every* deterministic stationary policy. A standard calculation shows that, under uniform w-geometric ergodicity, we can obtain Assumption C by letting

$$\bar{h} = \frac{R\bar{c}}{1-d}(1 + w(x_0))$$

for any fixed $x_0 \in X$. Hence, in what follows we will not suppose Assumption C and, instead, we will use Assumption E.

PROPOSITION 5.2. Suppose that Assumptions A, B, D and E hold. Let $n \geq 1$, $\mathfrak{d} > 0$ and $\omega \in F_n(\mathfrak{c}_3)$, and define $\tilde{R} = 1 + 4(w(x^*) + b)/(1-d)$. Under these conditions, we have that the control model $\mathcal{M}_{n,\mathfrak{d}}$ is $(\tilde{R}, (1+d)/2)$-uniformly w-geometrically ergodic on $\Phi_{\mathfrak{d}}$. This means that, for each $\varphi \in \Phi_{\mathfrak{d}}$, the Markov chain on X with transition kernel given by $Q_n(\cdot|x,\varphi)$ has a unique invariant probability measure $\mu_\varphi^{n,\mathfrak{d}}$ on X with $\mu_\varphi^{n,\mathfrak{d}}(w) < \infty$, and that

$$\sup_{\varphi\in\Phi_{\mathfrak{d}}} |E^{\varphi,x}[u(x_t)] - \mu_\varphi^{n,\mathfrak{d}}(u)| \leq \tilde{R}\left(\frac{1+d}{2}\right)^t \cdot \|u\|_w w(x) \quad \text{for all } t \in \mathbb{N},\ u \in \mathbb{B}_w(X), \text{ and } x \in X.$$

Proof. Let us prove that the control model $\mathcal{M}_{n,\mathfrak{d}}$ satisfies Assumption E. Since $\omega \in F_n(\mathfrak{c}_2)$, from Proposition 3.6 we have $Q_n w(x,a) \leq ((1+d)/2)w(x) + 2b$ for all $(x,a) \in \mathbb{K}_{\mathfrak{d}}$, and so the first condition in Assumption E is satisfied for any $a^* \in A_{\mathfrak{d}}(x^*)$ by letting

$$\tilde{w}^* := \frac{1+d}{2} w(x^*) + 2b.$$

Concerning the second statement in Assumption E, given (x,a) and (x',a') in $\mathbb{K}_{\mathfrak{d}}$ we have

$$\int_X w(y)|Q_n(\mathrm{d}y|x,a) - Q_n(\mathrm{d}y|x',a')| = \int_X w(y)\left|\frac{q(y|x,a)}{\beta_n(x,a)} - \frac{q(y|x',a')}{\beta_n(x',a')}\right|\mu_n(\mathrm{d}y)$$

(this is because, from Definition 3.3, the kernel $Q_n(\cdot|x,a)$ has density $q(y|x,a)/\beta_n(x,a)$ with respect to μ_n). By Assumption (B2), the function

$$y \mapsto w(y)\left|\frac{q(y|x,a)}{\beta_n(x,a)} - \frac{q(y|x',a')}{\beta_n(x',a')}\right|$$

is Lipschitz-continuous on X with Lipschitz constant given by

$$L_{wq}\cdot\left(\frac{w(x)}{\beta_n(x,a)} + \frac{w(x')}{\beta_n(x',a')}\right) \leq 2L_{wq}\cdot(w(x)+w(x'))$$

because on $F_n(1/2L_q)$ we have $\beta_n(x,a) \geq 1/2$ and $\beta_n(x',a') \geq 1/2$. Therefore,

$$\begin{aligned}\int_X w(y)|Q_n(\mathrm{d}y|x,a) - Q_n(\mathrm{d}y|x',a')| &\leq 2L_{wq}\cdot(w(x)+w(x'))W_1(\mu,\mu_n)\\ &\quad + \int_X w(y)\left|\frac{q(y|x,a)}{\beta_n(x,a)} - \frac{q(y|x',a')}{\beta_n(x',a')}\right|\mu(\mathrm{d}y).\end{aligned}$$

Observe now that $w(y)|(q(y|x,a))/(\beta_n(x,a)) - (q(y|x',a'))/(\beta_n(x',a'))|$ is less than

$$\begin{aligned}&\frac{|w(y)q(y|x,a) - w(y)q(y|x',a')|}{\beta_n(x,a)} + \frac{w(y)q(y|x',a')}{\beta_n(x,a)\beta_n(x',a')}|\beta_n(x,a) - \beta_n(x',a')|\\ &\quad\leq \frac{|w(y)q(y|x,a) - w(y)q(y|x',a')|}{\beta_n(x,a)} + 8w(y)q(y|x',a')L_qW_1(\mu,\mu_n)\end{aligned}$$

by recalling Equation (3.4). Since $Qw(x',a') \leq dw(x') + b \leq (d+b)(w(x)+w(x'))$ we obtain that

$$\begin{aligned}&\int_X w(y)|Q_n(\mathrm{d}y|x,a) - Q_n(\mathrm{d}y|x',a')|\\ &\qquad\leq \left[\frac{d}{1-L_qW_1(\mu,\mu_n)} + 2W_1(\mu,\mu_n)[L_{wq} + 4L_q(d+b)]\right][w(x)+w(x')]\\ &\qquad\leq \big[d + 2W_1(\mu,\mu_n)(L_{wq} + L_q[1+4(d+b)])\big][w(x)+w(x')].\end{aligned}$$

Finally, we have established that for all (x,a) and (x',a') in $\mathbb{K}_{\mathfrak{d}}$

$$\int_X w(y)|Q_n(\mathrm{d}y|x,a) - Q_n(\mathrm{d}y|x',a')| \leq \tilde{d}_n(w(x)+w(x'))$$

where

$$\tilde{d}_n = d + 2W_1(\mu,\mu_n)\big[L_{wq} + L_q[1+4(d+b)]\big].$$

We note that, on $F_n(\mathfrak{c}_3)$, we have $\tilde{d}_n \leq (1+d)/2$.

Summarizing, the control model $\mathcal{M}_{n,\mathfrak{d}}$ satisfies

$$Q_nw(x^*,a^*) \leq \frac{1+d}{2}w(x^*) + 2b < \infty \quad \text{for some } (x^*,a^*) \in \mathbb{K}_{\mathfrak{d}}$$

and, for all (x, a) and (x', a') in $\mathbb{K}_{\mathfrak{d}}$,

$$\int_X w(y)|Q_n(\mathrm{d}y|x,a) - Q_n(\mathrm{d}y|x',a')| \leq \left(\frac{1+d}{2}\right)\cdot(w(x)+w(x')). \tag{5.1}$$

It follows from Lemma 5.1 that the control model $\mathcal{M}_{n,\mathfrak{d}}$ is $(\tilde{R}, (1+d)/2)$-uniformly w-geometrically ergodic on $\mathbb{F}_{\mathfrak{d}}$. It remains to show that $\mathcal{M}_{n,\mathfrak{d}}$ is $(\tilde{R}, (1+d)/2)$-uniformly w-geometrically ergodic on $\Phi_{\mathfrak{d}}$. To see this, fix $\varphi \in \Phi_{\mathfrak{d}}$ and note that $Q_n(\cdot|x,\varphi)$ has density

$$y \mapsto \int_{A_{\mathfrak{d}}(x)} \frac{q(y|x,a)}{\beta_n(x,a)}\,\varphi(\mathrm{d}a|x)$$

with respect to μ_n. Therefore

$$\begin{aligned}
&\int_X w(y)|Q_n(\mathrm{d}y|x,\varphi) - Q_n(\mathrm{d}y|x',\varphi)|\mu_n(\mathrm{d}y)\\
&\quad= \int_X w(y)\left|\int_{A_{\mathfrak{d}}(x)\times A_{\mathfrak{d}}(x')}\left[\frac{q(y|x,a)}{\beta_n(x,a)} - \frac{q(y|x',a')}{\beta_n(x',a')}\right]\varphi(\mathrm{d}a|x)\times\varphi(\mathrm{d}a'|x')\right|\mu_n(\mathrm{d}y)\\
&\quad\leq \int_{A_{\mathfrak{d}}(x)\times A_{\mathfrak{d}}(x')}\int_X w(y)\left|\frac{q(y|x,a)}{\beta_n(x,a)} - \frac{q(y|x',a')}{\beta_n(x',a')}\right|\mu_n(\mathrm{d}y)\varphi(\mathrm{d}a|x)\times\varphi(\mathrm{d}a'|x')\\
&\quad\leq \left(\frac{1+d}{2}\right)\cdot(w(x)+w(x')),
\end{aligned}$$

by (5.1). Recalling ([18], Theorem 7.3.14), the result follows. □

Let us now comment on Assumption E. Assumptions A, B and C on the control model $\mathcal{M}$ ensure that the corresponding ACOE has a solution (g^*, h). Besides, we have obtained explicit bounds on the norm of h. To approximate an optimal policy, we need a solution to the ACOE for control model $\mathcal{M}_{n,\mathfrak{d}}$. In general, imposing Assumption C on $\mathcal{M}$ does not imply that a similar condition is satisfied by $\mathcal{M}_{n,\mathfrak{d}}$. On the other hand, as seen in Proposition 5.2, Assumption E for $\mathcal{M}$ *does imply* the same condition for $\mathcal{M}_{n,\mathfrak{d}}$. Moreover, it has the advantage that we have explicit values for the constants involved in the uniform w-geometric ergodicity. This will allow to establish the ACOE for $\mathcal{M}_{n,\mathfrak{d}}$ and to obtain bounds on the norms of the involved functions.

