Adaptive Control of Stochastic Systems with Unknown Disturbance Distribution: Discounted Criteria*

Nadine Hilgert† and J. Adolfo Minjárez-Sosa‡

March 2002

Abstract

We consider a class of discrete-time stochastic control systems, with Borel state and action spaces, and possibly unbounded costs. The processes evolve according to the equation $x_{t+1} = F(x_t, a_t, \xi_t)$, $t = 0, 1, \ldots$, where the ξ_t are i.i.d. random vectors whose common distribution is *unknown*. Assuming observability of $\{\xi_t\}$, we use the empirical estimator of its distribution to construct adaptive policies which are asymptotically discounted cost optimal .
AMS 1991 subject classifications: 93E20, 90C40.

Key Words: Distribution estimation; discrete-time stochastic systems; discounted cost criteria; optimal adaptive policy.

1 Introduction

We consider a class of discrete-time Markov control processes evolving according to the equation

$$x_{t+1} = F(x_t, a_t, \xi_t), \quad t = 0, 1, \ldots, \tag{1}$$

*Work supported partially by Consejo Nacional de Ciencia y Tecnología (CONACyT) under Grant 37239-E.

†Laboratoire d'Analyse des Systèmes et de Biométrie, INRA-ENSA.M, 2 place Viala, 34060 Montpellier Cedex 1, France. (hilgert@ensam.inra.fr).

‡Departamento de Matemáticas, Universidad de Sonora, Rosales s/n, Col. Centro, 83000, Hermosillo, Sonora, México. (aminjare@gauss.mat.uson.mx)

where x_t, a_t and ξ_t are the state, action and random disturbance at time t respectively, taking values on Borel spaces. F is a known measurable function. Moreover, $\{\xi_t\}$ is an observable sequence of independent and identically distributed (i.i.d.) random vectors with unknown distribution θ. The actions that can be applied at any given time are selected according to rules known as control policies directed to optimize a performance index. The optimal control problem we are dealing with in this paper is to determine a control policy that minimizes an α-discounted cost criterion. This criterion is expressed from some one stage cost functions c, possibly unbounded, and depends on the unknown distribution θ. However, since θ is unknown, the controller has to combine the actions selection with a statistical estimation procedure of θ, and the resulting policy is called *adaptive.*

The main contribution of the paper is as follows: using the empirical distribution of the disturbance process $\{\xi_t\}$ to estimate θ, we construct two adaptive policies which are asymptotically discounted cost optimal for the system (1). The first adaptive policy is obtained via the so-called Principle of Estimation and Control (PEC), which was described by Mandl in [17] as the method of substituting the estimates into optimal stationary controls (see also [16]). The PEC-policy is also known in the literature of the stochastic control theory as *certainty equivalence controller* or *naive feedback controller* (see [2]). We also construct an iterative adaptive policy which is a slight extension of "The Non-stationary Value Iteration" policy introduced in [13].

The problem of constructing asymptotically discounted cost optimal adaptive policies for the system (1), when the distribution of the disturbance process is unknown, has been studied in several contexts. For instance, this problem is studied in [4, 9, 10, 13] considering either bounded costs or compact state spaces. In recent papers [8, 14], these results are extended to the cases of unbounded costs and general state and action spaces, but considering that the unknown disturbance distribution θ is absolutely continuous (with respect to the Lebesgue measure on $\Re^k$, the space of the disturbances ξ_t), this implies the existence of an unknown density function. Hence, the estimation of θ is obtained by means of an estimator of its density function. However, the unboundedness assumption on the cost in [8, 14] makes difficult the implementation of the density estimation process. For instance, the estimator is defined by the projection (of an auxiliar estimator) on some special set of density functions to ensure good properties of the estimated model. Beyond the complexity of the estimation procedure, the assumption of ab-

solutely continuity excludes the case of discrete distributions, which appears in some inventory-production and queueing systems.

The construction of adaptive policies using the empirical distribution is very general in the sense that the distribution θ can be arbitrary. Therefore our work extends the results of above mentioned papers [4, 8, 9, 10, 13].

The paper is organized as follows. In Section 2 we introduce the Markov control models we are concerned with and the definition of asymptotic discount optimality. In Section 3 we construct the adaptive policies and state the optimality in the main result, Theorem 3.8. It is proved in Section 4. Finally, the assumptions and the results are illustrated in Section 5.

Remark 1.1 *Given a Borel space X (that is, a Borel subset of a complete and separable metric space) its Borel sigma-algebra is denoted by $\mathcal{B}(X)$, and "measurable", for either sets or functions, means "Borel measurable". The space of probability measures on X is denoted by $\mathbb{P}(X)$. Let X and Y be Borel spaces. Then a stochastic kernel $Q(dx \mid y)$ on X given Y is a function such that $Q(\cdot \mid y)$ is a probability measure on X for each fixed $y \in Y$, and $Q(B \mid \cdot)$ is a measurable function on Y for each fixed $B \in \mathcal{B}(X)$.*

2 Markov control models

We consider a class of discrete-time Markov control models

$$\mathcal{M} := (X, A, \{A(x) \subset A | x \in X\}, S, F, \theta, c) \tag{2}$$

associated to the system (1), satisfying the following conditions. The state space X, the action space A and the disturbance space S are Borel spaces endowed with their Borel σ-algebras (See Remark 1.1). For each state $x \in X$, $A(x)$ is a nonempty Borel subset of A denoting the set of admissible controls when the system is in state x. The set

$$\mathbb{K} = \{(x, a) : x \in X, a \in A(x)\}$$

of admissible state-action pairs is assumed to be a Borel subset of the Cartesian product of X and A. The function $F : X \times A \times S \to X$, as in (1), is a given (known) measurable function and represents the dynamics of the system. Moreover, $\theta \in \mathbb{P}(S)$ denotes the common–but unknown– distribution of the i.i.d. disturbances ξ_t in (1), which are S-valued random vectors

defined on an underlying probability space $(\Omega, \mathcal{F}, P)$. Thus

$$\theta(B) = P\left(\xi_t \in B\right),\ t \in \mathbb{N},\ B \in \mathcal{B}(S). \tag{3}$$

Finally, the cost-per-stage $c(x, a)$ is a nonnegative measurable real-valued function on $\mathbb{K}$, possibly unbounded.

We denote by Q the stochastic kernel representing the transition law corresponding to (1), that is, for all $t \in \mathbb{N}$, $(x, a) \in \mathbb{K}$ and $B \in \mathcal{B}(X)$,

$$\begin{aligned} Q(B|x,a) :&= \text{Prob}\left[x_{t+1} \in B | x_t = x, a_t = a\right] \\ &= \int_S 1_B\left(F(x,a,s)\right)\theta(ds) \\ &= \theta\left(\{s \in S\ :\ F(x,a,s) \in B\}\right), \end{aligned}$$

where $1_B(.)$ denotes the indicator function of the set B.

Throughout the paper, the probability space $(\Omega, \mathcal{F}, P)$ is fixed and $a.s.$ means *almost surely with respect to* P. In addition, we assume that the realizations $\xi_0, \xi_1, ...$ of the disturbance process and the states $x_0, x_1, ...$ are completely observable.

