

DOI: 10.11830/ISSN.1000-5013.202403035



概率预测强化学习下非结构环境 机械臂变阻抗力跟踪控制

董梓呈¹, 胡伟石², 邵辉¹, 郭霖¹

(1. 华侨大学 信息科学与工程学院, 福建 厦门 361021;

2. 华侨大学 实验室与设备管理处, 福建 厦门 361021)

摘要: 针对非结构环境下末端实时移动机械臂阻抗控制力跟踪问题, 通过动态调节阻尼系数以应对接触环境的不确定性。为确保阻抗策略的高效搜索, 利用机械臂与接触环境交互产生状态-动作序列构建概率预测模型 (PPM)。学习过程中, 机械臂仅需与非结构接触环境进行少量交互即可获得最优变阻抗策略, 这使得该过程在真实机械臂上直接训练成为可能。仿真实验表明, 在几种非结构环境下, 所提出的方法使力跟踪动态和稳态性能均明显优于传统阻抗控制和自适应变阻抗控制。

关键词: 变阻抗控制; 机械臂力跟踪; 强化学习; 非结构环境; 概率预测模型

中图分类号: TP 273

文献标志码: A

文章编号: 1000-5013(2024)04-0461-10

Probability Prediction Reinforcement Learning for Variable Impedance Force Tracking Control of Robotic Arms in Unstructured Environments

DONG Zicheng¹, HU Weishi², SHAO Hui¹, GUO Lin¹

(1. College of Information Science and Engineering, Huaqiao University, Xiamen 361021, China;

2. Department of Laboratory and Device Management, Huaqiao University, Xiamen 361021, China)

Abstract: Aiming at the real-time impedance control force tracking problems of the end mobile robotic arm in a unstructured environment, the damping coefficient is dynamically adjusted to cope with the uncertainty of the contact environment. To ensure efficient search of the impedance strategy, a probabilistic prediction model (PPM) is constructed by utilizing the interaction between the robotic arm and the contact environment to generate state-action sequences. During the learning process, the robotic arm only needs to interact minimally with the unstructured contact environment to obtain the optimal variable impedance strategy. This makes it possible to directly train the process on a real robotic arm. Simulation results show that in several unstructured environments, the proposed method significantly outperforms the traditional impedance control and adaptive variable impedance control in both dynamic and steady-state force tracking performance.

Keywords: variable impedance control; robotic arm force tracking; reinforcement learning; unstructured environment; probability prediction model

收稿日期: 2024-03-23

通信作者: 邵辉 (1973-), 女, 副教授, 博士, 主要从事机器人运动规划与控制的研究。E-mail: shaohuihu11@163.com。

基金项目: 福建省自然科学基金资助项目 (2021J01291); 华侨大学研究生教育教学改革研究项目 (22YJG006)

机械臂已经被广泛应用于各类接触式任务,如人机协作^[1]、货物装卸^[2]、外科手术^[3]等。这些场景中,除了高精度的运动控制外,还需考虑末端接触力的跟踪控制,以保证机械臂作业效果和交互安全性。阻抗控制是一种机械臂力控制的经典方法,然而,实际中的接触环境往往是动态且未知的,经典阻抗控制缺乏适应环境变化的能力,因此,难以实现精确力控制。

一些国内外学者研究了非结构环境下的阻抗控制力跟踪方法,目前主要方法可归结为参考轨迹自适应和变阻抗控制两类。参考轨迹自适应通过辨识环境信息或直接根据接触力来预测机器人的参考轨迹。Li 等^[4]用李雅普诺夫理论对接触动力学进行分析,提出一种迭代学习控制器,调节参考轨迹使接触力保持在所需范围,控制性能优于传统阻抗控制,但所需迭代次数较多。刘胜遂等^[5]提出基于卡尔曼滤波的自适应阻抗控制方法,对机械臂接触环境的位置和刚度进行估计,但仍存在一定力跟踪误差。李振等^[6]在基于环境参数估计自适应生成参考轨迹的方法上,采用遗传算法补偿接触力误差,提高了接触力跟踪精度。Roveda 等^[7]关注阻抗控制接触力过冲的问题,采用扩展卡尔曼滤波对环境刚度进行连续自适应估计,避免接触过程的力超调和不稳定,但该方法的响应速度较慢且跟踪精度有限。此类方法依赖于环境信息的辨识精度,对辨识误差力控精度有较大影响。变阻抗控制是一种更简单有效的自适应力控制方法,对环境特性的估计误差不敏感,关键在于设计控制性能良好而通用的变阻抗策略以应对复杂的接触环境。Jung 等^[8]和 Duan 等^[9]提出的自适应变阻抗控制算法具有等价的形式,根据机械臂末端接触力实时调节阻尼系数,能够在未知刚度和几何形状的曲面上实现力跟踪,但该方法的跟踪精度受限于采样频率和初始阻抗参数,在控制器和力传感器的采样频率足够高时,才能获得较好的控制效果。Cao 等^[10]对该自适应变阻抗方法进行改进,提出一种自适应更新率策略,但力控精度提升有限。Hamedani 等^[11]提出了基于小波神经网络的智能变阻抗算法来自动调节阻尼系数,但这种方法在斜面和复杂曲面上的力跟踪精度不高,且动态性能不佳。此类变阻抗方法难以较好地平衡力跟踪动态性能和稳态误差,综合控制性能仍存在提升空间。

人工智能的快速发展为机械臂控制问题提供了新思路,例如,利用强化学习,机械臂能够通过试错的方式优化自身行为,而不需要本体和环境的先验信息^[12-13]。Buchli 等^[14]提出一种基于策略函数的强化学习算法 PI^2 ,将此方法运用于机器人的自适应阻抗控制中,并证明其最优性。Li 等^[15]提出一种强化学习变阻抗方法,通过仿真和实验证明机器人与环境只需少量交互即可成功学习出力控制策略。Wu 等^[16]研究了人机协作最优阻抗问题,用 Q-Learning 设计自适应阻抗控制律,使机器人能够根据接触力在线估测人的示教轨迹,实现人机平顺交互。Du 等^[17]将虚拟阻尼项引入传统阻抗控制中,使用模糊强化学习对虚拟阻尼进行调整,提升了手术机器人的力跟踪性能,并保证能量消耗最优。

然而,目前大多数基于强化学习的变阻抗方法主要关注任务本身而忽略了数据效率,机械臂需与环境进行大量交互以采集足量的训练样本,这在实际机械臂系统中存在安全问题,且交互过程通常非常耗时,因此,数据效率低下成为严重限制强化学习在实际机器人系统中应用的主要原因之一^[18-20]。基于此,本文提出一种概率预测强化学习下非结构环境机械臂变阻抗力跟踪控制 (PPM-VIC) 方法。

1 问题描述

笛卡尔空间中,阻抗控制利用质量-弹簧-阻尼模型维持机械臂运动状态与外力之间的动态关系,使机械臂末端呈现期望的柔顺性。基于位置的阻抗控制,如图 1 所示。图 1 中: F_d, F_e 分别表示期望力和实际接触力, $F_d, F_e \in R^k, k$ 为受力数; X_r, X_d 分别表示参考轨迹和期望轨迹, $X_r, X_d \in R^n$, 在位置控制精度足够高的情况下可近似认为机械臂末端实际轨迹与期望轨迹相等,即 $X = X_d$ 。阻抗模型将力跟踪误差转化为运动补偿量,与参考轨迹叠加后得到期望轨迹,机械臂末端跟踪期望轨迹可实现力跟踪。

