SIAM J. CONTROL OPTIM.
Vol. 35, No. 6, pp. 2070–2085, November 1997

010

ASYMPTOTIC OPTIMIZATION OF A NONLINEAR HYBRID SYSTEM GOVERNED BY A MARKOV DECISION PROCESS*

EITAN ALTMAN† AND VLADIMIR GAITSGORY‡

Abstract. We consider in this paper a continuous time stochastic hybrid control system with finite time horizon. The objective is to minimize a nonlinear function of the state trajectory. The state evolves according to a nonlinear dynamics. The parameters of the dynamics of the system may change at discrete times $l\epsilon$, $l = 0, 1, ...$, according to a controlled Markov chain which has finite state and action spaces. Under the assumption that ϵ is a small parameter, we justify an averaging procedure allowing us to establish that our problem can be approximated by the solution of some deterministic optimal control problem.

Key words. hybrid stochastic systems, asymptotic optimality, nonlinear dynamics, Markov decision processes, averaging

AMS subject classifications. 49B10, 49B50

PII. S0363012995279985

1. Introduction and statement of the problem.

Consider the following hybrid stochastic control system. The state $Z_t \in \mathbb{R}^n$ evolves according to the following dynamics:

$$\frac{d}{dt}Z_t = f(Z_t, Y_t), \qquad t \in [0,1], \qquad Z_0 = z, \tag{1}$$

where $Y_t \in \mathbb{R}^k$ is the "control" to be specified later and z is the initial state. f is assumed to be linear in the second argument (for each value of the first argument), i.e.,

$$f(z, y) = f^1(z) + f^2(z)y, \tag{2}$$

where f^1 is an n-dimensional vector and f^2 is an $n \times k$ matrix; $f^2(z)y$ is the multiplication between the matrix $f_2(z)$ and the vector y. The functions $f^1(z)$ and $f^2(z)$ are supposed to be bounded and to satisfy the Lipschitz condition

$$\left\| f^i(z) - f^i(z') \right\|_1 \le C_1 \left\| z - z' \right\|_1 \qquad \forall z, z', \tag{3}$$

$$\left\| f^i(z) \right\|_1 \le C_2, \tag{4}$$

where z, z' are from a sufficiently large domain which contains all possible trajectories of (1), C_1 and C_2 are constants, and $\|\cdot\|_1$ stands for the L_1 norm in the finite-dimensional space. That is, $\|q\|_1 = \max_{i=1,...,k} |q_i|$ for the vector $q = \{q_i\}, i = 1, ..., k$, and $\|A\|_1 = \max_{\|q\|_1=1} \|Aq\|_1$ for the matrix $A(n \times k)$.

It is assumed in what follows that there exists a bounded domain containing all the trajectories of (1), and, thus, (4), in fact, is implied by (3).

---

*Received by the editors January 13, 1995; accepted for publication (in revised form) September 11, 1996. The research undertaken in this paper was supported by the Australian Research Council (ARC).

http://www.siam.org/journals/sicon/35-6/27998.html

†INRIA, BP93, 2004 Route des Lucioles, 06902 Sophia Antipolis Cedex, France (altman@martingale.inria.fr).

‡School of Mathematics, University of South Australia, The Levels, Pooraka, South Australia 5095, Australia (mavg@lux.levels.unisa.edu.au).

Y_t is not chosen directly by the controller but is obtained as a result of controlling the following underlying stochastic discrete event system. Let ϵ be the basic time unit. Time is discretized; i.e., transitions occur at times $t = n\epsilon,\ n = 0, 1, 2, ..., \lfloor \epsilon^{-1} \rfloor$, where $\lfloor x \rfloor$ stands for the greatest integer which is smaller than or equal to x. There is a finite state space $\mathbf{X} = \{1, ..., N\}$ and a finite action space $\mathbf{A}$. If a state is v and an action a is chosen, then the next state is w with the probability P_{vaw}. A policy $u = \{u_0, u_1, ...\}$ in the set of policies U is a sequence of probability measures on $\mathbf{A}$; at each time $t = n\epsilon$ the controller chooses u_n based on the history of all previous states and actions, as well as the present state. Thus, u_n is a function that maps histories of the form $h_n = (x_0, a_0, x_1, a_1, ..., x_{n-1}, a_{n-1}, x_n)$ to probability measures on $\mathbf{A}$.

We shall be especially interested in the following classes of policies:

- the Markov policies, denoted by $\mathcal{M}$, i.e., policies for which u_t depends only on the current state and does not depend on previous states and actions.
- the stationary policies, denoted by $\mathcal{S}$, i.e., policies for which u_t depends only on the current state and does not depend on previous states and actions nor on the time.

The stochastic process $\{X_n, A_n\}$ is known as a controlled Markov chain, or Markov decision process (MDP); see Derman [11, pp. 2–4]. We assume throughout the paper that under any stationary policy, the state space forms an aperiodic Markov chain such that all states communicate (regular Markov chain). The results of the paper hold, in fact, under weaker ergodicity assumptions; however, the restricted assumption makes the presentation clearer.

Denote by $\mathbf{H}$ the set of all possible states and actions histories which can be observed until time $\lfloor \epsilon^{-1} \rfloor$:

$$\mathbf{H} = \bigcup \{h\}, \qquad h = \left\{(x_n, a_n), n = 0, 1, ..., \lfloor \epsilon^{-1} \rfloor\right\}.$$

Let $\mathcal{F}$ be the σ-algebra of all subsets of $\mathbf{H}$. Each policy u and initial state x determines a probability measure on $\mathcal{F}$, on which the stochastic state and action process $H = \left\{X_n, A_n, n = 0, 1, ..., \lfloor \epsilon^{-1} \rfloor\right\}$ is defined. Denote by P_x^u and E_x^u the probability measure and mathematical expectation that correspond to an initial state $X_0 = x$ and a policy u. Sometimes we shall assume an initial distribution ξ on X_0, instead of a fixed initial state. In that case P_ξ^u, E_ξ^u denote the corresponding probability measure and mathematical expectation.

Let $y : \mathbf{X} \times \mathbf{A} \to \mathbb{R}^k$, $j = 1, ..., k$, be some given vector-valued function. Then Y_t in (1) is given by

$$Y_t = y(X_{\lfloor t/\epsilon \rfloor}, A_{\lfloor t/\epsilon \rfloor}). \tag{5}$$

The system (1) with thus-defined Y_t is called hybrid, first, because Y_t changes its values via some random jumps whereas Z_t is a smooth (differentiable) function of time and, second, because, as follows from the consideration below, Y_t being controlled "statistically" through controlling the transition probabilities plays by itself the role of a "direct" control with respect to Z_t.

Let $g : \mathbb{R}^n \to \mathbb{R}$ be some operating cost related to the process Z_t. We assume that it is Lipschitz continuous; i.e.,

$$\|g(z) - g(z')\|_1 \leq C_1 \|z - z'\|_1 .$$

We consider the following control problem with ϵ and x fixed.

$\mathbf{Q}_\epsilon$: find a policy u that achieves $F^\epsilon(z, x) = \inf_{u \in U} E_x^u g(Z_1)$, where Z_1 is obtained through (1).