We define the spaces of admissible histories up to time t by $\mathbb{H}_0 := X$ and $\mathbb{H}_t := (\mathbb{K}\times S)^t \times X$, $t \geq 1$. A generic element of $\mathbb{H}_t$ is written as $h_t = (x_0, a_0, \xi_0, ..., x_{t-1}, a_{t-1}, \xi_{t-1}, x_t)$. A control policy (randomized, history-dependent) is a sequence $\pi = \{\pi_t\}$ of stochastic kernels π_t on A given $\mathbb{H}_t$ such that $\pi_t(A(x_t) \mid h_t) = 1$, for all $h_t \in \mathbb{H}_t$, $t \in \mathbb{N}$. Let Π be the set of all control policies and $\mathbb{F} \subset \Pi$ the subset of stationary policies. If necessary, see for example [5, 6, 8, 9, 11, 12, 14, 15, 22] for further information on those policies. As usual, each stationary policy $\pi \in \mathbb{F}$ is identified with a measurable function $f : X \to A$ such that $\pi_t(\cdot \mid h_t)$ is concentrated at $f(x_t) \in A(x_t)$ for all $h_t \in \mathbb{H}_t$, $t \in \mathbb{N}$, so that π is of the form $\pi = \{f, f, f, ...\}$. In this case we denote π by f. For each $f \in \mathbb{F}$, we write

$$c(x, f) := c(x, f(x)) \quad \text{and} \quad F(x, f, s) := F(x, f(x), s)$$

for all $x \in X$ and $s \in S$.

Let $V(\pi, x)$ be the α-discounted cost using the policy $\pi \in \Pi$, given the initial state $x_0 = x$. That is,

$$V(\pi, x) := E_x^\pi \left[\sum_{t=0}^{\infty} \alpha^t c(x_t, a_t)\right], \tag{4}$$

where $\alpha \in (0,1)$ is the so-called discount factor, and E_x^π denotes the expectation operator with respect to the probability measure P_x^π induced by the policy π, given the initial state $x_0 = x$. The corresponding value (or optimal cost) function is

$$V^*(x) := \inf_{\pi \in \Pi} V(\pi, x), \ x \in X. \tag{5}$$

A policy $\pi^* \in \Pi$ is said to be α-discount optimal (or simply α-optimal) for the control model $\mathcal{M}$ if

$$V^*(x) = V(\pi^*, x) \text{ for all } \ x \in X. \tag{6}$$

Since θ is unknown, we combine suitable statistical estimation methods of θ and control procedures in order to construct the adaptive policy. That is, we use the observed history of the system to estimate θ and then adapt the decision or control to the available estimate. On the other hand, as the discounted cost depends heavily on the controls selected at the first stages (precisely when the information about the distribution θ is poor or deficient), we can't ensure the existence of an α-optimal adaptive policy (see e.g. [9]). Thus the α-optimality of an adaptive policy will be understood in the following asymptotical sense:

Definition 2.1 *a) [20] A policy $\pi \in \Pi$ is said to be asymptotically discount optimal for the control model $\mathcal{M}$ if*

$$\left|V^{(k)}(\pi, x) - E_x^\pi \left[V^*(x_k)\right]\right| \to 0 \ \ as \ \ k \to \infty, \ for \ all \ x \in X,$$

where

$$V^{(k)}(\pi, x) := E_x^\pi \left[\sum_{t=k}^{\infty} \alpha^{t-k} c(x_t, a_t)\right]$$

is the expected total discounted cost from stage k onward and $a_t = \pi_t(h_t)$.

b) Let $\delta \geq 0$. A policy π is δ-asymptotically discount optimal for the control model $\mathcal{M}$ if

$$\limsup_{k \to \infty} \left|V^{(k)}(\pi, x) - E_x^\pi \left[V(x_k)\right]\right| \leq \delta, \quad \forall x \in X.$$

Clearly, discount optimality implies asymptotic discount optimality, which in turn implies δ-asymptotic discount optimality.

3 Main result

To estimate θ we use the empirical distribution $\{\theta_t\} \subset I\!P(S)$ of the disturbance process $\{\xi_t\}$, defined as follows. Let $\nu \in I\!P(S)$ be a given probability measure. Then

$$\begin{aligned} \theta_0 &:= \nu, \\ \theta_t(B) &:= \frac{1}{t}\sum_{i=0}^{t-1} 1_B(\xi_i), \quad \text{for all } t \geq 1 \text{ and } B \in \mathcal{B}(S). \end{aligned} \tag{7}$$

Lemma 3.1 *(See [7].) θ_t converges weakly to θ a.s., that is,*

$$\int u d\theta_t \to \int u d\theta \quad a.s. \quad as\ t \to \infty,$$

for every real-valued, continuous and bounded function u on S. Equivalently, if u is only lower semicontinuous (l.s.c.) and bounded below, then

$$\liminf_{t\to\infty} \int u d\theta_t \geq \int u d\theta \quad a.s.$$

Assumption 3.2 *a) For each $x \in X$, the set $A(x)$ is σ-compact.*
b) For each $x \in X$ the function $a \to c(x,a)$ is l.s.c. on $A(x)$. Moreover, there exists a l.s.c. function $W : X \to [\bar{w}, \infty)$ such that $\sup_{a\in A(x)} c(x,a) \leq W(x)$ for all $x \in X$, where $\bar{w}$ is a positive constant. (Recall that c is assumed to be nonnegative.)
c)There exist three constants $p > 1$, $\beta_0 < 1$ and $b_0 < +\infty$ such that for all $x \in X$, $a \in A(x)$ and $t \geq 1$, the empirical distribution θ_t satisfies

$$\int_S W^p\left(F(x,a,s)\right)\theta_t(ds) = \frac{1}{t}\sum_{i=0}^{t-1} W^p\left(F(x,a,\xi_i)\right) \leq \beta_0 W^p(x) + b_0 \quad a.s. \tag{8}$$

Remark 3.3 *Concerning the function W in Assumption 3.2(b), we require it to be l.s.c. It is generally supposed to be only measurable [8]. This stronger hypothesis on the cost function c is the price we have to pay for nothing assuming on the unknown distribution θ, see the proof of Lemma 4.1.*

In some applications it suffices to take $W(x) := \sup_{a\in A(x)} c(x,a)$, provided that it is l.s.c. In general, one can try an exponential function, say $W(x) := \beta e^{\gamma x}$, for some suitable values of $\beta > 0$ and $\gamma > 0$ (See §5.2 bellow).

We denote by L_W^∞ the normed linear space of all measurable functions $u : X \to \Re$ with a finite norm $\|u\|_W$ defined as

$$\|u\|_W := \sup_{x \in X} \frac{|u(x)|}{W(x)}. \tag{9}$$

The statement of our main result, Theorem 3.8, requires as background the following proposition, which is proved in [8] (see also [11, 14, 18].)

Proposition 3.4 *Suppose that Assumption 3.2 holds. Then*
a) For all $\pi \in \Pi$ and $x \in X$, $V(\pi, x) \leq CW(x)/(1-\alpha)$, for some constant $C > 0$. Hence, $V^*(x) \leq CW(x)/(1-\alpha)$ for all $x \in X$, and so V^* is in L_W^∞.

