采用强化学习的多轴运动系统时间最优轨迹优化

张铁，廖才磊，邹焱飚，康中强
（华南理工大学机械与汽车工程学院，510641，广州）

摘要：为实现多轴运动系统高速运动并解决电机动载荷过载的问题，提出了一种采用强化学习的时间最优轨迹优化方法。使用改进状态-动作-奖励-状态-动作（SARSA）算法和迭代交互法来寻找时间最优轨迹：通过改进 SARSA 算法与基于运动学模型建立的强化学习环境进行交互学习，找到满足运动学约束的初始策略轨迹；通过迭代交互法与真实环境进行交互学习，从而将电机动态载荷约束引入到强化学习环境中并对策略轨迹进行修正；最终得到满足电机动态载荷约束的时间最优轨迹。在自行搭建的两轴运动系统上进行验证，结果表明，改进 SARSA 算法优化得到的策略轨迹的速度和加速度曲线均在约束范围内，且经过 10 次迭代后的轨迹实际测量力矩曲线也在电机动载荷约束范围内，所提方法能够得到同时满足运动学约束和动力学约束的时间最优轨迹。

关键词：多轴运动系统；电机动载荷过载；时间最优轨迹优化；强化学习
中图分类号：TP273.1 **文献标志码**：A
DOI：10.7652/xjtuxb202107004 **文章编号**：0253-987X(2021)07-0033-08

OSID 码

Time-Optimal Trajectory Optimization of Multi-Axis Motion System by Reinforcement Learning

ZHANG Tie, LIAO Cailei, ZOU Yanbiao, KANG Zhongqiang
(School of Mechanical and Automotive Engineering, South China University of Technology, Guangzhou 510641, China)

Abstract: To realize high-speed motion of the multi-axis motion system and solve the problem of dynamic overload for motor, a time-optimal trajectory optimization scheme with reinforcement learning is proposed. This scheme chooses an improved state-action-reward-state-action (SARSA) algorithm and iterative interaction to find the time-optimal trajectory. The agent interacts with the reinforcement learning environment established by the kinematics model via the improved SARSA algorithm to find the initial trajectory satisfying the kinematics constraints. Then the agent interacts with the real environment by iterative interaction to introduce the motor dynamic load constraints into the reinforcement learning environment and modify the strategy trajectory so as to obtain the time-optimal trajectory satisfying motor dynamic load constraints. The time-optimal trajectory is verified on a self-built two-axis motion system. The results show that the speed and acceleration curves of the trajectory optimized by the improved SARSA algorithm are within the constraint ranges and the actual measured torque curve of the trajectory after 10 iterations is also within the dynamic load constraint ranges of motor, thus the time-optimal trajectory satisfying both kinematic constraints and dynamic load constraints of motor is obtained by this way.

收稿日期：2020-11-25。 作者简介：张铁（1968—），男，教授，博士生导师。 基金项目：国家重点研发计划资助项目（2020YFC2007603）。
网络出版时间：2021-03-19 网络出版地址：http://kns.cnki.net/kcms/detail/61.1069.T.20210318.1604.002.html

Keywords: multi-axis motion system; motor dynamic overload; time-optimal trajectory optimization; reinforcement learning

在多轴运动系统的自动化装配、搬运等生产过程中,往往要求系统能够高速运动以提高生产效率,但是系统的高速运动可能会导致电机动载荷过载的问题,这一现象在重载运动时尤为明显,不仅对电机产生损害,还降低了生产质量。

为实现多轴运动系统的高速运动,需要以时间最优为目标对轨迹进行优化[1-3]。为了进一步解决高速运动过程中的电机动载荷过载问题,需要将电机动态载荷视为约束条件引入到时间最优轨迹优化数学模型中,如文献[4]通过对系统建立动力学模型,并在轨迹优化时增加电机动态载荷约束,从而得到理论计算力矩,满足电机动载荷约束的时间最优轨迹。但是,由于电机动子电流和定子磁场畸变以及机械参数变化等扰动,所建立的动力学模型和实际模型无法完全匹配,最终实际测量力矩仍会超出电机动载荷限制。因此,文献[5]在动力学模型中增加补偿项并利用迭代学习方法来更新补偿项,以此提高动力学模型精度。虽然能够在一定程度上减小模型不匹配度,但其仍是对理论计算力矩进行约束,无法从根本上解决高速运动中电机动载荷过载的问题。

近年来,强化学习[6]在控制领域已经有了广泛的研究应用,例如移动路径规划[7-8]、步态生成[9-10]、避障[11-12]和装配[13-14]等。这些研究用马尔可夫决策过程来描述实际物理问题,然后利用强化学习方法进行学习求解。本文也基于马尔可夫决策过程对时间最优轨迹优化问题进行研究,利用强化学习中智能体与环境进行交互学习的特性,提出了一种采用强化学习的时间最优轨迹优化方法:首先,将时间最优轨迹优化问题描述为马尔可夫决策问题;其次,对经典的状态-动作-奖励-状态-动作(SARSA)算法[15]进行改进,使其能适用于时间最优轨迹优化;最后,通过迭代交互法与真实环境交互学习,从而获得满足运动学约束和动力学约束的时间最优轨迹。本文方法无需建立动力学模型,而是直接对电机实际力矩进行约束,避免了动力学模型和实际模型不匹配的问题,从根本上解决了多轴运动系统在高速运动中电机动载荷过载的问题。