Our model is characterized by the fact that ϵ is supposed to be a small parameter and our objective is to construct a policy (depending, in general, on ϵ) which is asymptotically optimal for $\mathbf{Q}_\epsilon$. That is, the difference between the cost under this policy and $F^\epsilon(z, x)$ converges to zero as $\epsilon \to 0$.

The type of model which we introduce is natural in the control of inventories or of production, where we deal with material whose quantity may "slowly" change in a continuous (linear) way. Breakdowns, repairs, and other control decisions yield the underlying MDP. Our model may also be used in the control of highly loaded queueing networks for which the fluid approximation holds (see Kleinrock [20, p. 56]). The slow variables Z_t may then represent the number of customers in the different queues, whereas the underlying MDP may correspond to routing, or flow control of, say, some long on/off traffic.

The fact that ϵ is chosen to be small means that the variables Y_t along with the MDP X_t can be considered to be fast with respect to the time scale t in which Z_t evolves. Indeed, Y_t and X_t may have large jumps between $t = m\epsilon$ and $t = (m+1)\epsilon$, whereas the corresponding change in Z_t in that period is of order ϵ. The problem is, thus, close in nature to stochastic singular perturbed control problems intensively studied in the literature (see, for example, [1], [5], [6], [7], [9], [10], [21], [23], [24], [25] and references therein). A common approach to this kind of problem is an application of singular perturbations or averaging techniques to the Hamilton–Jacobi–Bellman (HJB) equation for problems in continuous time (as in [5], [6], [21]) or to the dynamic programming equation for singularly perturbed MDPs [1], [7], [9], [10], [24], [25]. In contrast to this approach, we, as in [23], apply an averaging method directly to the "slow" stochastic equation. Our model differs, however, from the ones in [23] in many respects—mainly in the type of fast motions involved, which implies the differences in both the technique used and the results obtained.

In our previous paper [2], we considered the problem similar to $\mathbf{Q}_\epsilon$ for the case of linear dynamics f and cost g and showed that an asymptotically optimal policy can be constructed via maximization of the Hamiltonian of some linear deterministic system. The technique we used was, however, strongly related to the linearity of the model, and it is not applicable to the case when the dynamics and/or the cost are nonlinear. As opposed to the linear case, the consideration for the nonlinear case is much more involved and based on an ergodicity-type result for MDPs obtained in this paper (see Theorem 4.1 below). Using this result we establish that the trajectories of stochastic hybrid system (1) are approximated by the trajectories of some nonlinear deterministic control system, and the problem $\mathbf{Q}_\epsilon$ is approximated by the corresponding deterministic optimal control problem allowing us, in particular, to construct an asymptotically optimal policy for $\mathbf{Q}_\epsilon$. Notice that this result can be viewed as an extension of the averaging technique for deterministic singularly perturbed control systems (see, e.g., [15]) to the stochastic case under consideration. On the other hand, it can be viewed as an extention of results on uncontrolled motions establishing that the solution of the original stochastic system is approximated by the solution of some deterministic system obtained via averaging over the fast random dynamics [16], [19], [22] to the case when this random dynamics is defined by the controlled Markov chain.

The paper consists of four sections. Section 1 is this introduction; section 2 describes the main results about the approximation of the problem of optimal control of the hybrid system by a deterministic optimal control problem. In section 3 we discuss ways that the solution of the deterministic optimal control problem can be characterized and how it can be used to obtain an asymptotically optimal policy. Section 4 contains the above-mentioned Theorem 4.1, as well as the proofs of some basic lemmas used in section 2.

2. Description of main results. Let

$$\mathbf{Y}(m,x) \stackrel{\text{def}}{=} \bigcup_{u\in U}\left\{(m+1)^{-1}\sum_{t=0}^{m} E_x^u Y_t\right\},$$

where the union is taken over all policies. As follows from Theorem 3 in [2], the set $\mathbf{Y}(m,x)$ converges in the Haussdorff metric to a set $\mathbf{Y}$ defined below:

$$\lim_{m\to\infty}\mathbf{Y}(m,x) = \mathbf{Y} \stackrel{\text{def}}{=} \bigcup_{u\in\mathcal{S}}\left\{\sum_{v,a}\eta(u;v,a)y(v,a)\right\}, \tag{6}$$

where the union is taken over all stationary policies, and $\eta(u) = \{\eta(u;v,a)\}$ is the vector of steady state probabilities of state-action pairs obtained when using a stationary policy u. That is,

$$\eta(u;v,a) = \lim_{n\to\infty} P_x^u(X_n = v, A_n = a). \tag{7}$$

Notice that due to the ergodicity assumption on our model, $\eta(u;v,a)$ does not depend on the initial distribution. Notice also that, since the set

$$W \stackrel{\text{def}}{=} \bigcup_{u\in\mathcal{S}}\{\eta(u)\} \tag{8}$$

is a polyhedron (see, for example, [11, pp. 93–95]), the set $\mathbf{Y}$ is a polyhedron as well.

Define now the averaged deterministic control system as

$$\frac{d}{dt}z_t = f(z_t, y_t), \qquad z_0 = z, \tag{9}$$

where y_t is a measurable function of t taking values in $\mathbf{Y}$. The set of such functions

$$y : [0,1] \to \mathbf{Y}$$

will be called the set of admissible controls.

Our claim is that the set of all random trajectories of (1) is approximated by the set of solutions of (9) obtained with all admissible controls. More specifically, we establish that there exists a function $\gamma(\epsilon)$ satisfying

$$\lim_{\epsilon\to 0}\gamma(\epsilon) = 0$$

such that the following holds.

LEMMA 2.1. *Corresponding to any admissible control $y = \{y_t, t \in [0,1]\}$, there exists a Markov policy $u_\epsilon(y)$ such that the random trajectory Z_t of (1), obtained with this policy $u_\epsilon(y)$, and the deterministic solution z_t^y of (9), obtained with y, satisfy the inequality*

$$\max_{t\in[0,1]} E_x^{u_\epsilon(y)}\|Z_t - z_t^y\|_1 \le \gamma(\epsilon). \tag{10}$$

LEMMA 2.2. *There exists a function* $\tilde{y}_t^\epsilon(h)$,

$$\tilde{y}^\epsilon : [0,1] \times \mathbf{H} \to \mathbf{Y},$$

such that (a) *for each* $h \in \mathbf{H}$, $\tilde{y}_t^\epsilon(h)$ *is a piecewise constant function of* t *and* (b) *for any policy* u,

$$\max_{t\in[0,1]} E_x^u \left\| Z_t - \tilde{z}_t^\epsilon(H) \right\|_1 \le \gamma(\epsilon), \tag{11}$$

where Z_t *is the solution of* (1), $\tilde{z}_t^\epsilon(H)$ *is the solution of* (9) *obtained with* $y_t = \tilde{y}_t^\epsilon(H)$, *and* H *is the random realization of the state-action trajectories.*

Notice that the quantity under the expectation sign in (11) is a random variable for any policy u since $\mathbf{H}$ is a finite set and $\mathcal{F}$ is the σ-algebra of all subsets of $\mathbf{H}$.