Moreover, V^* satisfies the α-discounted cost optimality equation

$$V^*(x) = \inf_{a \in A(x)} \left(c(x,a) + \alpha \int_S V^*(F(x,a,s))\theta(ds) \right), \quad x \in X. \tag{10}$$

b) For each $t \in \mathbb{N}$, there exists a function $V_t \in L_W^\infty$ such that

$$V_t(x) = \inf_{a \in A(x)} \left(c(x,a) + \alpha \int_S V_t(F(x,a,s))\theta_t(ds) \right) \ a.s. \,, \quad x \in X, \tag{11}$$

and $V_t(x) \leq CW(x)/(1-\alpha)$.
c) For each $t \in \mathbb{N}$ and $\hat{\delta}_t > 0$, there exists a stationary policy $\hat{f}_t \in \mathbb{F}$ such that

$$c(x, \hat{f}_t) + \alpha \int_S V_t(F(x, \hat{f}_t, s))\theta_t(ds) \leq V_t(x) + \hat{\delta}_t \ a.s. \quad \forall x \in X. \tag{12}$$

d) Let $\left\{ \bar{V}_t \right\}$ be a sequence of functions defined as: $\bar{V}_0 \equiv 0$ and

$$\bar{V}_t\,(x) = \inf_{a \in A(x)} \left(c(x,a) + \alpha \int_S \bar{V}_{t-1}\,(F(x,a,s))\theta_t(ds) \right) \ a.s. \,, \quad x \in X, \ t \geq 1. \tag{13}$$

Then, for each $t \in \mathbb{N}$ and $\bar{\delta}_t > 0$, there exists $\bar{f}_t \in \mathbb{F}$ such that

$$c(x, \bar{f}_t) + \alpha \int_S \bar{V}_{t-1} \left[F(x, \bar{f}_t, s) \right] \theta_t(ds) \leq \bar{V}_t\,(x) + \bar{\delta}_t, \ \ x \in X. \tag{14}$$

Definition 3.5 *Let* $\{\hat{\delta}_t\}$ *and* $\left\{\bar{\delta}_t\right\}$ *be arbitrary convergent sequences of positive numbers, and let* $\hat{\delta} := \lim_{t\to\infty} \hat{\delta}_t$ *and* $\bar{\delta}$:$= \lim_{t\to\infty} \bar{\delta}_t$. *In addition, let* $\{\hat{f}_t\}$ *and* $\left\{\bar{f}_t\right\}$ *be sequences of functions in* $I\!F$ *satisfying (12) and (14) respectively.*
a) The adaptive policy $\hat{\pi} = \{\hat{\pi}_t\}$ *is defined as*

$$\hat{\pi}_t(h_t) = \hat{\pi}_t(h_t; \theta_t, \hat{\delta}_t) := \hat{f}_t(x_t),\ t \in IN.$$

b) The iterative adaptive policy $\bar{\pi}$= $\{\bar{\pi}_t\}$ *is define as*

$$\bar{\pi}_t\ (h_t) =\bar{\pi}_t\ (h_t; \theta_t, \bar{\delta}_t) :=\bar{f}_t\ (x_t),\ t \in IN. \tag{15}$$

In (a) and (b), $\hat{\pi}_0(x)$ *and* $\bar{\pi}_0\ (x)$ *are any fixed action in* $A(x)$.

Remark 3.6 *a) Observe that, by a result of Schäl [19], there is a policy* $\hat{f}_\infty \in I\!F$ *such that, for each* $x \in X$, $\hat{f}_\infty(x) \in A(x)$ *is an accumulation point of* $\{\hat{f}_t(x)\}$.

b) The construction of the adaptive control policy $\hat{\pi}$ *requires the calculation of* V_t *at each stage* $t \geq 0$ *(i.e., solving an optimality equation for each* $t \geq 0$*) as opposed to the construction of* $\bar{\pi}$ *which is obtained recursively. This is an obvious advantage from the point of view of the numerical implementation. Our main goal is to prove that both policies* $\hat{\pi}$ *and* $\bar{\pi}$ *are asymptotically discounted optimal, which is stated in Theorem 3.8.*

We define

$$V_W^*(x) := \frac{V^*(x)}{W(x)},\ \ x \in X, \tag{16}$$

and

$$\mathcal{V}_W := \{V_W^*\left(F(x, a, .)\right) : (x, a) \in \mathrm{K}\}.$$

To prove our main result, Theorem 3.8, we need the equicontinuity on S of the family of functions $\mathcal{V}_W$. This is supposed in the following.

Assumption 3.7 *The family of functions* $\mathcal{V}_W$ *is equicontinuous on* S.

This assumption is discussed below in Section 5, where we first give two different sets of sufficient hypotheses for Assumption 3.7. Both Assumptions 3.2 and 3.7 are then proved to be true for the example of an inventory-production system.

With the above notation, we may now state our main result as follows.

Theorem 3.8 *Under Assumptions 3.2 and 3.7, we have*

a) $\|V_t - V^*\|_W \to 0$ *a.s. as* $t \to \infty$.

b) The adaptive policies $\hat{\pi}$ *and* $\bar{\pi}$ *are respectively* $\hat{\delta}$ *and* $\bar{\delta}$*-asymptotically discount optimal. In particular, if* $\hat{\delta} = 0$ *(resp.* $\bar{\delta} = 0$*), then the policy* $\hat{\pi}$ *(resp.* $\bar{\pi}$*) is asymptotically discount optimal.*

c) If moreover F *is continuous in* $a \in A(x)$ *for all* $x \in X$*, and* $\hat{\delta} = 0$*, then* $\hat{f}_\infty$ *is* α*-discount optimal for* $\mathcal{M}$.

It is well known that the existence of minimizer of the discounted cost optimality equation (10) implies the existence of discounted cost optimal stationary policies. Thus, it can happen that the assumptions made in this paper (especially Assumption 3.2) are not sufficient to prove the existence of a stationary optimal policy with a known distribution θ of the r.v. ξ_t (see [12]). However Theorem 3.8(c) guarantees the existence of such a policy while considering θ unknown.

4 Proofs

4.1 Preliminary lemmas

Before proving the theorem itself, we shall state some preliminary facts.

Lemma 4.1 *Suppose that Assumption 3.2 holds. Then:*
a) For all $x \in X$ *and* $a \in A(x)$*,*

$$\int_S W^p\left(F(x,a,s)\right)\theta(ds) \le \beta_0 W^p(x) + b_0. \tag{17}$$

b) Letting $\beta := \beta_0^{1/p}$ *and* $b := b_0^{1/p}$*, we have for all* $x \in X, a \in A(x)$*, and* $t \in I\!N$*,*

$$\int_S W\left(F(x,a,s)\right)\theta_t(ds) \le \beta W(x) + b \quad a.s. \tag{18}$$

$$\int_S W\left(F(x,a,s)\right)\theta(ds) \le \beta W(x) + b. \tag{19}$$

c) For all $x \in X$ *and* $\pi \in \Pi$*, we have*

$$\sup_{t\ge 1} E_x^\pi\left[W^p(x_t)\right] < \infty \;\; \textit{and} \;\; \sup_{t\ge 1} E_x^\pi\left[W(x_t)\right] < \infty.$$