1 时间最优轨迹优化的强化学习环境

时间最优轨迹优化问题实质上是寻找满足约束条件的电机运动轨迹,使其以尽可能大的速度工作,从而达到时间最优。为了采用强化学习方法解决时间最优轨迹优化问题,首先需要基于运动学约束条件来构造强化学习环境,动力学约束的引入需要通过与真实环境的迭代交互来完成,具体将在 2.2 小节研究。

1.1 时间最优轨迹优化的运动学约束

多轴运动系统包括串联结构和并联结构,本文主要研究串联结构,如数控机床、串联工业机器人等。图 1 是多轴运动系统示意,可以看出,轴的运动方式主要分为转动和移动两种,前 3 轴为转动轴,第 4 轴为移动轴。结构的选用及组合需要根据应用场景来设计,本文暂不考虑具体结构,只针对电机进行描述,记电机转角关于时间 t 的函数为 $\boldsymbol{q}(t)=[q_1(t)\ q_2(t)\ \cdots]^{\mathrm{T}}$。

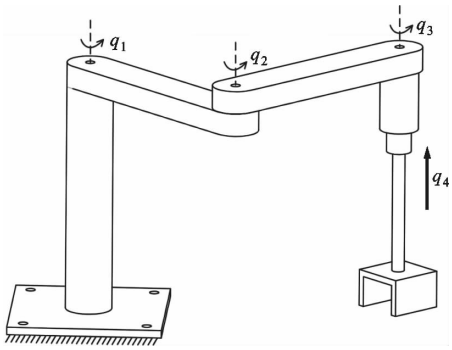


图 1 多轴运动系统示意

Fig. 1 Schematic diagram of the multi-axis motion system

为了减少优化过程中的变量,轨迹优化往往在相平面上进行[16]。假设路径标量 $s\in[0,1]$[17]与时间 t 存在函数映射关系 $s(t)$,则有

$$\boldsymbol{q}(t)=\boldsymbol{q}(s) \tag{1}$$

进一步地,可得到电机角速度和角加速度关于标量 s 的表达式

$$\dot{\boldsymbol{q}}(t)=\dot{\boldsymbol{q}}(s)=\boldsymbol{q}'(s)\dot{s} \tag{2}$$

$$\ddot{\boldsymbol{q}}(t)=\ddot{\boldsymbol{q}}(s)=\boldsymbol{q}'(s)\ddot{s}+\boldsymbol{q}''(s)\dot{s}^2 \tag{3}$$

式中:$\dot{s}=\mathrm{d}s/\mathrm{d}t$;$\ddot{s}=\mathrm{d}^2s/\mathrm{d}t^2$;$\boldsymbol{q}'(s)=\partial\boldsymbol{q}(s)/\partial(s)$;$\boldsymbol{q}''(s)=\partial^2\boldsymbol{q}(s)/\partial s$。

给定电机转速限制 $\dot{\boldsymbol{q}}_{\max}$ 和 $\dot{\boldsymbol{q}}_{\min}$,则可得到速度约束不等式

$$\dot{\boldsymbol{q}}_{\min}\leqslant\dot{\boldsymbol{q}}(s)\leqslant\dot{\boldsymbol{q}}_{\max} \tag{4}$$

将式(4)代入式(2)中,得到标量速度 $\dot{s}$ 的不等式约束

$$\dot{\boldsymbol{q}}_{\min}/\boldsymbol{q}'(s)\leqslant\dot{\boldsymbol{s}}\leqslant\dot{\boldsymbol{q}}_{\max}/\boldsymbol{q}'(s) \tag{5}$$

给定电机的加速度限制 $\ddot{\boldsymbol{q}}_{\max}$ 和 $\ddot{\boldsymbol{q}}_{\min}$，则可得到加速度约束不等式

$$\ddot{\boldsymbol{q}}_{\min} \leqslant \ddot{\boldsymbol{q}}(s) \leqslant \ddot{\boldsymbol{q}}_{\max} \tag{6}$$

将式(6)代入式(3)，得到标量加速度 $\ddot{s}$ 的不等式约束

$$\frac{(\ddot{\boldsymbol{q}}_{\min} - \boldsymbol{q}''(s)\dot{s}^2)}{\boldsymbol{q}'(s)} \leqslant \ddot{s} \leqslant \frac{(\ddot{\boldsymbol{q}}_{\max} - \boldsymbol{q}''(s)\dot{s}^2)}{\boldsymbol{q}'(s)} \tag{7}$$

