基于参数探索的期望最大化策略搜索

程玉虎 [1] 冯涣婷 [1] 王雪松 [1]

摘 要 针对随机探索易于导致梯度估计方差过大的问题, 提出一种基于参数探索的期望最大化 (Expectation-maximization, EM) 策略搜索方法. 首先, 将策略定义为控制器参数的一个概率分布. 然后, 根据定义的概率分布直接在控制器参数空间进行多次采样以收集样本. 在每一幕样本的收集过程中, 由于选择的动作均是确定的, 因此可以减小采样带来的方差, 从而减小梯度估计方差. 最后, 基于收集到的样本, 通过最大化期望回报函数的下界来迭代地更新策略参数. 为减少采样耗时和降低采样成本, 此处利用重要采样技术以重复使用策略更新过程中收集的样本. 两个连续空间控制问题的仿真结果表明, 与基于动作随机探索的策略搜索强化学习方法相比, 本文所提方法不仅学到的策略最优, 而且加快了算法收敛速度, 具有较好的学习性能.

关键词 策略搜索, 强化学习, 参数空间, 探索, 期望最大化, 重要采样

DOI 10.3724/SP.J.1004.2012.00038

Expectation-maximization Policy Search with Parameter-based Exploration

CHENG Yu-Hu[1] FENG Huan-Ting[1] WANG Xue-Song[1]

Abstract In order to reduce large variance of gradient estimation resulted from stochastic exploration strategy, a kind of expectation-maximization policy search reinforcement learning with parameter-based exploration is proposed. At first, a probability distribution over the parameters of a controller is used to define a policy. Secondly, samples are collected by directly sampling in the controller parameter space according to the probability distribution for several times. During the sample-collection procedure of each episode, because the selected actions are deterministic, sampling from the defined policy leads to a small variance in the samples, which can reduce the variance of gradient estimation. At last, based on the collected samples, policy parameters are iteratively updated by maximizing the lower bound of the expected return function. In order to reduce the time-consumption and to lower the cost of sampling, an importance sampling technique is used to repeatedly use samples collected from policy update process. Simulation results on two continuous-space control problems illustrate that the proposed policy search method can not only obtain the most optimal policy but also improve the convergence speed as compared with several policy search reinforcement learning methods with action-based stochastic exploration, thus has a better learning performance.

Key words Policy search, reinforcement learning, parameter space, exploration, expectation-maximization (EM), importance sampling

作为一类求解模型未知的 Markov 决策问题的有效方法, 强化学习已在城市交通信号控制[1]、车间调度[2] 和机器人技术[3−6] 等领域有了成功的应用. 目前, 用于处理连续状态和动作空间问题的策略搜索强化学习方法主要有基于梯度的方法[7] (例如策略梯度方法、自然策略梯度方法、有限差分梯度等) 和期望最大化 (Expectation-maximization, EM) 策略搜索方法两种. 在利用基于梯度的方法更新策略参数时, 如果参数在每次更新中的调整量较大, 则将导致学习系统的期望回报减小. 为了保证系统获得的期望回报是单调增大的, 通常的做法是通过设置较小的学习率参数来减小策略参数的调整量, 但是, 过小的学习率参数设置又会导致系统的在线学习速度较为缓慢, 影响了算法的学习性能. EM 算法是一种求参数极大似然估计的方法, 它可以从非完整数据集中对参数进行极大似然估计, 是一种非常简单实用的学习算法. 为了能够增大策略参数在每次更新中的调整量, Dayan 等[8] 将强化学习问题映射为一种极大似然概率密度估计问题, 并利用 EM 算法估计该概率密度, 提出了一种 EM 强化学习框架. 随后, 在 EM 强化学习基础上, Peters 等[9] 将强化学习简化为一种加权回报非线性回归问题, 提出一种基于加权回报回归的 EM 策略搜索算法, 并将其用于解决机械臂的关节控制问题. 与基于梯度的策略搜索强化学习方法相比, EM 策略搜索不

收稿日期 2011-05-24 录用日期 2011-08-30
Manuscript received May 24, 2011; accepted August 30, 2011
国家自然科学基金 (60804022, 60974050, 61072094), 教育部新世纪优秀人才支持计划 (NCET-08-0836, NCET-10-0765), 霍英东教育基金会青年教师基金 (121066) 资助
Supported by National Natural Science Foundation of China (60804022, 60974050, 61072094), Program for New Century Excellent Talents in University (NCET-08-0836, NCET-10-0765), and Fok Ying-Tung Education Foundation for Young Teachers (121066)
本文责任编委 方海涛
Recommended by Associate Editor FANG Hai-Tao
1. 中国矿业大学信息与电气工程学院 徐州 221116
1. School of Information and Electrical Engineering, China University of Mining and Technology, Xuzhou 221116

仅可以避免学习率参数的调节问题, 而且可以加快算法的学习速度.

不同于可以获得教师信号的监督学习, 强化学习需要利用 "试错法" 来发现哪个动作能带来更大的回报. 探索作为强化学习的一个重要组成部分, 影响智能体尝试学习的次数和学习结果的质量[10]. 目前, 大部分策略搜索强化学习方法都依赖于一个随机探索策略, 即在每一个时间步给确定性动作添加一个高斯噪声扰动, 然后利用似然比方法估计梯度, 例如 REINFORCE[11], 但是, 该方法估计梯度的方差过大, 导致算法收敛速度过慢. 这主要是因为当收集一幕 (Episode) 样本时, 智能体在每一个时间步都需要对动作进行随机扰动, 从而给梯度估计过程引入噪声, 且随着幕的长度增长梯度估计的方差也相应地增大. 为此, 针对随机探索容易导致梯度估计方差过大的问题, Rückstieß 等[12] 提出状态相关探索 (State-dependent exploration, SDE) 方法, 在每一个时间步给动作添加一个与状态有关的探索, 于是在一幕样本中, 对于任意给定的状态都返回相同的动作, 从而降低了每一幕样本的方差. 结合 SDE 方法, Peters 等将 EM 策略搜索方法从立即回报强化学习形式[8−9] 扩展到批量式强化学习中[13], 提出一种关于回报加权探索的策略学习方法. 但是, SDE 方法要求所学的控制器 (表示为一个参数化的函数) 关于其参数必须具有可微性, 对于复杂控制器问题来说, 求解控制器关于其参数的导数非常困难. 为此, 从探索的另一个角度出发, 直接对控制器参数进行探索来减小梯度估计方差, Sehnke 等[14] 提出了基于参数探索的策略梯度 (Policy gradients with parameter-based exploration, PGPE) 方法, 并证明了该方法的性能优于基于动作随机探索的强化学习方法. 与 SDE 方法相比, PGPE 方法不需要求解控制器关于其参数的导数, 对所学的控制器没有任何限制, 可以处理任意非可微控制器问题.

