A Hybrid (Differential-Stochastic) Zero-Sum Game with Fast Stochastic Part

Eitan Altman
Projet MISTRAL, INRIA, BP93
2004 Route des Lucioles
06902 Sophia Antipolis Cedex
France

Vladimir Gaitsgory
School of Mathematics, Univ. of South Australia
the Levels, Pooraka, South Australia 5095
Australia

Abstract

We consider in this paper a continuous time stochastic hybrid system with a finite time horizon, controlled by two players with opposite objectives (zero-sum game). Player one wishes to maximize some linear function of the expected state trajectory, and player two wishes to minimize it. The state evolves according to a linear dynamics. The parameters of the state evolution equation may change at discrete times according to a MDP, i.e. a Markov chain that is directly controlled by both players, and has a countable state space. Each player has a finite action space. We use a procedure similar in form to the maximum principle; this determines a pair of stationary strategies for the players, which is asymptotically a saddle point, as the number of transitions during the finite time horizon grows to infinity.

Keywords: Hybrid stochastic systems, stochastic games, asymptotic optimality, linear dynamics, Markov decision processes, finite horizon.

1 Introduction and statement of the problem.

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

$$\frac{d}{dt}Z_t = AZ_t + BY_t, \qquad t \in [0,1], \qquad Z_0 = z \tag{1}$$

where $Y_t \in \mathbb{R}^k$ is the "control" and $A(n \times n)$ and $B(n \times k)$ are matrices of real numbers. Y_t is not chosen directly by the controllers, 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 or equal to x. There is a countable state space $\mathbf{X} = \mathbb{N}$ and two players having finite action spaces $\mathbf{A}_1$ and $\mathbf{A}_2$ respectively. Let $\mathbf{A} = \mathbf{A}_1 \times \mathbf{A}_2$. If the state is v and actions $a = (a_1, a_2)$ are chosen by the players, then the next state is w with probability P_{vaw}. Denote $\mathcal{P} = \{P_{vaw}\}$. A policy $u^i = \{u_0^i, u_1^i, ...\}$ in the set of policies U^i for player i, $i = 1, 2$ is a sequence of probability measures on $\mathbf{A}_i$ conditioned on the history of all previous states and actions of both players, as well as the current state. More precisely, define the set of histories: $\mathbf{H} := \cup_l \mathbf{H}_l$, where

$$\mathbf{H}_l := \{(x_0, a_0^1, a_0^2, x_1, a_1^1, a_1^2, ..., x_l)\}$$

are the sets of all sequences of $3l + 1$ elements describing the possible samples of previous states and actions prior to l as well as the currents state at stage l (i.e. at time $l\epsilon$). (The range of l will be either $l = 0, 1, ..., \lfloor \epsilon^{-1} \rfloor$, or, in other contexts, all nonnegative integers, depending on whether we consider finite or infitine horizon problems). The policy at stage l for player i, u_l^i, is a map from $\mathbf{H}_l$ to the set of probability measures over the action space $\mathbf{A}_i$. (Hence at each time $t = l\epsilon$, player i, observing the history h_l, chooses action a_i with probability $u_l^i(a_i|h_l)$). Let $\mathcal{F}_l$ be the discrete σ-Algebra of subsets of $\mathbf{H}_l$. Each initial distribution ξ and policy pair u for the players uniquely define a probability measure P_ξ^u over the space of samples $\mathbf{H}$ (equiped with the discrete σ-algebra), see e.g. [**?**]. Denote by E_ξ^u the corresponding expectation operator. On the above probability space are now defined the random processes X_l and $A_l = (A_l^1, A_l^2)$, denoting the state and actions processes. When the initial distribution is concentrated on a single state x, we shall denote the corresponding probability measure and expectation by P_x^u and E_x^u.

Let $y^j : \mathbf{X} \times \mathbf{A} \to \mathbb{R}$, $j = 1, ..., k$ be some given bounded functions. Then Y_t in (**??**) is given by

$$Y_t = y(X_{\lfloor t/\epsilon \rfloor}, A_{\lfloor t/\epsilon \rfloor}). \tag{2}$$

Y_t and thus Z_t are well defined stochastic processes, and are both $\mathcal{F}_{\lfloor \epsilon^{-1} \rfloor}$ measurable.

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

(i) The Markov policies $\mathcal{M}_1, \mathcal{M}_2$: these are policies where u_l^i depends only on the current state (at time $t = l\epsilon$) and on l, and does not depend on previous states and actions. If a Markov policy $u^i \in \mathcal{M}_i$ is used by player i, we shall denote

$$u_l^i(a|x) : \text{ the probability under } u^i \text{ of choosing } a \in \mathbf{A}_i \text{ in state } x \text{ at stage } l. \tag{3}$$

Denote $\mathcal{M} = \mathcal{M}_1 \times \mathcal{M}_2$.