Concerning constants $\mathfrak{c}_i$ introduced in Definition 3.2, their interpretation is now clear. On $F_n(\mathfrak{c}_1)$ the control model $\mathcal{M}_{n,\mathfrak{d}}$ is well defined; on $F_n(\mathfrak{c}_2)$ the optimal average cost function $J^*_{n,\mathfrak{d}}$ is finite; on $F_n(\mathfrak{c}_3)$ the control model $\mathcal{M}_{n,\mathfrak{d}}$ is uniformly w-geometrically ergodic.

We recall the definition of the ACOE. Given $n \geq 1$, $\mathfrak{d} > 0$ and $\omega \in F_n(\mathfrak{c}_1)$, we say that the pair $(g, h) \in \mathbb{R}\times\mathbb{B}_w(X)$ is a solution to ACOE for the control model $\mathcal{M}_{n,\mathfrak{d}}$ if

$$g + h(x) = \min_{a\in A_{\mathfrak{d}}(x)}\left\{c(x,a) + \int_X h(y)Q_n(\mathrm{d}y|x,a)\right\} \quad \text{for all } x \in X. \tag{5.2}$$

We say that $f \in \mathbb{F}_{\mathfrak{d}}$ is an $\mathcal{M}_{n,\mathfrak{d}}$-canonical policy if it attains the minimum in (5.2), that is,

$$g + h(x) = c(x, f(x)) + \int_X h(y) Q_n(\mathrm{d}y|x, f(x)) \quad \text{for all } x \in X.$$

THEOREM 5.3. Suppose that Assumptions A, B, D and E are satisfied. Given $n \geq 1$ and $\mathfrak{d} > 0$, consider $\omega \in F_n(\mathfrak{c}_3)$.

(i) There exists a solution $(g^*_{n,\mathfrak{d}}, h_{n,\mathfrak{d}}) \in \mathbb{R} \times \mathbb{B}_w(X)$ to the ACOE for the control model $\mathcal{M}_{n,\mathfrak{d}}$. This solution satisfies

$$g^*_{n,\mathfrak{d}} = J^*_{n,\mathfrak{d}}(x) \quad \text{for all } x \in X, \quad \text{and} \quad \|h_{n,\mathfrak{d}}\|_w \leq H,$$

where

$$H = \frac{2\tilde{R}\bar{c}}{1-d}(1 + w(x_0)).$$

(ii) If (g', h') is any other solution to the ACOE for $\mathcal{M}_{n,\mathfrak{d}}$ then $g' = g^*_{n,\mathfrak{d}}$ and functions h' and $h_{n,\mathfrak{d}}$ differ on X by a constant.

(iii) The set of $\mathcal{M}_{n,\mathfrak{d}}$-canonical policies is non-empty and it does not depend on the particular solution $h_{n,\mathfrak{d}}$ of the ACOE. Moreover, any $\mathcal{M}_{n,\mathfrak{d}}$-canonical policy is average optimal for $\mathcal{M}_{n,\mathfrak{d}}$.

Proof.

(i) According to the results in Section 3, Assumptions 3.1 and 3.2 in [15] are satisfied. Therefore, for every $0 < \alpha < 1$, from ([15], Lemma 3.2), the discounted value function $V^*_{\alpha,n,\mathfrak{d}}$ belongs to $\mathbb{B}_w(X)$ and, besides, it verifies

$$\left|V^*_{\alpha,n,\mathfrak{d}}(x)\right| \leq \frac{2\bar{c}}{1-d}\left[w(x) + \frac{2b}{1-\alpha}\right] \tag{5.3}$$

and satisfies the discounted optimality equation

$$V^*_{\alpha,n,\mathfrak{d}}(x) = \min_{a \in A_{\mathfrak{d}}(x)} \left\{ c(x,a) + \alpha \int_X V^*_{\alpha,n,\mathfrak{d}}(y) Q_n(\mathrm{d}y|x,a) \right\}. \tag{5.4}$$

Now, consider $x_0 \in X$ and define the function $h_{\alpha,n,\mathfrak{d}}$ as $h_{\alpha,n,\mathfrak{d}}(x) = V^*_{\alpha,n,\mathfrak{d}}(x) - V^*_{\alpha,n,\mathfrak{d}}(x_0)$ for $x \in X$. If $f^* \in \mathbb{F}_{\mathfrak{d}}$ attains the minimum in DCOE (5.4), then

$$\begin{aligned} |h_{\alpha,n,\mathfrak{d}}(x)| &\leq \sum_{t=0}^{\infty} \alpha^t \left| E^{f^*,x}_{n,\mathfrak{d}}[c(x_t, f^*(x_t))] - E^{f^*,x_0}_{n,\mathfrak{d}}[c(x_t, f^*(x_t))] \right| \\ &\leq \sum_{t=0}^{\infty} \alpha^t \left(\frac{1+d}{2}\right)^t \tilde{R}\bar{c}[w(x) + w(x_0)] \end{aligned}$$

by using Proposition 5.2 and so, for every $0 < \alpha < 1$

$$\|h_{\alpha,n,\mathfrak{b}}\|_w \leq H. \tag{5.5}$$

Introduce $g_{\alpha,n,\mathfrak{b}} = (1-\alpha)V^*_{\alpha,n,\mathfrak{b}}(x_0)$. The discounted optimality equation can be re-written as

$$g_{\alpha,n,\mathfrak{b}} + h_{\alpha,n,\mathfrak{b}}(x) = \min_{a\in A_{\mathfrak{b}}(x)}\left\{c(x,a) + \alpha\int_X h_{\alpha,n,\mathfrak{b}}(y)Q_n(\mathrm{d}y|x,a)\right\}.$$

We observe that, as a consequence of (5.3), the sequence $\{g_{\alpha,n,\mathfrak{b}}\}$ is bounded when $0 < \alpha < 1$. By using similar arguments as in Lemma 2.3 and Proposition 3.5, it follows that

$$|h_{\alpha,n,\mathfrak{b}}(x) - h_{\alpha,n,\mathfrak{b}}(x')| \leq \left[L_c + 2HL_q\left[\mu(w) + \frac{L_w}{2L_q}\right]\left[1 + 2\bar{q}w(x)\right]\right](1 + L^*_\Psi)\rho_X(x,x'). \tag{5.6}$$

From the previous equation, it follows that the family of functions $\{h_{\alpha,n,\mathfrak{b}} : \alpha \in (0,1)\}$ is equicontinuous. Now, by using inequality (5.5) and Ascoli's theorem, we obtain that there exist a sequence $\{\alpha_k\}$, a constant $g^*_{n,\mathfrak{b}} \in \mathbb{R}$ and a function $h_{n,\mathfrak{b}} \in \mathbb{B}_w(X)$ such that

$$\alpha_k \to 1, \quad g_{\alpha_k,n,\mathfrak{b}} \to g^*_{n,\mathfrak{b}} \quad \text{and} \quad h_{\alpha_k,n,\mathfrak{b}}(x) \to h_{n,\mathfrak{b}}(x) \quad \text{for any } x \in X$$

as $k \to \infty$. Clearly, from (5.5) we have $\|h_{n,\mathfrak{b}}\|_w \leq H$, while (5.6) implies that $h_{n,\mathfrak{b}}$ is locally Lipschitz continuous on X.

By using standard arguments (in particular, we make use of the extended Fatou lemma in ([18], Lemma 8.3.7)), we obtain that

$$g^*_{n,\mathfrak{b}} + h_{n,\mathfrak{b}}(x) = \min_{a\in A_{\mathfrak{b}}(x)}\left\{c(x,a) + \int_X h_{n,\mathfrak{b}}(y)Q_n(\mathrm{d}y|x,a)\right\}$$

for all $x \in X$. The fact that $g^*_{n,\mathfrak{b}} = J^*_n(x)$ for $x \in X$ follows in a straightforward way.

(ii) Let (g_1,h_1) and (g_2,h_2) in $\mathbb{R}\times\mathbb{B}_w(X)$ be two solutions to the ACOE for the control model $\mathcal{M}_{n,\mathfrak{b}}$. The fact that $g_1 = g_2 = g^*_{n,\mathfrak{b}}$ follows from standard arguments, such as those used in the proof of item (i) above.