根据式(7)，可以得到运动学加速度约束下的标量加速度上下极限 $\alpha(s,\dot{s})$ 和 $\beta(s,\dot{s})$。由文献[18]，相平面$(s,\dot{s})$上满足

$$\alpha(s,\dot{s}) = \beta(s,\dot{s}) \tag{8}$$

的曲线即为加速度约束下的最大速度限制曲线 $\dot{s}_{\lim,\mathrm{acc}}$。再进一步考虑由速度不等式约束式(5)得到的最大速度限制曲线 $\dot{s}_{\lim,\mathrm{vel}}$，最终得到满足运动学约束的最大速度限制曲线 $\dot{s}_{\lim}=\min(\dot{s}_{\lim,\mathrm{acc}},\dot{s}_{\lim,\mathrm{vel}})$。

1.2　基于运动学约束的强化学习环境

对任一条运行时间为 t_{f} 的完整电机轨迹，不失一般性地，假设

$$s(0) = 0 \leqslant s(t) \leqslant 1 = s(t_{\mathrm{f}}) \tag{9}$$

此外，由于轨迹的起始点和终止点的速度均为0，即 $\dot{\boldsymbol{q}}(0)=\dot{\boldsymbol{q}}(t_{\mathrm{f}})=\boldsymbol{0}$，则根据式(2)可得

$$\dot{s}(0) = \dot{s}(t_{\mathrm{f}}) = 0 \tag{10}$$

根据式(9)(10)，时间最优轨迹优化问题可描述为：在相平面$(s,\dot{s})$上寻找一条从起始点(0,0)出发到目标点(1,0)结束的满足 $\dot{s}_{\lim}$ 约束的使标量速度 $\dot{s}$ 最大的轨迹。因此，可构建时间最优强化学习环境，如图 2 所示，其中，相平面中未超过 $\dot{s}_{\lim}$ 的部分为可行区域，其他部分为不可行区域即障碍。此外，为了让最优轨迹最终能够到达目标点(1,0)，需要将直线 $s=1$ 上除了(1,0)外的其他点全部设置为不可行点。

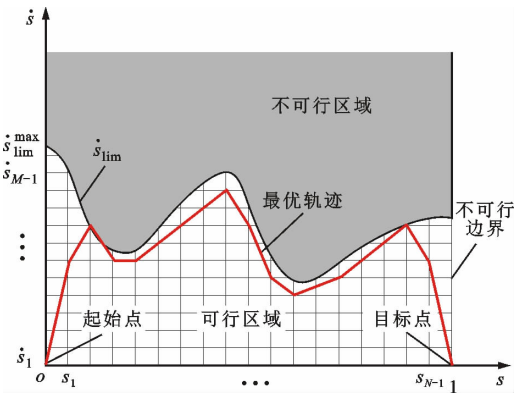


图 2　时间最优强化学习环境示意

Fig. 2　Schematic diagram of time-optimal reinforcement learning environment

进一步地，为了采用强化学习方法解决时间最优轨迹优化问题，需要将相平面划分成多个网格。s 轴方向依据路径标量曲率变化率划分，将标量位移区间[0,1]等分为 N 段，网格点依次为 $s_0,s_1,\cdots,s_N$，其中 $s_0=0,s_N=1$。$\dot{s}$ 轴方向采用均匀划分，记 $\dot{s}_{\lim}^{\max}$ 为最大速度限制曲线 $\dot{s}_{\lim}$ 上的最大标量速度，则区间 $[0,\dot{s}_{\lim}^{\max}]$ 等分为 M 段，网格点依次为 $\dot{s}_0,\dot{s}_1,\cdots,\dot{s}_M$，其中 $\dot{s}_0=0,\dot{s}_M=\dot{s}_{\lim}^{\max}$。最终，相平面被划分成 $N\times M$ 的网格，而一个网格点$(s_k,\dot{s}_k)$即为一个状态 S_k。

2　时间最优轨迹优化的强化学习算法

2.1　改进 SARSA 算法

SARSA 算法[19]是一种经典的强化学习算法，其中，状态是智能体在环境中的位置，动作是智能体从一个状态转移到另一个状态所采取的行动，奖励是环境对智能体的一个反馈。记 $Q(S_k,A_k)$ 为智能体在状态 S_k 采取动作 A_k 时获得的平均奖励，则用于时间最优轨迹优化的 SARSA 算法的学习过程如图 3 所示。智能体从状态 S_0 开始，在状态 S_k 根据 $Q(S_k,A_k)$ 采取一个动作 A_k，之后从环境中获得奖励 R_{k+1}，更新 $Q(S_k,A_k)$ 并到达下一个状态 S_{k+1}，这一连串状态动作的组合称为一个情节。

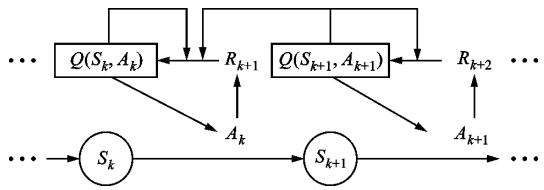


图 3　SARSA 学习过程

Fig. 3　SARSA learning procedure

2.1.1　动作　SARSA 算法中，动作的选择采用 ε 贪心法，公式为

$$A_{k+1} = \begin{cases} \text{随机动作}, & \lambda < \varepsilon \\ \arg\max\limits_{A_{k+1}} Q(S_{k+1},A_{k+1}), & \lambda > \varepsilon \\ (1,0), & \text{状态点}(1,0)\text{在动作范围内} \end{cases} \tag{11}$$