(ii) The stationary policies, denoted by $\mathcal{S}_1$, for player 1, and $\mathcal{S}_2$, for player 2. A policy u is called stationary if u_l depends only on the current state, and does

not depend on previous states and actions nor on the time. Let $\mathcal{S} := \mathcal{S}_1 \times \mathcal{S}_2$. If a stationary policy f is used, we shall denote by $f_x(a)$ the probability under f of choosing action a when in state x. When stationary policies $f = (f^1, f^2)$ are used by the players, we set

$$P_{vfw} = P_{vf^1f^2w} = \sum_{a^1,a^2} P_{va^1a^2w} f_v^1(a^1) f_v^2(a^2),$$

$$y(v,f) = y(v,f^1,f^2) = \sum_{a^1,a^2} y(v,a^1,a^2) f_v^1(a^1) f_v^2(a^2).$$

Let $P_f = \{P_{vfw}\}$ be the transition probabilities of the the Markov chain induced by a stationary policy pair f, and let $P_f^l = \{[P_f^l]_{vw}\}$ be the l step transition probabilities under f.

We make throughout the following assumption, which is a strong version of the Simultaneous Doeblin Condition, introduced in [**?**] Section 11.1, with a communicating condition.

(A1): There exists a state $x^* \in \mathbf{X}$ and a positive real number q_0 such that

$$P_{xfx^*} \geq q_0, \qquad \forall x \in \mathbf{X} f \in \mathcal{S}.$$

Let c be an n-dimensional vector representing the (linear) operating cost related to the process Z_t. Define the cost:

$$J_x^z(u^1,u^2) = E_x^{(u^1,u^2)} c^T Z_1, \qquad Z_0 = z$$

when policies u^1, u^2 are used by the players, and the initial state of the linear system is z, and the initial state of the controlled Markov chain is x. In our dynamic game, player 1 wishes to maximize $J_x^z(u^1,u^2)$ and player 2 wants to minimize it. More precisely, define the following problems:

$\mathbf{Q1}_I^\epsilon$: find a policy $u^1 \in U^1$ that achieves

$$F_I^\epsilon(x) = \sup_{u^1 \in U^1} \inf_{u^2 \in U^2} J_x^z(u^1,u^2)$$

where Z_1 is obtained through (**??**). If such a policy exists, then it is called optimal for $\mathbf{Q1}_I^\epsilon$. If for some δ and $u^1 \in U^1$,

$$F_I^\epsilon(x) \leq \inf_{u^2 \in U^2} J_x^z(u^1,u^2) + \delta$$

then u^1 is called δ-optimal for $\mathbf{Q1}_I^\epsilon$. One may consider also:

$\mathbf{Q1}_{II}^\epsilon$: find a policy $u^2 \in U^2$ that achieves

$$F_{II}^\epsilon(x) = \inf_{u^2 \in U^2} \sup_{u^1 \in U^1} J_x^z(u^1,u^2).$$

Define similarly optimality and δ-optimality of policies for $\mathbf{Q1}^\epsilon_{II}$. We clearly have $F^\epsilon_{II}(x) \geq F^\epsilon_I(x)$. If there exist some $u = (u^1, u^2)$ and δ such that

$$F^\epsilon_I(x) + \delta \geq E^{(u^1,u^2)}_x c^T Z_1 \geq F^\epsilon_{II}(x) - \delta,$$

then u is called δ-saddle point, or δ-equilibrium strategy pair for $\mathbf{Q1}^\epsilon$ (we need not specify $\mathbf{Q1}^\epsilon_I$ or $\mathbf{Q1}^\epsilon_{II}$). If this holds for $\delta = 0$, then u is called saddle point or equilibrium strategy for $\mathbf{Q}^\epsilon$.

Remarks:

(i) $\mathbf{Q1}^\epsilon_I$ is equivalent to the problem: find a policy $u^1 \in U^1$ that achieves $\sup_{u^1 \in U^1} \inf_{u^2 \in U^2} c^T \overline{Z}_1$, where $\overline{Z}_t \in \mathbb{R}^n$ is given by

$$\frac{d}{dt}\overline{Z}_t = A\overline{Z}_t + BE^{(u^1,u^2)}_x Y_t, \qquad t \in [0,1], \qquad \overline{Z}_0 = z \tag{4}$$

The same holds for $\mathbf{Q1}^\epsilon_{II}$.

(ii) By solving the problem $\mathbf{Q1}^\epsilon_I$, one can also solve a problem with an integral cost function, i.e. to find a policy u that achieves

$$\sup_{u^1 \in U^1} \inf_{u^2 \in U^2} E^{(u^1,u^2)}_x \int_0^1 c^T Z_t dt.$$

This is obtained by using a new variable R_t defined by $dR_t/dt = c^T Z_t$.