Hence, it remains to show that h_1 and h_2 differ by a constant, i.e. there exists $\gamma \in \mathbb{R}$ such that $h_1(x) - h_2(x) = \gamma$ for all $x \in X$. Let the policies f_1 and f_2 in $\mathbb{F}_{\mathfrak{b}}$ attain the minimum in the ACOE for h_1 and h_2, respectively; that is, we have for all $x \in X$

$$g^*_{n,\mathfrak{b}} + h_1(x) = c(x,f_1) + \int_X h_1(y)Q_n(\mathrm{d}y|x,f_1) \tag{5.7}$$

$$\leq c(x,f_2) + \int_X h_1(y)Q_n(\mathrm{d}y|x,f_2), \tag{5.8}$$

$$g^*_{n,\mathfrak{d}} + h_2(x) = c(x,f_2) + \int_X h_2(y) Q_n(\mathrm{d}y|x,f_2) \tag{5.9}$$

$$\leq c(x,f_1) + \int_X h_2(y) Q_n(\mathrm{d}y|x,f_1), \tag{5.10}$$

with $c(x,f_1(x)) = c(x,f_1)$ and $Q_n(\mathrm{d}y|x,f_1(x)) = Q_n(\mathrm{d}y|x,f_1)$. Consider now the transition matrices of the policies f_i, for $i = 1,2$, when restricted to the set of states $\Gamma_n = \{Y_k(\omega)\}_{1\leq k\leq n}$. The corresponding transition matrix will be denoted by P_i, with $P_i(x,y) = Q_n(\{y\}|x,f_i(x))$ for $x,y \in \Gamma_n$ (recall Definition 3.3). By Proposition 5.2, the transition matrix P_i has a unique invariant probability measure. Therefore, Γ_n can be partitioned into an irreducible class of positive recurrent states R_i and a (possibly empty) class of transient states T_i. So, let us write the transition matrix P_i in block form

$$P_i = \begin{pmatrix} P_i(R_i,R_i) & \mathbf{0} \\ P_i(T_i,R_i) & P_i(T_i,T_i) \end{pmatrix},$$

where we highlight the transitions between the classes of recurrent and transient states. It is a well known fact that

$$(\mathbf{I} - P_i(T_i,T_i))^{-1} = \mathbf{I} + \sum_{k=1}^{\infty} (P_i(T_i,T_i))^k \quad \text{and} \quad (\mathbf{I} - P_i(T_i,T_i))^{-1} P_i(T_i,R_i)\mathbf{1} = \mathbf{1}. \tag{5.11}$$

By (5.7) and (5.10) we have that function $h_1 - h_2$ on Γ_n (that will be interpreted as a column vector) is subharmonic for P_1, meaning that $h_1 - h_2 \geq P_1(h_1 - h_2)$. As a consequence, for each $x \in \Gamma_n$ and $t \geq 0$

$$E^{f_1,x}_{n,\mathfrak{d}}[(h_1 - h_2)(x_t)] \leq h_1(x) - h_2(x).$$

By Proposition 5.2, this implies that $\mu^{n,\mathfrak{d}}_{f_1}(h_1 - h_2) \leq h_1(x) - h_2(x)$ for all $x \in \Gamma_n$, with $\mu^{n,\mathfrak{d}}_{f_1}$ being the invariant probability measure associated with the policy f_1. Consequently, $h_1 - h_2$ is constant on R_1. Similarly, using (5.8) and (5.9) we obtain that $h_1 - h_2$ is superharmonic for P_2, and it follows that $h_1 - h_2$ is constant on R_2. Summarizing, there exist γ_1 and $\gamma_2 \in \mathbb{R}$ such that

$$h_1(x) - h_2(x) = \gamma_1 \ \text{ for all } \ x \in R_1 \ \text{ and } \ h_1(x) - h_2(x) = \gamma_2 \ \text{ for all } \ x \in R_2. \tag{5.12}$$

Now, by (5.8), for every $x \in T_2$,

$$h_1(x) - \sum_{y\in T_2} P_2(x,y)h_1(y) \leq -g^*_{n,d} + c(x,f_2) + \sum_{y\in R_2} P_2(x,y)h_1(y)$$

or, in matrix notation and denoting, generically, by $u(S)$ the column vector u restricted to the components of $S \subseteq \Gamma_n$,

$$(\mathbf{I} - P_2(T_2,T_2))h_1(T_2) \leq -g^*_{n,\mathfrak{d}}\mathbf{1} + c(T_2,f_2) + P_2(T_2,R_2)h_1(R_2). \tag{5.13}$$

Also, from (5.9),

$$(\mathbf{I} - P_2(T_2, T_2))h_2(T_2) = -g^*_{n,\mathfrak{d}}\mathbf{1} + c(T_2, f_2) + P_2(T_2, R_2)h_2(R_2). \tag{5.14}$$

Since the matrix $(\boldsymbol{I} - P_2(T_2, T_2))^{-1}$ is non-negative (recall (5.11)), it follows from (5.13) that

$$h_1(T_2) \le (\mathbf{I} - P_2(T_2, T_2))^{-1} \cdot \Big(-g^*_{n,\mathfrak{d}}\mathbf{1} + c(T_2, f_2) + P_2(T_2, R_2)h_1(R_2) \Big).$$

However, we have $h_1(R_2) = h_2(R_2) + \gamma_2\mathbf{1}$, and so

$$\begin{aligned} h_1(T_2) &\le (\mathbf{I} - P_2(T_2, T_2))^{-1} \cdot \Big(-g^*_{n,\mathfrak{d}}\mathbf{1} + c(T_2, f_2) + P_2(T_2, R_2)(h_2(R_2) + \gamma_2\mathbf{1}) \Big) \\ &= h_2(T_2) + \gamma_2\mathbf{1}, \end{aligned}$$

where we have applied (5.11) and (5.14). By following a similar argument, we obtain that $h_1(T_1) - h_2(T_1) \ge \gamma_1\mathbf{1}$. So far, we have established that

$$\begin{cases} h_1(x) - h_2(x) = \gamma_1 & \text{on } R_1, \\ h_1(x) - h_2(x) \ge \gamma_1 & \text{on } T_1, \end{cases} \quad \text{and} \quad \begin{cases} h_1(x) - h_2(x) = \gamma_2 & \text{on } R_2, \\ h_1(x) - h_2(x) \le \gamma_2 & \text{on } T_2. \end{cases}$$

In particular, $\gamma_1 \le h_1(x) - h_2(x) \le \gamma_2$ for all $x \in \Gamma_n$.

Suppose, for the moment, that $\gamma_1 < \gamma_2$. Then we necessarily have $R_1 \cap R_2 = \emptyset$. Define the policy $\bar{f} \in \mathbb{F}_{\mathfrak{d}}$ as follows: $\bar{f}(x) = f_1(x)$ if $x \in R_1$, and $\bar{f}(x) = f_2(x)$ if $x \in R_2$ (the definition outside $R_1 \cup R_2$ is not relevant). We have that R_1 and R_2 are two disjoint recurrent classes for $\bar{f}$. In this case, it is not possible that $\{Q_n(\cdot|x, \bar{f})\}$ has a unique invariant probability measure (recall Proposition 5.2). This leads to a contradiction, and so $\gamma_1 = \gamma_2 =: \gamma$.

Once we know that $h_1(x) = h_2(x) + \gamma$ for all $x \in \Gamma_n$, the equality is extended to $X - \Gamma_n$ by noting that $h_i(x)$ for $x \in X - \Gamma_n$ is uniquely determined by the values of h_i in Γ_n; namely,

$$g^*_{n,\mathfrak{d}} + h_i(x) = \min_{a \in A_{\mathfrak{d}}(x)} \left\{ c(x, a) + \sum_{y \in \Gamma_n} h_i(y) Q_n(\{y\}|x, a) \right\},$$

and $h_1(x) - h_2(x) = \gamma$ for $x \in X - \Gamma_n$ follows.

(iii) This statement easily follows from (ii) and standard arguments. □

Under the hypotheses of Theorem 5.3, we note that the optimal average cost $g^*_{n,\mathfrak{d}}$ of $\mathcal{M}_{n,\mathfrak{d}}$ is such that $|g^*_{n,\mathfrak{d}}| \le 4\bar{c}b/(1-d)$. We also note that the solution $h_{n,\mathfrak{d}}$ to the ACOE that has been constructed in the proof of Theorem 5.3 is in fact the unique solution to the ACOE such that $h_{n,\mathfrak{d}}(x_0) = 0$.

LEMMA 5.4. For $n \geq 1$, $\mathfrak{d} > 0$ and $\omega \in F_n(\mathfrak{c}_3)$, let $h_{n,\mathfrak{d}} \in \mathbb{B}_w(X)$ be the solution in the ACOE for the control model $\mathcal{M}_{n,\mathfrak{d}}$ constructed in Theorem 5.3. Define for $x \in X$

$$\tilde{h}_{n,\mathfrak{d}}(x) = \min_{a \in A_{\mathfrak{d}}(x)} \{c(x,a) + \beta_n(x,a) Q_n h_{n,\mathfrak{d}}(x,a)\} = \min_{a \in A_{\mathfrak{d}}(x)} \left\{ c(x,a) + \int_X h_{n,\mathfrak{d}}(y) q(y|x,a) \mu_n(\mathrm{d}y) \right\}.$$

Function $\tilde{h}_{n,\mathfrak{d}}$ is in $\mathbb{L}_w(X)$, it verifies

$$\|g^*_{n,\mathfrak{d}} + h_{n,\mathfrak{d}} - \tilde{h}_{n,\mathfrak{d}}\|_w \leq H L_q \left(\frac{1+d}{2} + 2b \right) W_1(\mu, \mu_n) \quad \text{and}$$
$$\|\tilde{h}_{n,\mathfrak{d}}\|_w \leq H \left(\frac{5+d}{4} + b \right) + \frac{4\bar{c}b}{1-d},$$

and its Lipschitz constant is

$$\left(1 + L^*_\Psi\right) \left(L_c + H L_q \left(\mu(w) + \frac{L_w}{2L_q} \right) \right).$$

