DOI: 10.19650/j.cnki.cjsi.J2413479

基于深度强化学习的变步长 LMS 算法*

徐君阳[1],张红梅[1],张 坤[2]
(1. 武汉大学电气与自动化学院 武汉 430072; 2. 长江水利委员会水文局长江口水文水资源勘测局 上海 200136)

摘 要:针对定步长 LMS 算法在收敛速度和稳态误差之间难以取得平衡的问题以及传统变步长算法对初始参数选择依赖程度高、工作量大且存在主观性的缺陷,提出了一种基于深度强化学习的变步长 LMS 算法。该算法对初始参数的依赖性小,规避了繁琐的调参流程。首先,构建了一个融合深度强化学习和自适应滤波的算法模型,该模型利用深度强化学习智能体控制步长因子的变化,代替了传统变步长算法中用于步长调整的非线性函数,从而规避了繁琐的实验调参流程,降低了算法使用的复杂性。其次,提出了基于误差的状态奖励和基于步长的动作奖励函数,引入动态奖励与负奖励机制,有效提升算法的收敛速度。此外,设计了基于欠完备编码器的网络结构,提高了强化学习策略的推理能力。通过实验验证,相较于其他较新的变步长算法,所提出的算法具有更快的收敛速度和更小的稳态误差,在不同初始参数下均能快速调整至合理的步长值,减少了实验调参的工作量。将训练完成的网络应用到系统辨识、信号去噪以及截流区龙口水域水位信号的滤波等实际领域中,均取得了良好的性能表现,证明了算法具有一定的泛化能力,并进一步证实了其有效性。

关键词: 变步长 LMS 算法;深度强化学习;自适应滤波;奖励函数
中图分类号: TN911.7 TH701 **文献标识码**: A **国家标准学科分类代码**: 510.40

A variable step size LMS algorithm based on deep reinforcement learning

Xu Junyang[1], Zhang Hongmei[1], Zhang Kun[2]
(*1. School of Electrical Engineering and Automation, Wuhan University, Wuhan 430072, China; 2. The Survey Bureau of Hydrology and Water Resources of Changjiang Estuary, Shanghai 200136, China*)

Abstract: This article proposes a variable step size LMS algorithm based on deep reinforcement learning to address the problem of the difficult balance between convergence speed and steady-state error in the fixed step size LMS algorithm, as well as the high dependence on initial parameter selection, heavy workload, and subjective defects of traditional variable step size algorithms. This algorithm has a low dependence on initial parameters and avoids the cumbersome parameter tuning process. Firstly, an algorithm model integrating deep reinforcement learning and adaptive filtering is constructed, which utilizes deep reinforcement learning agents to control the change of step size factors, replacing the nonlinear function used for step size adjustment in traditional variable step size algorithms, thereby avoiding the cumbersome experimental parameter tuning process and reducing the complexity of algorithm use. Secondly, the error-based state reward and step size-based action reward functions are proposed. Dynamic rewards and negative reward mechanisms are introduced, which effectively improves the convergence speed of the algorithm. In addition, a network architecture based on incomplete encoders is designed to improve the inference ability of reinforcement learning strategies. Through experimental verification, compared with other newer variable step size algorithms, the algorithm proposed in this article can quickly adjust to a reasonable step size value under different initial parameters and reduce the workload of experimental parameter tuning, obtaining faster convergence speed and smaller steady-state error. The trained network has been applied to practical fields, such as system identification, signal denoising, and filtering of water level signals at the closure gap, and has achieved good performance, further confirming the generalization ability and effectiveness of the algorithm.

Keywords: variable step LMS; deep reinforcement learning; adaptive filter; reward function

收稿日期:2024-11-11 Received Date: 2024-11-11
*基金项目:国家自然科学基金(42176186)项目资助

0 引　　言

随着计算机技术的飞速发展以及处理器性能的显著提升,自适应滤波器在系统辨识、噪声消除、逆模型构建、波束形成等众多领域中获得了日益广泛的应用[1-2]。其应用范围广泛,涵盖了水文参数的测量[3]、医学诊断[4]、电力系统的回波检测[5]等多个方面。最小均方自适应滤波算法(least mean square, LMS)最初由 Widrow 等[6]提出,该算法以其广泛的适用场景、较低的硬件需求、强大的鲁棒性以及易于理解和实现的特点而著称。在缺乏先验知识的情况下,LMS 算法能够持续调整滤波器结构并跟踪信号的变化,以实现最佳的滤波效果。

定步长 LMS 算法由于其步长因子固定不变,在理论分析中难以实现收敛速度与稳态误差之间的优化平衡[7]。因此,众多研究者对 LMS 算法的步长调整策略进行了深入探讨[8-9]。覃景繁等[10]提出了基于 Sigmoid 函数的变步长 LMS 算法,有效缓解了传统 LMS 算法在收敛性能与稳态误差之间的矛盾。张红梅等[11]在前述算法基础上进行了深入研究与改进,提出了一种性能更优的变步长 LMS 算法。火元莲等[12]提出了基于反双曲正切函数的变步长 LMS 算法,也取得了显著成效。作者提出了一种基于对数函数的变步长 LMS 算法,相比反正切函数的 LMS 算法该算法有更快的收敛速度,但同时作者也指出,非线性函数的参数对算法的性能起决定性作用,通过大量仿真实验遴选最佳参数的过程极为耗时,怎样快速和更好的选择参数是一个值得深入研究的方向。

综上,基于非线性函数的变步长方法普遍存在涉及众多参数的问题,如何实现参数的快速甄选,同时削减算法对参数的过度依赖,成为了值得深入研究的课题。针对这一问题,解本巨等[14]引入深度学习方法,提出了一种依托反向传播(back propagation, BP)神经网络的 LMS 变步长算法,该算法通过深度学习模块完成当前步长的智能计算,一定程度上降低了对参数的依赖程度。然而,BP 神经网络在训练阶段对监督数据要求高,且此类数据的获取难度较大,致使其在实际应用中的表现并不理想。

鉴于传统变步长 LMS 算法存在的参数依赖性强,实验调参工作量大等弊端,本研究在传统变步长算法框架的基础上,创新性地引入了深度强化学习前沿方法[15-18],提出了一种基于近端策略优化(proximal policy optimization, PPO)的 LMS 变步长新算法。此算法的核心在于构建马尔可夫决策过程,通过针对性训练运用强化学习网络输出精准动作 a、灵活控制步长 μ 的增减,取代了传统算法中的非线性函数模块,规避了繁琐的实验调参流程,降低了算法使用的复杂性。新算法在系统辨识、信号滤波以及截流区龙口水位提取实验中均能取得较好性能。

1 定步长 LMS 算法及常规变步长 LMS 算法

自适应滤波器的原理如图 1 所示。其中,$x(n)$ 和 $y(n)$ 分别代表在 n 时刻系统的输入信号和自适应滤波器的输出信号。$d(n)$ 是 n 时刻的期望信号,$e(n)$ 是 n 时刻的期望信号与滤波器输出信号之差,称作误差信号。自适应算法根据误差信号 $e(n)$ 修正自适应滤波器权值 $\boldsymbol{W}(n)$,使得 $y(n+1)$ 更加逼近期望信号。