Note that the controllers do not require knowledge of the initial value z of Z_0, which may be assumed to be zero. More precisely, due to the linearity of the system (**??**), if a control strategy is optimal (or δ-optimal) for a given Z_0, then it is optimal (or δ-optimal, respectively) for any other value of Z_0.

Our model is characterized by the fact that ϵ is supposed to be a small parameter. We construct a set of Markov policies $\overline{u}^\epsilon = (\overline{u}^{1,\epsilon}, \overline{u}^{2,\epsilon})$ such that $\overline{u}^\epsilon$ is $\gamma(\epsilon)$-equilibrium for $\mathbf{Q1}^\epsilon$ where $\lim_{\epsilon \to 0} \gamma(\epsilon) = 0$. This implies, in particular, that the game has the value in the limit as $\epsilon \to 0$ and we call the mentioned above sequence of Markov policies asymptotically saddle-point.

This paper is a continuation and generalization of our previous work [**?**] which solves a hybrid problem restricted to a single controller and to a finite state space. As in [1], the fact that ϵ is small means that the variables Y_t can be considered to be fast with respect to Z_t, since, by (2), they may have a finite (not tending with ϵ to zero) change at each interval of the length ϵ. This along with the linearity of the system (1) allow to decompose the game into stochastic subgames on a sequence of intervals which are short with respect to the variables Z_t (in the sense that Z_t remain almost unchanged on these intervals) and which are long enough with respect to Y_t (so that the corresponding stochastic subgames show on these intervals there limit properties).

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 change in a continuous (linear) way. Breakdowns, repairs and other control decisions yield

the underlying controlled Markov chain. In particular, repair, or preventive maintenance decisions are typical actions of a player that minimizes costs. If there is some unknown parameter (disturbance) of the dynamics of the system (e.g. the probability of breakdowns) which may change in a way that depends on the current and past states in a way that is unknown and unpredictable by the minimizer, we may formulate this situation as a zero-sum game, where the minimizer wishes to guarantee the best performance (lowest expected cost) under the worst case behavior of nature. Nature may then be modeled as the maximizing player. (This yields $\mathbf{Q1}^{\epsilon}_{II}$.)

Our model may also be used in the control of highly loaded queueing networks for which the fluid approximation holds (see Kleinrock [?] p. 56). The quantities Z_t may then represent the number of customers in the different queues whereas the underlying controlled Markov chain may correspond to routing, or flow control of, say, some on-off traffic, with again, nature controlling some disturbances in quantities such as service rates.

The structure of the paper is as follows. In Section **??** we present the main result; we construct the sequence of non-stationary policy for the hybrid control problems $\mathbf{Q1}^{\epsilon}$. We prove in Section **??** that the sequence of policies introduced in Section **??** is indeed asymptotically saddle-point as ϵ tends to zero. Proofs of some technical lemmas are left to the Appendix.

Below, B^T will denote the transpose of a matrix (or of a column vector) B, and $\|B\|$ will denote the sum of absolute values of the components of B.

2 Construction of ϵ-equilibrium Markov strategies

Consider a family of infinite horizon stochastic games, all with the same state and action spaces $\mathbf{X}$ and $\mathbf{A}$ as above, and the same transition probabilities $\mathcal{P}$, parametrized by a vector $\lambda \in \mathbb{R}^n$. Let $r : \mathbb{R}^n \times \mathbf{X} \times \mathbf{A} \to \mathbb{R}$ be the immediate cost, i.e. $r(\lambda, x, a)$ is the cost in the MDP λ, when at state x and the actions chosen are a. r is given by

$$r(\lambda, x, a) = \lambda^T B y(x, a).$$

The definition of policies $U = (U^1, U^2)$ is as in Section **??**. Define the following cost functions. The finite horizon total expected cost:

$$\sigma^m(\lambda, \xi, u) := E^u_{\xi} \sum_{i=0}^{m-1} r(\lambda, X_i, A_i); \tag{5}$$

The infinite horizon expected average cost:

$$\overline{\sigma}(\lambda, \xi, u) := \underline{\lim}_{m \to \infty} \frac{\sigma^m(\lambda, \xi, u)}{m}$$

Remark: The results of the paper are unchanged if the liminf is replaced by a limsup in the definition of the infinite horizon average cost.

