文章编号:1671-6833(2022)02-0015-07

基于多估计器平均值的深度确定性策略梯度算法

李 琳[1,2], 李玉泽[1], 张钰嘉[1], 魏 巍[1,2]

(1.山西大学 计算机与信息技术学院,山西 太原 030006; 2.山西大学 计算智能与中文信息处理教育部重点实验室,山西 太原 030006)

摘 要: 为了解决强化学习行动者-评论家框架下双延迟深度确定性策略梯度算法的低估计问题,提出了一种基于多估计器平均值的深度确定性策略梯度(DDPG-MME)算法。基于多估计器平均值的确定性策略梯度算法包含一个行动者和 k(k > 3)个评论家,该算法首先计算 2 个评论家输出值的最小值和剩余(k-2)个评论家输出值的平均值,再取两者的平均值作为最终值来计算 TD 误差,最后根据 TD 误差来更新评论家网络,行动者网络则根据第 1 个评论家输出的值进行更新。DDPG-MME 算法的加权操作缓解了双延迟深度确定性策略梯度算法的低估计问题,并在一定程度上降低了估计方差,实现了更准确的 Q 值估计。在理论上对基于多估计器平均值的确定性策略梯度算法、深度确定性策略梯度算法和双延迟深度确定性策略梯度算法估值误差的期望和方差进行分析,证明了所提算法估值的准确性和稳定性。在 Reacher-v2、HalfCheetah-v2、InvertedPendulum-v2 和 InvertedDoublePendulum-v2 4 个MuJoCo 连续控制环境下对算法的性能进行测试,结果表明:在与对比算法相同的超参数(网络结构、奖励函数、环境参数、批次大小、学习率、优化器和折扣系数)设置下,所提算法的最终性能和稳定性均显著优于对比算法。

关键词: 强化学习; 行动者-评论家; 低估计; 多估计器; 策略梯度

中图分类号: TP391 **文献标志码**: A **doi**:10.13705/j.issn.1671-6833.2022.02.013

0 引言

强化学习是机器学习中比较特殊的一类,区别于传统的监督学习和无监督学习,强化学习没有预先给出标签,而是通过和环境交互获得的奖励来判断自己决策的优劣程度[1-3],进而获得最优策略。

Q 学习(Q-learning)是一种经典的基于值的强化学习算法,可用于解决马尔科夫决策过程问题(MDPs)[4]。然而,Q-learning 估计 Q 值时存在高估问题。针对高估问题,一些方法试图通过求平均值来直接减小每个时间步的偏差[5-6],或是使用平滑的方式减少偏差[7]。Hasselt[8] 在 2015 年提出了双重 Q 学习(double Q-learning),该方法在一定程度上缓解了过估计问题。

DQN(deep Q-Network)[9-10] 作为代表性的深度强化学习算法,由 Mnih 等提出。DQN 算法在多个游戏中的表现已超越了人类玩家,但仍存在样本利用率低的问题。针对这些不足,一些研究者提出了优先经验回放[11]、竞争深度 Q 网络[12]和权重平均值的深度双 Q 网络[13]等算法。同时,因 DQN 采用了与 Q 学习类似的更新机制,同样存在高估问题。针对 DQN 高估问题,Van HASSELT 等[14] 在 2015 年提出了 Double DQN。与 Double Q-learning 的思想类似,Double DQN 使用 2 个估计器,在计算目标 Q 值时同时利用主网络和目标网络来计算 Q 值,达到了避免 DQN 中 Q 值高估的目的,使得 Q 值的估计更加接近真实值,并在 Atari 游戏上的表现超越了传统的 DQN 算法。

然而,基于值函数的强化学习方法存在着不能应用到连续动作空间的问题,研究者们提出了 AC(actor critic)框架[15]。AC 框架分离了行动策略和评估策略,能够更快输出连续动作,但这也只是解决了连续动作空间的问题。Lillicrap 等[16]、Silver 等[17] 在确定性策略的基础上,结合 AC 框

收稿日期:2021-10-09;**修订日期**:2021-11-20

基金项目:国家自然科学基金资助项目(61772323);山西省自然科学基金资助项目(201801D221165);山西省高校科技创新项目(2019L0057)

作者简介:李琳(1986—),女,山西运城人,山西大学讲师,博士,主要从事强化学习、车联网、无人驾驶研究,E-mail:lilynn1116@ sxu.edu.cn。

