(19)中华人民共和国国家知识产权局

(12)发明专利

(10)授权公告号 CN 108803321 B
(45)授权公告日 2020.07.10

(21)申请号 201810535773.8

(22)申请日 2018.05.30

(65)同一申请的已公布的文献号
申请公布号 CN 108803321 A

(43)申请公布日 2018.11.13

(73)专利权人 清华大学
地址 100084 北京市海淀区清华园1号

(72)发明人 宋士吉 石文杰

(74)专利代理机构 北京清亦华知识产权代理事务所(普通合伙) 11201
代理人 廖元秋

(51)Int.Cl.
G05B 13/04(2006.01)

(56)对比文件
CN 107102644 A,2017.08.29,
US 2012188365 A1,2012.07.26,
CN 107856035 A,2018.03.30,
CN 107368076 A,2017.11.21,
KR 101545731 B1,2015.08.20,
CN 107065881 A,2017.08.18,
马琼雄等.基于深度强化学习的水下机器人最优轨迹控制.《华南师范大学(自然科学版)》.2018,第50卷(第1期),第118-123页.
段勇等.进化强化学习及其在机器人路径跟踪中的应用.《控制与决策》.2009,第24卷(第4期),第532-536、541页.
Runsheng Yu等.Deep Reinforcement Learning Based Optimal Trajectory Tracking Control of Autonomous Underwater Vehicle.《Proceedings of the 36th Chinese Control Conference》.2017,第4958-4965页.
Li Zhou等.AUV Based Source Seeking with Estimated Gradients.《Journal of Systems Science & Complexity》.2018,(第1期),第262-275页.

审查员 孔璐璐

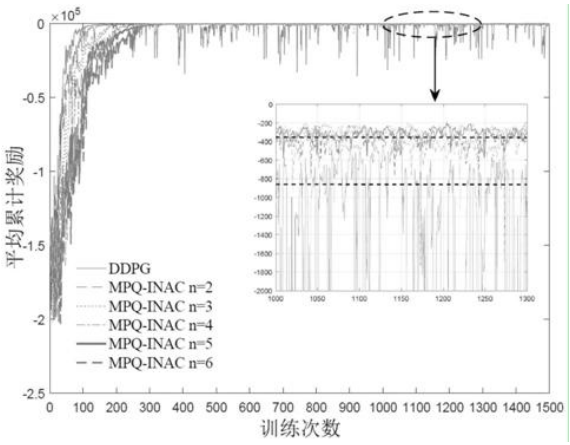
权利要求书4页 说明书12页 附图2页

(54)发明名称

基于深度强化学习的自主水下航行器轨迹跟踪控制方法

(57)摘要

本发明提出了一种基于深度强化学习的自主水下航行器轨迹跟踪控制方法,属于深度强化学习和智能控制领域。首先定义AUV轨迹跟踪控制问题;然后建立AUV轨迹跟踪问题的马尔科夫决策过程模型;接着构建混合策略-评价网络,该网络由多个策略网络和评价网络构成;最后由构建的混合策略-评价网络求解AUV轨迹跟踪控制的目标策略,对于多个评价网络,通过定义期望贝尔曼绝对误差来评估每个评价网络的性能,在每个时间步只更新性能最差的一个评价网络,对于多个策略网络,在每个时间步随机选择一个策略网络,并采用确定性策略梯度进行更新,最终学习到的策略为所有策略网络的均值。本发明不易受到恶劣AUV历史跟踪轨迹的影响,精度高。



CN 108803321 B

1.一种基于深度强化学习的自主水下航行器轨迹跟踪控制方法,其特征在于,该方法包括以下步骤:

1)定义自主水下航行器AUV轨迹跟踪控制问题

定义AUV轨迹跟踪控制问题包括四个部分:确定AUV系统输入、确定AUV系统输出、定义轨迹跟踪控制误差和建立AUV轨迹跟踪控制目标;具体步骤如下:

1-1)确定AUV系统输入

令AUV系统输入向量为$\tau_k=[\xi_k,\delta_k]^T$,其中$\xi_k$、$\delta_k$分别为AUV的螺旋桨推力和舵角,下标k表示第k个时间步;ξ_k、δ_k的取值范围分别为$[0,\overline{\xi}]$和$[-\overline{\delta},\overline{\delta}]$,$\overline{\xi}$、$\overline{\delta}$分别为最大的螺旋桨推力和最大舵角;

1-2)确定AUV系统输出

令AUV系统输出向量为$\eta_k=[x_k,y_k,\psi_k]^T$,其中$x_k$、$y_k$分别为第k个时间步AUV在惯性坐标系I-XYZ下沿X、Y轴的坐标,ψ_k为第k个时间步AUV前进方向与X轴的夹角;

1-3)定义轨迹跟踪控制误差

根据AUV的行驶路径选取参考轨迹$d_k=[x_k^d,y_k^d]^T$,定义第k个时间步的AUV轨迹跟踪控制误差为:

$$\mathbf{e}_k=[x_k-x_k^d,y_k-y_k^d]^T$$

1-4)建立AUV轨迹跟踪控制目标

对于步骤1-3)中的参考轨迹d_k,选择如下形式的目标函数:

$$P_k(\tau)=\sum_{i\geq k}\gamma^{i-k}(\mathbf{e}_i^T\mathbf{e}_i+\tau_i^T\mathbf{H}\tau_i)$$

其中,γ是折扣因子,H为权重矩阵;

建立AUV轨迹跟踪控制的目标为找到一个最优系统输入序列τ^*使得初始时刻的目标函数$P_0(\tau)$最小,计算公式如下:

$$\tau^*=\arg\min_{\tau}P_0(\tau)$$

2)建立AUV轨迹跟踪问题的马尔科夫决策过程模型

对步骤1)中的AUV轨迹跟踪问题进行马尔科夫决策过程建模,具体步骤如下:

2-1)定义状态向量

定义AUV系统的速度向量为$\Phi_k=[u_k,v_k,x_k]^T$,其中u_k、v_k分别为第k个时间步AUV沿前进方向、垂直于前进方向的线速度,x_k为第k个时间步AUV环绕前进方向的角速度;

根据步骤1-2)确定的AUV系统输出向量η_k和步骤1-3)定义的参考轨迹,定义第k个时间步的状态向量如下:

$$s_k=[\eta_k^T,\phi_k^T,d_k^T,d_{k+1}^T]^T$$

2-2)定义动作向量

定义第k个时间步的动作向量为该时间步的AUV系统输入向量,即$a_k=\tau_k$;

2-3)定义奖励函数

第k个时间步的奖励函数用于刻画在状态s_k采取动作a_k的执行效果,根据步骤1-3)定义的轨迹跟踪控制误差e_k和步骤2-2)定义的动作向量a_k,定义第k个时间步的AUV奖励函数如

下：

$$r_{k+1}=r(s_k,a_k)=-(\mathbf{e}_k^T\mathbf{e}_k+a_k^T\mathbf{H}a_k)$$

2-4)将步骤1-4)建立的AUV轨迹跟踪控制的目标τ*转换为强化学习框架下的AUV轨迹跟踪控制目标

定义策略π为在某一状态下选择各个可能动作的概率，则定义动作值函数如下：

$$Q^{\pi}(s_k,a_k)=E_{r_{i>k},s_{i>k},a_{i>k}\sim\pi}\left[\sum_{i=k}^{K}\gamma^{i-k}r_{i+1}\,|\,s_k,a_k\right]$$

其中，$E_{r_{i>k},s_{i>k},a_{i>k}\sim\pi}$表示对奖励函数、状态和动作的期望值；K为最大时间步；

该动作值函数用于描述在当前及之后所有状态下均采取策略π时的期望累计折扣奖励，故在强化学习框架下，AUV轨迹跟踪控制目标是通过与AUV所处环境的交互来学习一个最优目标策略π*，使得初始时刻的动作值最大，计算公式如下：

$$\pi^*=\arg\max_{\pi}E_{p(s_0),a_0\sim\pi}Q^{\pi}(s_0,a_0)$$

其中，$p(s_0)$为初始状态s_0的分布；a_0为初始动作向量；

将步骤1-4)建立的AUV轨迹跟踪控制的目标τ*的求解转换为π*的求解；

2-5)简化强化学习框架下的AUV轨迹跟踪控制目标

通过如下迭代贝尔曼方程来求解步骤2-4)中的动作值函数：

$$Q^{\pi}(s_k,a_k)=E_{r_{k+1},s_{k+1}}\left[r_{k+1}+\gamma E_{a_{k+1}\sim\pi}\left[Q^{\pi}(s_{k+1},a_{k+1})\right]\right]$$