对于 n 自由度的机械臂系统,阻抗控制的

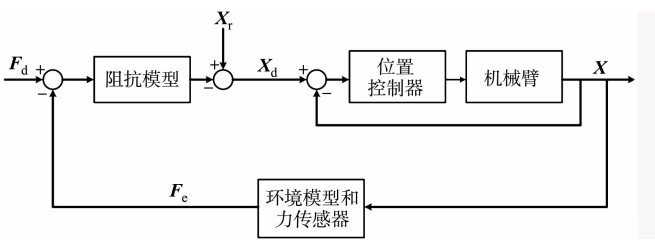


图 1 基于位置的阻抗控制

Fig. 1 Position based impedance control

一般形式可表示为

$$\mathbf{M}(\ddot{\mathbf{X}}_r - \ddot{\mathbf{X}}_d) + \mathbf{B}(\dot{\mathbf{X}}_r - \dot{\mathbf{X}}_d) + \mathbf{K}(\mathbf{X}_r - \mathbf{X}_d) = \mathbf{F}_d - \mathbf{F}_e. \quad (1)$$

式(1)中: $\mathbf{M}, \mathbf{B}, \mathbf{K}$ 分别为质量、阻尼和刚度矩阵, 它们直接决定了阻抗模型的动力学特性, $\mathbf{M}, \mathbf{B}, \mathbf{K} \in \mathbf{R}^{n \times n}$ 。

纯刚性接触环境 \mathbf{F}_e 定义为

$$\mathbf{F}_e = \begin{cases} \mathbf{K}_e(\mathbf{X}_e - \mathbf{X}), & \mathbf{X}_e \geq \mathbf{X}, \\ \mathbf{0}, & \mathbf{X}_e < \mathbf{X}. \end{cases} \quad (2)$$

式(2)中: \mathbf{K}_e 为环境刚度, $\mathbf{K}_e \in \mathbf{R}^{k \times n}$; \mathbf{X}_e 为环境位置, $\mathbf{X}_e \in \mathbf{R}^n$; $\mathbf{X}_e \geq \mathbf{X}$ 为机器人末端处于接触状态。

为简化分析, 假设阻抗模型在笛卡尔空间各方向上是解耦的, 以一维力跟踪为例, 设 $f_e, f_d, m, b, k, k_e, x_e$ 分别表示 $\mathbf{F}_e, \mathbf{F}_d, \mathbf{M}, \mathbf{B}, \mathbf{K}, \mathbf{K}_e, \mathbf{X}_e$ 中的元素。

根据文献[9, 11]的分析, 若环境刚度未知或时变, 可设力控方向的刚度为零, 以实现力跟踪无偏差, 故设 $k=0$ 。非结构环境中, 环境位置 x_e 通常难以精确获取, 因此, 可用常值估计量 \hat{x}_e 代替, 假设估计误差 $\delta x_e = \hat{x}_e - x_e$ 。令 $e = x_e - x_d = x_e - x$, 则 $\hat{e} = e + \delta x_e$, 用 \hat{e} 代替式(1)中的偏差项 e , 有

$$\Delta f = f_d - f_e = m\ddot{\hat{e}} + b\dot{\hat{e}} = m(\ddot{e} + \delta\ddot{x}_e) + b(\dot{e} + \delta\dot{x}_e) = m(\ddot{x}_e - \ddot{x} + \delta\ddot{x}_e) + b(\dot{x}_e - \dot{x} + \delta\dot{x}_e). \quad (3)$$

机械臂末端执行器在未知几何形状的接触面上实时移动时, 末端实际轨迹、真实环境轨迹和估计误差可能随时间连续变化, 即 $x, \dot{x}, \ddot{x}, x_e, \dot{x}_e, \ddot{x}_e, \delta x_e, \delta \dot{x}_e, \delta \ddot{x}_e$ 时变, 因此, 式(3)的跟踪误差 Δf 将始终存在。自适应阻抗参数可以补偿环境位置时变产生的跟踪误差, 而质量系数的变化容易引起系统震荡^[9]。

2 强化学习变阻抗策略

最优变阻抗策略 $\pi^*(s_t)$ 在任意时刻状态 s_t 满足跟踪误差 $\Delta f_e = 0$ 。无模型强化学习(如深度确定性策略梯度算法(DDPG)、近端策略优化算法(PPO)等)通常要求智能体与环境进行大量交互以收集足够的训练样本, 但过多的交互可能会对环境和机器人造成潜在的损伤, 在实际机器人应用中并不可取。强化学习可根据交互数据建立接触状态转移模型 $p(s_{t+1} | s_t)$, 从而显著提高数据利用效率。

为消除未知动态环境引起的力跟踪误差, 引入高斯过程建立接触状态转移概率模型, 借助该模型近似预测未来一段时间内的状态分布, 并采用价值函数 $V^\pi(s)$ 进行评估, 最后使用 BFGS(broyden-fletcher-goldfarb-shanno)算法更新参数, 以逐步逼近最优变阻抗策略。

2.1 策略学习框架

考虑机械臂移动方向和受力方向, 设连续状态 $s = [p_y, p_z, f_z, \Delta f_z]^T$, 其中, p_y, p_z 分别表示世界坐标系下机械臂末端位置在 y 和 z 方向的对应分量, f_z 为力控方向上的实际接触力, Δf_z 为力跟踪误差, 自适应调整量 u 为力控方向的阻尼系数。定义自适应阻抗策略 $\pi(s_t, \theta)$, θ 为待学习的策略参数。自适应阻抗策略由两部分构成。

1) 径向基(RBF)神经网络, 将状态映射到初始控制量 $u' = \pi'(s)$ 。

2) 饱和函数表达式为 $u = \text{Sat}(u')$, 将控制量限制在合理范围内。

RBF 神经网络等价于 N 个高斯核的线性组合, 即

$$\pi'(s_t) = \sum_{i=1}^N \beta_i k_\pi(c_i, s_t) = \beta_\pi^T k_\pi(C_\pi, s_t); \quad (4)$$

$$k_\pi(c_i, s_t) = \exp\left(-\frac{1}{2}(c_i - s_t)^T \Lambda_\pi^{-1}(c_i - s_t)\right). \quad (5)$$

式(4), (5)中: $\beta_\pi = (\mathbf{K}_\pi + \sigma_\pi^2 \mathbf{I})^{-1} \mathbf{y}_\pi$ 表示加权向量, \mathbf{K}_π 是由元素 $k_\pi(c_i, c_j)$, $i, j = 1, 2, \dots, N$ 构成的 Gram 矩阵, \mathbf{I} 为单位矩阵, \mathbf{y}_π 为训练目标, $\mathbf{y}_\pi = \pi'(C_\pi) + \boldsymbol{\eta}_\pi$, $\boldsymbol{\eta}_\pi \sim N(0, \sigma_\pi^2 \mathbf{I})$ 表示测量噪声, σ_π^2 代表噪声方差; Λ_π 为加权欧式权重矩阵; C_π 表示核函数的中心点, $C_\pi = [c_1, c_2, \dots, c_N]^T$ 。