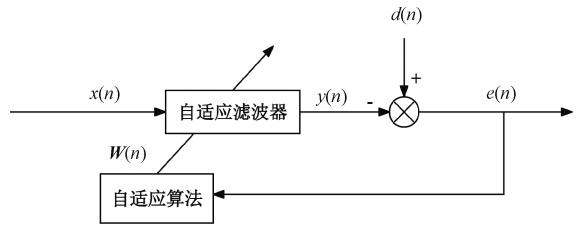


图 1　自适应滤波器结构

Fig. 1　Architecture of the adaptive filter

LMS 算法的公式可以表示为:

$$e(n) = d(n) - \boldsymbol{X}^{\mathrm{T}}(n)\boldsymbol{W}(n) \tag{1}$$

$$\boldsymbol{W}(n+1) = \boldsymbol{W}(n) + 2\mu\boldsymbol{X}(n)e(n) \tag{2}$$

式中:参数 μ 为自适应滤波器的固定步长因子,应当满足收敛条件 $0 < \mu < \lambda_{\max}$;$\lambda_{\max}$ 为输入信号的自相关矩阵的最大特征值;对于 m 阶滤波器,$\boldsymbol{X}(n)$ 是由 m 个时刻输入信号组成的向量。

变步长 LMS 算法的设计原则在于,初始阶段采用较大步长以实现快速收敛,而在收敛至稳态后则减小步长以降低稳态误差。传统变步长策略通常通过将固定的步长参数 μ 转化为可变函数来实现步长的自适应调节。具体实现方式为:

$$\begin{cases} e(n) = d(n) - \boldsymbol{X}^{\mathrm{T}}(n)\boldsymbol{W}(n) \\ \mu(n) = f(e(n)) \\ \boldsymbol{W}(n+1) = \boldsymbol{W}(n) + 2\mu(n)\boldsymbol{X}(n)e(n) \end{cases} \tag{3}$$

式中:$\mu(n)$ 为可变步长,一般与误差信号 $e(n)$ 存在函数关系 f;f 中通常包含多个参数,实际应用中一般通过大量实验,根据实验结果人为选取较优的一组参数。

2 基于 PPO 的变步长 LMS 算法

2.1 强化学习问题建模

强化学习的核心构成要素包括 5 个主要部分[19],分别为:环境 (env)、智能体($agent$)、动作(a)、状态($\boldsymbol{s}$)、奖励(r)。如图 2 所示,在 t 时刻智能体 $agent$ 基于

接收的状态信息 s_t，自动选取相应的动作 a_t。环境 *env* 在接收到动作 a_t 后，过渡到下一个状态 s_{t+1} 并向系统提出相应的奖励信号 r_t。随后，智能体 *agent* 根据环境所提供的动作奖励反馈调整自身策略，并输出下一个时刻的动作 a_{t+1}。

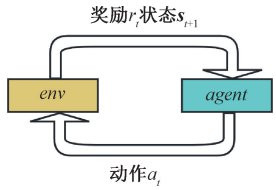


图 2 强化学习框架

Fig. 2 Architecture of reinforcement learning

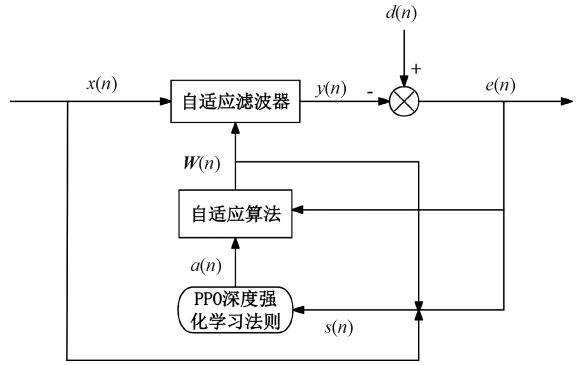


图 3 PPO_LMS 自适应滤波器

Fig. 3 PPO_LMS adaptive filter

针对 LMS 算法中步长因子 μ 的选取问题，强化学习需要构建马尔科夫决策过程(Markov decision process, MDP)[20]，该过程可以用一个四元组 (S,A,P,R) 来表述。在自适应滤波算法步长选择问题中，状态空间 S 对应滤波器权值 $\boldsymbol{W}$、误差信号 e 和步长 μ；动作空间 A 对应于步长调整指令 a；状态转移空间 P 对应于式(1)和(2)；奖励空间 R 对应于自适应滤波器在每个时刻 t 的奖励 r_t，通过误差信号等要素决定。

通过深度强化学习方法，自适应滤波器与外界环境进行交互，寻找步长调整方法的最佳策略 π_θ^*，从而获得最大的奖励。对于任一步长调整策略 π_θ，用长期奖励期望表示其目标函数。

$$J(\pi_\theta)=E_{\tau\sim\pi_\theta}[R(\tau)]=E_{\tau\sim\pi_\theta}\left[\sum_{t=0}^{\infty}\gamma^t r_t\right] \tag{4}$$

式中：$\tau=(s_0,a_0,r_0,\cdots,s_T,a_T,r_T)$ 是自适应滤波器在策略 π_θ 下产生的一条轨迹；初始状态 s_0 由环境决定；折扣因子 $\gamma\in[0,1]$ 表示奖励随时间的衰减；最大化目标函数 J 等价于求出最佳策略 π_θ^* 对应的深度网络参数 θ^*。

$$\theta^*=\arg\max_\theta E_{\tau\sim\pi_\theta}[R(\tau)] \tag{5}$$

至此，自适应滤波算法的变步长问题就被转化为了一个可以通过深度强化学习方法训练求解的 MDP 模型。

2.2 PPO_LMS 算法流程

结合式(3)所述步长调整原则与上文所建立的 MDP 模型，提出了一种基于 PPO 的 LMS 变步长算法，简称 PPO_LMS 算法，其原理如图 3 所示。

PPO_LMS 算法相较于传统 LMS 算法，融合了深度强化学习机制。该机制涉及一个经过训练的步长调整策略模块 π_θ^*。对于任意给定的初始步长 μ_0，该策略模块能够依据环境状态对步长进行动态增减。在 n 时刻，策略模块能够依据当前环境状态 $\boldsymbol{s}(n)$ 输出最优动作指令 $a(n)$，自适应算法依据该指令自动调整步长。该模块取代了传统算法中的非线性函数模块，从而去除了实验调参的步骤，极大程度降低了遴选参数的工作量，以及算法使用的复杂性。PPO_LMS 算法的数学表达式可以表述为：

$$\begin{cases} e(n)=d(n)-\boldsymbol{X}^{\mathrm{T}}(n)\boldsymbol{W}(n) \\ a(n)=\pi_\theta^*(\boldsymbol{s}(n)) \\ \mu(n)=f(a(n),\mu(n-1)) \\ \boldsymbol{W}(n+1)=\boldsymbol{W}(n)+2\mu(n)\boldsymbol{X}(n)e(n) \end{cases} \tag{6}$$

