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

EITAN ALTMAN† AND VLADIMIR GAITSGURY‡

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 \( t_\ell, \ell = 0, 1, \ldots \), according to a controlled Markov chain which has finite state and action spaces. Under the assumption that \( \epsilon \) 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. S0363-0129(95)27998-5

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:

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

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.,

\[
f(z, y) = f^1(z) + f^2(z)y,
\]

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

\[
\|f^i(z) - f^i(z')\|_1 \leq C_1 \|z - z'\|_1 \quad \forall z, z',
\]

\[
\|f^2(z)\|_1 \leq C_2,
\]

where \( z, z' \) are from a sufficiently large domain which contains all possible trajectories of (1), \( C_1 \) and \( C_2 \) are constants, and \( \|\| \|_1 \) stands for the \( L_1 \) norm in the finite-dimensional space. That is, \( \|q\|_1 = \max_{i=1,\ldots,k} |q_i| \) for the vector \( q = \{q_i\}, i = 1, \ldots, k, \) and \( \|A\|_1 = \max_{i=1,\ldots,k} |Aq| \) 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).
is a finite state space \( X \) unit. Time is discretized; i.e., transitions occur at times \( t = n\epsilon \), \( n = 0,1,2,...,[\epsilon^{-1}] \), where \([x]\) stands for the greatest integer which is smaller than or equal to \( x \). There is a finite state space \( X = \{1,...,N\} \) and a finite action space \( A \). If a state is \( v \) and an action \( a \) is chosen, then the next state is \( w \) with the probability \( P_{vw} \). A policy \( u = \{u_0,u_1,\ldots\} \) in the set of policies \( U \) is a sequence of probability measures on \( A \); 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 \( A \).

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

- the Markov policies, denoted by \( M \), 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 \( S \), 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 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 decision process (MDP); see Derman [11, pp. 2–4].

**Notation**

Let \( \xi \) be the set of all possible states and actions histories which can be observed until time \( [\epsilon^{-1}] \):

\[
H = \bigcup \{h\}, \quad h = \{(x_n,a_n), n = 0,1,...,[\epsilon^{-1}]\}.
\]

Let \( \mathcal{F} \) be the \( \sigma \)-algebra of all subsets of \( H \). Each policy \( u \) and initial state \( x \) determines a probability measure on \( \mathcal{F} \), on which the stochastic state and action process \( H = \{X_n,A_n, n = 0,1,...,[\epsilon^{-1}]\} \) 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 \( \xi \) on \( X_0 \), instead of a fixed initial state. In that case \( P_x^\xi \), \( E_x^\xi \) denote the corresponding probability measure and mathematical expectation.

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

\[
Y_t = y(X_{\{t/\epsilon\}},A_{\{t/\epsilon\}}).
\]  

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 \( \epsilon \) and \( x \) fixed.
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 \epsilon is supposed to be a small parameter and our objective is to construct a policy (depending, in general, on \epsilon) which is asymptotically optimal for 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 \epsilon 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 \epsilon. 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 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 Q_\epsilon is approximated by the corresponding deterministic optimal control problem allowing us, in particular, to construct an asymptotically optimal policy for 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 extension 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

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

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

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

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,

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

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

$$W \overset{\text{def}}{=} \bigcup_{u \in S} \{ \eta(u) \}$$

is a polyhedron (see, for example, [11, pp. 93-95]), the set $Y$ is a polyhedron as well.

Define now the averaged deterministic control system as

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

where $y_t$ is a measurable function of $t$ taking values in $Y$. The set of such functions $y : [0, 1] \to Y$ 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