令 $g(u') = [9\sin(u') + \sin(3u')]/8 \in [-1, 1]$, 饱和函数 $\text{Sat}(u')$ 把控制量限制在 u_{\max} 与 u_{\min} 之间, 其表达式为

$$\text{Sat}(u') = \frac{1}{2}(u_{\max} - u_{\min})g(u') + \frac{1}{2}(u_{\max} + u_{\min}). \quad (6)$$

代价函数设计为有界的形式,设目标状态 $\mathbf{s}_{\text{tar}} = [0, 0, f_d, 0]^T$, \mathbf{A}_L 为对角权重矩阵,与位置相关的元素为零,常数 λ 用于控制代价函数达到饱和时的状态偏差度。

代价函数 $L(\mathbf{s}_t) \in [0, 1]$ 为

$$L(\mathbf{s}_t) = 1 - \exp\left(-\frac{1}{2\lambda^2}(\mathbf{s}_t - \mathbf{s}_{\text{tar}})^T \mathbf{A}_L (\mathbf{s}_t - \mathbf{s}_{\text{tar}})\right). \quad (7)$$

2.2 接触状态概率预测模型

假设连续状态 $\mathbf{s} \in \mathbf{R}^E$ 、连续控制量 $u \in \mathbf{R}^1$ 及随机噪声 $\boldsymbol{\omega}$ 均服从高斯分布,则预测模型为高斯过程,即 $h \sim \text{GP}(m(\cdot), k(\cdot, \cdot))$ 。接触状态概率预测模型表达式为

$$\mathbf{s}_t = h(\mathbf{s}_{t-1}, u_{t-1}, \boldsymbol{\omega}). \quad (8)$$

机械臂在特定变阻抗策略作用下沿接触环境表面的运动过程中,以固定频率对数据采样,构成预测模型的训练输入 $\mathbf{X} = [\tilde{\mathbf{s}}_1, \tilde{\mathbf{s}}_2, \dots, \tilde{\mathbf{s}}_n]^T$ 及训练目标 $\mathbf{Y} = [\mathbf{\Delta}_1, \mathbf{\Delta}_2, \dots, \mathbf{\Delta}_n]^T$, 其中, $\tilde{\mathbf{s}}_t = (\mathbf{s}_t, u_t) \in \mathbf{R}^{E+1}$ 表示状态-动作二元组, $\mathbf{\Delta}_t = \mathbf{s}_{t+1} - \mathbf{s}_t \in \mathbf{R}^E$ 为相邻时刻的状态变化量。

协方差函数 $k(\cdot, \cdot)$ 与式(5)有相似的形式,即

$$k(\tilde{\mathbf{s}}, \tilde{\mathbf{s}}') = \sigma_f^2 \exp\left(-\frac{1}{2}(\tilde{\mathbf{s}} - \tilde{\mathbf{s}}')^T \mathbf{A}^{-1}(\tilde{\mathbf{s}} - \tilde{\mathbf{s}}')\right) + \delta \sigma_\omega^2. \quad (9)$$

式(9)中: δ 在 $\tilde{\mathbf{s}}$ 与 $\tilde{\mathbf{s}}'$ 相等时为 1, 否则为 0; $\mathbf{A} = \text{diag}(l_1^2, l_2^2, \dots, l_E^2)$ 是由尺度 l 组成的权重矩阵,与信号方差 σ_f^2 、噪声方差 σ_ω^2 共同构成预测模型的超参数(利用第二类最大似然估计^[19]获取)。

每个独立的预测模型分别对应每一维状态分量。由于高斯分布经非线性映射通常会变成非高斯分布,对于任一输入 $\tilde{\mathbf{s}}_{t-1} \sim N(\boldsymbol{\mu}_{\tilde{\mathbf{s}}_{t-1}}, \boldsymbol{\Sigma}_{\tilde{\mathbf{s}}_{t-1}}) \in \mathbf{R}^{E+1}$, 利用矩匹配法近似预测输出 $\mathbf{\Delta}_t \sim N(\boldsymbol{\mu}_{\mathbf{\Delta}_t}, \boldsymbol{\Sigma}_{\mathbf{\Delta}_t}) \in \mathbf{R}^E$, 故预测均值为

$$\boldsymbol{\mu}_{\mathbf{\Delta}_t} = [\boldsymbol{\beta}_1^T \mathbf{q}_1, \dots, \boldsymbol{\beta}_E^T \mathbf{q}_E]^T. \quad (10)$$

式(10)中: $\boldsymbol{\beta}_a = (\mathbf{K}_a + \sigma_{\omega_a}^2 \mathbf{I})^{-1} \mathbf{y}_a, a \in [1, 2, \dots, E]$, $\mathbf{K}_a, \sigma_{\omega_a}, \mathbf{y}_a$ 分别表示第 a 个预测模型的 Gram 矩阵、噪声方差及训练目标;向量 $\mathbf{q}_a = [q_{a_1}, q_{a_2}, \dots, q_{a_n}]^T \in \mathbf{R}^n$,

$$q_{a_i} = \frac{\sigma_{f_a}^2 \exp\left(-\frac{1}{2}(\tilde{\mathbf{s}}_i - \boldsymbol{\mu}_{\tilde{\mathbf{s}}_{t-1}})^T (\boldsymbol{\Sigma}_{\tilde{\mathbf{s}}_{t-1}} + \mathbf{A}_a)^{-1}(\tilde{\mathbf{s}}_i - \boldsymbol{\mu}_{\tilde{\mathbf{s}}_{t-1}})\right)}{\sqrt{|\boldsymbol{\Sigma}_{\tilde{\mathbf{s}}_{t-1}} \mathbf{A}_a^{-1} + \mathbf{I}|}}. \quad (11)$$

式(11)中: $\sigma_{f_a}, \mathbf{A}_a$ 分别为对应预测模型的信号方差和权重矩阵。

预测协方差 $(\boldsymbol{\Sigma}_{\mathbf{\Delta}_t})$ 为

$$\boldsymbol{\Sigma}_{\mathbf{\Delta}_t} = \begin{bmatrix} \text{var}[h_1(\tilde{\mathbf{s}}_{t-1})] & \dots & \text{cov}[h_1(\tilde{\mathbf{s}}_{t-1}), h_E(\tilde{\mathbf{s}}_{t-1})] \\ \vdots & & \vdots \\ \text{cov}[h_E(\tilde{\mathbf{s}}_{t-1}), h_1(\tilde{\mathbf{s}}_{t-1})] & \dots & \text{var}[h_E(\tilde{\mathbf{s}}_{t-1})] \end{bmatrix}. \quad (12)$$

式(12)中: 对角线元素 $\text{var}[h_a(\tilde{\mathbf{s}}_{t-1})]$ 为第 a 个预测模型对 $\tilde{\mathbf{s}}_t$ 的预测方差,非对角线元素 $\text{cov}[h_a(\tilde{\mathbf{s}}_{t-1}), h_b(\tilde{\mathbf{s}}_{t-1})], b \in [1, 2, \dots, E]$ 为不同预测模型对同一输入 $\tilde{\mathbf{s}}_t$ 的预测混合协方差。

