文章编号: 1000−8152(2011)08−1187−06

基于执行器−评价器学习的自适应PID控制

陈学松[1,2], 杨宜民[2]

(1. 广东工业大学 应用数学学院, 广东 广州 510006; 2. 广东工业大学 自动化学院, 广东 广州 510006)

摘要: 针对传统PID控制器无法在线自整定参数的不足, 提出了一种基于执行器−评估器(Actor-Critic, AC)学习的自适应PID控制器结构与学习算法. 该控制器利用AC学习实现PID参数的自适应整定, 采用一个径向基函数网络同时对Actor的策略函数和Critic的值函数进行逼近. 径向基函数网络的输入为系统误差、误差的一次差分和二次差分, Actor实现系统状态到PID参数的映射, Critic则对Actor的输出进行评判并且生成时序差分(temporal difference, TD)误差信号. 基于AC学习的体系结构和TD误差性能指标, 给出了控制器设计的步骤流程图. 两个仿真实验表明: 与传统的PID控制器相比, 基于AC学习的PID控制器在响应速度和自适应能力方面要优于传统PID控制器.

关键词: 强化学习; 执行器−评价器; 自适应PID控制

中图分类号: TP273　　**文献标识码**: A

A novel adaptive PID controller based on Actor-Critic learning

CHEN Xue-song[1,2], YANG Yi-min[2]

(1. Faculty of Applied Mathematics, Guangdong University of Technology, Guangzhou Guangdong 510006, China;
2. Faculty of Automation, Guangdong University of Technology, Guangzhou Guangdong 510006, China)

Abstract: Owing to the lack of the self-tuning for PID parameters in typical PID(T−PID) controllers, a self tuning PID control strategy using Actor-Critic learning(AC−PID) is proposed. Actor-Critic learning is used to tune PID parameters of the controller in an adaptive way. The policy function of Actor and the value function of Critic are approximated by a simple radial basis function neural network. The system error, the first and the second-order differences of system error are employed as inputs to the radial basis function network. The mapping from the system states to PID parameters is realized by the Actor, and the temporal difference(TD) error is evaluated by the Critic. Based on the structure of Actor-Critic learning and TD error performance index, the block diagram of the controller is developed. Two simulation results show that the proposed controller is efficient and perfectly adaptable with fast responses, providing better performances than the typical PID controller.

Key words: reinforcement learning; Actor-Critic; adaptive PID control

1　引言(Introduction)

在各种控制器设计方法中, PID控制由于具有实现简单、鲁棒性较好等优点, 已经成为一种在工程中广泛应用的控制器设计方法[1]. 但是在控制过程中, PID参数一旦确定下来, 就无法在线调整, 为了解决这个问题, 研究者提出了众多PID参数整定的方法. 根据发展阶段的划分, 可以分为常规PID参数整定方法和智能PID整定方法; 按照被控对象个数来划分, 可以分为单个变量PID参数整定方法及多变量PID参数整定方法[2]. 其中, 自适应PID控制器的设计得到了普遍的关注和研究. 自适应PID控制器的一般设计思路是, 通过根据系统状态的变化而改变PID参数, 以获得更好的控制效果. 目前已提出的自适应PID控制方法主要有: 基于模糊逻辑的自适应PID控制[3], 这种设计方法往往对先验知识要求较多, 并且存在参数优化的问题; 基于神经网络的自适应PID控制[4], 该设计方法一般采用监督学习进行参数的优化, 而监督学习中的教师信号难以获取; 基于进化算法的自适应PID控制器设计方法[5], 虽然对先验知识要求较少, 但是存在计算时间较长, 难以进行实时控制; 基于即时学习的非线性系统自适应PID控制[6], 利用等价多项式的方法, 推导出PID形式的控制律, 从而达到控制精度和计算效率的效果, 但是该方法需要建模数据在时间和空间上具有相邻性, 这在许多系统中无法满足这个要求.

Barto等提出的执行器−评估器(Actor-Critic, AC)学习算法, 也称为自适应启发式评价算法(adaptive heuristic critic, AHC), 是一种重要的强化学习算法[7], 在人工智能和智能控制等领域得到广泛应用. 在AC学习中, 同时对马氏决策过程中的值函数

收稿日期: 2010−05−26; 收修改稿日期: 2010−10−27.
基金项目: 国家自然科学基金资助项目(60974019); 广东省自然科学基金资助项目(9451009001002686).