Proof: a) As W is l.s.c., there exists an increasing sequence $\{u_k\}$ of continuous and bounded functions such that $u_k(x) \uparrow W^p(x)$ for all $x \in X$. Choose arbitrary $x \in X$ and $a \in A(x)$. Then, by Assumption 3.2(c), for each k and t in $\mathbb{N}$,

$$\int_S u_k\left(F(x,a,s)\right)\theta_t(ds) \leq \int_S W^p\left(F(x,a,s)\right)\theta_t(ds) \leq \beta_0 W^p(x) + b_0,$$

and letting $t \to \infty$, Lemma 3.1 yields

$$\liminf_{t\to\infty} \int_S u_k\left(F(x,a,s)\right)\theta_t(ds) \geq \int_S u_k\left(F(x,a,s)\right)\theta(ds).$$

Thus, for each k in $\mathbb{N}$,

$$\int_S u_k\left(F(x,a,s)\right)\theta(ds) \leq \beta_0 W^p(x) + b_0,$$

and (17) follows by letting $k \to \infty$.
b) See [8].
c) This part follows from (17) and (19). See [8]. $\diamond$

Lemma 4.2 *Under Assumptions 3.2 and 3.7, we have*

$$\lim_{t\to\infty} \sup_{(x,a)\in IK} \left| \int_S V^*\left(F(x,a,s)\right)\theta_t(ds) - \int_S V^*\left(F(x,a,s)\right)\theta(ds) \right| = 0 \quad a.s. \tag{20}$$

Proof: Choose and fix arbitrary $x \in X$ and $a \in A(x)$. Let μ_t and μ be two measures on S defined as

$$\mu_t(B) := \int_B W\left(F(x,a,s)\right)\theta_t(ds), \tag{21}$$

$$\mu(B) := \int_B W\left(F(x,a,s)\right)\theta(ds), \tag{22}$$

for all $B \in \mathcal{B}(S)$. Observe that $\mu(B) = E\left[1_B(\xi_0) W\left(F(x,a,\xi_0)\right)\right]$. Thus, from Lemma 4.1(b), $\mu(B) \leq \beta W(x) + b < \infty$ as x is fixed. We can then

apply the law of large numbers to $\mu_t(B)$:

$$\begin{aligned}\lim_{t\to\infty}\mu_t(B) &= \lim_{t\to\infty}\frac{1}{t}\sum_{i=0}^{t-1}1_B(\xi_i)W\left(F(x,a,\xi_i)\right)\\ &= E\left[1_B(\xi_0)W\left(F(x,a,\xi_0)\right)\right]\\ &= \mu(B)\quad \text{a.s.},\end{aligned}$$

that is, μ_t converges setwise to μ a.s., which of course implies that μ_t converges weakly to μ ($\mu_t \xrightarrow{w} \mu$).

On the other hand, from Assumption 3.7, the family of functions $\mathcal{V}_W$ is equicontinuous at each point $s \in S$. It is also uniformly bounded (by the definition (16) of V_W^* and Proposition 3.4(a)). Therefore $\mathcal{V}_W$ is a μ-uniformity class (see [3]), that is, since $\mu_t \xrightarrow{w} \mu$, we have, as $t \to \infty$,

$$\sup_{(x,a)\in K}\left|\int_S \frac{V^*\left(F(x,a,s)\right)}{W\left(F(x,a,s)\right)}\mu_t(ds) - \int_S \frac{V^*\left(F(x,a,s)\right)}{W\left(F(x,a,s)\right)}\mu(ds)\right| \to 0 \quad \text{a.s.}$$

Thus, from the definitions (21) and (22) of μ_t and μ, we get (20). $\diamond$

We also need the following characterization of asymptotic discount optimality.

Lemma 4.3 *[20, 11] A policy $\pi \in \Pi$ is asymptotically discount optimal for the control model $\mathcal{M}$ if and only if, for $x \in X$,*

$$E_x^\pi\left[\Phi(x_t,a_t)\right] \to 0 \text{ as } t \to \infty,$$

where

$$\Phi(x,a) := c(x,a) + \alpha\int_S V^*\left(F(x,a,s)\right)\theta(ds) - V^*(x),\ (x,a) \in I\!K, \qquad (23)$$

is the so-called discounted discrepancy function. (By (10), Φ is nonnegative.)

Remark 4.4 *For $\delta \geq 0$, a policy $\pi \in \Pi$ is δ-asymptotically discount optimal for the control model $\mathcal{M}$ if*

$$\limsup_{t\to\infty} E_x^\pi\left[\Phi(x_t,a_t)\right] \leq \delta,\ x \in X.$$

4.2 Proof of Theorem 3.8

a) Let us define the operators

$$Tu(x) := \inf_{a\in A(x)} \left\{ c(x,a) + \alpha \int_S u\left(F(x,a,s)\right) \theta(ds) \right\}, \tag{24}$$

$$T_t u(x) := \inf_{a\in A(x)} \left\{ c(x,a) + \alpha \int_S u\left(F(x,a,s)\right) \theta_t(ds) \right\}, \tag{25}$$

for all $x \in X$ and $u \in L^\infty_W$. From Assumption 3.2 and Lemma 4.1(b), T and T_t map L^∞_W to itself.

Now we fix an arbitrary number $\gamma \in (\alpha, 1)$ and define the function $\bar{W}(x) := W(x) + d$ for $x \in X$, where $d := b\left(\gamma/\alpha - 1\right)^{-1}$. Let $L^\infty_{\bar{W}}$ be the space of measurable functions $u : X \to \Re$ with norm

$$\|u\|_{\bar{W}} := \sup_{x\in X} \frac{|u(x)|}{\bar{W}(x)} < \infty.$$

Observe that the norms $\|.\|_W$ and $\|.\|_{\bar{W}}$ are equivalent because

$$\|u\|_{\bar{W}} \le \|u\|_W \le (1+d)\ \|u\|_{\bar{W}}. \tag{26}$$

A consequence of [21, Lemma 2] is that the inequalities (18) and (19) imply respectively that the operators T_t and T, $t \in \mathbb{N}$, are contractions with modulus γ, with respect to the norm $\|.\|_{\bar{W}}$, i.e. for all $u, v \in L^\infty_{\bar{W}}$:

$$\|Tv - Tu\|_{\bar{W}} \le \gamma \|v - u\|_{\bar{W}}, \tag{27}$$

$$\|T_t v - T_t u\|_{\bar{W}} \le \gamma \|v - u\|_{\bar{W}} \text{ a.s.} \tag{28}$$

Thus, by (10) and (11), V^* and V_t are fixed points in $L^\infty_{\bar{W}}$ of the operators T and T_t, respectively, i.e.

$$TV^* = V^* \qquad \text{and} \qquad T_t V_t = V_t \text{ a.s. } \forall t \in \mathbb{N}. \tag{29}$$

Hence

$$\|V^* - V_t\|_{\bar{W}} \le \|TV^* - T_t V^*\|_{\bar{W}} + \gamma \|V^* - V_t\|_{\bar{W}} \quad \text{a.s.},$$

which implies that a.s.

$$\begin{aligned} \|V^* - V_t\|_{\bar{W}} &\le \tfrac{1}{1-\gamma} \|TV^* - T_t V^*\|_{\bar{W}} \qquad (30)\\ &\le \frac{1}{\bar{w}(1-\gamma)} \sup_{(x,a)\in IK} \left| \int_S V^*\left(F(x,a,s)\right) \theta_t(ds) - \int_S V^*\left(F(x,a,s)\right) \theta(ds) \right|. \end{aligned}$$