预测协方差各元素为

$$\left. \begin{aligned} &\sigma_{f_a}^2 - \text{tr}[(\mathbf{K}_a + \sigma_{\omega_a}^2 \mathbf{I})^{-1} \mathbf{Q}] + \boldsymbol{\beta}_a^T \mathbf{Q} \boldsymbol{\beta}_a - (\mu_{\mathbf{\Delta}_t}^a)^2, & a=b, \\ &\boldsymbol{\beta}_a^T \mathbf{Q} \boldsymbol{\beta}_b - \mu_{\mathbf{\Delta}_t}^a \mu_{\mathbf{\Delta}_t}^b, & a \neq b. \end{aligned} \right\} \quad (13)$$

令 $\mathbf{P} = \boldsymbol{\Sigma}_{\tilde{\mathbf{s}}_{t-1}} (\mathbf{A}_a^{-1} + \mathbf{A}_b^{-1}) + \mathbf{I}, \boldsymbol{\rho}_i = \tilde{\mathbf{s}}_i - \boldsymbol{\mu}_{\tilde{\mathbf{s}}_{t-1}}, \boldsymbol{\rho}_j = \tilde{\mathbf{s}}_j - \boldsymbol{\mu}_{\tilde{\mathbf{s}}_{t-1}}, \mathbf{z}_{i,j} = \mathbf{A}_a^{-1} \boldsymbol{\rho}_i + \mathbf{A}_b^{-1} \boldsymbol{\rho}_j, i, j \in [1, 2, \dots, n]$ 。

矩阵 $\mathbf{Q} \in \mathbf{R}^{n \times n}$ 的元素为

$$Q_{i,j} = \frac{\sigma_{f_a} \sigma_{f_b}}{\sqrt{|\mathbf{P}|} \exp\left\{\frac{1}{2}[\boldsymbol{\rho}_i^T \mathbf{A}_a^{-1} \boldsymbol{\rho}_i + \boldsymbol{\rho}_j^T \mathbf{A}_b^{-1} \boldsymbol{\rho}_j - \mathbf{z}_{i,j}^T \mathbf{P}^{-1} \mathbf{z}_{i,j}]\right\}}. \quad (14)$$

2.3 状态预测及策略评估

相邻时刻的状态概率分布为

$$p(s_{t-1}) \xrightarrow{\text{RBF}} p(u'_{t-1}) \rightarrow p(u_{t-1}) \rightarrow p(\tilde{s}'_{t-1}) \rightarrow p(\tilde{s}_{t-1}) \xrightarrow{\text{GP}} p(\Delta_t) \rightarrow p(s_t). \quad (15)$$

假设前一时刻的状态概率分布 $p(s_{t-1})$ 已知, 可得出初始控制量概率分布 $p(u'_{t-1})$, 其均值和协方差分别为

$$\left. \begin{aligned} \mu_{u'_{t-1}} &= \beta_\pi^\top q_\pi, \\ \Sigma_{u'_{t-1}} &= \beta_\pi^\top Q_\pi \beta_\pi - (\beta_\pi^\top q_\pi)^2. \end{aligned} \right\} \quad (16)$$

根据正弦函数期望和方差的性质, 容易计算限幅后的控制量概率分布 $p(u_{t-1})$, 继而初始联合概率分布 $p(s_{t-1}, u'_{t-1}) = p(\tilde{s}'_{t-1})$, $p(\tilde{s}'_{t-1})$ 计算式为

$$p(\tilde{s}'_{t-1}) = N \left[\begin{bmatrix} \mu_{s_{t-1}} \\ \mu_{u'_{t-1}} \end{bmatrix}, \begin{bmatrix} \Sigma_{s_{t-1}} & \Sigma_{s_{t-1}, u'_{t-1}} \\ \Sigma_{s_{t-1}, u'_{t-1}}^\top & \Sigma_{u'_{t-1}} \end{bmatrix} \right]. \quad (17)$$

非对角线元素 $(\Sigma_{s_{t-1}, u'_{t-1}})$ 的计算式为

$$\Sigma_{s_{t-1}, u'_{t-1}} = \sum_{i=1}^N \beta_{\pi_i} q_{\pi_i} \Sigma_{s_{t-1}} (\Sigma_{s_{t-1}} + \Lambda_\pi)^{-1} (s_i - \mu_{s_{t-1}}). \quad (18)$$

利用正弦函数期望和方差的性质, 可以得到联合概率分布 $p(\tilde{s}_{t-1})$, 根据当前预测模型及矩匹配法, 可预测状态变化量的概率分布 $p(\Delta_t)$, 考虑到 $\Delta_t = f(s_{t-1}, u_{t-1}, \omega) - s_{t-1}$, $p(s_t)$ 计算式为

$$p(s_t) = \begin{cases} \mu_{s_t} = \mu_{s_{t-1}} + \mu_{\Delta_{t-1}}, \\ \Sigma_{s_t} = \Sigma_{s_{t-1}} + \Sigma_{s_{t-1}, \Delta_t} + \Sigma_{s_{t-1}, \Delta_t}^\top + \Sigma_{\Delta_t}. \end{cases} \quad (19)$$

式(19)中: $\mu_{s_{t-1}}$ 和 $\Sigma_{s_{t-1}}$ 分别为上一时刻的状态分布; $\mu_{\Delta_{t-1}}$ 和 Σ_{Δ_t} 分别为状态变化量的预测分布; $\Sigma_{s_{t-1}, \Delta_t}$ 分别为交叉协方差项。

重复式(15), 得到虚拟状态序列 $[s_0, s_1, \dots, s_H]$, 以此实现策略评估, 序列的价值函数 $(V^\pi(s_0))$ 为

$$V^\pi(s_0) = \sum_{t=0}^H E[L(s_t)] = \sum_{t=0}^H \int L(s_t) p(s_t) ds_t. \quad (20)$$

2.4 策略参数更新

待学习的策略参数 $\theta = [C_\pi, y_\pi, \Lambda_\pi, \sigma_\pi^2]$ 。最优变阻抗策略为

$$\pi^*(s, \theta^*) = \arg \min_{\theta} V^\pi(s_0). \quad (21)$$

为保证价值函数最小, 需计算策略参数的梯度, 即

$$\frac{dV^\pi(s_0)}{d\theta} = \sum_{t=1}^H \frac{d}{d\theta} E[L(s_t)]. \quad (22)$$

代价函数 $L(s_t)$ 依赖状态概率分布 $p(s_t) \sim N(\mu_{s_t}, \Sigma_{s_t})$, 利用链式法则, 有

$$\frac{dE[L(s_t)]}{d\theta} = \frac{\partial E[L(s_t)]}{\partial \mu_{s_t}} \cdot \frac{d\mu_{s_t}}{d\theta} + \frac{\partial E[L(s_t)]}{\partial \Sigma_{s_t}} \cdot \frac{d\Sigma_{s_t}}{d\theta}. \quad (23)$$

