第 32 卷第 2 期
2015 年 2 月

控 制 理 论 与 应 用
Control Theory & Applications

Vol. 32 No. 2
Feb. 2015

DOI: 10.7641/CTA.2015.14005

具有未知动态的线性系统二人零和博弈问题在线学习方案

富　月†, 柴天佑

(东北大学 流程工业综合自动化国家重点实验室, 辽宁 沈阳 110819)

摘要: 针对具有未知动态线性系统的二人零和博弈问题, 本文提出了一种新的基于单环迭代方法的在线学习方案. 为保证单环迭代方法的收敛性, 给出了一种新的分析方法. 在系统内部矩阵A, 控制输入矩阵B以及干扰输入矩阵D均未知的情况下, 通过在线迭代策略, 同步得到了博弈代数黎卡提方程的近似解, 以及控制和干扰策略. 仿真结果表明了所提方法的有效性.

关键词: 二人零和博弈; 策略迭代; 博弈代数黎卡提方程

中图分类号: TP273　　　**文献标识码**: A

Online solution of two-player zero-sum games for linear systems with unknown dynamics

FU Yue†, CHAI Tian-you

(State Key Laboratory of Synthetical Automation for Process Industries, Northeastern University, Shenyang Liaoning 110819, China)

Abstract: For two-player zero-sum games of continuous-time linear systems with unknown dynamics, we present an online adaptive learning algorithm based on the policy iteration (PI) scheme with only one loop. A new analytical method to prove the convergence of the PI scheme is presented. An approximate solution to the generalized game algebraic Riccati equation without using a priori knowledge of the system matrices is developed. Simulation results illustrate the effectiveness of the proposed method.

Key words: two-player zero-sum game; policy iterations; game algebraic Riccati equation

1　引言(Introduction)

二人零和博弈问题是指参与博弈的双方的利益之和为零或一个常数, 即一方有所得, 另一方必有所失. 实际上, 二人零和博弈问题可以看作最小最大优化问题, 其中控制输入为使某一指定指标最小的参与者, 干扰输入为使某一指标最大的参与者.

对于连续时间非线性系统, 二人零和博弈问题的求解依赖于解Hamilton-Jacobi-Isaacs(HJI)方程. 通常, 即使系统动态完全已知, HJI方程也是很难求解的. 将文献[1]中的Kleinman's算法推广到连续时间非线性二人零和博弈问题, 可通过离线迭代求解HJI方程. 例如, 文献[2]提出了一种在内环对控制进行迭代, 在外环对干扰进行迭代的双环迭代求解方法, 文献[3–5]提出了另外一种双环迭代方法, 即在内环对干扰进行迭代, 在外环对控制迭代. 尽管这些迭代方法能够得到HJI方程的近似解, 但很难在每一迭代过程获得值函数的解. 为克服这一局限, 通过采用沿着系统轨迹的数据, 文献[6]针对非线性连续时间二人零和博弈问题, 提出了一种基于迭代和神经网络的在线求解方法. 上述方法都要求系统的动态完全已知.

对于连续时间线性系统, 二人零和博弈问题的求解依赖于解博弈代数黎卡提方程(GARE). 文献[7]针对连续时间线性系统, 基于在内环对控制进行迭代的双环迭代方法, 在系统内部矩阵 A 未知的情况下, 提出了一种在线求解 GARE 的方案. 文献[8–9]提出了一种单环迭代方法, 并基于该方法, 对上述方案进行了改善. 文献 [8–9] 采用“Frechet derivative”, “Gateaux derivative”, “Kantorovtich's Theorem”等理论对单环迭代方法进行了收敛性分析. 鉴于这些概念比较复杂, 本文提出了一种新的相对直观、简单、容易理解的收敛性分析方法. 在系统内部矩阵A, 控制输入矩阵B和干扰输入矩阵D均未知的情况下, 基于单环

收稿日期: 2014−01−09; 录用日期: 2014−10−18.
†通信作者. E-mail: fuyue@mail.neu.edu.cn.
国家自然科学基金项目(61374042), 中央高校基本科研业务费基金项目(N130408003, N130108001)资助.
Supported by Natural Science Foundation of China (61374042) and Fundamental Research Funds for the Central Universities (N130408003, N130108001).

迭代策略, 针对连续时间线性系统二人零和博弈问题, 提出了一种在线求解方案, 同步得到了GARE的近似解, 以及控制和干扰策略.