式中：f 为自适应算法中的步长调整方法，它通过动作 $a(n)$ 和上一步的步长 $\mu(n-1)$，结合步长最大值和最小值的范围，得到当前步长。当动作指令 $a=1$ 时，降低下一时刻步长；当动作指令 $a=2$ 时，维持原步长不变；当动作指令 $a=3$ 时，增加下一时刻步长。f 的具体表达为：

$$\begin{cases} \mu=\begin{cases}\mu(n-1)\times\beta_{down}, & a=1\\ \mu(n-1), & a=2\\ \mu(n-1)\times\beta_{up}, & a=3\end{cases} \\ \mu(n)=\begin{cases}\mu_{\max}, & \mu>\mu_{\max}\\ \mu_{\min}, & \mu<\mu_{\min}\\ \mu, & \text{其他}\end{cases} \end{cases} \tag{7}$$

式中：$0<\beta_{down}<1$ 是步长下降系数；$1<\beta_{up}<2$ 是步长上升系数；$\mu_{\max}$ 是最大步长，其值小于 $\lambda_{\max}$；$\mu_{\min}$ 是最小步长。由于 $\mu<\mu_{\max}<\lambda_{\max}$，该算法收敛。

2.3 网络训练流程

为了获得深度强化学习步长调整模块 π_θ^*，使用 PPO 算法[21]进行训练。PPO 是一种基于策略的强化学习方法，它在信赖域策略优化算法(trust region policy optimization, TRPO)[22]基础上引入了多种机制，提高了强化学习训练的稳定性和数据利用率。根据式(4)可以得到策略函数优化的梯度表达式。

$$\nabla_\theta J(\pi_\theta)=E_\tau\left[\sum_{t=0}^{T}\nabla_\theta\log\pi_\theta(a_t|\boldsymbol{s}_t)R(\tau)\right] \tag{8}$$

PPO 算法为了解决数据利用率低的问题，引入了中间策略网络 $\pi_{\theta'}$ 和重要性权重 $p_t(\theta)=\pi_\theta(a_t|\boldsymbol{s}_t)/\pi_{\theta'}(a_t|\boldsymbol{s}_t)$，使

用中间网络与环境进行交互,重复利用生成的数据训练策略网络 π_θ。同时,引入了 Critic 价值网络 Q,并通过价值网络计算优势函数,实现了单步更新,进一步提高了效率。式(8)可以进一步改写为:

$$E_{\pi_{\theta'}}[p_t(\theta)\nabla_\theta \log \pi_\theta(a_t|\boldsymbol{s}_t)A_t^{GAE}(\boldsymbol{s}_t,a_t)] \tag{9}$$

为使得中间策略与策略网络之间的差异不超过优化范围,TRPO 算法利用相对熵进行距离约束,缓解了优化过程中可能存在的梯度消失问题。PPO 算法则在 TRPO 的基础上进一步优化,采用了梯度裁剪策略限制了策略的更新幅度,降低了算法实现的复杂度,其目标函数可以表示为:

$$J^{PPO}(\pi_\theta)=\sum_{\boldsymbol{s}_t,a_t}\min[p_t(\theta)A_t^{GAE}(\boldsymbol{s}_t,a_t),\ clip(p_t(\theta),1-\epsilon,1+\epsilon)A_t^{GAE}(\boldsymbol{s}_t,a_t)] \tag{10}$$

当优势函数 $A_t^{GAE}(\boldsymbol{s}_t,a_t)>0$ 时,一方面为了最大化该目标函数,需要增大在 $\boldsymbol{s}_t$ 状态下执行 a_t 动作的概率,重要性权重 $p_t(\theta)$ 随之增大。另一方面,对于式(10)有:

$$J^{PPO}(\pi_\theta)\leqslant\sum_{\boldsymbol{s}_t,a_t}clip(p_t(\theta),1-\epsilon,1+\epsilon)A_t^{GAE}(\boldsymbol{s}_t,a_t)\leqslant\sum_{\boldsymbol{s}_t,a_t}(1+\epsilon)A_t^{GAE}(\boldsymbol{s}_t,a_t) \tag{11}$$

目标函数存在最大值上限,当重要性权重 $p_t(\theta)$ 增大至 $1+\epsilon$ 时达到峰值。此时,$p_t(\theta)$ 继续增大不会使目标函数继续增加。类似的,当优势函数 $A_t^{GAE}(\boldsymbol{s}_t,a_t)<0$ 时,需要降低在 $\boldsymbol{s}_t$ 状态下执行 a_t 动作的概率,重要性权重 $p_t(\theta)$ 随之降低。当重要性权重 $p_t(\theta)$ 降低至 $1-\epsilon$ 时达到峰值。此时,$p_t(\theta)$ 继续降低不会使目标函数继续增加。这样,重要性权重 $p_t(\theta)$ 就被限制在了 $1-\epsilon$ 与 $1+\epsilon$ 之间,从而保证了中间策略与策略网络的差异不超过优化范围,本研究中 ϵ 取 0.02。

PPO_LMS 自适应滤波器训练的流程如图 4 所示。其核心在于将中间策略网络 $\pi_{\theta'}$ 采样得到的结果保存至缓冲区 E,进行多个 batch 的梯度计算,用以优化策略 π_θ,从而提高数据质量与数据利用率,减少最终的控制误差。训练流程图相对应的算法伪代码如算法 1 所示。