令 $\Psi = \Lambda_L (I + \Sigma_{s_t} \Lambda_L)^{-1}$, 由式(7), 期望 $E[L(s_t)]$ 为

$$E[L(s_t)] = \int L(s_t) p(s_t) ds_t = 1 - \frac{\exp \left[-\frac{1}{2} (\mu_{s_t} - s_{\text{tar}})^\top \Psi (\mu_{s_t} - s_{\text{tar}}) \right]}{\sqrt{|I + \Sigma_{s_t} \Lambda_L|}}. \quad (24)$$

则偏导数为

$$\frac{\partial E[L(s_t)]}{\partial \mu_{s_t}} = -E[L(s_t)] (\mu_{s_t} - s_{\text{tar}})^\top \Psi, \quad (25)$$

$$\frac{\partial E[L(s_t)]}{\partial \Sigma_{s_t}} = \frac{1}{2} E[L(s_t)] [\Psi (\mu_{s_t} - s_{\text{tar}}) (\mu_{s_t} - s_{\text{tar}})^\top - I] \Psi. \quad (26)$$

当前时刻的状态概率分布 $p(s_t)$ 由前一时刻的状态概率分布 $p(s_{t-1})$ 通过策略 $\pi(s_{t-1}, \theta)$ 及高斯过程模型 $h(\cdot)$ 预测得到。因此, 再次利用链式法则, 有

$$\frac{d\mu_{s_t}}{d\theta} = \frac{\partial \mu_{s_t}}{\partial \mu_{s_{t-1}}} \cdot \frac{d\mu_{s_{t-1}}}{d\theta} + \frac{\partial \mu_{s_t}}{\partial \Sigma_{s_{t-1}}} \cdot \frac{d\Sigma_{s_{t-1}}}{d\theta} + \frac{\partial \mu_{s_t}}{\partial \theta}, \quad (27)$$

$$\frac{d\Sigma_{s_t}}{d\theta} = \frac{\partial \Sigma_{s_t}}{\partial \mu_{s_{t-1}}} \cdot \frac{d\mu_{s_{t-1}}}{d\theta} + \frac{\partial \Sigma_{s_t}}{\partial \Sigma_{s_{t-1}}} \cdot \frac{d\Sigma_{s_{t-1}}}{d\theta} + \frac{\partial \Sigma_{s_t}}{\partial \theta}. \quad (28)$$

显然,这是一个迭代计算的过程, $\frac{d\boldsymbol{\mu}_{s_{t-1}}}{d\boldsymbol{\theta}}$ 和 $\frac{d\boldsymbol{\Sigma}_{s_{t-1}}}{d\boldsymbol{\theta}}$ 由前次计算中得出,利用链式法则,有

$$\frac{\partial \boldsymbol{\mu}_{s_t}}{\partial \boldsymbol{\theta}} = \frac{\partial \boldsymbol{\mu}_{\Delta_t}}{\partial \boldsymbol{\mu}_{u_{t-1}}} \cdot \frac{\partial \boldsymbol{\mu}_{u_{t-1}}}{\partial \boldsymbol{\theta}} + \frac{\partial \boldsymbol{\mu}_{\Delta_t}}{\partial \boldsymbol{\Sigma}_{u_{t-1}}} \cdot \frac{\partial \boldsymbol{\Sigma}_{u_{t-1}}}{\partial \boldsymbol{\theta}}, \tag{29}$$

$$\frac{\partial \boldsymbol{\Sigma}_{s_t}}{\partial \boldsymbol{\theta}} = \frac{\partial \boldsymbol{\Sigma}_{\Delta_t}}{\partial \boldsymbol{\mu}_{u_{t-1}}} \cdot \frac{\partial \boldsymbol{\mu}_{u_{t-1}}}{\partial \boldsymbol{\theta}} + \frac{\partial \boldsymbol{\Sigma}_{\Delta_t}}{\partial \boldsymbol{\Sigma}_{u_{t-1}}} \cdot \frac{\partial \boldsymbol{\Sigma}_{u_{t-1}}}{\partial \boldsymbol{\theta}}. \tag{30}$$

由价值函数算出策略参数的梯度,使用 BFGS 算法更新策略参数,当 $V^\pi(\mathbf{s}_0)$ 趋于零时,训练收敛。

3 仿真实验及分析

仿真实验基于 MATLAB/Simulink 设计,用 Robotic Toolbox 搭建 PUMA560 机械臂模型,期望充分体现机械臂动力学特性。PUMA560 型机械臂可视化模型,如图 2 所示。轨迹生成和接触环境模型通过 S-Function 实现,机械臂位置内环可达较高控制精度,满足验证要求。仿真和策略训练过程在搭载 Core i7-10700 型工作站中完成,无 GPU 加速。

3.1 训练设置

策略网络模型,如图 3 所示。输入层由当前状态 s_t 构成,隐藏层神经元个数 N 根据实际情况而定,其中的高斯核函数对输入信息进行空间映射变换,输出层对隐藏层神经元的信息进行线性加权求和,得到初始控制量 u'_t ,经连续可微的饱和函数 Sat 限幅到合理的范围内,得到最终控制量 u_t 。训练时基于 BFGS(broyden-fletcher-goldfarb-shanno)算法更新策略。

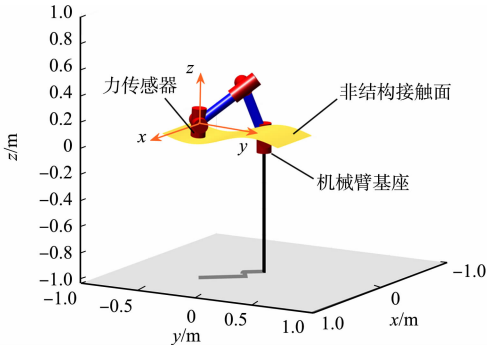


图 2 PUMA560 型机械臂可视化模型

Fig. 2 PUMA560 type manipulator visualization model

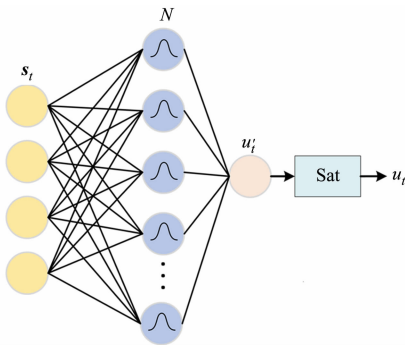


图 3 策略网络模型

Fig. 3 Policy network model

3.2 训练过程

假设接触环境刚度 $k_e=5\,000\text{ N}\cdot\text{m}^{-1}$,环境为余弦曲面(图 4),力控方向(z 方向)的期望力 $f_d=10\text{ N}$,在机械臂运动过程中,保持末端姿态不变。选择合适的质量系数 $m=0.2\text{ kg}$ 和刚度系数 $k=0\text{ N}\cdot\text{m}^{-1}$,阻尼系数由策略网络动态调整。机械臂末端在 y 方向上运动速度为 $0.16\text{ m}\cdot\text{s}^{-1}$, x 方向位置保持不变,机械臂从接触面的起点运动至终点需 6 s 。

