Hybrid Control of a prototypical nonholonomic system

Xiao-Song Yang[1,2], Quan Yuan[2]
[1]Department of Mathematics, Huazhong University of Science and Technology ,Wuhan, 430074, China
[2]Institute for Nonlinear Systems, Chongqing University of Posts and Telecomm,Chongqing 400065, China

Abstract: In this paper we present a simple hybrid control for stabilization of a prototypical nonholonomic system, the computer simulation is provided to verify the efficiency of this hybrid control.

Key words: hybrid control, nonholonomic system.

1 Introduction

Nonholonomic systems most commonly arise in finite dimensional mechanical systems where constrains are imposed on the motion that are not integrable in the sense that the constant can not be written as time derivatives of some smooth function of the generalized coordinates. Nonholonomic control systems result from formulations of nonholonomic systems including control inputs, which have been studied extensively in the literature. An important feature of such nonlinear control systems is that such control problem is not amenable to methods of linear control theory, and more than that, even the conventional theory of nonlinear control systems fails to treat some of nonholonomic control problems. For instance, quite a lot of nonholonomic systems can not be stabilized by time invariant smooth state feedback due to a famous result on necessary condition for stabilization by Brockett [1]. To overcome this problem, several approaches have been proposed to stabilize the origin of (1) in the literature. Typical among these are discontinuous time invariant feedback control, time varying feedback control, and logic-based hybrid feedback control methods [2-6].

In this paper we consider the following well known nonholonomic integrator which is derived from wheeled mobile robot model:

$$\begin{cases} \dot{x}_1 = u_1 \\ \dot{x}_2 = u_2 \\ \dot{x}_3 = x_1 u_2 \end{cases} \tag{1}$$

Our hybrid control is simpler comparative to the hybrid and periodic feedback control proposed by [5] and the logic-base switching proposed by [6]

2 A hybrid control design for non-holonomic integrator

The purpose of this paper is to design a feedback hybrid control law $u = [u_1, u_2]^T$ to stabilize the origin of (1), where $u = [u_1, u_2]^T$ is designed to be of the following form

$$u_{p(t)} = \mathrm{A}_{p(t)} x, \quad p \in \{1,2,3\}, \tag{2}$$

where $p(t^+) = \zeta(x(t), p(t))$ is transition or switching function between the given control laws, which is right continuous. Our control designed is in fact a logic based hybrid control.

Let the control matrices be of the form

$$A_1 = \begin{bmatrix} 0 & 1 \\ -1 & 0 \end{bmatrix}, \ A_2 = \begin{bmatrix} \alpha & -\beta \\ \beta & \alpha \end{bmatrix}, \ \alpha < 0,$$

$$\mathrm{A}_3 = \begin{bmatrix} 0 & -1 \\ 1 & 0 \end{bmatrix}, \tag{3}$$

Then switching among these three control matrices is determined by the transition

function defined as follows:

$p(t^+)=\zeta(x(t),p(t))=1$, if $p(t)=2$ and

$$z \geq 4\|x\|^2 \tag{4a}$$

$p(t^+)=\zeta(x(t),p(t))=2$, if $p(t)=1$ or 3,

and $$-\|x\|^2 \leq z \leq \|x\|^2 \tag{4b}$$

$p(t^+)=\zeta(x(t),p(t))=3$, if $p(t)=2$ and

$$z \leq -4\|x\|^2 \tag{4c}$$

Note that for initial conditions (x,z) satisfying $x=0$ the above control with the transition law (4a)-(4c) can not stabilize the origin, therefore for such initial conditions one first design a control law to drives the initial conditions away from the z -axis. For instance, one can let $u=[c_1,c_2]^T \neq 0$ for a while then convert to the transition law (4a)-(4c)

Finally, for initial conditions $(x(0),z(0))$ satisfying $z(0) \geq 4\|x(0)\|^2$, the initial discrete state is taken to be $p(0)=1$; for initial conditions $(x(0),z(0))$ satisfying

$$-4\|x(0)\|^2 < z(0) < 4\|x(0)\|^2,$$

the initial discrete state is taken to be $p(0)=2$; and for $(x(0),z(0))$ satisfying $z(0) \leq -4\|x(0)\|^2$ the initial discrete state is taken to be $p(0)=3$.

To see this hybrid control works well, let us proceed with the following discussions.

1. If $z \geq 4\|x\|^2$, take $u_1 = \mathrm{A}_1 x$ then we have

$$\begin{cases} \dot{x}_1 = x_2 \\ \dot{x}_2 = -x_1 \\ \dot{z} = -x_1^2 \end{cases}$$

The general solution with initial condition (c_0,c_1,c_2) is

$$x_1 = c_0\cos t + c_1 \sin t, \quad x_2 = -c_0 \sin t + c_1 \cos t,$$

$$z = c_0 c_1 \cos^2 t - \tfrac{1}{4}(c_0^{\,2} - c_1^{\,2})\sin 2t$$
$$-\tfrac{1}{2}(c_0^{\,2} + c_1^{\,2})t - c_0 c_1 + c_2$$

2. If $-\|x\|^2 \leq z \leq \|x\|^2$, take $u_2 = \mathrm{A}_2 x$, then we have

$$\begin{cases} \dot{x}_1 = \alpha x_1 - \beta x_2 \\ \dot{x}_2 = \beta x_1 + \alpha x_2 \\ \dot{z} = x_1(\beta x_1 + \alpha x_2) \end{cases}$$

The general solution with initial condition (c_0,c_1,c_2) is

$$x_1 = c_0 e^{\alpha t}\cos\beta t - c_1 e^{\alpha t}\sin\beta t,$$

$$x_2 = c_0 e^{\alpha t}\sin\beta t + c_1 e^{\alpha t}\cos\beta t, \quad \text{and}$$

$$z = \tfrac{1}{2} c_0 c_1 e^{2\alpha t}\cos 2\beta t + \frac{(c_0^{\,2} + c_1^{\,2})\beta}{4\alpha}(e^{2\alpha t} - 1)$$

$$-\frac{c_0 c_1}{2} + c_2$$

3. If $z \leq -4\|x\|^2$, take $u_3 = \mathrm{A}_3 x$, we have

$$\begin{cases} \dot{x}_1 = -x_2 \\ \dot{x}_2 = x_1 \\ \dot{z} = x_1^2 \end{cases}$$

and the general solution is

$$x_1 = c_0\cos t - c_1 \sin t, \quad x_2 = c_0 \sin t + c_1 \cos t$$

$$z = c_0 c_1 \cos^2 t + \tfrac{1}{4}(c_0^{\,2} - c_1^{\,2})\sin 2t$$
$$+\tfrac{1}{2}(c_0^{\,2} + c_1^{\,2})t - c_0 c_1 + c_2$$

From above formulas one can prove the asymptotic

stability of the closed loop system when hybrid control is applied, the proof overloaded with details is omitted here. For details see [7].

3 Computer Simulation

Note that the four surfaces $z = 4\|x\|^2$, $z = \|x\|^2$, $z = -\|x\|^2$ and $z = -4\|x\|^2$ play the role of making transition among the different control laws. To demonstrate that the hybrid control designed does work well, let us take initial value $(x(0), z(0))$ = (1,1,10), $p(0) = 1$ for instance, and let a =-1,ß= 1, the simulation is illustrated in the following figures.

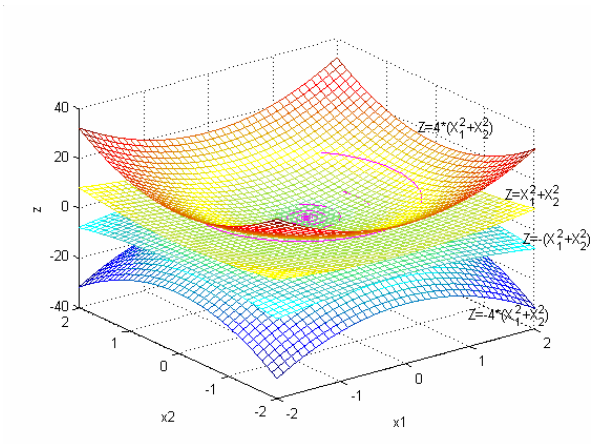

Fig.1 The four transition manifolds (surfaces)

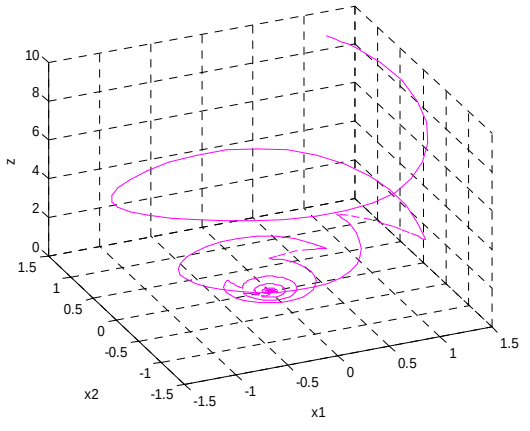

Fig.2 The trajectory tends to the origin in state space.