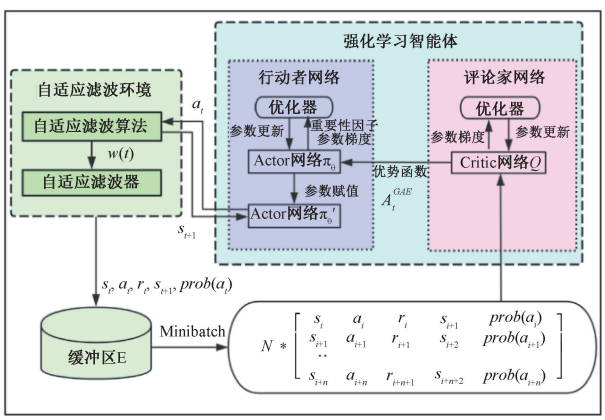


图 4　PPO_LMS 自适应滤波器训练流程

Fig. 4　Training process of PPO_LMS adaptive filter

算法 1　基于 PPO_LMS 的自适应滤波算法

1. 初始化:Actor 网络 π_θ,Critic 网络 Q,仿真环境;将 π_θ 赋值给中间策略网络 $\pi_{\theta'}$;
2. **For** *episode* = 1,…,M:
3. 　输入自适应滤波器初始状态 $\boldsymbol{W}(0)$,μ_0;初始化经验池 E;
4. 　**For** t = 1,…,T:
5. 　　确定强化学习当前时刻状态 $\boldsymbol{s}_t$
6. 　　将当前状态 $\boldsymbol{s}_t$ 输入中间策略网络 $\pi_{\theta'}$,并采样得到动作 a_t 以及动作概率 $prob_{\theta'}(a_t)$;
7. 　　将动作 a_t 作用仿真环境自适应滤波器输入,获得下一时刻状态 $\boldsymbol{s}_{t+1}$,通过奖励函数得到 r_t;
8. 　　将本次交互结果 $<\boldsymbol{s}_t,a_t,r_t,\boldsymbol{s}_{t+1},prob(a_t)>$ 存入缓冲区 E;
9. 　　**If** 缓冲区>经验池阈值 $E_{\max}$
10. 　　对经验池中所有样本状态 $\boldsymbol{s}_t,\boldsymbol{s}_{t+1}$ 用 Critic 网络估计状态价值并计算优势值 A_t^{GAE} 和目标状态价值 v_{target}^t 用于计算更新梯度,计算 a_t 在当前策略下的概率 $prob_\theta(a_t)$;
11. 　　使用 Adam 优化方法,根据采样数据对目标函数式(9)进行反向梯度传播优化,用以更新策略网络 π_θ;
12. 　　使用本批次的 Critic 网络评估 $\boldsymbol{s}_t$ 价值与 v_{target}^t 取均方误差。使用 Adam 优化方法更新 Critic 网络;
13. 　　**End if** 批量数据策略和价值网络更新完成;
14. 　　将 π_θ 赋值给中间策略网络 $\pi_{\theta'}$;
15. 　**End for**
16. **End for**
17. **Return** 最优策略 π_θ^*

2.4　网络结构与奖励函数

在 PPO_LMS 算法的训练过程中需要构建动作空间与状态空间并选择合适的网络结构以及奖励函数从而得到策略函数 π_θ 和价值函数 Q。

网络的状态空间 $\boldsymbol{s}$ 由输出信号 y,误差信号 e,当前步长 μ,输入信号 $\boldsymbol{X}$,以及滤波器参数 $\boldsymbol{W}$ 组成,具体可表示为:

$$\boldsymbol{s}_n=[y_n,e_n,\mu_{n-1},\boldsymbol{X}_n,\boldsymbol{W}_n] \tag{12}$$

动作空间 A 是一个离散型的动作空间,由 3 种常见的动作指令构成:步长下降,步长不变,步长上升。分别对应了 a = 1,2,3 的情况。

如图 5 所示,PPO 的价值网络和策略网络使用了欠完备自编码器作为网络结构的一部分。欠完备自编码器通过瓶颈状的神经网络结构提取重要信息并忽略掉次要

信息,可有效提高强化学习的泛化能力。对于策略网络 Actor,输入层与隐藏层、隐藏层与隐藏层之间均使用全连接网络,激活函数为线性整流单元(rectified linear unit, Relu)函数;隐藏层和输出层之间则选择 Softmax 函数进行连接得到每个动作执行的概率,最后通过采样得到 a。对于价值网络 Critic,其输出层只含有一个神经元和隐藏层通过全连接网络连接,输出当前状态的价值,其余部分与策略网络结构一致。

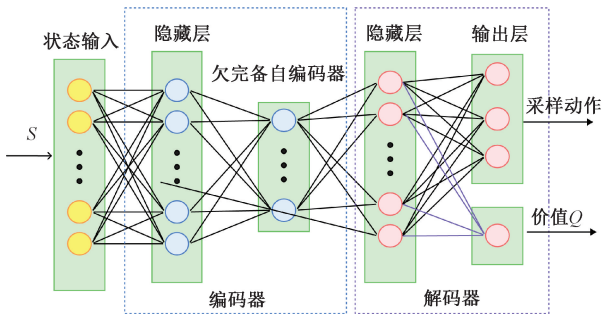


图 5 价值网络和策略网络结构

Fig. 5 Value network and policy network architecture

奖励函数是强化学习的重要组成部分,目的是激励 LMS 算法在尽可能少的迭代次数内使均方误差尽可能小。考虑到自适应滤波器任务特性,将奖励 r 分为误差奖励 r_e 和步长奖励 r_μ 两部分。

$$r = r_e + r_\mu \tag{13}$$

误差奖励 r_e 是基于误差 e 的状态奖励,其设计原则是围绕滤波效果进行优化。具体来说,当滤波后的信号与原始真实信号之间的误差越小,给予的奖励值越高。同时,为提升算法性能,引入了动态奖励机制,根据算法运行阶段动态调整奖励。

误差奖励形式上表现为一个随误差增大单调递减的分段函数:在起始阶段,因均方误差变化幅度大,设置较大的奖励系数,此时均方误差越小获得奖励值越大,能有效促进算法前期快速收敛;进入过渡阶段,均方误差变化幅度变小,相应调小奖励系数,引导算法平稳过渡,防止 PPO 算法陷入局部最优;在收敛阶段,算法达到目标状态,给与固定奖励值,达到此状态时的迭代次数越少则后续获得的奖励值越高,以此提高算法的收敛速度。通过这种奖励机制能引导算法朝着减小误差的方向不断进化,具体形式为:

$$r_e = \begin{cases} -k_1 \times e^2, & 0.5 < e^2 \\ -k_2 \times e^2, & 0.01 < e^2 \leqslant 0.5 \\ k_3, & e^2 \leqslant 0.01 \end{cases} \tag{14}$$

其中,k_1,k_2,k_3 均为正常数,分别是起始阶段奖励系数、过渡阶段奖励系数、收敛阶段奖励系数。

步长奖励 r_μ 是基于步长 μ 的动作奖励,目的是满足在误差大的时候增加步长加快收敛速度,误差小的时候降低步长从而减少稳态误差的基本步长调整原则。在设置步长奖励时引入了负奖励的机制:当误差较大时,若智能体输出动作指令 a 降低步长 μ,或是误差较小时,若智能体输出动作指令 a 增加步长 μ 时,奖励函数将给出一个负奖励值,对这样的不良指令进行有效惩罚,保证算法性能逐步优化。具体形式为:

$$r_\mu = \begin{cases} -k_{up}, & \mu_n > \mu_{n-1}, \quad e^2 < 0.01 \\ k_{up}, & \mu_n > \mu_{n-1}, \quad e^2 > 0.5 \\ k_{down}, & \mu_n < \mu_{n-1}, \quad e^2 < 0.01 \\ -k_{down}, & \mu_n < \mu_{n-1}, \quad e^2 > 0.5 \end{cases} \tag{15}$$

式中:k_{up}、k_{down} 均为正常数,分别表示步长增加与降低时的奖励系数。

3 仿真实验及分析

本研究对提出的 PPO_LMS 算法在系统辨识场景下进行训练并在多种场景下进行了性能评估,与经典的定步长 LMS 算法和变步长 LMS 算法进行各类实验对比分析。所有实验均在 CPU 为 Intel i5-13600KF,显卡为 RTX4070 的计算机硬件配置上进行。所有算法测试程序均在 MATLAB R2023a 上进行。