2 问题描述和预备知识(Problem descriptions and preliminaries)

2.1 问题描述(Problem formulation)

考虑如下连续时间线性系统:

$$\dot{x} = Ax + Bu + Dd, \tag{1}$$

其中: $x \in \mathbb{R}^n$是系统的状态, 对于反馈控制器的设计是完全可测的; $u \in \mathbb{R}^m$是控制输入; $d \in \mathbb{R}^q$为干扰输入; $A \in \mathbb{R}^{n\times n}$, $B \in \mathbb{R}^{n\times m}$和$D \in \mathbb{R}^{n\times q}$分别为未知的系统内部矩阵、控制输入矩阵和干扰输入矩阵. 此外, 假设系统是可镇定的. 定义如下性能指标:

$$J(x(0),u,d)=\int_0^\infty(x^{\mathrm T}Qx+u^{\mathrm T}Ru-\gamma^2\|d\|^2)\mathrm d\tau \equiv \int_0^\infty r(x,u,d)\mathrm d\tau, \tag{2}$$

其中: $Q = Q^{\mathrm T} \geqslant 0, R = R^{\mathrm T} > 0, \gamma > \gamma^* > 0$. 对于给定的反馈控制输入$u(x)$和干扰输入$d(x)$, 定义如下值函数:

$$V(x(t))=\int_t^\infty(x^{\mathrm T}Qx+u^{\mathrm T}Ru-\gamma^2\|d\|^2)\mathrm d\tau \equiv \int_t^\infty r(x,u,d)\mathrm d\tau \tag{3}$$

可以证明, 值函数(3)可以进一步表示为如下关于状态的二次型形式:

$$V(x) = x^{\mathrm T}Sx, \tag{4}$$

其中$S \geqslant 0$为非负定矩阵. 若值函数是有限的, 式(3)与下面的类李雅普诺夫贝尔曼方程微分等价:

$$0 = r(x,u,d) + (\nabla V)^{\mathrm T}\dot{x},\ V(0) = 0, \tag{5}$$

其中$\nabla V = \dfrac{\partial V}{\partial x} \in \mathbb{R}^n$为值函数的梯度. 此时, 哈密顿方程为

$$H(x,\nabla V,u,d)=r(x,u,d)+(\nabla V)^{\mathrm T}\dot{x},\ V(0)=0. \tag{6}$$

选择状态反馈控制输入$u = -Kx$, 干扰输入$d = Lx$, 并将它们代入贝尔曼方程(5)中, 得到

$$S(A - BK + DL) + (A - BK + DL)^{\mathrm T}S+ Q + K^{\mathrm T}RK - \gamma^2L^{\mathrm T}L = 0. \tag{7}$$

式(7)为针对反馈增益K和L的关于S的李雅普诺夫方程.

下面, 针对动态约束(1), 定义如下二人零和博弈:

$$V^*(x(0)) = \min_u \max_d J(x(0),u,d) = \min_u \max_d \int_0^\infty (x^{\mathrm T}Qx + u^{\mathrm T}Ru - \gamma^2\|d\|^2)\mathrm d\tau. \tag{8}$$

实际上, 二人零和博弈问题就是选择使性能指标最小的参与者$u = -Kx$和最大的参与者$d = Lx$. 文献[10]指出, 如果满足下面的纳什条件, 则二人零和博弈问题有唯一的解:

$$\min_u \max_d J(x(0),u,d)=\max_d \min_u J(x(0),u,d). \tag{9}$$

纳什条件的一个必要条件是Isaacs条件, 即对所有的参与者或策略u, d, 下式成立:

$$\min_u \max_d H(x,\nabla V,u,d) = \max_d \min_u H(x,\nabla V,u,d), \tag{10}$$

因此, 在平衡点附近, 有

$$\frac{\partial H}{\partial u} = 0,\ \frac{\partial H}{\partial d} = 0. \tag{11}$$

由式(6)和式(11), 得到如下控制和干扰策略:

$$u(x) = -R^{-1}B^{\mathrm T}Sx = -Kx, \tag{12}$$

$$d(x) = \frac{1}{\gamma^2}D^{\mathrm T}Sx = Lx. \tag{13}$$

将式(12)和式(13)代入式(7), 得到GARE:

$$A^{\mathrm T}S + SA + Q - SBR^{-1}B^{\mathrm T}S + \frac{1}{\gamma^2}SDD^{\mathrm T}S = 0. \tag{14}$$

为解决二人零和博弈问题, 本文需要解GARE, 从而得到S的最优值. 如果(A, B)是可镇定的, $(A, \sqrt{Q})$是可观测的, 并且$\gamma > \gamma^* > 0$, 则一定存在非负定矩阵$S \geqslant 0$满足式(14).

2.2 单环迭代算法(PI algorithm with only one loop)

由于关于S, 式(14)是非线性的, 因此直接根据式(14)求解S通常比较困难, 尤其是针对维数较大的矩阵. 为了得到S的近似解, 文献[2–5]提出了双环迭代算法, 文献[8]提出了单环迭代算法. 本文采用如下单环迭代算法:

算法 1　1) 初始化: 选择初始反馈增益矩阵$K_0 \in \mathbb{R}^{m\times n}$, $L_0 \in \mathbb{R}^{p\times n}$使得$A - BK_0 + DL_0$稳定.