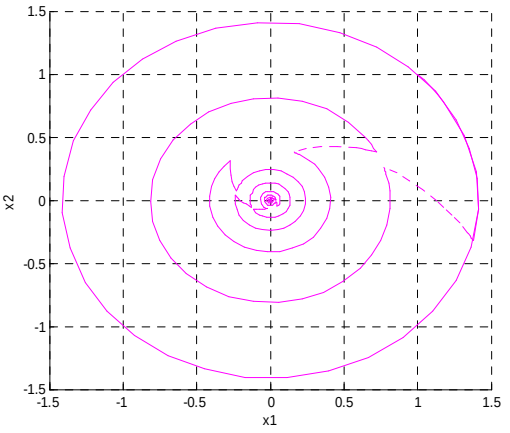

Fig.3 The projection of the trajectory onto the x_1 x_2 plane.

The solid line indicates the dynamics of the controlled system with $u_1 = A_1 x$, and the dashed line indicates the dynamics of the control system with control $u_2 = A_2 x$.

4 Conclusion

We have given a logic-based hybrid control law which is simpler than the other control laws designed for this nonholonomic integrator. We provide a computer simulation to demonstrate the efficiency of this control law , As for the rigorous proof, we are developing a general theory [7]that includes the result of this paper as a special case.

References

1 R.W. Brockett, Asymptotic stability and feedback stabilization, in R.W. Brockett, R.S. Milman and H.J. Sussmann, editors, Differential Geometric Control Theory, Birkhauser, Boston, 1983, 181-191.

2 I. Kolmanovsky and N.H. McClamroch, Developments in nonholonomic control problems, IEEE Control Systems,1995, 20-36.

3 N.H. McClamroch, I. Kolmanovsky, and M. Reyhanoglu, Hybrid closed loop systems: A nonlinear control perspective, Proc. of the 36the IEEE CDC, San Diego, 1997, 114-119.

4 Y.P. Tian and S.H. Li, Smooth exponential stabilization of nonholonomic systems via time-varying feedback , IEEE Conference on Decision and Control, Sydney, 2000, 1912-1917.

5 N.H. McClamroch, and I. Kolmanovsky, Performance benefits of hybrid control designe for linear and nonlinear systems, Proceedings of the IEEE, 88(7) (2000), 1083-1096.

6 J.P. Hespanha and A. Stephen Morse，Stabilization of Nonholonomic Integrators via Logic-Based Switching ? The13th World Congress of the Int. Federation of Automat. Contr., June 1996.

7 Xiao-Song Yang, Hybrid controller for a class of Chaplygin type nonholonomic systems, in preparation.

Time Separation for Cyclic Event Rule Systems *

Qianchuan Zhao and Jianfeng Mao

Department of Automation, Tsinghua University, Beijing 100084

Abstract. The analysis of the time separation of events is a fundamental problem in design and evaluation of asynchronous circuits and other real-time systems. Important progresses have been made based on the event rule system model. In this paper, we report a sufficient and necessary structural condition for the boundedness of time separation of events in cyclic event rule systems with both min and max constraints, i.e. uniformity.

Keywords Event rule systems, uniform system, cyclic timing constraint graph, time separation.

1. Introduction

Analysis of time separation for event rules systems plays a key role in developing asynchronous circuits. Since the framework of event rule systems was introduced for performance analysis of asynchronous circuits by Burns in [1], many advances have been made.

For acyclic event rule systems, the problems of interests include:

A1. Develop algorithms to determine the exact time separations of given pairs of events.

A2. If the exact determination of separation is hard, develop fast approximation algorithms.

For **A1**, in [2], a polynomial algorithm to calculate exact time separations was developed for acyclic event rule systems with max-only constraints. It was also shown in [2] that the exact time separations for general acyclic event rule systems with both min and max constraints are NP-complete to determine. In [3] and [4], algorithms are developed for analyzing acyclic systems with max and linear constraints. The algorithms are conjectured to run in polynomial-time. The exact complexity of the max and linear problem is unfortunately, unknown. For **A2**, in [5], McMillan and Dill's time separation algorithm was extended to acyclic event rule systems with both min and max constraints.

For cyclic event rule systems, the problems of interest include:

C1. Establish conditions under which the long-term time separations of events are bounded.

C2. Develop algorithms to determine time separations of events or their upper bounds.

For **C1**, based on a modification of McMillan and Dill's time separation algorithm and deep algebraic observations, [6] established that for a class of cyclic max-only event rule systems known as well-formed (See [7] for the original definition of well-formed systems.) strongly connected systems, the long-term time separations are bounded. [5] extended the notion of strong connectivity for well-formed max-only event rule systems to the condition of tightly coupled for well-formed event rule systems with both min and max constraints and proved that the long-term time separations are bounded for well-formed tightly coupled systems. For **C2**, in [8], a polynomial-time algorithm was described for approximate timing analysis of max-only systems with repeated events. [6] proposed an exact algorithm to determine the exact time separations for cyclic well-formed strongly connected max-only event rule systems. In [5], pseudopolynomial algorithm was proposed to determine an upper bound for the long-term time separations of events for cyclic event rules systems with both min and max constraints.

In this paper, we concern **C1** and present a new boundedness condition on long-term separations for cyclic event rule systems with min and max constraints. The new condition is captured by the notion

*This work was supported in part by NSFC(Grant No.60074012,60274011), National Key Project of China, Fundamental Research Funds from Tsinghua University and Chinese Scholarship Council, Ministry of Eduction of China. Email:zhaoqc@tsinghua.edu.cn

of uniformity which was introduced in [9] in the context of stochastic min-max systems. We first present the formal definition of uniform systems and show that the tightly coupled systems proposed in [5] are a special class of uniform systems.

2. Basic definitions

We follow the notations of [5] on the cyclic timing constraint graph representations of event rule systems and the notations of [10] on min-max systems.

A. Cyclic Timing Constraint Graphs

A cyclic timing constraint graph is a directed, labelled graph $G = (V, E)$. In general, the graph has two components—an acyclic component modelling the behavior of the system immediately after it is powered up or reset, and a cyclic component modelling the subsequent repeated behavior. A vertex in the acyclic component represents a single occurrence of an event, whereas a vertex in the cyclic component represents infinite occurrences of an event. As in [5] and [7], let us denote the k-th occurrence of event v by v_k and call k the *occurrence index* of v_k. v_1 is understood as the first occurrence of v. For convenience, we do not distinguish between a vertex and the event represented by the vertex. It is assumed that there is a unique RESET event (also known as the *root*) which has no incoming edges. The set of vertices V is divided into two disjoint subsets: the set of min vertices (shown as circles in figures) and the set of max vertices (shown as squares in figures). The types of timing constraints are different for min and max vertices. More specifically, every edge, $\langle u, v\rangle$, from u to v represents a timing constraint caused by u on the occurrence of v which is quantified by the interval label $[d_{u,v}, D_{u,v}]$ on the edge. The delay, denoted by $\delta_{u,v}$, in the propagation of the effect of event u to the component that generates event v, can take any value within the fixed lower bound $d_{u,v}$ and upper bound $D_{u,v}$. For the bounds, we assume that $0 \le d_{u,v} \le D_{u,v} < +\infty$. Let τ_{v_k} denote the time of occurrence of event v_k. The mathematical formulation of the timing constraint for min and max vertices can be given as follows. If v_k is a max vertex, then

$$\tau_{v_k} = \max_{u\in \mathrm{preds}(v_k)} (\tau_u + \delta_{u,v_k}), \quad d_{u,v_k} \le \delta_{u,v_k} \le D_{u,v_k}, \tag{1}$$

where as usual, $\mathrm{preds}(v_k)$ is the set of vertices having an edge to v_k. Sometimes, we use $\tau_{v_k} = \max_{u\in \mathrm{preds}(v_k)}(\tau_u + [d_{u,v_k}, D_{u,v_k}])$ for short. If v_k is a min vertex, the timing constraint for v_k is

$$\tau_{v_k} = \min_{u\in \mathrm{preds}(v_k)} (\tau_u + \delta_{u,v_k}), \quad d_{u,v_k} \le \delta_{u,v_k} \le D_{u,v_k}. \tag{2}$$

and we write $\tau_{v_k} = \min_{u\in \mathrm{preds}(v_k)}(\tau_u + [d_{u,v_k}, D_{u,v_k}])$ for short. The set of edges is also divided into two disjointed subsets: the set of *marked* edges and the set of *unmarked* edges. The timing constraint on a marked edge from u to v is effective for the occurrence time of u_k on the occurrence time of v_{k+1} but for unmarked edge from u to v, it is effective for the occurrence time of u_k on the occurrence time of v_k. Now, we see if every delay is assigned a value within the corresponding interval, the time of occurrence of every event is uniquely determined. As pointed out in [5], the timing constraint graphs cannot model systems with conflict or choice.

Definition 1 *[5][7] A timing constraint graph is well-formed if every cycle has at least one marked edge and for every event v in the cyclic component, there exists at least one cycle with exactly one marked edge.*

Example 1 *Figure 1 is an example of cyclic timing constraint graphs. It is inspired by an example in [11].*

A well-formed cyclic timing constraint graph G can be equivalently represented by an infinite acyclic graph known as its unfolded graph G^*(See [7] or [5] for details). As an example, the unfolded graph G^* of the graph G in Figure 1 is shown in Figure 2.