Proof. It is clear from our continuity hypotheses that $\tilde{h}_{n,\mathfrak{d}}$ is measurable (see, e.g. ([16], Proposition D.5.(b))), and so also $\tilde{h}_{n,\mathfrak{d}} \in \mathbb{B}_w(X)$. In addition, by the ACOE in Theorem 5.3, for all $x \in X$

$$g^*_{n,\mathfrak{d}} + h_{n,\mathfrak{d}}(x) = \min_{a \in A_{\mathfrak{d}}(x)} \left\{ c(x,a) + \int_X h_{n,\mathfrak{d}}(y) Q_n(\mathrm{d}y|x,a) \right\}$$
$$= \min_{a \in A_{\mathfrak{d}}(x)} \left\{ c(x,a) + \int_X h_{n,\mathfrak{d}}(y) q(y|x,a) \mu_n(\mathrm{d}y) + (1 - \beta_n(x,a)) \int_X h_{n,\mathfrak{d}}(y) Q_n(\mathrm{d}y|x,a) \right\}.$$

Since we have

$$\left| (1 - \beta_n(x,a)) \int_X h_{n,\mathfrak{d}}(y) Q_n(\mathrm{d}y|x,a) \right| \leq H L_q \left(\frac{1+d}{2} + 2b \right) W_1(\mu, \mu_n) w(x),$$

it follows by using Proposition 3.6 that

$$\left\| g^*_{n,\mathfrak{d}} + h_{n,\mathfrak{d}} - \tilde{h}_{n,\mathfrak{d}} \right\|_w \leq H L_q \left(\frac{1+d}{2} + 2b \right) W_1(\mu, \mu_n),$$

where $g^*_{n,\mathfrak{d}}$ is interpreted as a constant function on X. The bound on $\|\tilde{h}_{n,\mathfrak{d}}\|_w$ easily follows from the previous inequality. Indeed, we have

$$\|\tilde{h}_{n,\mathfrak{d}}\|_w \leq H L_q \left(\frac{1+d}{2} + 2b \right) W_1(\mu, \mu_n) + \|g^*_{n,\mathfrak{d}}\|_w + \|h_{n,\mathfrak{d}}\|_w.$$

Now, we obtain the desired bound by recalling that $|g^*_{n,\mathfrak{d}}| \leq 4\bar{c}b/(1-d)$ and $\|h_{n,\mathfrak{d}}\|_w \leq H$ (see item (i) of Theorem 5.3) and by observing that for $\omega \in F_n(\mathfrak{c}_3) \subseteq F_n(\mathfrak{c}_1)$ we have $W_1(\mu, \mu_n) \leq 1/2L_q$ (see Definition 3.2).

For notational convenience, let us introduce

$$\Delta_n(x,a,x',a') := |c(x,a) - c(x',a')| + |\beta_n(x,a)Q_n h_{n,\mathfrak{d}}(x,a) - \beta_n(x',a')Q_n h_{n,\mathfrak{d}}(x',a')|$$

for (x,a) and (x',a') in $\mathbb{K}_{\mathfrak{d}}$ (cf. the definition of $\Delta_\alpha(x,a,x',a')$ in the proof of Lemma 2.3). Then, proceeding as in the proof of Lemma 2.3 we have

$$\tilde{h}_{n,\mathfrak{d}}(x) \le \tilde{h}_{n,\mathfrak{d}}(x') + \sup_{a' \in A_{\mathfrak{d}}(x')} \inf_{a \in A_{\mathfrak{d}}(x)} \{\Delta_n(x,a,x',a')\},$$

and, symmetrically,

$$\tilde{h}_{n,\mathfrak{d}}(x') \le \tilde{h}_{n,\mathfrak{d}}(x) + \sup_{a \in A_{\mathfrak{d}}(x)} \inf_{a' \in A_{\mathfrak{d}}(x')} \{\Delta_n(x,a,x',a')\}.$$

So, for all x and x' in X

$$|\tilde{h}_{n,\mathfrak{d}}(x) - \tilde{h}_{n,\mathfrak{d}}(x')| \le \sup_{a \in A_{\mathfrak{d}}(x)} \inf_{a' \in A_{\mathfrak{d}}(x')} \{\Delta_n(x,a,x',a')\} \vee \sup_{a' \in A_{\mathfrak{d}}(x')} \inf_{a \in A_{\mathfrak{d}}(x)} \{\Delta_n(x,a,x',a')\}.$$

Now, from Assumption (A3), item (i) of Theorem 5.3, (2.4) and (3.6) we easily obtain that

$$\Delta_n(x,a,x',a') \le \left[L_c + HL_q\left(\mu(w) + \frac{L_w}{2L_q}\right)\right][\rho_X(x,x') + \rho_A(a,a')]. \tag{5.15}$$

With Assumption (D3), this yields the Lipschitz constant of $\tilde{h}_{n,\mathfrak{d}}$. □

This means that function $h_{n,\mathfrak{d}}$ in the average optimality equation for $\mathcal{M}_{n,\mathfrak{d}}$ is not necessarily Lipschitz continuous (in fact, it is locally Lipschitz continuous), but it can be approximated by a Lipschitz continuous function $\tilde{h}_{n,\mathfrak{d}}$, with an approximation error which is controlled in the w-norm by the Wasserstein distance $W_1(\mu_n,\mu)$. We note that in Proposition 2.4 we could derive that h in the average optimality equation for $\mathcal{M}$ is Lipschitz continuous.

THEOREM 5.5. Suppose that Assumptions A, B, D and E are satisfied. There exist constants $\mathbf{H}_1$ and $\mathbf{H}_2$ such that given $n \ge 1$, $\mathfrak{d} > 0$ and $\omega \in F_n(\mathfrak{c}_3)$, any $\mathcal{M}_{n,\mathfrak{d}}$-canonical policy $\tilde{f}_{n,\mathfrak{d}} \in \mathbb{F}_{\mathfrak{d}} \subseteq \mathbb{F}$ verifies

$$J(\tilde{f}_{n,\mathfrak{d}},x) \le g^* + \mathbf{H}_1 W_1(\mu,\mu_n) + \mathfrak{d}\mathbf{H}_2 \quad \text{for } x \in X.$$

Proof. Let $h_{n,\mathfrak{d}}$ be the solution of the ACOE for $\mathcal{M}_{n,\mathfrak{d}}$ constructed in Theorem 5.3, and let $\tilde{f}_{n,\mathfrak{d}}$ attain the minimum in the corresponding ACOE, that is,

$$g^*_{n,\mathfrak{d}} + h_{n,\mathfrak{d}}(x) = c(x,\tilde{f}_{n,\mathfrak{d}}(x)) + Q_n h_{n,\mathfrak{d}}(x,\tilde{f}_{n,\mathfrak{d}}(x)) \quad \text{for } x \in X. \tag{5.16}$$

Our first step in this proof is to 'replace' $h_{n,\mathfrak{d}}$ with $\tilde{h}_{n,\mathfrak{d}}$ in (5.16). To this end, regarding the left-hand side of (5.16), note that by Lemma 5.4,

$$g^*_{n,\mathfrak{d}} + h_{n,\mathfrak{d}}(x) \le \tilde{h}_{n,\mathfrak{d}}(x) + HL_q\left(\frac{1+d}{2} + 2b\right) W_1(\mu,\mu_n) w(x) \quad \text{for each } x \in X,$$

while, for the right-hand side,

$$h_{n,\mathfrak{d}}(y) \geq \tilde{h}_{n,\mathfrak{d}}(y) - g^*_{n,\mathfrak{d}} - HL_q\left(\frac{1+d}{2}+2b\right)W_1(\mu,\mu_n)w(y) \quad \text{for each } y \in X,$$

and so

$$Q_n h_{n,\mathfrak{d}}(x,a) \geq Q_n\tilde{h}_{n,\mathfrak{d}}(x,a) - g^*_{n,\mathfrak{d}} - HL_q\left(\frac{1+d}{2}+2b\right)W_1(\mu,\mu_n)Q_n w(x,a)$$
$$\text{for each } (x,a) \in \mathbb{K}_{\mathfrak{d}},$$

whence, by Proposition 3.6,

$$Q_n h_{n,\mathfrak{d}}(x,\tilde{f}_{n,\mathfrak{d}}(x)) \geq Q_n\tilde{h}_{n,\mathfrak{d}}(x,\tilde{f}_{n,\mathfrak{d}}(x)) - g^*_{n,\mathfrak{d}} - HL_q\left(\frac{1+d}{2}+2b\right)^2 W_1(\mu,\mu_n)w(x)$$
$$\text{for each } x \in X.$$