架提出了深度确定性策略梯度(deep deterministic policy gradient, DDPG)算法,DDPG 算法最终将状态空间和动作空间都拓展到了连续空间。与 DQN 最大化操作类似,DDPG 算法基于梯度上升来更新估计器,因此 DDPG 估计器的估值也存在过估计问题。为此,Fujimoto 等[18]提出了双延迟深度确定性策略梯度(twin delayed deep deterministic policy gradient, TD3)算法,该算法采用了 2 个估计器网络,通过选取一对估计器网络输出的较小的 Q 值作为目标值更新估计器网络参数,在一定程度上缓解了 DDPG 过估计问题。相比 DDPG 算法,TD3 算法估计出的 Q 值更加接近真实 Q 值,并在多个游戏中的表现优于 DDPG 算法。

然而,上述算法都是在解决强化学习算法的 Q 值高估问题。TD3 算法在使用双估计器缓解高估问题的同时,由于其选取 2 个估计器输出的较小的 Q 值作为目标更新的 Q 值,导致 TD3 算法估计的 Q 值存在低估现象。这种低估使得估计器对 Q 值的估值低于真实值,而累积下来的低估会导致次优策略的发生,进而使算法性能下降。并且,TD3 算法 Q 值的估计方差较大,估计值的波动幅度较大,这也影响到了算法的性能和稳定性。

据此,本文提出了一种基于多估计器平均值的深度确定性策略梯度(deep deterministic policy gradient based on mean of multiple estimators, DDPG-MME)算法,该算法通过平均加权来缓解上述的低估问题和稳定性问题。该算法在 TD3 的基础上,通过增加高估的估计器网络来平衡 TD3 的低估偏差,使估计器估计的 Q 值更加接近于真实 Q 值。本文分析了 TD3、DDPG-MME 算法估计 Q 值的偏差和方差,从理论上证明 DDPG-MME 算法估计出的 Q 值更加接近真实 Q 值且 DDPG-MME 算法 Q 值的估计方差低于 TD3 算法,具有更好的稳定性。

1 相关工作

1.1 强化学习基本概念

强化学习(reinforcement learning, RL)是一种从环境状态映射到动作的学习过程,其目标是使智能体(Agent)在与环境的交互过程中获得最大累积奖赏[19]。Agent 在某一状态下采取某个动作,通过获得奖励或者惩罚来学习是否未来在此状态下继续采取该动作,不断重复该过程从而逼近最优策略。强化学习过程可以抽象为马尔科夫决策过程(Markov decision process, MDP),并用马尔科夫决策过程建模。

马尔科夫决策模型为四元组$<S,A,P,R>$,其中:

(1)S 表示 Agent 所处状态的合集;

(2)A 表示 Agent 所能够采取的动作的合集;

(3)R 表示 Agent 的奖励函数;

(4)P 表示转移概率函数,表示在 s_t 状态下采取动作 a_t 后转移到 s_{t+1} 的概率。

转移概率函数:

$$P[S=s_{t+1} \mid S=s_t, A=a_t]。 \tag{1}$$

在 RL 中,一般假设未来越远的奖励对当前的影响越小,因此引入了折扣因子 $\gamma \in [0,1]$。t 时刻到 T 时刻的累积折扣奖励表示为

$$R_t = \sum_{i=t}^{T} \gamma^{i-t} r(s_i, a_i)。 \tag{2}$$

1.2 基于值函数的强化学习方法

在基于值函数的方法中,通常采用状态价值或状态动作价值来进行动作估值,状态动作价值定义为

$$Q_\pi(S,A) = E_\pi(R_t \mid S_t = s, A_t = a)。 \tag{3}$$

式中:π 代表智能体的参数化策略函数。

状态动作价值用于评估策略的好坏程度,引导 Agent 优化策略去获得最大的累积奖励。因而在基于值函数的方法中,价值评估的准确与否至关重要。

求解目标 Q 值主要采用贝尔曼等式:

$$Q_\pi(s,a) = r + \gamma E_\pi(s',a')。 \tag{4}$$

式中:s 代表当前状态;a 代表当前状态采取的动作;s' 代表状态 s 下采取动作 a 后转移到的新状态;a' 代表 s' 状态下采取的动作。

为了解决状态空间过大而导致的 Q 学习失效的问题,Allen 等[6]、Nachum 等[7]将卷积神经网络和 Q 学习结合起来,提出了深度 Q 网络(deep Q-network)算法。用参数 θ 表示神经网络的参数,并使用下式更新网络:

$$L(\theta) = E_{s,a,r,s'}[(y^{\mathrm{DQN}} - Q(s,a;\theta))^2]; \tag{5}$$

$$y^{\mathrm{DQN}} = r + \gamma \max_{a' \in A} Q(s',a';\theta')。 \tag{6}$$

DQN 算法主要有 2 个贡献,一是引入了经验回放机制,打破了采样数据之间的关联性;二是引入目标网络机制,稳定网络的更新。虽然 DQN 算法在当时表现优异,在多个游戏环境中甚至超越了人类玩家,但 DQN 算法因为更新估值时的最大化操作而造成了对真实 Q 值的高估,进而导致次优策略。