和策略函数进行逼近, 由于值函数估计的精度直接影响Actor的行为选择和策略梯度的估计, 因此, Critic的值函数预测性能对于算法求解学习控制问题的效率具有关键作用. 本文利用AC学习的无模型在线学习能力, 对PID参数进行自适应调整, 提出一种基于AC学习的自适应PID(AC–PID)控制器设计方法, 该方法可以解决传统PID(T–PID)控制器不易在线实时整定参数的不足, 且具有响应速度快, 自适应能力强等优点.

2 AC学习的自适应PID控制(Adaptive PID controller based on Actor-Critic learning)

2.1 AC–PID控制器结构(The structure of AC–PID)

本文提出的基于AC学习的PID控制器结构如图1所示, 其设计思想来源于增量式PID控制器(incremental PID controller)和AC学习[8]. PID控制器由以下方程设计:

$$\begin{aligned}&u(t)=u(t-1)+\Delta u(t)=\\&u(t-1)+K(t)x(t)=\\&u(t-1)+k_{\rm I}(t)x_1(t)+k_{\rm P}(t)x_2(t)+k_{\rm D}(t)x_3(t)=\\&u(t-1)+k_{\rm I}(t)e(t)+k_{\rm P}(t)\Delta e(t)+k_{\rm D}(t)\Delta^2(t),\end{aligned}\tag{1}$$

其中:

$$\begin{aligned}x(t)&=[x_1(t)\ \ x_2(t)\ \ x_3(t)]^{\rm T}=\\&\quad[e(t)\ \ \Delta e(t)\ \ \Delta^2e(t)]^{\rm T},\\e(t)&=y_{\rm d}(t)-y(t),\\\Delta e(t)&=e(t)-e(t-1),\\\Delta^2e(t)&=2e(t-1)+e(t-2)\end{aligned}$$

分别表示系统的输出误差、误差的一次差分和二次差分; $K(t)=[k_{\rm I}(t)\ \ k_{\rm P}(t)\ \ k_{\rm D}(t)]$为PID参数向量.

如图1, 误差$e(t)=y_{\rm d}-y(t)$经状态转换器(state vector, SV)转换成为AC学习所需要的状态向量$x(t)$. 这里AC学习由3个模块组成: Actor, Critic和随机动作修正器(stochastic action modifier, SAM), Actor用于进行策略估计, 将系统状态变量映射为PID参数$K'(t)=[k'_{\rm I}(t)\ \ k'_{\rm P}(t)\ \ k'_{\rm D}(t)]$, Actor输出的参数$K'(t)$不会直接参与到PID控制器的设计之中, 而是由SAM根据Critic提供的值函数估计信息$V(t)$进行随机修正, 从而得到实际的PID参数值$K(t)=[k_{\rm I}(t)\ \ k_{\rm P}(t)\ \ k_{\rm D}(t)]$. Critic接受系统的状态变量和如式(2)定义的回报函数r_t, 对强化学习每个时间段产生评判, 经过处理之后产生TD误差$\delta_{\rm TD}(t)$及值函数的估计$V(t)$, 其中$\delta_{\rm TD}(t)$直接提供给Actor和Critic, 用于Actor和Critic更新各种参数的重要依据.