2) 求解S_k: 解下面的李雅普诺夫方程:

$$S_k(A-BK_k+DL_k)+(A-BK_k+DL_k)^{\mathrm T}S_k+ Q + K_k^{\mathrm T}RK_k - \gamma^2L_k^{\mathrm T}L_k = 0, \tag{15}$$

其中$K_k, L_k(k = 1, 2, \cdots)$根据下式递推得到:

$$L_k = \frac{1}{\gamma^2}D^{\mathrm T}S_{k-1}, \tag{16}$$

$$K_k = R^{-1}B^{\mathrm T}S_{k-1}. \tag{17}$$

3) 令$k = k + 1$. 如果$\|S_k - S_{k-1}\| \leqslant \varepsilon(\varepsilon > 0$为事先指定的小的正实数), 停止迭代, 并输出$S_k$; 否则返回到2), 继续迭代.

定理 1　如果选择的正定矩阵R和常数γ满足γ^2

$\geqslant \|R\|\cdot\|D\|^2/\|B\|^2$, 并且$A-BK_i+DL_i$稳定, S_i根据式(15)求解, K_{i+1}和L_{i+1}根据式(16)–(17)计算, 则$A-BK_{i+1}+DL_{i+1}$也稳定, 并且对任意的$i=0, 1,\cdots$,

$$0\leqslant S_{i+1}\leqslant S_i. \tag{18}$$

此外, 如果式(14)的解S^*存在, 则$\{S_{i+1}\}_{i=0}^{\infty}$, $\{K_i\}_{i=0}^{\infty}$和$\{L_i\}_{i=0}^{\infty}$收敛到S^*, K^*和L^*, 其中:

$$K^*=R^{-1}B^{\mathrm{T}}S^*,\ L^*=\gamma^{-2}D^{\mathrm{T}}S^*.$$

证 定理1的证明在附录A中.

注 1 实际上, 文献[8–9]已经对上述单环迭代算法的收敛性进行了分析. 然而, 由于文献[8–9]采用的"Frechet derivative", "Gateaux derivative", "Kantorovtich's Theorem"等理论比较复杂, 本文提出了一种新的相对直观、简单、容易理解的收敛性分析方法.

3 具有未知动态的二人零和博弈在线学习方案(Online solution of linear two-player zero-sum games without using system dynamics)

针对线性二人零和博弈问题, 本文给出一种不依赖于系统内部矩阵A, 控制输入矩阵B和干扰输入矩阵D的一种新的在线学习方案. 首先, 本文假设初始反馈增益矩阵K_0, L_0已知; 其次, 对每一个$k\in\mathbb{Z}_+$, 根据式(15)求解对称非负定矩阵S_k; 最后, 根据式(16)和式(17)得到反馈增益矩阵$K_{k+1}\in\mathbb{R}^{m\times n}$和$L_{k+1}\in\mathbb{R}^{p\times n}$.

为此, 本文将系统(1)重新描述为

$$\dot{x}=A_kx+B(K_kx+u)-D(L_kx-d), \tag{19}$$

其中$A_k=A-BK_k+DL_k$.

于是, 根据式(15)–(17)和式(19), 有

$$\begin{aligned}&x^{\mathrm{T}}(t+\Delta t)S_kx(t+\Delta t)-x^{\mathrm{T}}(t)S_kx(t)=\\&-\int_t^{t+\Delta t}x^{\mathrm{T}}Q_kx\mathrm{d}\tau+\\&2\int_t^{t+\Delta t}(K_kx+u)^{\mathrm{T}}RK_{k+1}\mathrm{d}\tau-\\&2\int_t^{t+\Delta t}(L_kx-d)^{\mathrm{T}}L_{k+1}x]\mathrm{d}\tau,\end{aligned} \tag{20}$$

其中$Q_k=Q+K_k^{\mathrm{T}}RK_k-\gamma^2L_k^{\mathrm{T}}L_k$.

下面, 本文将给出, 在某些条件下, 无需已知A, B和D, 当给定反馈增益矩阵K_k和L_k时, 可以唯一确定满足式(15)–(17)的矩阵(S_k,K_{k+1},L_{k+1}), 其中$S_k=S_k^{\mathrm{T}}\geqslant 0$. 为此, 定义下面的两个运算:

$$\begin{aligned}&S\in\mathbb{R}^{n\times n}\to\hat{S}\in\mathbb{R}^{\frac12 n\times(n+1)},\\&x\in\mathbb{R}^n\to\bar{x}\in\mathbb{R}^{\frac12 n\times(n+1)},\end{aligned}$$