3.1 网络与算法参数设置

为了验证新算法的泛化能力,设计了如下实验。首先,在系统辨识场景下采集数据、搭建神经网络并且用采集到的数据对网络进行训练。在此基础上,将训练完成的网络直接迁移至其他场景使用,并与传统的变步长 LMS 算法进行性能比较。

系统辨识实验仿真环境将在 Simulink 中搭建。自适应滤波器为六阶线性滤波器,未知系统的滤波器权值为 $\boldsymbol{W}_0$,在第 500 个采样点时刻变为 $\boldsymbol{W}_1$,算法初始步长为 μ_0,输入信号 $\boldsymbol{X}(n)$ 是均值为 0、方差为 1 的高斯白噪声,干扰噪声 $\boldsymbol{V}(n)$ 是均值为 0,方差为 0.000 1 的高斯白噪声;输入信号和噪声信号互不相干。采样频率为 100 Hz,每轮实验采样点数为 1 000。自适应滤波器的参数设置如表 1 所示。

表 1 PPO_LMS 自适应滤波器参数设置

Table 1 PPO_LMS adaptive filter parameters

参数名称	参数设置
β_{down}	0.9
β_{up}	1.1
$\mu_{\min}$	0.001
$\mu_{\max}$	0.08

训练过程采取了 10 组不同的滤波器权值 $\boldsymbol{W}_0$、$\boldsymbol{W}_1$,和初始步长 μ_0,提高算法在不同情况下的泛化能力。在 10 万个时间步长上运行共 100 次自适应滤波任务从而获得 PPO 的网络参数。PPO 算法的参数如表 2 所示。

表 2　深度强化学习算法参数设置

Table 2　Deep reinforcement learning parameters

参数名称	参数设置
折扣因子 γ	0.997
学习率 Lr	1×10^{-4}
裁剪率 ϵ	0.02
回合步长	1 000
经验池大小	600
Epoch 数	500
批训练大小	100
Critic 网络参数	[15;128;8;128;3]
Actor 网络参数	[15;128;8;128;1]
k_1	100
k_2	10
k_3	30
k_{up}	100
k_{down}	50

图 6 为 PPO 算法的学习曲线,策略网络在整个时间步长上的奖励值逐步提高。由图 6 可知,该算法在前 20 次训练中性能提升速度较快,并在 30 次训练左右收敛。

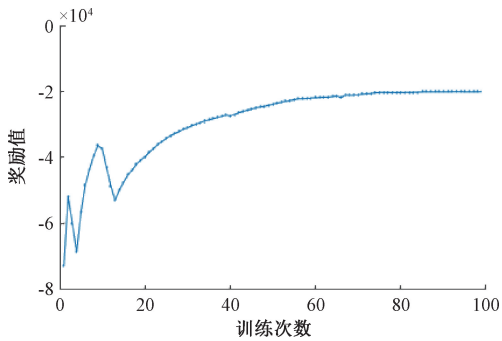


图 6　PPO_LMS 算法学习曲线

Fig. 6　Learning curve of PPO_LMS

3.2　变步长算法在系统辨识中的性能对比

为了检验本研究算法的性能,将其与固定步长 LMS 算法和较新的变步长 LMS 算法进行比较。

文献[11]提出的基于 Sigmoid 函数的变步长函数为:

$$\mu(n)=\beta\frac{2}{1+e^{-\alpha(|e(n)^3|)}}-1 \tag{16}$$

文献[12]提出的基于反双曲正切函数的变步长函数为:

$$\mu(n)=\beta\times\operatorname{arctanh}(\alpha\times[|e(n)|]^{\lambda}) \tag{17}$$

实验仿真环境设置同 3.1 节选择一致:初始权值 $\boldsymbol{W}_0=[0.8\quad 0.6]^{\mathrm{T}}$,第 500 个采样点时权值 $\boldsymbol{W}_1=[1.4\quad 1.2]^{\mathrm{T}}$。该仿真环境下,对每种算法重复 100 次实验。各算法的参数选取如表 3 所示。

表 3　各算法参数表

Table 3　Parameters of each algorithm

算法	α	β	γ	μ_0
文献[11]	0.02	4	1	
文献[12]	0.5	0.02		
固定步长				0.001
本文算法				0.001

由图 7 可知在上述实验条件下,为了取得较小的稳态误差,固定步长方法所需迭代次数较多。而文献[11-12]所提出的变步长方法仍需要 30 次左右迭代才能收敛。本研究提出的 PPO_LMS 算法能在 20 次迭代左右收敛,并且相较于其他算法稳态误差更小。

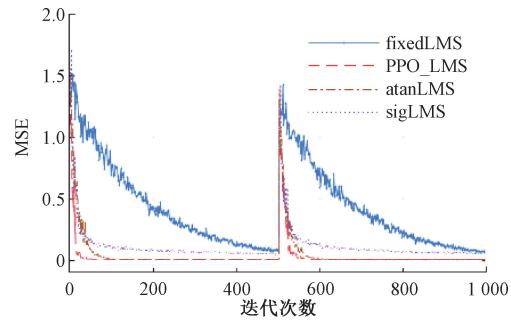


图 7　各算法性能比较

Fig. 7　Performance of the algorithms

此外,如前所示,传统的变步长方法涉及较多参数且对结果影响显著,需要进行多次实验才能人为选取较好的参数,降低了算法的自动化程度,而所提算法只需在最大步长与最小步长之间任取初始步长 μ_0,极大简化了参数选取的流程,最大程度降低了人为主观因素的干扰。

为了验证新算法对初始参数的低依赖特性,设计了算法性能的验证实验,通过设置不同数量级的初始参数,将传统算法与新算法进行了稳定性与适应性比较。

图 8 为文献[11]提出的传统变步长 LMS 算法在不同初始参数下的运行结果。由图可知,当初始参数 α 由 0.5 变为 0.05 以及初始参数 β 由 0.02 变为 0.2 时,由于初始参数的改变限制了步长变化的性能,致使收敛速度

放缓,稳态误差增大,算法的综合性能呈现不同程度的下降。

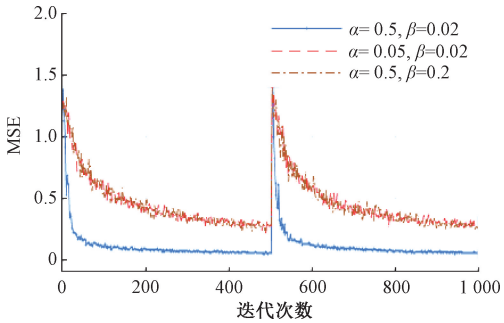


图 8 不同参数下传统变步长 LMS 算法性能比较

Fig. 8 Performance of the traditional LMS algorithm with different parameters

反观图 9 为应用本研究算法在不同初始步长 μ_0 下的仿真结果。当μ_0 分别=0.001、0.005、0.02、0.05 时,算法均表现出了良好的性能。当初始步长为 0.001 时算法在约 30 次左右迭代时收敛,当初始步长为 0.05 时算法约在 20 次左右迭代收敛,算法的收敛速度相近,远远优于传统算法。