A policy pair $u^\lambda = (u^{1,\lambda}, u^{2,\lambda}) \in U$ is said to be a saddle point or an equilibrium policy pair for problem λ with infinite horizon expected average cost criterion, if for all $u^1 \in U^1, u^2 \in U^2$,

$$\overline{\sigma}(\lambda, \xi, u^1, u^{2,\lambda}) \leq \overline{\sigma}(\lambda, \xi, u^{1,\lambda}, u^{2,\lambda}) \leq \overline{\sigma}(\lambda, \xi, u^{1,\lambda}, u^2). \tag{6}$$

Let $f^\lambda = (f^{1,\lambda}, f^{2,\lambda})$, where $f^{1,\lambda} \in \mathcal{S}_1, f^{2,\lambda} \in \mathcal{S}_2$, be some stationary equilibrium policy pair for the expected average problem. The existence of such stationary equilibrium policy pair (under assumption (A1)) is well known, see e.g. [**?**].

$$\overline{\sigma}(\lambda) := \overline{\sigma}(\lambda, \xi, f^{1,\lambda}, f^{2,\lambda}) \tag{7}$$

is then defined to be the value of the λ stochastic game, and is known to be independent on ξ (which we shall thus omit from the notation). It can be computed using value iteration, see e.g. [**?**].

Let $\lambda(t) \in \mathbb{R}^n$, $t \in [0,1]$ be the solution of

$$\frac{d}{dt}\lambda_t = -A^T\lambda_t, \quad \lambda_1 = -c \tag{8}$$

i.e.,

$$\lambda(t) = e^{A^T(1-t)}c.$$

Define the following:

- $\Delta(\epsilon)$:= a function of ϵ such that

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

  $\Delta(\epsilon)$ will be the length of sub-intervals of [0,1] during which we shall use fixed stationary policies. (In each new sub-interval, a new stationary policy has to be computed).

- $\tau_l := l\Delta(\epsilon)$, $l = 0, 1, 2, ..., \lfloor \Delta(\epsilon)^{-1} \rfloor$, is the instant at which the lth sub-interval begins.

- $M_\epsilon := \lfloor \Delta(\epsilon)^{-1} \rfloor$ is the number of sub-intervals.

- $\tau_{M_\epsilon + 1} := 1$.

- $m_l := \lfloor (l+1)\Delta(\epsilon)\epsilon^{-1} \rfloor - \lfloor (l)\Delta(\epsilon)\epsilon^{-1} \rfloor$.

- $\overline{u}^\epsilon := (\overline{u}^{1,\epsilon}, \overline{u}^{2,\epsilon})$ is a pair of Markov policies defined by the players as follows: each player $i = 1, 2$ defines $\overline{u}^{i,\epsilon}$ by applying $f^{i,\lambda(\tau_l)}$, $i = 1, 2$ during $n = \lfloor \tau_l/\epsilon \rfloor, \lfloor \tau_l/\epsilon \rfloor + 1, ..., \lfloor \tau_{l+1}/\epsilon \rfloor - 1; l = 0, 1, ..., M_\epsilon$, where $f^{i,\lambda(\tau_l)}$ is defined in the paragraph above (**??**), and by choosing an arbitrary action at $\lfloor \epsilon^{-1} \rfloor$.

Theorem 2.1 *$\overline{u}^\epsilon$ is an asymptotically saddle point, i.e. for every ϵ there exists some $\gamma(\epsilon)$ with $\lim_{\epsilon\to 0}\gamma(\epsilon)=0$, such that $\overline{u}^\epsilon$ is $\gamma(\epsilon)$-equilibrium for problem* **Q1**$^\epsilon$*:*

$$J_x^z(u^1,\overline{u}^{2,\epsilon})-\gamma(\epsilon)\le J_x^z(\overline{u}^{1,\epsilon},\overline{u}^{2,\epsilon})\le J_x^z(\overline{u}^{1,\epsilon},u^2)+\gamma(\epsilon),\qquad \forall u^1\in U^1,u^2\in U^2. \tag{9}$$

Moreover,

$$J_x^z(\overline{u}^{1,\epsilon},\overline{u}^{2,\epsilon})=\lambda^T(0)z+\int_0^1\overline{\sigma}(\lambda(t))dt+O(\gamma(\epsilon)), \tag{10}$$

*where $\overline{\sigma}(\lambda)$ was defined in (**??**).*

Remark: As follows from the proof below, one can choose

$$\gamma(\epsilon)=O\left(\max\left\{\Delta(\epsilon),\frac{\epsilon}{\Delta(\epsilon)}\right\}\right),$$

so taking $\Delta(\epsilon)=\epsilon^{1/2}$, one obtains $\gamma(\epsilon)=O(\epsilon^{1/2})$.

3 Proof of main result

The proof is based on the following Lemmas, whose proof is provided in the appendix.