For each $t \in \mathbb{N}$ let

$$n_t := \sup_{(x,a)\in K} \left| \int_S V^* \left(F(x,a,s)\right) \theta_t(ds) - \int_S V^* \left(F(x,a,s)\right) \theta(ds) \right|. \tag{31}$$

Then, by (26),

$$\|V^* - V_t\|_W \leq \frac{(1+d)n_t}{\bar{w}(1-\gamma)} \quad \text{a.s.} \tag{32}$$

Part (a) of Theorem 3.8 is then a consequence of (20).

b)Optimality of $\hat{\pi}$. For each $t \in \mathbb{N}$, we consider the approximate discrepancy functions $\hat{\Phi}_t : \mathbb{K} \to \mathbb{R}$, given, as in (23), by

$$\hat{\Phi}_t(x,a) := c(x,a) + \alpha \int_S V_t \left(F(x,a,s)\right) \theta_t(ds) - V_t(x) \tag{33}$$

for all $(x,a) \in \mathbb{K}$. Now, from the Definition 3.5 of the adaptive policy $\hat{\pi}$, we have that $\hat{\Phi}_t \left(., \hat{\pi}_t(.)\right) \leq \hat{\delta}_t$ for each $t \in \mathbb{N}$. Thus

$$\begin{aligned}
\Phi\left(x_t, \hat{\pi}_t(h_t)\right) &\leq \left| \Phi\left(x_t, \hat{\pi}_t(h_t)\right) - \hat{\Phi}_t\left(x_t, \hat{\pi}_t(h_t)\right) + \hat{\delta}_t \right| \\
&\leq \sup_{a\in A(x_t)} \left| \Phi(x_t,a) - \hat{\Phi}_t(x_t,a) \right| + \hat{\delta}_t \\
&\leq W(x_t) \sup_{x\in X} W(x)^{-1} \sup_{a\in A(x)} \left| \Phi(x,a) - \hat{\Phi}_t(x,a) \right| + \hat{\delta}_t. \qquad (34)
\end{aligned}$$

Moreover, from the definitions (23) and (33) of Φ and $\hat{\Phi}_t$, we get, by adding and subtracting the term $\alpha \int_S V^* \left(F(x,a,s)\right) \theta_t(ds)$,

$$\begin{aligned}
\left| \hat{\Phi}_t(x,a) - \Phi(x,a) \right| &\leq |V^*(x) - V_t(x)| \\
&\quad + \alpha \int_S \left| V_t\left(F(x,a,s)\right) - V^*\left(F(x,a,s)\right) \right| \theta_t(ds) \\
&\quad + \alpha \left| \int_S V^* \left(F(x,a,s)\right) \theta_t(ds) - \int_S V^* \left(F(x,a,s)\right) \theta(ds) \right|,
\end{aligned}$$

which combined with Lemma 4.1(b) yields

$$\begin{aligned}
\left| \hat{\Phi}_t(x,a,) - \phi(x,a) \right| &\leq \|V^* - V_t\|_W W(x) \\
&\quad + \alpha \|V^* - V_t\|_W \left(\beta W(x) + b\right) + n_t \quad \text{a.s.}
\end{aligned}$$

for all $(x,a) \in \mathbb{K}$ and $t \in \mathbb{N}$. Thus, using (32), we obtain

$$\begin{aligned}
&\sup_{x\in X} W(x)^{-1} \sup_{a\in A(x)} \left|\Phi(x,a) - \hat{\Phi}_t(x,a)\right| \\
&\qquad \le \|V^* - V_t\|_W + \alpha\|V^* - V_t\|_W \left(\beta + \tfrac{b}{\bar{w}}\right) + \tfrac{n_t}{\bar{w}} \\
&\qquad \le \left(1 + \alpha\left(\beta + \tfrac{b}{\bar{w}}\right)\right)\|V^* - V_t\|_W + \tfrac{n_t}{\bar{w}} \\
&\qquad \le \left(1 + \alpha\left(\beta + \tfrac{b}{\bar{w}}\right)\right)\tfrac{n_t(1+d)}{\bar{w}(1-\gamma)} + \tfrac{n_t}{\bar{w}} \\
&\qquad \le B_0 n_t \quad \text{a.s.}
\end{aligned} \tag{35}$$

where $B_0 := \left(1 + \alpha(\beta + \frac{b}{\bar{w}})\right)\left(\frac{1+d}{\bar{w}(1-\gamma)}\right) + \frac{1}{\bar{w}}$. Combining (34) and (35), we have

$$\Phi\left(x_t, \hat{\pi}_t(h_t)\right) \le B_0 W(x_t) n_t + \hat{\delta}_t \text{ a.s.}$$

Hence, to complete the proof of optimality of $\hat{\pi}$, it only remains to show that

$$E_x^{\hat{\pi}}\left(W(x_t)n_t\right) \to 0 \quad \text{as } t \to \infty. \tag{36}$$

To do this, first observe from (20) that $\sup_{t\ge 1} n_t \le B_1 < \infty$ for some constant B_1. Furthermore, since θ_t doesn't depend on $\hat{\pi}$ and x, from (20) we have

$$n_t \longrightarrow 0 \quad P_x^{\hat{\pi}} - \text{a.s.} \quad \text{as } t \to \infty, \tag{37}$$

whereas from Lemma 4.1(c),

$$\sup_{t\ge 1} E_x^{\hat{\pi}}\left(W(x_t)n_t\right)^p \le B_1^p \sup_{t\ge 1} E_x^{\hat{\pi}}\left(W^p(x_t)\right) < \infty.$$

Therefore, using a general result on the uniform integrability of sequences (see for example Lemma 7.6.9 in [1]), we conclude that $\{W(x_t)n_t\}$ is $P_x^{\hat{\pi}}$-uniformly integrable.

On the other hand, for arbitrary positive numbers ρ and l we have

$$P_x^{\hat{\pi}}\left(W(x_t)n_t > \rho\right) \le P_x^{\hat{\pi}}\left(n_t > \frac{\rho}{l}\right) + P_x^{\hat{\pi}}\left(W(x_t) > l\right).$$

Thus Chebyshev's inequality yields

$$P_x^{\hat{\pi}}\left(W(x_t)n_t > \rho\right) \le P_x^{\hat{\pi}}\left(n_t > \frac{\rho}{l}\right) + \frac{E_x^{\hat{\pi}}\left(W(x_t)\right)}{l},$$

which together with Lemma 4.1(c) and (37) gives that $\{W(x_t)n_t\}$ converges to zero in probability, i.e.

$$W(x_t)n_t \xrightarrow{P_x^{\hat{\pi}}} 0 \quad \text{as } t \to \infty. \tag{38}$$

Finally, the L^1 convergence (36) holds from (38) and the fact that $\{W(x_t)n_t\}$ is $P_x^{\hat{\pi}}$-uniformly integrable.

Optimality of $\bar{\pi}$. First observe that

$$\bar{V}_t = T_t\, \bar{V}_{t-1}, \;\; t \geq 1,$$

with T_t as in (25).