设策略π是确定性的，即从AUV的状态向量空间到AUV的动作向量空间是一一映射的关系，并记为μ，则将上述迭代贝尔曼方程简化为：

$$Q^{\mu}(s_k,a_k)=E_{r_{k+1},s_{k+1}}\left[r_{k+1}+\gamma Q^{\mu}(s_{k+1},\mu(s_{k+1}))\right]$$

对于确定性的策略μ，将步骤2-4)中的最优目标策略π*简化为确定性最优目标策略μ*：

$$\mu^*=\arg\max_{\mu}E_{p(s_0)}Q^{\mu}(s_0,\mu(s_0))$$

3)构建混合策略-评价网络

通过构建混合策略-评价网络来分别估计确定性最优目标策略μ*和对应的最优动作值函数Q^{μ^*}，构建混合策略-评价网络包括三部分：构建策略网络、构建评价网络和确定目标策略，具体步骤如下：

3-1)构建策略网络

混合策略-评价网络结构通过构建n个策略网络$\mu_{\theta_p}(s_k)$来估计确定性最优目标策略μ*；其中，θ_p为第p个策略网络的权重参数，p＝1,…,n；各策略网络均分别使用一个全连接的深度神经网络来实现，各策略网络均分别包含一个输入层、两个隐藏层和一个输出层；各策略网络的输入为状态向量s_k，各策略网络的输出为动作向量a_k；

3-2)构建评价网络

混合策略-评价网络结构通过构建m个评价网络$Q_{w_q}(s_k,a_k)$来估计最优动作值函数Q^{μ^*}；其中，w_q为第q个评价网络的权重参数，q＝1,…,m；各评价网络均分别使用一个全连接的深

度神经网络来实现，各评价网络均分别包含一个输入层、两个隐藏层和一个输出层；各评价网络的输入为状态向量s_k和动作向量a_k，其中状态向量s_k从输入层输入到各评价网络，动作向量a_k从第一个隐藏层输入到各评价网络，各评价网络输出为在状态向量s_k下采取动作向量a_k的动作值；

3-3)确定目标策略

根据所构建的混合策略-评价网络，将第k个时间步学习到的AUV轨迹跟踪控制的目标策略$\mu_f(s_k)$定义为n个策略网络输出的均值，计算公式如下：

$$a_k=\mu_f(s_k)=\frac{1}{n}\sum_{p=1}^{n}\mu_{\theta_p}(s_k)$$

4)求解AUV轨迹跟踪控制的目标策略$\mu_f(s_k)$，具体步骤如下：

4-1)参数设置

分别设置最大迭代次数M、每次迭代的最大时间步K、经验回放抽取的训练集大小N、各评价网络的学习率α_ω、各策略网络的学习率α_θ、折扣因子γ和奖励函数中的权重矩阵H；

4-2)初始化混合策略-评价网络

随机初始化n个策略网络$\mu_{\theta_p}(s_k)$和m个评价网络$Q_{w_q}(s_k,a_k)$的权重参数θ_p和w_q；从n个策略网络中随机选择第d个策略网络记为μ_{θ_d}，d＝1,…,n；

构建经验列队集合R，设该经验列队集合R的最大容量为B，并初始化为空；

4-3)迭代开始，对混合策略-评价网络进行训练，初始化迭代次数episode＝1；

4-4)设置当前时间步k＝0，随机初始化AUV的状态变量s_0，令当前时间步的状态变量s_k＝s_0；并产生一个探索噪声$Noise_k$；

4-5)根据n个当前策略网络μ_{θ_p}和探索噪声$Noise_k$确定当前时间步的动作向量a_k为：

$$a_k=\frac{1}{n}\sum_{p=1}^{n}\mu_{\theta_p}(s_k)+Noise_k$$

4-6)AUV在当前状态s_k下执行动作a_k，根据步骤2-3)得到奖励函数r_{k+1}，并观测到一个新的状态s_{k+1}；记e_k＝(s_k,a_k,r_{k+1},s_{k+1})为一个经验样本；如果经验列队集合R的样本数量已经达到最大容量B，则先删除最先加入的一个样本，再将经验样本e_k存入经验列队集合R中；否则直接将经验样本e_k存入经验列队集合R中；

从经验列队集合R中选取A个经验样本，具体如下：当经验列队集合R中样本数量不超过N时，则选取该经验列队集合R中的所有经验样本；当经验列队集合R超过N时，则从该经验列队集合R中随机选取N个经验样本(s_l,a_l,r_{l+1},s_{l+1})；

4-7)根据选取的A个经验样本计算每个评价网络的期望贝尔曼绝对误差$EBAE_q$，用于表征每个评价网络的性能，公式如下：

$$EBAE_q=\frac{1}{A}\sum_{l}\left|Q_{w_q}(s_l,a_l)-r_{l+1}-\gamma Q_{w_q}(s_{l+1},\mu_{\theta_d}(s_{l+1}))\right|,\quad q=1,\cdots,m$$

选择性能最差的评价网络，通过以下公式求得该性能最差的评价网络的序号，记为c：

$$c=\arg\max_{q} EABE_q$$

4-8)由第c个评价网络Q_{w_c},通过如下次贪婪策略得到每个经验样本在下一时间步的动作向量:

$$a_{l+1}=\arg\max_{a} Q_{w_c}(s_{l+1},a),\quad s.t.\ a\in\left\{\mu_{\theta_p}(s_{l+1}),\ p=1,\cdots,n\right\}$$

4-9)通过多个准Q学习方法计算第c个评价网络的目标值Y_{l+1}^{c},公式如下:

$$Y_{l+1}^{c}=r_{l+1}+\frac{\gamma}{m-1}\sum_{q=1,q\neq c}^{m}Q_{w_q}(s_{l+1},a_{l+1})$$

4-10)计算第c个评价网络的损失函数L(wc),公式如下:

$$L(w_c)=\frac{1}{A}\sum_{l}\left(Q_{w_c}(s_l,a_l)-Y_{l+1}^{c}\right)^2$$

4-11)通过损失函数L(wc)对权重参数wc的导数来更新第c个评价网络的权重参数,公式如下:

$$w_c=w_c+\alpha_{\omega}\nabla_{w_c}L(w_c)$$

其余评价网络的权重参数保持不变;

4-12)从n个策略网络中随机选择一个策略网络来重置第d个策略网络μ_{θ_d};

4-13)根据更新后的第c个评价网络计算第d个策略网络μ_{θ_d}的确定性策略梯度$\nabla_{\theta_d}J$并以此更新第d个策略网络μ_{θ_d}的权重参数θ_d,计算公式分别如下:

$$\nabla_{\theta_d}J=\frac{1}{A}\sum_{l}\nabla_{a}Q_{w_c}(s_l,a)|_{a=\mu_{\theta_d}(s_l)}\nabla_{\theta_d}\mu_{\theta_d}(s_l)$$

$$\theta_d=\theta_d-\alpha_{\theta}\nabla_{\theta_d}J$$

其余策略网络的权重参数保持不变;

4-14)令k=k+1并对k进行判定:如k<K,则重新返回步骤4-5),AUV继续跟踪参考轨迹;否则,进入步骤4-15);

4-15)令episode=episode+1并对episode进行判定:如episode<M,则重新返回步骤4-4),AUV进行下一个迭代过程;否则,进入步骤4-16);

4-16)迭代结束,终止混合策略-评价网络的训练过程,将迭代终止时的n个策略网络的输出值通过步骤3-3)中的计算公式得到最终AUV轨迹跟踪控制的目标策略$\mu_f(s_k)$,由该目标策略实现对AUV的轨迹跟踪控制。