Notice also that a construction of a policy $u_\epsilon(y)$ which allows an estimate (10) in Lemma 2.1 is described below in section 3. This is just a stationary policy when the deterministic control y is a constant function of time, and it consists of a finite number of stationary policies (and thus is not stationary itself) when y is piecewise constant.

Define the "deterministic" optimal control problem $\mathbf{Q_0}$ as follows.

$\mathbf{Q_0}$: Find an admissible control y which minimizes the cost function

$$F^0(z) \stackrel{\text{def}}{=} \inf_y g(z_1)$$

over the trajectories z of system (9). The following theorem about approximation of $\mathbf{Q}_\epsilon$ by $\mathbf{Q_0}$ is then easily established on the basis of Lemmas 2.1 and 2.2.

THEOREM 2.1. *The values* $F^\epsilon(z,x)$ *of the original problem* $\mathbf{Q}_\epsilon$ *converge to the value* $F^0(z)$ *of the problem* $\mathbf{Q_0}$, *as* $\epsilon \to 0$. *More precisely,*

$$\left| F^\epsilon(z,x) - F_x^0(z) \right| \le C_1 \gamma(\epsilon).$$

If y^* *is an optimal control for* $\mathbf{Q_0}$, *then the Markov policy* $u_\epsilon(y^*)$ *allowing estimate* (10) *with* $y = y^*$ *satisfies the inequality*

$$\left| E_x^{u_\epsilon(y^*)} g(Z_1) - F^\epsilon(z,x) \right| \le C_1 \gamma(\epsilon).$$

That is, $u_\epsilon(y^*)$ *is asymptotically optimal for* $\mathbf{Q}_\epsilon$.

Remark 2.1. In the linear case studied in [2], γ can be chosen such that

$$\lim_{\epsilon\to 0} \epsilon^{-(1/2)} \gamma(\epsilon) = 0.$$

Hence, for the linear case, simple bounds on the rate of convergence are available for Lemmas 2.1 and 2.2 as well as for Theorem 2.1.

Proof of Theorem 2.1. Let u be an arbitrary policy and $\tilde{y}^\epsilon(h) \in \mathbf{Y}$ be the function defined in Lemma 2.2. Then

$$\left| E_x^u g(Z_1) - E_x^u g(\tilde{z}_1^\epsilon(H)) \right| \le C_1 E_x^u \left\| Z_1 - \tilde{z}_1^\epsilon(H) \right\|_1 \le C_1 \gamma(\epsilon), \tag{12}$$

where C_1 is defined in (3). Being piecewise constant, the function $\tilde{y}^\epsilon$ is measurable in t. Hence,

$$g(\tilde{z}_1^\epsilon(h)) \ge F^0(z) \qquad \forall h \in \mathbf{H},$$

which implies

$$E_x^u g(\tilde{z}_1^\epsilon(H)) \geq F^0(z)$$

for any policy u. From the last inequality and (12), it follows that

$$E_x^u g(Z_1) \geq F^0(z) - C_1\gamma(\epsilon),$$

so that

$$F^\epsilon(z,x) = \inf_u E_x^u g(Z_1) \geq F^0(z) - C_1\gamma(\epsilon). \tag{13}$$

Now let y^* be an optimal control in $\mathbf{Q_0}$. By (10),

$$\left| E_x^{u_\epsilon(y^*)} g(Z_1) - F^0(z) \right| = \left| E_x^{u_\epsilon(y^*)} g(Z_1) - g(z_1^{y^*}) \right| \leq C_1 E_x^{u_\epsilon(y)} \left\| Z_1 - z_1^{y^*} \right\|_1 \leq C_1\gamma(\epsilon).$$

Hence

$$E_x^{u_\epsilon(y^*)} g(Z_1) \leq F^0(z) + C_1\gamma(\epsilon). \tag{14}$$

Since $E_x^{u_\epsilon(y^*)} g(Z_1) \geq F^\epsilon(z,x)$, the inequalities (13) and (14) conclude the proof of the theorem. $\square$

3. Construction of an asymptotically optimal policy. Let y be an arbitrary admissible control for $\mathbf{Q_0}$. We show below how to construct the policy $u_\epsilon(y)$ (appearing in Lemmas 2.1 and 2.2 and in Theorem 2.1). Choose a function $\Delta = \Delta(\epsilon)$ in such a way that

$$\lim_{\epsilon\to 0} \Delta(\epsilon) = 0, \qquad \lim_{\epsilon\to 0} \frac{\Delta(\epsilon)}{\epsilon} = \infty, \tag{15}$$

and set $\tau_l = \tau(l,\epsilon) := l\Delta(\epsilon),\ l = 0,1,2,...,\ell(\epsilon)$, where $\ell(\epsilon) := \lfloor \Delta(\epsilon)^{-1} \rfloor$. Let

$$r_l^\epsilon(y) \stackrel{\text{def}}{=} (\Delta(\epsilon))^{-1} \int_{\tau_l}^{\tau_{l+1}} y_t dt, \qquad l = 0,1,...,\ell(\epsilon)-1. \tag{16}$$

Since $\mathbf{Y}$ is a convex set, $r_l^\epsilon(y) \in \mathbf{Y}$. Hence there exists a stationary policy $s_l^\epsilon(y)$ such that

$$r_l(\epsilon) = \sum_{v,a} \eta(s_l^\epsilon(y); v, a) y(v,a). \tag{17}$$

Now construct $u_\epsilon(y)$ as the Markov policy obtained by applying $s_l^\epsilon(y)$ during $n = \lfloor \tau_l/\epsilon \rfloor, \lfloor \tau_l/\epsilon \rfloor + 1, ..., \lfloor \tau_{l+1}/\epsilon \rfloor - 1$, where $l = 0,1,...,\ell(\epsilon)-1$, and by applying an arbitrary stationary policy during $\lfloor \tau_{\ell(\epsilon)}/\epsilon \rfloor, \lfloor \tau_{\ell(\epsilon)}/\epsilon \rfloor + 1, ..., \lfloor \epsilon^{-1} \rfloor$. In the proof of Lemma 2.1 it is established that the policy $u_\epsilon(y)$ thus constructed satisfies inequality (10).

As follows from Theorem 2.1, the described procedure for obtaining the policy $u_\epsilon(y^*)$, on the basis of a control y_t^* which is optimal for the deterministic problem $\mathbf{Q_0}$, yields an asymptotically optimal policy for problems $\mathbf{Q}_\epsilon$. The optimal control y_t^* can by itself be characterized by necessary and sufficient optimality conditions. To formulate these, let us consider a parametrized set $\mathcal{L} = \{L(z,\lambda)\}$ of MDPs, $(z,\lambda) \in \mathbb{R}^n \times \mathbb{R}^n$, all of which have $\mathbf{X}$ and $\mathbf{A}$ as state and action spaces, and $\mathcal{P} = \{P_{vaw}, v, w \in$

$\mathbf{X}, a \in \mathbf{A}\}$ as transition probabilities. They differ by the immediate cost, which is given by

$$r(z, \lambda; v, a) = \lambda^T f(z, y(v, a)) = \lambda^T f^1(z) + \lambda^T f^2(z) y(v, a).$$

Consider the problem of minimization of the infinite horizon expected average cost related to an initial distribution ξ over $\mathbf{X}$:

(18)

$$J_\xi(z, \lambda) \stackrel{\text{def}}{=} \inf_u J_\xi(z, \lambda; u), \qquad J_\xi(z, \lambda; u) \stackrel{\text{def}}{=} \varliminf_{m \to \infty} \frac{1}{m+1} E_\xi^u \sum_{j=0}^{m} r(z, \lambda; X_j, A_j).$$

It is well known (see Derman [11, section 6]) that

a) The optimal value of the above problem does not depend on the initial distribution ξ, and it is equal to the optimal value of the following linear programming problem:

$$\begin{aligned} (19) \qquad J_\xi(z, \lambda) = J(z, \lambda) &\stackrel{\text{def}}{=} \min_\eta \left\{ \sum_{v,a} r(z, \lambda; v, a) \eta(v, a) | \eta = \{\eta(v, a)\} \in W \right\} \\ &= \lambda^T f^1(z) + \min_\eta \left\{ \lambda^T f^2(z) \sum_{v,a} y(v, a) \eta(v, a) | \eta = \{\eta(v, a)\} \in W \right\}. \end{aligned}$$

b) There is a one-to-one correspondence between optimal stationary policies of $L(z, \lambda)$ and the optimal solutions of (19).

The following statement describes necessary optimality conditions for $\mathbf{Q_0}$.

THEOREM 3.1. *Let y_t^* be an optimal control in $\mathbf{Q_0}$ and let z_t^* be the solution of* (9) *obtained with y^*. That is,*

$$(20) \qquad \frac{d}{dt} z_t^* = f(z_t^*, y_t^*), \qquad z_0 = z.$$

Then, for almost all $t \in [0, 1]$,

$$y_t^* = \sum_{v,a} \eta(z_t^*, \lambda_t; v, a) y(v, a),$$

where $\eta(z, \lambda) = \{\eta(z, \lambda; v, a\}_{v,a}$ stands for a solution of (19) *and λ_t is the solution of the conjugate system*

$$(21) \qquad \frac{d}{dt} \lambda_t = -f_z(z_t^*, y_t^*) \lambda_t, \qquad \lambda_1 = g_z(1);$$

f_z and g_z are $n \times n$ and $n \times 1$ matrices of the partial derivatives of f and g, respectively, over the components of z.

Proof. The proof follows from a direct application of the Pontryagin maximum principle [8, 13] to problem $\mathbf{Q_0}$. ☐

Notice that if the solution of (19) with $z = z_t^*$ and $\lambda = \lambda_t$ is unique for all $t \in [0, 1]$ except for a finite number of switching points and, thus, for all these $t \in [0, 1]$, the corresponding stationary policy $u(z_t^*, \lambda_t)$ achieving inf in (18) with $z = z_t^*$ and $\lambda = \lambda_t$ is unique, then an asymptotically optimal policy for $\mathbf{Q}_\epsilon$ can be defined by simply applying $u(z_{\tau_l}^*, \lambda_{\tau_l})$ during $\lfloor \tau_l/\epsilon \rfloor, \lfloor \tau_l/\epsilon \rfloor + 1, \lfloor \tau_{l+1}/\epsilon \rfloor - 1$, where $l = 0, 1, ..., \ell(\epsilon) - 1$.

Another way to characterize the optimal control in the problem $\mathbf{Q_0}$ is related to the HJB equation written for this problem in the form

$$B^0_t(z,t) + \min_{y\in\mathbf{Y}}\{(B^0_z(z,t))^T f(z,y)\} = 0, \qquad B^0(z,1) = g(z), \tag{22}$$

where $B^0_t(z,t)$, $B^0_z(z,t)$ stand for the partial derivatives of $B^0(z,t)$ over t and components of z, respectively. By (2), (6), (8), for any z and λ,

$$\min_{y\in\mathbf{Y}} \lambda^T f(z,y) = \lambda^T f^1(z) + \min_y\{\lambda^T f^2(z)y | y\in\mathbf{Y}\} = \lambda^T f^1(z)$$
$$+ \min_\eta \left\{ \sum_{v,a} \lambda^T f^2(z) y(v,a)\eta(v,a) | \eta = \{\eta(v,a)\} \in W \right\} = J(z,\lambda),$$

where $J(z,\lambda)$ is the optimal value of (19). Hence, HJB equation (22) can be rewritten in the form

$$B^0_t(z,t) + J(z, B^0_z(z,t)) = 0, \qquad B^0(z,1) = g(z). \tag{23}$$

This equation allows us to construct both necessary and sufficient conditions of optimality for $\mathbf{Q_0}$ and, in particular, to verify whether a given admissible control y_t and the corresponding solution z_t of (9) are optimal in $\mathbf{Q_0}$ (see details in [8]). On the other hand, the viscosity solution of (23) (see, e.g., [14]) defines the optimal value of the problem $\mathbf{Q_0}$ on the interval $[s,1]$ subject to the initial condition $z_s = z$, which provides an approximation for the optimal value $B^\epsilon(z,x,s)$ of the problem $\mathbf{Q}_\epsilon$ on the same interval $[s,1]$ subject to the same initial condition $z_s = z$ and with the initial state of the MDP being x. More precisely, since, by definition, $B^\epsilon(z,x,0) = F^\epsilon(z,x)$ and $B^0(z,0) = F^0(z)$, from Theorem 2.1 it follows that

$$\lim_{\epsilon\to 0} B^\epsilon(z,x,0) = B^0(z,0).$$

As in this theorem, one can also establish that

$$\lim_{\epsilon\to 0} B^\epsilon(z,x,s) = B^0(z,s),$$

with the convergence being uniform with respect to $s\in[0,1]$, $x\in\mathbf{X}$, and $z\in\mathcal{Z}$, where $\mathcal{Z}$ is a compact subset of $\mathbb{R}^n$.

Notice that the described approach has a decomposition structure. It consists of two phases. First is the optimization of the fast motions which is achieved via the solution of (18) with fixed "slow variables" z and λ. Second is the "slow optimization" achieved via the solution of HJB (23). Notice also that in a general case the solution of equation (23) can be quite complicated. If, however,

$$f(z,y) = Az + By, \qquad g(z) = c^T z, \tag{24}$$

where $A(n\times n)$, $B(n\times k)$, and $c(n\times 1)$ are matrices (that is, if as in [2], $\mathbf{Q_0}$ is a linear optimal control problem), then the solution of (23) is obvious:

$$B^0(z,s) = \lambda_s^T z + \int_s^1 J(\lambda(t))dt,$$

where $J(\lambda) \stackrel{\text{def}}{=} J(z,\lambda) - \lambda^T Az$ and λ_t is the solution of (21) under assumption (24).

4. Proof of Lemmas 2.1 and 2.2.

LEMMA 4.1. *Let $y_t^i(h)$, $i = 1, 2$, be functions of time t and state-action histories h. Let $z_t^i(h)$ be the solution of* (9) *obtained with $y_t^i(h)$ (h is fixed), $i = 1, 2$. Then there exists a constant L such that for any policy u and any initial state x,*

$$\begin{aligned} &\max_{t\in[0,1]} E_x^u \left\| z_t^1(H) - z_t^2(H) \right\|_1 \\ &\quad \le L\left(\Delta(\epsilon) + (\Delta(\epsilon))^{-1} \max_{l=0,\dots,\ell(\epsilon)-1} E_x^u \left\| \int_{\tau_l}^{\tau_{l+1}} [y_t^1(H) - y_t^2(H)]dt \right\|_1 \right), \end{aligned} \tag{25}$$

where H is the random realization of the state-action trajectories.