$$\max_{t \in [0, 1]} E_x^u(y) \| Z_t - z_t^y \|_1 \leq \gamma(\epsilon).$$
LEMMA 2.2. There exists a function \( \tilde{y}_t^*(h) \),
\[
\tilde{y}^* : [0, 1] \times \mathbf{H} \rightarrow \mathbf{Y},
\]
such that (a) for each \( h \in \mathbf{H} \), \( \tilde{y}_t^*(h) \) is a piecewise constant function of \( t \) and (b) for any policy \( u \),
\[
\max_{t \in [0, 1]} E_x^u \| Z_t - \tilde{z}_t^*(H) \|_1 \leq \gamma(\epsilon),
\]
where \( Z_t \) is the solution of (1), \( \tilde{z}_t^*(H) \) is the solution of (9) obtained with \( y_t = \tilde{y}_t^*(H) \), 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 \( \sigma \)-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
\[
F^0(z) \overset{\text{def}}{=} \inf_y g(z_1)
\]
over the trajectories \( z \) of system (9). The following theorem about approximation of \( \mathbf{Q}_0 \) by \( \mathbf{Q}_\epsilon \) 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,
\[
|F^\epsilon(z, x) - F^0(z)| \leq C_1 \gamma(\epsilon).
\]
If \( y^* \) is an optimal control for \( \mathbf{Q}_\epsilon \), then the Markov policy \( u_\epsilon(y^*) \) allowing estimate (10) with \( y = y^* \) satisfies the inequality
\[
|E_x^u(y^*) g(Z_1) - F^\epsilon(z, x)| \leq C_1 \gamma(\epsilon).
\]
That is, \( u_\epsilon(y^*) \) is asymptotically optimal for \( \mathbf{Q}_\epsilon \).

Remark 2.1. In the linear case studied in [2], \( \gamma \) 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 \( \tilde{y}_t^*(h) \in \mathbf{Y} \) be the function defined in Lemma 2.2. Then
\[
|E_x^u g(Z_1) - E_x^u g(\tilde{z}_1^*(H))| \leq C_1 E_x^u \| Z_1 - \tilde{z}_1^*(H) \|_1 \leq C_1 \gamma(\epsilon),
\]
where \( C_1 \) is defined in (3). Being piecewise constant, the function \( \tilde{y}^* \) is measurable in \( t \). Hence,
\[
g(\tilde{z}_t^*(h)) \geq F^0(z) \quad \forall h \in \mathbf{H},
\]
Lemma 2.1 it is established that the policy
in such a way that
\[ E^u_x g(z^1_H) \geq F^0(z) \]
for any policy \( u \). From the last inequality and (12), it follows that
\[ E^u_x g(Z_1) \geq F^0(z) - C_1 \gamma(\epsilon), \]
so that
\[ F^\epsilon(z, x) = \inf_u E^u_x g(Z_1) \geq F^0(z) - C_1 \gamma(\epsilon). \]

Now let \( y^* \) be an optimal control in \( Q_0 \). By (10),
\[ |E^u_x(y^*) g(Z_1) - F^0(z)| = |E^u_x(y^*) g(Z_1) - g(y^*)| \leq C_1 E^u_x(y) \|Z_1 - z_1^y\| \leq C_1 \gamma(\epsilon). \]
Hence
\[ E^u_x(y^*) g(Z_1) \leq F^0(z) + C_1 \gamma(\epsilon). \]
Since \( E^u_x(y^*) g(Z_1) \geq F^\epsilon(z, x) \), 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 \( Q_0 \). We show below how to construct the policy \( u_*(y) \) (appearing in Lemmas 2.1 and 2.2 and in Theorem 2.1). Choose a function \( \Delta = \Delta(\epsilon) \) in such a way that
\[ \lim_{\epsilon \to 0} \Delta(\epsilon) = 0, \quad \lim_{\epsilon \to 0} \frac{\Delta(\epsilon)}{\epsilon} = \infty, \]
and set \( \tau_l = \tau(l, \epsilon) := l \Delta(\epsilon), \ l = 0, 1, 2, \ldots, \ell(\epsilon) \), where \( \ell(\epsilon) := \lfloor \Delta(\epsilon)^{-1} \rfloor \). Let
\[ r^*_l(y) = (\Delta(\epsilon))^{-1} \int_{\tau_l}^{\tau_{l+1}} y_t dt, \quad l = 0, 1, \ldots, \ell(\epsilon) - 1. \]
Since \( Y \) is a convex set, \( r^*_l(y) \in Y \). Hence there exists a stationary policy \( s_l^*(y) \) such that
\[ r_l(\epsilon) = \sum_{v,a} \eta(s_l^*(y); v, a) y(v, a). \]