On the other hand, it is easy to see that under Assumption 3.2 there exists a positive constant $\bar{C}$ such that, for all $t \in \mathbb{N}$,

$$\left\|\bar{V}_t\right\|_W \leq \bar{C}\ . \tag{39}$$

Now from (25), (27)-(31) and (13), we have

$$\left\|V^* - \bar{V}_{t+1}\right\|_{\bar{W}} \leq \frac{n_t}{\bar{w}} + \gamma \left\|V^* - \bar{V}_t\right\|_{\bar{W}} \;\; \text{a.s.} \tag{40}$$

Letting $\lambda := \limsup_{t\to\infty} \left\|V^* - \bar{V}_t\right\|_{\bar{W}} < \infty$ (see Proposition 3.4 (a) and (39)) and taking the limit supremum in both sides of (40), from (20) we have that $\lambda \leq \gamma\lambda$, which implies (since $0 < \gamma < 1$) that $\lambda = 0$. Thus (see (26)) $\lim_{t\to\infty} \left\|V^* - \bar{V}_t\right\|_W = 0$ a.s.

Now, defining

$$\bar{\Phi}_t\,(x,a) := c(x,a) + \alpha \int_S \bar{V}_{t-1}\,(F(x,a,s))\,\theta_t(ds) - \bar{V}_t\,(x),\; x \in X,$$

from (13), (14) and (15), we get (see (34))

$$\Phi\,(x, \bar{\pi}_t\,(h_t)) \leq W(x_t) \sup_{x\in X} W(x)^{-1} \sup_{a\in A(x)} \left|\Phi(x,a) - \bar{\Phi}_t\,(x,a)\right| + \bar{\delta}_t\ . \tag{41}$$

Moreover, for some constants $\bar{B}_0$ and $\bar{B}_1$ (see (35)),

$$\sup_{x\in X} W(x)^{-1} \sup_{a\in A(x)} \left|\Phi(x,a) - \bar{\Phi}_t\,(x,a)\right| \leq \bar{B}_0 \left\|V^* - \bar{V}_{t-1}\right\|_W + \bar{B}_1\,(n_t + n_{t-1}) := \bar{n}_t \;\; \text{a.s.} \tag{42}$$

Combining (41) and (42) we obtain $\Phi\left(x_t, \bar{\pi}_t\left(h_t\right)\right) \leq W(x_t)\, \bar{n}_t + \bar{\delta}_t$ a.s. The convergence $E_x^{\bar{\pi}}\left(W(x_t) n_t\right) \to 0 \quad \text{as } t \to \infty$ is then proved as in (36)-(38). $\diamond$

Before proving part c) of Theorem 3.8, let us give a last technical result:

Lemma 4.5 *Under Assumptions 3.2 and 3.7, we have, for each $x \in X$ and $a \in A(x)$*

$$\lim_{t\to\infty} \left| \int_S V_t\left(F(x,a,s)\right) \theta_t(ds) - \int_S V^*\left(F(x,a,s)\right) \theta(ds) \right| = 0 \quad a.s.$$

Proof: Observe that

$$\begin{aligned}
&\int_S V_t\left(F(x,a,s)\right)\theta_t(ds) - \int_S V^*\left(F(x,a,s)\right)\theta(ds) \\
&\qquad = \int_S \Big[V_t\left(F(x,a,s)\right) - V^*\left(F(x,a,s)\right) \Big] \theta_t(ds) \\
&\qquad\quad + \int_S V^*\left(F(x,a,s)\right)\theta_t(ds) - \int_S V^*\left(F(x,a,s)\right)\theta(ds) \\
&\qquad \leq \left\| V_t - V^* \right\|_W \left(\beta W(x) + b\right) \\
&\qquad\quad + \left| \int_S V^*\left(F(x,a,s)\right)\theta_t(ds) - \int_S V^*\left(F(x,a,s)\right)\theta(ds) \right|.
\end{aligned}$$

The last inequality follows from (18). Thus, Lemma 4.5 holds thanks to part a) of Theorem 3.8 and (20). $\diamond$

Proof of part c) of Theorem 3.8 For each $x \in X$, $\hat{f}_\infty(x)$ is an accumulation point of $\{\hat{f}_t(x)\}$. That is to say, for each $x \in X$, there is a subsequence $\{t_i(x)\}$ of $\{t\}$ such that

$$\hat{f}_{t_i(x)}(x) \to \hat{f}_\infty(x) \quad \text{as } i \to \infty.$$

Now we fix an arbitrary $x \in X$ and replace t with $t_i(x)$ in (12). We get

$$c(x, \hat{f}_{t_i(x)}) + \alpha \int_S V_{t_i(x)}(F(x, \hat{f}_{t_i(x)}, s))\theta_{t_i(x)}(ds) \leq V_{t_i(x)}(x) + \delta_{t_i(x)} \text{ a.s.} \quad (43)$$

Before taking the limit infimum in (43), first note that

$$\liminf_{i\to\infty} \int_S V_{t_i(x)}\left(F(x, \hat{f}_{t_i(x)}, s)\right) \theta_{t_i(x)}(ds) \geq \int_S V^*\left(F(x, \hat{f}_\infty, s)\right) \theta(ds). \quad (44)$$

Indeed,

$$\int_S V_{t_i(x)}\left(F(x,\hat{f}_{t_i(x)},s)\right)\theta_{t_i(x)}(ds)$$
$$=\left[\int_S V_{t_i(x)}\left(F(x,\hat{f}_{t_i(x)},s)\right)\theta_{t_i(x)}(ds)-\int_S V^*\left(F(x,\hat{f}_{t_i(x)},s)\right)\theta(ds)\right]$$
$$+\int_S V^*\left(F(x,\hat{f}_{t_i(x)},s)\right)\theta(ds),$$

which yields, by Lemma 4.5, that

$$\liminf_{i\to\infty}\int_S V_{t_i(x)}\left(F(x,\hat{f}_{t_i(x)},s)\right)\theta_{t_i(x)}(ds)\geq\liminf_{i\to\infty}\int_S V^*\left(F(x,\hat{f}_{t_i(x)},s)\right)\theta(ds).$$

Thus (44) follows by applying Fatou's Lemma and as function F is continuous in $a\in A(x)$.

Now, taking the limit infimum in (43), we obtain

$$c(x,\hat{f}_\infty)+\alpha\int_S V^*(F(x,\hat{f}_\infty,s))\theta(ds)\leq V^*(x),\tag{45}$$

and so, by (10), equality holds in (45). In fact, as x was arbitrary, equality holds in (45) for every $x\in X$, and therefore, $\hat{f}_\infty$ is α-discount optimal for $\mathcal{M}$. $\diamond$

5 Examples

5.1 Sufficient sets of conditions for Assumption 3.7

An obvious sufficient condition for Assumption 3.7 is that the disturbance set S is countable (with the discrete topology). We next present other, less obvious sufficient conditions.

Assumption 5.1 *a)* $(X,||.||)$ *is a complete, separable, normed vector space.*
b) The function $V_W^*(x)=\frac{V^*(x)}{W(x)}$ *is convex.*
c)The family of functions $\{F(x,a,.):(x,a)\in I\!K\}$ *is equicontinuous on* S.