B. Uniform Systems

In the unfolded graph $G^* = (V^*, E^*)$, define a path from u to v as a sequence of one or more events $(u, \ldots, x, y, \ldots, v)$ such that $\langle x, y\rangle$ is an edge in E^* for each pair of consecutive events x and y in the sequence. We assume that (u) is a path from event u to itself. Event v is said to be reachable from u if there is a path from u to v. v is said to be reachable from a set of events D, if there is a u in D such that v is reachable from u. The set of events that are reachable from D is denoted $\mathcal{R}(D)$. The set of events that are not reachable from D is denoted $\mathcal{R}'(D)$. A finite set of events D is called a cutset of G^* if every path from RESET to every event reachable from D pass through at least one event in D. We introduce the following mild assumption on the structure of G^* as used in the definition of tightly coupled systems [5].

Assumption 1 *There exists a sequence of cutsets C_i for all $i \ge 1$, that satisfies the following properties:*

P1 there are finitely many events that are not reachable from C_1.

P2 for all $i \ge 1$,

$$C_i = \{v_i : v_1 \in C_1\}.$$

P3 for all i different from j, C_i is disjoint from C_j.

In the rest of this paper, we only consider systems satisfying Assumption 1. Let $G_i = (V_i, E_i)$ denote the subgraph between C_i and C_{i+1}, namely, $V_i = \mathcal{R}(C_i) \cap (\mathcal{R}'(C_{i+1}) \cup C_{i+1})$ and $E_i = \{\langle u, v\rangle : u, v \in V_i \text{ and } \langle u, v\rangle \in E^*\}$.

One obvious but extremely useful fact about the structure of timing constraint graphs with min and max events is that for each timing constraint graph G, we can associate a unique Monotone Boolean network $\bar{G}$ which we refer to as the *skeleton* of G in the following. By monotone, we mean only two logical operations OR and AND are involved. For each event v in G, if v is a max event, then v is assigned as an OR vertex in $\bar{G}$; if v is a min event, then v is assigned as an AND vertex in $\bar{G}$. In $\bar{G}$, all delay labelled on the edges of G are simply dropped and keep the marked edges as marked and unmarked edges unmarked. The notions of acyclic component, cyclic component and the unfolded graph are defined in the obvious way. In the unfolded graph $\bar{G}^*$, we attach a value σ_{v_k} to each vertex v_k. These values are determined in the following way. If v_k is an OR vertex, then

$$\sigma_{v_k} = \vee_{u \in \text{preds}(v_k)} \sigma_u, \tag{3}$$

Here $\vee$ is the logical operation OR. Similarly, if v is an AND vertex,

$$\sigma_{v_k} = \wedge_{u \in \text{preds}(v_k)} \sigma_u. \tag{4}$$

Here $\wedge$ is the logical operation AND. Note if we regard the Boolean space as a subset of real space, $\vee$ can be interpreted as max and $\wedge$ can be interpreted as min.

Denote $X_i = (\sigma_{u_i}, u_i \in C_i)_{n\times 1} \in \mathbb{B}^n$ where $n = |C_i|$. It is evident from the construction of $\bar{G}^*$ that the values X_i can be expressed as a monotone Boolean function of X_1. Denote this function $\bar{F}^{i-1}$. By definition $\bar{F}^0$ is the identity function. In terms of the composition of functions, it is easy to verify that $\bar{F}^i = \bar{F}(\bar{F}^{i-1})$, for all $i \geq 1$. For the example in Figure 1, we have

$$X_2 = \bar{F}(X_1) = \begin{pmatrix} \sigma_{a_1} \vee (\sigma_{b_1} \wedge \sigma_{c_1}) \\ \sigma_{a_1} \vee \sigma_{b_1} \vee \sigma_{c_1} \\ \sigma_{a_1} \vee \sigma_{b_1} \vee \sigma_{c_1} \end{pmatrix},$$

$$X_3 = \bar{F}^2(X_1) = \begin{pmatrix} \sigma_{a_1} \vee \sigma_{b_1} \vee \sigma_{c_1} \\ \sigma_{a_1} \vee \sigma_{b_1} \vee \sigma_{c_1} \\ \sigma_{a_1} \vee \sigma_{b_1} \vee \sigma_{c_1} \end{pmatrix}$$

where $X_i = (\ \sigma_{a_i} \quad \sigma_{b_i} \quad \sigma_{c_i}\ )^\tau$.

Definition 2 *A well-formed cyclic timing constraint graph G satisfying Assumption 1 is said to be uniform, if there exists a non-negative integer $r \geq 1$ and a monotone Boolean function $\bar{g}(X_1) : \mathbb{B}^n \to \mathbb{B}$ such that for all $X_1 \in \mathbb{B}^n$,*

$$\bar{F}^r(X_1) = \bar{g}(X_1)\mathbf{1}, \tag{5}$$

where $\bar{g}(X_1)\mathbf{1}$ is an n-dimension vector whose elements are uniformly $\bar{g}(X_1)$. n is the size of C_1. The minimal such integer r is called the height of G.

The system in Figure 1 is uniform since Assumption 1 is satisfied and $\bar{F}^2(X_1) = \bigvee_{u_1 \in C_1} \sigma_{u_1}\mathbf{1}$. Recall the definition of tightly coupled systems introduced in [12] and [5].

Definition 3 *[12][5] A timing constraint graph is tightly coupled, if the unfolded graph has a sequence of cutsets, C_i for all $i \geq 1$, such that Assumption 1 and the following condition is satisfied:*

P4 for every u in C_i and every v in C_{i+1}, there exists at least one path from u to v, such that all events along the path are associated with max constraints.

Proposition 1 *Tightly coupled systems are uniform systems with height $r = 1$.*

Proof. Omitted.

From Proposition 1, we can find that tightly coupled system is a special case of uniform system. And not all uniform systems are tightly coupled. In the system of Figure 1, there is no path from b_1 and c_1 in C_1 to a_2 containing only max events, so it is not tightly coupled. But as we have shown before, it is a uniform system.

C. The problem

The time separation problem for general timing constraint graphs is determining upper bounds on the time separations of all ordered pairs of events in G_i, maximized over all $i \geq 0$.

3. The main results

Before we show the boundedness results, we need define some convenient notations. Assume $|C_i| = n$ and $|G_i| = m$. Denote $\Delta_i(u_i, v_i)$ as the exact bound of time separation from the event u_i to v_i in C_i and let $\Delta_i = (\Delta_i(u_i, v_i); u_i, v_i \in C_i)_{n\times n}$. Similarly, define $\Omega_i(u_i, v_i)$ as the exact bound of time separation from the event u_i to v_i in G_i and $\Omega_i = (\Omega_i(u_i, v_i); u_i, v_i \in G_i)_{m\times m}$. Denote $\Pi(u)$ as the set of paths from events in cutset C_1 to u. The shortest and longest path lengths to u, denoted $l(u)$ and $L(u)$, are defined as follows: $l(u) = \min_{P\in\Pi(u)}(\sum_{\langle x,y\rangle \text{along } P} d_{x,y})$ and $L(u) =$

$\max_{P\in\Pi(u)}(\sum_{\langle x,y\rangle \text{along } P} D_{x,y})$. Let $\Delta_i^*(u_i,v_i) = L(v_i) - l(u_i)$ and $\Delta_i^* = (\Delta_i^*(u_i,v_i); u_i, v_i \in C_i)_{n\times n}$.

The problem in Section II can be decomposed into two problems below:

1. Compute the upper bounds of Ω_i for all $i = 0, ..., K_0$;
2. Compute the upper bounds of Ω_i for all $i \geq K_0 + 1$.

The first problem relates to the initial behavior of the system while the second one relates to the long term behavior. Since the first problem corresponds to a finite unfolding graph, its upper bounds must be finite for the systems with finite edge delays. So the second problem becomes the key part to determine whether the upper bounds of $\Omega_i(u_i, v_i)$ are finite or not. Furthermore, if the upper bounds of $\Delta_i(u_i, v_i)$ are finite, the bounds of $\Omega_i(u_i, v_i)$ must be finite for the system with finite edge delays. So the key point of boundedness condition is to study the long term behavior of $\Delta_i(u_i, v_i)$. For uniform systems, the upper bounds of $\Delta_i(u_i, v_i)$ are finite as shown in the following theorem.

Theorem 1 *Let G be a uniform system with height r. The value of $\Delta_i(u_i, v_i)$ is bounded above by $\Delta^*_{r+1}(u_{r+1}, v_{r+1})$ for all $i \geq r + 1$ and $u_i, v_i \in C_i$ regardless of the value of Δ_1.*

Proof. Omitted.

From Theorem 1, $\Omega_i(u, v)$ $(i \geq r + 1)$ must also be finite for all uniform systems. A nature question is raised here, i.e., how about the long term behavior of non-uniform system. It can be answered by the theorem below.

Theorem 2 *For all non-uniform systems, there must exist at least one pair of events $u_i, v_i \in C_i$ such that $\Delta_i(u, v)$ becomes $+\infty$ with some special edge delays when i goes to $+\infty$.*

Proof. Omitted.

From Theorem 1 and 2, we can conclude the uniformity is a sufficient and necessary structural condition for the boundedness of time separation of events in cyclic event rule systems with both min and max constraints. It should be noted that the non-uniform system may still have bounded time separations for specific set of edge delays.

4. Concluding remarks

In this paper, our main contribution is identifying a sufficient and necessary structural condition for Problem **C1**. For Problem **C2**, based on Theorem 1, we will develop algorithms to get finite upper bounds for all uniform systems.