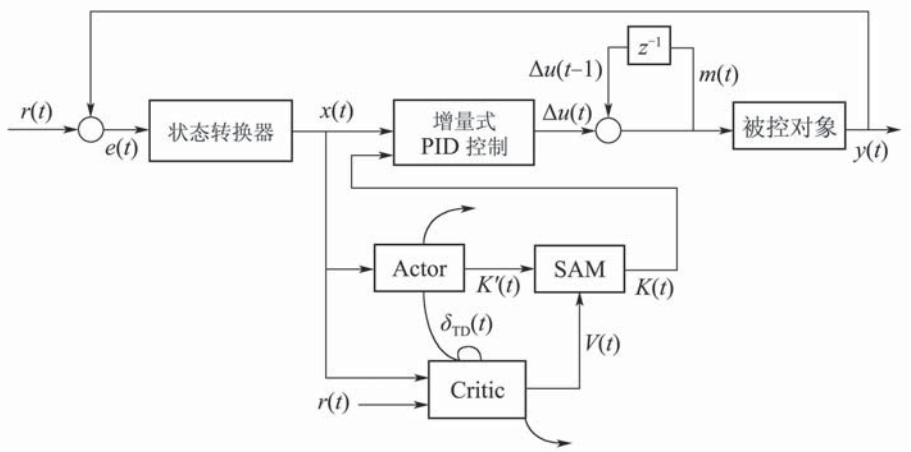


图 1 基于Actor-Critic学习的自适应PID控制结构框图

Fig. 1 The structure of an adaptive PID controller based on Actor-Critic learning

考虑到系统误差和误差变化率对系统控制性能的影响, 回报函数定义为

$$r_t=\alpha_1r_1(t)+\alpha_2r_2(t),\tag{2}$$

其中: α_1,α_2分别为误差变化率的强化信号系数, $r_1(t),r_2(t)$分别为误差和误差变化率的强化信号, 且定义为

$$r_1(t)=\begin{cases}0, & |e(t)|\leqslant\epsilon,\\-1, & \text{其他;}\end{cases}\tag{3}$$

$$r_2(t)=\begin{cases}0, & |e(t)|\leqslant e(t-1),\\-1, & \text{其他.}\end{cases}\tag{4}$$

其中ϵ为容许的误差带.

2.2 径向基函数网络的Actor-Critic学习(Actor-Critic learning based on RBF network)

径向基函数(RBF)网络是一种多层前向神经网络, 和其他前向网络相比, 它具有全局逼近能力强, 结构简单, 训练方法快速易行的优点. 由于

Actor和Critic的输入均来自于环境的状态变量, 两者的区别在于输出不同. 因此, 本文采用一个RBF网络同时实现Actor的策略函数和Critic的值函数学习, 如图2所示. Actor和Critic共享RBF网络的输入层和隐层的资源, 这样不仅可以降低学习系统对存储空间的要求, 同时还可以避免隐层节点输出的重复计算, 从而提高系统的学习效率.

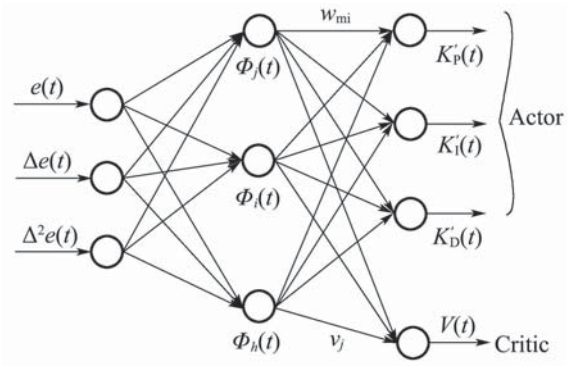


图 2 基于RBF网络的Actor-Critic学习

Fig. 2 Actor-Critic learning based on RBF network

第1层　输入层. 该层的每个输入节点代表状态向量

$$x(t)=[x_1\ \ x_2(t)\ \ x_3(t)]^{\mathrm{T}}=[e(t)\ \ \Delta e(t)\ \ \Delta^2 e(t)]^{\mathrm{T}}\in\mathbb{R}^3$$

的分量.

第2层　隐层. RBF隐层节点的基函数选用高斯型核函数, 第j个隐节点的输出为

$$\Phi_j(t)=\exp(-\frac{||x(t)-\mu_j(t)||^2}{2\sigma_j^2}),\ j=1,\cdots,h, \tag{5}$$

其中:

$$\mu_j=[\mu_{1j}\ \ \mu_{2j}\ \ \mu_{3j}]^{\mathrm{T}}$$

为第j个节点的中心向量, σ_j为第j个节点的宽度参数, h为隐节点个数.