为减少训练时间,将 Simulink 仿真步长固定为 0.005 s ,决策频率为 0.05 s ,采样频率 0.05 s ,阻尼为 $0.01\sim150.00$,预测时间域为 120 ,隐藏层神经元数 N 为 200 ,代价函数饱和系数 λ 为 5 。

在每一次训练迭代中,机械臂在当前阻抗策略(第 1 次迭代使用随机策略)的作用下从接触面的起点运动到终点,同时,以特定频率状态和控制量进行采样。完成一次交互后,采样的数据用于估计高斯过程预测模型的超参数。机械臂根据当前策略与该预测模型进行虚拟交互,产生虚拟状态-动作序列,并以此虚拟数据计算价值函数。最后,计算价值函数的梯度,更新策略的参数。随着迭代次数的增加,用于训练

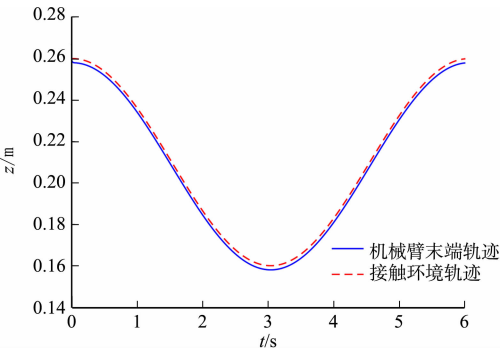


图 4 用于训练的接触环境

Fig. 4 Contact environment for training

预测模型的数据集不断扩充,模型趋于准确,预测不确定性趋于降低。

训练过程中的代价,如图 5 所示。图 5 中: L 为价值;蓝色曲线是机械臂与预测模型进行虚拟交互时的预测代价,其宽度表示预测过程的不确定性;红色曲线为机械臂与真实环境交互的实际代价,直接反应了力控制效果。

由图 5 可知:在训练初期,由于数据集较小,高斯过程模型的预测是不准确的,方差很大,随着迭代次数的增多,预测模型趋于准确,不确定性变得很小;最终,预测代价与实际代价都趋于零,机械臂获得最优变阻抗策略 $\pi^*(s, \theta^*)$ 。

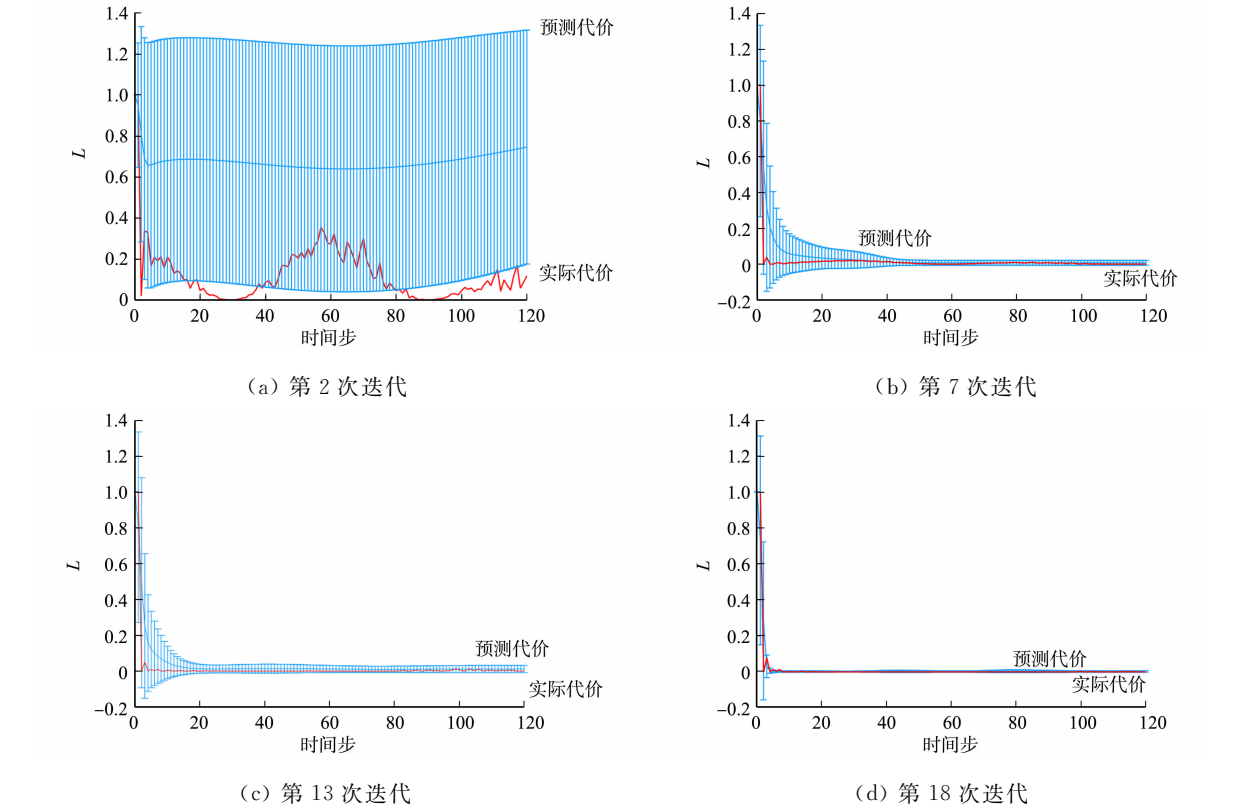


图 5 训练过程中的代价
Fig. 5 Cost during training process

图 6 为训练结果。对比训练结果与传统阻抗控制、自适应变阻抗控制(初始阻尼系数设为 $20\text{ N}\cdot\text{m}^{-1}$,更新率为 0.02)^[9]可知:参考轨迹不准确导致传统阻抗控制在非结构环境中无法实现恒力跟踪,接触力稳态误差随环境位置的变化而呈现周期性变化,最大稳态跟踪误差为 1.6 N ;相对而言,自适应变阻抗控制稳态精度更高,但动态过程较差,其稳态跟踪误差与初始阻尼系数、更新率及采样频率有关,更高精度的力跟踪参数易导致更差的动态过程^[8],在此场景下最大稳态误差约为 0.25 N 。因此,提出的 PPM-VIC 方法具有很小的超调和更高的稳态跟踪精度。

3.3 接触环境的对比测试

为了验证训练的变阻抗策略是否适用于其他类型的接触环境,设计斜面环境恒力跟踪、复杂曲面环境恒力跟踪和复杂曲面环境变力跟踪 3 种非结构环境任务场景,初始环境刚度均为 $k_e=5\,000\text{ N}\cdot\text{m}^{-1}$ 。对机械臂而言,环境信息未知。

设置机械臂的作业环境为斜率未知的斜面,则机械臂末端实时移动过程中 \dot{x}_e 为非零常值, $\ddot{x}_e=0$ 。斜面环境恒力跟踪,如图 7 所示。