基于深度强化学习的自主水下航行器轨迹跟踪控制方法

技术领域

[0001] 本发明属于深度强化学习和智能控制领域,涉及一种基于深度强化学习的自主水下航行器(AUV)轨迹跟踪控制方法。

背景技术

[0002] 深海海底科学的发展高度依赖于深海探测技术和装备,由于深海环境复杂、条件极端,目前主要采用深海作业型自主水下航行器代替或辅助人对深海进行探测、观察和采样。而针对海洋资源探索、海底调查和海洋测绘等人类无法到达现场操作的任务场景,保证AUV水下运动的自主性和可控性是一项最基本且重要的功能要求,是实现各项复杂作业任务的前提。然而,AUV的许多离岸应用(例如轨迹跟踪控制、目标跟踪控制等)极具挑战性,这种挑战性主要由AUV系统以下三方面的特性导致。第一,AUV作为一种多输入多输出系统,其动力学和运动学模型(以下简称模型)复杂,具有高度非线性、强耦合、存在输入或状态约束和时变等特点;第二,模型参数或水动力环境存在不确定性,导致AUV系统建模较为困难;第三,当前大部分AUV属于欠驱动系统,即自由度大于独立执行器的数量(各独立执行器分别对应一个自由度)。通常,通过数学物理机理推导、数值模拟和实物实验相结合的方法来确定AUV的模型及参数,并合理刻画模型中的不确定部分。复杂的模型导致AUV的控制问题也非常复杂。而且,随着AUV应用场景的不断扩展,人们对其运动控制的精度、稳定性都提出更高的要求,如何提高AUV在各种运动场景下的控制效果已成了重要的研究方向。

[0003] 在过去的几十年中,针对轨迹跟踪、路径点跟踪、路径规划和编队控制等不同应用场景,研究者们设计了各种AUV运动控制方法并验证了其有效性。其中具有代表性的是Refsnes等人提出的基于模型的输出反馈控制方法,该控制方法采用了两个解耦的系统模型:一个用于刻画海流负载的三自由度海流诱导船体模型和一个用于描述系统动态的五自由度模型。另外,Healey等人设计了一种基于状态反馈的跟踪控制方法,该控制方法采用固定的前向运动速度并对系统模型进行线性化处理,同时该控制方法采用了三个解耦的模型:纵荡模型、水平导向模型(横荡和艏摇)和垂向模型(垂荡和纵摇)。然而,这些方法都对系统模型进行了解耦或线性化处理,因此很难满足AUV在特定应用场景下的高精度控制要求。

[0004] 由于上述经典运动控制方法的局限性以及强化学习强大的自学习能力,近几年,研究者们对以强化学习为代表的智能控制方法表现出了极大的研究兴趣。而各种基于强化学习技术(例如Q学习、直接策略搜索、策略-评价网络和自适应强化学习)的智能控制方法也是不断地被提出并成功应用到不同的复杂应用场景中,如机器人运动控制、无人机飞行控制、高超音速飞行器跟踪控制以及道路信号灯控制等。基于强化学习的控制方法的核心思想是在无先验知识的前提下实现控制系统的性能优化。对于AUV系统,不少研究者已经设计出各种基于强化学习的控制方法并实际验证了其可行性。针对自主水下缆线跟踪控制问题,EI-Fakdi等人采用直接策略搜索技术来学习状态/动作映射关系,但是该方法仅适用于状态和动作空间都是离散的情况;而对于连续的动作空间,Paula等人采用径向基网络来近

似策略函数，然而由于径向基网络的函数近似能力较弱，该控制方法无法保证较高的跟踪控制精度。

[0005] 近年来，随着批学习、经验回放和批正则化等深度神经网络(DNN)训练技术的发展，深度强化学习在机器人运动控制、自主地面车辆运动控制、四旋翼控制和自动驾驶等复杂任务中表现出了优异性能。尤其是近期提出的深度Q网络(DQN)在许多极具挑战性的任务中都表现出人类水平的控制精度。然而DQN不能处理同时具有高维状态空间和连续动作空间的问题。在DQN的基础上，深度确定性策略梯度(DDPG)算法被进一步提出并实现了连续控制。然而DDPG使用目标评价网络来估计评价网络的目标值，使得评价网络不能有效地评价由策略网络学习到的策略，且学习到的动作值函数存在较大的方差，因此当DDPG应用于AUV轨迹跟踪控制问题时，无法满足较高的跟踪控制精度和稳定学习的要求。

发明内容

[0006] 本发明的目的是提出一种基于深度强化学习的AUV轨迹跟踪控制方法，该方法采用一种混合策略-评价网络结构，并采用多个准Q学习和确定性策略梯度来分别训练评价网络和策略网络，克服以往基于强化学习的方法控制精度较低、无法实现连续控制和学习过程不稳定等问题，实现高精度的AUV轨迹跟踪控制和稳定学习。

[0007] 为了实现上述目的，本发明采用如下技术方案：

[0008] 一种基于深度强化学习的自主水下航行器轨迹跟踪控制方法，该方法包括以下步骤：

[0009] 1)定义自主水下航行器AUV轨迹跟踪控制问题

[0010] 定义AUV轨迹跟踪控制问题包括四个部分：确定AUV系统输入、确定AUV系统输出、定义轨迹跟踪控制误差和建立AUV轨迹跟踪控制目标；具体步骤如下：

[0011] 1-1)确定AUV系统输入

[0012] 令AUV系统输入向量为$\tau_k=[\xi_k,\delta_k]^T$，其中$\xi_k$、$\delta_k$分别为AUV的螺旋桨推力和舵角，下标k表示第k个时间步；ξ_k、δ_k的取值范围分别为$[0,\overline{\xi}]$和$[-\overline{\delta},\overline{\delta}]$，$\overline{\xi}$、$\overline{\delta}$分别为最大的螺旋桨推力和最大舵角；

[0013] 1-2)确定AUV系统输出

[0014] 令AUV系统输出向量为$\eta_k=[x_k,y_k,\psi_k]^T$，其中$x_k$、$y_k$分别为第k个时间步AUV在惯性坐标系I-XYZ下沿X、Y轴的坐标，ψ_k为第k个时间步AUV前进方向与X轴的夹角；

[0015] 1-3)定义轨迹跟踪控制误差

[0016] 根据AUV的行驶路径选取参考轨迹$d_k=[x_k^d,y_k^d]^T$，定义第k个时间步的AUV轨迹跟踪控制误差为：

[0017] $$\mathbf{e}_k=[x_k-x_k^d,y_k-y_k^d]^T$$

[0018] 1-4)建立AUV轨迹跟踪控制目标

[0019] 对于步骤1-3)中的参考轨迹d_k，选择如下形式的目标函数：

[0020] $$P_k(\tau)=\sum_{i\geq k}\gamma^{i-k}(\mathbf{e}_i^T\mathbf{e}_i+\tau_i^T\mathbf{H}\tau_i)$$

[0021] 其中，γ是折扣因子，H为权重矩阵；

[0022] 建立AUV轨迹跟踪控制的目标为找到一个最优系统输入序列τ^*使得初始时刻的目标函数$P_0(\tau)$最小，计算公式如下：

[0023] $$\tau^* = \arg\min_{\tau} P_0(\tau)$$

[0024] 2)建立AUV轨迹跟踪问题的马尔科夫决策过程模型

[0025] 对步骤1)中的AUV轨迹跟踪问题进行马尔科夫决策过程建模，具体步骤如下：

[0026] 2-1)定义状态向量

[0027] 定义AUV系统的速度向量为$\Phi_k=[u_k,v_k,x_k]^T$，其中u_k、v_k分别为第k个时间步AUV沿前进方向、垂直于前进方向的线速度，x_k为第k个时间步AUV环绕前进方向的角速度；

