控制与决策
Control and Decision

概率推理学习控制方法的不确定性来源分析及量化

曹瑞, 吕慧涛

引用本文:
曹瑞, 吕慧涛. 概率推理学习控制方法的不确定性来源分析及量化[J]. 控制与决策, 2026, 41(2): 494-504.

在线阅读 View online: https://doi.org/10.13195/j.kzyjc.2025.0153

概率推理学习控制方法的不确定性来源分析及量化

曹 瑞[1,2†]，吕慧涛[3]

(1. 扬州大学 信息与人工智能学院，江苏 扬州 225101；
2. 北京理工大学 自主智能无人系统全国重点实验室，北京 100081；
3. 沈阳飞机设计研究所扬州协同创新研究院有限公司，江苏 扬州 225006)

摘 要: 研究概率推理学习控制 (PILCO) 方法如何在决策过程中使用不确定性, 并进一步探索不确定性对 PILCO 优化性能的影响. 首先, 对给定策略下函数模型中不同不确定性的来源进行梳理和量化; 然后, 使用方差作为不确定性度量, 并采用总方差定律将总成本不确定性分解为两个组成部分; 最后, 引入一个金标准蒙特卡洛方案, 通过传播轨迹来近似 PILCO 的动力学模型, 从而分别量化成本中的随机不确定性和认知不确定性. 仿真结果表明, 当 PILCO 有效学习时, 它选择的是认知成本不确定性在总成本不确定性中具有高占比相关的策略.

关键词: 概率推理学习控制；不确定性；蒙特卡洛；量化；学习策略；演化分析

中图分类号: TP273 **文献标志码:** A

DOI: 10.13195/j.kzyjc.2025.0153

引用格式: 曹瑞, 吕慧涛. 概率推理学习控制方法的不确定性来源分析及量化 [J]. 控制与决策, 2026, 41(2): 494-504.

Analysis and quantification of uncertainty sources in probabilistic inference learning control

CAO Rui[1,2†]，*LV Hui-tao*[3]

(1. School of Information and Artificial Intelligence，Yangzhou University，Yangzhou 225101，China；2. State Key Lab of Autonomous Intelligent Unmanned Systems，Beijing Institute of Technology，Beijing 100081，China；3. Yangzhou Collaborative Innovation Research Institute Co. Ltd，Shenyang Aircraft Design and Research Institute，Yangzhou 225006，China)

Abstract: This paper examines how the probabilistic inference for learning control (PILCO) method incorporates uncertainty into its decision-making process and investigates the impact of uncertainty on its optimization performance. First, the various sources of uncertainty within the function model under a given policy are identified and quantified. Then, variance is adopted as the measure of uncertainty, and the law of total variance is applied to decompose the overall cost uncertainty into two components. Finally, a gold-standard Monte Carlo scheme is introduced to approximate PILCO's dynamics model by propagating trajectories, thereby enabling the separate quantification of aleatoric and epistemic uncertainties in the cost. Simulation results indicate that effective PILCO learning is associated with the selection of policies in which the epistemic cost uncertainty constitutes a high proportion of the total cost uncertainty.

Keywords: probabilistic inference for learning control； uncertainty； Monte Carlo； quantification； evolution analysis；learning strategy

0 引 言

强化学习[1-2]是一种通用的顺序决策框架, 学习从环境到行动的映射, 以最大化奖励函数. 在许多情况下, 智能体所采取的行动不仅会影响瞬时奖励, 还会影响下一步状态, 从而影响所有未来的奖励. 为了在给定的一系列环境中找到一组最佳行动, 智能体必须考虑行动对所有未来奖励的影响. 因此, 试错搜索和延迟奖励是强化学习最显著的两个特征[3-4].

收稿日期: 2025-02-13；**录用日期:** 2025-08-07.

基金项目: 江苏省自然科学基金项目 (BK20230560)；国家自然科学基金项目 (62303400, 92371116, 52272369)；自主智能无 人系统全国重点实验室开放课题项目 (ZZKF2025-3-1)；江苏省高等学校面上项目 (23KJB590003)；2022 年度扬州市“绿扬金凤计划”优秀博士项目 (YZLYJFJH2022YXBS109).

责任编委: 董久祥.

†通信作者. E-mail: stdio@yzu.edu.cn.

概率推理学习控制 (PILCO)[5-6] 是一种基于模型的连续状态-动作的动态系统间接策略搜索方法. PILCO 假设输入上的分布是高斯分布来评估特定策略, 并通过其在特定策略下的系统运动学概率表示传播该分布. 这种方式使得 PILCO 了解当前学习策略可以访问的各种状态, 进而推断该策略在实现低成本方面的有效性. 然而, PILCO 完成这些操作并没有事先进行任何有意的探索, 也就是说, 它是贪婪的[7].

研究者们试图提高 PILCO 的学习效率, 这需要更多关于任务的信息性先验知识, 或者从可用数据中提取更多相关信息. 许多学习控制方法通过在动作选择中引入随机性来实现探索. 例如, 一些方法使控制器随机生成动作, 然后不停探索能获取最大奖励的动作[8]. 然而, 在探索的初始阶段, 智能体可能难以适应纯粹随机的环境变化. 为了处理这一问题, 其他研究者使用ε-贪婪策略[9], 该方法在大多数情况下利用具有最大预期回报的行动, 但有一定概率ε会选择随机行动[3]. 虽然ε-贪婪方法鼓励学习, 但对于现实世界的系统而言, 选择随机动作可能会反复将系统引导到不理想或危险的环境[10]. 此外, 探索是无方向的, 因此随机动作选择效率低下, 其经常会导致智能体重复返回状态空间中已经充分探索过的区域.

目前, PILCO 通过系统运动学 (过渡函数) 的概率模型级联不确定输入来评估特定策略. 这种做法使得它可以了解该策略下可能产生的各种状态, 进而在实现低成本方面的有效性做出更明智的决定[11]. 然而, 这种方法只允许 PILCO 根据模型量化的总不确定性做出决策, 而实际上系统中存在两个不确定性来源. 不确定性的第 1 个来源是随机性, 即代表每次智能体遇到的不同未知量, 例如动态系统中的测量误差或混沌运动; 另一个不确定性来源是认知, 产生于智能体, 即原则上可以知道但目前尚未获取的信息. 因此, 通过观察更多的数据可以减少认知不确定性, 而任意不确定性是不可减少的. 这意味着在某些情况下, PILCO 可能会反复选择与高随机不确定性相对应的策略, 这将阻碍学习.

为了进一步探索两种不确定性对 PILCO 优化性能的影响, 本文研究了 PILCO 如何在决策过程中使用不确定性, 特别是对给定策略下过渡函数模型中的不同不确定性来源进行了梳理和量化. 本文使用方差作为不确定性的度量, 并采用总方差定律将总成本不确定性分解为两个组成部分, 同时引入一个金标准蒙特卡洛方案, 通过传播轨迹来近似 PILCO 的动力学模型, 从而分别估计成本中的随机不确定性和认知不确定性. 最后, 通过仿真结果表明, 当 PILCO 进行有效学习时, 它选择的是认知成本不确定性在总成本不确定性占比较高的相关策略. 本文的主要贡献如下:

1) 提出方差分解, 将 PILCO 中的总成本不确定性解耦为随机分量与认知分量, 克服了现有工作[5-7]仅量化总不确定性的局限.

2) 设计金标准蒙特卡洛方案, 通过双重采样架构高效分离并量化两类不确定性, 较传统蒙特卡洛方法[12] 计算效率有所提升.

3) 现有策略搜索方法 (如ε-贪婪[9]) 依赖启发式探索, 缺乏理论指导. 本文分析发现认知不确定性占比与 PILCO 学习效率的正相关性, 为基于不确定性的主动探索策略提供了新的理论支撑.

1　PILCO 不确定性分析及量化

本节描述了如何对 PILCO 决策过程中的不确定性来源进行分解和量化. 首先简要介绍 PILCO 算法; 然后引入一个有限参数三角贝叶斯回归模型来近似 PILCO 的概率高斯过程 (GP) 动力学模型, 将三角模型与全 GP 联系起来, 表明它可以被解释为可以近似任何全 GP 的稀疏谱 GP; 接着讨论转换函数中的不确定性来源及其对成本函数的影响, 并提出方差分解, 将不确定性分离并量化为随机和认知成分; 最后提出一个金标准蒙特卡洛方案来估计随机和认知不确定性.

1.1　PILCO 描述

PILCO[6] 是一种基于模型的间接策略搜索方法, 用于描述连续状态$\boldsymbol{x} \in \mathbb{R}^D$和动作$\boldsymbol{u} \in \mathbb{R}^F$动力系统, 由下式描述:

$$\boldsymbol{x}_{t+1} = f(\boldsymbol{x}_t, \boldsymbol{u}_t) + \boldsymbol{\eta},\ \boldsymbol{\eta} \sim \mathcal{N}(0, \Sigma_{\boldsymbol{\eta}}). \tag{1}$$

其中: 系统动力学f是未知的, $\boldsymbol{\eta}$是均值为 0 的高斯噪声, $\Sigma_{\boldsymbol{\eta}}$代表噪声方差. PILCO 方法的目标是找到一个确定性控制策略$\Pi : \boldsymbol{x} \mapsto \pi(\boldsymbol{x}, \boldsymbol{\theta}) = \boldsymbol{u}$, 使得预期成本

$$J^{\pi}(\boldsymbol{\theta}) = \sum_{t=0}^{T} \mathbb{E}_{\boldsymbol{x}_t}[\mathbb{C}(\boldsymbol{x}_t)],\ \boldsymbol{x}_0 \sim \mathcal{N}(\boldsymbol{\mu}_0, \Sigma_{\boldsymbol{\eta}}), \tag{2}$$

在遵循Π进行T阶运算后, 其值最小化[7]. 该策略由$\boldsymbol{\theta}$参数化, $\boldsymbol{\theta}$是非线性径向基网络的权重和特征参数. PILCO 方法中时间t处状态$\boldsymbol{x}$的成本函数被定义为广义二元饱和函数[7]

$$\mathbb{C}(\boldsymbol{x}_t) = 1 - \exp\left(-\frac{1}{2\sigma_c^2}\|\boldsymbol{x}_t - \boldsymbol{x}_{\text{target}}\|^2\right) \in [0, 1]. \tag{3}$$

该函数用于惩罚从当前状态$\boldsymbol{x}_t$到目标状态$\boldsymbol{x}_{\text{target}}$的欧几里德距离$||\boldsymbol{x}_t - \boldsymbol{x}_{\text{target}}||$. 因此, $\mathbb{C}(\boldsymbol{x}_t)$是局部二

次的. 此外, 若当前状态与目标状态之间差异较大, $\mathbb{C}(\boldsymbol{x}_t)$会达到饱和数值 1, 饱和距离由参数$\sigma_c$控制.