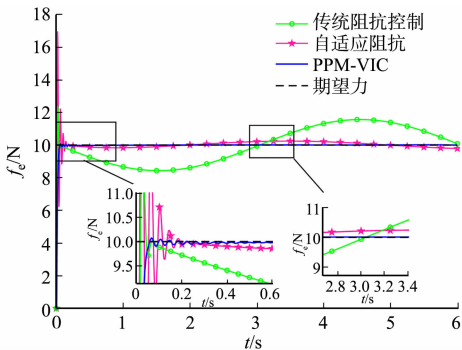


图 6 训练结果
Fig. 6 Training result

由图 7(b)可知:在斜面环境下,传统阻抗控制在刚度系数 $k=0$ 时始终存在恒定的稳态误差;自适应阻抗控制在接触初期会产生较大的超调,需要约 0.7 s 才能使接触力稳定至期望值,动态性能较差,但稳态时可实现高精度力跟踪;PPM-VIC 方法在刚发生接触时存在微小抖震,但超调量明显小于另外两种控制方式,稳定后跟踪精度优于自适应变阻抗。接触环境刚度突变时,3 种控制方法都表现出不同程度的超调和震荡,但 PPM-VIC 方法表现出更优的控制效果。

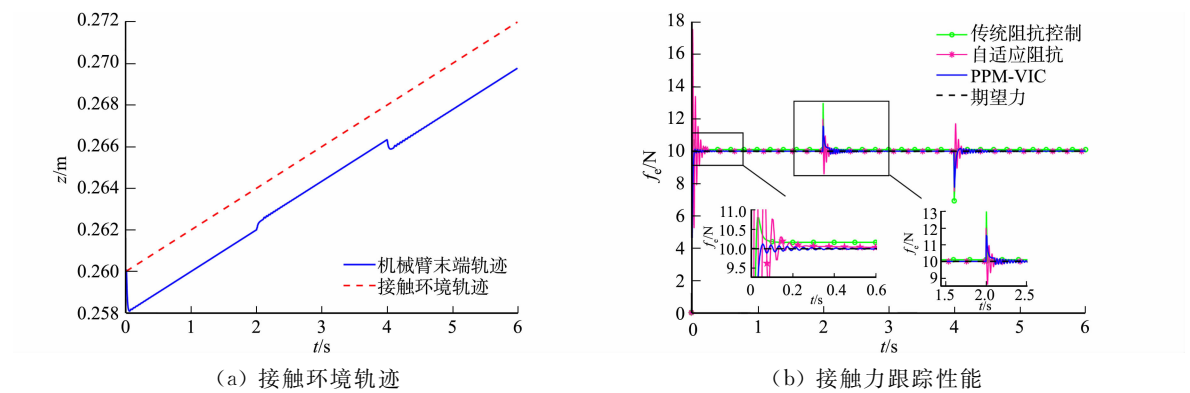


图 7 斜面环境恒力跟踪

Fig. 7 Constant force tracking on slope surrounding

斜面环境恒力跟踪性能对比,如表 1 所示。
对于未知表达式的复杂接触曲面,显然 $\dot{x}_e, \ddot{x}_e, \ddot{x}_e \neq 0$ 且始终随着时间变化。设期望力 $f_d=10$ N,复杂曲面环境恒力跟踪,如图 8 所示。

表 1 斜面环境恒力跟踪性能对比

Tab. 1 Comparison of constant force tracking performance on slope environment

控制方法	超调量/%	调节时间/s	稳态误差/N
传统阻抗	8.0	0.15	0.170 00
自适应变阻抗	68.0	0.70	0.000 25
PPM-VIC	1.0	0.55	0.000 07

由图 8(b)可知:接触环境起伏对传统阻抗控制的影响最大,跟踪误差与环境位置变化速度有关,2 s 后环境变化速度明显变大,力跟踪误差也随之增大。自适应变阻抗控制的动态性能较差,但稳态误差优于传统阻抗控制。PPM-VIC 方法几乎不受环境位置变化的影响,能够以较高的精度跟踪恒定期望力。

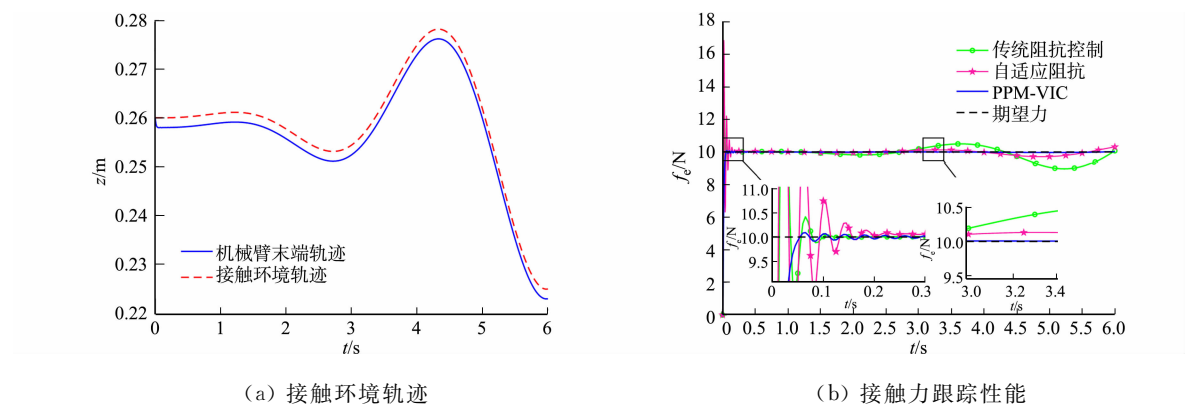


图 8 复杂曲面环境恒力跟踪

Fig. 8 Constant force tracking on complex surface

复杂曲面环境恒力跟踪性能对比,如表 2 所示。复杂曲面环境变力跟踪性能对比,如表 3 所示。

表 2 复杂曲面环境恒力跟踪性能对比

Tab. 2 Comparison of constant force tracking performance on a complex surface environment

控制方法	超调量/%	调节时间/s	稳态误差/N
传统阻抗	37.2	0.15	≤ 1.02
自适应变阻抗	68.5	0.25	≤ 0.29
PPM-VIC	0.9	0.30	≤ 0.03

表 3 复杂曲面环境变力跟踪性能对比

Tab. 3 Comparison of variable force tracking performance on complex surface environment

控制方法	超调量/%	调节时间/s	稳态误差/N
传统阻抗	38.0	0.15	≤ 0.880
自适应变阻抗	69.0	0.30	≤ 0.239
PPM-VIC	4.0	0.40	≤ 0.014