[0028] 根据步骤1-2)确定的AUV系统输出向量η_k和步骤1-3)定义的参考轨迹，定义第k个时间步的状态向量如下：

[0029] $$s_k = [\eta_k^T, \phi_k^T, d_k^T, d_{k+1}^T]^T$$

[0030] 2-2)定义动作向量

[0031] 定义第k个时间步的动作向量为该时间步的AUV系统输入向量，即$a_k=\tau_k$；

[0032] 2-3)定义奖励函数

[0033] 第k个时间步的奖励函数用于刻画在状态s_k采取动作a_k的执行效果，根据步骤1-3)定义的轨迹跟踪控制误差e_k和步骤2-2)定义的动作向量a_k，定义第k个时间步的AUV奖励函数如下：

[0034] $$r_{k+1} = r(s_k, a_k) = -(\mathbf{e}_k^T \mathbf{e}_k + a_k^T \mathbf{H} a_k)$$

[0035] 2-4)将步骤1-4)建立的AUV轨迹跟踪控制的目标τ^*转换为强化学习框架下的AUV轨迹跟踪控制目标

[0036] 定义策略π为在某一状态下选择各个可能动作的概率，则定义动作值函数如下：

[0037] $$Q^{\pi}(s_k, a_k) = E_{r_{i>k}, s_{i>k}, a_{i>k}\sim\pi}\left[\sum_{i=k}^{K} \gamma^{i-k} r_{i+1} \mid s_k, a_k\right]$$

[0038] 其中，$E_{r_{i>k}, s_{i>k}, a_{i>k}\sim\pi}$表示对奖励函数、状态和动作的期望值；K为最大时间步；

[0039] 该动作值函数用于描述在当前及之后所有状态下均采取策略π时的期望累计折扣奖励，故在强化学习框架下，AUV轨迹跟踪控制目标是通过与AUV所处环境的交互来学习一个最优目标策略π^*，使得初始时刻的动作值最大，计算公式如下：

[0040] $$\pi^* = \arg\max_{\pi} E_{p(s_0), a_0\sim\pi} Q^{\pi}(s_0, a_0)$$

[0041] 其中，$p(s_0)$为初始状态s_0的分布；a_0为初始动作向量；

[0042] 将步骤1-4)建立的AUV轨迹跟踪控制的目标τ^*的求解转换为π^*的求解；

[0043] 2-5)简化强化学习框架下的AUV轨迹跟踪控制目标

[0044] 通过如下迭代贝尔曼方程来求解步骤2-4)中的动作值函数：

[0045] $$Q^{\pi}(s_k, a_k) = E_{r_{k+1}, s_{k+1}}\left[r_{k+1} + \gamma E_{a_{k+1}\sim\pi}\left[Q^{\pi}(s_{k+1}, a_{k+1})\right]\right]$$

[0046] 设策略π是确定性的，即从AUV的状态向量空间到AUV的动作向量空间是一一映射的关系，并记为μ，则将上述迭代贝尔曼方程简化为：

[0047] $$Q^{\mu}(s_k,a_k)=E_{r_{k+1},s_{k+1}}\left[r_{k+1}+\gamma Q^{\mu}(s_{k+1},\mu(s_{k+1}))\right]$$

[0048] 对于确定性的策略μ，将步骤2-4)中的最优目标策略π*简化为确定性最优目标策略μ*：

[0049] $$\mu^{*}=\arg\max_{\mu}E_{p(s_0)}Q^{\mu}(s_0,\mu(s_0))$$

[0050] 3)构建混合策略-评价网络

[0051] 通过构建混合策略-评价网络来分别估计确定性最优目标策略μ*和对应的最优动作值函数$Q^{\mu^{*}}$，构建混合策略-评价网络包括三部分：构建策略网络、构建评价网络和确定目标策略，具体步骤如下：

[0052] 3-1)构建策略网络

[0053] 混合策略-评价网络结构通过构建n个策略网络$\mu_{\theta_p}(s_k)$来估计确定性最优目标策略μ*；其中，θ_p为第p个策略网络的权重参数，p＝1,…,n；各策略网络均分别使用一个全连接的深度神经网络来实现，各策略网络均分别包含一个输入层、两个隐藏层和一个输出层；各策略网络的输入为状态向量s_k，各策略网络的输出为动作向量a_k；

[0054] 3-2)构建评价网络

[0055] 混合策略-评价网络结构通过构建m个评价网络$Q_{w_q}(s_k,a_k)$来估计最优动作值函数$Q^{\mu^{*}}$；其中，w_q为第q个评价网络的权重参数，q＝1,…,m；各评价网络均分别使用一个全连接的深度神经网络来实现，各评价网络均分别包含一个输入层、两个隐藏层和一个输出层；各评价网络的输入为状态向量s_k和动作向量a_k，其中状态向量s_k从输入层输入到各评价网络，动作向量a_k从第一个隐藏层输入到各评价网络，各评价网络输出为在状态向量s_k下采取动作向量a_k的动作值；

[0056] 3-3)确定目标策略

[0057] 根据所构建的混合策略-评价网络，将第k个时间步学习到的AUV轨迹跟踪控制的目标策略$\mu_f(s_k)$定义为n个策略网络输出的均值，计算公式如下：

[0058] $$a_k=\mu_f(s_k)=\frac{1}{n}\sum_{p=1}^{n}\mu_{\theta_p}(s_k)$$

[0059] 4)求解AUV轨迹跟踪控制的目标策略$\mu_f(s_k)$，具体步骤如下：

[0060] 4-1)参数设置

[0061] 分别设置最大迭代次数M、每次迭代的最大时间步K、经验回放抽取的训练集大小N、各评价网络的学习率α_ω、各策略网络的学习率α_θ、折扣因子γ和奖励函数中的权重矩阵H；

[0062] 4-2)初始化混合策略-评价网络

[0063] 随机初始化n个策略网络$\mu_{\theta_p}(s_k)$和m个评价网络$Q_{w_q}(s_k,a_k)$的权重参数θ_p和w_q；从n个策略网络中随机选择第d个策略网络记为μ_{θ_d}，d＝1,…,n；

[0064] 构建经验列队集合R，设该经验列队集合R的最大容量为B，并初始化为空；

[0065] 4-3)迭代开始，对混合策略-评价网络进行训练，初始化迭代次数episode＝1；

[0066] 4-4)设置当前时间步k＝0,随机初始化AUV的状态变量s_0,令当前时间步的状态变量s_k＝s_0;并产生一个探索噪声$Noise_k$;

[0067] 4-5)根据n个当前策略网络μ_{θ_p}和探索噪声$Noise_k$确定当前时间步的动作向量a_k为:

[0068] $$a_k=\frac{1}{n}\sum_{p=1}^{n}\mu_{\theta_p}(s_k)+Noise_k$$

[0069] 4-6)AUV在当前状态s_k下执行动作a_k,根据步骤2-3)得到奖励函数r_{k+1},并观测到一个新的状态s_{k+1};记e_k＝(s_k,a_k,r_{k+1},s_{k+1})为一个经验样本;如果经验列队集合R的样本数量已经达到最大容量B,则先删除最先加入的一个样本,再将经验样本e_k存入经验列队集合R中;否则直接将经验样本e_k存入经验列队集合R中;

[0070] 从经验列队集合R中选取A个经验样本,具体如下:当经验列队集合R中样本数量不超过N时,则选取该经验列队集合R中的所有经验样本;当经验列队集合R超过N时,则从该经验列队集合R中随机选取N个经验样本(s_l,a_l,r_{l+1},s_{l+1});

[0071] 4-7)根据选取的A个经验样本计算每个评价网络的期望贝尔曼绝对误差$EBAE_q$,用于表征每个评价网络的性能,公式如下:

[0072] $$EBAE_q=\frac{1}{A}\sum_{l}\left|Q_{w_q}(s_l,a_l)-r_{l+1}-\gamma Q_{w_q}(s_{l+1},\mu_{\theta_d}(s_{l+1}))\right|,\quad q=1,\cdots,m$$

[0073] 选择性能最差的评价网络,通过以下公式求得该性能最差的评价网络的序号,记为c:

[0074] $$c=\arg\max_{q} EABE_q$$

[0075] 4-8)由第c个评价网络Q_{w_c},通过如下次贪婪策略得到每个经验样本在下一时间步的动作向量:

[0076] $$a_{l+1}=\arg\max_{a} Q_{w_c}(s_{l+1},a),\quad s.t.\ a\in\left\{\mu_{\theta_p}(s_{l+1}),\ p=1,\cdots,n\right\}$$

[0077] 4-9)通过多个准Q学习方法计算第c个评价网络的目标值Y_{l+1}^{c},公式如下:

[0078] $$Y_{l+1}^{c}=r_{l+1}+\frac{\gamma}{m-1}\sum_{q=1,q\neq c}^{m}Q_{w_q}(s_{l+1},a_{l+1})$$

[0079] 4-10)计算第c个评价网络的损失函数L(w_c),公式如下:

[0080] $$L(w_c)=\frac{1}{A}\sum_{l}\left(Q_{w_c}(s_l,a_l)-Y_{l+1}^{c}\right)^2$$

[0081] 4-11)通过损失函数L(w_c)对权重参数w_c的导数来更新第c个评价网络的权重参数,公式如下:

[0082] $$w_c=w_c+\alpha_{\omega}\nabla_{w_c}L(w_c)$$

[0083] 其余评价网络的权重参数保持不变;

[0084] 4-12)从n个策略网络中随机选择一个策略网络来重置第d个策略网络μ_{θ_d};

[0085] 4-13)根据更新后的第c个评价网络计算第d个策略网络μ_{θ_d}的确定性策略梯度

$\nabla_{\theta_d} J$ 并以此更新第d个策略网络 μ_{θ_d} 的权重参数θ_d,计算公式分别如下:

[0086] $$\nabla_{\theta_d} J = \frac{1}{A}\sum_{l} \nabla_a Q_{w_c}(s_l, a)|_{a=\mu_{\theta_d}(s_l)} \nabla_{\theta_d} \mu_{\theta_d}(s_l)$$

[0087] $$\theta_d = \theta_d - \alpha_\theta \nabla_{\theta_d} J$$

[0088] 其余策略网络的权重参数保持不变;

[0089] 4-14)令k=k+1并对k进行判定:如k<K,则重新返回步骤4-5),AUV继续跟踪参考轨迹;否则,进入步骤4-15);

[0090] 4-15)令episode=episode+1并对episode进行判定:如episode<M,则重新返回步骤4-4),AUV进行下一个迭代过程;否则,进入步骤4-16);

[0091] 4-16)迭代结束,终止混合策略-评价网络的训练过程,将迭代终止时的n个策略网络的输出值通过步骤3-3)中的计算公式得到最终AUV轨迹跟踪控制的目标策略$\mu_f(s_k)$,由该目标策略实现对AUV的轨迹跟踪控制。

[0092] 本发明的特点及有益效果:

[0093] 本发明提出的方法采用了多个策略网络和评价网络。对于多个评价网络,通过定义期望贝尔曼绝对误差来评估每个评价网络的性能,在每个时间步只更新性能最差的一个评价网络,不同于已有基于强化学习的控制方法,本发明提出多个准Q学习方法来计算更为准确的评价网络目标值,该方法可以解决动作值函数过估计问题,并且可以在不借助目标评价网络的前提下稳定学习过程。对于多个策略网络,在每个时间步随机选择一个策略网络,并采用确定性策略梯度进行更新。最终学习到的策略为所有策略网络的均值。

[0094] 1)本发明提出的AUV轨迹跟踪控制方法不依赖于模型,通过AUV在行驶过程中的采样数据,来自主学习出使得控制目标达到最优的目标策略,该过程不需要对AUV模型做任何假设,尤其适用于在复杂深海环境下工作的AUV,有很高的实际应用价值。

[0095] 2)本发明方法采用多个准Q学习来得到比已有方法更加准确的评价网络目标值,既减小了由评价网络近似得到的动作值函数的方差,还解决了动作值函数过估计问题,从而得到更优的目标策略,实现高精度的AUV轨迹跟踪控制。

[0096] 3)本发明方法基于期望贝尔曼绝对误差来决定每个时间步该更新哪一个评价网络,这种更新规则可以减弱较差评价网络的影响,从而保证学习过程的快速收敛。

[0097] 4)本发明方法由于采用了多个评价网络,其学习过程不易受到恶劣的AUV历史跟踪轨迹的影响,鲁棒性好,学习过程稳定。

[0098] 5)本发明方法将强化学习与深度神经网络相结合,具有很强的自学习能力,能够在不确定的深海环境中实现对AUV的高精度自适应控制,在AUV轨迹跟踪、水下避障等场景中有着很好的应用前景。

附图说明

[0099] 图1是本发明提出方法与现有DDPG方法的性能对比图;其中,图(a)为学习曲线对比图,图(b)为AUV轨迹跟踪效果对比图。

[0100] 图2是本发明提出方法与神经网络PID方法的性能对比图;其中,图(a)为AUV沿X、Y方向的坐标轨迹跟踪效果对比图,图(b)为AUV在X、Y方向的跟踪误差对比图。

具体实施方式

[0101] 本发明提出的一种基于深度强化学习的自主水下航行器轨迹跟踪控制方法，下面结合附图和具体实施例进一步详细说明如下。

[0102] 本发明提出了一种基于深度强化学习的自主水下航行器跟踪控制算法，主要包括四个部分：定义AUV轨迹跟踪控制问题、建立AUV轨迹跟踪问题的马尔科夫决策过程模型、构建混合策略-评价网络结构和求解AUV轨迹跟踪控制的目标策略。

[0103] 1)定义AUV轨迹跟踪控制问题

[0104] 定义AUV轨迹跟踪控制问题包括四个组成部分：确定AUV系统输入、确定AUV系统输出、定义轨迹跟踪控制误差和建立AUV轨迹跟踪控制目标；具体步骤如下：

[0105] 1-1)确定AUV系统输入

[0106] 令AUV系统输入向量为$\tau_k=[\xi_k,\delta_k]^T$，其中$\xi_k$、$\delta_k$分别为AUV的螺旋桨推力和舵角，下标k表示第k个时间步即时刻k·t的取值，其中t为时间步长，下同；ξ_k、δ_k的取值范围分别为$[0,\bar{\xi}]$和$[-\bar{\delta},\bar{\delta}]$，其中$\bar{\xi}$、$\bar{\delta}$分别为最大的螺旋桨推力和最大舵角，根据AUV所采用的螺旋桨型号确定。

[0107] 1-2)确定AUV系统输出

[0108] 令AUV系统输出向量为$\eta_k=[x_k,y_k,\psi_k]^T$，其中$x_k$、$y_k$分别为第k个时间步AUV在惯性坐标系I-XYZ下沿X、Y轴的坐标，ψ_k为第k个时间步AUV前进方向与X轴的夹角。

[0109] 1-3)定义轨迹跟踪控制误差

[0110] 根据AUV的行驶路径选取参考轨迹$d_k=[x_k^d,y_k^d]^T$，定义第k个时间步的AUV轨迹跟踪控制误差为：

[0111] $$\mathbf{e}_k=[x_k-x_k^d,y_k-y_k^d]^T$$

[0112] 1-4)建立AUV轨迹跟踪控制目标

[0113] 对于步骤1-3)中的参考轨迹d_k，选择如下形式的目标函数：

[0114] $$P_k(\tau)=\sum_{i\geq k}\gamma^{i-k}(\mathbf{e}_i^T\mathbf{e}_i+\tau_i^T\mathbf{H}\tau_i)$$

[0115] 其中，γ是折扣因子，H为权重矩阵；

[0116] 建立AUV轨迹跟踪控制的目标为找到一个最优系统输入序列τ^*使得初始时刻的目标函数$P_0(\tau)$最小，计算公式如下：

[0117] $$\tau^*=\arg\min_{\tau}P_0(\tau)$$

[0118] 2)建立AUV轨迹跟踪问题的马尔科夫决策过程模型

[0119] 马尔科夫决策过程(MDP)是强化学习理论的基础，因此需要对步骤1)中的AUV轨迹跟踪问题进行MDP建模。强化学习的主要元素包括智能体、环境、状态、动作和奖励函数，智能体的目标是通过与AUV所处环境的交互来学习一个最优动作(或控制输入)序列来最大化累计奖励(或最小化累计跟踪控制误差)，进而实现AUV轨迹跟踪目标的求解。具体步骤如下：