第3层　输出层. 由Actor和Critic两部分组成. Actor第m个输出节点的权值和到Critic的输出值函数$V(t)$分别由下面式子计算:

$$K'_{\mathrm{m}}(t)=\sum_{j=1}^{h}w_{\mathrm{m}j}(t)\Phi_j(t),\ m=1,2,3,\quad j=1,2,\cdots,h, \tag{6}$$

$$V(t)=\sum_{j=1}^{h}V_j(t)\Phi_j(t), \tag{7}$$

其中: $w_{\mathrm{m}j}, v_j$分别对应于隐层第j个节点到Actor的权值和到Critic输出节点的权值.

如图1所示, Actor的输出并不直接传递给PID控制器, 而是在其输出的PID参数$K'(t)$上叠加一个高斯干扰η_k. 高斯干扰的大小取决于$V(t)$, $V(t)$大则干扰小, 反之则大[8]. 具体方法如下:

$$K(t)=K'(t)+\eta_k(0,\sigma_V(t)), \tag{8}$$

其中方差

$$\sigma_V(t)=\frac{1}{1+e^{2V(t)}}.$$

在AC学习中的误差δ_{TD}由状态转移中相邻状态值函数的时序差分来计算:

$$\delta_{\mathrm{TD}}=r(t)+\gamma V(t+1)-V(t), \tag{9}$$

其中$0<\gamma<1$为折扣因子. 误差δ_{TD}反映了所选动作的优劣程度, 定义系统的学习性能指标为

$$E(t)=\frac{1}{2}\delta_{\mathrm{TD}}^2(t). \tag{10}$$

采用梯度下降法对Actor网络的权值进行迭代更新, 即Actor网络t时刻的权值加上Actor学习率乘以一个策略梯度作为$t+1$ 时刻的权值, 其迭代公式为

$$w_{\mathrm{m}j}(t+1)=w_{\mathrm{m}j}(t)+\alpha_{\mathrm{A}}\frac{\partial E}{\partial w}, \tag{11}$$

其中: α_{A}为Actor的学习率, $\dfrac{\partial E}{\partial w}$为策略梯度, 而又易得

$$\frac{\partial E}{\partial \delta_{\mathrm{TD}}}=\frac{\partial(\frac{1}{2}\delta_{\mathrm{TD}}^2)}{\partial\delta_{\mathrm{TD}}}=\delta_{\mathrm{TD}}. \tag{12}$$

由于δ_{TD}是TD误差, 且Actor的PID参数$K'(t)$上叠加一个方差为$\sigma_V(t)$的高斯误差, 因此可采用如下的近似策略梯度估计算法:

$$\frac{\partial\delta_{\mathrm{TD}}}{\partial w}\approx\frac{K_{\mathrm{m}}(t)-K'_{\mathrm{m}}(t)}{\sigma_V(t)}\Phi_j(t). \tag{13}$$

故由偏导数的链式法则有

$$\frac{\partial E}{\partial w}=\frac{\partial E}{\partial\delta_{\mathrm{TD}}}\frac{\partial\delta_{\mathrm{TD}}}{\partial w}=\delta_{\mathrm{TD}}\frac{\partial\delta_{\mathrm{TD}}}{\partial w}\approx\delta_{\mathrm{TD}}\frac{K_{\mathrm{m}}(t)-K'_{\mathrm{m}}(t)}{\sigma_V(t)}\Phi_j(t). \tag{14}$$

把式(14)代入式(11)则有

$$w_{\mathrm{m}j}(t+1)=w_{\mathrm{m}j}(t)+\alpha_{\mathrm{A}}\delta_{\mathrm{TD}}(t)\frac{K_{\mathrm{m}}(t)-K'_{\mathrm{m}}(t)}{\sigma_V(t)}\Phi_j(t). \tag{15}$$

同理可以得到

$$v_j(t+1)=v_j(t)+\alpha_{\mathrm{C}}\delta_{\mathrm{TD}}(t)\Phi_j(t), \tag{16}$$

其中$\alpha_{\mathrm{A}}, \alpha_{\mathrm{C}}$分别为Actor和Critic的学习率.