Acknowledgement

The authors would like to thank Prof. Da-Zhong Zheng for his encouragement in the study of min-max systems. The first author would like to thank Prof. Bruce Krogh for his hospitality during the first author's visit to ECE department of Carnegie Mellon University as a visiting scholar.

References

[1] S. Burns, "Performance analysis and optimization of asynchronous circuits," Ph.D. dissertation, California Inst. Technol., Pasadena, 1991.

[2] K. McMillan and D. Dill, "Algorithm for interface timing verification," in *Proc. IEEE Int. Conf. Computer Design: VLSI in Computers and Processors*, 1992, pp. 48–51.

[3] E. Walkup, "Optimization of linear max-plus systems with application to timing analysis," Ph.D. dissertation, Univ. Washington, Seatle, 1995.

[4] T.-Y. Yen, A. Ishii, A. Casavant, and W. Wolf, "Efficient algorithms for interface timing verification," *Formal Meth. Syst. Design*, vol. 12, no. 3, pp. 241–265, 1998.

[5] S. Chakraborty, K. Y. Yun, and D. L. Dill, "Timing analysis of asynchronous systems using time separation of events," *IEEE Transactions on Computer-aided Design of Integerated Circuits and Systems*, vol. 18, no. 8, pp. 1061–1076, 1999.

[6] H. Hulgaard, S. M. Burns, T. Amon, and G. Borriello, "An algorithm for exact bounds on the time separation of events in concurrent systems," *IEEE Transactions on Computers*, vol. 44, no. 11, pp. 1306–1317, 1995.

[7] T. Amon, H. Hulgaard, G. Borriello, and S. Burns, "Timing anaylsis of concurrent systems," Univ. Washington, Seattle, Tech. Rep. UW-CS-TR-92-11-01, 1992.

[8] C. Myers and T.-Y. Meng, "Synthesis of timed asynchronous circuits," *IEEE Trans. VLSI Syst.*, vol. 1, no. 2, pp. 106–119, 1993.

[9] Q. Zhao, D.-Z. Zheng, and X. Zhu, "Structure properties of min-max systems and existence of global cycle time," *IEEE Trans. Automat. Contr.*, vol. 46, no. 1, pp. 148–151, 2001.

[10] J. Gunawardena, "Min-max functions," *J. DEDS*, vol. 4, pp. 377–406, 1994.

[11] S. Gaubert and J. Gunawardena, "Existence of eigenvectors for monotone homogeneous functions," HP Lab, Tech. Rep. HPL-BRIMS-99-08, 1999.

[12] S. Chakraborty, "Polynomial-time techniques for approximate timing analysis of asynchronous systems," Ph.D. dissertation, Stanford Univ., Stanford, CA, 1998.

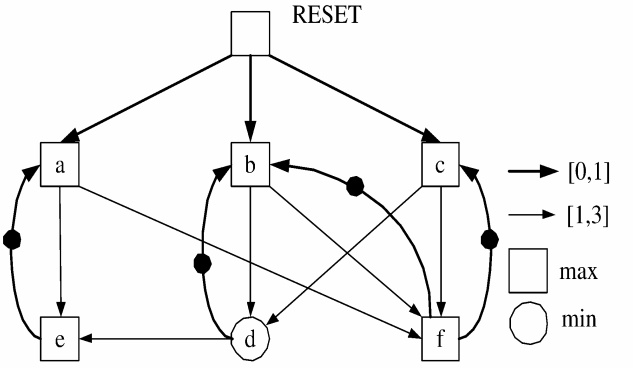

Figure 1: A cyclic timing constraint graph

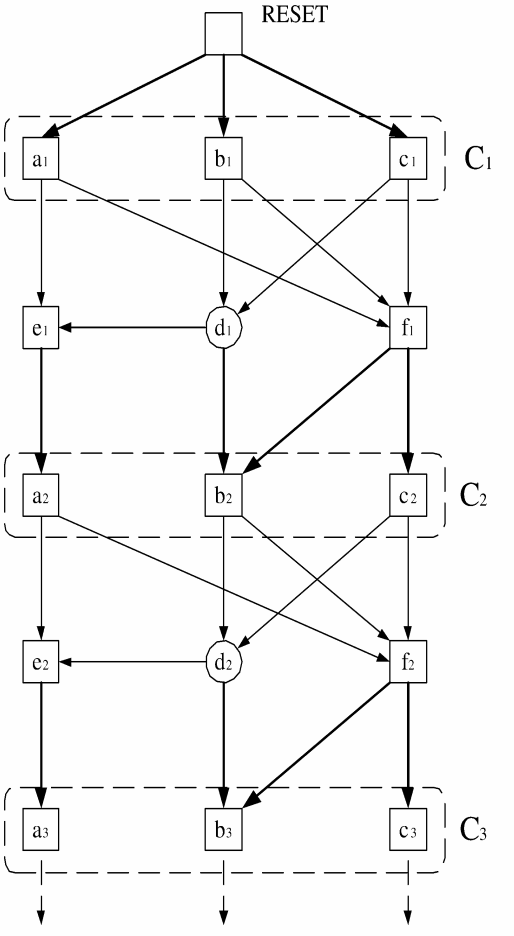

Figure 2: The unfolded graph G^* for the graph in Figure 1.

半马尔可夫控制过程基于全局优化的最优鲁棒控制策略求解

刘 春 唐 昊 高 隽
(合肥工业大学 计算机与信息学院，安徽合肥 230009)
E-mail: tangh@ustc.edu

摘要：论文考虑一类不确定性半马尔可夫控制过程（SMCP）在参数相关或不相关时的鲁棒控制问题，重点研究了全局优化方法在其最优鲁棒控制策略求解过程中的应用。针对参数（策略）空间离散和连续两种情况，分别采用了不同的全局优化算法，即模拟退火算法和填充函数法，进行求解。数值例子说明，全局优化方法保证了平均准则和折扣准则下的计算结果之间当折扣因子趋近于零时的极限关系成立。

关键词：半马尔可夫控制过程；最优鲁棒控制策略；全局优化；模拟退火算法；填充函数法

Solution of Optimal Robust Control Policy for Semi-Markov Control Processes Based on Global Optimization

Liu Chun Tang Hao Gao Jun
(School of Computer and Information, Hefei University of Technology, Heifei, 230009)

Abstract: In this paper, we consider robust control problems for a class of semi-Markov control processes (SMCPs) with dependent or independent uncertain parameters, and focus on the application of global optimization methods to derive the optimal robust control policy. Different global optimization methods, such as simulated annealing and filled function approaches, are adopted respectively in the cases of both discrete and continuous parameter (or policy) spaces. The numerical examples show that, with the application of these global optimization methods, an average-cost problem is the limitation of a discounted problem as the discount factor goes to zero.

Key words: semi-Markov control processes; robust control policy; global optimization; simulated annealing; filled function

1 引 言

实际中的许多序贯决策问题，都可模型化为半 Markov 控制过程（SMCP），其性能分析和优化是当前离散事件动态系统（DEDS）领域的一个研究热点。近来，文献[1]首次提出并建立了 SMCP 的性能势概念和理论，给出了 SMCP 等价无穷小生成子的定义，指出 SMCP 通过适当转换可以看作是一个等价 Markov 控制过程（MCP），并证明 SMCP 的平均模型是折扣模型当折扣因子趋于零时的极限。殷保群博士随后也给出了类似的结果[2-3]。于是，可以发展一些基于性能势的优化算法，如策略迭代和梯度方法等[1-3]。但是，实际的半 Markov 系统有时面临两个问题：一是系统模型参数确定，但不全知；二是由于干扰摄动等原因，系统模型参数本身不确定。这两种参数不确定性，最终导致等价无穷小转移速率不确定。因此，基于系统精确参数的优化方法不再适用。尽管实际系统存在不确定性，但系统参数的变化范围一般是有界并可知的，因而考虑其鲁棒控制问题是一项有意义的研究工作，即考虑在系统参数变化有可能取最坏值的情况下如何采取最优控制策略问题，以得到极坏情况下的最优性能值。

文献[4]在文献[5]的基础上，讨论了一类遍历的 MCP 基于性能势的策略迭代算法，以求解不确定参数不相关条件下平均代价 MCP 的最优鲁棒控制策略，根据文献[1,2]，其结果可拓展到平均或折扣

*合肥工业大学中青年科技创新群体计划和安徽省优秀青年科技基金资助(04042044)

准则 SMCP 在参数不相关时的鲁棒控制问题；文献[6]讨论了参数相关情况下，SMCP 在平均准则和折扣准则下的鲁棒控制问题。理论上，仅用局部优化求解多极值鲁棒控制问题一般得不到较好的结果，也不能反映文献[1]中描述的平均准则和折扣准则之间当折扣因子趋近于零时的极限关系。因此，本文结合文献[4,6]，运用全局优化算法，分析了参数在相关和不相关时，两种准则下 SMCP 基于性能势的最优鲁棒控制策略求解过程，通过实例说明了全局优化在研究折扣准则和平均准则之间关系中的重要作用。