Proof. For the sake of brevity, we omit H from the notation below and write Δ and ℓ instead of $\Delta(\epsilon)$ and $\ell(\epsilon)$. By definition,

$$z_{\tau_{l+1}}^i = z_{\tau_l}^i + \int_{\tau_l}^{\tau_{l+1}} f(z_t^i, y_t^i)dt.$$

Hence, denoting

$$\delta_l := E_x^u \left\| z_{\tau_l}^1 - z_{\tau_l}^2 \right\|_1$$

and taking into account (2), one can write

$$\begin{aligned} \delta_{l+1} \le \delta_l &+ \int_{\tau_l}^{\tau_{l+1}} E_x^u \left\| f(z_t^1, y_t^1) - f(z_{\tau_l}^1, y_t^1) \right\|_1 dt \\ &+ E_x^u \left\| \int_{\tau_l}^{\tau_{l+1}} \left[f(z_{\tau_l}^1, y_t^1) - f(z_{\tau_l}^1, y_t^2) \right] dt \right\|_1 \\ &+ \int_{\tau_l}^{\tau_{l+1}} E_x^u \left\| f(z_{\tau_l}^1, y_t^2) - f(z_{\tau_l}^2, y_t^2) \right\|_1 dt + \int_{\tau_l}^{\tau_{l+1}} E_x^u \left\| f(z_{\tau_l}^2, y_t^2) - f(z_t^2, y_t^2) \right\|_1 dt \\ &\le \delta_l + L_1 \Delta E_x^u \left\| \frac{1}{\Delta} \int_{\tau_l}^{\tau_{l+1}} (y_t^1 - y_t^2)dt \right\|_1 + L_3 \Delta \delta_l + L_1 \Delta^2, \end{aligned}$$

where L_i are constants defined by C_1 and C_2 in (3) and (4) (and thus do not depend on H). Applying now Proposition 5.1 of Gaitsgory [15], one obtains that for any $K = 0, 1, ..., \ell$,

$$\delta_K \le \tilde{L}\left(\Delta + \max_{l=0,\dots,\ell-1} E_x^u \left\| \frac{1}{\Delta} \int_{\tau_l}^{\tau_{l+1}} (y_t^1 - y_t^2)dt \right\|_1 \right), \tag{26}$$

where $\tilde{L}$ is a constant. Since

$$\left\| z_t^i - z_{\tau_l}^i \right\|_1 \le L_4 \Delta \qquad \forall t \in [\tau_l, \tau_{l+1}]$$

for some constant L_4, (26) implies (25) with $L = \tilde{L} + 2L_4$. ∎

We need another general result on MDPs that establishes the uniform convergence of the state-action frequencies to their limits. More precisely, consider arbitrary integers m and K, and define the random variables

$$\psi_m^K(v, a) = \psi_m^K(H; v, a) := \frac{1}{K} \sum_{n=m+1}^{m+K} 1\{X_n = v, A_n = a\}.$$

Let $\psi_m^K := \{\psi_m^K(v,a)\}_{v,a}$ denote the vector of state-action frequencies. Denote

$$d_K^1 = \text{dist}\{\psi_0^K, W\} = \inf_{\eta\in W} \left\|\psi_0^K - \eta\right\|_1 .$$

It follows from Derman [11, Chapter 8, p. 98] (see also [3, section 3]) that for any policy u and initial distribution ξ,

$$\lim_{K\to\infty} d_K^1 = 0, \qquad P_\xi^u \text{ a.s.} \tag{27}$$

This implies, by the bounded convergence theorem, that

$$\lim_{K\to\infty} E_\xi^u d_K^1 = 0. \tag{28}$$

For any stationary policy $u \in \mathcal{S}$ the limit

$$\psi_0 := \lim_{K\to\infty} \psi_0^K$$

exists (P_ξ^u a.s.), and it does not depend on the initial distribution ξ (in fact, $\psi_0(v,a) = \eta(u;v,a)$). Define

$$d_K^2 = \left\|\psi_0^K - \psi_0\right\|_1 .$$

THEOREM 4.1. *The following holds:*

$$\lim_{K\to\infty} \sup_\xi \sup_{u\in U} E_\xi^u d_K^1 = 0, \tag{29}$$

$$\lim_{K\to\infty} \sup_\xi \sup_{u\in \mathcal{S}} E_\xi^u d_K^2 = 0. \tag{30}$$

Proof. In order to prove the theorem, we define some operations on policies. A k-shift $v = \Theta^k u$ of a policy u is defined to be a sequence $v = \{v_k, v_{k+1}, ...\}$, where

$$v_{n+k}(x_0,a_0,x_1,a_1,...,x_{n+k-1},a_{n+k-1},x_{n+k})$$

$$= u_n(x_k,a_k,x_{k+1},a_{k+1},...,x_{n+k-1},a_{n+k-1},x_{n+k}).$$

A policy w is defined to be a concatenation of u and v from time k if

$$w_n = \begin{cases} u_n, & n < k, \\ (\Theta^k v)_n, & n \ge k. \end{cases}$$

We then denote this policy by $w = [u\{k\}v]$. We similarly define a concatenation of a sequence of policies u^i with times t^i, and denote it by $[u^1\{t^1\}u^2\{t^2\}...]$ (where policy u^i is used for a duration of t^i time units).

Assume (29) does not hold. Then there exist sequences of initial distribution over the states $\xi(i) = \{\xi_1(i),...,\xi_N(i)\}$, of strictly increasing times $t(i)$ and of policies $u(i)$, and a constant $\alpha_1 > 0$ such that for all i,

$$E_{\xi(i)}^{u(i)} d_{t(i)}^1 \ge \alpha_1. \tag{31}$$

It follows that there exist sequences of strictly increasing times $t'(i)$ and of policies $u'(i)$, and a constant $\alpha_2 > 0$ such that for all i,

$$E_{\xi'}^{u'(i)} d_{t'(i)}^1 \ge \alpha_2 \tag{32}$$

for any initial distribution ξ'. Indeed, fix $t'(i) = t(i) + N, i = 1, 2, ...$ (N is the number of states). Fix some stationary policy s and let $u'(i)$ be the policy $[s\{N\}\Theta^N u(i)]$, i.e., the policy obtained by using s during the first N steps, and then using a shifted policy $\Theta^N u(i)$. Due to the unichain and aperiodicity assumption, the Markov chain induced by the stationary policy s is regular, and it follows (see [18]) that there exists some $\alpha_3 > 0$ such that $P^s_{\xi'}(X_N = z) > \alpha_3$ for any z and ξ'. (31) then implies that (32) holds for all i sufficiently large and ξ' with $\alpha_2 = \alpha_1\alpha_3/2$. Indeed, let i be such that