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

As follows from Theorem 2.1, the described procedure for obtaining the policy \( u_*(y^*) \), on the basis of a control \( y^*_l \) which is optimal for the deterministic problem \( Q_0 \), yields an asymptotically optimal policy for problems \( Q_\epsilon \). The optimal control \( y^*_l \) can by itself be characterized by necessary and sufficient optimality conditions. To formulate these, let us consider a parametrized set \( L = \{ L(z, \lambda) \} \) of MDPs, \( (z, \lambda) \in \mathbb{R}^n \times \mathbb{R}^n \), all of which have \( X \) and \( A \) as state and action spaces, and \( P = \{ P_{vw}, v, w \in \mathbb{R}^n \} \).
Consider the problem of minimization of the infinite horizon expected average cost related to an initial distribution $\xi$ over $X$:

\begin{equation}
J_\xi(z, \lambda) \overset{\text{def}}{=} \inf_u J_\xi(z, \lambda; u), \quad J_\xi(z, \lambda; u) \overset{\text{def}}{=} \lim_{m \to \infty} \frac{1}{m+1} E_\xi \sum_{j=0}^{m} r(z, \lambda; X_j, A_j).
\end{equation}

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 $\xi$, and it is equal to the optimal value of the following linear programming problem:

\begin{equation}
J(z, \lambda) = J(z, \lambda; u) \overset{\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} g(v, a)\eta(v, a) | \eta = \{\eta(v, a)\} \in W \right\}.
\end{equation}

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 $Q_0$.

**Theorem 3.1.** Let $y^*_t$ be an optimal control in $Q_0$ and let $z^*_t$ be the solution of (9) obtained with $y^*$. That is,

\begin{equation}
\frac{d}{dt} z^*_t = f(z^*_t, y^*_t), \quad z_0 = z.
\end{equation}

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

\[ y^*_t = \sum_{v, a} \eta(z^*_t, \lambda_t; v, a)g(v, a), \]

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

\begin{equation}
\frac{d}{dt} \lambda_t = -f(z^*_t, y^*_t)\lambda_t, \quad \lambda_1 = g_z(1);
\end{equation}

$f_z$ 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 $Q_0$.  

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 $Q_\epsilon$ can be defined by simply applying $u(z^*_t, \lambda_t)$ during $[t_l/\epsilon, [t_l+1]/\epsilon]$ and $[t_{l+1}/\epsilon, 1]$, where $l = 0, 1, ..., \ell(\epsilon) - 1$. 


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

\begin{equation}
B^0_t(z,t) + \min_{y \in Y} \{ (B^0_t(z,t))^T f(z,y) \} = 0, \quad B^0(z,1) = g(z),
\end{equation}

where $B^0_t(z,t)$, $B^0_z(t)$ 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 $\lambda$,

\[
\min_{y \in Y} \lambda^T f(z,y) = \lambda^T f^1(z) + \min_y \{ \lambda^T f^2(z) y | y \in Y \} = \lambda^T f^1(z)
\]

\[
+ \min_{\eta} \left\{ \sum_{v,a} \lambda^T f^2(z)v(a)\eta(v,a) | \eta = \{ \eta(v,a) \} \in W \right\} = J(z, \lambda),
\]

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

\begin{equation}
B^0_t(z,t) + J(z, B^0_z(t)) = 0, \quad B^0(z,1) = g(z).
\end{equation}

This equation allows us to construct both necessary and sufficient conditions of optimality for $Q_0$ and, in particular, to verify whether a given admissible control $y_\epsilon$ and the corresponding solution $z_\epsilon$ of (9) are optimal in $Q_0$ (see details in [8]). On the other hand, the viscosity solution of (23) (see, e.g., [14]) defines the optimal value of the problem $Q_0$ 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 $Q_\epsilon$ 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 $s \in [0, 1]$, $x \in X$, and $z \in Z$, where $Z$ 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 $\lambda$. 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,

\begin{equation}
f(z, y) = Az + By, \quad g(z) = c^T z,
\end{equation}

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

\[
B^0(z, s) = \lambda^T_s z + \int_s^1 J(\lambda(t)) dt,
\]

where $J(\lambda) \stackrel{\text{def}}{=} J(z, \lambda) - \lambda^T A z$ and $\lambda_s$ is the solution of (21) under assumption (24).
4. Proof of Lemmas 2.1 and 2.2.