2 问题描述和基本理论

五元组 $X(\theta)=(X_t,\Phi,D,Q^{\nu}(t,\theta),f^{\nu})$ 表示一个 SMDP，其中 f^{ν} 的定义同文献[7]，其它各符号的定义同文献[6]。在 $X(\theta)$ 的一条样本轨道上，对任意 $t\in[T_n,T_{n+1})$，令 $Y_t=X_{n+1}$，则 $X(\theta)$ 的平均性能准则定义为无穷水平平均代价的期望值：

$$\eta^{\nu}(\theta)=\lim_{T\to\infty}\frac{1}{T}E\left\{\int_0^T f(X_t,Y_t,v(X(t)))dt\right\},v\in\Omega_s.$$

由文献[1]，$X(\theta)$ 的折扣性能准则定义为如下无穷水平总折扣代价的期望值：

$$\eta_{\alpha}^{\nu}(i,\theta)=\alpha\lim_{N\to\infty}E\left[\int_0^T e^{-\alpha t}f(X_t,Y_t,v(X_t))dt\mid X_0=i\right]$$

这里 $\alpha>0$ 是折扣因子，不失一般性，本文中令 $\alpha\in[0,a],a>0$。记 $\eta_{\alpha}^{\nu}(\theta)=(\eta_{\alpha}^{\nu}(1,\theta),\mathrm{L},\eta_{\alpha}^{\nu}(M,\theta))^{\tau}$。“$\tau$”在本文中表示转置。对任意 $\alpha\ge0$，定义矩阵 $A_{\alpha}^{\nu}(\theta)=\alpha I-(H_a^{\nu}(\theta))^{-1}(I-Q_{\alpha}^{\nu}(\theta))$[1]，其中各符号的定义同文献[6]。$A_{\alpha}^{\nu}(\theta)$ 即为 SMDP 的等价无穷小生成子。定义 $f_{\alpha}^{\nu}=(\Psi_{\alpha}^{\nu}(\theta)\mathrm{e}\ f^{\nu})e$，$\alpha>0$，其中各符号的定义同文献[7]。

假设在任意 $v\in\Omega_s$ 下，$X(\theta)$ 都是遍历的，且 $X(\theta)$ 的稳态分布为向量 $\pi_{\alpha}^{\nu}(\theta)$，则 $\pi_{\alpha}^{\nu}(\theta)$ 为方程

$$\pi_{\alpha}^{\nu}(\theta)A_{\alpha}^{\nu}(\theta)=0,\pi_{\alpha}^{\nu}(\theta)e=1 \quad (1)$$

的唯一解。定义 $\lambda^{\nu}(\theta)=\max_{i\in\Phi,\alpha\in[0,a]}(h_{\alpha}^{\nu}(i,\theta))^{-1}$，其中 $h_{\alpha}^{\nu}(i,\theta)$ 的定义同文献[6]。对任一 $\alpha\ge0$，定义

$$(\alpha I-A_{\alpha}^{\nu}(\theta)+\lambda^{\nu}(\theta)e\pi_{\alpha}^{\nu}(\theta))g_{\alpha}^{\nu}(\theta)=f_{\alpha}^{\nu} \quad (2)$$

为 SMCP 在策略 v 下的 Poisson 方程[7]，$g_{\alpha}^{\nu}(\theta)$ 是性能势向量。根据文献[1,7]，有

$$\eta_{\alpha}^{\nu}(\theta)=f_{\alpha}^{\nu}+A_{\alpha}^{\nu}(\theta)g_{\alpha}^{\nu}(\theta),\alpha\ge0 \quad (3)$$

在一些实际的半 Markov 系统中，对一个固定的策略 v，参数 θ 是固定且未知的或是随时间慢变化的。假设决策者确知在任何策略 $v\in\Omega_s$ 下，参数 θ 的实际变化范围在集合 Θ^{ν} 上，他需要考虑当 θ 的取值最不利于系统获得最佳性能值时选择策略 v^*，以满足 $v^*\in\arg\min_{v\in\Omega_s}\max_{\theta\in\Theta^{\nu}}\eta_{\alpha}^{\nu}(\theta)$ 或 $v^*\in\arg\min_{v\in\Omega_s}\max_{\theta\in\Theta^{\nu}}\eta^{\nu}(\theta)$。

3 全局优化在最优鲁棒控制中的应用

鉴于平均准则是折扣准则的特例，以下仅讨论折扣准则最优鲁棒控制问题。

对于一个具有 M 个状态的系统，求解最优鲁棒控制策略问题可分为两个子问题。

问题一是对给定的策略 $v\in\Omega_s$，确定一个最不利于系统获得最佳性能值的参数 $\theta^{\nu}\in\Theta^{\nu}$，满足

$$\theta^{\nu}\in\arg\max_{\theta\in\Theta^{\nu}}\eta_{\alpha}^{\nu}(\theta) \quad (4)$$

通过方程(4)建立策略 v 与 $\eta_{\alpha}^{\nu}(\theta^{\nu})$ 的一一对应关系。记 $\eta_{\alpha}^{\nu}(\theta^{\nu})$ 为 $\hat{\eta}_{\alpha}^{\nu}$。

问题二是根据求解问题一得到的性能值，寻求一最优鲁棒控制策略 v^*，满足 $v^*\in\arg\min_{v\in\Omega_s}\hat{\eta}_{\alpha}^{\nu}$。

在求解最优鲁棒控制策略问题的两个子问题中涉及寻求满足最优值条件的参数（或策略）。若是单极值问题，可用局部优化算法求解；若是多极值问题，必须采用全局优化算法求解，否则无法寻得较为理想的策略，其对应的性能值也无法反映文献[1]中描述的平均准则和折扣准则之间的关系。对于离散参数（或策略）空间，可采用模拟退火算法进行全局优化[8]；对于连续参数（或策略）空间，可采用填充函数法进行全局优化[9]。

3.1 不相关不确定参数最优鲁棒控制策略求解

在不相关不确定参数条件下，根据文献[4]，最优鲁棒控制策略可通过策略迭代算法求解。首先构造下列算法，记为算法一。

[2] 定义过程根据文献[2]，并且易证其定义同曹希仁教授的定义本质上等价[1]。

(1) 任选一可行的 $\bar{\theta} \in \Theta^{v}$，求得其对应的无穷小转移速率矩阵 $\bar{A}_{\alpha}^{v}$。

(2) 利用 $\bar{A}_{\alpha}^{v}$ 根据方程(1)，求解对应的平稳分布 $\bar{\pi}_{\alpha}^{v}$，再根据方程(2)来求解对应的性能势 $\bar{g}_{\alpha}^{v}$。

(3) 求一 $\tilde{\theta} \in \arg\max_{\theta \in \Theta^{v}} \{A_{\alpha}^{v} \bar{g}_{\alpha}^{v}\}$，并求得其对应的转移速率矩阵 $\tilde{A}_{\alpha}$。

(4) 若 $\tilde{A}_{\alpha} = \bar{A}_{\alpha}^{v}$，则停止；否则令 $\bar{A}_{\alpha}^{v} = \tilde{A}_{\alpha}$，转(2)。

再构造下列策略迭代算法，记为算法二。

(1) 任选一初始策略 $v_0 \in \Omega_s$，调用算法一，得到与 $\hat{\eta}_{\alpha}^{v_0}$ 对应的 $\hat{A}_{\alpha}^{v_0}$ 和 $\hat{g}_{\alpha}^{v_0}$，并计算 $\hat{\eta}_{\alpha}^{v_0}$；令 $k = 0$。

(2) 求一个 $v_{k+1} \in \arg\min_{v \in \Omega_s} \max_{\theta \in \Theta^{v}} \{f_{\alpha}^{v} + A_{\alpha}^{v} \hat{g}_{\alpha}^{v_k}\}$。

(3) 针对策略 v_{k+1} 调用算法一，得到与 $\hat{\eta}_{\alpha}^{v_{k+1}}$ 对应的 $\hat{A}_{\alpha}^{v_{k+1}}$ 和 $\hat{g}_{\alpha}^{v_{k+1}}$，并计算 $\hat{\eta}_{\alpha}^{v_{k+1}}$。

(4) 若 $\hat{\eta}_{\alpha}^{v_{k+1}} = \hat{\eta}_{\alpha}^{v_k}$，则停止；否则，令 $v_k := v_{k+1}$，$k := k+1$，转(2)。

在算法一的(3)和算法二的(2)中，若涉及全局优化问题，在离散参数（或策略）空间上，可以用模拟退火算法(SA)来求解。

SA 算法是基于 Monte Carlo 迭代求解策略的一种随机寻优算法。若将 SA 应用在算法二第(2)步的求解中，其一般步骤是：

(1) 给定初温 t_0，随机产生初始策略 v_0，令 $k = 0$；

(2) 由策略 v_k 以随机方式在策略空间中产生新策略 v_j；

(3) 若 $\min\{1, \exp[-(\hat{\eta}_{\alpha}^{v}(i) - \hat{\eta}_{\alpha}^{v_k}(i))/t_k]\} \geq random[0,1]$ 满足，则令 $v_k := v_j$。其中，$\hat{\eta}_{\alpha}^{v}(i)$ 表示性能值向量的第 i 分量；