由于Actor和Critic共享RBF网络的输入层和隐层资源, 因此, 在网络的学习阶段, 只要对隐层节点的中心和宽度做一次更新即可. 具体更新如下:

$$\mu_{mj}(t+1)=\mu_{ij}(t)+\alpha_\mu\delta_{\rm TD}(t)v_j(t)\Phi_j(t)\frac{x_i(t)-\mu_{ij}(t)}{\sigma_j^2(t)},\tag{17}$$

$$\sigma_j(t+1)=\sigma_j(t)+\alpha_\sigma\delta_{\rm TD}(t)v_j(t)\Phi_j(t)\frac{||x(t)-\mu_j(t)||^2}{\sigma_j^3(t)},\tag{18}$$

其中α_μ,α_σ分别为中心和宽度的学习率.

2.3 控制器的设计流程图(The block diagram of AC–PID)

根据上面的分析给出基于AC学习的自适应PID控制器的设计流程图如下:

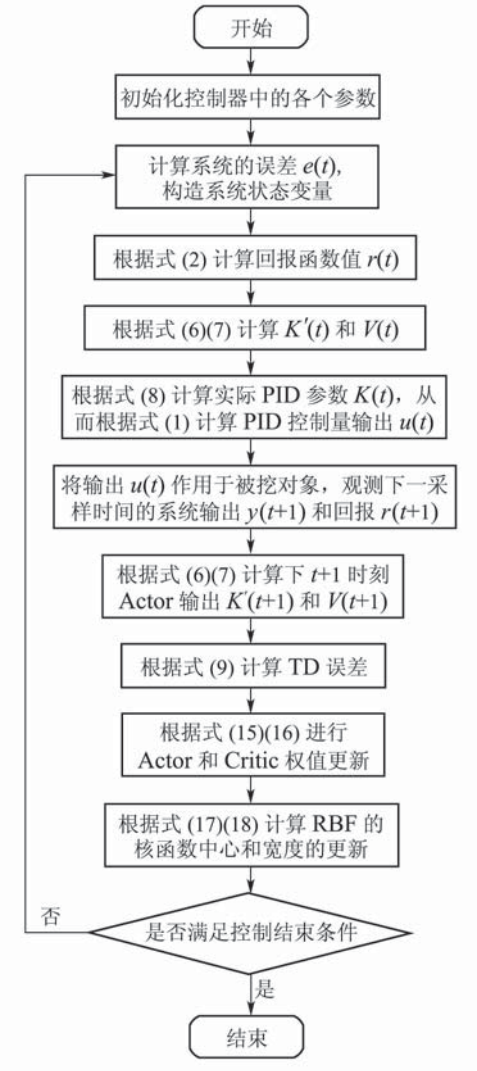


图 3 基于Actor-Critic学习的自适应PID控制器设计流程

Fig. 3 The block diagram of AC–PID controller

3 仿真实验(Simulations)

3.1 非线性系统仿真(Simulation on nonlinear system)

为了能够说明系统在动态特性发生变化时, PID参数能够实现动态的自整定. 设被控对象为

$$\frac{{\rm d}^2y}{{\rm d}t^2}+20\frac{{\rm d}y}{{\rm d}t}+25y(t)=133u(t).\tag{19}$$

定义其对应的参考模型为

$$\frac{{\rm d}^2y_{\rm m}}{{\rm d}t^2}+20\frac{{\rm d}y_{\rm m}}{{\rm d}t}+25y_{\rm m}(t)=133r(t).\tag{20}$$

式(19)(20)中: $y(t)$为系统输出, $u(t)$为控制输入, $y_{\rm m}(t)$为模型输出, $r(t)$为系统指令输入. 定义信号误差为

$$e(t)=y_{\rm m}(t)-y(t),\tag{21}$$

则由式(19)~(21)可得到误差动态方程:

$$\frac{{\rm d}^2e}{{\rm d}t}+20\frac{{\rm d}e}{{\rm d}t}+30e(t)=50r(t)-133u(t)-5\theta(t).\tag{22}$$

仿真过程中, 被控对象初始状态取$[0,0]$, 参考模型初始状态取$[0,0]$, 设采样时间为1 s, 指令信号分两种, 当$S=1$时为方波, 当$S=2$时为正弦波. 采用本文提出的基于AC学习的自适应PID控制, 各个参数设置为:

$$\alpha_1=0.75,\ \alpha_2=0.25,$$
$$\alpha_{\rm A}=0.02,\ \alpha_{\rm C}=0.04,$$
$$\alpha_\mu=0.025,\ \alpha_\sigma=0.02,$$
$$\epsilon=0.001,\ \gamma=0.99.$$

仿真结果如图4~7所示, 图4和图5分别为参考模型位置跟踪及跟踪误差, 图6为控制器输出, 图7为PID控制器的自适应变换过程. 从仿真结果图可以看出: 基于AC学习的自适应PID控制具有响应速度快, 自适应能力强等优点.