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

\[
\max_{t \in [0, 1]} E_x^u \left\| z^1_t(H) - z^2_t(H) \right\|_1
\]

\[
\leq L \left( \Delta(\epsilon) + (\Delta(\epsilon))^{-1} \max_{t = 0, \ldots, \ell(\epsilon) - 1} E_x^u \left\| \int_{\tau_i}^{\tau_{i+1}} [y^1_t(H) - y^2_t(H)] dt \right\|_1 \right),
\]

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 \( \Delta \) and \( \ell \) instead of \( \Delta(\epsilon) \) and \( \ell(\epsilon) \). By definition,

\[
z^i_{\tau_{i+1}} = z^i_{\tau_i} + \int_{\tau_i}^{\tau_{i+1}} f(z^i_t, y^i_t) dt.
\]

Hence, denoting

\[
\delta_i := E_x^u \left\| z^1_{\tau_i} - z^2_{\tau_i} \right\|_1
\]

and taking into account (2), one can write

\[
\delta_{i+1} \leq \delta_i + \int_{\tau_i}^{\tau_{i+1}} E_x^u \left\| f(z^1_t, y^1_t) - f(z^2_t, y^2_t) \right\|_1 dt + E_x^u \left\| \int_{\tau_i}^{\tau_{i+1}} \left[ f(z^1_t, y^1_t) - f(z^2_t, y^2_t) \right] dt \right\|_1
\]

\[
+ \int_{\tau_i}^{\tau_{i+1}} E_x^u \left\| f(z^1_t, y^1_t) - f(z^2_t, y^2_t) \right\|_1 dt + \int_{\tau_i}^{\tau_{i+1}} E_x^u \left\| f(z^2_t, y^2_t) - f(z^2_t, y^2_t) \right\|_1 dt
\]

\[
\leq \delta_i + L_3 \Delta E_x^u \left\| \frac{1}{\Delta} \int_{\tau_i}^{\tau_{i+1}} (y^1_t - y^2_t) dt \right\|_1 + L_3 \Delta \delta_i + L_3 \Delta^2,
\]

where \( L_3 \) 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 \( K = 0, 1, \ldots, \ell \),

\[
\delta_K \leq \tilde{L} \left( \Delta + \max_{t = 0, \ldots, \ell - 1} E_x^u \left\| \frac{1}{\Delta} \int_{\tau_t}^{\tau_{t+1}} (y^1_t - y^2_t) dt \right\|_1 \right),
\]

where \( \tilde{L} \) is a constant. Since

\[
\left\| z^1_{\tau_i} - z^2_{\tau_i} \right\|_1 \leq L_4 \Delta \quad \forall t \in [\tau_i, \tau_{i+1}]
\]

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

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

\[
\psi^K_m(v, a) = \psi^K_m(H; v, a) := \frac{1}{K} \sum_{n=m+1}^{m+K} 1\{X_n = v, A_n = a\}.
\]
ASYMPTOTIC OPTIMIZATION OF A HYBRID SYSTEM GOVERNED BY MDP

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

$$d^K_K = \text{dist}\{\psi^K_0, W\} = \inf_{\eta \in W} ||\psi^K_0 - \eta||_1.$$  