Lemma 3.1 *There exists some constant L such that for any initial distributions ξ, ζ and η on the initial state X_0, and any m,*

$$\begin{aligned}\sigma^m(\lambda,\xi,u^1,f^{2,\lambda})-L &\le \sigma^m(\lambda,\zeta,f^{1,\lambda},f^{2,\lambda}) &(11)\\ &\le \sigma^m(\lambda,\eta,f^{1,\lambda},u^2)+L,\quad \forall u^1\in U^1,u^2\in U^2, &(12)\end{aligned}$$

and

$$|\sigma^m(\lambda,\xi,f^\lambda)-m\overline{\sigma}(\lambda)|\le L, \tag{13}$$

*where f^λ are defined below (**??**), and λ belongs to a bounded set containing $\lambda(t)$, $t\in[0,1]$.*

Lemma 3.2 *The value functions $\overline{\sigma}$, defined in (**??**), are continuous functions of λ.*

Proof of Theorem ??: We first note that for each fixed ϵ, the hybrid dynamic game problem can be formulated as a finite-horizon non-stationary zero-sum stochastic game (see e.g. Nowak [**?**, **?**]), with bounded immediate cost, a countable state space and a finite number of actions. Although we do not pursue this direction, we conclude, that both players may restrict to Markov policies, so that it suffices in (**??**) to restrict to Markov policies u^1 and u^2 (this follows e.g. from Remark 2.1 in [**?**] or Lemma 3.5 in [**?**]).

Due to the linearity of the system, for any $u \in \mathcal{M}$, one can write the value of the hybrid game

$$J_x^z(u) = \lambda^T(0)z + \int_0^1 \lambda^T(t)BE_x^u Y(t)dt$$

which implies the inequality

$$\left| J_x^z(u) - \lambda^T(0)z - \sum_{l=0}^{M_\epsilon - 1} E_x^u \left\{ \lambda^T(\tau_l) B \int_{\tau_l}^{\tau_{l+1}} Y(t)dt \right\} \right| \leq L_1 \Delta(\epsilon), \tag{14}$$

where L_1 is some constant (that does not depend on u, x and z). By (??) we have,

$$E_x^u \left| \lambda^T(\tau_l) B \int_{\tau_l}^{\tau_{l+1}} Y(t)dt - \epsilon \sum_{i=\lfloor \tau_l \epsilon^{-1} \rfloor}^{\lfloor \tau_{l+1} \epsilon^{-1} \rfloor - 1} \lambda^T(\tau_l) B y(X_i, A_i) \right| \leq L_2 \epsilon \tag{15}$$

where L_2 is some constant (that does not depend on u, x and z).

We define for any Markov policy u^i for player i the s-step *shifted strategy* $\theta^j u^i$ by

$$\left(\theta^j u^i\right)_l (a|x) = u_{j+l}(a|x), \qquad \forall l, x, a \in \mathbf{A}_i$$

(we used (??) for the notation of a Markov policy). When both players use Markov policies $u = (u^1, u^2)$, we shall use the notation $\theta^j u = (\theta^j u^1, \theta^j u^2)$. For any Markov policy pair u,

$$E_x^u \left\{ \sum_{i=\lfloor \tau_l \epsilon^{-1} \rfloor}^{\lfloor \tau_{l+1} \epsilon^{-1} \rfloor - 1} \lambda^T(\tau_l) B y(X_i, A_i) \right\} = E_x^u \left\{ \sigma^{m_l} \left(\lambda(\tau_l), X(\lfloor \tau_l \epsilon^{-1} \rfloor), \theta^{\lfloor \tau_l \epsilon^{-1} \rfloor} u \right) \right\} \tag{16}$$

(where σ^m is defined in (??). Notice that by definition of the policies $\overline{u}^\epsilon$,

$$E_x^{\overline{u}^\epsilon} \left\{ \sum_{i=\lfloor \tau_l \epsilon^{-1} \rfloor}^{\lfloor \tau_{l+1} \epsilon^{-1} \rfloor - 1} \lambda^T(\tau_l) B y(X_i, A_i) \right\} = E_x^u \left\{ \sigma^{m_l} \left(\lambda(\tau_l), X(\lfloor \tau_l \epsilon^{-1} \rfloor), f^{\lambda(\tau_l)} \right) \right\} \tag{17}$$

By (??), for any distributions ξ, ζ and η on the state space, and $\forall u^1 \in U^1, u^2 \in U^2$,

$$\begin{aligned} & \sigma^{m_l} \left(\lambda(\tau_l), \xi, \theta^{\lfloor \tau_l \epsilon^{-1} \rfloor} u^1, f^{2,\lambda} \right) - L \\ \leq \quad & \sigma^{m_l} \left(\lambda(\tau_l), \zeta, f^{1,\lambda}, f^{2,\lambda} \right) \\ \leq \quad & \sigma^{m_l} \left(\lambda(\tau_l), \eta, f^{1,\lambda}, \theta^{\lfloor \tau_l \epsilon^{-1} \rfloor} u^2 \right) + L, \quad \forall u^1 \in \mathcal{M}_1, u^2 \in \mathcal{M}_2. \end{aligned}$$