Proposition 5.2 *Under Assumptions 3.2 and 5.1, Assumption 3.7 –and therefore Theorem 3.8– holds.*

Proof: From [6], Assumptions 5.1(a), (b) imply that V_W^* is Lipschitz: there exists a constant L such that, for all x_1 and x_2 in X,

$$|V_W^*(x_1) - V_W^*(x_2)| \leq L \|x_1 - x_2\|.$$

Let $\varepsilon > 0$. By Assumption 5.1 (c), there exists $\delta > 0$ such that $d_S(s_1, s_2) < \delta$ implies

$$\|F(x, a, s_1) - F(x, a, s_2)\| < \varepsilon, \text{ for all } (x, a) \in \mathbb{K},$$

where d_S is the metric on S. Then, for all $(x, a) \in \mathbb{K}$,

$$|V_W^*(F(x, a, s_1)) - V_W^*(F(x, a, s_2))| \leq L \|F(x, a, s_1) - F(x, a, s_2)\| \leq L\varepsilon.$$

Thus Assumption 3.7 is verified and Theorem 3.8 follows. $\diamond$

We define the weighted total variation norm of a signed measure m on $\mathcal{B}(X)$ as follows:

$$\|m\|_{T_W} := \int_X W(x) |m| (dx),$$

where $|m|$ denotes the variation of the measure m.

Assumption 5.3 *a) The family of functions $\{F(x, a, .) : (x, a) \in \mathbb{K}\}$ is equicontinuous on S.*
b) Let d_X be the metric on X. There exist three constants L_0, L_1 and L_2 such that for every (x_1, a_1), $(x_2, a_2) \in \mathbb{K}$ the following holds:

$$\begin{aligned} \left| \frac{c(x_1, a_1)}{W(x_1)} - \frac{c(x_2, a_2)}{W(x_2)} \right| &\leq L_0 d_X(x_1, x_2), \\ \|Q(dy|x_1, a_1) - Q(dy|x_2, a_2)\|_{T_W} &\leq L_1 d_X(x_1, x_2), \\ |W(x_1) - W(x_2)| &\leq L_2 d_X(x_1, x_2). \end{aligned} \tag{46}$$

Proposition 5.4 *Under Assumptions 3.2 and 5.3, Assumption 3.7 –and therefore Theorem 3.8– holds.*

Proof: First observe that for all $x \in X$ and $\pi \in \Pi$,

$$\begin{aligned} V(\pi, x) :=& E_x^\pi \left[\sum_{t=0}^{\infty} \alpha^t c(x_t, a_t) \right] \\ =& \int_A \left[c(x, a) + \alpha \int_X V(\pi', y) Q(dy|x, a) \right] \pi_0(da|x), \end{aligned}$$

where $\pi' = \{\pi'_t\}$ is the "shifted" policy $\pi'_t(\cdot|h_t) := \pi_{t+1}(\cdot|x, a, h_t)$ for all $t = 0, 1, ...$ Thus, for arbitrary x_1 and x_2 in X, we have

$$\begin{aligned}\left|\frac{V(\pi,x_1)}{W(x_1)} - \frac{V(\pi,x_2)}{W(x_2)}\right| &= \left|\int_A \left[\frac{c(x_1,a)}{W(x_1)} + \alpha\int_X \frac{V(\pi',y)}{W(x_1)}Q(dy|x_1,a)\right]\pi_0(da|x_1)\right. \\ &\quad \left. - \int_A \left[\frac{c(x_2,a)}{W(x_2)} + \alpha\int_X \frac{V(\pi',y)}{W(x_2)}Q(dy|x_2,a)\right]\pi_0(da|x_2)\right| \\ &\leq \sup_{a_1,a_2}\left|\frac{c(x_1,a_1)}{W(x_1)} - \frac{c(x_2,a_2)}{W(x_2)}\right| \\ &\quad + \sup_{a_1,a_2}\left|\int_X \frac{V(\pi',y)}{W(x_1)}Q(dy|x_1,a_1) - \int_X \frac{V(\pi',y)}{W(x_2)}Q(dy|x_2,a_2)\right|, \end{aligned} \tag{47}$$

where the supremum is over all $a_1 \in A(x_1)$ and $a_2 \in A(x_2)$.

Now, for each $a_1 \in A(x_1)$ and $a_2 \in A(x_2)$, adding and subtracting the term $\int_X \frac{V(\pi',y)}{W(x_2)}Q(dy|x_1,a_1)$, we have

$$\begin{aligned} I &:= \left|\int_X \frac{V(\pi',y)}{W(x_1)}Q(dy|x_1,a_1) - \int_X \frac{V(\pi',y)}{W(x_2)}Q(dy|x_2,a_2)\right| \\ &\leq \int_X \left|\frac{V(\pi',y)}{W(x_1)} - \frac{V(\pi',y)}{W(x_2)}\right| Q(dy|x_1,a_1) \\ &\quad + \int_X \frac{V(\pi',y)}{W(x_2)}\left|Q(dy|x_1,a_1) - Q(dy|x_2,a_2)\right| \\ &\leq \left|\frac{1}{W(x_1)} - \frac{1}{W(x_2)}\right| \int_X V(\pi',y)Q(dy|x_1,a_1) \\ &\quad + \frac{1}{W(x_2)}\int_X V(\pi',y)\left|Q(dy|x_1,a_1) - Q(dy|x_2,a_2)\right|. \end{aligned}$$

Moreover, from Proposition 3.4 (a), (19) and Assumption 5.3 (b),

$$\begin{aligned}
I \leq & \frac{C\,|W(x_1)-W(x_2)|}{(1-\alpha)W(x_1)W(x_{2)}}\int_X W(y)Q(dy|x_1,a_1) \\
& + \frac{C/(1-\alpha)}{W(x_2)}\int_X W(y)\,|Q(dy|x_1,a_1)-Q(dy|x_2,a_2)| \\
\leq & \frac{C\,|W(x_1)-W(x_2)|}{(1-\alpha)W(x_1)W(x_{2)}}(\beta W(x_1)+b) \\
& + \frac{C/(1-\alpha)}{W(x_2)}\,\|Q(\cdot|x_1,a_1)-Q(\cdot|x_2,a_2)\|_{T_W} \\
\leq & \frac{C\beta\,|W(x_1)-W(x_2)|}{(1-\alpha)W(x_{2)}} + \frac{Cb\,|W(x_1)-W(x_2)|}{(1-\alpha)W(x_1)W(x_{2)}} \\
& + \frac{C/(1-\alpha)}{W(x_2)}\,\|Q(\cdot|x_1,a_1)-Q(\cdot|x_2,a_2)\|_{T_W} \\
\leq & \frac{C\beta L_2}{(1-\alpha)\,\bar{w}}d(x_1,x_2) + \frac{CbL_2}{(1-\alpha)\,\bar{w}^2}d_X(x_1,x_2) + \frac{C}{\bar{w}\,(1-\alpha)}L_1 d_X(x_1,x_2) \\
= & L' d_X(x_1,x_2), \qquad (48)
\end{aligned}$$

where $L' := \frac{C}{\bar{w}(1-\alpha)}\left[L_2\left(\beta+\frac{b}{\bar{w}}\right)+L_1\right]$.