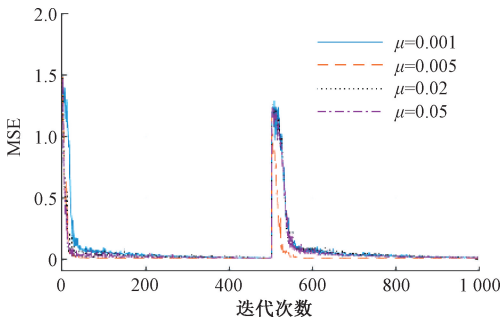


图 9 不同初始步长下本研究算法性能比较

Fig. 9 Performance of the algorithm with different steps

图 10 为新算法在前 100 次迭代中的步长动态变化曲线,分析图中曲线可知,即使提供给新算法一个看似欠佳的初始值,通过内置的深度强化学习模块,算法仍能快速自我调整,迅速趋近合理的步长值,从而保证整体性能。上述实验表明,新算法对初始参数的依赖程度低,可为实际的应用场景提供更高效的解决方案。

3.3 信号去噪

为了进一步检验所提方法的有效性和泛化能力,设计了信号去噪实验。将由 3.1 节训练得到的深度强化学习网络直接运用到信号去噪中。其余参数的设置包括策略网络均与 3.1 节保持一致。

为了探究新算法在复杂噪声环境下的性能,构建了一种包含高斯白噪声与脉冲噪声的混合噪声信号。其中,高斯白噪声均值设定是 0、方差为 1,脉冲噪声的

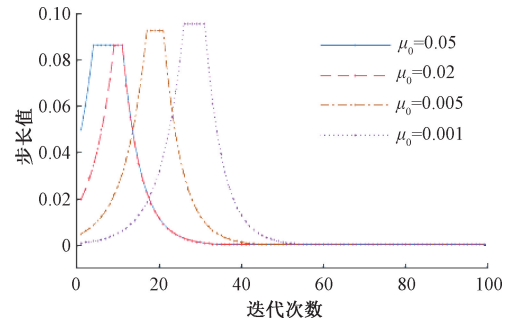


图 10 不同初始值下本研究算法步长变化

Fig. 10 Step changes of the algorithm with different initial values

脉冲幅度于-2~2 区间内随机取值,且脉冲出现的概率为 0.1。图 11 呈现了噪声信号与带噪信号经新算法滤波后的结果,分析滤波后的信号图可知,新算法能高效地将混合噪声从正弦信号中滤除。上述实验将在系统辨识场景中训练得到的深度强化学习模块直接运用到信号去噪中,并取得良好效果,证明了新算法具备一定的泛化能力,能够突破场景局限,灵活应对不同任务。

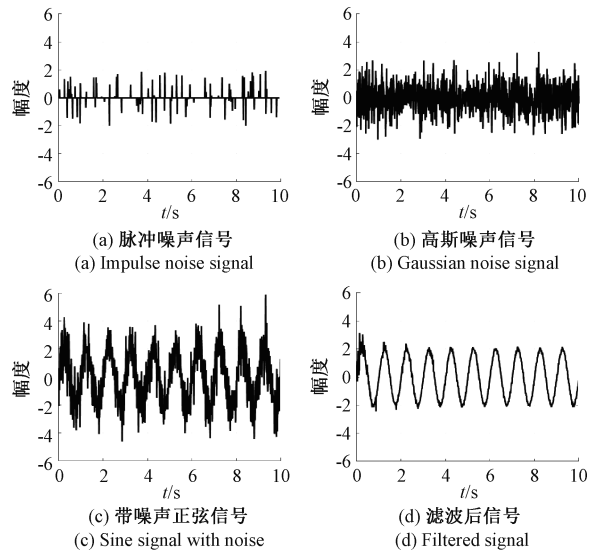


图 11 信号去噪效果示意

Fig. 11 The results of denoising signal

图 12 对新算法与传统变步长 LMS 算法的降噪性能进行直观对比,由收敛曲线可知本文算法对比文献[11]算法收敛性能有明显提升,对比文献[12]算法虽然稳态误差水平相近但是新算法收敛速度明显加快。

3.4 新算法在截流区龙口水位信号滤波中的应用

受波浪等噪声、以及龙口截流等多种因素影响,截流区龙口水位变化幅度大,上下起伏剧烈,须对监测水位信息进行快速有效滤波,才能提取出真实水位信息。为进

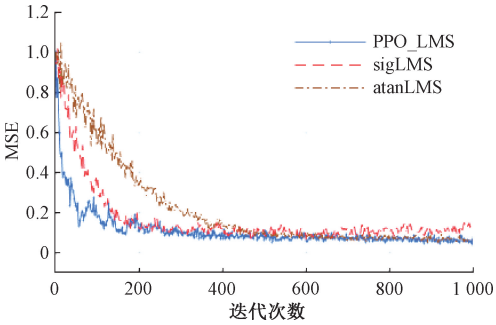


图 12　不同算法性能对比

Fig. 12　Comparison of denoising signal results

一步验证所提出的 PPO_LMS 算法性能,将该算法应用于截流水域龙口水位信号的滤波中。

如图 13 所示为实际监测得到的一段龙口水域水位信号以及经过滤波后得到的水位信号。监测信号的采样频率设定为 20 Hz,采样点数达到 10 000 点。新算法的滤波器长度设置为 200,初始步长设置为 0.002。由图 13 可知,相比滤波前的原始水位信号,经新算法滤波后的信号平滑,大量噪声被很好地滤除,使得水位信号得以准确提取。

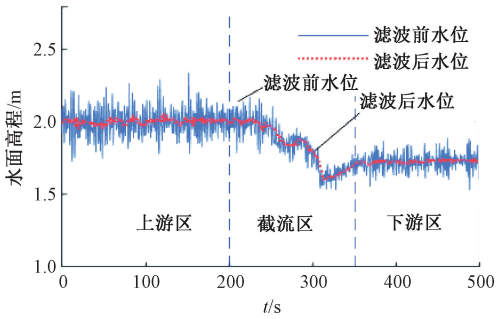


图 13　龙口水位信号

Fig. 13　Closure water level signal

此外,为进一步验证新算法在实际工程中的优越性,特选取了文献[12]变步长算法并将其同步应用于龙口水位信号的滤波处理之中。图 14 为经过滤波处理后的截流区水位效果对比图,分析图中数据可知:在 280~320 s 的截流区,传统算法在对水位信号进行滤波时,出现了较为明显的时延问题,导致其无法跟上水位信号的快速变化。而新算法凭借强化学习网络提供的快速步长调整能力仍能实现水位的快速跟踪和精准提取,满足了实际需求。