式中 λ 为随机数。

ε 贪心法基于概率 $\varepsilon(0\leqslant\varepsilon<1)$ 对探索和利用进行折中：每次学习时，以 ε 的概率进行探索，即以均匀概率随机选择一个动作，以发现可以获得更大奖励的动作；以 $1-\varepsilon$ 的概率进行利用，即选择当前状态下最大的 Q 所对应的动作，以尽可能多地获得奖励。此外，当处于倒数第二个状态时，如果通过计算发现终止目标状态点(1,0)刚好在可选择的动作范围内，那么这个终止目标状态为唯一可选的动作。

随机动作选择范围的计算需要遵循一定的运动规则。当强化学习智能体在状态 S_k 时，根据约束条件式(7)可以求出此状态下的最大标量加速度 $\dot{s}_{k,\max}$ 和最小标量加速度 $\dot{s}_{k,\min}$。由匀加速运动方程

$$\dot{s}_{k+1}=\sqrt{2\dot{s}_{k,\max/\min}(s_{k+1}-s_k)+\dot{s}_k^2} \tag{12}$$

可以求出状态 S_{k+1} 的标量速度范围 $[\dot{s}_{k+1,\min},\dot{s}_{k+1,\max}]$。因此，在直线 $s=s_{k+1}$ 上，标量速度范围 $[\dot{s}_{k+1,\min},\dot{s}_{k+1,\max}]$ 即为动作范围，范围内的网格点即为可供选取的动作。

2.1.2 *奖励* 为了获得具有最大标量速度的时间最优轨迹，需要根据强化学习智能体在环境中的状态给予奖励或惩罚，以规范学习行为。图 4 是时间最优轨迹优化问题的奖励和惩罚示意。可以看出，在最大速度限制曲线之下的可行区域内，越往上的区域应该给予越多的奖励。当智能体位于这些区域时，虽然越往上意味着可以获得越多的奖励，但同时也意味着智能体越靠近限制边界，即可能受到惩罚的概率也应随着增加。为了体现这一特点，需要将奖励和惩罚都设置为与标量速度 $\dot{s}$ 相关。时间最优强化学习问题的奖励和惩罚设置为

$$R_{k+1}=\begin{cases}\dot{s}_k+\dot{s}_{k+1}, & \text{奖励}\\ -(\dot{s}_k+\dot{s}_{k+1}), & \text{惩罚}\end{cases} \tag{13}$$

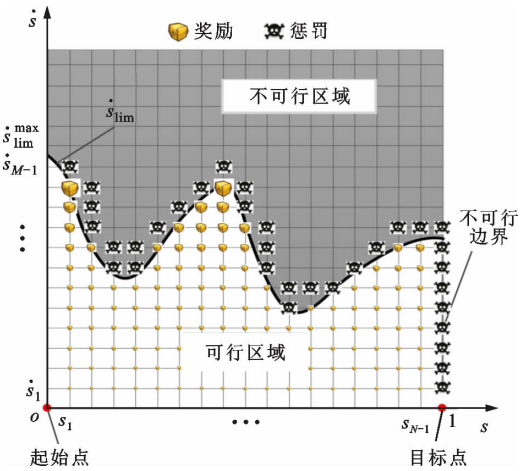


图 4 时间最优轨迹优化问题的奖励和惩罚示意

Fig. 4 Schematic diagram of rewards and penalties for the time-optimal trajectory optimization problem

2.1.3 *Q 表* 在经典的 SARSA 算法中，状态动作函数为

$$Q(S_k,A_k)\leftarrow Q(S_k,A_k)+\alpha(R_{k+1}+\gamma Q(S_{k+1},A_{k+1})-Q(S_k,A_k)) \tag{14}$$

式中：α 为学习系数，$0\leqslant\alpha<1$，α 越大则表示靠后的积累奖励越重要；γ 表示折扣因子，$0\leqslant\gamma<1$。

由于 SARSA 算法是一个单步更新的强化学习算法，当智能体到达一个不可行状态点并受到惩罚时，它只能通过单步更新，将惩罚往前回溯到前一步。因此，需要相当多的时间去更新 Q 才能使不可行状态点的 Q 小于 0，以保证智能体不再经过这些状态点。为了加速这个惩罚的传导过程，应该在智能体经历过一段失败的情节时，便对该段情节上的所有动作进行惩罚，使该段情节上的所有动作都能从这次失败的情节中学到经验。改进的状态动作函数通过在状态动作后面增加一个惩罚项用来加速惩罚，公式为

$$Q(S_k,A_k)\leftarrow Q(S_k,A_k)+\alpha(R_{k+1}+\gamma Q(S_{k+1},A_{k+1})-Q(S_k,A_k))+\rho^{K-k}R_{K+1} \tag{15}$$

式中：K 表示这段失败的情节一共经历的状态数，$0\leqslant k\leqslant K$；R_{K+1} 表示智能体在到达第一个不可行状态点时受到的惩罚；ρ 表示惩罚折扣因子，$0<\rho<1$，ρ 使越靠近不可行区域的动作受到的惩罚越大。