在基于值函数的方法中,价值估计上任何微小的变化都可能导致次优策略。Sutton 等[20]提出了另一种强化学习算法——策略梯度方法,该方法通过输出一个动作概率的分布避免了上述问题,使得最优策略更加稳定。然而,随机策略的更新需要对状态空间和动作空间同时进行积分,会导致效率较低等问题,因此, Silver 等[17]提出了确定性梯度策略算法,同时引入高斯噪声平衡探索和利用。相比随机策略梯度,确定性策略梯度训练所需的样本更少,收敛速度更快。

DDPG 采用 actor-critic 框架,它包含一个参数化的策略函数 π,其参数用 ϕ 表示,该函数通过将状态映射到确定的动作来表明当前的行动策略。估计器的参数用 θ 表示, π'、ϕ'、Q'、θ' 分别描述目标行动者网络及其参数和目标估计器网络及其参数。

估计器的更新与 Q-learning 类似,使用贝尔曼等式更新:

$$L(\theta)=E_{s,a,r,s'}\{[y^{\mathrm{DDPG}}-Q(s,a;\theta)]^2\};\quad(7)$$

$$y^{\mathrm{DDPG}}=r+\gamma Q'(s',a';\theta')。\quad(8)$$

策略网络的更新基于估计器参数 θ,并通过梯度传播的链式规则进行更新:

$$\nabla_{\phi}J(\phi)=E_{s\sim\rho\pi}[\nabla_a Q(s,a;\theta)\mid_{a=\pi(s;\phi)}\nabla_{\phi}\pi(s;\phi)]。\quad(9)$$

式中: $Q(s,a;\theta)=E_{\pi}(R_t\mid s,a)$。

由于 DDPG 算法的估计器采取了与 Q-learning 相似的更新方法,因此也会产生与 Q-learning 相同的高估问题,进而导致次优策略的发生。

1.3　双延迟深度确定性策略梯度

双延迟深度确定性策略梯度(TD3)算法是对 DDPG 算法的改进[18]。TD3 在 DDPG 的基础上引入了 2 个估计器网络,通过选取 2 个网络估值中的较小 Q 值更新估计器网络,达到降低高估影响的目的,即

$$\min_{i=1,2}Q'[s',\pi(s';\phi');\theta'_i]。\quad(10)$$

TD3 算法还加入了延迟更新的策略,在估计器网络更新 d 次之后更新一次行动者网络,在 TD3 算法中 d 的取值为 2,本文中也使用此设定。尽管 TD3 算法缓解了估计器对 Q 值高估问题,但是所采取的最小化 Q 值操作又导致了估计的 Q 值低于真实值,这种低估将会导致次优策略。根据式(10),TD3 算法的更新过程可以表示为

$$Q(s,a)\leftarrow r+\gamma\min_{i=1,2}Q'_i[s',\pi(s';\phi');\theta'_i]。\quad(11)$$

为了更好地解释最小化的影响,假设 Q_1^{approx} 和 Q_2^{approx} 分别代表从 2 个独立估计器中估计的 Q 值, Q^{true} 为真实值。估计器在进行函数逼近时会引起噪声,因此对于 2 个独立的估计器,存在 2 个偏差项 $Z_1=Q_1^{\mathrm{approx}}-Q^{\mathrm{true}}$ 和 $Z_2=Q_2^{\mathrm{approx}}-Q^{\mathrm{true}}$ 独立同分布于 $[-\mu,\mu]$ 的均匀分布。

Sutton 等[20]指出,若 $E[\min_{i=1,2}Z_i]=0$,表明最小化操作后的估计仍然是无偏估计;若 $E[\min_{i=1,2}Z_i]\neq 0$,说明最小化操作后的估计是有偏估计。对于 2 个独立同分布于均匀分布于 $[-\mu,\mu]$ 的随机变量 Z_1 和 Z_2,经过最小化操作后 $\min_{i=1,2}Z_i$ 的期望为 $-\mu/3$,因此,可以得出 TD3 算法在进行最小化操作后偏差的期望小于 0,说明即使估计器是无偏的,最小化操作也可能在每次更新时引入负的预期偏差,导致低估计问题。