其中:

$$\hat{S}=[s_{11}\ 2s_{12}\ \cdots\ 2s_{1n}\ s_{22}\ 2s_{23}\ \cdots\ 2s_{n-1\ n}\ s_{nn}]^{\mathrm{T}},$$
$$\bar{x}=[x_1^2\ x_1x_2\ \cdots\ x_1x_n\ x_2^2\ x_2x_3\ \cdots\ x_{n-1}x_n\ x_n^2]^{\mathrm{T}}.$$

此外, 根据Kronecker积的表示形式, 有

$$\begin{aligned}&x^{\mathrm{T}}Q_kx=(x^{\mathrm{T}}\otimes x^{\mathrm{T}})\mathrm{vec}(Q_k),\\&(u+K_kx)^{\mathrm{T}}RK_{k+1}x=\\&[(x^{\mathrm{T}}\otimes x^{\mathrm{T}})(I_n\otimes K_k^{\mathrm{T}}R)+\\&(x^{\mathrm{T}}\otimes u^{\mathrm{T}})(I_n\otimes R)]\mathrm{vec}(K_{k+1}),\\&(L_kx-d)^{\mathrm{T}}L_{k+1}x=\\&[(x^{\mathrm{T}}\otimes x^{\mathrm{T}})(I_n\otimes L_k^{\mathrm{T}})-(x^{\mathrm{T}}\otimes d^{\mathrm{T}})]\mathrm{vec}(L_{k+1}).\end{aligned}$$

进一步的, 对正整数l, 定义矩阵

$$\begin{aligned}&\delta_{\mathrm{xx}}\in\mathbb{R}^{l\times\frac12 n(n+1)},\ I_{\mathrm{xx}}\in\mathbb{R}^{l\times n^2},\\&I_{\mathrm{xu}}\in\mathbb{R}^{l\times nm},\ I_{\mathrm{xd}}\in\mathbb{R}^{l\times np},\end{aligned}$$

使得

$$\begin{aligned}\delta_{\mathrm{xx}}=&[\bar{x}(t_1)-\bar{x}(t_0),\bar{x}(t_2)-\bar{x}(t_1),\cdots,\\&\bar{x}(t_l)-\bar{x}(t_{l-1})]^{\mathrm{T}},\\I_{\mathrm{xx}}=&[\int_{t_0}^{t_1}x\otimes x\mathrm{d}\tau,\int_{t_1}^{t_2}x\otimes x\mathrm{d}\tau,\cdots,\\&\int_{t_{l-1}}^{t_l}x\otimes x\mathrm{d}\tau]^{\mathrm{T}},\\I_{\mathrm{xu}}=&[\int_{t_0}^{t_1}x\otimes u\mathrm{d}\tau,\int_{t_1}^{t_2}x\otimes u\mathrm{d}\tau,\cdots,\\&\int_{t_{l-1}}^{t_l}x\otimes u\mathrm{d}\tau]^{\mathrm{T}},\\I_{\mathrm{xd}}=&[\int_{t_0}^{t_1}x\otimes d\,\mathrm{d}\tau,\int_{t_1}^{t_2}x\otimes d\,\mathrm{d}\tau,\cdots,\\&\int_{t_{l-1}}^{t_l}x\otimes d\,\mathrm{d}\tau]^{\mathrm{T}},\end{aligned}$$

其中$0\leqslant t_0<t_1<\cdots<t_l$.

于是, 对任意给定的反馈增益矩阵K_k, L_k, 式(20)可重新表示成下面的线性矩阵等式形式:

$$X_k\begin{bmatrix}\hat{S}_{k+1}\\\mathrm{vec}(K_{k+1})\\\mathrm{vec}(L_{k+1})\end{bmatrix}=Y_k, \tag{21}$$

其中$X_k\in\mathbb{R}^{l\times[\frac12 n(n+1)+mn+pn]}$和$Y_k\in\mathbb{R}^l$的定义如下:

$$\begin{aligned}X_k=&[\delta_{\mathrm{xx}}\ \ -2I_{\mathrm{xx}}(I_n\otimes K_k^{\mathrm{T}}R)-2I_{\mathrm{xu}}(I_n\otimes R)\\&2\gamma^2I_{\mathrm{xx}}(I_n\otimes L_k^{\mathrm{T}})-2\gamma^2I_{\mathrm{xd}}],\\Y_k=&-I_{\mathrm{xx}}\mathrm{vec}(Q_k).\end{aligned}$$

由于式(21)的解可能不存在, 在这种情况下, 可通过下面的式(22)求解. 式(22)可以看作是式(21)的最小二乘解. 实际上, 如果X_k是列满秩的, 则式(21)可以直接通过下式求解:

$$\begin{bmatrix}\hat{S}_{k+1}\\\mathrm{vec}(K_{k+1})\\\mathrm{vec}(L_{k+1})\end{bmatrix}=(X_k^{\mathrm{T}}X_k)^{-1}X_k^{\mathrm{T}}Y_k. \tag{22}$$