至此，得到改进 SARSA 算法的学习框架，如图 5 所示。初始化 Q 表，利用改进 SARSA 算法进行情节学习，在经历过一段成功的情节之后，将贪婪策略的贪婪因子 ε 设置为 0，用于对学习到的经验进行开采以获得最优策略，并保存这段最优策略轨迹；重新设置一个在 0～1 之间的贪婪因子用于探索，并在获得另一个成功的情节之后，重新将贪婪因子 ε 设置为 0 进行开采，获得新的最优策略轨迹；如果新获得的最优策略轨迹和保存的最优策略轨迹一样，则表明智能体可能已经遍历了所有可能的情况，并且算法收敛于最优策略。

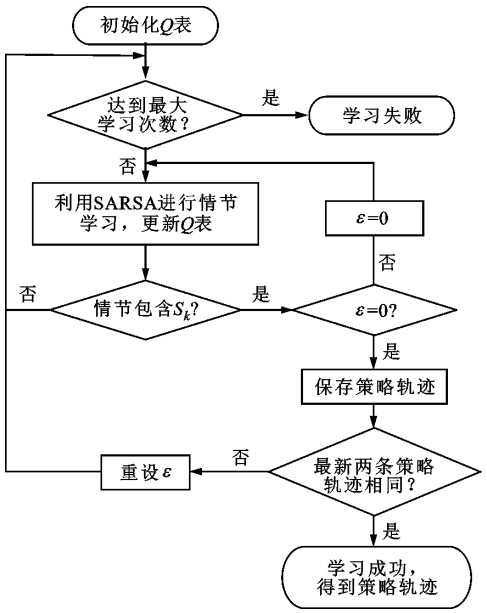


图 5 改进 SARSA 算法的框架

Fig. 5 Framework of improved SARSA algorithm

2.2　迭代交互法

根据 3.1 小节中的改进 SARSA 算法，可以学习到一条满足运动学约束的初始策略轨迹，但是由于算法并未考虑动力学约束，因此当执行这条轨迹时，电机实际测量力矩可能超限。所以，需要通过与真实环境进行迭代交互来引入动力学约束，并对初始策略轨迹进行修正，使电机实际测量力矩最终能够满足动载荷约束。时间最优轨迹优化问题的迭代交互法框架如图 6 所示，主要包括以下 3 步。

第 1 步　将时间最优轨迹优化的运动学约束条件构建成强化学习环境，将时间最优轨迹优化问题构建成强化学习模型。

第 2 步　使用改进 SARSA 算法学习得到策略轨迹。

第 3 步　在真实环境中运行获得策略轨迹，并采集真实的测量力矩，判断是否超限，如果超限则将那些超限部分的状态点都设置为不可行状态点，从而将动力学约束更新到强化学习环境中。

重复第 2、3 步，直到测量力矩不再超限，便可获得一条同时满足运动学和动力学约束的时间最优轨迹。

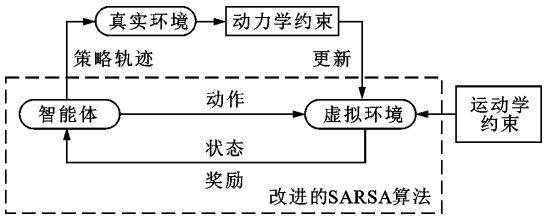


图 6　迭代交互法框架

Fig. 6　Framework of iterative interaction

3　仿真与实验

3.1　改进 SARSA 算法仿真

为了验证改进 SARSA 算法的有效性，在 MATLAB R2018b 对改进 SARSA 算法进行仿真验证。考虑到实际工况中路径的曲率多变性，本文用 NURBS 曲线插补器[20]生成一条曲率多变的星形曲线轨迹，并对其进行优化。星形曲线的参数中，次数 $k=2$，节点矢量 $\boldsymbol{U}=[0,0,0,1/9,2/9,3/9,4/9,5/9,6/9,7/9,8/9,1,1,1]$，顶点参数如表 1 所示。插补生成的轨迹如图 7 所示。

根据 1.2 小节中的网格划分方法将相平面网格化：路径标量曲率变化率设为 0.01，则星形轨迹的相平面沿 s 轴被划分为3 014段；再用均匀划分法将 $\dot{s}$ 轴方向划分为4 000段。改进 SARSA 算法的折扣因子 α 设为 0.6，衰减因子 γ 以及惩罚衰减因子 ρ 均设为 0.8，贪婪算法的贪婪因子 ε 设置为 0.6，最大学习数设为100 000。

表 1　星形曲线顶点参数

Table 1　Apex parameters of star curve

顶点序号	控制点坐标/mm	权因子
1	(0,0)	1.0
2	(48.0,24.0)	1.0
3	(48.0,64.0)	1.0
4	(96.0,32.0)	1.0
5	(144.0,40.0)	0.7
6	(108.0,0)	1.0
7	(144.0,−40.0)	0.7
8	(96.0,−32.0)	1.0
9	(48.0,−64.0)	1.0
10	(48.0,−24.0)	1.0