2　基于多估计器平均值的深度确定性策略梯度算法

本节针对 TD3 的低估问题,提出一种基于多估计器平均值的深度确定性策略梯度算法(DDPG-MME)。该算法在 TD3 的基础上增加了多个估计器网络来降低低估偏差和估计方差,可以在一定程度上提高算法的性能和稳定性。

2.1　DDPG-MME 的 Q 值估计偏差

作为单个估计器的强化学习算法,DDPG 会产生 Q 值高估偏差,而 TD3 算法中双估计器之间的最小化操作又带来了低估偏差。本文通过将这 2 种相反的估计偏差结合,实现更准确的 Q 值估计。具体而言,本文在 TD3 和 DDPG 的基础上,提出了 DDPG-MME 算法,该算法共使用 k 个估计器,通过对其中 2 个估计器输出的 Q 值之间的最小值和另外(k-2)个估计器输出 Q 值的平均值进行平均加权来获得目标,更新 Q 值,如下式所示:

$$y^{\mathrm{DDPG\text{-}MME}}=r+\gamma(Q'_m+Q'_n)/2。\quad(12)$$

式中: $Q'_m=\min_{i=1,2}Q'_i[s',\pi(s';\phi');\theta'_i]$;

$Q'_n=\{Q'_3[s',\pi(s';\phi');\theta'_3]+Q'_4[s',\pi(s';\phi');\theta'_4]+\cdots+Q'_k[s',\pi(s';\phi');\theta'_k]\}/(k-2)$,并且 $Q'_1,Q'_2,\cdots,Q'_k$ 之间相互独立。行动者网络参数 ϕ 的更新方式为

$$\nabla_{\phi}J(\phi)=N^{-1}\sum\nabla_a Q_{\theta_1}(s,a)\mid_{a=\pi_{\phi}(s)}\nabla_{\phi}\pi_{\phi}(s)。\quad(13)$$

估计器的目标网络参数的更新方式为

$$\theta'_i\leftarrow\tau\theta_i+(1-\tau)\theta'_i,\phi'\leftarrow\tau\phi+(1-\tau)\phi'。\quad(14)$$

为分析 DDPG-MME 算法导致的 Q 值估计偏

差,本文给出如下定理。

定理1 给定$k(k>3)$个独立同分布于区间为$[-\mu,\mu]$的均匀分布的变量$Z_1,Z_2,\cdots,Z_k$,则经过平均加权后的$\frac{1}{2}\left[\min\limits_{i=1,2}Z_i+\frac{1}{k-2}(Z_3+Z_4+\cdots+Z_k)\right]$的期望为

$$E\left\{\frac{1}{2}\left[\min_{i=1,2}Z_i+\frac{1}{k-2}(Z_3+Z_4+\cdots+Z_k)\right]\right\}=-\frac{1}{6}\mu。\tag{15}$$

证明 根据已知条件可知,对$i=1,2,\cdots,k$,Z_i独立同分布于$[-\mu,\mu]$的均匀分布,且$E[\min\limits_{i=1,2}Z_i]=-\mu/3$,则:

$$E\left\{\frac{1}{2}\left[\min_{i=1,2}Z_i+\frac{1}{k-2}(Z_3+Z_4+\cdots+Z_k)\right]\right\}=-\frac{1}{6}\mu。$$

证毕。

根据定理1,可得DDPG-MME算法的估计偏差低于TD3算法,较低的估计偏差有助于提高算法稳定性,提升算法性能。

2.2 DDPG-MME算法的Q值估计偏差的方差

估计偏差的方差反映了偏差的波动程度,方差越小,估计Q值的波动程度越低,算法稳定性越高,进而可以得到更好的策略。

定理2 给定$k(k>3)$个独立同分布于均匀分布$[-\mu,\mu]$的随机变量$Z_1,Z_2,\cdots,Z_k$,则经过平均加权后得到的$\frac{1}{2}\left[\min\limits_{i=1,2}Z_i+\frac{1}{k-2}(Z_3+Z_4+\cdots+Z_k)\right]$的方差:

$$Var\left\{\frac{1}{2}\left[\min_{i=1,2}Z_i+\frac{1}{k-2}(Z_3+Z_4+\cdots+Z_k)\right]\right\}=\frac{1}{18}\mu^2+\frac{1}{12(k-2)}\mu^2。\tag{16}$$