(4) 若抽样稳定准则满足，转到(5)，否则，转到(2)；

(5) 若算法终止准则满足，则停止；否则，由 t_k 退温至 t_{k+1}，并令 $k := k+1$，转到(2)。

3.2 相关不确定参数最优鲁棒控制策略求解

首先定义性能指标 $\bar{\eta}_{\alpha}^{v}(\theta) = \omega_{\alpha}^{v}(\theta)\eta_{\alpha}^{v}(\theta)$。这里 $\omega_{\alpha}^{v}(\theta) = (\omega_{\alpha}^{v}(1,\theta), \mathrm{L}, \omega_{\alpha}^{v}(M,\theta))$，其中 $\omega_{\alpha}^{v}(i,\theta) \in [0,1]$ 且 $\omega_{\alpha}^{v}(\theta)e = 1$。即 $\bar{\eta}_{\alpha}^{v}(\theta)$ 是每个状态代价的加权平均。用 $\bar{\eta}_{\alpha}^{v}(\theta)$ 替代上述问题一、二中的 $\eta_{\alpha}^{v}(\theta)$，直接进行极小极大优化。对于问题一和问题二中可能出现的全局优化情况，在连续参数（或策略）空间上，我们考虑用填充函数法求解。

填充函数是在获得一个局部极小点后根据该点构造的一个辅助函数。若在问题二的求解中应用填充函数法，其一般步骤是：由任一初始策略 v_0 出发对 $\hat{\eta}_{\alpha}^{v}$ 局部极小化，得局部极小点 v_1^*。记 $S_1(v_1^*) = \{v \in \Omega_s \mid \hat{\eta}_{\alpha}^{v} \geq \hat{\eta}_{\alpha}^{v_1^*}\}$，$S_2(v_1^*) = \{v \in \Omega_s \mid \hat{\eta}_{\alpha}^{v} < \hat{\eta}_{\alpha}^{v_1^*}\}$。构造填充函数 $U(v)$ 具备下述性质：(a) v_1^* 是 $U(v)$ 的一个极大点；(b) $U(v)$ 在 $S_1(v_1^*)$ 中没有极小点或鞍点；(c) $U(v)$ 在 $S_2(v_1^*)$ 中肯定有极小点。以 v_1^* 的邻近点为初始点极小化 $U(v)$，得 $\bar{v} \in S_2(v_1^*)$。再从 $\bar{v}$ 出发极小化 $\hat{\eta}_{\alpha}^{v}$，得到 v_2^*。从 v_2^* 点出发重复上述过程，直至算法终止条件满足，最终求得全局最优点 v^*。

4 实例分析

考虑一个具有 M 个状态的系统，其过程在任一状态下的逗留时间服从 K 阶超指数分布。定义下标 $i, j \in \{1,2,\mathrm{L},M\}$，$l \in \{1,2,\mathrm{L},K\}$。策略 $v = [\mu_{il}]$，其中 $\mu_{il} \in [0.5, 20]$。参数 $\theta = [\theta_{il}]$，且 $\sum_{l=1}^{K} \theta_{il} = 1$。$F_{ij}^{v}(t, \theta_{il}) = 1 - \theta_i \exp(T_i t)e$ 为状态 i 下的逗留时间分布，其中 $T_i = diag(-\mu_{i1}, -\mu_{i2}, \mathrm{L}, -\mu_{iK})$，$\theta_i = (\theta_{i1}, \theta_{i2}, \mathrm{L}, \theta_{iK})$。嵌入链的转移矩阵 $P^{v}(\theta) = \left[p_{ij}^{v}(\theta) \right]$，其中

$$p_{ij}^{v}(\theta) = \begin{cases} \dfrac{\exp(-\sum_{l=1}^{K} \mu_{il} / j)}{M(1 + \exp(-\sum_{l=1}^{K} \mu_{il}))}, & j \neq i+1 \\ 1 - \sum_{j \neq i+1} p_{ij}, & j = i+1 \end{cases}。$$

性能函数 $f^{v} = [f(i, v(i))]$，其中

$$f(i, v(i)) = \ln[(1+i)\sum_{l=1}^{K} \mu_{il}] + \sqrt{i} / (2 \times \sum_{l=1}^{K} \mu_{il})。$$

最优鲁棒控制策略的两个子问题，在本例中，问题一是单极值优化，问题二是多极值优化。

a、参数不相关计算结果

取 $M=5, K=3$，参数 θ 中每一行 θ_i 相互独立，$\theta_{i1} \in [0.1, 0.3]$，$\theta_{i2} \in [0.2, 0.4]$，$\theta_{i3} \in [0.4, 0.6]$，用策略迭代法求解最优鲁棒控制策略。在算法一中用梯度投影法进行优化计算；在算法二中用 SA 法

进行优化计算，为此将策略 v 的取值空间按步长 0.1 离散化，即 $\mu_{ij} \in \{0.5, 0.6, \mathrm{L}\ , 20\}$。结果如图 1。

b、参数相关计算结果

取 $M=5, K=3$，要求参数 θ 中 $\theta_i = \theta_j, i \neq j$，$\theta_{i1} \in [0.1, 0.3]$，$\theta_{i2} \in [0.2, 0.4]$，$\theta_{i3} \in [0.4, 0.6]$。在问题一中用梯度投影法进行优化计算；在问题二中用填充函数法进行优化计算，构造填充函数

$$U(v) = B\exp(-\frac{\| v - v_n^* \|}{\sqrt{B}}) \cdot \arctan[B(\hat{\eta}_\alpha^v - \hat{\eta}_\alpha^{v_n^*} + h)]^{[8]},$$

其中 B 是一适当大的正数，h 是一小正数，v_n^* 是某个局部极小点。当取权向量分别为 $\omega_\alpha^v(\theta) = (0.5, 0.5)$ 和 $\omega_\alpha^v(\theta) = (1, 0)$ 时，结果如图 2。

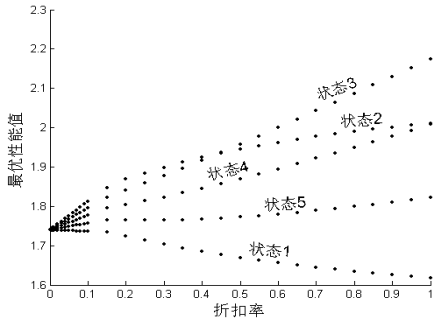

图 1 参数不相关计算结果

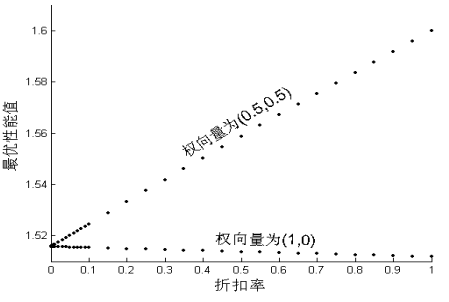

图 2 参数相关计算结果

上述在相关和不相关不确定参数条件下的计算结果表明，通过在最优鲁棒控制策略求解过程中合理地应用全局优化算法，通常可以得到较为理想的结果，能够很好地反映文献[1]中描述的平均准则和折扣准则之间关于折扣率的变化关系。

5 结 论

本文针对一类具有相关或不相关参数的不确定性 SMCP 鲁棒控制问题，分析了全局优化方法在其最优鲁棒控制策略求解过程中的重要作用，并通过实例进行了验证。由于半 Markov 模型可用来描述实际生活中的很大一类离散事件动态系统，本文关于应用全局优化方法求解最优鲁棒控制策略问题的分析对这类实际系统的设计和可靠运行具有一定的参考价值，有利于用来提高系统的管理水平和生产效率。

参考文献

1. Cao X R. Semi-Markov decision problems and performance sensitivity analysis. IEEE Trans. on Automatic Control, 2003, 48(5): 758-769.
2. Yin B Q, Li Y J, Zhou Y P, Xi H S. Semi-Markov decision problems with discounted-cost performance criteria. Automatica (submitted), 2003.
3. 殷保群，李衍杰，唐昊等。半 Markov 决策过程折扣模型与平均模型之间的关系。应用数学学报（已投稿），2003。
4. 唐昊，韩江洪，高隽。连续时间 Markov 控制过程的平均代价鲁棒控制策略。中国科学技术大学学报，2004, 34(2): 219-225。
5. Kalyanasundaram S, Chong E K P, Shroff N B. Markov decision processes with uncertain transition rates: sensitivity and robust control. Proceedings of the 41th IEEE Conference on Decision and Control, Las Vegas, Nevada USA, 2002, vol.4: 3799-3804.
6. Tang H, Liang X J, Gao J, and Liu C. Robust control policy for semi-Markov decision processes with dependent uncertain parameters. The 5th World Congress on Intelligent Control and Automation, Hangzhou, China, June, 2004.
7. 唐昊，袁继彬，陆阳，程文娟。SMDP 基于 α 一致化 Markov 链的神经元策略迭代优化，自动化学报，2004（已录用）。
8. 王凌。《智能优化算法及其应用》。清华大学出版社，2001 年 10 月。
9. 粟塔山。《最优化计算原理与算法程序设计》。国防科技大学出版社，2001 年 1 月。

马尔可夫决策过程基于 TD(0)学习和性能势的 NDP 优化 [1)]