式 (1) 所描述的系统动力学在 PILCO 中被建模为概率高斯过程. D维状态向量和F维动作向量的元组$(\boldsymbol{x}_t, \boldsymbol{u}_t) \in \mathbb{R}^{D+F}$作为训练输入, 差$\Delta_t = \boldsymbol{x}_{t+1} - \boldsymbol{x}_t + \boldsymbol{\varepsilon} \in \mathbb{R}^D$, $\boldsymbol{\varepsilon} \sim \mathcal{N}(0, \Sigma_{\boldsymbol{\eta}})$作为训练目标. 有条件独立的 GP 针对每个目标维度进行训练[6]

$$p(\boldsymbol{x}_{t+1}|\boldsymbol{x}_t, \boldsymbol{u}_t) = \mathcal{N}(\boldsymbol{x}_t|\boldsymbol{\mu}_t, \Sigma_t), \tag{4}$$

$$\boldsymbol{\mu}_{t+1} = \boldsymbol{x}_t + \mathbb{E}_f(\Delta_t), \tag{5}$$

$$\Sigma_{t+1} = \mathrm{var}_f[\Delta_t], \tag{6}$$

其中状态$\boldsymbol{x}_{t+1}$的平均值$\boldsymbol{\mu}$是通过将当前状态$\boldsymbol{x}_t$与单步差分预测$\mathbb{E}_f[\Delta_t]$相加而获得的. GP 由均值函数$m(\cdot)$和半正定协方差函数$k(\cdot,\cdot)$定义; PILCO 将均值先验函数设置为零, 并使用了平稳各向异性协方差函数

$$k(\tilde{\boldsymbol{x}}_i, \tilde{\boldsymbol{x}}_j) = \sigma_0^2 \exp\left(-\frac{1}{2}(\tilde{\boldsymbol{x}}_i - \tilde{\boldsymbol{x}}_j)^{\mathrm{T}} \boldsymbol{\Lambda}^{-1} (\tilde{\boldsymbol{x}}_i - \tilde{\boldsymbol{x}}_j)\right). \tag{7}$$

其中: $\tilde{\boldsymbol{x}} = [\boldsymbol{x}, \boldsymbol{u}]^{\mathrm{T}}$是状态-动作向量, $\boldsymbol{\Lambda} = \mathrm{diag}[\ell_1^2, \ldots, \ell_{D+F}^2]$取决于特征长度尺度$\ell$, σ_0是隐函数方差[7]. 长度尺度决定了协方差随输入间距离衰减的速度[12].

1.2 近似 PILCO 的 GP 模型

为了使用蒙特卡洛方法近似 PILCO 成本函数的不确定性, 有必要使用从 PILCO 动力学模型的后验分布中采样的函数进行单步预测. 在 PILCO 的 GP 中使用平方指数协方差函数, 其对应于具有无限个基函数的贝叶斯线性回归模型[13]. 这意味着描述一个代表性函数需要无限数量的权重, 因为每个基函数对应一个权重值. 为了克服这个问题, 本文使用有限参数平稳三角贝叶斯回归模型[12] 来近似 GP. 除了解决采样问题外, 该模型还具有其他优点. 首先, 三角基函数的周期性意味着不需要明确指定每个基函数的位置. 其次, 直接 GP 实现在计算与内存要求方面存在实际限制, 要求分别为$\mathcal{O}(n^2)$和$\mathcal{O}(n^3)$; 而三角模型具有$\mathcal{O}(nm^2)$的计算要求与$\mathcal{O}(nm)$的存储要求, 其中$m << n$ [12, 14], 这能够降低对计算和存储量的需求. 后续 1.2.1 节将采用三角贝叶斯回归模型近似的 GP 模型与 PILCO 的单步预测联系起来; 1.2.2 节将表明该近似 GP 模型可以被视为一个稀疏的平稳 GP, 能够近似任何完整的 GP.

三角贝叶斯回归模型由三角函数的线性组合组成[12]

$$f(\tilde{\boldsymbol{x}}) = \sum_{r=1}^{m} a_r \cos(2\pi \boldsymbol{s}_r^{\mathrm{T}} \tilde{\boldsymbol{x}}) + b_r \sin(2\pi \boldsymbol{s}_r^{\mathrm{T}} \tilde{\boldsymbol{x}}). \tag{8}$$

其中: $\boldsymbol{s}_r$是每对基函数共享的谱频向量, 维数为$D+F$; a_r、b_r是每个基函数独立的振幅参数 (谱频率选择见第 1.2.2 节). 振幅是具有线性缩放方差的独立高斯先验, 即

$$a_r \sim \mathcal{N}\left(0, \frac{\sigma_0^2}{m}\right),\ b_r \sim \mathcal{N}\left(0, \frac{\sigma_0^2}{m}\right), \tag{9}$$

其中m是基函数的数目. 谱频率作为确定性参数, 振幅以贝叶斯方式处理. 为此, 系统模型被打包为振幅集与基函数之间的点积

$$f(\tilde{\boldsymbol{x}}, \boldsymbol{w}) = \boldsymbol{w}^{\mathrm{T}} \varphi(\tilde{\boldsymbol{x}}). \tag{10}$$

其中: $\boldsymbol{w} = [a_1, b_1, \ldots, a_m, b_m]^{\mathrm{T}}$是模型权重向量, 并且

$$\varphi(\tilde{\boldsymbol{x}}) = [\cos(2\pi \boldsymbol{s}_1^{\mathrm{T}} \tilde{\boldsymbol{x}}) \quad \sin(2\pi \boldsymbol{s}_1^{\mathrm{T}} \tilde{\boldsymbol{x}}) \quad \ldots \quad \rightarrow \quad \leftarrow \quad \cos(2\pi \boldsymbol{s}_m^{\mathrm{T}} \tilde{\boldsymbol{x}}) \quad \sin(2\pi \boldsymbol{s}_m^{\mathrm{T}} \tilde{\boldsymbol{x}})]^{\mathrm{T}}. \tag{11}$$

与 PILCO 类似, 与目标状态的差$\boldsymbol{\Delta}$被用作训练目标, 假设这些目标由函数$f(\tilde{\boldsymbol{x}}, \boldsymbol{w})$生成, 并受到恒定方差为$\sigma_n^2$的加性高斯噪声的破坏, 即

$$\boldsymbol{\Delta} = f(\tilde{\boldsymbol{x}}, \boldsymbol{w}) + \varepsilon,\ \varepsilon \sim \mathcal{N}(0, \sigma_n^2), \tag{12}$$

其中差$\boldsymbol{\Delta}$的概率分布可以写为

$$p(\boldsymbol{\Delta}|\tilde{\boldsymbol{x}}, \boldsymbol{w}, \sigma_n^2) = \mathcal{N}(\boldsymbol{\Delta}|f(\tilde{\boldsymbol{x}}, \boldsymbol{w}), \sigma_n^2). \tag{13}$$

对于一个输入数据集$\tilde{\boldsymbol{X}} = \{\tilde{\boldsymbol{x}}_1, \ldots, \tilde{\boldsymbol{x}}_N\}$, 对应的目标值为$\boldsymbol{\Delta} = \{\Delta_1, \ldots, \Delta_N\}$. 假设这些数据点是从$\mathcal{N}(\boldsymbol{\Delta}|f(\tilde{\boldsymbol{x}}, \boldsymbol{w}), \sigma_n^2)$中独立抽取的, 则$\boldsymbol{\Delta}$的似然函数为

$$p(\boldsymbol{\Delta}|\tilde{\boldsymbol{x}}, \boldsymbol{w}, \sigma_n^2) = \prod_{i=1}^{N} \mathcal{N}(\Delta_i|\boldsymbol{w}^{\mathrm{T}} \varphi(\tilde{\boldsymbol{x}}_i), \sigma_n^2). \tag{14}$$

根据式 (14) 可知, 模型权重上的后验分布与似然函数和先验的乘积成正比. 由于高斯先验是共轭的, 模型权重的后验分布也是高斯的, 有

$$q(\boldsymbol{w}|\boldsymbol{\Delta}, \tilde{\boldsymbol{X}}, \sigma_0^2, \sigma_n^2) = \mathcal{N}(\boldsymbol{w}|\mu_{\boldsymbol{w}}, \boldsymbol{A}^{-1}). \tag{15}$$

其中: 后验均值$\mu_{\boldsymbol{w}}$和精度矩阵$\boldsymbol{A}$分别为

$$\mu_{\boldsymbol{w}} = \frac{1}{\sigma_n^2} \boldsymbol{A}^{-1} \boldsymbol{\Phi} \boldsymbol{y}, \tag{16}$$

$$\boldsymbol{A} = \frac{1}{\sigma_n^2} \boldsymbol{\Phi} \boldsymbol{\Phi}^{\mathrm{T}} + \frac{m}{\sigma_0^2} \boldsymbol{I}_{2m}, \tag{17}$$

$\boldsymbol{\Phi} = [\varphi(\tilde{\boldsymbol{x}}_1), \ldots, \varphi(\tilde{\boldsymbol{x}}_N)]$是$2m \times N$的设计矩阵.