鉴于 PGPE 的特性, 为减小梯度估计方差以及提高算法的学习收敛速度, 不同于目前已有的研究工作, 本文尝试将直接在参数空间进行探索的策略应用到 EM 策略搜索方法中, 提出一种基于参数探索的 EM 策略搜索算法. 学习过程中, 智能体首先在每一幕样本收集前根据控制器参数的概率分布对参数进行采样; 然后, 根据定义的控制器选择动作, 这样在每一幕样本中选择的动作都是确定的; 最后, 根据策略更新过程中收集的样本, 利用重要采样技术来更新策略参数.

1　策略搜索强化学习

1.1　强化学习框架

在强化学习中, 往往把问题建模为 Markov 决策过程 (Markov decision process, MDP)[15]. 在离散时间 t, 智能体观察状态 $\boldsymbol{s}_t \in S$, 根据一个随机策略 $\pi(a_t|\boldsymbol{s}_t,\boldsymbol{\theta})$ 选择动作 $a_t \in A$, 然后转移到状态 $\boldsymbol{s}_{t+1} \sim p(\boldsymbol{s}_{t+1}|\boldsymbol{s}_t,a_t)$, 并接收一个立即回报 $r_t(\boldsymbol{s}_t,a_t,\boldsymbol{s}_{t+1})$. 随机策略 $\pi(a_t|\boldsymbol{s}_t,\boldsymbol{\theta})$ 表示在给定状态 $\boldsymbol{s}_t$ 和策略参数 $\boldsymbol{\theta}$ 时, 选择动作 a_t 的概率. $p(\boldsymbol{s}_{t+1}|\boldsymbol{s}_t,a_t)$ 表示转移到下一个状态 $\boldsymbol{s}_{t+1}$ 的概率, 且仅依赖于前一时刻的状态 –动作对 $(\boldsymbol{s}_t,a_t)$. 智能体与环境反复交互收集状态、动作和回报样本, 记为一幕 $h=\{\boldsymbol{s}_1,a_1,r_1,\boldsymbol{s}_2,\cdots,\boldsymbol{s}_T,a_T,r_T\}$, 其中, T 为幕的长度.

设 $R(h)=\sum_{t=1}^{T}\gamma^{t-1}r_t(\boldsymbol{s}_t,a_t,\boldsymbol{s}_{t+1})$ 表示一幕的折扣累积回报, 其中, $\gamma\in[0,1]$ 为折扣因子, $p(h|\boldsymbol{\theta})$ 为智能体根据初始状态概率 $p(\boldsymbol{s}_1)$, 状态转移概率 $p(\boldsymbol{s}_{t+1}|\boldsymbol{s}_t,a_t)$ 和随机策略 $\pi(a_t|\boldsymbol{s}_t,\boldsymbol{\theta})$ 得到一幕 h 的概率, 有 $p(h|\boldsymbol{\theta})=p(\boldsymbol{s}_1)\times\prod_{t=1}^{T}p(\boldsymbol{s}_{t+1}|\boldsymbol{s}_t,a_t)\pi(a_t|\boldsymbol{s}_t,\boldsymbol{\theta})$. 在以上设置下, 智能体的期望回报 $J(\boldsymbol{\theta})$ 表示为

$$J(\boldsymbol{\theta})=\mathrm{E}\left\{R(h)\right\}=\int_H p(h|\boldsymbol{\theta})R(h)\mathrm{d}h \tag{1}$$

强化学习的目标是学习一个最优策略参数 $\boldsymbol{\theta}^*$ 以最大化期望回报 $J(\boldsymbol{\theta})$:

$$\boldsymbol{\theta}^*=\arg\max_{\boldsymbol{\theta}} J(\boldsymbol{\theta}) \tag{2}$$

1.2　EM 策略搜索学习

一般来说, $J(\boldsymbol{\theta})$ 具有较强的非线性, 通过直接最大化 $J(\boldsymbol{\theta})$ 来求解最优策略参数 $\boldsymbol{\theta}^*$ 比较困难. 为此, 借鉴监督学习优化目标函数下界的思想, EM 策略搜索方法通过最大化目标函数的下界来迭代地更新策略参数[16]. 假设 $\boldsymbol{\theta}_l$ 是当前的策略参数, 其中, l 表示迭代次数. 首先, 根据 Jensen 不等式可以得到归一化期望回报 $\log(J(\boldsymbol{\theta})/J(\boldsymbol{\theta}_l))$ 的下界:

$$\begin{aligned}\log\frac{J(\boldsymbol{\theta})}{J(\boldsymbol{\theta}_l)}=&\log\int_H\frac{p(h|\boldsymbol{\theta})R(h)}{J(\boldsymbol{\theta}_l)}\mathrm{d}h=\\&\log\int_H\frac{p(h|\boldsymbol{\theta}_l)R(h)}{J(\boldsymbol{\theta}_l)}\frac{p(h|\boldsymbol{\theta})}{p(h|\boldsymbol{\theta}_l)}\mathrm{d}h\geq\\&\int_H\frac{p(h|\boldsymbol{\theta}_l)R(h)}{J(\boldsymbol{\theta}_l)}\log\frac{p(h|\boldsymbol{\theta})}{p(h|\boldsymbol{\theta}_l)}\mathrm{d}h=J_l(\boldsymbol{\theta})\end{aligned} \tag{3}$$

上式在应用 Jensen 不等式时, 将 $\frac{p(h|\boldsymbol{\theta}_l)R(h)}{J(\boldsymbol{\theta}_l)}$ 看作为一个概率密度函数, 因此, 此处 $R(h)$ 必须具有非负性. 然后, 在策略改进阶段通过最大化下界 $J_l(\boldsymbol{\theta})$ 来迭代地更新参数 $\boldsymbol{\theta}$, 有:

$$\boldsymbol{\theta}_{l+1}=\arg\max_{\boldsymbol{\theta}} J_l(\boldsymbol{\theta}) \tag{4}$$

在求解式 (4) 时, 下界 $J_l(\boldsymbol{\theta})$ 的最大化参数 $\boldsymbol{\theta}_{l+1}$ 需要满足方程:

$$\frac{\partial J_l(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}}\Big|_{\boldsymbol{\theta}=\boldsymbol{\theta}_{l+1}} = \int_H \frac{p(h|\boldsymbol{\theta}_l)R(h)}{J(\boldsymbol{\theta}_l)}\frac{\partial \log p(h|\boldsymbol{\theta})}{\partial \boldsymbol{\theta}}\Big|_{\boldsymbol{\theta}=\boldsymbol{\theta}_{l+1}} \mathrm{d}h = \int_H \frac{p(h|\boldsymbol{\theta}_l)R(h)}{J(\boldsymbol{\theta}_l)}\sum_{t=1}^{T}\frac{\partial \log \pi(a_t|\boldsymbol{s}_t,\boldsymbol{\theta})}{\partial \boldsymbol{\theta}}\Big|_{\boldsymbol{\theta}=\boldsymbol{\theta}_{l+1}} \mathrm{d}h = 0 \tag{5}$$