证明 由已知条件可知,$Z_1,Z_2,\cdots,Z_k$独立同分布,则

$$Var\left\{\frac{1}{2}\left[\min_{i=1,2}Z_i+\frac{1}{k-2}(Z_3+Z_4+\cdots+Z_k)\right]\right\}=\frac{1}{4}Var(\min_{i=1,2}Z_i)+\frac{1}{4}Var\left[\frac{1}{k-2}(Z_3+Z_4+\cdots+Z_k)\right]=\frac{1}{4}E[(\min_{i=1,2}Z_i)^2]-\frac{1}{4}[E(\min_{i=1,2}Z_i)]^2+\frac{1}{4(k-2)^2}Var(Z_3+Z_4+\cdots+Z_k)=\frac{1}{18}\mu^2+\frac{1}{12(k-2)}\mu^2。$$

证毕。

在得到$\frac{1}{2}\left[\min\limits_{i=1,2}Z_i+\frac{1}{k-2}(Z_3+Z_4+\cdots+Z_k)\right]$的方差后下面对TD3的估计偏差的方差进行分析。

定理3 给定2个独立同分布于均匀分布$[-\mu,\mu]$的随机变量Z_1、Z_2,经过最小化操作得到的$\min\limits_{i=1,2}Z_i$的方差为

$$Var(\min_{i=1,2}Z_i)=\frac{2}{9}\mu^2。\tag{17}$$

证明 由已知条件可知,Z_1、Z_2独立同分布,则

$$Var(\min_{i=1,2}Z_i)=E[(\min_{i=1,2}Z_i)^2]-[E(\min_{i=1,2}Z_i)]^2=\frac{2}{9}\mu^2。$$

证毕。

结合定理2和定理3的结论,可以得出,对于任意$k(k>3)$:

$$Var\left\{\frac{1}{2}\left[\min_{i=1,2}Z_i+\frac{1}{k-2}(Z_3+Z_4+\cdots+Z_k)\right]\right\}\leqslant Var(\min_{i=1,2}Z_i)。\tag{18}$$

即DDPG-MME算法的估计方差低于TD3算法的估计方差,这说明与TD3算法相比,DDPG-MME算法得到的估计值更加稳定。

此外,从定理2的结论可以看出,随着k的增大,$\frac{1}{12(k-2)}\mu^2$会逐步下降。在算力允许的情况下,估计器的个数越多,估计偏差的方差则越低。当k趋于无穷大时,估计偏差的方差趋近于$\mu^2/18$。需要指出的是,当采用$k=4$的设置时,根据定理2可以得到估计偏差的方差为

$$Var\left\{\frac{1}{2}\left[\min_{i=1,2}Z_i+\frac{1}{2}(Z_3+Z_4)\right]\right\}=\frac{7}{72}\mu^2。\tag{19}$$

这时DDPG-MME算法的估计偏差的方差仍低于TD3算法。

2.3 DDPG-MME算法

在TD3算法框架的基础上,引入k个独立的估计器,形成DDPG-MME算法的基本框架。DDPG-MME算法的输入为初始状态s,输出为行动a。首先,初始化k个估计器$\theta_1,\theta_2,\cdots,\theta_k$和1个行动者主网络的参数$\phi$,并将$(k+1)$个网络的参数$\theta_1,\theta_2,\cdots,\theta_k,\phi$复制到目标网络当中。其次,Agent通过$\pi_\phi(s)+\varepsilon,\varepsilon\sim N(0,\sigma)$策略选择并执行相应动作$a$得到即时奖励$r$和下个状态$s'$,并将转移样本$(s,a,r,s')$存储到经验回放池

B 中。在训练估计器和行动者网络时,从经验回放池 B 中抽取 N 个转移样本 (s,a,r,s'),按照策略网络平滑正则化的方式 $a' \leftarrow \pi_\phi'(s) + \varepsilon, \varepsilon \sim clip[N(0,\sigma'), -c, c]$ 选取动作 a',并根据式(12)来计算目标 Q 值,随后根据 $\theta_i \leftarrow \arg\min N^{-1} \cdot \sum (y^{\text{MME-DDPG}} - Q_{\theta_i}(s,a))^2$ 更新估计器网络参数。最后,根据延迟策略更新行动者网络参数和目标网络参数。