要对新的输入值$\tilde{\boldsymbol{x}}_*$进行预测Δ_*, 需要对预测分布进行评估, 其定义如下:

$$p(\Delta_*|\boldsymbol{\Delta}, \tilde{\boldsymbol{X}}, \tilde{\boldsymbol{x}}_*, \sigma_0^2, \sigma_n^2) = \int p(\Delta_*|\tilde{\boldsymbol{x}}_*, \boldsymbol{w}, \sigma_n^2) q(\boldsymbol{w}|\boldsymbol{\Delta}, \tilde{\boldsymbol{X}}, \sigma_0^2, \sigma_n^2) \mathrm{d}\boldsymbol{w}. \tag{18}$$

其中: 目标变量的条件分布$p(\Delta_*|\tilde{\boldsymbol{x}}_*, \boldsymbol{w}, \sigma_n^2)$由式 (13) 定义, 权重的后验分布由式 (15) 给出. 单个输入的预

测分布可通过下式[12] 计算:

$$p(\Delta_*|\boldsymbol{\Delta}, \tilde{\boldsymbol{X}}, \tilde{\boldsymbol{x}}_*, \sigma_0^2, \sigma_n^2) = \mathcal{N}(\Delta_*|\mu_{\Delta_*}, \sigma_{\Delta_*}^2). \quad (19)$$

其中

$$\mu_{\Delta_*} = \frac{1}{\sigma_n^2}\varphi(\tilde{\boldsymbol{x}}_*)^{\mathrm{T}}\boldsymbol{A}^{-1}\boldsymbol{\Phi}\boldsymbol{\Delta}, \quad (20)$$

$$\sigma_{\Delta_*}^2 = \sigma_n^2 + \varphi(\tilde{\boldsymbol{x}}_*)^{\mathrm{T}}\boldsymbol{A}^{-1}\varphi(\tilde{\boldsymbol{x}}_*). \quad (21)$$

由根据式 (21) 计算的预测方差可以看出, 该方差由两项组成, 第 1 项是数据上存在的噪声, 第 2 项是与模型权重$\boldsymbol{w}$相关的不确定性[15]. 第 2 项为本文研究的重点, 将在第 1.3 节中进行扩展. 随着观察到更多的数据, 学习策略会对系统模型越来越有信心, 使得第 2 项的数值收缩, 后验分布变窄.

1.2.1　单步预测

PILCO 方法中 GP 动力学模型生成的单步预测是通过后验预测分布生成的. 为此, 首先从式 (15) 所示的模型权重后验分布$\boldsymbol{w} \sim q(\boldsymbol{w})$中提取一组维数为$2m$的权重, 为简洁起见, 此处省略条件变量. 然后用以下公式计算单步预测:

$$\boldsymbol{x}_{t+1} = \boldsymbol{x}_t + \Delta_t = \boldsymbol{x}_t + \boldsymbol{w}^{\mathrm{T}}\varphi(\tilde{\boldsymbol{x}}_t) + \boldsymbol{\varepsilon},$$
$$\boldsymbol{\varepsilon} \sim \mathcal{N}(0, \Sigma_{\boldsymbol{\varepsilon}}). \quad (22)$$

1.2.2　全 GP 的稀疏近似

本节遵循 Lázaro-Gredilla 等[12] 的处理方法, 通过创建平稳协方差函数的稀疏表示, 将三角基函数模型与完整 GP 联系起来. GP 回归是一种概率、非参数贝叶斯方法, 完全由均值函数和协方差函数指定[13], 即

$$m(\tilde{\boldsymbol{x}}) = \mathbb{E}[f(\tilde{\boldsymbol{x}})],$$
$$k(\tilde{\boldsymbol{x}}_i, \tilde{\boldsymbol{x}}_j) = \mathbb{E}[(f(\tilde{\boldsymbol{x}}_i) - m(\tilde{\boldsymbol{x}}_i))(f(\tilde{\boldsymbol{x}}_j) - m(\tilde{\boldsymbol{x}}_j))]. \quad (23)$$

先考虑均值函数为零, 即$m(\tilde{\boldsymbol{x}}) = 0$; 协方差函数为平稳平方指数协方差函数, 表达式如下:

$$k(\tilde{\boldsymbol{x}}_i, \tilde{\boldsymbol{x}}_j) = k(\tilde{\boldsymbol{x}}_i - \tilde{\boldsymbol{x}}_j) = k(\boldsymbol{\tau}), \quad (24)$$

其中平稳协方差函数是$\boldsymbol{\tau} = \tilde{\boldsymbol{x}}_i - \tilde{\boldsymbol{x}}_j$的函数, 仅取决于输入之间的差. 考虑式 (8) 所示三角基函数模型, 在先验条件下, 函数上的分布是均值为零的高斯分布, 方差为由三角基函数模型表示的平稳协方差[12], 即

$$k(\tilde{\boldsymbol{x}}_i, \tilde{\boldsymbol{x}}_j) = \frac{\sigma_0^2}{m}\varphi(\tilde{\boldsymbol{x}}_i)^{\mathrm{T}}\varphi(\tilde{\boldsymbol{x}}_i) =$$
$$\frac{\sigma_0^2}{m}\sum_{r=1}^{m}\cos(2\pi\boldsymbol{s}_r^{\mathrm{T}}(\tilde{\boldsymbol{x}}_i - \tilde{\boldsymbol{x}}_j)). \quad (25)$$

下面将描述如何通过稀疏化 GP 近似上式的具体过程.

定理 1 (Bochner 定理)[16]　若$\mathbb{R}^D$上的复值函数k是$\mathbb{R}^D$上弱平稳均方连续复值随机过程的协方差函数, 则它可以表示为

$$k(\boldsymbol{\tau}) = \int_{\mathbb{R}} \mathrm{e}^{2\pi i\boldsymbol{s}\cdot\boldsymbol{\tau}}\mathrm{d}\mu_{fm}(\boldsymbol{s}), \quad (26)$$

其中μ_{fm}是一个正的有限度量.

如果μ_{fm}具有概率密度$S(\boldsymbol{s})$, 则S是与k相关的谱密度[13], 与概率度量$p_S(\boldsymbol{s})$成正比, 即$S(\boldsymbol{s}) \propto p_S(\boldsymbol{s})$. 比例常数可以通过$\boldsymbol{\tau} = 0$时计算方程 (24) 中的协方差函数来获得, 即

$$S(\boldsymbol{s}) = k(\boldsymbol{0})p_S(\boldsymbol{s}) = \sigma_0^2 p_S(\boldsymbol{s}). \quad (27)$$

如果$S(\boldsymbol{s})$存在, 根据维纳-辛钦定理[17], 则协方差函数和谱密度形成傅里叶对

$$k(\boldsymbol{\tau}) = \int_{\mathbb{R}} S(\boldsymbol{s})\mathrm{e}^{2\pi i\boldsymbol{s}\cdot\boldsymbol{\tau}}\mathrm{d}\boldsymbol{s},$$
$$S(\boldsymbol{s}) = \int k(\boldsymbol{\tau})\mathrm{e}^{-2\pi i\boldsymbol{s}\cdot\boldsymbol{\tau}}\mathrm{d}\boldsymbol{\tau}. \quad (28)$$

由于$S(\boldsymbol{s})$与$\boldsymbol{s}$中的多元概率密度成正比, 协方差函数可以重写为对概率度量$p_S(\boldsymbol{s})$的期望

$$k(\boldsymbol{\tau}) = \int_{\mathbb{R}^D} \mathrm{e}^{2\pi i\boldsymbol{s}^{\mathrm{T}}(\tilde{\boldsymbol{x}}_i - \tilde{\boldsymbol{x}}_j)}S(\boldsymbol{s})\mathrm{d}\boldsymbol{s} =$$
$$\delta_0^2\int_{\mathbb{R}^D} \mathrm{e}^{2\pi i\boldsymbol{s}^{\mathrm{T}}\tilde{\boldsymbol{x}}_i}(\mathrm{e}^{2\pi i\boldsymbol{s}^{\mathrm{T}}\tilde{\boldsymbol{x}}_j})^* p_S(\boldsymbol{s})\mathrm{d}\boldsymbol{s} =$$
$$\delta_0^2\mathbb{E}_{p_s(\boldsymbol{s})}[\mathrm{e}^{2\pi i\boldsymbol{s}^{\mathrm{T}}\tilde{\boldsymbol{x}}_i}(\mathrm{e}^{2\pi i\boldsymbol{s}^{\mathrm{T}}\tilde{\boldsymbol{x}}_j})^*], \quad (29)$$

其中上标“$*$”表示复数共轭. 式 (29) 中的结果可以通过简单的蒙特卡洛来估计. 由于谱密度关于零对称, 采样频率对$\{\boldsymbol{s}_r, -\boldsymbol{s}_r\}$保留了虚项抵消的精确展开式

$$k(\tilde{\boldsymbol{x}}_i, \tilde{\boldsymbol{x}}_j) \simeq$$
$$\frac{\delta_0^2}{2m}\sum_{r=1}^{m}[\mathrm{e}^{2\pi i\boldsymbol{s}_r^{\mathrm{T}}\tilde{\boldsymbol{x}}_i}(\mathrm{e}^{2\pi i\boldsymbol{s}_r^{\mathrm{T}}\tilde{\boldsymbol{x}}_j})^* + (\mathrm{e}^{2\pi i\boldsymbol{s}_r^{\mathrm{T}}\tilde{\boldsymbol{x}}_i})^*\mathrm{e}^{2\pi i\boldsymbol{s}_r^{\mathrm{T}}\tilde{\boldsymbol{x}}_j}] =$$
$$\frac{\delta_0^2}{2m}\mathrm{Re}\left[\sum_{r=1}^{m}\mathrm{e}^{2\pi i\boldsymbol{s}_r^{\mathrm{T}}\tilde{\boldsymbol{x}}_i}(\mathrm{e}^{2\pi i\boldsymbol{s}_r^{\mathrm{T}}\tilde{\boldsymbol{x}}_j})^*\right] =$$
$$\frac{\delta_0^2}{2m}\sum_{r=1}^{m}\cos(2\pi\boldsymbol{s}_r^{\mathrm{T}}(\tilde{\boldsymbol{x}}_i - \tilde{\boldsymbol{x}}_j)). \quad (30)$$

其中: Re 是复数的实部; $\boldsymbol{s}_r$为蒙特卡洛频率的有限集 (称为谱点), 其从$p_S(\boldsymbol{s})$中采样. 该结果表明, 通过稀疏化全 GP 的平稳协方差函数, 可以恢复式 (25), 这意味着上述的三角基函数模型确实是一个稀疏谱高斯过程.