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

$$\lim_{K \to \infty} d^K_1 = 0, \quad P^u_\xi \text{ a.s.} \tag{27}$$

This implies, by the bounded convergence theorem, that

$$\lim_{K \to \infty} E^u_\xi d^K_1 = 0. \tag{28}$$

For any stationary policy $$u \in S$$ the limit

$$\psi_0 := \lim_{K \to \infty} \psi^K_0$$

exists ($$P^u_\xi \text{ a.s.}$$), and it does not depend on the initial distribution $$\xi$$ (in fact, $$\psi_0(v, a) = \eta(u; v, a)$$). Define

$$d^K_2 = ||\psi^K_0 - \psi_0||_1.$$  

THEOREM 4.1. The following holds:

$$\lim_{K \to \infty} \sup_{\xi} \sup_{u \in U} E^u_\xi d^K_1 = 0, \tag{29}$$

$$\lim_{K \to \infty} \sup_{\xi} \sup_{u \in S} E^u_\xi d^K_2 = 0. \tag{30}$$

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 $$v = \{v_k, v_{k+1}, \ldots\}$$, where

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

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

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

$$w_n = \begin{cases} u_n, & n < k, \\ \Theta^k u, & n \geq k. \end{cases}$$

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 $$[u^i\{t^i\}]w^2\{t^2\}$$ (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 $$\xi(i) = \{\xi_1(i), \ldots, \xi_N(i)\}$$, of strictly increasing times $$t(i)$$ and of policies $$u(i)$$, and a constant $$\alpha_1 > 0$$ such that for all $$i$$,

$$E^{u(i)}_{\xi(i)} d^K_1 \geq \alpha_1.$$  

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 $$i$$,

$$E^{u'(i)}_{\xi'} d^K_1 \geq \alpha_2.$$
for any initial distribution $\xi'$. Indeed, fix $t'(i) = t(i) + N, i = 1, 2, \ldots$ \((N\text{ is the number of states})\). 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 $P_{\xi'}^z(X_N = z) > \alpha_3$ for any $z$ and $\xi'$. (31) then implies that (32) holds for all $i$ sufficiently large and $\xi'$ 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

$$E_{\xi'}^u(i) d^1_{t'(i)} = \sum_z P^z_{\xi'}(X_N = z) \left[ E_{\xi'}^{u'(i)} d^1_{t'(i)} | X_N = z \right]$$

$$= \sum_z P^z_{\xi'}(X_N = z) \left[ E_{\xi'}^{u'(i)} d^1_{t'(i)} | X_N = z \right]$$

$$\geq \sum_z P^z_{\xi'}(X_N = z) E^z_{\xi'}(d^1_{t'(i)}) - \frac{\alpha_1 \alpha_3}{2} \geq \alpha_3 \sum_z E^z_{\xi'}(d^1_{t'(i)}) - \frac{\alpha_1 \alpha_3}{2} \geq \alpha_3 \frac{\alpha_1 \alpha_3}{2}.$$ 

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 $P^z_{\xi'}(X_N = z)$. 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

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

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

$$\lim_{K \to \infty} E_{\xi'} \hat{u} d^1_K \geq \frac{\alpha_2}{2} > 0$$

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

$$|E_{\xi'} \hat{u} d^1_K |X(K) = z| - E_{\xi'}^{u'(i+1)} d^1_{t'(i+1)}| \leq 2 \sum_{i=1}^{t'(i)} t'(l) / t'(i+1) \leq \frac{\alpha_2}{2},$$

which implies that

$$E_{\xi'} \hat{u} d^1_K = \sum_z P^z_{\xi'}(X(K) = z) E_{\xi'} \hat{u} d^1_K |X(K) = z$$

$$\geq \sum_z P^z_{\xi'}(X(K) = z) E_{\xi'}^{u'(i+1)} d^1_{t'(i+1)} - \frac{\alpha_2}{2} \geq \frac{\alpha_2}{2}.$$ 


This, however, contradicts (28) for \( u = \tilde{u} \). We thus conclude that the convergence in (28) is uniformly in \( \xi \) 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 \( u(i) \in S \), 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 \( i \),

\[
E^s(i)_{\xi} \geq \alpha_4
\]

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

\[
\lim_{i \to \infty} \eta(s(i)) = \eta(s^*)
\]

(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 \( \xi \) and for any stationary policy \( s(i) \), we have

\[
\psi_0 = \eta(s(i)), \quad P_{\xi}^s(i) \text{ a.s.,}
\]

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

\[
\lim_{i \to \infty} E^\tilde{s}(\xi) \left\| \psi_0^{t(i)} - \eta(s(i)) \right\|_1 > 0
\]

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

\[
\lim_{i \to \infty} E^\tilde{s}(\xi) \left\| \psi_0^{t(i)} - \eta(s^*) \right\|_1 > 0
\]

for any initial distribution \( \xi \).