算法 1　基于多估计器平均值的深度确定性策略梯度算法(DDPG-MME)。

输入:观测状态 s;

输出:行动 a。

① 用随机参数 $\theta_1, \theta_2, \cdots, \theta_k, \phi$ 初始化 k 个估计器网络 $Q_{\theta_1}, Q_{\theta_2}, \cdots, Q_{\theta_k}$ 和 1 个 Actor 网络 π_ϕ;

② 初始化目标网络参数 $\theta_1', \theta_2', \cdots, \theta_k', \phi'$;

③ 初始化经验回放池 B;

④ For $t=1$ to T do

⑤ 根据 $a \sim \pi_\phi(s) + \varepsilon, \varepsilon \sim N(0,\sigma)$ 选择并执行行动 a,得到奖励 r 和新的状态 s';

⑥ 将转移样本 (s,a,r,s') 存储到 B 中;

⑦ 从 B 中抽取 N 个转移样本 (s,a,r,s');

⑧ $a' \leftarrow \pi_\phi'(s) + \varepsilon, \varepsilon \sim clip(N(0,\sigma'), -c, c)$;

⑨ 根据式(12)计算目标 Q 值;

⑩ 对于 $i = 1,2,\cdots,k$, 更新 Critic 网络参数 $\theta_i \leftarrow \arg\min N^{-1} \sum [y^{\text{MME-DDPG}} - Q_{\theta_i}(s,a)]^2$;

⑪ If t mod d then

⑫ 根据式(13)更新 ϕ;

⑬ 根据式(14)更新目标网络参数;

⑭ End if

⑮ End for

3　实验结果及分析

本节主要介绍评估 DDPG-MME 算法性能所用的实验环境、实验设置、实验结果及分析。

3.1　实验环境

为了评估 DDPG-MME 算法,本文在 OpenAI 开发的 Gym 平台上测量了算法的性能,所用的实验环境为 4 个 MuJoCo 连续控制任务环境。

3.2　实验设置

本文在 4 个 MuJoCo 连续控制环境中对 DDPG-MME、DDPG 和 TD3 算法进行对比。实验中每个算法用 10 个随机种子运行,并且每 5 000 个时间步长评估 10 次算法性能。在每个实验环境下,本文都进行了 10^6 步长的训练,为了获得更好的经验,消除对初始参数的依赖,本文在前 10 000步长中采用纯探索策略。DDPG-MME 算法的基本网络架构和 TD3 算法的相似,均采用两层全连接神经网络,分别由 400 个和 300 个隐藏节点组成。在每一层之间都有 Relu 激活函数,最后一层采用 tanh 单元输出动作。

由于 DDPG-MME 算法中估计器的个数 k 设置为 4 时,算法估计偏差的方差已经小于 TD3 算法,并且此时算力的需求最小,因此本文在对比实验中使用了 $k=4$ 的设置,其余超参数(奖励函数、环境参数、批次大小、学习率、优化器、折扣系数)均与 TD3 算法的参数一致。为了最小化 Bellman 损失,网络使用 Adam 优化器进行更新。此外,每次更新网络时,都选取一批数量为 100 的样本来训练网络参数,实验中学习率设定为10^{-3}。

通过在目标动作网络添加正则项 $\varepsilon \sim clip(N(0,0.2), -0.5, 0.5)$ 平滑动作的输出,并加入延迟更新机制,每当估计器网络更新 d 步(本文中 d 设置为 2)之后,更新 1 次行动者网络。估计器和行动者网络参数通过软更新来进行更新,软更新参数 $\tau = 0.005$。

3.3　实验结果分析

本文在 Reacher-v2、HalfCheetah-v2、Inverted-Pendulum-v2、InvertedDoublePendulum-v2 中对比了 DDPG-MME、DDPG 和 TD3 算法的表现。

图 1 展示了 DDPG-MME、DDPG 和 TD3 算法在不同训练次数后的平均得分,其中,实线部分代表 10 次评估的平均得分,实线的高低程度反映了算法性能的优劣,阴影部分代表标准差的一半,阴影范围越大,算法稳定性越差。由图 1 可得,在 4 种不同的环境中,在经历了大约 20 000 步长之后 DDPG-MME 的得分便开始逐步高于 TD3 和 DDPG 算法,并在最终的平均得分上超越上述 2 种算法。