So far, we have established that

$$HL_q\left(\frac{1+d}{2}+2b\right)\left(\frac{3+d}{2}+2b\right)W_1(\mu,\mu_n)w(x) + g^*_{n,\mathfrak{d}} + \tilde{h}_{n,\mathfrak{d}}(x)$$
$$\geq c(x,\tilde{f}_{n,\mathfrak{d}}(x)) + Q_n\tilde{h}_{n,\mathfrak{d}}(x,\tilde{f}_n(x))$$

for all $x \in X$. Now we use the fact that $\tilde{h}_{n,\mathfrak{d}} \in \mathbb{L}_w(X)$ – Lemma 5.4 – and Proposition 4.1 to derive

$$\left(\mathcal{C}_{\tilde{h}_{n,\mathfrak{d}}} + HL_q\left(\frac{1+d}{2}+2b\right)\left(\frac{3+d}{2}+2b\right)\right)W_1(\mu,\mu_n)w(x) + g^*_{n,\mathfrak{d}} + \tilde{h}_{n,\mathfrak{d}}(x)$$
$$\geq c(x,\tilde{f}_{n,\mathfrak{d}}(x)) + Q\tilde{h}_{n,\mathfrak{d}}(x,\tilde{f}_{n,\mathfrak{d}}(x))$$

for all $x \in X$. Taking into account the bounds on $\|\tilde{h}_{n,\mathfrak{d}}\|_w$ and its Lipschitz constant in Lemma 5.4, we obtain that there exists a constant $\mathbf{G}$ that depends only on the constants in Assumptions A, B, E and L^*_Ψ in (D3) (and not on n nor on $\mathfrak{d}$) such that

$$\mathbf{G}w(x)W_1(\mu,\mu_n) + g^*_{n,\mathfrak{d}} + \tilde{h}_{n,\mathfrak{d}}(x) \geq c(x,\tilde{f}_{n,\mathfrak{d}}(x)) + Q\tilde{h}_{n,\mathfrak{d}}(x,\tilde{f}_{n,\mathfrak{d}}(x))$$

for all $x \in X$. Iterating this inequality and using (by now) standard arguments, we derive

$$\mathbf{G}\frac{4b}{1-d}W_1(\mu,\mu_n) + g^*_{n,\mathfrak{d}} \geq J(\tilde{f}_{n,\mathfrak{d}},x) \quad \text{for } x \in X.$$

By Theorem 4.3, we conclude that there are constants $\mathbf{H}_i$ depending on the parameters in our assumptions (and not on n nor on $\mathfrak{d}$) such that

$$g^* \leq J(\tilde{f}_{n,\mathfrak{d}},x) \leq g^* + \mathbf{H}_1 W_1(\mu,\mu_n) + \mathfrak{d}\mathbf{H}_2 \quad \text{for } x \in X.$$

This completes the proof. □

We have that $J(\tilde{f}_{n,\mathfrak{d}},x)$ is defined for $\omega \in F_n(\mathfrak{c}_3)$. We extend this definition as we did with $J^*_{n,\mathfrak{d}}(x)$ in the previous section. If the control model $\mathcal{M}_{n,\mathfrak{d}}$ is well defined and if a

solution to the corresponding ACOE exists, let $J(\tilde{f}_{n,\mathfrak{d}}, x)$ be the average cost (for the control model $\mathcal{M}$) of any canonical policy (in particular, these statements hold on $F_n(\mathfrak{c}_3)$). Otherwise, define $J(\tilde{f}_{n,\mathfrak{d}}, x)$ arbitrarily.

THEOREM 5.6. Suppose that Assumptions A, B, D and E are satisfied. There exists $\varepsilon_0 > 0$ such that for any $0 < \varepsilon \leq \varepsilon_0$ there exist $\mathfrak{d} > 0$ and constants $\mathcal{S}'', \mathcal{T}'' > 0$ such that

$$\mathbb{P}^*\{J(\tilde{f}_{n,\mathfrak{d}}, x) - g^* > \varepsilon\} \leq \mathcal{S}'' \exp\{-\mathcal{T}'' n\}$$

for all $n \geq 1$ and $x \in X$.

Proof. The proof is similar to that of Theorem 4.4. Just define $\varepsilon_0 = 2\mathbf{H}_1\gamma_0$, where the constants $\mathbf{H}_1$ and γ_0 are taken from Theorems 5.5 and 3.1. Fix ε such that $0 < \varepsilon \leq \varepsilon_0$, and let

$$\mathfrak{c} = \mathfrak{c}_3 \wedge \frac{\varepsilon}{2\mathbf{H}_1} \quad \text{and} \quad \mathfrak{d} = \frac{\varepsilon}{2\mathbf{H}_2}.$$

Note now that the set $\{J(\tilde{f}_{n,\mathfrak{d}}, x) > g^* + \varepsilon\}$ is contained in $\{W_1(\mu, \mu_n) > c\}$ and proceed as in the proof of Theorem 4.4. □

6. Numerical approximations

Suppose that Assumptions A, B, D and E are satisfied. Given $n \geq 1$ and $\mathfrak{d} > 0$, consider $\omega \in F_n(\mathfrak{c}_3)$. Moreover, we will suppose that the sets $A_{\mathfrak{d}}(x)$ are finite for every $x \in X$. We recall the notation $\Gamma_n = \{Y_k(\omega)\}_{1 \leq k \leq n}$. In what follows, we consider the control model $\mathcal{M}_{n,\mathfrak{d}}$ restricted to the states in Γ_n. We can do so because we have $Q_n(\Gamma_n | x, a) = 1$ for all $x \in \Gamma_n$ and $a \in A_{\mathfrak{d}}(x)$. Therefore, in what follows we will be dealing with a unichain, finite state and action average cost MDP.

LP formulation of the average cost problem. Consider the following (primal) LP problem (**P**)

$$\text{minimize} \quad \sum_{x \in \Gamma_n} \sum_{a \in A_{\mathfrak{d}}(x)} c(x, a) z(x, a)$$

subject to

$$\sum_{a \in A_{\mathfrak{d}}(x)} z(x, a) = \sum_{x' \in \Gamma_n} \sum_{a' \in A_{\mathfrak{d}}(x')} z(x', a') Q_n(\{x\} | x', a') \quad \text{for all } x \in \Gamma_n,$$

$$\sum_{x \in \Gamma_n} \sum_{a \in A_{\mathfrak{d}}(x)} z(x, a) = 1 \quad \text{and} \quad z(x, a) \geq 0 \quad \text{for all } x \in \Gamma_n \text{ and } a \in A_{\mathfrak{d}}(x).$$

Here, the variable $z(x, a)$ is interpreted as a state-action limiting frequency.

It is known that the minimal value of (**P**) equals $g^*_{n,\mathfrak{d}}$, the optimal average cost of the control model $\mathcal{M}_{n,\mathfrak{d}}$. Moreover, if $\{z^*(x, a)\}$ is an optimal solution of (**P**), letting $X^* = \{x \in \Gamma_n : \sum_{a \in A_{\mathfrak{d}}(x)} z^*(x, a) > 0\}$, define the randomized stationary policy

$\varphi^* \in \Phi_{\mathfrak{d}}$ as

$$\varphi^*(\{a\}|x) = \frac{z^*(x,a)}{\sum_{a' \in A_{\mathfrak{d}}(x)} z^*(x,a')} \quad \text{for } a \in A_{\mathfrak{d}}(x) \text{ and } x \in X^*, \tag{6.1}$$

and define it arbitrarily for states in $\Gamma_n - X^*$. We have that the policy φ^* is average cost optimal for the control model $\mathcal{M}_{n,\mathfrak{d}}$. Besides,

$$\mu_{\varphi}^{n,\mathfrak{d}}(x) = \sum_{a \in A_{\mathfrak{d}}(x)} z^*(x,a) \quad \text{for } x \in \Gamma_n$$

is its unique invariant probability measure. In particular, $R(\varphi^*) := X^*$ is the set of recurrent states for φ^*.

Observe that our procedure in Section 5 to approximate an optimal control policy for $\mathcal{M}$ is concerned with a *canonical* policy $\tilde{f}_{n,\mathfrak{d}}$ for $\mathcal{M}_{n,\mathfrak{d}}$. The policy φ^* determined above is average optimal for $\mathcal{M}_{n,\mathfrak{d}}$ but it might not be canonical. Hence, to use our method in Section 5, we *must solve the ACOE* for $\mathcal{M}_{n,\mathfrak{d}}$ in order to find a canonical policy.

The dual problem of (**P**) is the LP problem (**D**) given by:

$$\text{maximize} \quad g \quad \text{subject to} \quad g + h(x) \le c(x,a) + \sum_{y \in \Gamma_n} Q_n(\{y\}|x,a)h(y)$$

$$\text{for all } x \in \Gamma_n \text{ and } a \in A_{\mathfrak{d}}(x), \quad g \in \mathbb{R} \quad \text{and} \quad h(x) \in \mathbb{R} \quad \text{for } x \in \Gamma_n.$$

Its optimal value is $g_{n,\mathfrak{d}}^*$ and, at optimality, we obtain a solution of the following inequalities

$$g_{n,\mathfrak{d}}^* + h(x) \le \min_{a \in A_{\mathfrak{d}}(x)} \left\{ c(x,a) + \sum_{y \in \Gamma_n} Q_n(\{y\}|x,a)h(y) \right\} \quad \text{for all } x \in \Gamma_n. \tag{6.2}$$

It is important to mention that, by solving (**D**), we might not obtain a solution to the ACOE for the control model $\mathcal{M}_{n,\mathfrak{d}}$, although we know from Theorem 3 that such a solution indeed exists.

Solving the ACOE by LP. Next we show how we can find a solution to the ACOE for the control model $\mathcal{M}_{n,\mathfrak{d}}$ by solving two linear programs.

LEMMA 6.1. Let $\{z^*(x,a)\}$ be an optimal solution of (**P**), let $\varphi^* \in \Phi_{\mathfrak{d}}$ be as in (6.1), and fix arbitrary $x^* \in R(\varphi^*)$. Let $h^* \in \mathbb{R}^n$ be the unique solution of the ACOE for $\mathcal{M}_{n,\mathfrak{d}}$ such that $h^*(x^*) = 0$, and let $h \in \mathbb{R}^{|\Gamma_n|}$, with $h(x^*) = 0$, verify the inequalities in (6.2). Then we have $h(x) \le h^*(x)$ for all $x \in \Gamma_n$.