袁继彬， 唐 昊，韩江洪
(合肥工业大学 计算机与信息学院， 合肥 230009)
E-mail: tangh@ustc.edu

摘要: 在 Markov 性能势基础上讨论了一种基于强化学习的马尔可夫决策过程（MDP）优化方法。本文通过 MDP 的一个一致化链和 Markov 性能势的神经元网络逼近，重点研究了 Critic 模式下的一种神经元动态规划（NDP）优化方法，给出了用于平均代价准则和折扣代价准则 MDP 优化的参数化 TD(0)学习规则和参数改进公式，并讨论了基于性能势的逼近策略迭代算法。文中最后给出一个数值仿真实例，实验结果表明平均准则下的 NDP 优化方法是折扣准则当折扣因子趋近于零的极限情况。

关键词：马尔可夫决策过程; 性能势; TD(0) **学习; 神经元动态规划**

The NDP Optimization of Markov Decision Processes Based on TD(0) Learning and Performance Potentials

Yuan Jibin, Tang Hao, Han Jianghong
(School of Computer and Information, Hefei University of Technology, Hefei 230009)

Abstract: We discuss the reinforcement learning-based optimization methods of Markov decision processes (MDPs) using the Markov performance potentials. By one uniformized chain of a MDP and the approximation representation of performance potentials with a neural network, we focus on the critic model of neuro-dynamic programming (NDP) methodology. We derive parameterized TD(0) learning rules and parameter-updating formula for both average criteria and discounted criteria problems, and discuss the potential-based approximate policy iteration algorithms. Finally, a numerical example is provided, and the results show that the average-cost criteria is the limitation of discounted-cost criteria as the discount factor goes to zero for our NDP optimization.

Key Words: Markov decision processes, performance potentials, TD(0) learning, neuro-dynamic programming

1 引 言

马尔可夫决策过程（MDP）可以用来描述一类序贯决策、控制问题，在系统模型参数已知时，其最优控制策略可以用数值迭代或策略迭代等数值算法来求解[1-2]。性能势理论的建立为马尔可夫决策过程的性能分析和优化开辟了一条新的途径[3-4]，定义在随机过程样本轨道上的性能势可以看作是 Poisson 方程的解，本质上同即时代价、相对代价以及 bias 的概念相同[1-3]，可以发展一些基于理论计算或仿真的优化算法[5-7]。当系统状态空间很大时，要利用计算机来精确地求解性能势或用查表法来表示性能势，无论是在计算时间上还是在存储空间上一般是不可行的。建立在强化学习（RL）基础上的神经元动态规划（NDP）方法为解决这类问题提供了一个有效的途径。强化学习是人工智能领域的一个重要研究方向，学习单元通过与环境接触取得控制经验并改善控制行为。$TD(\lambda)$ 学习是强化学习中一种常用的学习方法，有着以概率 1 收敛以及在线学习速度快的优点。$TD(0)$ 作为 $TD(\lambda)$ 学习算法的一种特例，文献[10]详细讨论了其性质。神经元动态规划由 MIT 的 Bertsekas D P 和 Tsitsiklis J N 于上个世纪九十年代正式提出，是解决大规模离散事件动态系统（DEDS）优化问题的一种有效方法。

文献 [8] 研究了 Critic 模式下，基于 Monte-Carlo 仿真和性能势逼近的半 Markov 决策过程（SMDP）优化问题；文献[9]讨论了 Actor

模式下，MDP 基于其一致化链的单样本轨道仿真和随机策略逼近的一种 NDP 优化方法；文献[7]提出了将 $TD(0)$ 学习用于估计性能势的思想。以此为基础，本文结合 NDP 方法和性能势理论，研究了平均代价准则和折扣代价准则下基于 $TD(0)$ 学习的 MDP 优化问题，同时给出了相应的优化算法和仿真实例。

2 问题描述

一个有限的 Markov 决策过程由五元组 $X=(X_t,\Phi,D,P^v(t),f^v)$ 定义。这里 $\{X_t;t\ge 0\}$ 是一个 Markov 过程；$\Phi=\{1,2,...M\}$ 为有限状态集；D 为紧致行动集，$v(i)$ 为在状态 i 时所采取的行动，$v(i)\in D$，$i\in\Phi$；所有确定性平稳策略集合定义为 $\Omega_s=\left\{v\middle|v=\left(v(1),v(2),\mathrm{L}\ ,v(M)\right),v(i)\in D\right\}$；$P^v(t)$ 为 X 的状态转移矩阵；定义策略 v 下的性能函数为 $f^v=\left[f(i,v(i))\right]$，$f(i,v(i))$ 表示过程在状态 i 采取行动 $v(i)$ 时的性能值。对固定策略 v，记 $X_n=X_{T_n},n=0,1,2,\mathrm{L}$，$T_0,T_1,\mathrm{L}\ ,T_n$ 为随机过程转移时刻，则 $\{X_n,n\ge 0\}$ 是 $\{X_t,t\ge 0\}$ 的嵌入 Markov 链，其转移概率矩阵记为 $P^v=[p_{ij}(v(i))]$，$p_{ij}(v(i))=P\{X_{n+1}=j\mid X_n=i,v(i)\in D\}$。$X$ 的平均性能准则定义为无穷时段平均代价的期望值

$$\eta^v=\lim_{T\to\infty}\frac{1}{T}E\{\int_0^T f(X_t,v(X(t)))dt\},v\in\Omega_s$$

X 的折扣性能准则定义为如下无穷时段总折扣代价的期望值

$$\eta_\alpha^v(i)=E[\int_0^\infty \alpha e^{-\alpha t}f(X_t,v(X_t))dt\mid X_0=i]$$

其中 $i\in\Phi,v\in\Omega_s$ 这里 $\alpha\ge 0$ 是 X 的折扣因子，记 $\eta_\alpha^v=[\eta_\alpha^v(1),\eta_\alpha^v(2),...,\eta_\alpha^v(M)]^\tau$，符号“ τ ”在文中表示转置。假设在任意 $v\in\Omega_s$ 下，X 都是不可约的，即遍历的。记 X 的无穷小转移速率矩阵为 A^v，则 $A^v=\mathrm{diag}(\lambda^v(1),\Lambda\ ,\lambda^v(M))\cdot(P^v-I)$，其中 I 为单位阵，$\lambda^v(i)>0$ 是 Markov 过程在状态 i 采用行动 $v(i)$ 时的平均转移率。令 $\lambda=\max_{i,v}\{\lambda^v(i)\}$，且 $\bar{P}=(1/\lambda)A^v+I$，可以定义 X 的一个一致化链 $\bar{X}=(X_n,\Phi,D,\bar{P},f^v)$[8]。$\bar{X}$ 的平均性能准则定义为

$$\bar{\eta}^v=\lim_{N\to\infty}\frac{1}{N}E(\sum_{l=0}^{N}f\ (X_l,v(X_l)))$$

折扣性能准则定义为[7]

$$\bar{\eta}_{\beta_\alpha}^v(i)=(1-\beta_\alpha)\lim_{N\to\infty}E[\sum_{n=0}^{N}\beta_\alpha^n f\ (X_n,v(X_n))\mid X_0=i]$$

这里，$\beta_\alpha=\lambda/(\lambda+\alpha)$ 为 $\bar{X}$ 的折扣因子，记

$$\bar{\eta}_{\beta_\alpha}^v=[\bar{\eta}_{\beta_\alpha}^v(1),\bar{\eta}_{\beta_\alpha}^v(2),...,\bar{\eta}_{\beta_\alpha}^v(M)]^\tau$$

则对于 MDP X 的优化问题可以转化为相应的一致化链 $\bar{X}$ 的优化问题来研究。优化目标是在 Ω_s 上寻找一个最优策略 v^* 满足

$$v^*\in\arg\min_{v\in\Omega_S}\bar{\eta}^v \text{ 或 } v^*\in\arg\min_{v\in\Omega_S}\bar{\eta}_{\beta_\alpha}^v。$$

3 基于 TD(0)学习的 NDP 优化

应用 NDP 方法解决 Markov 决策过程的优化问题主要是利用神经网络结构来逼近性能值或逼近策略，当逼近性能值时是把神经网络作为评判器（Critic 模型）；当逼近策略时是把神经网络作为行动器（Actor 模型）。NDP 方法的基本思想是选择参数数目比系统状态数少的网络逼近结构，以节省计算机存储空间，并利用样本轨道的仿真来估计有关数值，对神经网络的逼近质量做出改进，再通过特定的优化机制来更新当前策略。本文采用的是 Critic 模型，用于逼近性能势。

性能势理论提出和完善为我们解决 MDP 的优化问题提供了新的手段。在性能势理论中，一个策略是最优的当且仅当下式成立：

$$v_{k+1}\in\arg\min_{v\in\Omega_S}\{f^v+\beta_\alpha\bar{P}g_{\beta_\alpha}^{v_k}\}\qquad(1)$$

相关推导见文献[7]，其中 $0<\beta_\alpha\le 1$ 为折扣因子，

$g_{\beta_\alpha}^{v_k}=[g_{\beta_\alpha}^{v_k}(1),g_{\beta_\alpha}^{v_k}(2),...,g_{\beta_\alpha}^{v_k}(M)]^\tau$ 为策略 v_k 下的性能势。当系统状态空间很大时，由于计算时间和空间上的原因无法精确求解性能势，人们研究了一些性能势的估计算法，如基于单样本轨道的估计方法等[5]。文献[7]讨论了在平均代价准则下用 $TD(0)$ 学习算法来估计性能势，其即时代价定义为 $d_n=f(X_n)-\eta+g(X_{n+1})-g(X_n)$ （2）