图 1 表明 DDPG-MME 在缓解 TD3 的低估问题后,获得了更优的策略和更高的分数,体现了 DDPG-MME 算法的优越性。此外,图 1 也表明 TD3 算法在引入目标策略平滑和软更新后,算法的稳定性得到了提升,估计方差得到了较好的控制。在 TD3 算法的基础上,DDPG-MME 算法的平均加权方法也使得算法的稳定性进一步得到提升。

表 1 对比了 3 种算法在 4 个 MuJoCo 连续控制环境下 10^6 个时间步长的 10 次试验中的最大平均回报。从表 2 可以得出,采取了平均加权操作后的 DDPG-MME 算法比 TD3 算法和 DDPG 算法具有更好的性能和稳定性。

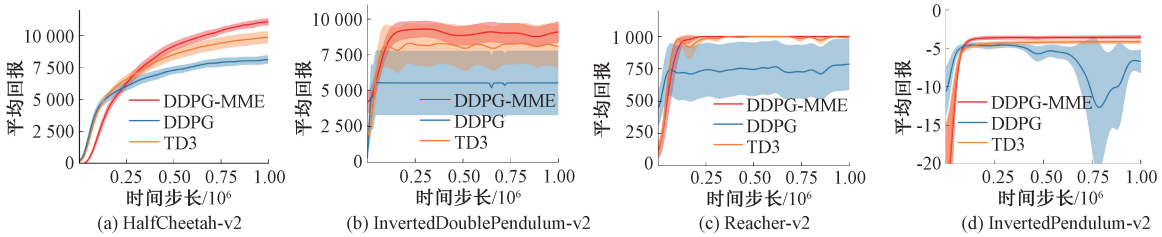

图1 MuJoCo环境下4种实验学习曲线

Figure 1 Learning curves for four experiments in the MuJoCo control environment

表1 DDPG-MME、TD3、DDPG算法的最大平均回报对比

Table 1 Comparison of the maximum average returns of the DDPG-MME, TD3, DDPG algorithms

环境	DDPG-MME	DDPG	TD3
InvertedPendulum-v2	1 000.00	780.25±392.28	992.98±21.07
InvertedDoublePendulum-v2	9 234.11±452.03	5 538.05±4 499.08	8 153.41±2 754.33
Reacher-v2	−3.65±0.51	−6.62±2.27	−4.17±0.57
HalfCheetah-v2	11 065.41±382.00	8 070.43±642.93	9 842.387±1 013.92

4 结论

为了对 Q 值的估计更加准确,本文提出了一种基于多估计器平均值的深度确定性策略梯度(DDPG-MME)算法。该算法通过平均加权的方式缓解了TD3算法的低估问题,提高了算法性能,结论如下:

(1)DDPG-MME算法在一定程度上缓解了TD3算法存在的低估问题;

(2)从理论上分析了DDPG-MME算法和TD3算法估值的偏差和方差,证明DDPG-MME算法 Q 值估计的偏差低于TD3算法,并且DDPG-MME算法 Q 值估计偏差的方差低于TD3算法;

(3)在4个MuJoCo连续控制环境中对DDPG-MME、TD3以及DDPG算法的性能和稳定性进行对比,验证了DDPG-MME算法的优越性。

参考文献:

[1] 陈兴国,俞扬.强化学习及其在电脑围棋中的应用[J].自动化学报,2016,42(5):685-695.

[2] 张凯峰,俞扬.基于逆强化学习的示教学习方法综述[J].计算机研究与发展,2019,56(2):254-261.

[3] 王丙琛,司怀伟,谭国真.基于深度强化学习的自动驾驶车控制算法研究[J].郑州大学学报(工学版),2020,41(4):41-45,80.

[4] BERTSEKAS D P, BERTSEKAS D P, BERTSEKAS D P, et al. Dynamic programming and optimal control[M]. Nashua,NH:Athena scientific, 1995.

[5] ANSCHEL O, BARAM N, SHIMKIN N,et al. Averaged-DQN: variance reduction and stabilization for deep rein-forcement learning[C]//Proceedings of the 34th International Conference on Machine Learning. New York:ACM,2017:176-185.

[6] ALLEN C,ASADI K,RODERICK M,et al.Mean actor critic[EB/OL]. (2017-06-11)[2021-08-04]. https://arxiv.org/abs/1709.00503.

[7] NACHUM O, NOROUZI M, TUCKER G, et al. Smoothed action value functions for learning Gaussian policies[EB/OL].(2018-10-11)[2021-08-04]. https://arxiv.org/abs/1803.02348.