Proof. We deduce from (6.2) that

$$g_{n,\mathfrak{d}}^* + h(x) \le c(x,\varphi^*) + \sum_{y \in \Gamma_n} Q_n(\{y\}|x,\varphi^*)h(y) \quad \text{for all } x \in \Gamma_n,$$

with $c(x,\varphi^*) = \int_{A_{\mathfrak{d}}(x)} c(x,a)\varphi^*(\mathrm{d}a|x)$. Since φ^* is average optimal, i.e.

$$\sum_{x \in \Gamma_n} c(x,\varphi^*)\mu_{\varphi^*}^{n,\mathfrak{d}}(x) = g_{n,\mathfrak{d}}^*,$$

the above inequality necessarily holds with equality for the recurrent states $R(\varphi^*)$ of φ^*:

$$g^*_{n,\mathfrak{d}} + h(x) = c(x,\varphi^*) + \sum_{y\in R(\varphi^*)} Q_n(\{y\}|x,\varphi^*)h(y) \quad \text{for all } x \in R(\varphi^*).$$

On the other hand, we deduce from the ACOE that

$$g^*_{n,\mathfrak{d}} + h^*(x) \le c(x,\varphi^*) + \sum_{y\in R(\varphi^*)} Q_n(\{y\}|x,\varphi^*)h^*(y) \quad \text{for all } x \in R(\varphi^*).$$

Therefore, function $h - h^*$, when restricted to the set of recurrent states $R(\varphi^*)$, is subharmonic for φ^* and, hence, constant. Since $h - h^*$ vanishes at $x^* \in R(\varphi^*)$, we conclude that $h(x) = h^*(x)$ for all $x \in R(\varphi^*)$.

Now, let $f^* \in \mathbb{F}_{\mathfrak{d}}$ be a canonical policy for $\mathcal{M}_{n,\mathfrak{d}}$, that is, it attains the minimum in the ACOE:

$$g^*_{n,\mathfrak{d}} + h^*(x) = c(x,f^*) + \sum_{y\in\Gamma_n} Q_n(\{y\}|x,f^*)h^*(y) \quad \text{for all } x \in \Gamma_n. \tag{6.3}$$

Since we also have, by (6.2),

$$g^*_{n,\mathfrak{d}} + h(x) \le c(x,f^*) + \sum_{y\in\Gamma_n} Q_n(\{y\}|x,f^*)h(y) \quad \text{for all } x \in \Gamma_n, \tag{6.4}$$

we obtain that $h - h^*$ is superharmonic for the kernel $Q_n(\cdot|\cdot,f^*)$. Hence, $h - h^*$ is constant on set R of recurrent states for f^*. Arguing as in the proof of Theorem 5.3(ii), we have that $R \cap R(\varphi^*)$ is not empty, and so $h(x) = h^*(x)$ for all $x \in R$.

Let us now write (6.4) in matrix form for the transient states $T = \Gamma_n - R$ of f^*:

$$h(T) \le c(T,f^*) - g^*_{n,\mathfrak{d}}\mathbf{1} + P_{f^*}(T,T)h(T) + P_{f^*}(T,R)h(R),$$

with

$$Q_n(\cdot|\cdot,f^*) = \begin{pmatrix} P_{f^*}(R,R) & \mathbf{0} \\ P_{f^*}(T,R) & P_{f^*}(T,T) \end{pmatrix}.$$

This implies, as in the proof of Theorem 5.3(ii), that

$$\begin{aligned} h(T) &\le (\mathbf{I} - P_{f^*}(T,T))^{-1}\Big(c(T,f^*) - g^*_{n,\mathfrak{d}}\mathbf{1} + P_{f^*}(T,R)h(R)\Big) \\ &= (\mathbf{I} - P_{f^*}(T,T))^{-1}\Big(c(T,f^*) - g^*_{n,\mathfrak{d}}\mathbf{1} + P_{f^*}(T,R)h^*(R)\Big) \end{aligned} \tag{6.5}$$

$$= h^*(T), \tag{6.6}$$

where (6.5) follows from the fact that $h = h^*$ on R, and (6.6) is deduced from (6.3). This completes the proof that $h(x) \le h^*(x)$ for all $x \in \Gamma_n$. □

Note that the proof that $h \le h^*$ mainly relies on the properties of the canonical policy f^*. Since our goal is, in fact, to solve the ACOE, such a canonical policy is not, therefore, 'available'. That is why this proof uses the policy φ^*, which can be explicitly determined

by solving (**P**), and then uses the link between them: $R \cap R(\varphi^*) \neq \emptyset$, deduced from the ergodicity property of $\mathcal{M}_{n,\mathfrak{d}}$ on $\Phi_{\mathfrak{d}}$ (recall Proposition 5.2).

As a consequence of Lemma 6.1, we have that h^* in the ACOE for $\mathcal{M}_{n,\mathfrak{d}}$, with $h^*(x^*) = 0$ is the maximal solution h of (6.2) with $h(x^*) = 0$. This leads to the definition of the LP problem (**D**′) as

$$\text{maximize} \quad \sum_{x\in\Gamma_n} h(x) \quad \text{subject to} \quad g^*_{n,d} + h(x) \leq c(x,a) + \sum_{y\in\Gamma_n} Q_n(\{y\}|x,a)h(y)$$

$$\text{for all } x \in \Gamma_n \text{ and } a \in A_{\mathfrak{d}}(x), \quad h(x^*) = 0 \quad \text{and} \quad h(x) \in \mathbb{R} \text{ for } x \in \Gamma_n.$$

It should be clear that h^*, the unique solution of the ACOE with $h^*(x^*) = 0$, is the unique optimal solution of (**D**′). Observe also that the LP problem (**D**′) is somehow 'parametrized' by the optimal value $g^*_{n,\mathfrak{d}}$ and the state x^*. We summarize these results in our next theorem.

THEOREM 6.2. Suppose that Assumptions A, B, D and E are satisfied. Given $n \geq 1$ and $\mathfrak{d} > 0$, consider $\omega \in F_n(\mathfrak{c}_3)$ and suppose that the sets $A_{\mathfrak{d}}(x)$ are finite for every $x \in X$. The following procedure allows to derive $g^*_{n,\mathfrak{d}} \in \mathbb{R}$ and the policy $\tilde{f}_{n,\mathfrak{d}} \in \mathbb{F}_{\mathfrak{d}}$ – the optimal average cost and a canonical policy for $\mathcal{M}_{n,\mathfrak{d}}$ – with the properties given in Theorems 4.4 and 5.6.

- Solve the LP problem (**P**). Let $g^*_{n,\mathfrak{d}} \in \mathbb{R}$ be its optimal value and let $\{z^*(x,a)\}$ be an optimal solution. Determine a state $x^* \in \Gamma_n$ with $\sum_{a\in A_{\mathfrak{d}}(x)} z^*(x^*,a) > 0$.
- For $g^*_{n,\mathfrak{d}}$ and x^* as above, solve the LP problem (**D**′) and determine $h^* \in \mathbb{R}^{|\Gamma_n|}$, which is a solution of the ACOE for $\mathcal{M}_{n,\mathfrak{d}}$. The canonical policy $\tilde{f}_{n,\mathfrak{d}}$ can now be obtained from h^*.

Therefore, we can find the solutions to the ACOE for $\mathcal{M}_{n,\mathfrak{d}}$ by solving two 'connected' LP problems (**P**) and (**D**′). These LP problems are connected in the sense that, first of all, we must solve (**P**) and then, with some data obtained from this solution, we solve (**D**′), which yields the solution to the ACOE.

7. Application to an inventory management system

The dynamics of the inventory management system is given by

$$x_{t+1} = \max\{x_t + a_t - \xi_t, 0\} \quad \text{for } t \in \mathbb{N}, \tag{7.1}$$

where x_t stands for the stock level at the beginning of period t, a_t is the amount ordered by the controller at the beginning of period t and ξ_t is the random demand at the end of period t. We suppose that $\{\xi_t\}_{t\in\mathbb{N}}$ are i.i.d. random variables taking values in $\mathbb{R}_+$, with density function f with respect to the Lebesgue measure and distribution function F. The capacity of the warehouse is given by $M > 0$. Therefore, we have

$$X = A = [0,M] \quad \text{and} \quad A(x) = [0, M - x] \quad \text{for } x \in X,$$

and so $\mathbb{K} = \{(x,a) \in [0,M] \times [0,M] : x + a \leq M\}$. The transition kernel is given by

$$Q(B|x,a) = (1 - F(x+a))\mathbf{I}_B(0) + \int_{[0,x+a]\cap B} f(x+a-y)\,\mathrm{d}y \tag{7.2}$$

for measurable $B \subseteq X$ and $(x, a) \in \mathbb{K}$. The controller incurs a (buying) cost of $b > 0$ for each ordered unit, a holding cost $h > 0$ for each period a unit spends in the warehouse and receives an amount of $p > 0$ for each unit that is sold. Therefore, the running cost function is

$$c(x, a) = ba + h(x + a) - pE[\min\{x + a, \xi\}],$$

where ξ has density f, that is,

$$c(x, a) = ba + h(x + a) - p\int_0^{x+a} sf(s)\,\mathrm{d}s - p(x + a)(1 - F(x + a)) \quad \text{for } (x, a) \in \mathbb{K}.$$

Due to the particular nature of the state $0 \in X$, which can be reached with positive probability, the state space X is endowed with the following metric:

$$\rho_X(x, x') = \begin{cases} |x - x'| & \text{if } x, x' \in (0, M] \\ 1 + x' & \text{if } x = 0 \text{ and } x' \in (0, M] \\ 0 & \text{if } x = x' = 0. \end{cases}$$

This consists in considering the usual topology on $(0,M]$, and letting 0 to be an isolated point identified with -1. On $A = [0,M]$ we consider the usual topology.