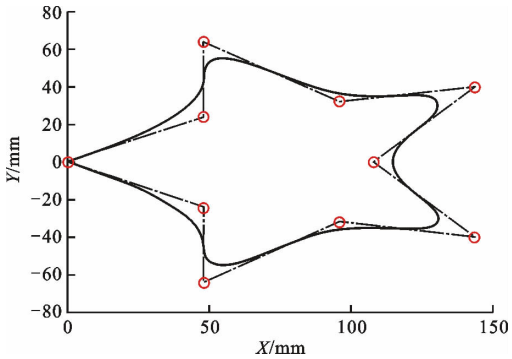


图 7　星形曲线轨迹

Fig. 7　Star curve trajectory

改进 SARSA 算法对星形轨迹进行优化得到的结果如图 8 和图 9 所示。图 8 为相平面上的优化结果。可以看出，改进 SARSA 算法优化得到的路径标量速度曲线均未超过 $\dot{s}_{\lim}$。图 9 为改进 SARSA 算法优化得到的速度和加速度曲线。可以看出，算法优化得到的速度和加速度曲线都在约束内，改进

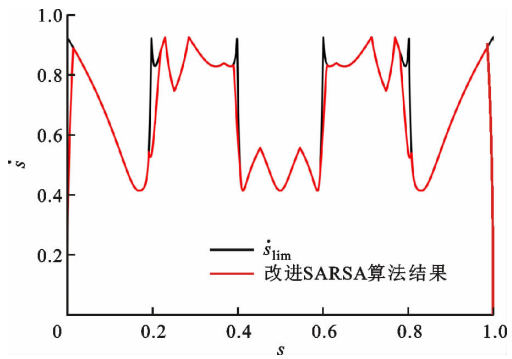
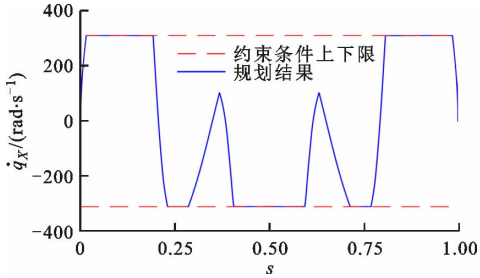


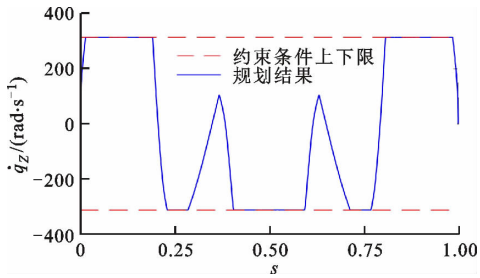
图 8　相平面轨迹优化结果

Fig. 8　Optimization results of phase plane trajectory

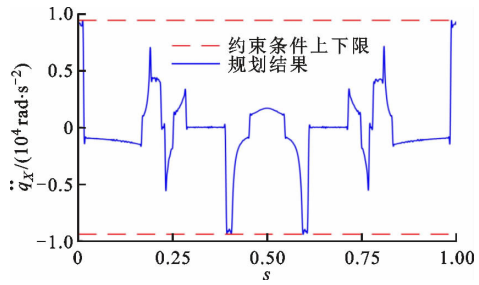
SARSA 算法优化得到的结果满足运动学约束，算法的有效性得到了验证。



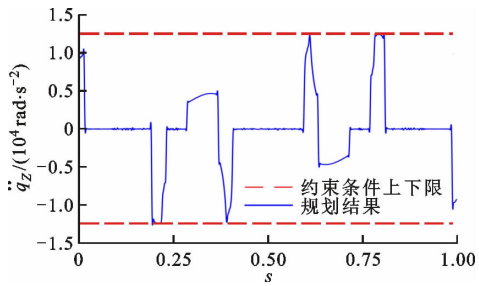
(a)X 轴速度曲线



(b)Z 轴速度曲线



(c)X 轴加速度曲线



(d)Z 轴加速度曲线

图 9 优化轨迹的速度和加速度曲线

Fig. 9 Speed and acceleration curves of optimized trajectory

3.2 迭代交互实验

时间最优轨迹优化的迭代交互实验平台如图 10 所示，平台由做独立运动的 X、Z 两轴串联组成。两轴的电机的型号均为 Delta ECMA-CA0604RS，由 Delta ASD-A2-0421-E 交流伺服驱动器进行驱动，再经过滚珠丝杠将旋转运动转换成工作台的直线运动。平台的控制系统搭建于 Windows7 64-bit 系统，并使用 EtherCAT 工业以太网进行通信，控制周期和采样周期均为 1 ms。作为主站的工控机型号为 DT-610P-ZQ170MA，CPU 为 Intel Core i7-4770，最高主频为 3.40 GHz，运行内存为 8 GB。