对 GP 动力学模型中使用的平方指数协方差函数进行傅里叶变换, 可以得到多元高斯形式的概率密度

$$p_S(\boldsymbol{s}) = \frac{1}{k(\boldsymbol{0})}\int_{\mathbb{R}^D} \mathrm{e}^{-2\pi i\boldsymbol{s}^{\mathrm{T}}\boldsymbol{\tau}}k(\boldsymbol{\tau})\mathrm{d}\boldsymbol{\tau} =$$
$$\sqrt{|2\pi\boldsymbol{\Lambda}|}\exp(-2\pi^2\boldsymbol{s}^{\mathrm{T}}\boldsymbol{\Lambda}\boldsymbol{s}). \quad (31)$$

之后从中绘制谱点, 用于$k(\tilde{\boldsymbol{x}}_i, \tilde{\boldsymbol{x}}_j)$的蒙特卡洛估计. 式中$\boldsymbol{\Lambda}$的定义与式 (7) 一致.

1.2.3 超参数优化

在对三角基函数模型进行完全贝叶斯处理时, 会对σ_0^2和σ_n^2引入先验分布, 并通过对超参数和权重$\boldsymbol{w}$进行边缘化来实现预测. 然而, 虽然对超参数或权重中的任意一个进行边缘化是可行的, 但对两者同时进行完全边缘化在解析上是不可行的[15]. 因此, 此处通过优化对数边缘似然来学习超参数, 即

$$\log p(\boldsymbol{\Delta}|\boldsymbol{\theta}) = -\frac{1}{2\sigma_n^2}\Big[\boldsymbol{\Delta}^{\mathrm{T}}\boldsymbol{\Delta} - \frac{1}{\sigma_n^2}\boldsymbol{\Delta}^{\mathrm{T}}\boldsymbol{\Phi}^{\mathrm{T}}\boldsymbol{A}^{-1}\boldsymbol{\Phi}\boldsymbol{\Delta}\Big] - \frac{1}{2}\log|\boldsymbol{A}| + m\log\frac{m}{\sigma_0^2} - \frac{n}{2}\log 2\pi\sigma_n^2. \tag{32}$$

针对超参数σ_0^2、σ_n^2以及控制每个输入维度的长度尺度$\{\ell_1,\dots,\ell_D\}$, 超参数的初始化方式为: σ_0^2设置为y_i的方差, σ_n^2初始化为$\sigma_0^2/4$, 长度尺度初始化为输入维度范围的一半[12]. 从式 (31) 中抽取100组$m\times D$谱点, 并对每组计算对数边缘似然值, 并保留具有最高对数边际似然值的谱点集.

1.3 分析不确定性来源

PILCO 通过 GP 动力学模型级联不确定输入, 以生成状态演化的长期预测$p(\boldsymbol{x}_1|\Pi),\dots,p(\boldsymbol{x}_T|\Pi)$, 从而评估并最小化预期回报$J^{\Pi}$. 这种方法可以揭示在特定策略下可能访问的多种状态, 使算法能够量化当前控制品质并选择下一步控制指令. 然而, 该方法仅考虑了模型预测分布所量化的不确定性总量, 实际上模型存在多种不确定性来源. 本文分析了模型中不确定性的来源, 并将其分解为随机不确定性和认知不确定性.

用于级联不确定输入的单步预测是通过高斯过程的预测分布生成的. 在第 1.2 节中, 这一预测分布是采用三角模型预测分布近似表示的, 这里以简化的符号重新展示为

$$p(\Delta_*|\tilde{\boldsymbol{x}}_*) = \int p(\Delta_*|\tilde{\boldsymbol{x}}_*,\boldsymbol{w})q(\boldsymbol{w})\mathrm{d}\boldsymbol{w}. \tag{33}$$

其中: $p(\Delta_*|\tilde{\boldsymbol{x}}_*,\boldsymbol{w})$是模型似然函数, $q(\boldsymbol{w})$是模型权重的后验分布. 对Δ_*的不确定性或随机性来源包括模型权重$\boldsymbol{w}\sim q(\boldsymbol{w})$、先验$a_r,b_r\sim\mathcal{N}(0,\sigma_0^2/m)$和加性噪声$\varepsilon\sim\mathcal{N}(0,\sigma_n^2)$[18]. 这意味着在$\Delta_*$的预测中混杂了两种类型的不确定性, 即随机不确定性和认知不确定性[19-20]. 随机不确定性来源于a_r、b_r和ε中存在的随机性, 这是不可减少的, 而认知不确定性与模型权重$\boldsymbol{w}$相关. 随着收集到更多数据, 后验分布$q(\boldsymbol{w})$收缩, 认知不确定性会随之减少.

由于 PILCO 在最小化式 (2) 中的J^{Π}时是直接考虑成本函数的, 本文工作可以只关注转移函数中的不确定性, 这些不确定性直接影响成本$\mathbb{C}(\boldsymbol{x})$的不确定性. 从另一角度来看, 对于复杂环境, 模型对状态-动作空间的表示可能非常庞大, 而在状态空间中探索与学习控制任务无关的区域, 会浪费时间和内存资源. 实际上, 人们也仅关心与轨迹相关的不确定性. 所以, 在接下来的处理过程中, 成本函数被指定为$\mathbb{C}(\boldsymbol{x}_t)$, 表示在策略$\Pi$下状态$\boldsymbol{x}$在时间$t$上的成本.

在转移函数中可以使用多种指标来表示不确定性. 在本文中, 成本的方差被用作主要衡量指标. 首先, 将成本的方差表示为随机不确定项, 有

$$\mathbb{V}_{\varepsilon}(\mathbb{C}^{\Pi}(\boldsymbol{x}_t)) = \mathbb{E}_{\varepsilon}[\mathbb{C}^{\Pi}(\boldsymbol{x}_t)^2] - \mathbb{E}_{\varepsilon}[\mathbb{C}^{\Pi}(\boldsymbol{x}_t)]^2. \tag{34}$$

其中所有的随机性来源都被归类到ε这一项中.

接着, 对模型权重的后验分布$q(\boldsymbol{w})$(该分布表示认知不确定性) 应用总期望定律, 可以得到

$$\mathbb{V}(\mathbb{C}^{\Pi}(\boldsymbol{x}_t)) = \mathbb{E}_{q(\boldsymbol{w})}[\mathbb{E}_{\varepsilon}[\mathbb{C}^{\Pi}(\boldsymbol{x}_t)^2]] - \mathbb{E}_{q(\boldsymbol{w})}[\mathbb{E}_{\varepsilon}[\mathbb{C}^{\Pi}(\boldsymbol{x}_t)]]^2. \tag{35}$$

引入后验分布的期望使得方程的左侧发生了变化, 使其能够表示总不确定性. 注意到$\mathbb{E}_{\varepsilon}[\mathbb{C}^{\Pi}(\boldsymbol{x}_t)^2]$可以被替换为方差和一阶矩的和, 从而得到

$$\mathbb{V}(\mathbb{C}^{\Pi}(\boldsymbol{x}_t)) = \mathbb{E}_{q(\boldsymbol{w})}[\mathbb{V}_{\varepsilon}(\mathbb{C}^{\Pi}(\boldsymbol{x}_t)) + \mathbb{E}_{\varepsilon}[\mathbb{C}^{\Pi}(\boldsymbol{x}_t)]^2] - \mathbb{E}_{q(\boldsymbol{w})}[\mathbb{E}_{\varepsilon}[\mathbb{C}^{\Pi}(\boldsymbol{x}_t)]]^2. \tag{36}$$

总期望值也是各期望值的和, 因此将第 1 项分离, 并将其与最后一项重新组合在一起, 可以得到

$$\mathbb{V}(\mathbb{C}^{\Pi}(\boldsymbol{x}_t)) = \mathbb{E}_{q(\boldsymbol{w})}[\mathbb{V}_{\varepsilon}(\mathbb{C}^{\Pi}(\boldsymbol{x}_t))] + (\mathbb{E}_{q(\boldsymbol{w})}[\mathbb{E}_{\varepsilon}[\mathbb{C}^{\Pi}(\boldsymbol{x}_t)]^2] - \mathbb{E}_{q(\boldsymbol{w})}[\mathbb{E}_{\varepsilon}[\mathbb{C}^{\Pi}(\boldsymbol{x}_t)]]^2). \tag{37}$$

注意到括号内的项是相对于后验分布的方差, 可以通过一阶矩和二阶矩来表示, 从而得到最终的不确定性分解

$$\underbrace{\mathbb{V}(\mathbb{C}^{\Pi}(\boldsymbol{x}_t))}_{\mathbb{V}_1} = \underbrace{\mathbb{E}_{q(\boldsymbol{w})}[\mathbb{V}_{\varepsilon}(\mathbb{C}^{\Pi}(\boldsymbol{x}_t))]}_{\mathbb{V}_2} + \underbrace{\mathbb{V}_{q(\boldsymbol{w})}(\mathbb{E}_{\varepsilon}[\mathbb{C}^{\Pi}(\boldsymbol{x}_t)])}_{\mathbb{V}_3}, \tag{38}$$

其中$\mathbb{V}_1$是总不确定性, $\mathbb{V}_2$是随机不确定性, $\mathbb{V}_3$是认知不确定性. 所有不确定性都是在策略Π下时间t时刻的状态$\boldsymbol{x}$的不确定性.