从式 (5) 可以看出, 计算参数 $\boldsymbol{\theta}_{l+1}$ 需要事先已知 $\pi(a_t|\boldsymbol{s}_t,\boldsymbol{\theta})$ 的模型. 目前, 在强化学习方法中, 策略模型都定义为一个参数化函数逼近器, 其输出表示执行不同动作的概率. 假设一个控制器输出的控制量 (动作) 表示为 $a = f(\boldsymbol{s},\boldsymbol{\omega}) + \xi$, 其中, $f(\boldsymbol{s},\boldsymbol{\omega})$ 表示一个带有参数向量 $\boldsymbol{\omega} \in \mathbf{R}^d$ 的控制器, ξ 为一个随机探索因子, 且 $\xi \sim \mathrm{N}(0,\sigma^2)$, 则策略 $\pi(a_t|\boldsymbol{s}_t,\boldsymbol{\theta})$ 的模型可定义为

$$\pi(a_t|\boldsymbol{s}_t,\boldsymbol{\theta}) = \frac{1}{\sqrt{2\pi}\sigma}\exp(-\frac{(a_t - f(\boldsymbol{s}_t,\boldsymbol{\omega}))^2}{2\sigma^2}) \tag{6}$$

其中, 策略参数 $\boldsymbol{\theta} = (\boldsymbol{\omega}^{\mathrm{T}},\sigma)^{\mathrm{T}}$. 根据式 (6) 和采样方法即可求解方程式 (5). 但是, 智能体在每一时间步均需根据式 (6) 选择动作, 这样每一时间步动作的采样都会给一幕样本引入噪声, 从而给梯度 $\frac{\partial J_l(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}}$ 的估计带来高方差, 导致参数 $\boldsymbol{\theta}_{l+1}$ 的计算不准确. 另外, 在更新参数时需要计算 $\frac{\partial f(\boldsymbol{s},\boldsymbol{\omega})}{\partial \boldsymbol{\omega}}$, 即要求控制器 $f(\boldsymbol{s},\boldsymbol{\omega})$ 关于参数 $\boldsymbol{\omega}$ 必须具有可微性.

2 基于参数探索的 EM 策略搜索学习

正如引言部分所描述, 策略搜索强化学习方法中梯度估计方差过大的原因在于采用随机探索策略对每一时间步的动作进行采样. 为了解决随机探索带来的高方差问题, Sehnke 等[14] 在策略梯度方法中提出了一种新的探索策略 — 基于参数探索的策略梯度方法. 为此, 借鉴文献 [14] 的思想, 针对式 (6) 会给梯度 $\frac{\partial J_l(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}}$ 的估计引入高方差问题, 此处采用参数探索策略选择动作, 则策略 $\pi(a_t|\boldsymbol{s}_t,\boldsymbol{\theta})$ 的模型定义为控制器参数的概率分布:

$$\pi(a_t|\boldsymbol{s}_t,\boldsymbol{\theta}) = \int_\Omega p(\boldsymbol{\omega}|\boldsymbol{\rho})\delta_{f(\boldsymbol{s}_t,\boldsymbol{\omega}),a_t}\mathrm{d}\boldsymbol{\omega} \tag{7}$$

其中, $\boldsymbol{\theta} = \boldsymbol{\rho}$, $\boldsymbol{\rho}$ 是确定 $\boldsymbol{\omega}$ 分布的参数, $f(\boldsymbol{s}_t,\boldsymbol{\omega})$ 表示一个由控制器选择的确定性动作, δ 是狄拉克函数. 在每一幕样本收集的开始阶段, 智能体首先根据概率分布函数 $p(\boldsymbol{\omega}|\boldsymbol{\rho})$ 选择参数 $\boldsymbol{\omega}$, 然后根据状态转移概率和式 (7) 采样, 这样每一幕样本的采样仅依赖于一个控制器参数 $\boldsymbol{\omega}$, 从而在每一个时间步中选择的动作都是确定的. 与式 (6) 定义的策略相比, 采用式 (7) 定义的策略学习一方面可以减小每一幕样本收集过程中带来的方差, 从而相应地减小梯度 $\frac{\partial J_l(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}}$ 估计的方差; 另一方面, 在估计 $\frac{\partial J_l(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}}$ 时, 不需要计算控制器关于其参数的导数 $\frac{\partial f(\boldsymbol{s},\boldsymbol{\omega})}{\partial \boldsymbol{\omega}}$, 因此, 该方法可以应用于非可微控制器问题中.

由以上分析可知, 当根据式 (7) 收集一幕样本 h 时, 给定参数 $\boldsymbol{\omega}$, h 条件独立于参数 $\boldsymbol{\theta}$, 因此 $p(h,\boldsymbol{\omega}|\boldsymbol{\theta}) = p(h|\boldsymbol{\omega})p(\boldsymbol{\omega}|\boldsymbol{\theta})$, 则式 (1) 改写为

$$J(\boldsymbol{\theta}) = \mathrm{E}\left\{R(h)\right\} = \int_\Omega\int_H p(h,\boldsymbol{\omega}|\boldsymbol{\theta})R(h)\mathrm{d}h\mathrm{d}\boldsymbol{\omega} \tag{8}$$

将式 (8) 代入式 (3) 中, 并应用 Jensen 不等式, 则归一化期望回报 $\log(J(\boldsymbol{\theta})/J(\boldsymbol{\theta}_l))$ 的下界可改写成如下形式:

$$\log\frac{J(\boldsymbol{\theta})}{J(\boldsymbol{\theta}_l)} = \log\int_\Omega\int_H \frac{p(h,\boldsymbol{\omega}|\boldsymbol{\theta})R(h)}{J(\boldsymbol{\theta}_l)}\mathrm{d}h\mathrm{d}\boldsymbol{\omega} = \log\int_\Omega\int_H \frac{p(h|\boldsymbol{\omega})R(h)}{J(\boldsymbol{\theta}_l)}\mathrm{d}h p(\boldsymbol{\omega}|\boldsymbol{\theta}_l)\frac{p(\boldsymbol{\omega}|\boldsymbol{\theta})}{p(\boldsymbol{\omega}|\boldsymbol{\theta}_l)}\mathrm{d}\boldsymbol{\omega} = \log\int_\Omega\int_H \frac{p(h|\boldsymbol{\omega})p(\boldsymbol{\omega}|\boldsymbol{\theta}_l)R(h)}{J(\boldsymbol{\theta}_l)}\mathrm{d}h\frac{p(\boldsymbol{\omega}|\boldsymbol{\theta})}{p(\boldsymbol{\omega}|\boldsymbol{\theta}_l)}\mathrm{d}\boldsymbol{\omega} \geq \int_\Omega\int_H \frac{p(h|\boldsymbol{\omega})p(\boldsymbol{\omega}|\boldsymbol{\theta}_l)R(h)}{J(\boldsymbol{\theta}_l)}\mathrm{d}h\log\frac{p(\boldsymbol{\omega}|\boldsymbol{\theta})}{p(\boldsymbol{\omega}|\boldsymbol{\theta}_l)}\mathrm{d}\boldsymbol{\omega} = J_l'(\boldsymbol{\theta}) \tag{9}$$

于是, 式 (4) 表示的参数 $\boldsymbol{\theta}$ 的更新规则可写为

$$\boldsymbol{\theta}_{l+1} = \arg\max_{\boldsymbol{\theta}} J_l'(\boldsymbol{\theta}) \tag{10}$$