其中，η 为平均代价估计，$g(X_n)$ 为状态 X_n 对应的性能势估计。这里，我们利用神经元网络来逼近性能势，并根据 $TD(0)$ 学习来构造神经元网络权系数的改进规则。由于神经元网络是利用其结构来保存信息，有着信息压缩等功能。网络输出 $g_{\beta_\alpha}^{v_k}(i,r)$ 是输入状态 i 和网络权值 r 的函数，近似表示在策略 v_k 下网络权值为 r 时状态 i 对应的性能势，记 $g_{\beta_\alpha}^{v_k}(r)=[g_{\beta_\alpha}^{v_k}(1,r,),g_{\beta_\alpha}^{v_k}(2,r),\mathrm{L}\ ,g_{\beta_\alpha}^{v_k}(M,r)]^\tau$，因此我们有如下逼近策略改进公式：

$$v_{k+1}\in\arg\min_{v\in\Omega_S}\{f^v+\beta_\alpha P^{\%}g_{\beta_\alpha}^{v_k}(r)\} \quad (3)$$

当 $0<\beta_\alpha<1$ 时，（3）式为折扣代价准则 MDP 的策略改进公式；当 $\beta_\alpha=1$ 时，（3）式为平均准则下的策略改进公式，记 $g^{v_k}(r)=g_{\beta_\alpha}^{v_k}(r)$。

在第 n 步学习过程中，在策略 v_k 下，平均代价估计值定义为 $\eta_n=(1-\delta_n)\eta_{n-1}+\delta_n f(X_n,v_k(X_n))$，$\delta_n$ 为第 n 步的学习步长。若取 $\delta_n=1/n$，则有

$$\eta_n=(1/N)\sum_{l=0}^{N}f(X_l,v_k(X_l))$$

η_n 的期望满足 $E(\eta_n)=\bar{\eta}^v$，所以这种取法是合理的。由（2）式及以上分析，我们可得

$$d_n=f(i_n,v_k(i_n))-\eta_n+g^{v_k}(i_{n+1},r_n)-g^{v_k}(i_n,r_n) \quad (4)$$

其中，$i_n\in\Phi$ 表示在时刻 n 系统所处的状态；$v_k(i_n)\in D$ 表示在策略 v_k 下在状态 i_n 时所采取的行动，$f(i_n,v_k(i_n))$ 表示在策略 v_k 下在状态 i_n 时的性能值。利用神经网络来逼近性能势时，不同的输入状态对应着相应的输出，其学习效果由网络权值保存。因此，我们可以定义网络权值修改规则如下：

$$r_{n+1}=r_n+\delta_n\bullet d_n\bullet\nabla_r g^{v_k}(r_n) \quad (5)$$

其中，$\nabla_r g^{v_k}(r)$ 是 $g^{v_k}(r)$ 对 r 的梯度。利用公式（3）、（4）、（5），我们可以对平均代价准则下的 Markov 决策过程进行优化。类似于文献[7]以及以上分析，我们可构造折扣代价准则下的 MDP 问题的即时代价学习公式和参数改进规则如下：

$$d_n=f(i_n,v_k(i_n))+\beta_\alpha g_{\beta_\alpha}^{v_k}(i_{n+1},r_n)-g_{\beta_\alpha}^{v_k}(i_n,r_n) \quad (6)$$

$$r_{n+1}=r_n+\delta_n\bullet d_n\bullet\nabla_r g_{\beta_\alpha}^{v_k}(r_n) \quad (7)$$

利用公式（3）、（6）、（7）我们可以对折扣代价准则下的 MDP 问题进行优化。

现在，我们给出学习优化算法。先设计一个神经网络来逼近性能势，神经网络的学习规则采用 $TD(0)$ 学习算法，学习成果保存在网络权值 r_n 中。然后再以性能势为基础对 Markov 决策过程进行优化。算法 1 用于逼近性能势，算法 2 用于策略寻优，具体如下：

算法 1：步骤 1、初始化 i_t，r_t，$t=0$。

步骤 2、根据 $P^{\%}$ 及 i_t 仿真产生下一状态 i_{t+1}。

步骤 3、若为平均代价准则，由公式（4）、（5）得到新权值，转步骤 4；若为折扣代价准则，由公式（6）、（7）得到新权值 r_{t+1}，转步骤 4。

步骤 4、判断某种指标是否满足，若不满足，则令 $i_t:=i_{t+1}$，$t:=t+1$，转步骤 2；否则，退出。

算法 2：步骤 1、$k=0$，随机产生一初始策略 v_k。

步骤 2、$v := v_k$，调用算法 1 计算性能势 $g_{\beta_\alpha}^{v_k}(r)$。

步骤 3、根据公式（3）求得 v_{k+1}。

步骤 4、判断停止条件是否满足。若不满足，令 $v_k := v_{k+1}$，$k := k+1$，转步骤 2；否则，退出。

算法 1 的结束条件可以选择为固定学习步数，神经网络的初始权值可以选择为随机数；根据问题的性质，算法 2 的结束条件可以按下式选择，其中 ε 为一个小的正数,

$$\| f^{v_{k+1}} + \beta_\alpha P^{v_{k+1}} g_{\beta_\alpha}^{v_k}(r) - f^{v_k} - \beta_\alpha P^{v_k} g_{\beta_\alpha}^{v_k}(r) \|_2 < \varepsilon$$

4 仿真实例

我们给出一个半马尔可夫过程的例子，通过定义一个 $\alpha-$ 一致化链可以容易地将其转化为 MDP 问题的优化[8]。所有实验结果在 MATLAB 平台下得到。

考虑一个 SMDP，状态空间 $\Phi = \{1, 2, ...31\}$，即 $M = 31$，行动集 D 为实数轴上的区间 $[0.5, 3.5]$，在策略 v 下过程的嵌入链 $P^v = [p_{ij}(v(i))]$ 定义为：

$$p_{ij}(v(i)) = \begin{cases} \dfrac{\exp(-v(i)/j)}{M(1+\exp(-v(i)))} & j \neq i+1 \\ 1 - \sum\limits_{j \neq i+1} p_{ij}(v(i)) & j = i+1 \end{cases}$$

过程在状态 i 的性能函数与下一状态无关，为 $f(i, v(i)) = \ln[(1+i)v(i)] + \sqrt{i}/(2v(i))$；过程在状态 i 的逗留时间服从三阶 Erlang 分布，且与下一状态无关，即

$$F_{ij}(t, v(i)) = F_i(t, v(i)) = 1 - e_1 \exp(T^{v(i)} t) e,$$

这里，$T^{v(i)} = 3v(i)\begin{bmatrix} -1 & 1 & 0 \\ 0 & -1 & 1 \\ 0 & 0 & -1 \end{bmatrix}$，

$e = [1,1,1]^\tau$，$e_1 = [1,0,0]$。

设计一个 5×3×1 *BP* 神经网络，权初始值、阈值随机生成，神经网络的转移函数为 Sigmoid 型函数。表一为平均代价准则下的一组实验数据及比较。在基于 $TD(0)$ 学习的 NDP 优化中，step 表示进行每次策略迭代前网络权值的学习步数；文献[8]中讨论的方法实质上是基于 Monte-Carlo 仿真的 NDP 优化，表中 step 表示每次用于性能势估计的样本轨道长度；理论性能值由一致化链的稳态分布计算得到。从表中可以看出，基于 $TD(0)$ 学习的 NDP 优化方法与基于 Monte-Carlo 仿真的 NDP 优化方法的优化结果接近。

表一 不同优化方法下的平均代价最优值及理论值

step	基于 TD(0)学习的 NDP 优化	理论性能值	Monte-Carlo 仿真的优化
1000	3.935394421	/	4.016900676
2000	3.935310321	/	3.977413391
5000	3.933643601	/	3.950680732
10000	3.927708269	3.717584845	3.927713473

在折扣准则下，取 $\varepsilon = 3e-5$，分别采用不同的学习步数和折扣因子对过程进行优化。图中各坐标点表示给定状态下停止策略对应的折扣代价 $\bar{\eta}_{\beta_\alpha}^{v}(i)$。图一为在确定的折扣因子（$\beta_\alpha$＝0.9999）下，采用不同学习步数进行优化的结果比较。图二为在确定的学习步数（step=10000）下，采用不同折扣因子进行优化的结果比较。由图一，折扣代价随学习步数的增加而逐步收敛。由图二，折扣代价随折扣因子趋近于 1 而趋于平均代价值。这说明平均代价下的 NDP 优化问题是折扣代价下的 NDP 优化的极限情况，同时也表明基于 $TD(0)$ 学习的 NDP 优化方法有效地对 MDP 进行了优化。

5 结 论

通过定义一个等价 $\alpha-$ 一致化链，算法可用于一类半 Markov 决策过程的最优控制策略求解[8]。在本算法中，采用了先学习、再规划的方法；若固定算法 1 的学习步数为 1，即可得到边学习、边规划的优化算法。另外，除了将 $TD(0)$ 学习、Critic 模型和性能势相结合用于 MDP 问题的优化外，亦可将 Critic-Actor 算法模型与性能势结合对 MDP 问题进行优化，作者正在做进一步的研究工作。