1.4 金标准蒙特卡洛不确定性估计

PILCO 策略评估方法存在一个根本性问题, 即它没有认识到实际情况中只有一个真实的底层动力学存在. 因此, 在预测状态演化时, 它会在每个时间点反复搜索所有可能的动力学. 这个问题可以通过采用蒙特卡洛 (MC) 采样对式 (38) 进行估计来解决. 因为 MC 方法可以对动力学单独采样, 将状态向前传

播, 然后再次采样动力学, 这允许以正确的顺序计算式 (38) 所需的平均值. 通过将式 (38) 转化为蒙特卡洛采样形式, 可以直观地理解在策略Π下, 随着轨迹在 PILCO 的状态-动作表示中的演化. 随后概述如何将成本中的随机不确定性和认知不确定性区分开来.

首先, 考虑$\mathbb{V}_2$表示随机不确定性的项. 从后验分布$\boldsymbol{w} \sim q(\boldsymbol{w})$中采样的一组权重, 根据式 (22) 使用它们, 在策略Π下通过转移函数对N条轨迹进行单步预测. 一旦抽取了一组权重并保持固定, 除了加性高斯噪声ε外, 单步预测结果将是确定的. 接下来, 计算N条轨迹的单步预测方差, 得到从$q(\boldsymbol{w})$中抽取的单个转移函数的随机不确定性的经验估计值. 然后, 这个过程重复M次, 并对从后验分布中抽取的M个转移函数的结果取经验平均值. 通过这种方法来计算出仅受噪声影响的转换平均方差. 只要N和M足够大, 上述过程就能提供在策略Π下单次状态转移中总随机不确定性的无偏估计值. 在通过转移函数的T个步骤过程中, 随机不确定性是每一步随机不确定性之和.

同样, 对于认知不确定性项$\mathbb{V}_3$. 从后验分布中抽取的每组权重$\boldsymbol{w}$保持固定, 根据式 (22) 进行单步预测. 对于N条轨迹, 除加性高斯噪声ε外, 预测是确定性的. 对于N条轨迹 (假设N足够大), 单步预测的期望是从后验分布中采样函数的平均值. 因为高斯噪声的均值为零, 这种期望能够消除随机不确定性. 然后, 对M个后验转移函数的方差进行计算. 在N和M足够大的情况下, 这提供了在策略Π下单次状态转移中认知不确定性的无偏估计值. 在通过转移函数的T个步骤后, 认知不确定性是每一步认知不确定性之和. 最终, 策略Π下成本的不确定性是随机不确定性和认知不确定性之和.

将这些步骤形式化, 便得到一个金标准蒙特卡洛方案, 用于根据式 (38) 估算策略Π下成本中存在的随机不确定性和认知不确定性. 具体而言, 蒙特卡洛估计可通过以下步骤生成: 1) 初始化轨迹: 初始化$M \times N$条轨迹, 每条轨迹的起始状态为$\boldsymbol{x}_0 \sim \mathcal{N}(\boldsymbol{\mu}_0, \Sigma_0)$. 2) 传播轨迹: 将这些轨迹在策略$\Pi$下通过近似 GP 动力学模型传播$T$步. 3) 抽取权重: 从后验分布$q(\boldsymbol{w})$中抽取$M$组权重$\boldsymbol{w}$. 4) 固定权重的滚动模拟: 对于每组权重$\boldsymbol{w}$, 执行$N$次$T$步滚动模拟, 其中对每次转移采样高斯加性噪声. 5) 控制计算: 在每个时间t, 根据动力学模型的需求, 计算确定性的控制动作$\boldsymbol{u}$. 6) 单步预测: 使用式 (22) 计算每次转移的单步预测. 7) 记录成本: 在每个时间步计算成本 (3), 并存储为形状为(M, N)的成本值数组$\mathbb{C}^{\Pi}(\boldsymbol{x}_t)$. 最后, 通过以下方法估计式 (38):

$$\mathbb{V}(\mathbb{C}^{\Pi}(\boldsymbol{x}_t)) = \sum_{t=0}^{T-1}\left\{\mathbb{V}_N\left(\frac{\mathbb{C}^{\Pi}(\boldsymbol{x}_t)\mathbf{1}}{N}\right) + \frac{\mathbf{1}^{\mathrm{T}}\mathbb{V}_M(\mathbb{C}^{\Pi}(\boldsymbol{x}_t))}{M}\right\}. \tag{39}$$

其中: $\mathbf{1}$是一个全为 1 的列向量, $\mathbb{V}_N$和$\mathbb{V}_M$分别是对应于形状为(M, N)的成本值$\mathbb{C}^{\Pi}(\boldsymbol{x}_t)$的行和列计算得到的经验方差.

本方案被称为金标准蒙特卡洛, 因其通过双重采样架构严格实现总方差分解 (式 (38)), 是量化分离不确定性的无偏估计器. 相较普通蒙特卡洛的混合采样, 金标准蒙特卡洛需模型权重和轨迹噪声双重收敛. 金标准属性源于: 1) Bishop[15] 证明的估计无偏性; 2) 与真实 GP 后验的等价采样 (文献 [12]).

2　仿真分析

所有仿真实验均在 Matlab 2023b/Simscape Multibody 虚拟环境中运行 (硬件环境 Apple M2, 64 GB RAM). 物理模型通过刚体动力学模块构建, 控制输入作用于系统质心, 仿真步长 5 ms (求解器 ode45). 训练时间对比见表 1, 可见所提方法在量化认知不确定性的同时仍保持显著速度优势, 所提金标准蒙特卡洛方案通过双重采样并行化提高了计算效率.

表1　单次迭代训练时间对比　　单位: s

方法	欠驱动机械臂	推车双摆
本文方法	38.2 ± 1.5	126.7 ± 3.8
原PILCO[6]	52.4 ± 2.1	205.3 ± 7.2

2.1　成本函数不确定性对 PILCO 学习效果影响分析

在本节中采用两种非线性系统对第 1 节研究内容进行仿真分析. 每次 PILCO 执行完成后, 都会在 PILCO 结束前的所有观测数据上训练三角基函数 GP 模型 (第 1.2 节), 其中输入是状态-动作元组, 目标是状态差异. 在所有情况下, 均使用 500 个基函数. 蒙特卡洛轨迹展开根据算法 1 在策略Π下执行 ($M = 100, N = 100, T = T_E$), 其中$T_E$和过渡噪声$\varepsilon$是特定于环境的. 然后, 按照第 1.3 节中详述的程序, 将总不确定性分解为整个事件在策略Π下成本$\mathbb{C}(\boldsymbol{x}_t)$的总随机不确定性和总认知不确定性.

2.1.1　欠驱动机械臂

欠驱动机械臂 (Pendubot), 如图 1 所示, 由内摆和外摆组成, 质量分别为m_1和m_2, 长度分别为l_1和l_2. 这两个摆相互连接, 且内摆与地面相连. 内摆角和外摆角分别为θ_2和θ_3, 从垂直位置逆时针测量. 智能体通过施加力矩$\boldsymbol{u}$于地面与内摆之间的关节来作

用系统, 但无法直接对两个摆之间的连接施加作用力. 系统的目标是通过对中心关节施加力矩, 使两个摆同时摆动到竖直向上的位置并在该位置保持平衡.

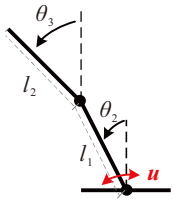


图1 双摆结构示意图

该非线性系统有 4 个连续的状态变量, $\boldsymbol{x}=[\theta_2, \theta_3, \dot{\theta}_2, \dot{\theta}_3]^{\mathrm{T}}$, θ_2和θ_3为摆的角度 (单位 rad); $\dot{\theta}_2$和$\dot{\theta}_3$为摆的角速度 (单位 rad/s), 其微分运动学表达式如下:

$$f(\boldsymbol{x})=\begin{bmatrix} l_1^2\left(\frac{m_1}{4}+m_2\right)+\frac{m_1 l_1^2}{12} & \frac{m_2}{2}l_1 l_2\cos(\theta_2-\theta_3) \\ \frac{m_2}{2}l_1 l_2\cos(\theta_2-\theta_3) & \frac{5m_2 l_2^2}{12} \end{bmatrix},$$

$$g(\boldsymbol{x},\boldsymbol{u})=\begin{bmatrix} gl_1\sin(\theta_2)(0.5m_1+m_2)- \\ 0.5m_2 l_1 l_2\dot{\theta}_3^2\sin(\theta_2-\theta_3)+u \\ 0.5m_2 l_2(l_1\dot{\theta}_2^2\sin(\theta_2-\theta_3)+g\sin(\theta_2)) \end{bmatrix},$$

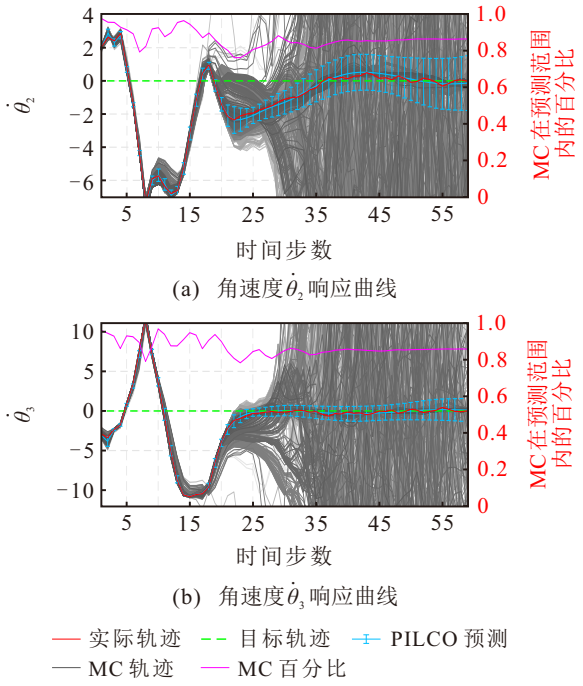
$$\dot{\boldsymbol{x}}=f(\boldsymbol{x})^{-1}g(\boldsymbol{x},\boldsymbol{u}). \tag{40}$$

其中: $l_1=l_2=0.5$ m, $m_1=m_2=0.5$ kg, $g=9.8\mathrm{m/s}^2$. 状态变量的训练目标为$\boldsymbol{x}_{\mathrm{target}}=[0,0,0,0]^{\mathrm{T}}$.