下面, 本文给出二人零和博弈问题在线学习方案:

算法 2 1) 在$[t_0,t_l]$时间区域内, 分别将$u=-K_0x+e_{\mathrm{u}}$和$d=L_0x+e_{\mathrm{d}}$作为控制和干扰输入, 其

中K_0, L_0为初始反馈增益矩阵, 满足$A-BK_0+DL_0$稳定, $e_{\rm u}$, $e_{\rm d}$为噪声. 计算$\delta_{\rm xx}, I_{\rm xx}, I_{\rm xu}, I_{\rm xd}$直到下面式(23)成立. 令$k=0$;

2) 根据式(22)求解S_k, K_{k+1}, L_{k+1};

3) 令$k\rightarrow k+1$, 重复步骤2)直到$\|S_k-S_{k-1}\|\leqslant\varepsilon$, 其中: $k\geqslant 1, \varepsilon>0$为事先指定的小正实数;

4) 将$u=-K_kx$, $d=L_kx$作为二人零和博弈问题近似解.

引理 1　如果存在正整数$l_0>0$, 使得对所有的$l\geqslant l_0$, 下面的等式成立:

$$\operatorname{rank}(I_{\rm xx}, I_{\rm xu}, I_{\rm xd})=\frac{n(n+1)}{2}+mn+pn, \quad (23)$$

那么对所有的$k\in\mathbb{Z}_+$, X_k是列满秩的.

定理 2　根据算法2得到的序列$\{S_k\}_{k=0}^{\infty}$, $\{K_i\}_{i=1}^{\infty}$, $\{L_j\}_{j=1}^{\infty}$分别收敛到它们的最优解S^*, K^*, L^*.

证　定理2的证明与文献[11]中定理7的证明类似, 这里不再赘述.

4　仿真(Simulations)

为了验证本文提出的方法的有效性, 考虑如下二阶连续时间线性系统:

$$\dot{x}=\begin{bmatrix}0 & 1\\ -1 & -3\end{bmatrix}x+\begin{bmatrix}0\\ 0.6\end{bmatrix}u+\begin{bmatrix}1\\ 4\end{bmatrix}d, \quad (24)$$

其中$x\in\mathbb{R}^2$, $u\in\mathbb{R}$和$d\in\mathbb{R}$分别为系统的状态、输入和干扰.

在求取近似最优解的设计过程中, 并没有用到系统的内部矩阵、控制输入矩阵和干扰输入矩阵. 首先选择如下的加权项:

$$Q=\begin{bmatrix}1 & 0\\ 0 & 1\end{bmatrix}, \ R=1, \ \gamma=11.$$

在原点附近随机选择状态变量的初始值. 选择$\varepsilon=0.01$, 并选择初始反馈增益矩阵为$K_0=[1, 0]$, $L_0=[0, 1]$, 易知$A-BK_0+DL_0$是稳定的. 将下面的噪声分别施加到控制和干扰的输入端:

$$e_{\rm u}=0.5\sin t, \ e_{\rm d}=\sin(10t). \quad (25)$$

仿真过程中, 收集状态和输入数据的时间间隔为1 s. 从$t=0$ s开始迭代, 经过8次迭代后, 有$\|S_k-S_{k-1}\|\leqslant 0.01$, 迭代结束, 得到如下性能和反馈增益矩阵:

$$S_8=\begin{bmatrix}1.8886 & 0.5166\\ 0.5166 & 0.3369\end{bmatrix},$$

$$K_8=[0.3100, 0.2022]^{\rm T}, \ L_8=[0.0327 \ \ 0.0154]^{\rm T}.$$

另一方面, 通过直接求解GARE, 即

$$A^{\rm T}S+SA+Q-SBR^{-1}B^{\rm T}S+\frac{1}{\gamma^2}SDD^{\rm T}S=0.$$

并利用式(12)–(13), 可以得到如下最优解:

$$S^*=\begin{bmatrix}1.8883 & 0.5166\\ 0.5166 & 0.3368\end{bmatrix},$$

$$K^*=[0.3100 \ \ 0.2021]^{\rm T}, \ L^*=[0.0327 \ \ 0.0154]^{\rm T}.$$

图1表明了序列S_k, K_k和L_k的收敛过程, 图2为S_k的组成部分的变化过程.