为了更好地比较新算法与传统算法的收敛速度,本文使用归一化偏差 F (normalized misalignment)[23] 作为衡量标准。

$$F = \frac{E\{ \| \boldsymbol{W}_0 - \boldsymbol{W}(k) \|^2 \}}{\| \boldsymbol{W}_0 \|^2} \tag{18}$$

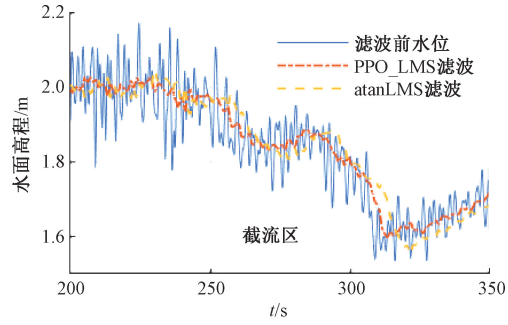


图 14　基于不同算法滤波的截流区水位对比

Fig. 14　Comparison of water levels in diversion areas based on different algorithmic filtering

式中:$\boldsymbol{W}_0$ 为最终时刻滤波器的权值向量;$\boldsymbol{W}(k)$ 代表 k 时刻滤波器的权值向量。由图 15 可知,相较于其他的变步长 LMS 自适应滤波算法,新算法的滤波器权值能更快的收敛至最终时刻权值,收敛速度大大提升。

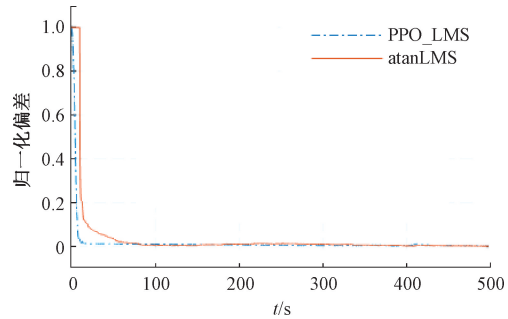


图 15　多种算法的归一化偏差曲线

Fig. 15　Normalized misalignment curves for various algorithms

3.5　算法复杂性分析

为了衡量算法复杂性对算法运行效率的影响,将本研究算法与文献[11-12]中阐述的变步长 LMS 算法进行对比研究。具体而言,在 3.2 节的系统辨识场景与 3.3 节的信号去噪场景下,分别针对各算法进行 100 次实验,并详细统计每次实验所耗费的时间,实验结果如表 4 所示。

表 4　各算法用时表

Table 4　Time consumption of each algorithm　(s)

算法	系统辨识用时	信号去噪用时
文献[11]	0.043	0.037
文献[12]	0.050	0.040
本文算法	0.072	0.066

由表 4 可知,尽管构建的 PPO_LMS 算法在传统的 LMS 算法的基础上创新性地引入了深度强化学习模

块,且在步长调整前需要借助 Actor 网络来计算动作 a,这一过程不可避免涉及额外的矩阵运算环节,致使该算法的耗时较传统算法有一定程度的增长。然而,与文献[11]所提出的算法比较,在系统辨识与信号去噪这两个场景下每 100 次实验的耗时仅增加 0.029 s,增幅极为有限。相较于文献[12]中的算法,在系统辨识场景下每 100 次实验增加 0.022 s,而在信号去噪场景下每 100 次实验仅增加 0.026 s。由此可以看出,其影响微乎其微。

此外,值得一提的是,新算法相较于传统算法具备一项突出优势,即能够有效规避繁琐的手动调参流程。在实际工程应用当中,手动调参往往需要耗费大量的时间与人力成本。从整体的算法应用来看,PPO_LMS 算法通过避免手动调参这一环节,实际上在很大程度上降低了算法使用的综合复杂度,为其在实际工程领域的广泛推广与高效应用奠定了坚实基础。

4 结 论

针对定步长以及传统变步长 LMS 算法存在的问题,将深度强化学习算法融入变步长 LMS 自适应滤波算法中,创新性地提出了一种基于深度强化学习的新的变步长 LMS 算法。构建了一个融合 PPO 算法和自适应滤波的算法模型,该模型通过训练并使用强化学习智能体,实现了自适应滤波算法步长因子自动调整,突破了步长参数调整受初始参数影响的瓶颈,克服了常规 LMS 算法的不足。此外,将深度强化学习方法与自适应滤波方法特性相结合,设计了基于误差与步长的奖励函数并使用了欠完备编码器作为网络结构,提高了强化学习智能体训练的收敛速度。在此基础上设计了多种场景的泛化性能验证实验,将在系统辨识场景下训练的智能体应用到信号去噪中,取得了良好的性能。仿真实验结果表明,该算法可根据环境状态自动调整步长因子,且在多种初始步长下都取得了良好的效果。在系统辨识与信号去噪应用中,该算法较传统算法有更快的收敛速度和更小的稳态误差。尽管本研究算法的耗时较传统算法略有增长,但并不会影响算法的整体性能,且新算法能够规避调参流程,降低使用的综合复杂度。最后,将新算法应用到了截流区龙口水位数据的滤波中,得到了准确的水位信号,满足了实际工程需求。

参考文献

[1] 刘 伟,郭尚尚,商世广. 用于 CZT 探测器前端的数字自校准 SAR-ADC 设计[J]. 电子测量与仪器学报,2022,36(9):167-173.
LIU W, GUO SH SH, SHANG SH G. Design of SAR-ADC with digital self-calibration for CZT detectors front-ends [J]. Journal of Electronic Measurement and Instrumentation, 2022,36(9):167-173.

[2] 张展,冷全超,王维, 等. 基于反余切函数的变步长 LMS 谐波检测算法[J]. 传感器与微系统,2022,41(9):144-147,155.
ZHANG ZH, LENG Q CH, WANG W, et al. Variable step size LMS harmonic detection algorithm based on anti-cotangent function [J]. Transducer and Microsystem Technologies, 2022,41(9):144-147,155.

[3] 宁小玲,张林森,刘志坤. 基于自适应混合能量参数的变步长 LMS 水声信道均衡算法[J]. 系统工程与电子技术, 2015,37(9):2141-2147.
NING X L, ZHANG L S, LIU ZH K. Variable step size LMS equalization algorithm based on adaptive mixed power parameter in underwater acoustic channels [J]. Systems Engineering and Electronics, 2015, 37 (9): 2141-2147.

[4] 庞宇,陈亚军,汪立宇. 一种改进的变步长最小均方算法滤除心电信号运动伪迹的研究[J]. 科学技术与工程, 2020,20(8):3083-3087.
PANG Y,CHEN Y J,WANG L Y. An improved variable step size LMS algorithm for filtering motion artifacts of ECG signals[J]. Science Technology and Engineering, 2020,20(8):3083-3087.

[5] 王亚军. 基于变步长 LMS 算法的谐波电流检测与治理研究[D]. 大庆: 东北石油大学,2021.
WANG Y J. Research on harmonic current detection and control based on variable step size LMS algorithm[D]. Daqing: Northeast Petroleum University, 2021.

[6] WIDROW B, HOFF M. Adaptive switching circuits[C]. IRE WESCON Convention Record, 1960:96-104.