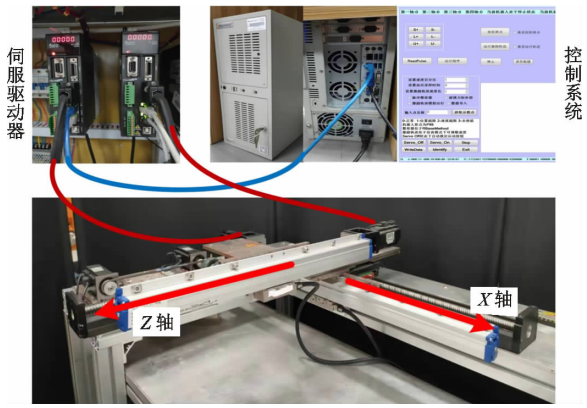


图 10 实验平台

Fig. 10 Experimental platform

利用 3.1 小节中改进 SARSA 算法优化得到的策略轨迹与真实环境进行迭代交互实验，结果如图 11 所示。可以看出，动载荷过载点的数量随着迭代的进行逐渐减少，且经过 10 次迭代之后，动载荷过载点的数量减少为 0。

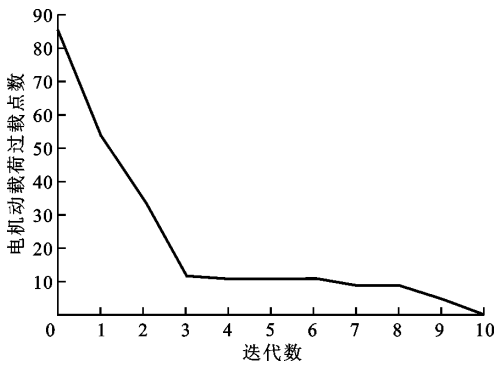
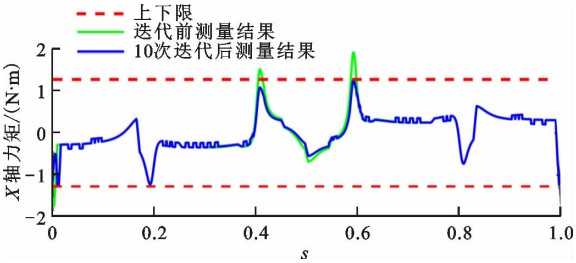


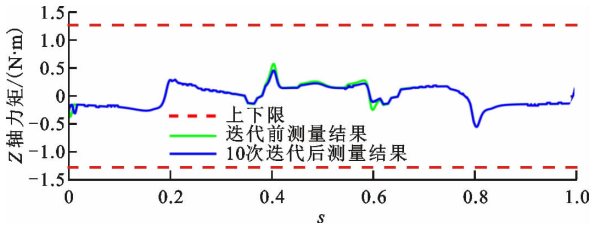
图 11 电机动载荷过载点数

Fig. 11 Motor dynamic load overload points

迭代前和经过 10 次迭代后的实际测量力矩曲线如图 12 所示。可以看出，迭代前 X 轴的部分实际测量力矩曲线超过了动载荷限制，而经过 10 次迭代后，X 轴和 Z 轴的实际力矩曲线均在动载荷限制内。这是因为改进 SARSA 算法并未考虑动力学约束，在运行迭代前的策略轨迹时，部分点的电机动载荷过载。与真实环境进行迭代交互后，动力学约束被引入，策略轨迹逐渐得到修正，最终轨迹的实际测量力矩满足动载荷限制。由此，迭代交互法的有效性得到了验证。



(a)X 轴力矩曲线



(b)Z 轴力矩曲线

图 12　迭代前后实际测量力矩曲线

Fig. 12　Actual measured torque curve before and after iteration

4　结　论

为实现多轴运动系统高速运动并解决电机动载荷过载的问题，本文将时间最优轨迹优化问题描述为马尔可夫决策问题，提出了一种采用强化学习的时间最优轨迹优化方法。该方法无需建立动力学模型，而是通过与现实环境的交互学习来直接对实际力矩进行约束，避免了动力学模型和实际模型不匹配的问题，从根本上解决了多轴运动系统在高速运动中动载荷过载的问题。

采用强化学习的时间最优轨迹优化方法根据时间最优轨迹优化问题的特性对经典 SARSA 算法进行改进，通过与基于运动学模型建立的强化学习环境进行交互学习，找到满足运动学约束的初始策略轨迹。通过迭代交互法与真实环境进行交互学习，从而将动力学约束引入到强化学习环境中并对策略轨迹进行修正。最终，获得同时满足运动学约束和动力学约束的时间最优轨迹。

实验结果显示，改进 SARSA 算法优化得到的策略轨迹的速度和加速度曲线均在约束内，且经过 10 次迭代后的轨迹实际测量力矩曲线也均在动载荷约束内，提出的采用强化学习的时间最优轨迹优化方法有效。

参考文献：

[1]　KAHN M E, ROTH B. The near-minimum-time control of open-loop articulated kinematic chains [J]. Journal of Dynamic Systems, Measurement, and Control, 1971, 93(3): 164-172.

[2]　PFEIFFER F, JOHANNI R. A concept for manipulator trajectory planning [J]. IEEE Journal on Robotics and Automation, 1987, 3(2): 115-123.

[3]　PHAM Q C. A general, fast, and robust implementation of the time-optimal path parameterization algorithm [J]. IEEE Transactions on Robotics, 2014, 30(6): 1533-1540.