图 2 和图 3 显示了 Pendubot 系统状态变量执行蒙特卡洛轨迹展开的随机样本, 其中图 2 描绘了摆角速度$\dot{\theta}_2$和$\dot{\theta}_3$的相关曲线, 摆角θ_2和θ_3展示在图 3 中. 图 2 和图 3 中: 左边纵轴表示状态变量, 绘制的曲线分别对应状态蒙特卡洛轨迹趋势、实际运行期间状态轨迹、状态变量目标以及事件运行前 PILCO 对状态演化的预测; 右轴表示落在 PILCO 的预测值均方误差内的蒙特卡洛轨迹占整个蒙特卡洛轨迹的百分比. 此外, 图 2 和图 3 中所示的轨迹来自第 30 次训练后的结果, 当时 PILCO 已经实现了系统目标, 因此其结果也代表了成熟策略Π下的轨迹.

由图 2 和图 3 可以看出, 在大约 20 s 的时间长度内, 蒙特卡洛轨迹是向内聚拢的, 此时也对应于 Pendubot 被操纵至目标位置的时期. 在此之后, 蒙特卡洛轨迹开始发散, 对应于智能体试图在目标位置平衡 Pendubot 的时期. 这一仿真结果表明, 在平衡阶段, 策略的不确定性比在上升阶段要大. 尽管蒙特卡洛轨迹存在较大的偏差, 但约 80% 以上的轨迹仍在 PILCO 预测范围的 2 个标准差内.

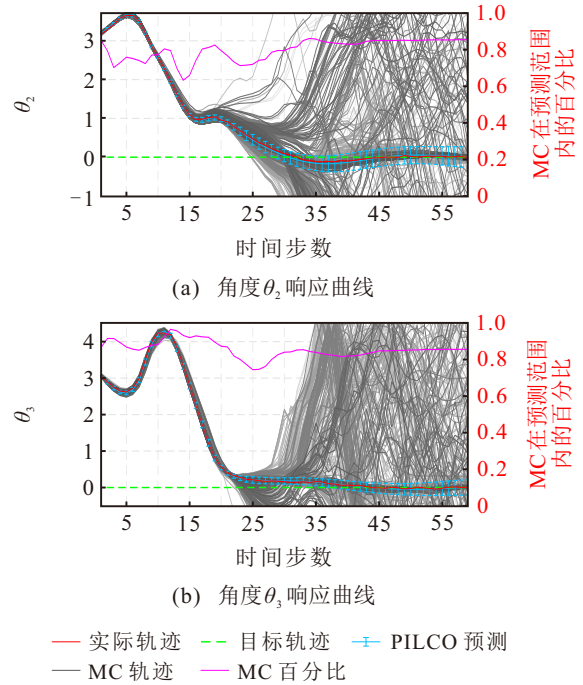
图 4 显示了学习过程中策略Π下每次迭代的平均成本值和成本不确定性分解. 其中, 图 4(a) 显示了



(a) 角速度$\dot{\theta}_2$响应曲线

(b) 角速度$\dot{\theta}_3$响应曲线

图2 Pendubot 系统的摆角速度蒙特卡洛展开



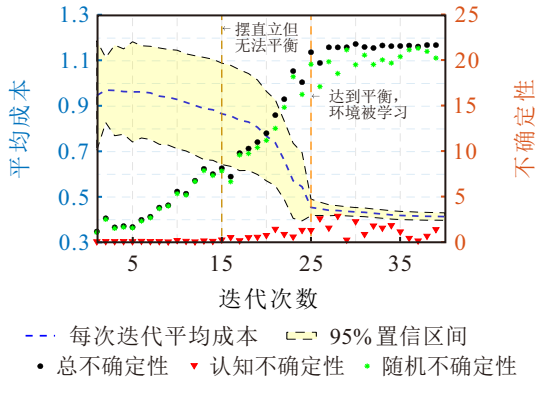
(a) 角度θ_2响应曲线

(b) 角度θ_3响应曲线

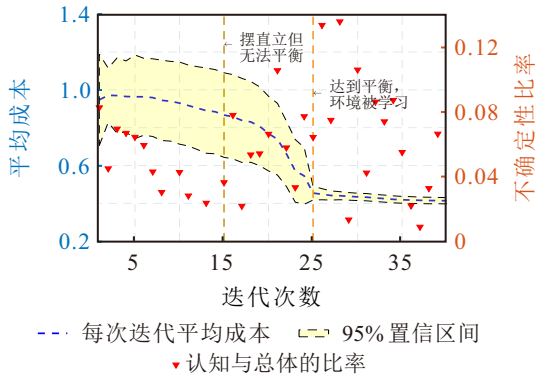
图3 Pendubot 系统的摆角蒙特卡洛展开

完整的不确定性分解, 图 4(b) 显示了认知不确定性占总不确定性的比率. 如图 4(a) 所示, 在前 15 次迭代中, 智能体学会了将双摆摆动到直立位置, 但无法保持平衡. 在此阶段, 平均成本值从其初始值下降, 并且成本中的总不确定性和随机不确定性增加, 而认知成本不确定性保持较低水平. 总成本不确定性的增加是由于智能体对 Pendubot 系统形成了初始认知 (没有认知就没有不确定性). 在接下来至第 25 次的迭代过程中, 智能体学会了保持双摆平衡, 并且在此期间能观察到平均成本下降, 以及随机、认知和总

成本不确定性增加. 超过 25 次迭代以后, 一旦系统被充分认知, 认知成本不确定性会降低, 随机成本不确定性也会随之稳定下来.



(a)　平均成本和不确定性分布



(b)　平均成本和认知不确定性占总不确定性比率

图4　每次迭代的平均成本和不确定性 (双摆)

从图 4(a) 的结果和上述表述中很难看出成本不确定性的细分与学习有何关系, 因此绘制图 4(b) 以进行进一步分析. 在图 4(b) 中, 红色下三角符号表示认知不确定性与总不确定性之比. 前 3 迭代的平均成本约为 0.95, 表明在此期间学习很少. 观察图 4(b) 右纵轴, 在这些迭代过程中, PILCO 选择了认知不确定性占总成本不确定性比率较低的策略. 第 15 次迭代平均成本从最初的 0.95 降低到 0.83, 这一学习步骤所对应的策略具有比前几次迭代更高的认知不确定性占比. 之后至第 25 次迭代, 平均成本有所下降, 均对应于认知不确定性占比较高的策略. 最后, 在第 25 次迭代系统认知得到解决后, 随着智能体完善策略和对模型认知信心的增加, 认知不确定性与总成本不确定性的比率下降, 并维持稳定.

2.1.2　推车双摆

推车双摆结构如图 5 所示, 其由一个质量为 m_1 的推车, 以及质量分别为 m_2 和 m_3、长度分别为 l_2 和 l_3 的内摆和外摆组成. 两个摆锤连接在一起, 且内部摆锤连接在推车上. 内摆角和外摆角分别由 θ_2 和 θ_3 给出. 控制输入 $\boldsymbol{u}$ 是作用于推车质量块 m_1 的水平方向外力. 该任务的目标是对推车施加水平力, 使双摆系统摆动到垂直位置并保持平衡, 同时将推车定位在系统中心 $x_{\text{car}}=0$ 处. 无约束推车双摆系统表现出复杂的动力学行为, 是一个混沌系统.

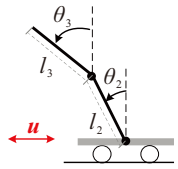


图5　推车双摆结构示意图

该非线性系统有 6 个连续的状态变量, $\boldsymbol{x}=[x_{\text{car}},\dot{x}_{\text{car}},\theta_2,\theta_3,\dot{\theta}_2,\dot{\theta}_3]^{\mathrm{T}}$, 分别为推车的位置 x_{car}、推车的速度 $\dot{x}_{\text{car}}$、双摆的角度 θ_2 和 θ_3(单位 rad)、双摆的角速度 $\dot{\theta}_2$ 和 $\dot{\theta}_3$(单位 rad/s). 微分运动学表达式如下:

$$
f(\boldsymbol{x})=\begin{bmatrix} 2(m_1+m_2+m_3) & -(m_2+2m_3)l_2\cos\theta_2 & -m_3l_3\cos\theta_3 \\ -(3m_2+6m_3)\cos\theta_2 & (2m_2+6m_3)l_2 & 3m_3l_3\cos(\theta_2-\theta_3) \\ -3\cos\theta_3 & 3l_2\cos(\theta_2-\theta_3) & 2l_3 \end{bmatrix}\begin{matrix}\rightarrow\\ \\ \leftarrow\end{matrix},
$$

$$
g(\boldsymbol{x},\boldsymbol{u})=\begin{bmatrix} 2u-2b\dot{x}_{\text{car}}-(m_2+2m_3)l_2\dot{\theta}_2^2\sin\theta_2-m_3l_3\dot{\theta}_3^2\sin\theta_3 \\ (3m_2+6m_3)g\sin\theta_2-3m_3l_3\dot{\theta}_3^2\sin(\theta_2-\theta_3) \\ 3l_2\dot{\theta}_2^2\sin(\theta_2-\theta_3)+3g\sin\theta_3 \end{bmatrix},
$$

$$
\dot{\boldsymbol{x}}=f(\boldsymbol{x})^{-1}g(\boldsymbol{x},\boldsymbol{u}).
$$

其中: $l_2=l_3=0.6\,\text{m}$, $m_1=m_2=m_3=0.5\,\text{kg}$, $b=0.1\,\text{N}\cdot\text{s/m}$, $g=9.8\,\text{m/s}^2$. 状态变量的训练目标为 $\boldsymbol{x}_{\text{target}}=[0,0,0,0,0,0]^{\mathrm{T}}$.