which, along with (??)-(??) implies that

$$\begin{aligned} & E_x^{(u^1,\overline{u}^{2,\epsilon})}\left\{\sum_{i=\lfloor\tau_l\epsilon^{-1}\rfloor}^{\lfloor\tau_{l+1}\epsilon^{-1}\rfloor-1}\lambda^T(\tau_l)By(X_i,A_i)\right\}-L \\ \leq\ & E_x^{(\overline{u}^{1,\epsilon},\overline{u}^{2,\epsilon})}\left\{\sum_{i=\lfloor\tau_l\epsilon^{-1}\rfloor}^{\lfloor\tau_{l+1}\epsilon^{-1}\rfloor-1}\lambda^T(\tau_l)By(X_i,A_i)\right\} \\ \leq\ & E_x^{(\overline{u}^{1,\epsilon},u^2)}\left\{\sum_{i=\lfloor\tau_l\epsilon^{-1}\rfloor}^{\lfloor\tau_{l+1}\epsilon^{-1}\rfloor-1}\lambda^T(\tau_l)By(X_i,A_i)\right\}+L \end{aligned}$$

This, in turn, leads via (??) to

$$\begin{aligned} & E_x^{(u^1,\overline{u}^{2,\epsilon})}\lambda^T(\tau_l)B\int_{\tau_l}^{\tau_{l+1}}Y(t)dt-(L+L_2)\epsilon \\ \leq\ & E_x^{(\overline{u}^{1,\epsilon},\overline{u}^{2,\epsilon})}\lambda^T(\tau_l)B\int_{\tau_l}^{\tau_{l+1}}Y(t)dt \\ \leq\ & E_x^{(\overline{u}^{1,\epsilon},u^2)}\lambda^T(\tau_l)B\int_{\tau_l}^{\tau_{l+1}}Y(t)dt+(L+L_2)\epsilon \end{aligned}$$

and this, via (??), to

$$\begin{aligned} & J_x^z(u^1,\overline{u}^{2,\epsilon})-L_1\Delta(\epsilon)-(L+L_2)\epsilon M_\epsilon \\ \leq\ & J_x^z(\overline{u}^{1,\epsilon},\overline{u}^{2,\epsilon}) \\ \leq\ & J_x^z(\overline{u}^{1,\epsilon},u^2)+L_1\Delta(\epsilon)+(L+L_2)\epsilon M_\epsilon \end{aligned}$$

This proves (??) with

$$\gamma(\epsilon)=L_1\Delta(\epsilon)+(L+L_2)\epsilon M_\epsilon=L_1\Delta(\epsilon)+(L+L_2)[\Delta(\epsilon)^{-1}].$$

Now, from (??) and (??) it follows that

$$\left|E_x^{\overline{u}^\epsilon}\left\{\sum_{i=\lfloor\tau_l\epsilon^{-1}\rfloor}^{\lfloor\tau_{l+1}\epsilon^{-1}\rfloor-1}\lambda^T(\tau_l)By(X_i,A_i)\right\}-m_l\overline{\sigma}(\lambda(\tau_l))\right|\leq L.$$

This, with (??) and (??), imply that

$$\left|J_x^z(\overline{u}^{1,\epsilon},\overline{u}^{2,\epsilon})-\lambda^T(0)z-\sum_{l=0}^{M_\epsilon-1}\overline{\sigma}(\lambda(\tau_l))\epsilon m_l\right|\leq L_1\Delta(\epsilon)+(L+L_2)\epsilon M_\epsilon. \quad (18)$$

By definition of m_l, we have

$$|\epsilon m_l-\Delta(\epsilon)|\leq 2\epsilon. \quad (19)$$

On the other hand, since by Lemma **??**, the function $\overline{\sigma}(\lambda)$ is continuous, it follows that it is uniformly continuous (since we need only consider a compact set of λ), so that

$$\left| \sum_{l=0}^{M_\epsilon - 1} \overline{\sigma}(\lambda(\tau_l))\Delta(\epsilon) - \int_0^1 \overline{\sigma}(\lambda(t))dt \right| = O(\Delta(\epsilon)).$$

This, along with (**??**) and (**??**) establishs (**??**). ■

4 Appendix

Before proving Lemma **??**, we introduce some definitions and quote some results from dynamic programming. Define the matrices $\Pi, D : \mathbf{X} \times \mathbf{X} \to \mathbb{R}$ parametrized by the stationary policies $f = (f^1, f^2)$:

$$\Pi^f(vw) := \lim_{l\to\infty} \frac{1}{l+1} \sum_{i=0}^{l} [P_f^i]_{vw}, \qquad D^f(v,w) = \sum_{l=0}^{\infty} \left([P_f^l]_{vw} - \Pi_{vw} \right).$$