类似于式 (4) 的求解过程, 此处根据 Monte-Carlo 采样方法求解式 (10). 假设在第 l 次策略迭代学习中, 智能体首先根据 $p(\boldsymbol{\omega}|\boldsymbol{\theta}_l)$ 选择 N 个参数 $\{\boldsymbol{\omega}^n\}_{n=1}^N$, 然后根据 $p(h|\boldsymbol{\omega}^n)$ 收集 N 幕训练样本 $D^l = \left\{h_n^l\right\}_{n=1}^N$, 利用训练样本 D^l 即可估计 $J_l'(\boldsymbol{\theta})$ 关于参数 $\boldsymbol{\theta}$ 的梯度:

$$\frac{\partial J_l'(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}}\Big|_{\boldsymbol{\theta}=\boldsymbol{\theta}_{l+1}} \approx \frac{1}{N}\sum_{n=1}^{N} R(h_n^l)\frac{\partial \log p(\boldsymbol{\omega}|\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} = 0 \tag{11}$$

假设参数 $\boldsymbol{\theta}$ 由一些均值 $\{\mu_i\}$ 和标准差 $\{\sigma_i\}$ 组成, $\boldsymbol{\omega}$ 的各个分量 ω_i 之间是相互独立的, 且 $\omega_i \sim \mathrm{N}(\mu_i,\sigma_i^2)$, 则 $\log p(\boldsymbol{\omega}|\boldsymbol{\theta})$ 关于 μ_i 和 σ_i 的导数可以通过解析法得出:

$$\begin{cases} \dfrac{\partial \log p(\boldsymbol{\omega}|\boldsymbol{\theta})}{\partial \mu_i} = \dfrac{\omega_i - \mu_i}{\sigma_i^2} \\ \dfrac{\partial \log p(\boldsymbol{\omega}|\boldsymbol{\theta})}{\partial \sigma_i} = \dfrac{(\omega_i - \mu_i)^2 - \sigma_i^2}{\sigma_i^3} \end{cases} \tag{12}$$

将式 (12) 代入方程式 (11) 中, 即可得出参数 $\boldsymbol{\theta}_{l+1} = (\{\mu_i^{l+1}\}_{i=1}^d, \{\sigma_i^{l+1}\}_{i=1}^d)^{\mathrm{T}}$ 的估计值:

$$\begin{cases} \hat{\mu}_i^{l+1} = (\sum\limits_{n=1}^{N} R(h_n^l))^{-1} \times (\sum\limits_{n=1}^{N} R(h_n^l)\omega_i^{l,n}) \\ (\hat{\sigma}_i^2)^{l+1} = (\sum\limits_{n=1}^{N} R(h_n^l))^{-1} \times \\ \qquad (\sum\limits_{n=1}^{N} R(h_n^l)(\omega_i^{l,n} - \hat{\mu}_i^{l+1})^2) \end{cases} \tag{13}$$

由采样方法可知, 当 $N \to \infty$ 时, 估计值 $\hat{\boldsymbol{\theta}}_{l+1}$ 收敛于真实值 $\boldsymbol{\theta}_{l+1}$. 因此, 为了降低在策略迭代学习中收集的样本数目 N, 本文考虑采用离策略学习方法, 根据重要采样技术[16] 重复使用以前策略迭代学习过程中收集的样本, 可以得到式 (13) 的一致估计:

$$\begin{cases} \hat{\mu}_i^{l+1,\mathrm{IW}} = (\sum\limits_{l'=1}^{l}\sum\limits_{n=1}^{N} R(h_n^{l'}) w_{l,l'}(h_n^{l'}))^{-1} \times \\ \qquad (\sum\limits_{l'=1}^{l}\sum\limits_{n=1}^{N} R(h_n^{l'}) w_{l,l'}(h_n^{l'}) \omega_i^{l',n}) \\ (\hat{\sigma}_i^2)^{l+1,\mathrm{IW}} = (\sum\limits_{l'=1}^{l}\sum\limits_{n=1}^{N} R(h_n^{l'}) w_{l,l'}(h_n^{l'}))^{-1} \times \\ \qquad (\sum\limits_{l'=1}^{l}\sum\limits_{n=1}^{N} R(h_n^{l'}) w_{l,l'}(h_n^{l'}) (\omega_i^{l',n} - \hat{\mu}_i^{l+1,\mathrm{IW}})^2) \end{cases} \tag{14}$$

其中, $w_{l,l'}(h) = \dfrac{p(h,\boldsymbol{\omega}|\boldsymbol{\theta}_l)}{p(h,\boldsymbol{\omega}|\boldsymbol{\theta}_{l'})} = \dfrac{p(\boldsymbol{\omega}|\boldsymbol{\theta}_l)}{p(\boldsymbol{\omega}|\boldsymbol{\theta}_{l'})}$ 表示重要权重 (Importance weight, IW).

3　算法步骤

根据上述分析, 给出基于参数探索的 EM 策略搜索强化学习算法的计算步骤如下:

步骤 1. 初始化参数, 包括策略参数 $\boldsymbol{\theta} = (\boldsymbol{\mu}^{\mathrm{T}}, \boldsymbol{\sigma}^{\mathrm{T}})^{\mathrm{T}}$, 折扣因子 γ, 策略学习精度 ε, 初始迭代次数 $l = 1$ 和最大迭代次数 L.

步骤 2. 首先, 根据 $\boldsymbol{\omega} \sim \mathrm{N}(\boldsymbol{\mu}, I\boldsymbol{\sigma}^2)$ 选择 N 个参数 $\{\boldsymbol{\omega}^n\}_{n=1}^N$; 然后, 根据初始状态概率 $p(\boldsymbol{s}_1)$, 状态转移概率 $p(\boldsymbol{s}_{t+1}|\boldsymbol{s}_t, a_t)$ 和式 (7) 定义的策略 $\pi(a_t|\boldsymbol{s}_t, \boldsymbol{\theta})$ 收集 N 幕样本, 且每幕样本里有 T 步转移, 记为 $D^l \equiv \{h_n^l\}_{n=1}^N$.

步骤 3. 计算重要权重, 并利用收集的样本 $\{D^{l'}\}_{l'=1}^l$, 按照式 (14) 更新策略参数 $\boldsymbol{\theta}_l$.

步骤 4. 若两个连续迭代步的策略参数变化小于指定的精度 ε, 即 $\|\boldsymbol{\theta}_{l+1} - \boldsymbol{\theta}_l\| < \varepsilon$, 或者 $l > L$, 则终止算法; 否则, 令 $l = l + 1$, 返回步骤 2.

4　仿真研究

为了验证本文所提基于参数探索的 EM 策略搜索算法 (Expectation-maximization policy search with parameter-based exploration, EMPE) 的有效性, 对具有连续状态和动作空间的两个控制任务分别进行仿真研究. 首先, 在一个简单的小球平衡学习[16] 问题上比较基于动作随机探索的 EM 策略搜索 (Expectation-maximization policy search with action-based stochastic exploration, EMASE)[16] 和 EMPE, 以分析转移步长对算法性能的影响. 然后, 针对具有四维连续状态空间和一维连续动作空间的倒立摆平衡问题[17], 比较 EMASE、EMPE、Vanilla 策略梯度 ("Vanilla" policy gradient, VPG)[5] 和批量式自然策略梯度 (Episodic natural actor-critic, eNAC)[5,18] 四种方法的性能.