图 6 ～ 图 8 显示了第 40 次迭代结果下 (代表了成熟策略 Π 下的结果), 推车双摆环境状态变量的蒙特卡洛轨迹展开的随机样本. 其中: 图 6 展示了推车位置 x_{car} 和速度 $\dot{x}_{\text{car}}$, 图 7 描述了双摆角速度 $\dot{\theta}_2$ 和 $\dot{\theta}_3$, 双摆角 θ_2 和 θ_3 被描述在图 8 中. 在这 3 幅图中: 左垂直轴均表示状态变量, 绘制的曲线分别对应状态蒙特卡洛轨迹趋势、实际运行期间状态轨迹、状态变量目标以及事件运行前 PILCO 对状态演化的预测; 右轴表示落在 PILCO 预测的蒙特卡洛轨迹占整个蒙特卡洛轨迹的百分比.

从图 6 ～ 图 8 可以看出, 在大约 30 s 的时间长度内蒙特卡洛轨迹具有内聚性, 此时对应于推车双摆被操纵至目标位置的时期. 在此之后, 蒙特卡洛轨迹开始发散, 对应于智能体试图在目标位置平衡推车双摆的时期. 一直到本次迭代结束, 大约有 60%

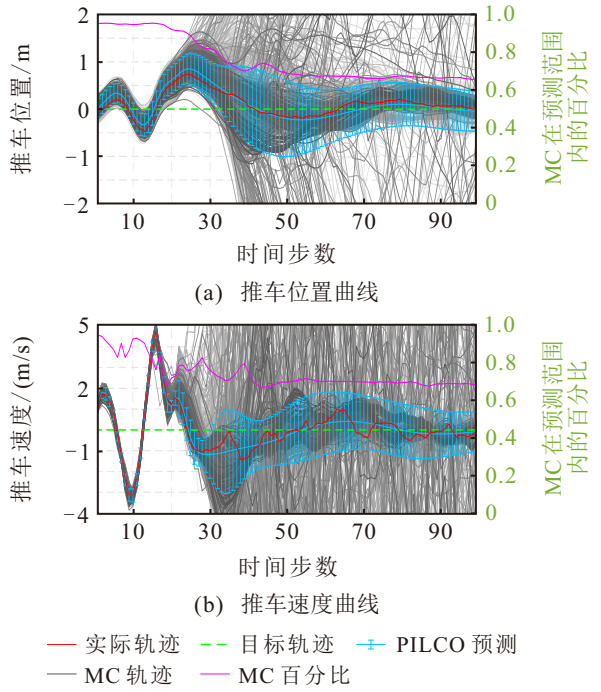


图6 推车位置和速度的蒙特卡洛展开

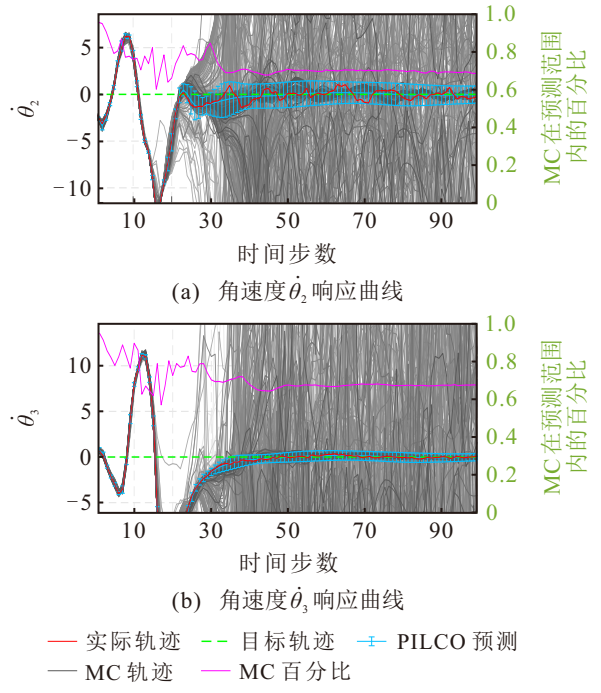


图7 双摆角速度的蒙特卡洛展开

以上的轨迹在 PILCO 预测范围内.

图 9 显示了学习过程中策略Π下的每次迭代产生的平均成本和成本不确定性分解. 图 9(a) 显示了完整的不确定性分解, 图 9(b) 显示了认知不确定性与总不确定性的比率. 如图 9(a) 所示, 在 10 次迭代后, 智能体已经学会将推车双摆到达目标位置, 但还不能将其保持在平衡状态. 在此阶段, 平均成本略有下降. 然后, 智能体再进过 10 次训练后, 可以将推车双摆保持在平衡但偏离目标的位置. 在这一阶段, 随着智能体对系统形成认知, 总成本和认知成本的不

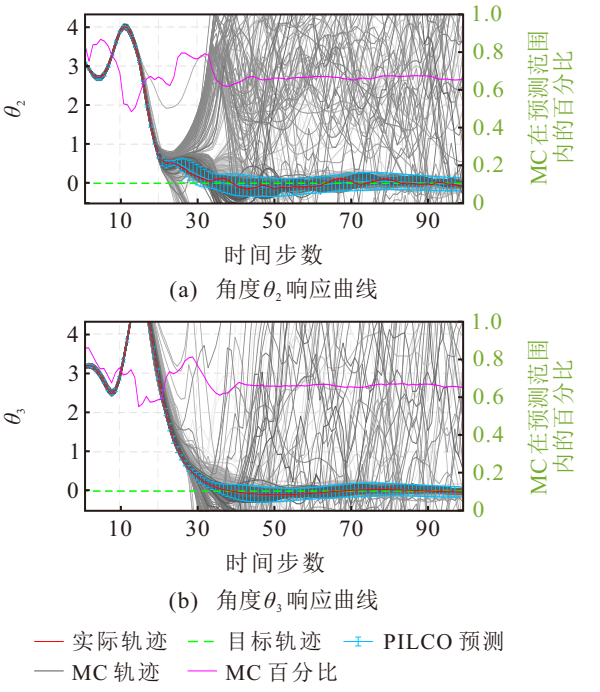


图8 第 40 次迭代下双摆角的蒙特卡洛展开

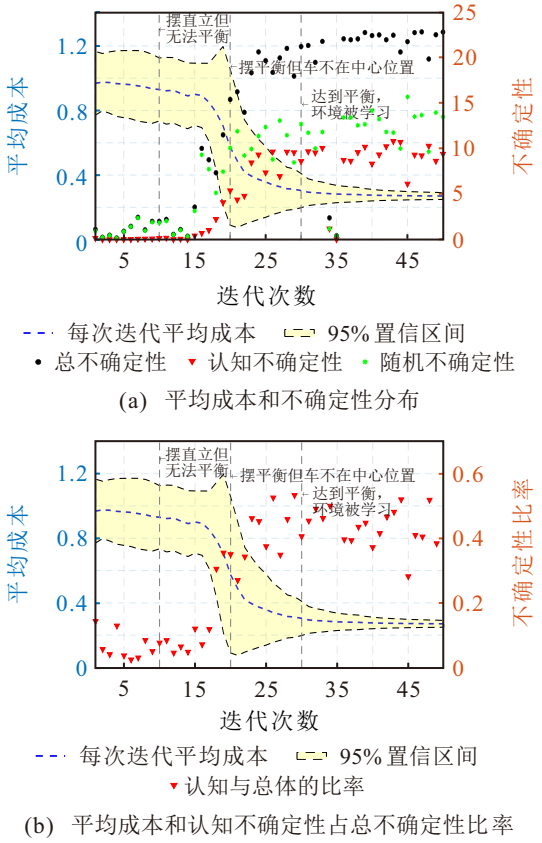


图9 每次迭代的平均成本和不确定性 (推车双摆)

确定性增加; 然而, 低认知成本不确定性相关的策略一直在被选择. 因此, 学习梯度是平缓的. 但是在第 20 次迭代之后, 由于认知成本不确定性的增加, 学习效率迅速增加.

如图 9(b) 所示, 认知成本不确定性与总不确定性之比的趋势更为明显. 接近 20 次迭代时, PILCO 选择认知成本不确定性占比越来越大的策略, 从而

增加了学习效率. 这里, 策略中存在的认知成本不确定性与平均成本之间存在明显的相关性. 这表明, 具有高认知成本不确定性的策略与更陡的学习梯度有关.

上述两个非线性系统的数值仿真结果表明, 在给定策略Π下, 认知不确定性与总成本不确定性之间的比值与 PILCO 的学习速率存在显著相关性. 具有较高认知与总成本不确定性比值的策略会具有更高效的学习, 而比值较低的策略学习速度则较慢.