Define the vector $h_\lambda^f = D^f r(\lambda, f)$. Consider a bounded vector of "terminating cost" $\alpha : \mathbf{X} \to \mathbb{R}$, and define the finite horizon expected average cost corresponding to α by

$$\sigma_\alpha^m(\lambda, \xi, u) := E_\xi^u \left[\left(\sum_{i=0}^{m-1} r(\lambda, X_i, A_i) \right) + \alpha(X_m) \right].$$

Define the optimal cost against policy $f^{2,\lambda}$:

$$\sigma_\alpha^m(\lambda, x, f^{2,\lambda}) = \sup_{u^1 \in U^1} \sigma_\alpha^m(\lambda, x, u^1, f^{2,\lambda}).$$

Lemma 4.1 *(i) Under any stationary policy pair f, Π^f is well defined and has identical rows equal to the unique steady state probability under f. Moreover,*

$$\overline{\sigma}(\lambda, f) := \overline{\sigma}(\lambda, v, f) = \sum_{w \in \mathbf{X}} \Pi^f(vw) r(\lambda, w, f),$$

and is independent of $v \in \mathbf{X}$.
(ii) D is well defined and $\sum_{w\in\mathbf{X}} |D^f(v,w)|$ are bounded by some constant $\mathcal{D}$, uniformly over all states v and all stationary policies of both players. Hence $|h_\lambda^f(v)|$ are bounded by some constant $\hat{h}$, uniformly in all stationary policies f, all states v and all λ in some compact set that contains $\lambda(t)$, $t \in [0,1]$.

(iii) The pair $(\overline{\sigma}(\lambda, f^\lambda), h^f_\lambda)$ *is the unique bounded solution (the uniqueness of* h^f_λ *is up to an additive constant) of the dynamic programming equation*

$$\overline{h}(v) + g = \max_{a^1 \in \mathbf{A}_1} \left\{ r(\lambda, a^1, f^{2,\lambda}) + P_{va^1 f^2 w} \overline{h}(w) \right\}.$$

(iv) $\sigma^m_\alpha(\lambda, v, f^{2,\lambda})$ *satisfies the following dynamic programming equation:*

$$\begin{aligned} \sigma^0_\alpha(\lambda, v, f^{2,\lambda}) &:= \alpha(v) \\ \sigma^m_\alpha(\lambda, v, f^{2,\lambda}) &= \max_{a_1 \in \mathbf{A}_1} \left\{ r(\lambda, v, a_1, f^{2,\lambda}) + \sum_{w \in \mathbf{X}} P_{va^1 f^{2,\lambda} w} \sigma^{m-1}_\alpha(\lambda, w, f^{2,\lambda}) \right\} \end{aligned}$$

for all $v \in \mathbf{X}$.

Proof: The proof of (i), (ii) and (iii) are given in Proposition 5.1 in [**?**] (by choosing $\mu = 1$ there). (iv) are well known, see e.g. [**?**] (Note that when player two restricts to a stationary policy, i.e. to $f^{2,\lambda}$, then player 1 is faced with a standard Markov decision process (MDP)). ■

Proof of Lemma ??: We prove the inequality

$$\sigma^m(\lambda, \xi, u^1, f^{2,\lambda}) - L \leq m\overline{\sigma}(\lambda, \xi, f^{1,\lambda}, f^{2,\lambda}).$$

The proof of the other one is the same. Consider the following terminating costs:

$$\alpha(v) = h^f_\lambda(v) + \hat{h}, \qquad v \in \mathbf{X},$$

where $\hat{h}$ is defined in Lemma **??** (ii). It follows from Lemma **??** (ii) that $\alpha \geq 0$. This implies that for any m,

$$\sigma^m_\alpha(\lambda, x, f^{2,\lambda}) \geq \sigma^m(\lambda, x, f^{2,\lambda}). \tag{20}$$

We now compute $\sigma^m_\alpha(\lambda, x, f^{2,\lambda})$ by Lemma **??** (iv):

$$\begin{aligned} \sigma^0_\alpha(\lambda, x, f^{2,\lambda}) &= \alpha(x) \\ \sigma^1_\alpha(\lambda, x, f^{2,\lambda}) &= \max_{a_1 \in \mathbf{A}_1} \left\{ r(\lambda, x, a_1, f^{2,\lambda}) + \sum_{w \in \mathbf{X}} P_{xa^1 f^{2,\lambda} w} \sigma^0_\alpha(\lambda, w, f^{2,\lambda}) \right\} \\ &= \max_{a_1 \in \mathbf{A}_1} \left\{ r(\lambda, x, a_1, f^{2,\lambda}) + \sum_{w \in \mathbf{X}} P_{xa^1 f^{2,\lambda} w} h^f_\lambda(w) \right\} + \hat{h} \\ &= h^f_\lambda(x) + \overline{\sigma}(\lambda, f^\lambda) + \hat{h}, \end{aligned}$$

where the last equality follows from Lemma **??** (iii). We can now establish by recursion that

$$\sigma^m_\alpha(\lambda, x, f^{2,\lambda}) = h^f_\lambda(x) + m\overline{\sigma}(\lambda, f^\lambda) + \hat{h} \leq m\overline{\sigma}(\lambda, f^\lambda) + 2\hat{h}. \tag{21}$$