4.1　小球平衡问题

小球平衡模型如图 1 所示. 智能体的目标是控制平板的角度, 使小球保持在平板的中间. 系统的一维连续动作空间由平板的角度 $\varphi \in (-45°, 45°)$ 组成, 状态空间由小球到平板中间的距离 $x\,(\mathrm{m})$ 和小球的速度 $\dot{x}\,(\mathrm{m/s})$ 两个连续变量组成. 变量 x 和 $\dot{x}$ 使用如下方程来描述:

$$\begin{cases} x_{t+1} = x_t + \dot{x}_{t+1}\Delta t \\ \dot{x}_{t+1} = \dot{x}_t + \Delta t(-\dfrac{q}{m}\dot{x}_t - 9.8\sin\varphi_t) \end{cases} \tag{15}$$

其中, $q = 0.5$ 为摩擦系数, $m = 3\,\mathrm{kg}$ 为小球的质量, φ_t 是智能体在 t 时刻选择的动作, $\Delta t = 0.05\,\mathrm{s}$ 为模拟的时间间隔. 立即回报 $r(\boldsymbol{s}_t, a_t, \boldsymbol{s}_{t+1})$ 定义为二次型函数:

$$r(\boldsymbol{s}_t, a_t, \boldsymbol{s}_{t+1}) = -\boldsymbol{s}_{t+1}^{\mathrm{T}} \begin{bmatrix} 1 & 0 \\ 0 & 0.5 \end{bmatrix} \boldsymbol{s}_{t+1} - 0.1a_t^2 \tag{16}$$

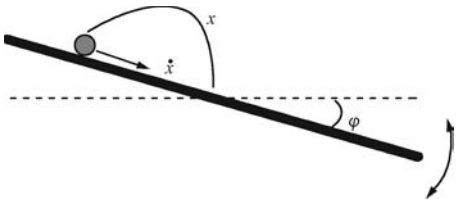


图 1　小球平衡模型

Fig. 1　Illustration of ball-balancing model

当小球停在平板中间 ($x = 0, \dot{x} = 0$) 时, 智能体得到的立即回报最大 (为 0). 根据上述二次型回报计算得到的 $R(h)$ 为负值, 这与在 EM 算法推导过程中 $R(h)$ 的非负性假设相矛盾 (参见第 1.2 节). 因此,

为了处理该问题, 在运行 EM 策略搜索算法时, 将 $R(h)$ 转换为 $R'(h) = -\frac{1}{10^{-5}+R(h)}$. 系统学习的控制器为 $f(\boldsymbol{s},\boldsymbol{\omega}) = \boldsymbol{s}^{\mathrm{T}} \cdot \boldsymbol{\omega}$, 其中 $\boldsymbol{\omega} \in \mathbf{R}^2$.

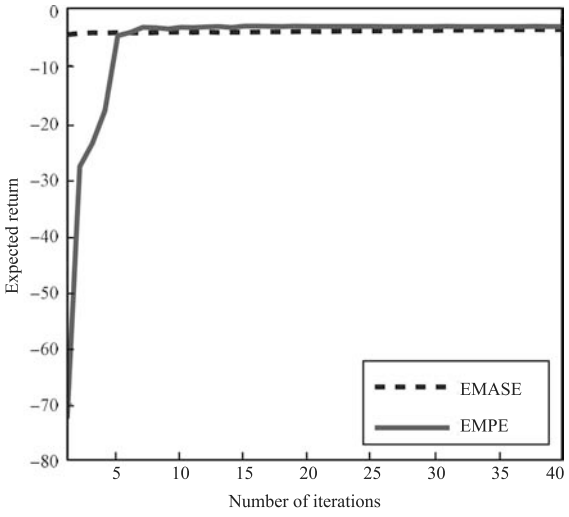
仿真过程中, 小球的初始状态取为 $x = -0.5$ 和 $\dot{x} = 0$, EMASE 和 EMPE 两种算法分别从随机初始策略开始学习, 最大迭代次数设为 $L = 40$, 在每次迭代学习中收集的训练样本数为 $N = 20$, $\varepsilon = 10^{-5}$, $\gamma = 0.95$. 为了验证文中所提方法是否能够有效地减小策略学习中的方差以及提高算法的学习速度, 分别将 EMASE 和 EMPE 算法使用不同转移步长 (即一幕的长度) 的训练样本进行学习, 两种算法均独立运行 30 次. 在策略的每次迭代学习中, 智能体根据测试样本 (其中, $N = 50$, $T = 150$) 评估学到的策略, 算法的性能由期望回报衡量, 图 2 和表 1 分别给出了相应的仿真结果.

图 2 (a) ∼ 2 (c) 分别为算法在转移步长为 $T = 20$, $T = 50$ 和 $T = 100$ 的训练样本下独立学习 30 次后期望回报的平均结果, 表 1 统计了算法在不同转移步长设置下学到的最大期望回报以及在最大迭代次数限制下达到收敛所需的迭代步数. 从图 2 和表 1 可以看出, 随着转移步长的增大, EMASE 算法学习的收敛速度有所提高, 但是, 学到的期望回报随之减小. 这是因为随着一幕样本转移步长的增大, 对动作进行随机探索的次数随之增多, 样本中的方差也相应地增大, 该方差超过了学习一个较优策略所需的样本数, 从而给策略学习过程中带来更大的方差, 此处也验证了在每一个时间步对动作进行随机探索是导致梯度估计方差过大的主要原因. 而在每一个时间步, 根据 EMPE 方法选择得到的动作均是确定的, 即使增大一幕样本的转移步长, 样本中的方差也不会发生改变, 因此, 转移步长越大, EMPE 算法的性能越优越. 以上分析说明了与 EMASE 算法相比, 利用本文所提方法不但能学到更优的策略, 而且能加快算法的学习速度.