[8] HASSELT H. Double Q-learning[C]//Advances in neural information processing systems 23. Boston: MIT,2010: 2613-2621.

[9] MNIH V, KAVUKCUOGLU K, SILVER D, et al. Human-level control through deep reinforcement learning[J].Nature,2015,518(7540):529-533.

[10] MNIH V, KAVUKCUOGLU K, SILVER D, et al. Playing atari with deep reinforcement learning[EB/OL]. (2013-06-11)[2021-08-04]. http://export.arxiv.org/pdf/1312.5602.

[11] LI A,LU Z Q,MIAO C L.Revisiting prioritized experience replay:a value perspective[EB/OL]. (2021-03-11)[2021-08-04].https://arxiv.org/abs/2102.03261.

[12] WANG Z,SCHAULT T, HESSEL M, et al. Dueling network archite ctures for deep reinforcement learning[C]//Proceedings of the 33rd International Conference on machine Learning. New York: ACM, 2016: 1995-2003.

[13] 吴金金,刘全,陈松,等.一种权重平均值的深度双Q网络方法[J].计算机研究与发展,2020,57(3):576-589.

[14] Van HASSELT H, GUEZ A, SILVER D. Deep reinforcement learning with double Q-Learning[C]//Proceedings of the 30th AAAI Conference on Artificial In-

telligence. Phoenix: AAAI, 2016:2094-2100.
[15] PETERS J,SCHAAL S.Natural actor-critic[J].Neurocomputing,2008,71(7/8/9):1180-1190.
[16] LILLICRAP T P, HUNT J J, PRITZEL A. Continuous control with deep reinforcement learning[EB/OL].(2019-06-05)[2021-09-09].https://arxiv.org/abs/1509.02971.
[17] SILVER D, LEVER G, HEESS N, et al. Deterministic policy gradient algorithms[C]//Proceedings of the 31st International Conference on International Conference on Machine Learning.New York: ACM,2014: 387-395.
[18] FUJIMOTO S, VAN HOOF H, MEGER D. Addressing function approximation error in actor-critic methods[EB/OL].(2018-03-11)[2021-08-04].https://arxiv.org/abs/1802.09477.
[19] 刘全,翟建伟,章宗长,等.深度强化学习综述[J].计算机学报,2018,41(1):1-27.
[20] SUTTON R S,MCALLESTER D,SINGH S,et al.Policy gradient methods for reinforcement learning with function approximation[C]//Advances in Neural Information Processing Systems 12.Boston:MIT,2000:1057-1063.

Deep Deterministic Policy Gradient Algorithm Based on Mean of Multiple Estimators

LI Lin[1,2], LI Yuze[1], ZHANG Yujia[1], WEI Wei[1,2]

(1.School of Computer and Information Technology, Shanxi University, Taiyuan 030006, China; 2.Key Laboratory of Computational Intelligence and Chinese Information Processing Ministry of Education, Shanxi University, Taiyuan 030006, China)

Abstract: In order to solve the underestimation problem of the twin delayed deep deterministic policy gradient algorithm in the reinforcement learning actor-critic framework, deep determinstic policy gradient based on mean of multiple estimators(DDPG-MME) was proposed. The method contained one actor and $k(k>3)$ critics, and the minimum of the output values of two critics and the mean value of the remaining $(k-2)$ critics was calculated first, and then the average of the two values as the final value was taken to calculate the TD error. Finally, we update the critic network based on the TD error, and the actor network is updated based on the value of the first critic. The weighting operation of the method could alleviate the underestimation problem of the twin delayed deep deterministic policy gradient algorithm and reduces the estimation variance to a certain extent to achieve more accurate Q-value estimation. The expectation and variance of the estimation error of our method, deep deterministic policy gradient was analyzed theoretically, and twin delayed deep deterministic policy gradient, and the accuracy and stability of the method was demonstrated. And the experimental results in four MuJoCo continuous control environments, such as Reacher-v2, HalfCheetah-v2, InvertedPendulum-v2 and InvertedDoublePendulum-v2, showed the superior final performance of the deep deterministic policy gradient based on mean of multiple estimators algorithm over TD3 and DDPG, and the results showed that the final performance and stability of our algorithm were significantly better than the comparison algorithms under the same hyperparameters (network structure, reward function, environment parameters, batch size, learning rate, optimizer and discount factor) settings as the comparison algorithms.

Keywords: reinforcement learning; actor-critic; underestimation; multiple estimators; policy gradient