Combining (??) with (??), we obtain

$$\sigma^m(\lambda, \xi, u^1, f^{2,\lambda}) - 2\hat{h} \leq m\overline{\sigma}(\lambda)$$

for any ξ. The reverse inequality

$$m\overline{\sigma}(\lambda) \leq \sigma^m(\lambda, \xi, u^1, f^{2,\lambda}) + 2\hat{h}$$

is obtained similarly. This implies both (??) and (??). ■

Proof of Lemma ??: From Lemma ?? (i), we have for any $f \in \mathcal{S}_1 \times \mathcal{S}_2$,

$$\overline{\sigma}(\lambda, f) = \sum_w \Pi^f(vw) r(\lambda, w, f)$$

(which in fact does not depend on v), so for any λ_1, λ_2 and any initial distribution ξ,

$$\begin{aligned} |\overline{\sigma}(\lambda_1, \xi, f) - \overline{\sigma}(\lambda_2, \xi, f)| &\leq \sum_v \sum_w \xi(v)\Pi^f(vw)|r(\lambda_1, w, f) - r(\lambda_2, w, f)| \\ &\leq \sup_{w,a} |r(\lambda_1, w, a) - r(\lambda_2, w, a)| \\ &\leq \|\lambda_1 - \lambda_2\| \sup_{w,a} \|By(w, a)\| \end{aligned}$$

Hence, for any initial distribution ξ,

$$\begin{aligned} \overline{\sigma}(\lambda_1) - \overline{\sigma}(\lambda_2) &= \overline{\sigma}(\lambda_1, \xi, f^{1,\lambda_1}, f^{2,\lambda_1}) - \overline{\sigma}(\lambda_2, \xi, f^{1,\lambda_2}, f^{2,\lambda_2}) \\ &\leq \overline{\sigma}(\lambda_1, \xi, f^{1,\lambda_1}, f^{2,\lambda_2}) - \overline{\sigma}(\lambda_2, \xi, f^{1,\lambda_1}, f^{2,\lambda_2}) \\ &\leq \|\lambda_1 - \lambda_2\| \sup_{w,a} \|By(w, a)\| \end{aligned}$$

and, in the same way we obtain

$$\overline{\sigma}(\lambda_2) - \overline{\sigma}(\lambda_1) \leq \|\lambda_1 - \lambda_2\| \sup_{w,a} \|By(w, a)\| .$$

Since y is bounded, we conclude that $\overline{\sigma}(\lambda)$ is continuous in λ. ■

REFERENCES

[1] E. Altman and V. A. Gaitsgory, "Control of a hybrid Stochastic System", *Systems and Control Letters* **20**, pp. 307-314, 1993.

[2] H.A.M. Couwenbergh, "Stochastic Games with metric state space", *J. of Game Theory*, **9**, issue 1, pp. 25-36, 1980.

[3] A. Federgruen, "On N-person stochastic Games with denumerable state space", *Adv. Appl. Prob.* **10**, pp. 452-471, 1978.

[4] K. Hinderer, *Foundations of Non-stationary Dynamic Programming with Discrete Time parameter*, Lecture Notes in Operations Research and Mathematical Systems **33**, Springer-Verlag, 1970.

[5] A. Hordijk, *Dynamic Programming and Markov Potential Theory*, Second Edition, Mathematical Centre Tracts 51, Mathematisch Centrum, Amsterdam, 1977.

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

[7] A.S. Nowak, "Approximation Theorems for zero-sum nonstationary stochastic games" *Proc. of the American Math. Soc.*, **92**, No. 3, pp. 418-424, 1984.

[8] U. Rieder "Non-Cooperative Dynamic Games with General Utility Functions", *Stochastic Games and related topics*, T.E.S. Raghavan et al (eds), pp. 161-174, Kluwer Academic Publishers, 1991.

[9] S. Ross, *Applied Probability Models with Optimization Applications*, Holden-Day, 1970.

[10] F. M. Spieksma, *Geometrically Ergodic Markov Chains and the Optimal Control of Queues*, Ph.D. thesis, University of Leiden, 1990.

[11] J. Van der Wal, "Successive Approximations for Average Reward Markov Games", *Int. J. of Game Theory*, **9**, issue 1, pp. 13-24.