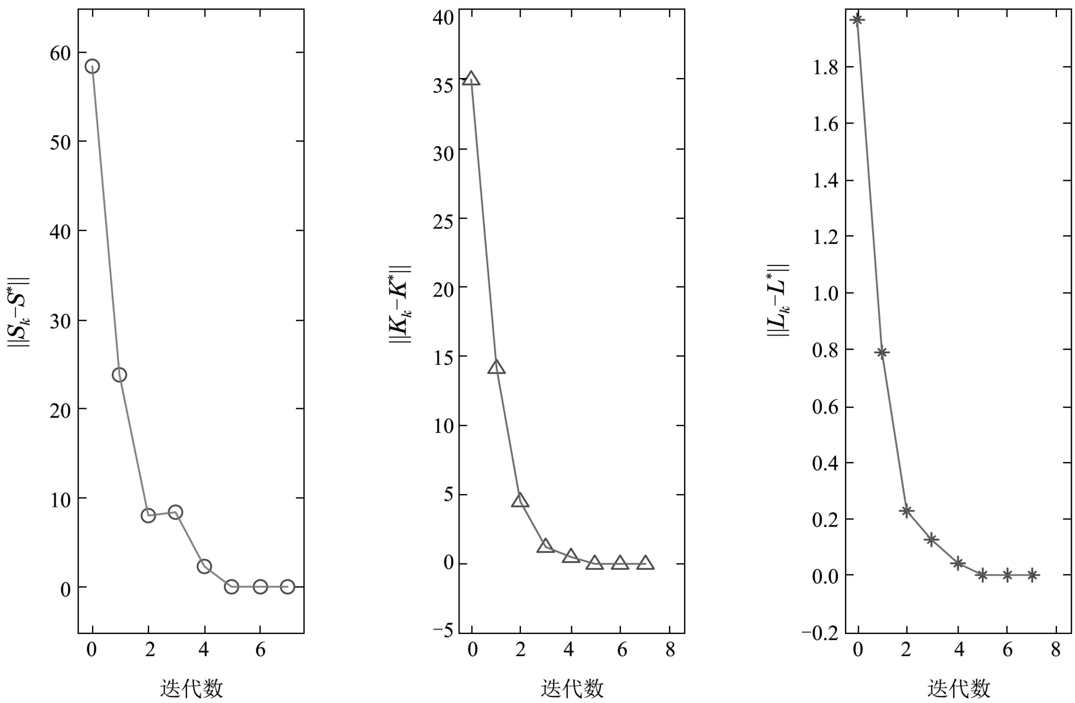

图 1　S_k, K_k和L_k收敛到S^*, K^*和L^*的变化过程

Fig. 1　Convergence of S_k, K_k, and L_k to their optimal values S^*, K^* and L^* during the learning process

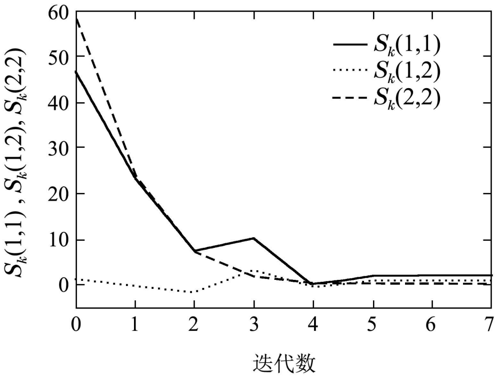

图 2 S_k组成元素的变化过程

Fig. 2 Evolution of the components of S_k

5 结论(Conclusions)

本文针对具有未知动态的连续时间线性二人零和博弈问题, 提出了一种新的在线求解方案. 首先提出了一种新的单环迭代策略的收敛性分析方法; 其次基于该单环迭代策略, 在系统内部矩阵、控制输入矩阵和干扰输入矩阵均未知的情况下, 通过利用系统的状态和输入数据, 采用最小二乘方法同步得到了博弈代数黎卡提方程的近似解, 以及控制和干扰策略; 最后通过仿真实验验证了所提方法的有效性.

参考文献(References):

[1] KLEINMAN D. On an iterative technique for Riccati equation computations [J]. *IEEE Transactions on Automatic Control* , 1968, 13(1): 114 – 115.

[2] FENG Y, ANDERSON B D, ROTKOWITZ M. A game theoretic algorithm to compute local stabilizing solutions to HJBI equations in nonlinear H_∞ control [J]. *Automatica* 2009, 45(4): 881 – 888.

[3] VAN DER SCHAFT AJ. L2-gain analysis of nonlinear systems and nonlinear state feedback H_∞ control [J]. *IEEE Transactions on Automatic Control*, 1992, 37(6): 770 – 784.

[4] ABU-KHALAF M, LEWIS F L. Neurodynamic programming and zero-sum games for constrained control systems [J]. *IEEE Transactions on Neural Networks*, 2008, 19(7): 1243 – 1252.

[5] ABU-KHALAF M, LEWIS F L, HUANG J. Policy iterations on the Hamilton–Jacobi–Isaacs equation for H_∞ state feedback control with input saturation [J]. *IEEE Transactions on Automatic Control*, 2006, 51(12): 1989 – 1995.