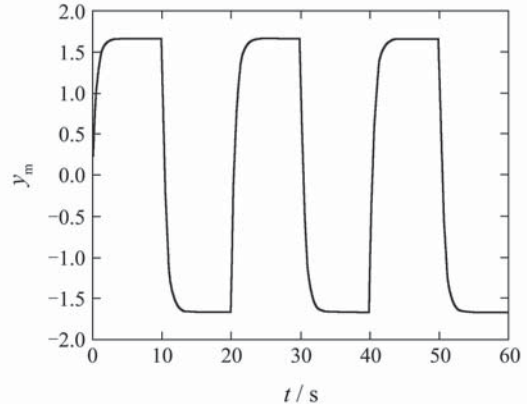


图 4 位置跟踪

Fig. 4 Position tracking

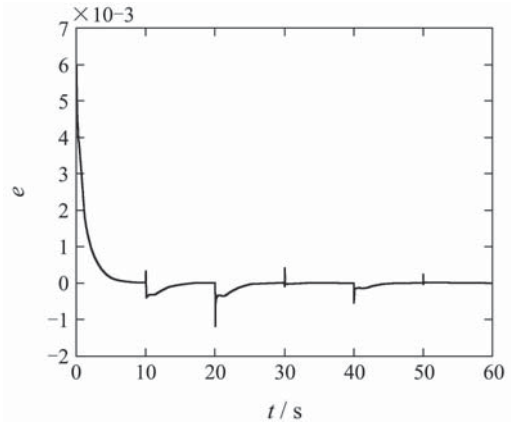


图 5 位置跟踪误差

Fig. 5 Position tracking error

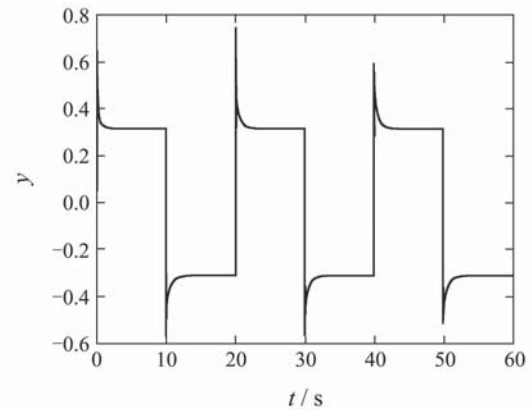


图 6 控制输出信号

Fig. 6 Control output

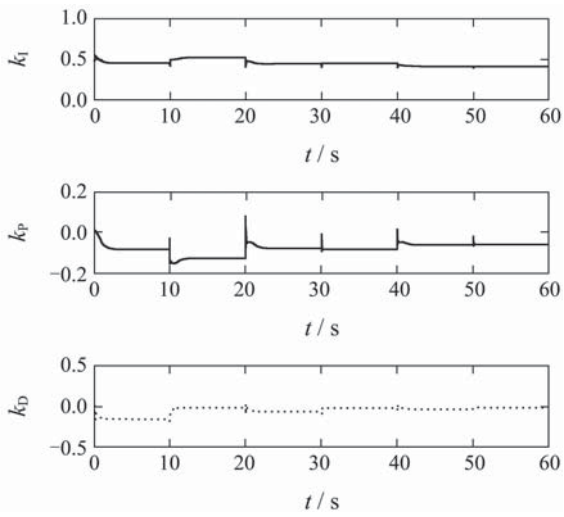


图 7 AC–PID控制中$k_{\rm I}, k_{\rm P}, k_{\rm D}$的整定结果

Fig. 7 The result of AC–PID parameter tuning on $k_{\rm I}, k_{\rm P}, k_{\rm D}$