[0120] 2-1)定义状态向量

[0121] 定义AUV系统的速度向量为$\Phi_k=[u_k,v_k,x_k]^T$，其中u_k、v_k分别为第k个时间步AUV沿前进方向、垂直于前进方向的线速度，x_k为第k个时间步AUV环绕前进方向的角速度。

[0122] 根据步骤1-2)确定的AUV系统输出向量η_k和步骤1-3)定义的参考轨迹,定义第k个时间步的状态向量如下:

[0123] $$s_k=[\eta_k^T,\phi_k^T,d_k^T,d_{k+1}^T]^T$$

[0124] 2-2)定义动作向量

[0125] 定义第k个时间步的动作向量为该时间步的AUV系统输入向量,即:$a_k=\tau_k$。

[0126] 2-3)定义奖励函数

[0127] 第k个时间步的奖励函数用于刻画在状态s_k采取动作a_k的执行效果,根据步骤1-3)定义的轨迹跟踪控制误差e_k和步骤2-2)定义的动作向量a_k,定义第k个时间步的AUV奖励函数如下:

[0128] $$r_{k+1}=r(s_k,a_k)=-(\mathbf{e}_k^T\mathbf{e}_k+a_k^T\mathbf{H}a_k)$$

[0129] 2-4)将步骤1-4)建立的AUV轨迹跟踪控制的目标τ^*转换为强化学习框架下的AUV轨迹跟踪控制目标

[0130] 定义策略π为在某一状态下选择各个可能动作的概率,则定义动作值函数如下:

[0131] $$Q^{\pi}(s_k,a_k)=E_{r_{i>k},s_{i>k},a_{i>k}\sim\pi}\left[\sum_{i=k}^{K}\gamma^{i-k}r_{i+1}\,|\,s_k,a_k\right]$$

[0132] 其中,$E_{r_{i>k},s_{i>k},a_{i>k}\sim\pi}$表示对奖励函数、状态和动作的期望值(下同);K为最大时间步;

[0133] 该动作值函数用于描述在当前及之后所有状态下均采取策略π时的期望累计折扣奖励,因此,在强化学习框架下,AUV轨迹跟踪控制目标(即智能体的目标)是通过与AUV所处环境的交互来学习一个最优目标策略π^*,使得初始时刻的动作值最大,即:

[0134] $$\pi^*=\arg\max_{\pi}E_{p(s_0),a_0\sim\pi}Q^{\pi}(s_0,a_0)$$

[0135] 其中,$p(s_0)$为初始状态s_0的分布;a_0为初始动作向量。

[0136] 因此,步骤1-4)建立的AUV轨迹跟踪控制的目标τ^*的求解可转换为π^*的求解。

[0137] 2-5)简化强化学习框架下的AUV轨迹跟踪控制目标

[0138] 类似于动态规划,许多强化学习方法使用如下迭代贝尔曼方程来求解步骤2-4)中的动作值函数:

[0139] $$Q^{\pi}(s_k,a_k)=E_{r_{k+1},s_{k+1}}\left[r_{k+1}+\gamma E_{a_{k+1}\sim\pi}\left[Q^{\pi}(s_{k+1},a_{k+1})\right]\right]$$

[0140] 假定策略π是确定性的,即从AUV的状态向量空间到AUV的动作向量空间是一一映射的关系,并记为μ,于是上述迭代贝尔曼方程可以简化为:

[0141] $$Q^{\mu}(s_k,a_k)=E_{r_{k+1},s_{k+1}}\left[r_{k+1}+\gamma Q^{\mu}(s_{k+1},\mu(s_{k+1}))\right]$$

[0142] 此外,对于确定性的策略μ,将步骤2-4)中的最优目标策略π^*简化为确定性最优目标策略μ^*:

[0143] $$\mu^*=\arg\max_{\mu}E_{p(s_0)}Q^{\mu}(s_0,\mu(s_0))$$

[0144] 3)构建混合策略-评价网络

[0145] 由步骤2-5)可知,利用强化学习求解AUV轨迹跟踪问题的核心是如何求解确定性

最优目标策略μ*和对应的最优动作值函数Q^{μ^*}。本发明方法采用一种混合策略-评价网络来分别估计μ*和Q^{μ^*}。构建混合策略-评价网络包括三部分：构建策略网络、构建评价网络和确定目标策略，具体步骤如下：

[0146] 3-1)构建策略网络

[0147] 混合策略-评价网络结构通过构建n(为了平衡本发明算法跟踪控制精度与网络训练速度，其取值不宜过大也不宜过小)个策略网络$\mu_{\theta_p}(s_k)$来估计确定性最优目标策略μ*。其中，θ_p为第p个策略网络的权重参数，p＝1,…,n；各策略网络均分别使用一个全连接的深度神经网络来实现，每个策略网络均分别包含一个输入层、两个隐藏层和一个输出层，各策略网络的输入为状态向量s_k，各策略网络输出为动作向量a_k，两个隐藏层分别含有400和300个单元。

[0148] 3-2)构建评价网络

[0149] 混合策略-评价网络结构通过构建m(评价网络数量的选取依据与上述策略网络数量的选取依据相同)个评价网络$Q_{w_q}(s_k,a_k)$来估计最优动作值函数Q^{μ^*}。其中，w_q为第q个评价网络的权重参数，q＝1,…,m；各评价网络均分别使用一个全连接的深度神经网络来实现，各评价网络均分别包含一个输入层、两个隐藏层和一个输出层，两个隐藏层分别含有400和300个单元；各评价网络的输入为状态向量s_k和动作向量a_k，其中状态向量s_k从输入层输入到各评价网络，动作向量a_k从第一个隐藏层输入到各评价网络，各评价网络输出为在状态向量s_k下采取动作向量a_k的动作值。

[0150] 3-3)确定目标策略

[0151] 根据所构建的混合策略-评价网络，将第k个时间步学习到的AUV轨迹跟踪控制的目标策略$\mu_f(s_k)$定义为n个策略网络输出的均值，计算公式如下：

[0152] $$a_k=\mu_f(s_k)=\frac{1}{n}\sum_{p=1}^{n}\mu_{\theta_p}(s_k)$$

[0153] 4)求解AUV轨迹跟踪控制的目标策略$\mu_f(s_k)$，具体步骤如下：

[0154] 4-1)参数设置

[0155] 分别设置最大迭代次数M、每次迭代的最大时间步K、经验回放抽取的训练集大小N、各评价网络的学习率α_ω、各策略网络的学习率α_θ、折扣因子γ和奖励函数中的权重矩阵H；本实施例中，M＝1500，K＝1000(每个时间步长t＝0.2s)，N＝64，各评价网络的α_ω＝0.01，各策略网络的α_θ＝0.001，γ＝0.99，H＝[0.001,0;0,0.001]；

[0156] 4-2)初始化混合策略-评价网络

[0157] 随机初始化n个策略网络$\mu_{\theta_p}(s_k)$和m个评价网络$Q_{w_q}(s_k,a_k)$的权重参数θ_p和w_q；从n个策略网络中随机选择第d(d＝1,…,n)个策略网络记为μ_{θ_d}；

[0158] 构建经验列队集合R，设该经验列队集合R的最大容量为B(本实施例B＝10000)，并初始化为空；

[0159] 4-3)迭代开始，对混合策略-评价网络进行训练，初始化迭代次数episode＝1；

[0160] 4-4)设置当前时间步k＝0，随机初始化AUV的状态变量s_0，令当前时间步的状态变

量$s_k=s_0$;并产生一个探索噪声$Noise_k$(本实施例采用奥恩斯坦-乌伦贝克(Ornstein-Uhlenbeck)探索噪声);

[0161] 4-5)根据n个当前策略网络μ_{θ_p}和探索噪声$Noise_k$确定当前时间步的动作向量a_k为:

[0162] $$a_k=\frac{1}{n}\sum_{p=1}^{n}\mu_{\theta_p}(s_k)+Noise_k$$

[0163] 4-6)AUV在当前状态s_k下执行动作a_k,根据步骤2-3)得到奖励函数r_{k+1},并观测到一个新的状态s_{k+1};记$e_k=(s_k,a_k,r_{k+1},s_{k+1})$为一个经验样本;如果经验列队集合R的样本数量已经达到最大容量B,则先删除最先加入的一个样本,再将经验样本e_k存入经验列队集合R中;否则直接将经验样本e_k存入经验列队集合R中;

[0164] 从经验列队集合R中选取A个经验样本,A≤N,具体如下:当经验列队集合R中样本数量不超过N时,则选取该经验列队集合R中的所有经验样本;当经验列队集合R超过N时,则从该经验列队集合R中随机选取N个经验样本$(s_1,a_1,r_{1+1},s_{1+1})$,1为被选择的经验样本所在的时间步;

[0165] 4-7)根据选取的A个经验样本计算每个评价网络的期望贝尔曼绝对误差$EBAE_q$,用于表征每个评价网络的性能,公式如下:

[0166] $$EBAE_q=\frac{1}{A}\sum_{l}\left|Q_{w_q}(s_l,a_l)-r_{l+1}-\gamma Q_{w_q}(s_{l+1},\mu_{\theta_d}(s_{l+1}))\right|,\quad q=1,\cdots,m$$

[0167] 选择性能最差的评价网络,通过以下公式求得该性能最差的评价网络的序号,记为c:

[0168] $$c=\arg\max_{q} EABE_q$$

[0169] 4-8)由第c个评价网络Q_{w_c},通过如下次贪婪策略得到每个经验样本在下一时间步的动作向量:

[0170] $$a_{l+1}=\arg\max_{a} Q_{w_c}(s_{l+1},a),\quad s.t.\ a\in\left\{\mu_{\theta_p}(s_{l+1}),\ p=1,\cdots,n\right\}$$

[0171] 4-9)通过多个准Q学习方法计算第c个评价网络的目标值Y_{l+1}^{c},公式如下:

[0172] $$Y_{l+1}^{c}=r_{l+1}+\frac{\gamma}{m-1}\sum_{q=1,q\neq c}^{m}Q_{w_q}(s_{l+1},a_{l+1})$$

[0173] 4-10)计算第c个评价网络的损失函数$L(w_c)$,公式如下:

[0174] $$L(w_c)=\frac{1}{A}\sum_{l}\left(Q_{w_c}(s_l,a_l)-Y_{l+1}^{c}\right)^2$$

[0175] 4-11)通过损失函数$L(w_c)$对权重参数w_c的导数来更新第c个评价网络的权重参数,公式如下:

[0176] $$w_c=w_c+\alpha_\omega\nabla_{w_c}L(w_c)$$

[0177] 其余评价网络的权重参数保持不变;

[0178] 4-12)从n个策略网络中随机选择一个策略网络来重置第d个策略网络μ_{θ_d};

[0179] 4-13)根据更新后的第c个评价网络计算第d个策略网络μ_{θ_d}的确定性策略梯度$\nabla_{\theta_d}J$并以此更新第d个策略网络μ_{θ_d}的权重参数θ_d,计算公式分别如下:

[0180] $$\nabla_{\theta_d}J=\frac{1}{A}\sum_{l}\nabla_{a}Q_{w_c}(s_l,a)|_{a=\mu_{\theta_d}(s_l)}\nabla_{\theta_d}\mu_{\theta_d}(s_l)$$

[0181] $$\theta_d=\theta_d-\alpha_\theta\nabla_{\theta_d}J$$

[0182] 其余策略网络的权重参数保持不变。

[0183] 4-14)令k=k+1并对k进行判定:如k<K,则重新返回步骤4-5),AUV继续跟踪参考轨迹;否则,进入步骤4-15)。

[0184] 4-15)令episode=episode+1并对episode进行判定:如episode<M,则重新返回步骤4-4),AUV进行下一个迭代过程;否则,进入步骤4-16)。

[0185] 4-16)迭代结束,终止混合策略-评价网络的训练过程,将迭代终止时的n个策略网络的输出值通过步骤3-3)中的计算公式得到最终AUV轨迹跟踪控制的目标策略$\mu_f(s_k)$,由该目标策略实现对AUV的轨迹跟踪控制。

[0186] 本发明实施例的有效性验证

[0187] 本发明所提出的基于深度强化学习的AUV轨迹跟踪控制方法(以下简称MPQ-DPG)的性能分析如下所示,所有对比实验均是基于广泛使用的REMUS自主无人航行器,其最大螺旋桨推力$\overline{\xi}$和舵角$\overline{\delta}$分别为86N和0.24rad;且采用如下参考轨迹:

[0188] $$x^d=(15-0.1t)\cos(\frac{\pi}{20}t),y^d=(15-0.1t)\sin(\frac{\pi}{20}t)$$

[0189] 此外,在本发明实施例中,评价网络数量m与策略网络数量n相同,后文统一记为n。

[0190] 1)MPQ-DPG与现有的DDPG方法对比分析

[0191] 图1为本发明提出的深度强化学习的AUV提出轨迹跟踪控制方法(MPQ-DPG)与现有DDPG方法在训练过程中的学习曲线和轨迹跟踪效果上的比较。其中,图(a)中的学习曲线是通过五次独立实验得到,图(b)中Ref表示参考轨迹。

[0192] 分析图1,可得如下结论:

[0193] a)相对于DDPG方法,MPQ-DPG的学习稳定性更好,这是由于MPQ-DPG采用多个评价网络和策略网络,可以降低差样本对学习稳定性的影响。

[0194] b)MPQ-DPG方法最终收敛的平均累计奖励明显高于DDPG方法,这说明了MPQ-DPG方法的跟踪控制精度要明显高于DDPG方法。

[0195] c)从图1(b)中可以观察到,MPQ-DPG方法得到的跟踪轨迹几乎与参考轨迹重合,说明MPQ-DPG方法可以实现高精度的AUV跟踪控制。

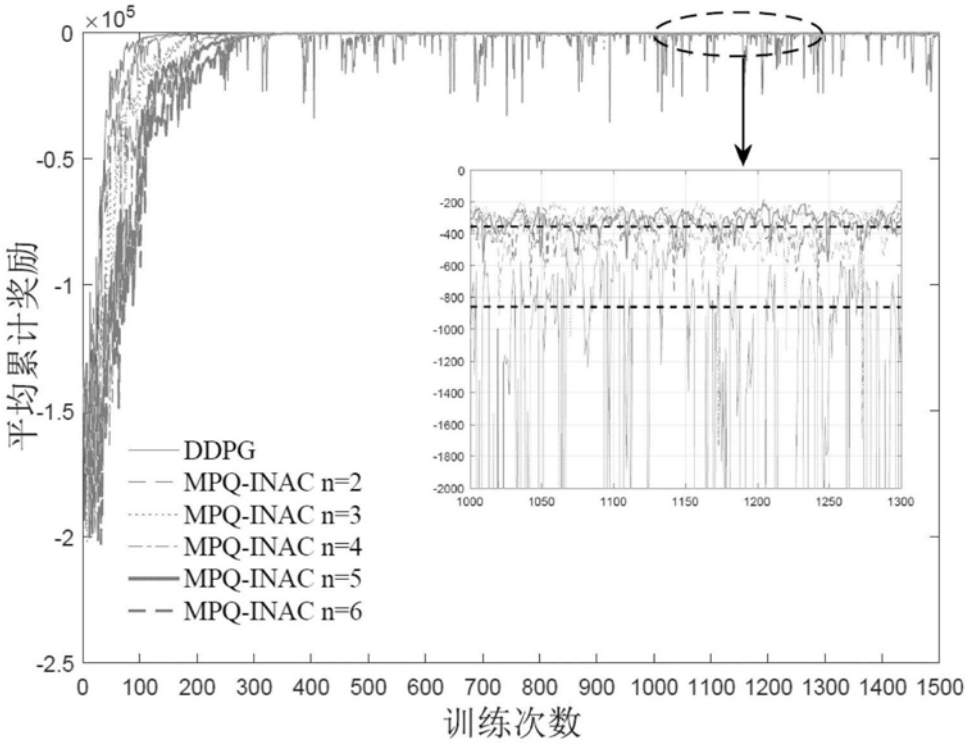
[0196] d)随着策略网络和评价网络数量的增大,MPQ-DPG方法的跟踪控制精度会逐渐提高,但提高的幅度在n>4之后将不再明显。