[6] VAMVOUDAKIS, K G, LEWIS F L. Online solution of nonlinear two-player zero-sum games using synchronous policy iteration [J]. *International Journal of Robust and Nonlinear Control*, 2012, 22(13): 1460 – 1483.

[7] VRABIE D, LEWIS F. Adaptive dynamic programming for online solution of a zero-sum differential game [J]. *Journal of Control Theory and Applications*, 2011, 9(3): 353 – 360.

[8] WU H N, LUO B. Simultaneous policy update algorithms for learning the solution of linear continuous-time H∞ state feedback control [J]. *Information Sciences*, 2013, 222: 472 – 485.

[9] LI H L, LIU D R, WANG D. Integral reinforcement learning for linear continuous-time zero-sum games with completely unknown dynamics [J]. *IEEE Transactions on Automation Science and Engineering*, 2014, 11(3): 706 – 714.

[10] LEWI F L, VRABIE D L, SYRMOS V L. *Optimal Control* [M]. New Jersey: Wiley, 2012.

[11] JIANG Y, JIANG Z P. Computational adaptive optimal control for continuous-time linear systems with completely unknown dynamics [J]. *Automatica*, 2012, 48(10): 2699 – 2704.

[12] VRABIE D, PASRAVANU O, ABU-KHALAF M. Adaptive optimal control for continuous-time linear systems based on policy iteration [J]. *Automatica*, 2009, 45(2): 477 – 484.

附 录 定 理 1 的 证 明(Appendix Proof of Theorem 1)

为证明定理1, 本文首先给出性能矩阵的定义.

定义 1 假设$u = -Kx$和$d = Lx$为任意的反馈策略. 如果将u和d施加到系统(1)上, 相应的性能为

$$V(x(t)) = x^{\mathrm T}Sx, \tag{A1}$$

其中S为与反馈增益矩阵K和L相对应的性能矩阵, 定义如下:

$$S = \int_0^\infty \mathrm{e}^{A-BK+DL}(Q+K^{\mathrm T}RK- \gamma^2L^{\mathrm T}L)\mathrm{e}^{A-BK+DL}\mathrm{d}\tau. \tag{A2}$$

S是有限的、非负定的当且仅当闭环系统特征矩阵$A - BK + DL$具有负实部, 并且$Q + K^{\mathrm T}RK - \gamma^2L^{\mathrm T}L \geqslant 0$.

证 令S_0为与K_0, L_0, S_0相对应的性能矩阵, 因此当$k = 0$时, 满足式(15). 令$K_1 = R^{-1}B^{\mathrm T}S_0$, $L_1 = \gamma^{-2}D^{\mathrm T}S_0$, 并令$S_1$为相应的性能矩阵.

记$A_0 = A - BK_0 + DL_0$, $A_1 = A - BK_1 + DL_1$, 则由式(15)可知

$$A_0^{\mathrm T}S_0 + S_0A_0 + Q + K_0^{\mathrm T}RK_0 - \gamma^2L_0^{\mathrm T}L_0 = 0. \tag{A3}$$

令$A_0 = A_1 - B(K_0 - K_1) + D(L_0 - L_1)$, 则

$$A_1^{\mathrm T}S_0 - (K_0 - K_1)^{\mathrm T}B^{\mathrm T}S_0 + (L_0 - L_1)^{\mathrm T}D^{\mathrm T}S_0+ S_0A_1 - S_0B(K_0 - K_1) + S_0D(L_0 - L_1)+ Q + K_0^{\mathrm T}RK_0 - \gamma^2L_0^{\mathrm T}L_0 = 0. \tag{A4}$$

另一方面, S_1满足

$$A_1^{\mathrm T}S_1 + S_1A_1 + Q + K_1^{\mathrm T}RK_1 - \gamma^2L_1^{\mathrm T}L_1 = 0. \tag{A5}$$

因此, 由式(A5)和式(A4)可知

$$A_1^{\mathrm T}(S_0 - S_1) + (S_0 - S_1)A_1 + K_0^{\mathrm T}RK_0 - K_1^{\mathrm T}RK_1- \gamma^2(L_0^{\mathrm T}L_0 - L_1^{\mathrm T}L_1) - (K_0 - K_1)^{\mathrm T}B^{\mathrm T}S_0+ (L_0 - L_1)^{\mathrm T}D^{\mathrm T}S_0 - S_0B(K_0 - K_1)+ S_0D(L_0 - L_1) = 0. \tag{A6}$$

从式(A6)的右边加和减下面的项

$$(K_0 - K_1)^{\mathrm T}RK_1 + K_1^{\mathrm T}R(K_0 - K_1)- \gamma^2(L_0 - L_1)^{\mathrm T}L_1 - \gamma^2L_1^{\mathrm T}(L_0 - L_1), \tag{A7}$$