Combining (47) and (48), and by Assumption 5.3 (b),

$$\left|\frac{V(\pi,x_1)}{W(x_1)} - \frac{V(\pi,x_2)}{W(x_2)}\right| \leq L_0 d(x_1,x_2) + L' d(x_1,x_2) = L^* d(x_1,x_2), \qquad (49)$$

where $L^* := L_0 + L_1$.

On the other hand, from (16) we have

$$\begin{aligned}
|V_W^*(x_1) - V_W^*(x_2)| &= \left|\frac{\inf_{\pi\in\Pi} V(\pi,x_1)}{W(x_1)} - \frac{\inf_{\pi\in\Pi} V(\pi,x_2)}{W(x_2)}\right| \\
&\leq \sup_{\pi\in\Pi}\left|\frac{V(\pi,x_1)}{W(x_1)} - \frac{V(\pi,x_2)}{W(x_2)}\right|,
\end{aligned}$$

which, together with (49) yields

$$|V_W^*(x_1) - V_W^*(x_2)| \leq L^* d(x_1,x_2).$$

Hence, V_W^* is Lipschitz. The proof is now completed as in the proof of Proposition 5.2. $\diamond$

5.2 An inventory-production system

We consider an inventory-production system of the form

$$x_{t+1} = (x_t + a_t - \xi_t)^+, \quad t = 0, 1, ..., \tag{50}$$

x_0 given, with state space $X = [0, \infty)$ and action set $A(x) = A = [0, a^*]$ for all $x \in X$, for some $a^* > 0$. In addition the random variables $\xi_0, \xi_1, ...,$ are i.i.d, having a discrete distribution with values in $S = \{0, 1, 2, ...\}$, and satisfying that

$$P[\xi_0 > a^*] = 1. \tag{51}$$

In (50), x_t represents the stock level at the beginning of period t, the control a_t is the quantity ordered or produced at the beginning of period t, and the random variable ξ_t is the demand during that period.

In general, for certain class of inventory systems, we can take the one-stage cost function of the form (see, for instance, [2, 5, 22]):

$$c(x, a) = G(x + a) + c_0 a, \ (x, a) \in \mathbb{K}, \tag{52}$$

where $c_0 > 0$ is the unit production (or purchasing) cost, and G is a convex function such that $\lim_{y\to\infty} G(y) = +\infty$, representing the cost for excess inventory and the holding cost. In the context of our example, we suppose in addition that the cost in (52) satisfies

$$\sup_{a\in A} c(x, a) \leq \bar{b}\, e^{\lambda x}, \text{ for all } x \in X,$$

where $\bar{b}$ and λ are arbitrary positive constants.

Clearly, the Assumptions 3.2 (a) and 3.7 are satisfied. The Assumptions 3.2 (b) and (c) follows taking $W(x) := \bar{b}\, e^{\lambda x}$ and from the following relations: for some $p > 1$, and all $x \in X, \ a \in A$,

$$\begin{aligned}
\frac{1}{t}\sum_{i=0}^{t-1} \bar{b}\, e^{p\lambda(x+a-\xi_i)^+} &= \frac{1}{t}\sum_{i=0}^{t-1} \bar{b}\, e^{p\lambda(x+a-\xi_i)^+} 1_{[\xi_i \geq x+a]} + \frac{1}{t}\sum_{i=0}^{t-1} \bar{b}\, e^{p\lambda(x+a-\xi_i)^+} 1_{[\xi_i < x+a]} \\
&\leq \bar{b} + \frac{\bar{b}\, e^{p\lambda x}}{t} \sum_{i=0}^{t-1} e^{p\lambda(a^*-\xi_i)} 1_{[\xi_i < x+a]} \\
&\leq \beta_0\, \bar{b}\, e^{p\lambda x} + b_0 \quad \text{a.s.},
\end{aligned}$$

where $b_0 := \bar{b}$ and $\beta_0 := \frac{1}{t}\sum_{i=0}^{t-1} e^{p\lambda(a^*-\xi_i)}$. Note that $\beta_0 < 1$ by (51).

References

[1] Ash R.B. (1972) Real Analysis and Probability. Academic Press, New York.

[2] Bertsekas D.P. (1987) Dynamic Programming: Deterministic and Stochastic Models. Prentice-Hall, Englewood Cliffs, N.J.

[3] Billingsley P. and Topsoe F. (1967) Uniformity in weak convergence. Z. Wahrsch. Verw. Geb. 7: 1-16.

[4] Cavazos-Cadena R. (1990) Nonparametric adaptive control of discounted stochastic systems with compact state space. J. Optim. Theory Appl.., 65: 191-207.

[5] Dynkin E.B. and Yushkevich A.A. (1979) Controlled Markov Processes. Springer-Verlag, New York.

[6] Fernández-Gaucherand E. (1994) A note on the Ross-Taylor Theorem. Applied Mathematics and Computation 64: 207–212.

[7] Gaenssler P. and Stute W. (1979) Empirical processes: a survey for i.i.d. random variables, Ann. Probab. 7: 193–243.

[8] Gordienko E.I. and Minjárez-Sosa J.A. (1998) Adaptive control for discrete-time Markov processes with unbounded costs: discounted criterion. Kybernetika 34: 217–234.

[9] Hernández-Lerma O. (1989) Adaptive Markov Control Processes. Springer-Verlag, New York.

[10] Hernández-Lerma O. and Cavazos-Cadena R. (1990) Density estimation and adaptive control of Markov processes: average and discounted criteria. Acata Appl. Math., 20: 285-307.

[11] Hernández-Lerma O. and Lasserre J.B. (1996) Discrete-Time Markov Control Processes: Basic Optimality Criteria. Springer-Verlag, New York.

[12] Hernández-Lerma O. and Lasserre J.B. (1999) Further Topics on Discrete-Time Markov Control Processes. Springer-Verlag, New York.

[13] Hernández-Lerma O. and Marcus S.I. (1985) Adaptive control of discounted Markov decision chains. J. Optim. Theory Appl., 46: 227-235.

[14] Hilgert N. and Minjárez-Sosa J.A. (2001) Adaptive policies for time-varying stochastic systems under discounted criterion. Math. Methods Oper. Res., 54, 3: 491-505.

[15] Hilgert N. and Hernández-Lerma O. (2000) Limiting optimal discounted-cost control of a class of time-varying stochastic systems. Syst. Control Lett., 40: 37–42.

[16] Kurano M. (1972) Discrete-time markovian decision processes with an unknown parameter-average return criterion, J. Oper. Res. Soc. Japan, 15: 67-76.

[17] Mandl P. (1974) Estimation and control in Markov chains. Adv. Appl. Probab. 6: 40–60.

[18] Rieder U. (1978) Measurable selection theorems for optimization problems. Manuscripta Math. 24: 115–131.

[19] Schäl M. (1975) Conditions for optimality and for the limit of n-stage optimal policies to be optimal. Z. Wahrs. Verw. Gerb. 32: 179–196.

[20] Schäl M. (1987) Estimation and control in discounted stochastic dynamic programming. Stochastics 20: 51–71.

[21] Van Nunen J.A.E.E. and Wessels J. (1978) A note on dynamic programming with unbounded rewards. Manag. Sci. 24: 576–580.

[22] Vega-Amaya O. and Montes-de-Oca R. (1998) Application of average dynamic programming to inventory systems. Math. Methods Oper. Res., 47: 451-471.