3.2 倒立摆仿真(Simulation on cart pole)

单级倒立摆的控制问题一直是控制研究中的一个典型问题. 控制的目的是通过给小车底部施加一个力u(控制量), 使小车停留在预定的位置, 并使杆不倒下, 即不超过一个预定的垂直偏离角度范围. 设小车质量为M, 摆的质量为m, 小车位置为x, 摆的角度为θ, 则可以推出单级倒立摆方程:

$$\ddot{\theta} = \frac{m(m+M)gl}{(m+M)I+mMl^2}\theta - \frac{ml}{(m+M)l+mMl^2}u, \tag{23}$$

$$\ddot{x} = \frac{m^2gl^2}{(m+M)I+mMl^2}\theta + \frac{I+ml^2}{(m+M)l+mMl^2}u. \tag{24}$$

其中: $I=\dfrac{1}{12}mL^2$, $l=\dfrac{1}{2}L$. 控制指标共有4个, 即单级倒立摆的摆角θ、摆速$\dot{\theta}$、小车位置x和小车速度$\dot{x}$.

仿真中, 倒立摆的参数为:

$$g=9.8\,\mathrm{m/s^2},\ M=10\,\mathrm{kg},\ m=0.1\,\mathrm{kg},$$

$$L=0.5\,\mathrm{m},\ \mu_{\rm c}=0.0005,\ \mu_{\rm p}=0.000002,$$

其中：$\mu_{\rm c}$为小车相对于导轨的摩擦系数, $\mu_{\rm p}$为杆相对于小车的摩擦系数. AC–PID控制器的各个参数设置为:

$$a_1=0.7,\ a_2=0.3,\ \alpha_{\rm A}=0.02,\ \alpha_{\rm C}=0.03,$$

$$\alpha_\mu=0.025,\ \alpha_\sigma=0.02,\ \epsilon=0.001,\ \gamma=0.99.$$

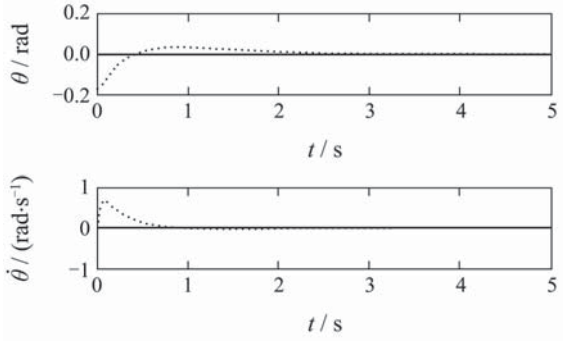
F为作用于小车的力, 即在$[-10, 10]$上连续取值的控制器输出. 采样周期为$T=10\,\mathrm{ms}$, 初始条件为:

$$\theta(0)=-10^o,\ \dot{\theta}(0)=0,\ x(0)=0,\ \dot{x}(0)=0,$$

期望状态为:

$$\theta(0)=0,\ \dot{\theta}(0)=0,\ x(0)=0,\ \dot{x}(0)=0.$$

仿真结果如图8和图9所示, 图8为5 s内, 单级倒立摆的摆角(angle)、摆速(angle rate)、小车位置(cart position)和小车速度(cart rate)的响应曲线, 该曲线表明了在AC–PID控制下, 倒立摆的4个控制指标都能快速获得稳定跟踪控制. 图9是传统PID控制器(T–PID)和AC–PID控制器的输出结果比较, 从图中的结果可以看出, AC–PID控制的效果要优于T–PID控制方法.