[0197] 2)MPQ-DPG方法与现有神经网络PID方法对比分析

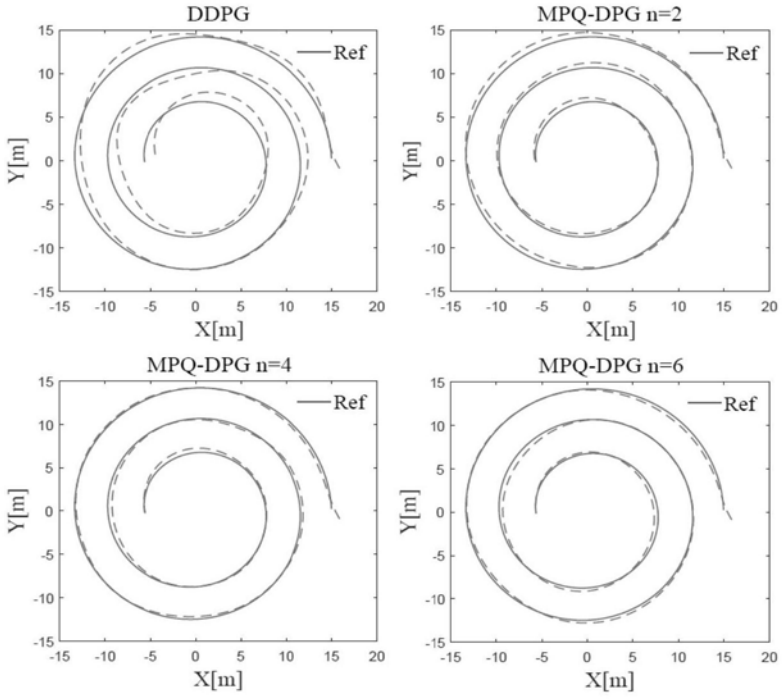
[0198] 图2为本发明为水下无人航行器轨迹跟踪控制提出的MPQ-DPG方法与神经网络PID方法在坐标轨迹跟踪曲线和坐标轨迹跟踪误差上的比较。图中Ref表示参考坐标轨迹,PIDNN表示神经网络PID算法,n=4。

[0199] 分析图2可得，神经网络PID控制方法的跟踪性能明显差于本发明提出的MPQ-DPG方法；此外，图2(b)中的跟踪误差表明，MPQ-DPG方法可以实现误差更快的收敛，特别是在起始阶段，MPQ-DPG方法仍然可以实现快速、高精度的跟踪性能，而神经网络PID方法的响应时间要明显长于MPQ-DPG方法，且跟踪误差的收敛性较差。

[0200] 上述实施例为本发明较佳的实施方式，但本发明的实施方式并不受上述实施例的限制，其他的任何未背离本发明的精神实质与原理下所作的改变、修饰、替代、组合、简化，均应为等效的置换方式，都包含在本发明的保护范围之内。

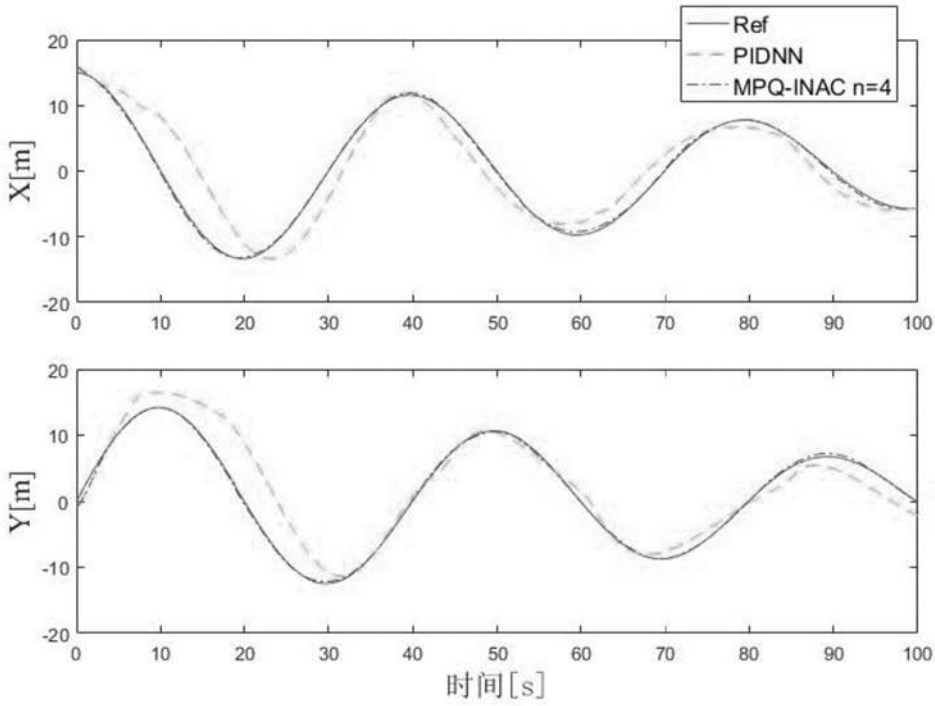

×10^5
平均累计奖励
训练次数
DDPG
MPQ-INAC n=2
MPQ-INAC n=3
MPQ-INAC n=4
MPQ-INAC n=5
MPQ-INAC n=6

（a）

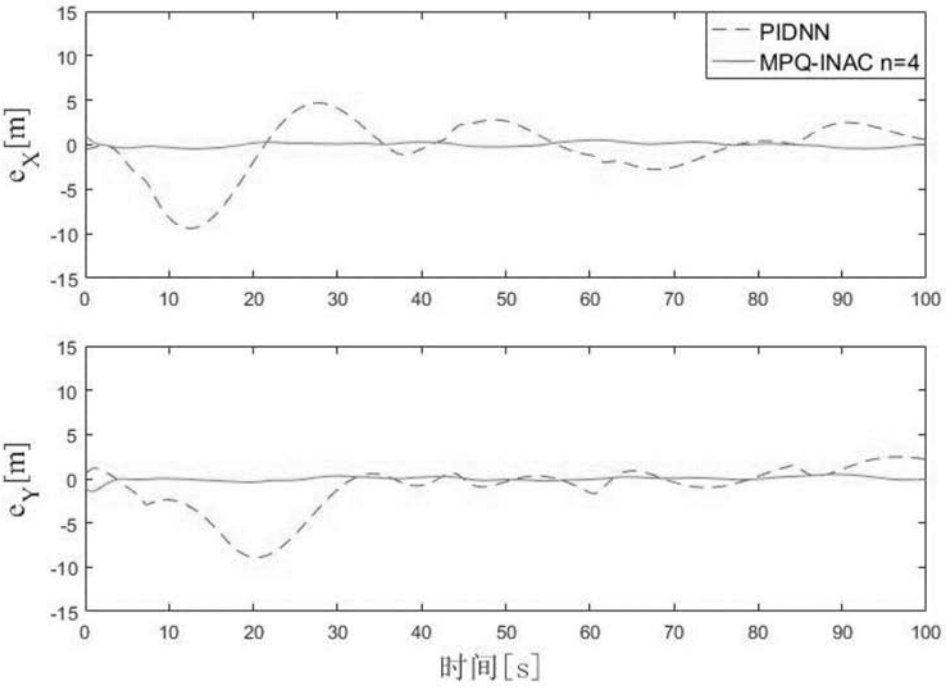

DDPG
MPQ-DPG n=2
MPQ-DPG n=4
MPQ-DPG n=6
Ref
X[m]
Y[m]

（b）

图1



（a）



（b）

图2