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

\[
\lim_{K \to \infty} \psi_0^K = \eta(s^*), \quad P_{\xi}^s \text{ a.s.}
\]

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

\[
\lim_{K \to \infty} E^\tilde{s}(\xi) \left\| \psi^K_0 - \eta(s^*) \right\|_1 = 0
\]

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

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

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

$$E_x^u(y) \|Z_t - z_t^y\|_1 \leq L \left( \Delta + \max_{t=0,\ldots,\ell-1} E_x^u(y) \left\| \frac{1}{\Delta} \int_{\tau_i}^{\tau_{i+1}} y_s(H) ds - \frac{1}{\Delta} \int_{\tau_i}^{\tau_{i+1}} y_s ds \right\|_1 \right)$$

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

$$\max_{t\in[0,1]} E_x^u(y) \|Z_t - z_t^y\|_1 \leq L \left( \Delta + \max_{t=0,\ldots,\ell-1} E_x^u(y) \left\| \frac{1}{\Delta} \int_{\tau_i}^{\tau_{i+1}} y_s(H) ds - \sum_{v,a} \eta(s_i^e(y); v, a) y(v, a) \right\|_1 \right).$$

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

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

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

$$E^s_\xi \left( \max_{v,a} |\psi^K_m(v, a) - \eta(s; v, a)| \right) \leq \mu(K).$$

Denote

$$K(\epsilon) \overset{\text{def}}{=} \min_{t=0,1,\ldots,\ell-1} \left( \frac{[\tau_{t+1}]}{\epsilon} - \frac{[\tau_t]}{\epsilon} - K(\epsilon) \right),$$

and notice that

$$2 \geq \frac{[\tau_{t+1}]}{\epsilon} - \frac{[\tau_t]}{\epsilon} - K(\epsilon) \geq 0, \quad \frac{K(\epsilon) - \Delta(\epsilon)}{\epsilon} \leq 1$$

$$\Rightarrow \left| \frac{1}{K(\epsilon)} - \frac{\epsilon}{\Delta(\epsilon)} \right| \leq \frac{\epsilon^2}{\Delta(\epsilon)^2} \left( \frac{1}{1 - \epsilon/\Delta(\epsilon)} \right).$$

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

$$\left\| \frac{1}{\Delta(\epsilon)} \int_{\tau_i}^{\tau_{i+1}} y_t(H) dt - \frac{\epsilon}{\Delta(\epsilon)} \sum_{n=[\tau_t]/\epsilon+1} y(X_n, A_n) \right\|_1 \leq L_1 \frac{\epsilon}{\Delta(\epsilon)},$$

$$\left\| \frac{\epsilon}{\Delta(\epsilon)} \sum_{n=[\tau_t]/\epsilon+1} y(X_n, A_n) - \frac{1}{K(\epsilon)} \sum_{n=[\tau_t]/\epsilon+1} y(X_n, A_n) \right\|_1 \leq L_2 \frac{\epsilon}{\Delta(\epsilon)}.$$
one can obtain, using (44), (46), and (47),

\[
E^u_x(y) \left\| \frac{1}{\Delta(\epsilon)} \int_{\tau_l}^{\tau_{l+1}} y_t(H) dt - \sum_{v,a} \eta(s'_t(y); v, a) y(v, a) \right\|_1 \\
\leq (L_1 + L_2) \frac{\epsilon}{\Delta(\epsilon)} + E^u_x(y) \left\{ E^{s_t(y)}_{x_{\tau_l/\epsilon}} \sum_{v,a} \left\| \psi^{K(\epsilon)}_{\tau_l/\epsilon}(H; v, a) - \eta(s'_t(y); v, a) \right\| y(v, a) \right\|_1 \right\}
\]

\[
\leq (L_1 + L_2) \frac{\epsilon}{\Delta(\epsilon)} + L_3\mu(K(\epsilon)),
\]

where

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

Substituting the last inequality in (42), one obtains

\[
\max_{t \in [0, 1]} E^u_x(y) \| Z_t - Z^y_t \|_1 \leq L \left\{ \Delta(\epsilon) + (L_1 + L_2) \frac{\epsilon}{\Delta(\epsilon)} + L_3\mu(K(\epsilon)) \right\},
\]