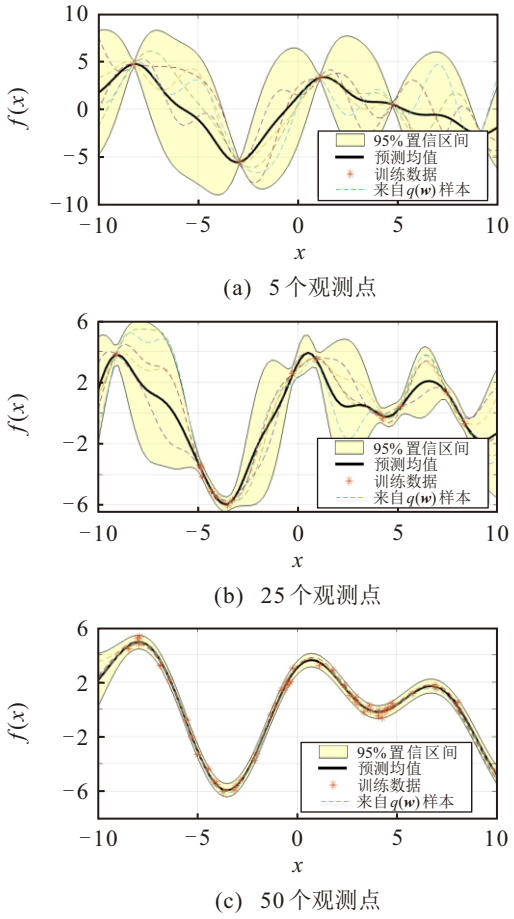
此外, 从图 2 ～ 图 9 可以看出, 当环境被学习达到平衡状态后, 其随机不确定性/总不确定性会升高并且趋于稳定, 相应平衡阶段的蒙特卡洛轨迹出现发散现象. 这是由于平衡点附近系统雅可比矩阵趋于奇异, 使得注入的探索噪声引发指数级状态漂移, 导致部分轨迹偏离平衡点. 这种偏差是算法主动探索引起的伴随效应, 若完全抑制探索, 将无法学习克服模型不确定性. 根据这一分析, 未来可以将本文算法纳入目标函数, 即通过修改目标函数增加认知不确定性奖励, 以加速平衡区数据收集, 实现平衡阶段的总不确定性回落.

2.2　成本函数不确定性的分解

一个目标噪声方差为 0.0625 的测量数据集被用来说明, 在策略Π下估计成本$\mathbb{C}(\boldsymbol{x})$中的随机性不确定性和认知不确定性的蒙特卡洛方案有效性. 这里使用一维的转移函数, 并未定义显式策略. 该实验可视为只有一个动作可供选择, 智能体每次都选择该动作, 使系统动力学能够自由随时间演化. 在该实验中, 智能体最初观察到 5 个转移数据点, 然后每次增加一个数据点, 直至观察 50 个数据点. 在每次观测时, 基于三角基函数的模型被用来对系统的转移动力学形成认知分布. 图 10 分别显示了在观察 5 个、25 个和 50 个数据点后, 对转移数据的预测分布以及从参数后验分布中抽取的 4 个函数. 在实验中, 刻意将谱点的长度尺度设置为较大的值, 以使模型对数据过拟合 (如图 10(a) 所示).

这样做的原因有两个: 一是为了人为地向模型中引入更多的认知性不确定性, 二是为了更好地模拟高维空间的情况. 在高维空间中, 单个数据点的观察为模型提供的信息比一维情况中更多. 图中结果显示, 随着观察到的数据点增多, 模型对参数的信心逐步增强. 从模型参数的后验分布中抽取的函数逐渐趋近于预测均值, 这表明模型的信心在不断增加.

每当观察到一个新数据点时, 会通过模型执行蒙特卡洛轨迹展开, 其按照第 1.4 节中描述的过程进行. 针对目标$\boldsymbol{x}=[0,0,0,0]^{\mathrm{T}}$计算成本, 并利用式 (4) 将成本的不确定性分解为随机不确定性和认知



(a)　5 个观测点

(b)　25 个观测点

(c)　50 个观测点

图10　不同观测数据点对应的样本预测分布

不确定性两部分. 蒙特卡洛展开的参数设置为$M=100, N=100, T=100$, 转移噪声方差为 0.09. 需要特别注意, 不确定性的来源在于模型内部以及通过模型的转移过程, 但此处是根据成本来衡量的.

图 11 展示了不确定性随数据点数量变化的分解结果. 随着数据点数量的增加, 模型对参数的信心逐步增强, 认知不确定性随之减少. 随着认知不确定性的降低, 总不确定性也相应减少, 而且随机不确定性与总不确定性之间的差距逐渐缩小. 这是因为每次进行轨迹展开时, 均基于相同的策略进行. 如果在每次新数据点被观测后策略发生变化 (这在实际应用中较为常见), 那么在不同策略下的展开会导致不同的不确定性, 2.1 节所示的 PILCO 实验验证了这一点. 最后, 虽然可以预期, 在相同策略下进行轨迹

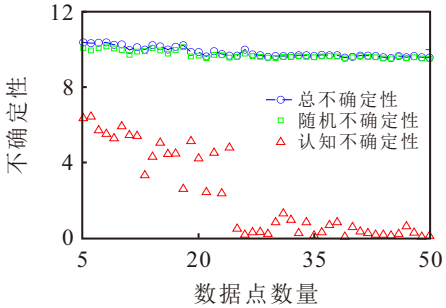


图11　不确定性分布随观测点数据变化

展开时, 随着更多数据的观测, 认知不确定性会单调减少, 但实际观察到的下降呈现一定的噪声波动. 这可能是蒙特卡洛方法的结果, 增加展开次数应该可以减小这种波动.

根据上述分析, 未来的工作可以将第 1.4 节中提出的金标准蒙特卡洛方法融入目标函数, 以在成本和认知与总成本不确定性比值之间建立平衡. 目标函数可以调整为优先寻找具有较高或较低认知与总成本不确定性比值的策略, 从而研究这种调整对学习效率的影响.

3 结 论

本文研究了 PILCO 如何在决策过程中使用不确定性. 为此, 本文中使用方差来度量不确定性, 并使用总方差定律将不确定性分解为随机不确定性和认知不确定性两部分. 此外, 本文还引入了一个金标准蒙特卡洛方案, 通过近似 PILCO 的动力学模型传播轨迹, 分别量化了成本函数中存在的随机和认知不确定性. 分析结果表明, 当 PILCO 处于有效学习时, 它选择的是认知成本不确定性与总成本不确定性的具有高比率的相关策略. 相比之下, 当 PILCO 处于学习缓慢时, PILCO 选择的是认知成本不确定性与总成本不确定性的低比率相关策略.

未来计划将金标准蒙特卡洛不确定量化方法纳入目标函数, 以在随机不确定性和认知不确定性与总不确定性的比率之间建立平衡. 然后, 可以通过可调超参数来调整目标函数, 使学习者能够专注于寻求认知与总成本不确定性比率较高或较低的策略, 并可以检查对学习效率的影响. 可调超参数可以通过类似于贪婪探索机制的方式处理, 调整探索量以在智能体中创建不同的行为特征.

参考文献 (References)

[1] Li X H, Liu Y, Zou S N. Optimal safety tracking control with prescribed performance based on variable barrier function and reinforcement learning[J]. Control and Decision, 2025, 40(3): 803-812.

[2] Guo W, Yao H, Zhang Z Z, et al. Optimization method for reservoir neuron selection based on reinforcement learning[J]. Control and Decision, 2024, 39(9): 2876-2884.

[3] Sutton R S, Barto A G. Reinforcement learning: An introduction[M]. Cambridge: The MIT Press, 2018: 36-70.

[4] Wen G H, Yang T, Zhou J L, et al. Reinforcement learning and adaptive/approximate dynamic programming: A survey from theory to applications in multi-agent systems[J]. Control and Decision, 2023, 38(5): 1200-1230.

[5] Amadio F, Dalla Libera A, Antonello R, et al. Model-based policy search using Monte Carlo gradient estimation with real systems application[J]. IEEE Transactions on Robotics, 2022, 38(6): 3879-3898.

[6] Deisenroth M, Rasmussen C. PILCO: A model-based and data-efficient approach to policy search[C]. Proceedings of the 28th International Conference on International Conference on Machine Learning. Bellevue, 2011: 465-472.

[7] 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.

[8] Thrun S, Möller K. Active exploration in dynamic environments[C]. Proceedings of the 5th International Conference on Neural Information Processing Systems. 1991, Denver: 531–538.

[9] Qian B, Liu D F, Hu R, et al. Hybrid iterated greedy algorithm for just in time distributed permutation flowshop scheduling problem[J]. Control and Decision, 2022, 37(11): 3042-3051.

[10] Wang L, Li Z Z. Adaptive greedy Gaussian segmentation algorithm based on multivariate time series[J]. Control and Decision, 2024, 39(2): 568-576.

[11] Zhang M M, Zhang S, Wu X Y, et al. Efficient reinforcement learning with the novel *N*-step method and *V*-network[J]. IEEE Transactions on Cybernetics, 2024, 54(10): 6048-6057.

[12] Lázaro-Gredilla M, Quiñonero-Candela J, Rasmussen C E, et al. Sparse spectrum Gaussian process regression[J]. Journal of Machine Learning Research, 2010, 11: 1865-1881.

[13] Rasmussen C E, Williams C K I. Gaussian processes for machine learning[M]. Cambridge: The MIT Press, 2005.

[14] Candela J Q, Rasmussen C. A unifying view of sparse approximate Gaussian process regression[J]. The Journal of Machine Learning Research, 2006(6): 1939-1959.

[15] Bishop M C. Pattern recognition and machine learning[M]. Cham: Spring, 2006.

[16] Stein L M. Interpolation of spatial data[M]. Cham: Springer-Verlag, 1999: 24.

[17] Chatfield C. The analysis of time series — An introduction[M]. The 4th edition. London: Chapman and Hall, 1989: 94-95.

[18] Depeweg S, Lobato M J, Doshi-Velez F, et al. Decomposition of uncertainty in bayesian deep learning for efficient and risk-sensitive learning[J]. Machine Learning, 2017: 1710.07283.

[19] Kiureghian D A, Ditlevsen O. Aleatory or epistemic? Does it matter?[J]. Structural Safety, 2009, 31(2): 105-112.

[20] Kendall A, Gal Y. What uncertainties do we need in bayesian deep learning for computer vision?[C]. Proceedings of the 31st International Conference on Neural Information Processing Systems. California, 2017: 5574-5584.

作者简介

曹瑞 (1994–), 女, 讲师, 博士, 主要研究方向为导航、制导与控制, E-mail: stdio@yzu.edu.cn;

吕慧涛 (1994–), 男, 工程师, 博士, 主要研究方向为飞行动力学与控制, E-mail: lht@nuaa.edu.cn.