PROPOSITION 7.1. Suppose that the distribution function F of the demand has a density function f which is Lipschitz continuous on $[0, M]$. In addition, assume that $f(0) = 0$ and $F(M) < 1$. Under these conditions, the inventory management system satisfies Assumptions A–E in this paper.

Proof. In this proof we will use the following fact. If $x, x' \in X$ then $|x - x'| \leq \rho_X(x, x')$. In particular, Lipschitz continuity of a function when X is endowed with the usual topology implies Lipschitz continuity with respect to ρ_X.

Assumption (A1) holds since $d_H(A(x), A(x')) = |x - x'|$ for $x, x' \in X$. Also, letting $w \equiv 1$, Assumption (A2) is satisfied by choosing any $0 < d < 1$ and $b > 0$ such that $d + b \geq 1$. Finally, it is clear that Assumption (A3) holds because f and F are Lipschitz continuous on $[0, M]$.

Concerning Assumption B, fix arbitrary $0 < \mathfrak{p} < 1$ and define the probability measure μ on X as follows:

$$\mu\{0\} = \mathfrak{p} \quad \text{and} \quad \mu(B) = \frac{1 - \mathfrak{p}}{M}\lambda(B) \quad \text{for measurable } B \subseteq (0, M],$$

where λ is the Lebesgue measure on $\mathbb{R}$. It then follows from (7.2) that the density function of $Q(\cdot|x, a)$ with respect to μ is given by

$$q(y|x, a) = \begin{cases} \dfrac{1}{\mathfrak{p}}(1 - F(x + a)) & \text{for } y = 0, \\ \dfrac{M}{1 - \mathfrak{p}} f(x + a - y) & \text{for } 0 < y \leq x + a, \\ 0 & \text{for } x + a \leq y \leq M. \end{cases}$$

The fact that q is Lipschitz continuous both in $y \in X$ and $(x,a) \in \mathbb{K}$ easily follows from Lipschitz continuity of f on $[0,M]$ and on the fact that $f(0) = 0$; so, Assumption (B2) holds. Assumption (B3) trivially holds, and therefore Assumption B is satisfied.

Now we turn to Assumption D. Given $\mathfrak{d} > 0$, let $q_{\mathfrak{d}} = 2 + [M/\mathfrak{d}]$, where $[M/\mathfrak{d}]$ is the integer part of $M/\mathfrak{d}$. For $x \in X$ define

$$A_{\mathfrak{d}}(x) = \left\{ \frac{(M-x)j}{q_{\mathfrak{d}} - 1} : j = 0, 1, \ldots, q_{\mathfrak{d}} - 1 \right\}.$$

This consists in choosing $q_{\mathfrak{d}}$ equally spaced points in $A(x) = [0, M-x]$. Clearly, Assumption D holds, while we have $d_H(A(x), A_{\mathfrak{d}}(x)) \leq M/(q_{\mathfrak{d}} - 1) \leq \mathfrak{d}$ (Assumption (D2)). Finally, we have that $x \mapsto A_{\mathfrak{d}}(x)$ is Lipschitz continuous with $L^*_{\Psi} = 1$.

Concerning Assumption E, which implies Assumption C, we observe that for (x,a) and (x',a') in $\mathbb{K}$

$$\begin{aligned}
\int_X |Q(\mathrm{d}y|x,a) - Q(\mathrm{d}y|x',a')| &= \int_X |q(y|x,a) - q(y|x',a')| \mu(\mathrm{d}y) \\
&= |F(x+a) - F(x'+a')| + \int_0^M |f(x+a-y) - f(x'+a'-y)| \mathrm{d}y \\
&\leq |F(x+a) - F(x'+a')| + F(x+a) + F(x'+a') \\
&= 2F(\max\{x+a, x'+a'\}) \leq 2F(M).
\end{aligned}$$

Consequently, Assumption E holds by letting $d = F(M) < 1$. □

Numerical experimentation. Define the density function of the demand as

$$f(x) = \frac{1}{\lambda^2} x \mathrm{e}^{-x/\lambda} \quad \text{for } x \geq 0$$

for some parameter $\lambda > 0$. In this case, we have

$$q(y|x,a) = \begin{cases} \dfrac{1}{p}\left(1 + \dfrac{x+a}{\lambda}\right) \cdot \mathrm{e}^{-(x+a)/\lambda} & \text{for } y = 0, \\ \dfrac{M}{1-p} \cdot \dfrac{x+a-y}{\lambda^2} \cdot \mathrm{e}^{-(x+a-y)/\lambda} & \text{for } 0 < y \leq x+a, \\ 0 & \text{for } y > x+a \end{cases}$$

and

$$c(x,a) = ba + h(x+a) - 2p\lambda\left(1 - \left(1 + \frac{x+a}{2\lambda}\right)\mathrm{e}^{-(x+a)/\lambda}\right).$$

We take the following values for the parameters of the control model:

$$M = 10; \quad b = 7; \quad h = 3; \quad p = 17; \quad \mathfrak{p} = \frac{1}{10}; \quad l = \frac{5}{2}.$$

For the approximation of the action sets we have chosen $q_{\mathfrak{d}} = 20$.

Table 1. Estimation of the optimal average cost g^*.

	$n = 50$	$n = 150$	$n = 300$	$n = 500$	$n = 700$	$n = 1000$
Mean	−26.8755	−26.4380	−26.2817	−26.1717	−26.1553	−26.1659
SD	2.2119	1.4578	1.0145	0.8104	0.6662	0.5734

To approximate the optimal average cost of the inventory management problem we have generated n (for $n = 50, 150, 300, 500, 700, 1000$) i.i.d. samples of the probability measure μ. For each such sample, the value $g^*_{n,\mathfrak{d}}$ has been determined by solving the primal LP problem (**P**) described in Section 6. Such computations have been repeated 500 times, thus yielding 500 observations of the random variable $g^*_{n,\mathfrak{d}}$. The results are summarized in Table 1.

We observe that the expected values converge very fast, with a decreasing variance. Also, we have displayed the density estimation for the 500 samples of $g^*_{n,\mathfrak{d}}$ for the above values of n; see Figure 1.

Regarding the approximation of an optimal policy, for the n ($n = 50, 150, 300, 500, 700, 1000$) i.i.d. samples of the probability measure μ we have solved the ACOE of the control model $\mathcal{M}_{n,\mathfrak{d}}$ by solving the linear programs (**P**) and (**D**′); recall Theorem 2. Once the solution h^* to the ACOE (on the states of Γ_n) is obtained, we can determine a canonical policy $\tilde{f}_{n,\mathfrak{d}} \in \mathbb{F}_{\mathfrak{d}}$. Namely, given arbitrary $x \in X$ we can explicitly determine an action $a^* = \tilde{f}_{n,\mathfrak{d}}(x) \in A_{\mathfrak{d}}(x)$ such that

$$\min_{a \in A_{\mathfrak{d}}(x)} \left\{ c(x,a) + \sum_{y \in \Gamma_n} Q_n(\{y\}|x,a)h(y) \right\} = c(x,a^*) + \sum_{y \in \Gamma_n} Q_n(\{y\}|x,a^*)h(y). \tag{7.3}$$

To evaluate the policy $\tilde{f}_{n,\mathfrak{d}}$ under the control model $\mathcal{M}$ we proceed as follows. If the system is in state $x \in X$, we determine the action $\tilde{f}_{n,\mathfrak{d}}(x)$ as in (7.3). Then we simulate a transition of the control model $\mathcal{M}$ under the dynamics (7.1). This procedure is repeated 2000 times starting from the initial state $x_0 = 5$. Then we compute the average value of the corresponding $c(x_t, a_t)$ for $t = 0, \ldots, 2000$. Therefore, to evaluate $J(\tilde{f}_{n,\mathfrak{d}}, x_0)$ we have performed a minimization as in (7.3) for each $t = 0, 1, \ldots, 2000$. This procedure is repeated 500 times, so that we obtain a sample of size 500 of the random variable $J(\tilde{f}_{n,\mathfrak{d}}, x_0)$ (we note that, by ergodicity, this average cost does not depend on the initial state x_0). Our results are displayed in Table 2, while in Figure 2 we display the corresponding density estimation.