[7] 李常虎,伍松,魏晟弘, 等. 箕舌线变步长 LMS 算法的分析与改进[J]. 广西科技大学学报, 2022,33(4):57-62.
LI CH H, WU S, WEI SH H, et al. Analysis and improvement of variable step size LMS algorithm based on tongue-like curve[J]. Journal of Guangxi University of Science and Technology, 2022, 33(4): 57-62.

[8] 谢炎江,原霞,刘锋, 等. 基于步长上界原理的变步长 FXLMS 算法研究[J]. 电子测量技术,2024,47(20):101-108.
XIE Y J, YUAN X, LIU F, et al. Research on variable

step size FXLMS algorithm based on step size upper limit ratio method[J]. Electronic Measurement Technology, 2024, 47(20): 101-108.

[9] 张鹏,李武,张超, 等. 超宽带信号带内平坦度的自适应补偿方法[J]. 国外电子测量技术, 2023, 42(1): 107-111.
ZHANG P, LI W, ZHANG CH, et al. Adaptive compensation method for in-band flatness of ultra-wideband signal[J]. Foreign Electronic Measurement Technology, 2023, 42(1): 107-111.

[10] 覃景繁,欧阳景正. 一种新的变步长 LMS 自适应滤波算法[J]. 数据采集与处理, 1997(3):171-174.
QIN J F, OUYANG J ZH. A novel variable step size LMS adaptive filter algorithm[J]. Journal of Data Acquisition and Processing, 1997(3): 171-174.

[11] 张红梅,韩万刚. 一种新的变步长 LMS 自适应滤波算法研究及其应用[J]. 仪器仪表学报,2015,36(8): 1822-1830.
ZHANG H M, HAN W G. A new variable step LMS algorithm and its application[J]. Chinese Journal of Scientific Instrument, 2015,36(8):1822-1830.

[12] 火元莲,安娅琦,巩琪, 等. 基于反双曲正切函数的变步长 LMS 算法[J]. 北京理工大学学报, 2022, 42(10):1051-1058.
HUO Y L, AN Y Q, GONG Q, et al. Variable step size LMS algorithm based on inverse hyperbolic tangent function[J]. Transactions of Beijing Institute of Technology, 2022,42(10):1051-1058.

[13] 茹国宝,黄燕,郭英杰, 等. 基于对数函数的新变步长 LMS 算法[J]. 武汉大学学报(理学版),2015,61(3): 295-298.
RU G B, HUANG Y, GUO Y J, et al. New variable step size LMS algorithm based on logarithmic function[J]. Wuhan University Journal of Natural Science, 2015, 61(3):295-298.

[14] 解本巨,王宁. 基于改进 BP-LMS 自适应滤波器算法的仿真研究[J]. 计算机与数字工程,2022,50(3): 481-485,585.
XIE B J, WANG N. Simulation research based on improved BP-LMS adaptive filter algorithm[J]. Computer and Digital Engineering, 2022, 50(3): 481-485,585.

[15] BAHRINI A, KHAMOSHIFAR M, ABBASIMEHR H, et al. ChatGPT: Applications, opportunities, and threats[C]. Systems and Information Engineering Design Symposium, 2023:274-279.

[16] LIU Y H, WANG H L, WU T C, et al. Attitude control for hypersonic reentry vehicles: An efficient deep reinforcement learning method[J]. Applied Soft Computing, 2022,123:108865.

[17] 杨傲雷,陈燕玲,徐昱琳. 基于强化学习的机器人手臂仿人运动规划方法[J]. 仪器仪表学报, 2021, 42(12): 136-145.
YANG AO L, CHEN Y L, XU Y L. Humanoid motion planning of robotic arm based on reinforcement learning[J]. Chinese Journal of Scientific Instrument, 2021, 42(12):136-145.

[18] 周志勇,莫非,赵凯, 等. 基于 PPO 的自适应 PID 控制算法研究[J]. 系统仿真学报,2024,36(6):1425-1432.
ZHOU ZH Y, MO F, ZHAO K, et al. Adaptive PID control algorithm based on PPO[J]. Journal of System Simulation, 2024,36(6):1425-1432.

[19] 李明阳,许可儿,宋志强, 等. 多智能体强化学习算法研究综述[J]. 计算机科学与探索, 2024, 18(8): 1979-1997.
LI M Y, XU K ER, SONG ZH Q, et al. Review of research on multi-agent reinforcement learning algorithms[J]. Journal of Frontiers of Computer Science and Technology, 2024,18(8):1979-1997.

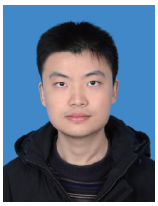
[20] 徐昊. 基于马尔可夫和 LSTM 神经网络的水质参数预测研究[D]. 长春: 长春工业大学, 2024.
XU H. Water quality parameters prediction based on Markov and LSTM neural networks[D]. Changchun: ChangChun University of Technology, 2024.

[21] GU Y, CHENG Y H, CHEN C L P, et al. Proximal policy optimization with policy feedback[J]. IEEE Transactions on Systems, Man, and Cybernetics Systems, 2022, 52(7): 4600-4610.

[22] SCHULMAN J, LEVINE S, ABBEEL P, et al. Trust region policy optimization[C]. International Conference on Machine Learning, PMLR, 2015:1889-1897.

[23] WANG W H, ZHANG H M. A new and effective nonparametric variable step-size normalized least-mean-square algorithm and its performance analysis[J]. Signal Processing, 2023,210:109060.

作者简介



徐君阳,2023 年于中南大学获得学士学位,现为武汉大学硕士研究生,主要研究方向为深度强化学习及其应用。

E-mail:2690646708@qq.com

Xu Junyang received his B. Sc. degree from Central South University in 2023. He is currently a master student at Wuhan University. His main research interests include deep reinforcement learning and its application.

张红梅(通信作者),1992 年和 1995 年于武汉水利电力大学分别获得学士和硕士学位,2003 年于武汉大学获得博士学位,现为武汉大学教授,主要研究方向为信号检测及处理技术、水下组合导航等。

E-mail:hmzhang@whu.edu.cn

Zhang Hongmei (Corresponding author) received her B. Sc. degree and M. Sc. degree both from Wuhan University of Hydraulic and Electric Engineering in 1992 and 1995, respectively, and received her Ph. D. degree from Wuhan University in 2003. She is currently a professor at Wuhan University. Her main research interests include measurement technology and signal processing, underwater integrated navigation technology, etc.

张坤,2020 年于山东科技大学获得学士学位,现为长江水利委员会水文局长江口水文水资源勘测局助理工程师,主要研究方向为河道勘测、水文测量及数据处理等。

E-mail:122741580@qq.com

Zhang Kun received his B. Sc. degree from Shandong University of Science and Technology in 2020. He is currently an associate engineer at the Survey Bureau of Hydrology and Water Resources of Changjiang Estuary. His main research interests include hydrographic surveying, geodesic survey in hydrology and their data processing.