设期望力为变力, 即 $f_d = 10 + 5\sin(t) N$, 复杂曲面环境变力跟踪, 如图 9 所示。

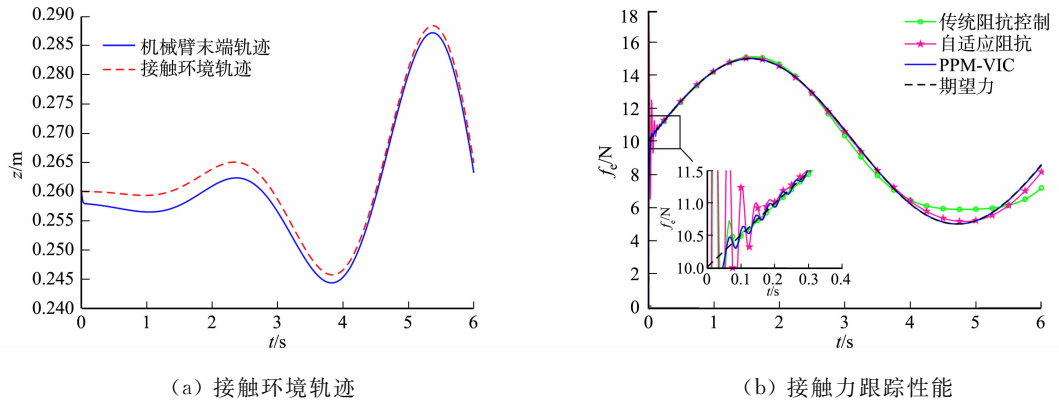


图 9 复杂曲面环境变力跟踪

Fig. 9 Variable force tracking on complex surface environment

由图 9 可知: 前 1.5 s 环境位置变化相对平缓, 3 种控制方法均可在稳定后较好地跟踪期望力; 自 2.5 s 开始, 接触环境变得陡峭, 传统阻抗控制和自适应变阻抗都出现了不同程度的跟踪误差, 但自适应变阻抗误差较小, PPM-VIC 方法仍然能以较高精度跟踪连续变化的期望力。

4 结束语

针对非结构环境下机械臂难以实现良好的力跟踪性能, 以及强化学习数据利用效率低的问题, 将机械臂力控制问题建模为马尔可夫决策过程, 提出一种基于概率预测强化学习的 PPM-VIC 方法。利用概率预测模型及矩匹配法预测未来时间域内的状态序列分布, 从而产生大量虚拟训练样本, 使机械臂仅需与环境交互 18 次即可获得良好的变阻抗策略。仿真结果表明, 提出的变阻抗策略适用于各种非结构接触环境, 其超调量、震荡幅度及稳态精度均显著优于传统阻抗控制和自适应变阻抗控制, 在期望力恒定和连续变化时均具备良好的跟踪性能。

参考文献:

[1] PETERNEL L, TSAGARAKIS N, CALDWELL D, *et al.* Robot adaptation to human physical fatigue in human-robot co-manipulation[J]. *Autonomous Robots*, 2018, 42(5): 1011-1021. DOI: 10.1007/s10514-017-9678.

[2] 倪涛, 黎锐, 缪海峰, 等. 船载机械臂末端位置实时补偿[J]. *吉林大学学报(工学版)*, 2020, 50(6): 2028-2035. DOI: 10.13229/j.cnki.jdxbgxb20190662.

[3] REN Qinyuan, ZHU Wenxin, ZHAO Feng, *et al.* Learning-based force control of a surgical robot for tool-soft tissue interaction[J]. *IEEE Robotics and Automation Letters*, 2021, 6(4): 6345-6352. DOI: 10.1109/LRA.2021.3093018.

[4] LI Y, GOWRISHANKAR G, NATHANAEL J, *et al.* Force, impedance, and trajectory learning for contact tooling and haptic identification[J]. *IEEE Transactions on Robotics*, 2018, 34(5): 1-13. DOI: 10.1109/TRO.2018.2830405.

[5] 刘胜遂, 李利娜, 熊晓燕, 等. 基于卡尔曼滤波的机器人自适应控制方法研究[J]. *机电工程*, 2023, 40(6): 936-944. DOI: 10.3969/j.issn.1001-4551.2023.06.017.

[6] 李振, 赵欢, 王辉, 等. 机器人磨抛加工接触稳态自适应力跟踪研究[J]. *机械工程学报*, 2022, 58(9): 200-209. DOI: 10.3901/JME.2022.09.200.

[7] ROVEDA L, IANNACCI N, VICENTINI F, *et al.* Optimal impedance force-tracking control design with impact formulation for interaction tasks[J]. *IEEE Robotics and Automation Letters*, 2016, 1(1): 130-136. DOI: 10.1109/LRA.2015.2508061.

[8] JUNG S, HSIA T C, BONITZ R G. Force tracking impedance control of robot manipulators under unknown environment[J]. *IEEE Transactions on Control Systems Technology*, 2004, 12(3): 474-483. DOI: 10.1109/TCST.2004.824320.

[9] DUAN Jinjun, GAN Yajui, CHEN Ming, *et al.* Adaptive variable impedance control for dynamic contact force tracking in uncertain environment[J]. *Robotics and Autonomous Systems*, 2018, 102: 54-65. DOI: 10.1016/j.robot.2018.01.009.

[10] CAO Hongli, CHEN Xiaolan, HE Ye, *et al.* Dynamic adaptive hybrid impedance control for dynamic contact force tracking in uncertain environments [J]. IEEE Access, 2019, 7: 83162-83174. DOI: 10. 1109/ACCESS. 2019. 2924696.

[11] HAMEDANI M H, SADEGHIAN H, ZEKRI M, *et al.* Intelligent impedance control using wavelet neural network for dynamic contact force tracking in unknown varying environments[J]. Control Engineering Practice, 2021, 113: 104840. DOI:10. 1016/J. CONENGPRAC. 2021. 104840.

[12] ANDRYCHOWICZ O M, BAKER B, CHOCIEJ M, *et al.* Learning dexterous in-hand manipulation[J]. The International Journal of Robotics Research, 2020, 39(1): 3-20. DOI:10. 1177/0278364919887447.

[13] LI Yunfei, KONG Tao, LI Lei, *et al.* Learning design and construction with varying-sized materials via prioritized memory resets[C]//International Conference on Robotics and Automation, Philadelphia: IEEE Press, 2022: 7469-7476. DOI:10. 1109/ICRA46639. 2022. 9811624.

[14] BUCHLI J, STULP F, THEODOROU E, *et al.* Learning variable impedance control[J]. The International Journal of Robotics Research, 2011, 30(7): 820-833. DOI:10. 1177/0278364911402527.

[15] LI Chao, ZHANG Zhi, XIA Guihua, *et al.* Efficient force control learning system for industrial robots based on variable impedance control[J]. Sensors, 2018, 18(8): 2539. DOI:10. 3390/s18082539.

[16] WU Min, HE Yanhao, LIU S. Adaptive impedance control based on reinforcement learning in a human-robot collaboration task with human reference estimation[J]. International Journal of Mechanics and Control, 2020, 21(1): 21-32. DOI:10. 1007/978-3-030-19648-6_12.

[17] DU Zhijiang, WANG Wei, YAN Zhiyuan, *et al.* Variable admittance control based on fuzzy reinforcement learning for minimally invasive surgery manipulator[J]. Sensors, 2017, 17(4): 844. DOI:10. 3390/s17040844.

[18] DEISENROTH M P, FOX D, RASMUSSEN C E. Gaussian processes for data-efficient learning in robotics and control[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2015, 37(2): 408-423. DOI:10. 1109/TPAMI. 2013. 218.

[19] RASMUSSEN C E, WILLIAMS C K I. Gaussian processes for machine learning[M]. Cambridge: MIT Press, 2005.

[20] DEISENROTH M P. Efficient reinforcement learning using Gaussian process[D]. Karlsruhe: Karlsruhe Institute of Technology, 2010. DOI:10. 5445/KSP/1000019799.

(责任编辑：陈志贤 英文审校：陈婧)