得到

$$A_1^{\mathrm T}(S_0 - S_1) + (S_0 - S_1)A_1+ (K_0 - K_1)^{\mathrm T}R(K_0 - K_1) - \gamma^2(L_0 - L_1)^{\mathrm T}(L_0 - L_1)+ (K_0-K_1)^{\mathrm T}(RK_1-B^{\mathrm T}S_0)+(K_1^{\mathrm T}R-S_0B)(K_0-K_1)-$$

$$(L_0-L_1)^{\mathrm{T}}(\gamma^2L_1-D^{\mathrm{T}}S_0)-(\gamma^2L_1^{\mathrm{T}}-S_0D)(L_0-L_1)=0. \tag{A8}$$

因此

$$\begin{aligned}S_0-S_1=\int_0^{\infty}\mathrm{e}^{A_1}[&(K_0-K_1)^{\mathrm{T}}R(K_0-K_1)-\\&\gamma^2(L_0-L_1)^{\mathrm{T}}(L_0-L_1)+\\&(K_0-K_1)^{\mathrm{T}}(RK_1-B^{\mathrm{T}}S_0)+\\&(RK_1^{\mathrm{T}}-S_0B)(K_0-K_1)-\\&(L_0-L_1)^{\mathrm{T}}(\gamma^2L_1-D^{\mathrm{T}}S_0)-\\&(\gamma^2L_1^{\mathrm{T}}-S_0D)(L_0-L_1)]\mathrm{e}^{A_1}\mathrm{d}t.\end{aligned} \tag{A9}$$

相应地, 本文也可以得到

$$\begin{aligned}S_0-S_1=\int_0^{\infty}\mathrm{e}^{A_0}[&(K_0-K_1)^{\mathrm{T}}R(K_0-K_1)-\\&\gamma^2(L_0-L_1)^{\mathrm{T}}(L_0-L_1)+\\&(K_0-K_1)^{\mathrm{T}}(RK_1-B^{\mathrm{T}}S_1)+\\&(K_1^{\mathrm{T}}R-S_1B)(K_0-K_1)-\\&(L_0-L_1)^{\mathrm{T}}(\gamma^2L_1-D^{\mathrm{T}}S_1)-\\&(\gamma^2L_1^{\mathrm{T}}-S_1D)(L_0-L_1)]\mathrm{e}^{A_0}\mathrm{d}t.\end{aligned} \tag{A10}$$

由式(A7), 可知

$$\begin{aligned}S_0-S_1=\int_0^{\infty}\mathrm{e}^{A_1}[&(K_0-K_1)^{\mathrm{T}}R(K_0-K_1)-\\&\gamma^2(L_0-L_1)^{\mathrm{T}}(L_0-L_1)]\mathrm{e}^{A_1}\mathrm{d}t.\end{aligned} \tag{A11}$$

考虑到定理1的条件, 易知

$$\gamma^2\|B\|^2\geqslant\|R\|\cdot\|D\|^2. \tag{A12}$$

因此

$$(K_0-K_1)^{\mathrm{T}}R(K_0-K_1)\geqslant\gamma^2(L_0-L_1)^{\mathrm{T}}(L_0-L_1). \tag{A13}$$

于是, 由式(A11)可知$S_0-S_1\geqslant 0$, 因此$S_1\leqslant S_0$. 此外, 由式(A10)有

$$\begin{aligned}S_1-S^*=\int_0^{\infty}\mathrm{e}^{A_1}[&(K_1-K^*)^{\mathrm{T}}R(K_0-K^*)-\\&\gamma^2(L_1-L^*)^{\mathrm{T}}(L_1-L^*)]\mathrm{e}^{A_1}\mathrm{d}t\geqslant 0.\end{aligned} \tag{A14}$$

因此S_1也是有界的. 这样当$k=1$时, A_1特征值具有负实部, 并且满足式(15). 对$k=1,2,\cdots$重复上面的步骤, 得到相应的结论.

由于单调收敛的序列一定存在极限, 因此$\lim\limits_{k\to\infty}S_k=S^*$存在. 当$k\to\infty$, 对式(14)两边求极限, 得到

$$\begin{aligned}&A^{\mathrm{T}}S_\infty+S_\infty A+Q-S_\infty BR^{-1}B^{\mathrm{T}}S_\infty+\\&\gamma^{-2}S_\infty DD^{\mathrm{T}}S_\infty=0.\end{aligned}$$

由于S^*是式(14)的唯一正定解, 因此$S_\infty=S^*$, 从而定理得证.　　证毕.

作者简介:

富　月　(1978–), 女, 副教授, 硕士生导师, 目前研究方向为自适应控制、智能解耦控制、自适应最优控制等, E-mail: fuyue@mail.neu.edu.cn;

柴天佑　(1947–), 男, 中国工程院院士, 教授, 博士生导师, 目前研究方向为全流程一体化控制等, E-mail: tychai@mail.neu.edu.cn.