[4]　SHILLER Z, LU H H. Computation of path constrained time-optimal motions with dynamic singularities [J]. Journal of Dynamic Systems, Measurement, and Control, 1992, 114(1): 34-40.

[5]　STEINHAUSER A, SWEVERS J. An efficient iterative learning approach to time-optimal path tracking for industrial robots [J]. IEEE Transactions on Industrial Informatics, 2018, 14(11): 5200-5207.

[6]　WALTZ M, FU K. A heuristic approach to reinforcement learning control systems [J]. IEEE Transactions on Automatic Control, 1965, 10(4): 390-398.

[7]　LOW E S, ONG P, CHEAH K C. Solving the optimal path planning of a mobile robot using improved Q-learning [J]. Robotics and Autonomous Systems, 2019, 115: 143-161.

[8]　KAREEM J M A, AL-ROUSAN M, QUADAN L. Reinforcement based mobile robot navigation in dynamic environment [J]. Robotics and Computer-Integrated Manufacturing, 2011, 27(1): 135-149.

[9]　KONAR A, GOSWAMI C I, SINGH S J, et al. A deterministic improved Q-learning for path planning of a mobile robot [J]. IEEE Transactions on Systems, Man, and Cybernetics: Systems, 2013, 43(5): 1141-1153.

[10]　WONG C C, LIU C C, XIAO S R, et al. Q-learning of straightforward gait pattern for humanoid robot based on automatic training platform [J]. Electronics, 2019, 8(6): 615.

[11]　ERDEN M S, LEBLEBICIOĞLU K. Free gait generation with reinforcement learning for a six-legged robot [J]. Robotics and Autonomous Systems, 2008, 56(3): 199-212.

[12]　DUGULEANA M, BARBUCEANU F G, TEIRELBAR A, et al. Obstacle avoidance of redundant manipulators using neural networks based reinforcement learning [J]. Robotics and Computer-Integrated Manufacturing, 2012, 28(2): 132-146.

[13]　SANGIOVANNI B, RENDINIELLO A, INCRE-

MONA G P, et al. Deep reinforcement learning for collision avoidance of robotic manipulators [C]// Proceedings of the 2018 European Control Conference (ECC). Piscataway, NJ, USA: IEEE, 2018: 2063-2068.

[14] INOUE T, DE MAGISTRIS G, MUNAWAR A, et al. Deep reinforcement learning for high precision assembly tasks [C]// Proceedings of the 2017 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS). Piscataway, NJ, USA: IEEE, 2017: 819-825.

[15] REN Tianyu, DONG Yunfei, WU Dan, et al. Learning-based variable compliance control for robotic assembly [J]. Journal of Mechanisms and Robotics, 2018, 10(6): 061008.

[16] SHIN K, MCKAY N. A dynamic programming approach to trajectory planning of robotic manipulators [J]. IEEE Transactions on Automatic Control, 1986, 31(6): 491-500.

[17] XIAO Yongqiang, DU Zhijiang, DONG Wei. Smooth and near time-optimal trajectory planning of industrial robots for online applications [J]. Industrial Robot: An International Journal, 2012, 39(2): 169-177.

[18] BOBROW J E, DUBOWSKY S, GIBSON J S. Time-optimal control of robotic manipulators along specified paths [J]. The International Journal of Robotics Research, 1985, 4(3): 3-17.

[19] RUMMERY G A, NIRANJAN M. On-line *Q*-learning using connectionist system [EB/OL]. [2020-10-01]. http://citeseer.ist.psu.edu/viewdoc/download;jsessionid=11751075B3B7CC74CF68CCE958731710?doi=10.1.1.17.2539&rep=rep1&type=pdf.

[20] CHENG M Y, TSAI M C, KUO J C. Real-time NURBS command generators for CNC servo controllers [J]. International Journal of Machine Tools and Manufacture, 2002, 42(7): 801-813.

[本刊相关文献链接]

张腾,张小栋,张英杰,陆竹风,朱文静,蒋永玉.引入深度强化学习思想的脑-机协作精密操控方法.2021,55(2):1-9.[doi:10.7652/xjtuxb202102001]

孙平,丁雨姗.全方向康复步行训练机器人具有死区补偿的反步有限时间控制.2020,54(7):1-8,74.[doi:10.7652/xjtuxb202007001]

张亮修,张铁柱,吴光强.考虑误差校正的智能车辆路径跟踪鲁棒预测控制.2020,54(3):20-27.[doi:10.7652/xjtuxb202003003]

颛孙少帅,杨俊安,刘辉,黄科举.未知拓扑无线自组网络多节点干扰决策算法.2018,52(6):91-97.[doi:10.7652/xjtuxb201806015]

张铁,林康宇,邹焱飚,刘晓刚.用于机器人末端残余振动控制的控制误差优化输入整形器.2018,52(4):90-97.[doi:10.7652/xjtuxb201804014]

颛孙少帅,杨俊安,刘辉,黄科举.采用双层强化学习的干扰决策算法.2018,52(2):63-69.[doi:10.7652/xjtuxb201802011]

(编辑 陶晴)