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

**Proof of Lemma 2.2.** Let \(h = \{x_0, a_0, \ldots, x_{\epsilon-1}, a_{\epsilon-1}\} \in H\) be some state-action trajectory, and define

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

As in (46)–(48), one obtains

\[
\left\| \frac{1}{\Delta(\epsilon)} \int_{\tau_l}^{\tau_{l+1}} y_t(h) dt - \sum_{v,a} \psi_{\tau_l/\epsilon}^{K(\epsilon)}(H; v, a) y(v, a) \right\|_1 \leq (L_1 + L_2) \frac{\epsilon}{\Delta(\epsilon)}.
\]

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

\[
\min_{\eta} \left\{ \left\| \psi_{\tau_l/\epsilon}^{K(\epsilon)}(H) - \eta \right\|_1 \mid \eta \in W \right\}.
\]

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 \(u\),

\[
E^u_x \\text{dist} \{\psi^K_m(H), W\} \leq \nu(K)
\]

where

\[
\text{dist} \{\psi^K_m(H), W\} \overset{\text{def}}{=} \min_{\eta} \left\{ \left\| \psi^K_m(H) - \eta \right\|_1 \mid \eta \in W \right\}.
\]

Hence,

\[
E^u_x \left\{ \max_{v,a} \left| \psi_{\tau_l/\epsilon}^{K(\epsilon)}(H; v, a) - \sigma_t(H; v, a) \right| \right\} \leq \nu(K(\epsilon)).
\]
Define the vectors \( y_l : \mathbf{H} \to \mathbb{R} \) as
\[
(52) \quad y_l(h) = \sum_{v,a} \sigma_l(h; v, a)y(v, a).
\]
Since, by definition, \( \sigma_l(h) \in W \), then
\[
y_l(h) \in \mathbf{Y} \quad \forall l = 0, 1, ..., \ell - 1.
\]
Define now the piecewise constant function \( \tilde{y}_l(h) \) as follows: for \( t \in [0, \ell \Delta) \), set \( \tilde{y}_l(h) := y_l(h) \) for \( t \in [\tau_l, \tau_{l+1}) \), \( l = 0, 1, ..., \ell - 1 \). For \( t \in [\ell \Delta, 1] \), set \( \tilde{y}_l(h) = \bar{y} \) where \( \bar{y} \) is an arbitrary element of \( \mathbf{Y} \). Let \( u \) be an arbitrary policy. Taking into account (49), (51), and (52), one obtains
\[
E_x^n \left\| \frac{1}{\Delta(\epsilon)} \int_{\tau_l}^{\tau_{l+1}} y_l(H)dt - \frac{1}{\Delta(\epsilon)} \int_{\tau_l}^{\tau_{l+1}} \tilde{y}_l(H)dt \right\|_1 \\
\leq (L_1 + L_2) \frac{\epsilon}{\Delta(\epsilon)} + E_x^n \max_{v,a} \psi^{(K(\epsilon))}_{(\tau_l/\epsilon)}(H; v, a) - \sigma_l(H; v, a) \sum_{v,a} \|y(v, a)\|_1 \\
\leq (L_1 + L_2) \frac{\epsilon}{\Delta(\epsilon)} + L_3 \nu(K(\epsilon)).
\]
Applying (25) one obtains
\[
\max_{t \in [0,1]} E_x^n \|Z_t - \tilde{z}_l^1(H)\| \leq L \left[ \Delta(\epsilon) + (L_1 + L_2) \frac{\epsilon}{\Delta(\epsilon)} + L_3 \nu(K(\epsilon)) \right],
\]
which completes the proof.

REFERENCES

ASYMPTOTIC OPTIMIZATION OF A HYBRID SYSTEM GOVERNED BY MDP