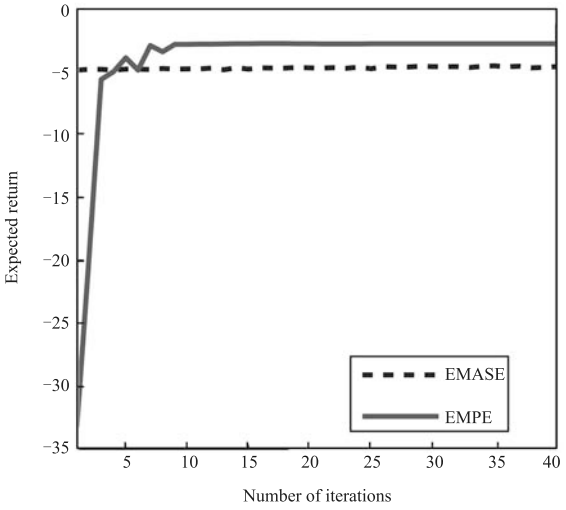
表 1 两种算法学到的最大期望回报和收敛性能比较

Table 1 Comparison of the largest expected returns and convergence performances learned by the two algorithms

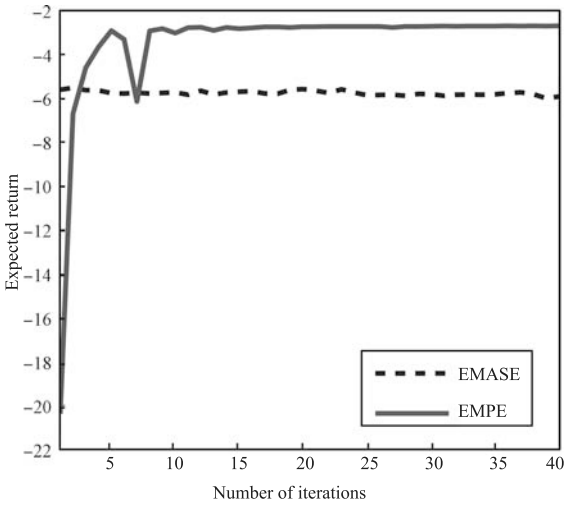
算法名称	转移步长 T	最大期望回报	在 40 次迭代内达到收敛的步数		
			最小步数	最大步数	平均步数
EMASE	20	−3.6437	40	40	40
	50	−4.6624	7	40	32
	100	−5.9176	5	40	24
EMPE	20	−3.1153	8	40	33
	50	−2.8472	8	40	29
	100	−2.6954	8	40	26



(a) $T = 20$



(b) $T = 50$



(c) $T = 100$

图 2 期望回报曲线

Fig. 2 Expected return curves

4.2 倒立摆平衡问题

图 3 给出了倒立摆系统平衡控制问题的示意图. 智能体控制系统的目标是通过移动车, 使摆杆和车能够以最长的时间分别稳定在 $[-12°, 12°]$ 和 $[-2.4\,\mathrm{m}, 2.4\,\mathrm{m}]$ 区域内. 系统的输入由摆杆的角度 α, 角速度 $\dot{\alpha}$, 车的位置 x 和运动速度 $\dot{x}$ 四个连续变量组成. 仿真中, 系统的非线性动力学特性由如下方程描述:

$$\begin{cases} \ddot{x} = \dfrac{f - m_{\mathrm{p}} l_{\mathrm{p}} (\ddot{\alpha} \cos\alpha - \dot{\alpha}^2 \sin\alpha)}{m_{\mathrm{c}} + m_{\mathrm{p}}} \\ \ddot{\alpha} = \dfrac{g \sin\alpha (m_{\mathrm{c}} + m_{\mathrm{p}}) - (f + m_{\mathrm{p}} l_{\mathrm{p}} \dot{\alpha}^2 \sin\alpha) \cos\alpha}{\dfrac{4}{3} l_{\mathrm{p}} (m_{\mathrm{c}} + m_{\mathrm{p}}) - m_{\mathrm{p}} l_{\mathrm{p}} \cos^2\alpha} \end{cases} \tag{17}$$

其中, $l_{\mathrm{p}} = 0.5\,\mathrm{m}$ 为摆杆的长度, $m_{\mathrm{c}} = 1.0\,\mathrm{kg}$ 为车的质量, $m_{\mathrm{p}} = 0.1\,\mathrm{kg}$ 为摆杆的质量, $g = 9.8\,\mathrm{m/s^2}$ 为重力常量. 仿真过程中, 采用四阶龙格 –库塔方法模拟动态系统, 模拟时间步设为 $0.02\,\mathrm{s}$, 摆杆的初始角度和车的初始位置分别设为 $-0.001\,\mathrm{rad}$ 和 $0.001\,\mathrm{m}$. 当摆杆的角度 $\alpha \notin [-12°, 12°]$ 或者车的位置 $x \notin [-2.4\,\mathrm{m}, 2.4\,\mathrm{m}]$ 时, 系统失败, 一幕运行停止, 智能体得到的立即回报定义为 $-2 \times (T - t)$, 其中, T 表示一幕样本的最大长度 (在此仿真中, 对于训练取 $T = 300$, 对于测试取 $T = 500$), t 表示运行的时间步. 在其余情况下, 智能体得到的立即回报为 0. 由于利用以上回报计算的 $R(h)$ 为负值, 因此, 在运行 EM 策略搜索算法时, 将 $R(h)$ 转换为 $R'(h) = -\frac{1}{10^{-8} + R(h)}$. 采用一个线性参数化策略表示控制器 $f(\boldsymbol{s}, \boldsymbol{\omega}) = \boldsymbol{s}^{\mathrm{T}} \cdot \boldsymbol{\omega}$, 其中 $\boldsymbol{\omega} \in \mathbf{R}^4$.

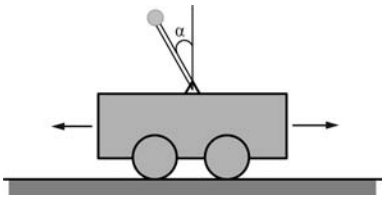


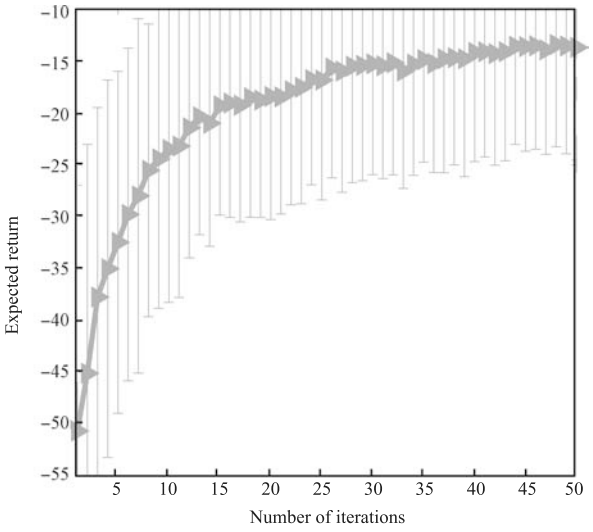
图 3　倒立摆系统模型

Fig. 3　Illustration of inverted pendulum model

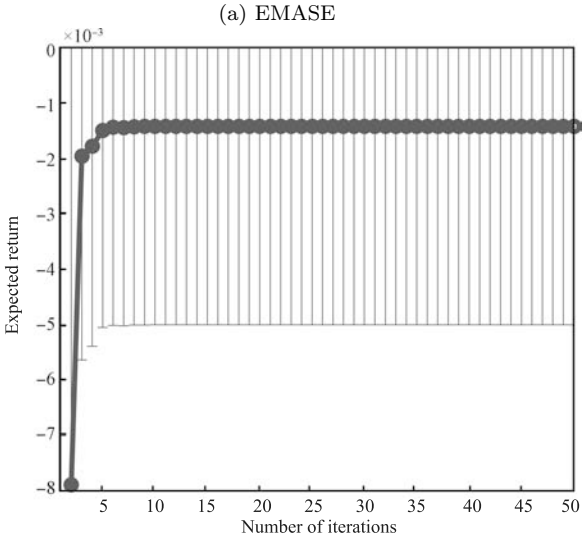
仿真中, 四种算法的最大迭代次数均设为 $L = 50$, 在每次迭代中收集的样本数设为 $N = 50$, 终止条件 $\varepsilon = 10^{-5}$, $\gamma = 0.95$. 在 VPG 和 eNAC 学习中, 根据经验法设置两种算法的参数, 其中, 探索参数分别设为 $\sigma_{\mathrm{VPG}} = 0.25$, $\sigma_{\mathrm{eNAC}} = 0.0125$, 学习率参数分别设为 $\beta_{\mathrm{VPG}} = 0.2$, $\beta_{\mathrm{eNAC}} = 2.0$. 针对本文所提算法, 初始探索参数设为 $\sigma_{\mathrm{init}} = 2.0$.