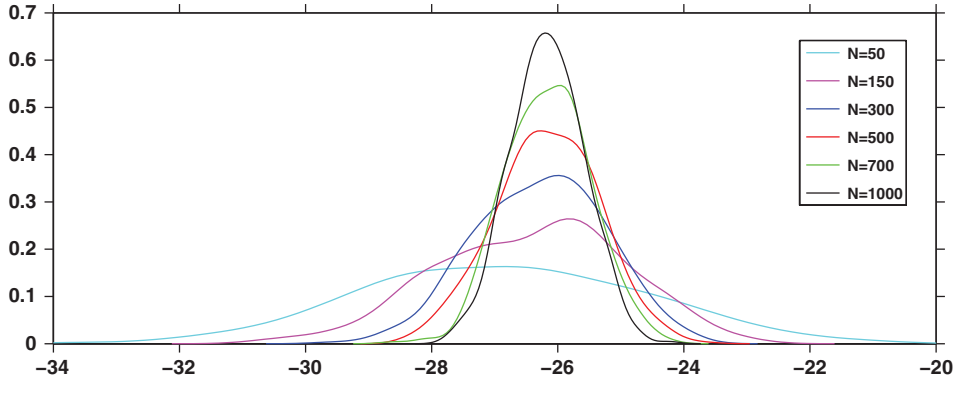

Figure 1. Density estimators for $g^*_{n,\mathfrak{d}}$.

Table 2. Estimation of the average cost of the policy $\tilde{f}_{n,\mathfrak{d}}$.

	$n = 50$	$n = 150$	$n = 300$	$n = 500$	$n = 700$	$n = 1000$
Mean	−25.6312	−25.8387	−25.9724	−26.0406	−26.0497	−26.0833
SD	0.7648	0.5394	0.3954	0.3387	0.3276	0.3133

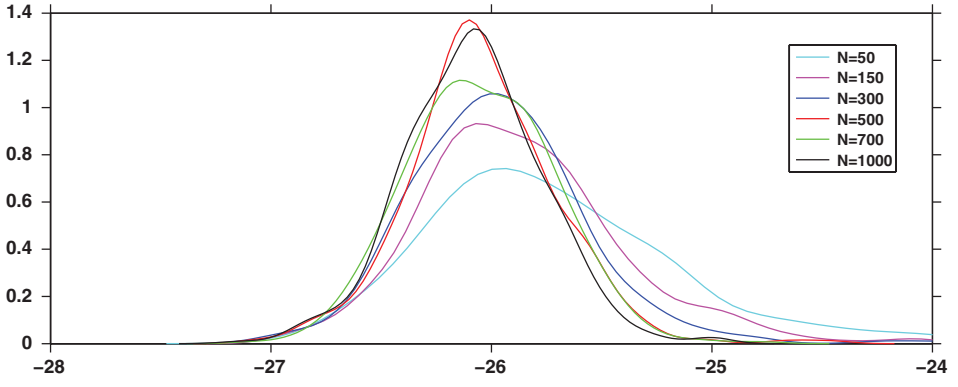

Figure 2. Density estimators for $J(\tilde{f}_{n,\mathfrak{d}}, x_0)$.

Table 3. Relative error of the policy $\tilde{f}_{n,\mathfrak{d}}$ in the control model $\mathcal{M}$.

$n = 50$	$n = 150$	$n = 300$	$n = 500$	$n = 700$	$n = 1000$
4.63%	2.27%	1.18%	0.50%	0.40%	0.32%

We also observe that the expected values converge very fast, that the variances become very small as well and also that these figures are very close to those given in Table 1.

For each n, define $\hat{g}_{n,\mathfrak{d}}$ as the mean value of the 500 observations of $g^*_{n,\mathfrak{d}}$ given in Table 1, and define also $\hat{J}_{n,\mathfrak{d}}$ as the mean of the 500 observations of $J(\tilde{f}_{n,\mathfrak{d}}, x_0)$. In Table 3 we display the relative error $|(\hat{J}_{n,\mathfrak{d}} - \hat{g}_{n,\mathfrak{d}})/\hat{g}_{n,\mathfrak{d}}|$, which we interpret as follows: taking $\hat{g}_{n,\mathfrak{d}}$ as the 'true' value of g^*, the relative error in Table 3 measures how far the policy $\tilde{f}_{n,\mathfrak{d}} \in \mathbb{F}$, constructed from the control model $\mathcal{M}_{n,\mathfrak{d}}$, is from optimality in the control model $\mathcal{M}$. The results in Table 3 show that our method to approximate an optimal policy is fairly accurate.

Funding

This research was supported by the Spanish Ministerio de Economía y Competitividad [grant number MTM2012-31393]. The first author was supported by the French National Agency of Research (ANR) [project number STOCH-MC ANR-13-BS02-0011-01].

Note

1. Email: dufour@math.u-bordeaux1.fr

References

[1] J. Abounadi, D. Bertsekas, and V.S. Borkar, *Learning algorithms for Markov decision processes with average cost*, SIAM J. Control Optim. 40(3) (2001), pp. 681–698.

[2] A. Arapostathis, V.S. Borkar, E. Fernández-Gaucherand, M.K. Ghosh, and S.I. Marcus, *Discrete-time controlled Markov processes with average cost criterion: A survey*, SIAM J. Control Optim. 31(2) (1993), pp. 282–344.
[3] D.P. Bertsekas and J.N. Tsitsiklis, *Neuro-Dynamic Programming*, Athena Scientific, Belmont, MA, 1996.
[4] V.I. Bogachev, *Measure Theory*, Vol. II, Springer, Berlin, 2007.
[5] E. Boissard, *Simple bounds for convergence of empirical and occupation measures in 1-Wasserstein distance*, Electron. J. Probab. 16 (2011), pp. 2296–2333.
[6] X.-R. Cao, Z. Ren, S. Bhatnagar, M. Fu, and S. Marcus, *A time aggregation approach to Markov decision processes*, Autom. J. IFAC 38(6) (2002), pp. 929–943.
[7] H.S. Chang, *A policy improvement method for constrained average Markov decision processes*, Oper. Res. Lett. 35(4) (2007), pp. 434–438.
[8] H.S. Chang, M.C. Fu, J. Hu, and S.I. Marcus, *Simulation-Based Algorithms for Markov Decision Processes*, Communications and Control Engineering Series, Springer, London, 2007.
[9] H.S. Chang and S.I. Marcus, *Approximation receding horizon approach for Markov decision processes: Average reward case*, J. Math. Anal. Appl. 286(2) (2003), pp. 636–651.
[10] W.L. Cooper, S.G. Henderson, and M.E. Lewis, *Convergence of simulation-based policy iteration*, Probab. Eng. Inform. Sci. 17(2) (2003), pp. 213–234.
[11] F. Dufour and T. Prieto-Rumeau, *Approximation of Markov decision processes with general state space*, J. Math. Anal. Appl. 388(2) (2012), pp. 1254–1267.
[12] F. Dufour and T. Prieto-Rumeau, *Finite linear programming approximations of constrained discounted Markov decision processes*, SIAM J. Control Optim. 51(2) (2013), pp. 1298–1324.
[13] V.F. Farias, C.C. Moallemi, B. Van Roy, and T. Weissman, *Universal reinforcement learning*, IEEE Trans. Inform. Theory 56(5) (2010), pp. 2441–2454.
[14] E. Gordienko, E. Lemus-Rodríguez, and R. Montes-de Oca, *Average cost Markov control processes: Stability with respect to the Kantorovich metric*, Math. Methods Oper. Res. 70(1) (2009), pp. 13–33.
[15] X. Guo and Q. Zhu, *Average optimality for Markov decision processes in Borel spaces: A new condition and approach*, J. Appl. Probab. 43(2) (2006), pp. 318–334.
[16] O. Hernández-Lerma and J.-B. Lasserre, *Discrete-Time Markov Control Processes: Basic Optimality Criteria*, Applications of Mathematics, Vol. 30, Springer, New York, 1996.
[17] O. Hernández-Lerma and J.B. Lasserre, *Policy iteration for average cost Markov control processes on Borel spaces*, Acta Appl. Math. 47(2) (1997), pp. 125–154.
[18] O. Hernández-Lerma and J.-B. Lasserre, *Further Topics On Discrete-Time Markov Control Processes*, Applications of Mathematics, Vol. 42, Springer, New York, 1999.
[19] V.R. Konda and J.N. Tsitsiklis, *On actor-critic algorithms*, SIAM J. Control Optim. 42(4) (2003), pp. 1143–1166.
[20] P. Marbach and J.N. Tsitsiklis, *Simulation-based optimization of Markov reward processes*, IEEE Trans. Automat. Control 46(2) (2001), pp. 191–209.
[21] P. Marbach and J.N. Tsitsiklis, *Approximate gradient methods in policy-space optimization of Markov reward processes*, Discrete Event Dyn. Syst. 13(1-2) (2003), pp. 111–148. Special issue on learning, optimization and decision making.
[22] W.B. Powell, *Approximate Dynamic Programming*, Wiley Series in Probability and Statistics, Wiley-Interscience [John Wiley & Sons], Hoboken, NJ, 2007.
[23] R.S. Sutton and A.G. Barto, *Reinforcement Learning: An Introduction*, MIT Press, Cambridge, MA, 1998.
[24] B. Van Roy, *Neuro-dynamic programming: Overview and recent trends*, in *Handbook of Markov Decision Processes*, International Series in Operations Research and Management Science, Vol. 40, Kluwer Academic, Boston, MA, 2002, pp. 431–459.
[25] H. Yu and D.P. Bertsekas, *Convergence results for some temporal difference methods based on least squares*, IEEE Trans. Automat. Control 54(7) (2009), pp. 1515–1531.