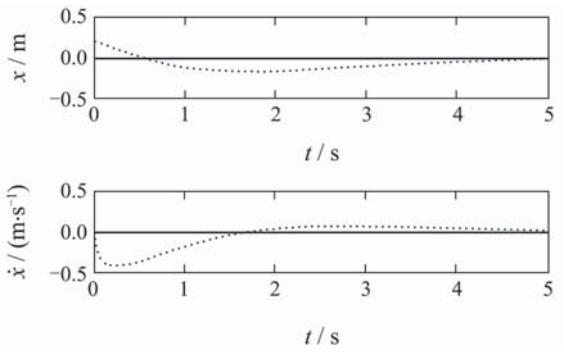


图 8 AC–PID控制下4个控制指标的响应结果

Fig. 8 The response of four index with AC–PID

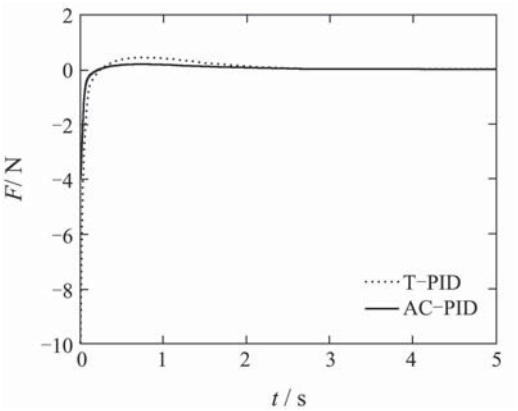


图 9 T–PID与AC–PID控制器输出对比

Fig. 9 The output of controller between T–PID and AC–PID

4 结论(Conclusion)

PID控制是过程控制中应用最为广泛的控制方法, 然而传统PID控制器无法在线自整定参数, 本文针对传统的PID控制的这个不足, 探讨了它与AC学习结合的可能性, 利用强化学习在线调节PID的参数, 提出了一种基于AC学习的自适应PID控制方法. 该方法中Actor和Critic共享径向基函数网络的工作机制, 不仅可以降低学习系统对存储空间的要求, 同时还可以避免隐层节点输出的重复计算, 从而达到减少计算量的目的. 仿真实验表明: AC–PID控制方法可以实现对非线性系统的稳定跟踪控制, 并且与传统的PID控制(T–PID)相比, 基于AC学习的自适应PID控制(AC–PID)具有响应速度快, 自适应能力强, 抗干扰能力强等优点.

参考文献(References):

[1] ANG K H, CHONG G, LI Y. PID control system analysis, design, and technology[J]. *IEEE Transactions on Control Systems Technology*, 2005, 13(4): 559 – 576.

[2] 王伟, 张晶涛, 柴天佑. PID参数先进整定方法综述[J]. 自动化学报, 2000, 26(3): 347 – 355.
(WANG Wei, ZHANG Jingtao, CHAI Tianyou. A survey of advanced pid parameter tuning methods[J]. *Acta Automatica Sinica*, 2000, 26(3): 347 – 355.)

[3] CARVAJAL J, CHEN G, OGMEN H. Fuzzy PID controller: design, performance evaluation, and stability analysis[J]. *Information Sciences*, 2000, 123(4): 249 – 270.

[4] CHEN J H, HUANG T C. Applying neural networks to on-line update PID controllers for nonlinear process control[J]. *Journal of Process Control*, 2004, 14(2): 211 – 230.

[5] WU L, WANG Y, ZHOU S. et al. Design of PID controller with incomplete derivation based on differential evolution algorithm[J]. *Journal of Systems Engineering and Electronics*, 2008, 19(3): 578 – 583.

[6] 潘天红, 李少远. 基于即时学习的非线性系统自适应PID控制[J]. 控制理论与应用, 2009, 26(10): 1180 – 1184.
(PAN Tianhong, LI Shaoyuan. Adaptive PID control for nonlinear systems based on lazy learning[J]. *Control Theory & Applications*, 2009, 26(10): 1180 – 1184.)

[7] BARTO A G, SUTTON R S, ANDERSON C W. Neuronlike adaptive elements that can solve difficult learning control problems[J]. *IEEE Transactions on Systems, Man and Cybernetics*, 1983, 13(5): 834 – 846.

[8] 王雪松, 程玉虎. 机器学习理论、方法及应用[M]. 北京: 科学出版社, 2009.
(WANG Xuesong, CHENG Yuhu. *Machine Learning in Theorem, Method and Application*[M]. Beijing: Science Press, 2009.)

作者简介:

陈学松 (1978—), 男, 讲师, 博士研究生, 目前研究方向为智能控制与人工智能, E-mail: chenxs@gdut.edu.cn;

杨宜民 (1945—), 男, 教授, 博士生导师, 目前研究方向为智能机器人, E-mail: yangy@gdut.edu.cn.