算法的性能由期望回报和倒立摆系统平衡的时间步数来评价. 仿真过程中, 在每一次迭代学习中更新算法相应的策略, 然后智能体根据测试样本评估学到的策略, 从起始状态开始运行, 当运行时间步数超过设定值 (仿真中设为 500) 或者达到系统失败的条件, 一次评估结束. 由于策略具有随机性, 以上评估过程重复进行 50 次, 以准确地估计期望回报和倒立摆系统在策略作用下平衡的时间步数. 四种算法分别从随机初始策略开始重复运行 30 次. 图 4 给出了四种算法独立运行 30 次后期望回报的平均结果, 表 2 显示的是四种算法学到的最大期望回报以及在最大迭代次数限制下达到收敛所需的迭代步数. 从图 4 和表 2 可以看出, 本文所提的 EMPE 算法不仅学到的期望回报最大, 而且在每次独立运行中都能以很快的速度达到收敛, 而其他三种算法在 50 次迭代中几乎都不能达到收敛, 学习速度比较慢, 学到的期望回报也较小.

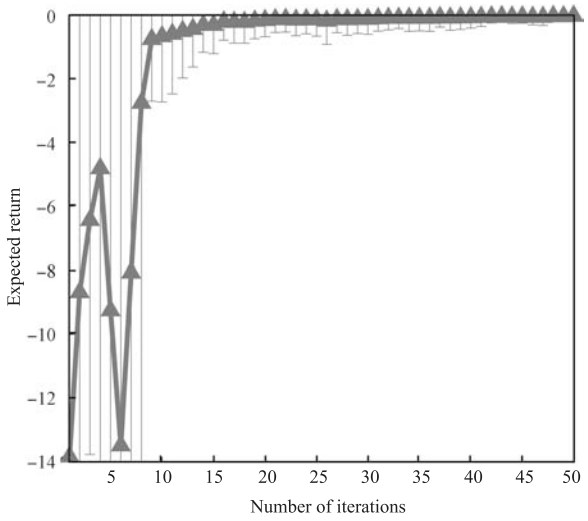
图 5 为倒立摆系统根据四种算法更新的策略在 30 次独立仿真中平衡时间步数的平均结果, 平衡时间步数越长, 说明算法的性能越好. 由图可知, EMPE



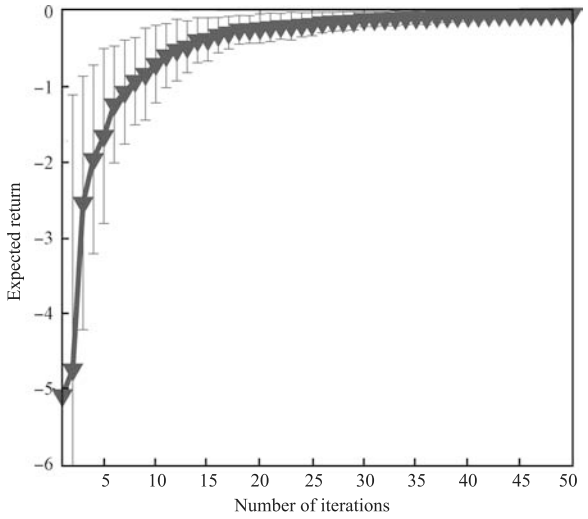
(a) EMASE



(b) EMPE



(c) VPG



(d) eNAC

图 4 期望回报曲线

Fig. 4 Expected return curves

表 2 四种算法学到的最大期望回报和收敛性能比较

Table 2 Comparison of the largest expected returns and convergence performances learned by the four algorithms

算法	最大期望回报	在 50 次迭代内达到收敛的步数		
		最小步数	最大步数	平均步数
EMASE	−13.4050	28	50	49
EMPE	**−0.0014**	**3**	**16**	**7**
VPG	−0.0486	50	50	50
eNAC	−0.0524	50	50	50

方法平衡的平均时间步数为 347, 而 VPG, eNAC 和 EMASE 方法平衡的平均时间步数分别为 215, 200 和 97. 图 6 给出了在学习过程中采用本文所提 EMPE 方法获得的最好和最差策略的性能, 从图中可以看出, 倒立摆系统在 5 次迭代后就能学到最好的策略, 且平衡的时间步数为 500, 而最差策略大约在 3 次迭代后学到, 平衡的时间步数为 208. 进一步说明了本文所提方法的性能优于其他三种算法.