$$t(i) \geq \frac{4N}{\alpha_1\alpha_3}.$$

It then follows that

$$|d^1_{t(i)} - d^1_{t'(i)}| \leq 2N/t(i) \leq \frac{\alpha_1\alpha_3}{2}.$$

This implies that

$$\begin{aligned}
E^{u'(i)}_{\xi'} d^1_{t'(i)} &= \sum_z P^{u'(i)}_{\xi'}(X_N = z)\left[E^{u'(i)}_{\xi'} d^1_{t'(i)} \Big| X_N = z\right] \\
&= \sum_z P^s_{\xi'}(X_N = z)\left[E^{u'(i)}_{\xi'} d^1_{t'(i)} \Big| X_N = z\right] \\
&\geq \sum_z P^s_{\xi'}(X_N = z) E^{u(i)}_z d^1_{t(i)} - \frac{\alpha_1\alpha_3}{2} \geq \alpha_3 \sum_z E^{u(i)}_z d^1_{t(i)} - \frac{\alpha_1\alpha_3}{2} \\
&\geq \alpha_3 \sum_z \xi_z(i) E^{u(i)}_z d^1_{t(i)} - \frac{\alpha_1\alpha_3}{2} = \alpha_3 E^{u(i)}_{\xi(i)} d^1_{t(i)} - \frac{\alpha_1\alpha_3}{2} \geq \frac{\alpha_1\alpha_3}{2}.
\end{aligned} \tag{33}$$

Equation (33) is due to the following. Policy $u'(i)$ behaves like the stationary policy s during the first N steps. So, at time N, we reach state z with probability $P^s_{\xi'}(X_N = z)$. Then the behavior during the interval $[N, t'(i)]$, according to policy $u'(i)$, is that of the policy u during the interval $[0, t'(i) - N] = [0, t(i)]$.

Consider now some subsequence $t'(i)$ for which (32) holds and for which

$$\frac{\sum_{l=1}^i t'(l)}{t'(i+1)} \leq \frac{\alpha_2}{4}.$$

Consider the concatenated policy $\tilde{u}$ defined as $\tilde{u} = [u'(1)\{t'(1)\}u'(2)\{t'(2)\}...]$. (32) implies that

$$\overline{\lim_{K\to\infty}} E^{\tilde{u}}_{\xi'} d^1_K \geq \frac{\alpha_2}{2} > 0 \tag{34}$$

for any initial distribution ξ'. Indeed, choose any integer n and define $K = \sum_{i=1}^n t'(i)$, $K' = \sum_{i=1}^{n+1} t'(i)$. Then

$$|E^{\tilde{u}}_{\xi'}[d^1_{K'}|X(K) = z] - E^{u'(i+1)}_z d^1_{t'(i+1)}| \leq \frac{2\sum_{l=1}^i t'(l)}{t'(i+1)} \leq \frac{\alpha_2}{2},$$

which implies that

$$\begin{aligned}
E^{\tilde{u}}_{\xi'} d^1_{K'} &= \sum_z P^{\tilde{u}}_{\xi'}(X(K) = z) E^{\tilde{u}}_{\xi'}[d^1_{K'}|X(K) = z] \\
&\geq \sum_z P^{\tilde{u}}_{\xi'}(X(K) = z) E^{u'(i+1)}_z d^1_{t'(i+1)} - \frac{\alpha_2}{2} \geq \frac{\alpha_2}{2}.
\end{aligned} \tag{35}$$

This, however, contradicts (28) for $u = \tilde{u}$. We thus conclude that the convergence in (28) is uniformly in ξ and $u \in U$.

Next, assume that (30) does not hold. Below, if u is stationary, we understand $u(a|x)$ to be the probability of choosing action a when in state x. The class of stationary policies is compact; i.e., for any sequence $u(i) \in \mathcal{S}$, there exists a subsequence $u(i_j)$ such that the policy $u^* = \lim_{j\to\infty} u(i_j)$ (i.e., the policy for which $u^*(a|x) = \lim_{j\to\infty} u(i_j)(a|x)$ for all a and x) is stationary.

It follows by arguments as in the first part of the proof that there exist sequences of times $t(i)$ and of stationary policies $s(i)$, and a constant $\alpha_4 > 0$ such that for all i,

$$E_\xi^{s(i)} d_{t(i)}^2 \geq \alpha_4 \tag{36}$$

for any initial distribution ξ. Moreover, due to the compactness of $\mathcal{S}$, $s(i)$ can be chosen to be a convergent sequence, with s^* its limit. It then follows that

$$\lim_{i\to\infty} \eta(s(i)) = \eta(s^*) \tag{37}$$

(see [17, p. 82]).

Consider now the Markov policy $\tilde{s}$ that follows policy $s(1)$ until time $t(1)$, then switches to $s(2)$ and uses that policy until $t(2)$, then switches to $s(3)$ and uses it until $t(3)$, and so on. Since for any initial distribution ξ and for any stationary policy $s(i)$, we have

$$\psi_0 = \eta(s(i)), \qquad P_\xi^{s(i)} \text{ a.s.}, \tag{38}$$

it follows by choosing the sequence of times $t(i)$ so that the intervals $t(i+1) - t(i)$ are sufficiently large, that (36) implies that

$$\overline{\lim_{i\to\infty}}\, E_\xi^{\tilde{s}} \left\| \psi_0^{t(i)} - \eta(s(i)) \right\|_1 > 0 \tag{39}$$

for any initial distribution ξ. It then follows from (37) and (39) that

$$\overline{\lim_{t\to\infty}}\, E_\xi^{\tilde{s}} \left\| \psi_0^{t} - \eta(s^*) \right\|_1 > 0 \tag{40}$$

for any initial distribution ξ.

Since $s(i)$ converges to s^*, it follows that $\tilde{s}$ is an asymptotically stationary policy (see (1.2) in [3]), and therefore,

$$\lim_{K\to\infty} \psi_0^K = \eta(s^*), \qquad P_\xi^{\tilde{s}} \text{ a.s.}$$

(see Lemma 6.3 in [3]; also see [4]). Hence

$$\lim_{K\to\infty} E_\xi^{\tilde{s}} \left\| \psi_0^K - \eta(s^*) \right\|_1 = 0 \tag{41}$$

for any initial distribution ξ. This contradicts (40), and thus (30) is established. ∎

Proof of Lemma 2.1. Let y_t be an admissible control for $\mathbf{Q_0}$ and let $u_\epsilon(y)$ be constructed as indicated in the beginning of section 3. Consider the policy $u_\epsilon(y)$ and a random realization of states and actions history $H \in \mathbf{H}$. The solution Z_t of (1) is the solution of (9) obtained with the random control

$$y_t(H) \stackrel{\text{def}}{=} y(X_{\lfloor t/\epsilon \rfloor}, A_{\lfloor t/\epsilon \rfloor}).$$

By Lemma 4.1, the mathematical expectation of the norm of the difference between Z_t and the solution z_t^y of (9) with the control y_t is bounded by

$$E_x^{u_\epsilon(y)} \|Z_t - z_t^y\|_1 \le L \left(\Delta + \max_{l=0,\dots,\ell-1} E_x^{u_\epsilon(y)} \left\| \frac{1}{\Delta} \int_{\tau_l}^{\tau_{l+1}} y_s(H) ds - \frac{1}{\Delta} \int_{\tau_l}^{\tau_{l+1}} y_s ds \right\|_1 \right)$$

for any $t \in [0,1]$. Hence, taking into account (16) and (17),

$$\begin{aligned} &\max_{t\in[0,1]} E_x^{u_\epsilon(y)} \|Z_t - z_t^y\|_1 \\ &\quad \le L \left(\Delta + \max_{l=0,\dots,\ell-1} E_x^{u_\epsilon(y)} \left\| \frac{1}{\Delta} \int_{\tau_l}^{\tau_{l+1}} y_s(H) ds - \sum_{v,a} \eta(s_l^\epsilon(y); v,a) y(v,a) \right\|_1 \right). \end{aligned} \tag{42}$$

To bound the right-hand side in (42), consider the state-action frequencies ψ_m^K corresponding to the realization H. It follows from Theorem 4.1 that there exists some $\mu : \mathbb{N} \to \mathbb{R}$ with

$$\lim_{K\to\infty} \mu(K) = 0 \tag{43}$$

such that for any stationary policy s applied during $n = m+1, ..., m+K$, and any probability distribution ζ over X_m,

$$E_\zeta^s \left(\max_{v,a} |\psi_m^K(v,a) - \eta(s;v,a)| \right) \le \mu(K). \tag{44}$$

Denote

$$K(\epsilon) \stackrel{\text{def}}{=} \min_{l=0,1,\dots,\ell-1} \left(\lfloor \tau_{l+1}/\epsilon \rfloor - \lfloor \tau_l/\epsilon \rfloor \right),$$

and notice that

$$\begin{aligned} &2 \ge \lfloor \tau_{l+1}/\epsilon \rfloor - \lfloor \tau_l/\epsilon \rfloor - K(\epsilon) \ge 0, \qquad \left| K(\epsilon) - \frac{\Delta(\epsilon)}{\epsilon} \right| \le 1 \\ &\Rightarrow \qquad \left| \frac{1}{K(\epsilon)} - \frac{\epsilon}{\Delta(\epsilon)} \right| \le \frac{\epsilon^2}{\Delta(\epsilon)^2} \left(\frac{1}{1 - \epsilon/\Delta(\epsilon)} \right). \end{aligned} \tag{45}$$

From (45) it follows that there exist constants L_1 and L_2 such that

$$\left\| \frac{1}{\Delta(\epsilon)} \int_{\tau_l}^{\tau_{l+1}} y_t(H) dt - \frac{\epsilon}{\Delta(\epsilon)} \sum_{n=\lfloor \tau_l/\epsilon \rfloor+1}^{\lfloor \tau_l/\epsilon \rfloor + K(\epsilon)} y(X_n, A_n) \right\|_1 \le L_1 \frac{\epsilon}{\Delta(\epsilon)}, \tag{46}$$

$$\left\| \frac{\epsilon}{\Delta(\epsilon)} \sum_{n=\lfloor \tau_l/\epsilon \rfloor+1}^{\lfloor \tau_l/\epsilon \rfloor + K(\epsilon)} y(X_n, A_n) - \frac{1}{K(\epsilon)} \sum_{n=\lfloor \tau_l/\epsilon \rfloor+1}^{\lfloor \tau_l/\epsilon \rfloor + K(\epsilon)} y(X_n, A_n) \right\|_1 \le L_2 \frac{\epsilon}{\Delta(\epsilon)}. \tag{47}$$

Since

$$\frac{1}{K(\epsilon)} \sum_{n=\lfloor \tau_l/\epsilon \rfloor+1}^{\lfloor \tau_l/\epsilon \rfloor + K(\epsilon)} y(X_n, A_n) = \sum_{v,a} \psi_{\lfloor \tau_l/\epsilon \rfloor}^{K(\epsilon)}(H; v,a) y(v,a), \tag{48}$$

one can obtain, using (44), (46), and (47),

$$
\begin{aligned}
&E_x^{u_\epsilon(y)} \left\| \frac{1}{\Delta(\epsilon)} \int_{\tau_l}^{\tau_{l+1}} y_t(H)dt - \sum_{v,a} \eta(s_l^\epsilon(y); v, a) y(v, a) \right\|_1 \\
&\le (L_1 + L_2) \frac{\epsilon}{\Delta(\epsilon)} + E_x^{u_\epsilon(y)} \left\{ E_{X_{\lfloor \tau_l/\epsilon \rfloor}}^{s_l^\epsilon(y)} \sum_{v,a} \left(\left| \psi_{\lfloor \tau_l/\epsilon \rfloor}^{K(\epsilon)}(H; v, a) - \eta(s_l^\epsilon(y); v, a) \right| \|y(v, a)\|_1 \right) \right\} \\
&\le (L_1 + L_2) \frac{\epsilon}{\Delta(\epsilon)} + L_3 \mu(K(\epsilon)),
\end{aligned}
$$

where

$$
L_3 = \sum_{v,a} \|y(v, a)\|_1 .
$$

Substituting the last inequality in (42), one obtains

$$
\max_{t \in [0,1]} E_x^{u_\epsilon(y)} \|Z_t - z_t^y\|_1 \le L \left[\Delta(\epsilon) + (L_1 + L_2) \frac{\epsilon}{\Delta(\epsilon)} + L_3 \mu(K(\epsilon)) \right],
$$

which, by (43), completes the proof of the lemma. □

Proof of Lemma 2.2. Let $h = \left\{ x_0, a_0, ..., x_{\lfloor \epsilon^{-1} \rfloor}, a_{\lfloor \epsilon^{-1} \rfloor} \right\} \in \mathbf{H}$ be some state-action trajectory, and define

$$
y_t(h) \stackrel{\text{def}}{=} y(x_{\lfloor t/\epsilon \rfloor}, a_{\lfloor t/\epsilon \rfloor}).
$$

As in (46)–(48), one obtains

$$
\left\| \frac{1}{\Delta(\epsilon)} \int_{\tau_l}^{\tau_{l+1}} y_t(h)dt - \sum_{v,a} \psi_{\lfloor \tau_l/\epsilon \rfloor}^{K(\epsilon)}(h; v, a) y(v, a) \right\|_1 \le (L_1 + L_2) \frac{\epsilon}{\Delta(\epsilon)}. \tag{49}
$$

Denote by $\sigma_l(H)$ the projection of $\psi_{\lfloor \tau_l/\epsilon \rfloor}^{K(\epsilon)}(H)$ on W; i.e., $\sigma_l(H) := \{\sigma_l(H; v, a)\}_{v,a}$ is the solution of

$$
\min_\eta \left\{ \left\| \psi_{\lfloor \tau_l/\epsilon \rfloor}^{K(\epsilon)}(H) - \eta \right\|_1 \middle| \eta \in W \right\}. \tag{50}
$$

It follows from Theorem 4.1 that there exists a function $\nu(K)$,

$$
\lim_{K \to \infty} \nu(K) = 0,
$$

such that for any policy u,

$$
E_x^u \text{dist} \left\{ \psi_m^K(H), W \right\} \le \nu(K)
$$

where

$$
\text{dist} \left\{ \psi_m^K(H), W \right\} \stackrel{\text{def}}{=} \min_\eta \left\{ \left\| \psi_m^K(H) - \eta \right\|_1 \middle| \eta \in W \right\}.
$$