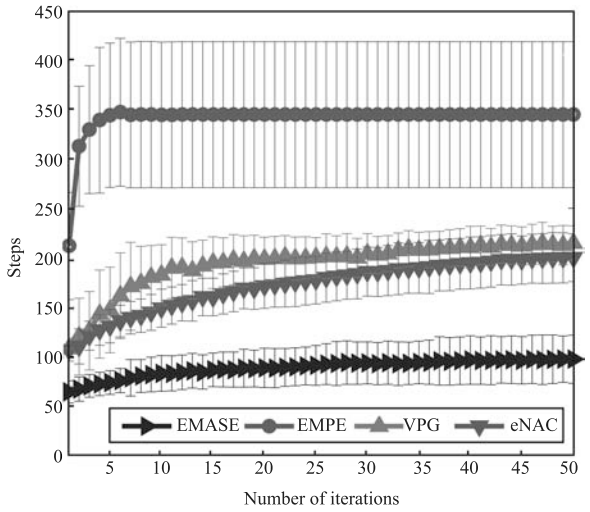


图 5 倒立摆系统平衡的时间步数

Fig. 5 Balancing time-steps of inverted pendulum system

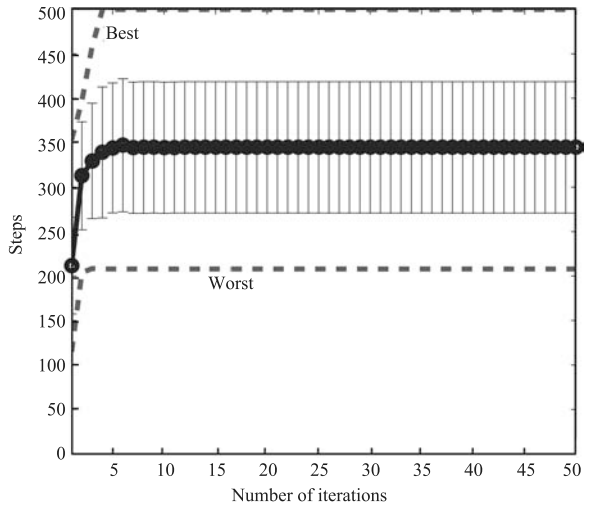


图 6 EMPE 算法学到的最好和最差策略

Fig. 6 The best and the worst policies learned by EMPE

5 结论

策略搜索强化学习方法能够有效地处理连续空间的控制问题, 但是, 在策略学习过程中采用的随机探索策略会导致算法学习速度过慢, 影响了策略搜索强化学习方法的进一步广泛应用. 为此, 从探索方面出发, 提出了一种基于参数探索的 EM 策略搜索算法. 该算法直接在策略的参数空间探索, 在收集的每一幕样本中选择的动作仅由一个事先采样好的控制器参数确定, 很大程度上减小了样本收集过程中带来的方差. 与三种基于动作随机探索的策略搜索

强化学习方法相比, 小球平衡和倒立摆系统两个控制问题的仿真结果表明, 本文所提算法不仅能提高算法的学习速度, 而且能够学到更优的策略, 且随着转移步长的增长, 其学习性能更为优越.

References

1 Zhao Dong-Bin, Liu De-Rong, Yi Jian-Qiang. An overview on the adaptive dynamic programming based urban city traffic signal optimal control. *Acta Automatica Sinica*, 2009, **35**(6): 676−681
(赵冬斌, 刘德荣, 易建强. 基于自适应动态规划的城市交通信号优化控制方法综述. 自动化学报, 2009, **35**(6): 676−681)

2 Zhang W, Dietterich T G. Value function approximation and job-shop scheduling. In: Proceedings of the Workshop on Value Function Approximation, Report Number CMU-CS-95-206, School of Computer Science, Carnegie-Mellon University, USA, 1995

3 Sugiyama M, Hachiya H, Towell C, Vijayakumar S. Value function approximation on non-linear manifolds for robot motor control. In: Proceedings of the IEEE International Conference on Robotics and Automation. Rome, Italy: IEEE, 2007. 1733−1740

4 Barto A G, Sutton R S, Anderson C W. Neuronlike adaptive elements that can solve difficult learning control problems. *IEEE Transactions on System, Man and Cybernetics*, 1983, **13**(5): 834−846

5 Peters J, Schaal S. Policy gradient methods for robotics. In: Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems. Beijing, China: IEEE, 2006. 2219−2225

6 Cheng Yu-Hu, Feng Huan-Ting, Wang Xue-Song. Policy iteration reinforcement learning based on geodesic Gaussian basis defined on state-action graph. *Acta Automatica Sinica*, 2011, **37**(1): 44−51
(程玉虎, 冯涣婷, 王雪松. 基于状态−动作图测地高斯基的策略迭代强化学习. 自动化学报, 2011, **37**(1): 44−51)

7 Wang Xue-Ning, Chen Wei, Zhang Meng, Xu Xin, He Han-Gen. A survey of direct policy search methods in reinforcement learning. *CAAI Transactions on Intelligent Systems*, 2007, **2**(1): 16−24
(王学宁, 陈伟, 张锰, 徐昕, 贺汉根. 增强学习中的直接策略搜索方法综述. 智能系统学报, 2007, **2**(1): 16−24)

8 Dayan P, Hinton G E. Using expectation-maximization for reinforcement learning. *Neural Computation*, 1997, **9**(2): 271−278

9 Peters J, Schaal S. Reinforcement learning by reward-weighted regression for operational space control. In: Proceedings of the 24th International Conference on Machine Learning. Corvallis, USA: ACM, 2007. 745−750

10 Wang Xue-Song, Tian Xi-Lan, Cheng Yu-Hu, Yi Jian-Qiang. Q-learning system based on cooperative least squares support vector machine. *Acta Automatica Sinica*, 2009, **35**(2): 214−219
(王雪松, 田西兰, 程玉虎, 易建强. 基于协同最小二乘支持向量机的 Q 学习. 自动化学报, 2009, **35**(2): 214−219)

11 Williams R J. Simple statistical gradient-following algorithms for connectionist reinforcement learning. *Machine Learning*, 1992, **8**(3−4): 229−256

12 Rückstieß T, Felder M, Schmidhuber J. State-dependent exploration for policy gradient methods. In: Proceedings of the European Conference on Machine Learning and Knowledge Discovery in Databases. Antwerp, Belgium: Springer, 2008. 234−249

13 Peters J, Kober J. Using reward-weighted imitation for robot reinforcement learning. In: Proceedings of the IEEE Symposium on Adaptive Dynamic Programming and Reinforcement Learning. Nashville, USA: IEEE, 2009. 226−232

14 Sehnke F, Osendorfer C, Rückstieß T, Graves A, Peters J, Schmidhuber J. Parameter-exploring policy gradients. *Neural Networks*, 2010, **23**(4): 551−559

15 Tang Hao, Wan Hai-Feng, Han Jiang-Hong, Zhou Lei. Coordinated look-ahead control of multiple CSPS system by multi-agent reinforcement learning. *Acta Automatica Sinica*, 2010, **36**(2): 289−296
(唐昊, 万海峰, 韩江洪, 周雷. 基于多 Agent 强化学习的多站点 CSPS 系统的协作 Look-ahead 控制. 自动化学报, 2010, **36**(2): 289−296)

16 Hachiya H, Peters J, Sugiyama M. Efficient sample reuse in EM-based policy search. In: Proceedings of the European Conference on Machine Learning and Knowledge Discovery in Databases. Bled, Slovenia: Springer, 2009. 469−484

17 Riedmiller M, Peters J, Schaal S. Evaluation of policy gradient methods and variants on the cart-pole benchmark. In: Proceedings of the IEEE Symposium on Approximate Dynamic Programming and Reinforcement Learning. Honolulu, USA: IEEE, 2007. 254−261

18 Peters J, Vijayakumar S, Schaal S. Natural actor-critic. In: Proceedings of the 16th European Conference on Machine Learning. Porto, Portugal: Springer, 2005. 280−291

程玉虎 中国矿业大学教授. 主要研究方向为机器学习, 智能优化与控制. 本文通信作者.
E-mail: chengyuhu@163.com
(**CHENG Yu-Hu** Professor at China University of Mining and Technology. His research interest covers machine learning, intelligent optimization and control. Corresponding author of this paper.)

冯涣婷 中国矿业大学硕士研究生. 主要研究方向为强化学习.
E-mail: fhtcumt@163.com
(**FENG Huan-Ting** Master student at China University of Mining and Technology. Her main research interest is reinforcement learning.)

王雪松 中国矿业大学教授. 主要研究方向为机器学习, 生物信息学.
E-mail: wangxuesongcumt@163.com
(**WANG Xue-Song** Professor at China University of Mining and Technology. Her research interest covers machine learning and bioinformatics.)