Hence,

$$
E_x^u \left\{ \max_{v,a} \left| \psi_{\lfloor \tau_l/\epsilon \rfloor}^{K(\epsilon)}(H; v, a) - \sigma_l(H; v, a) \right| \right\} \le \nu(K(\epsilon)). \tag{51}
$$

Define the vectors $y_l : \mathbf{H} \to \mathbb{R}$ as

$$y_l(h) = \sum_{v,a} \sigma_l(h; v, a) y(v, a). \tag{52}$$

Since, by definition, $\sigma_l(h) \in W$, then

$$y_l(h) \in \mathbf{Y} \qquad \forall l = 0, 1, ..., \ell - 1.$$

Define now the piecewise constant function $\tilde{y}^\epsilon_t(h)$ as follows: for $t \in [0, \ell\Delta)$, set $\tilde{y}^\epsilon_t(h) := y_l(h)$ for $t \in [\tau_l, \tau_{l+1})$, $l = 0, 1, ..., \ell - 1$. For $t \in [\ell\Delta, 1]$, set $\tilde{y}^\epsilon(h) = \overline{y}$ where $\overline{y}$ is an arbitrary element of $\mathbf{Y}$. Let u be an arbitrary policy. Taking into account (49), (51), and (52), one obtains

$$\begin{aligned}
&E^u_x \left\| \frac{1}{\Delta(\epsilon)} \int_{\tau_l}^{\tau_{l+1}} y_t(H) dt - \frac{1}{\Delta(\epsilon)} \int_{\tau_l}^{\tau_{l+1}} \tilde{y}^\epsilon_t(H) dt \right\|_1 \\
&\leq (L_1 + L_2) \frac{\epsilon}{\Delta(\epsilon)} + E^u_x \max_{v,a} \left| \psi^{K(\epsilon)}_{\lfloor \tau_l/\epsilon \rfloor}(H; v, a) - \sigma_l(H; v, a) \right| \sum_{v,a} \|y(v, a)\|_1 \\
&\leq (L_1 + L_2) \frac{\epsilon}{\Delta(\epsilon)} + L_3 \nu(K(\epsilon)).
\end{aligned}$$

Applying (25) one obtains

$$\max_{t\in[0,1]} E^u_x \|Z_t - \tilde{z}^\epsilon_t(H)\|_1 \leq L \left[\Delta(\epsilon) + (L_1 + L_2) \frac{\epsilon}{\Delta(\epsilon)} + L_3 \nu(K(\epsilon)) \right],$$

which completes the proof. □

REFERENCES

[1] M. Abbad and J. A. Filar, *Perturbation and stability theory for Markov control problems*, IEEE Trans. Automat. Control, 37 (1992), pp. 1415–1420.

[2] E. Altman and V. A. Gaitsgory, *Control of a hybrid stochastic system*, Systems Control Lett., 20 (1993), pp. 307–314.

[3] E. Altman and A. Shwartz, *Adaptive control of constrained Markov chains: Criteria and policies*, Ann. Oper. Res., 28 (1991), Special issue on "Markov Decision Processes," O. Hernandez-Lerma and J. B. Lasserre, eds., pp. 101–134.

[4] E. Altman and O. Zeitouni, *Rate of convergence of empirical measures and costs in controlled Markov chains and transient optimality*, Math. Oper. Res., 19 (1994), pp. 955–974.

[5] A. Bensoussan, *Perturbation Methods in Optimal Control Problems*, John Wiley, New York, 1989.

[6] A. Bensoussan and G.L. Blankenship, *Singular perturbations in stochastic control*, in Singular Perturbations and Asymptotic Analysis in Control Systems, P. Kokotovic, A. Bensoussan, and G. Blankenship, eds., Lecture Notes in Control and Inform. Sciences 90, Springer-Verlag, New York, 1987, pp. 171–260.

[7] D. Bielecki and J.A. Filar, *Singularly perturbed Markov control problem: Limiting average cost*, Ann. Oper. Res., 28 (1991), pp. 153–168.

[8] F. H. Clarke, *Optimization and Nonsmooth Analysis*, John Wiley, New York, 1983.

[9] F. Delebecque and J. Quadrat, *Optimal control of Markov chains admitting strong and weak interactions*, Automatica J. IFAC, 17 (1980), pp. 281–296.

[10] F. Delebecque and J. Quadrat, *Contribution of stochastic control singular perturbation averaging and team theories to an example of large scale systems: Management and hydropower production*, IEEE Trans Automat. Control, AC-23 (1978), pp. 209–222.

[11] C. Derman, *Finite State Markovian Decision Processes*, Academic Press, New York, 1970.

[12] C. Derman and R. E. Strauch, *A note on memoryless rules for controlling sequential control processes*, Ann. Math. Stat., 37 (1966), pp. 276–278.

[13] W. H. Fleming and R. W. Rishel, *Deterministic and Stochastic Optimal Control*, Springer-Verlag, Berlin, Heidelberg, New York, 1975.

[14] W. H. Fleming and H. M. Soner, *Controlled Markov Processes and Viscosity Solutions*, Springer-Verlag, New York, 1993.

[15] V. Gaitsgory, *Suboptimization of singularly perturbed control systems*, SIAM J. Control Optim., 30 (1992), pp. 1228–1249.

[16] I. I. Gikhman, *Po povodu odnoi teoremi N.I. Bogolubova*, Ukrain. Mat. Zh., 4 (1952), pp. 215–218.

[17] A. Hordijk, *Dynamic Programming and Markov Potential Theory*, 2nd ed., Mathematical Centre Tracts 51, Mathematisch Centrum, Amsterdam, 1977.

[18] J. G. Kemeny and J. L. Snell, *Finite Markov Chains*, D. Van Nostrand, New York, 1960.

[19] R. Z. Khas'minsky, *Ustoichivost' Sistem Differencial'nih Uravnenii Pri Sluchainih Vozmuscheniah Ikh Parametrov*, Nauka, Moskva, 1969. English translation: *Stochastic Stability of Differential Equations*, 2nd ed., Sijthoff and Noordhoff, Alphen aan den Rijn, the Netherlands, Rockville, MD, 1980.

[20] L. Kleinrock, *Queuing Systems, Volume* II*: Computer Applications*, John Wiley, New York, 1976.

[21] P. V. Kokotovic, H. Khalil, and J. O'Reilly, *Singular Perturbations in Control Analysis and Design*, Academic Press, New York, 1986.

[22] M. A. Krasnosel'skii and S.G. Krein, *O principe usrednenia v nelineinoi mechanike*, Uspekhi Mat. Nauk, 10 (1955), pp. 147–152.

[23] H. Kushner, *Weak Convergence and Singularly Perturbed Stochastic Control and Filtering Problems*, Birkhäuser, Boston, 1990.

[24] R. G. Philips and P.V. Kokotovic, *A singular perturbation approach to modeling and control of Markov chains*, IEEE Trans. Automat. Control, AC-26 (1981), pp. 1087–1094.

[25] A. A. Pervozvansky and V. Gaitsgory, *Theory of Suboptimal Decisions*, Kluwer, Dordrecht, the Netherlands, 1988.