多智能体强化学习驱动的主动声呐发射参数联合优化

生雪莉[1,2,3,4], 穆梦飞[1,2,3,4], 毕耀[1,2,3], 高远[1,2,3,4], 石冰玉[1,2,3]

(1. 哈尔滨工程大学 水声技术全国重点实验室, 黑龙江 哈尔滨 150001; 2. 极地海洋声学与技术应用教育部重点实验室(哈尔滨工程大学) 教育部, 黑龙江 哈尔滨 150001; 3. 哈尔滨工程大学 水声工程学院, 黑龙江 哈尔滨 150001; 4. 哈尔滨工程大学 三亚南海创新发展基地, 海南 三亚 572024)

摘 要:针对传统固定发射策略的主动声呐在水声信道中面临环境适配性不足,导致探测稳定性差的问题,本文提出一种基于多智能体强化学习的主动声呐发射波形与声源级的联合优化方法。采用多智能体协作学习方法,将发射波形优化与声源级优化解耦为多个智能体任务。引入奖励塑形方法,抑制多峰信道频谱引起的奖励信号噪声,提升智能体寻优能力,并避免子脉冲频点冲突。此外,使用双深度 Q 网络(double deep q-network),降低智能体 Q 值估计偏差并提升决策稳定性。在基于南海实测声速梯度重构的典型深海信道场景下进行了数值验证,结果表明:经所提算法优化后的信道适配度与回波信噪比调控准确性均优于对比算法,为构建具备环境自适应能力的智能主动声呐系统提供了一种可行的技术途径。

关键词:主动声呐; 水下探测; 信道适配; 发射参数联合优化; 多智能体; 强化学习; 奖励塑形; 双深度 Q 网络

DOI:10. 11990/jheu. 202507001

网络出版地址:https://link. cnki. net/urlid/23. 1390. U. 20250707. 1336. 002

中图分类号:TB566 **文献标志码**:A **文章编号**:1006-7043(2025)08-1557-09

Joint optimization of transmit parameters of active sonar driven by multiagent reinforcement learning

SHENG Xueli[1,2,3,4], MU Mengfei[1,2,3,4], BI Yao[1,2,3], GAO Yuan[1,2,3,4], SHI Bingyu[1,2,3]

(1. National Key Laboratory of Underwater Acoustic Technology, Harbin Engineering University, Harbin 150001, China; 2. Key Laboratory for Polar Acoustics and Application of Ministry of Education (Harbin Engineering University), Ministry of Education, Harbin 150001, China; 3. College of Underwater Acoustic Engineering, Harbin Engineering University, Harbin 150001, China; 4. Sanya Nanhai Innovation and Development Base of Harbin Engineering University, Sanya, 572024, China)

Abstract: Inadequate environmental adaptability of traditional fixed transmission strategies in active sonar systems leads to poor detection stability in underwater acoustic channels. To address this issue, this paper proposes a joint optimization method for active sonar transmission waveform and source level based on multiagent reinforcement learning. First, a multiagent collaborative learning approach was adopted to decouple waveform optimization and source level optimization into multiple agent tasks. Then, a reward-shaping method was introduced to suppress reward signal noise induced by multipeak channel spectra, enhancing the optimization capability of the agents while avoiding subpulse frequency conflicts. Furthermore, a double deep Q-network was employed to reduce Q-value estimation bias and improve decision stability. Finally, numerical validation was conducted in a typical deep-sea channel scenario reconstructed using measured sound speed gradients from the South China Sea. The results demonstrate that the proposed algorithm outperforms baseline methods in terms of both channel adaptability and echo signal-to-noise ratio control accuracy, providing a viable technical approach for constructing intelligent active sonar systems with environmental self-adaptation capabilities.

Keywords: active sonar; underwater detection; channel adaptation; joint optimization of transmit parameters; multi-agent; reinforcement learning; reward shaping; double-deep Q network

海洋的资源勘探、环境监测与国防安全等领域对高稳健性水下探测技术提出了迫切需求[1-7]。传统主动声呐的固定发射策略在水声信道中引发探测性能波动与能量效率低下问题,亟需发展具备环境感知与自主决策能力的智能化声呐系统实现动态参数优化。Haykin[8]源于蝙蝠回声探测过程中的认知

收稿日期:2025-07-01. **网络出版日期**:2025-07-07.
基金项目:国家重点研发计划(2022YFC2807804).
作者简介:生雪莉, 女, 教授,博士生导师;
穆梦飞,男,博士研究生.
通信作者:穆梦飞,E-mail:18846130480@ hrbeu. edu. cn.

机理,在雷达领域提出“认知”理念。随后,学者们逐渐应用于声呐系统设计,形成了“认知声呐”(cognitive sonar, CS)[9-10]。文献[11]描述了认知声呐的概念和结构,并使用多普勒处理和实时插值来调整认知声呐发射波束形成器。文献[12]开发了一种基于认知范式的集中式多基地目标连续跟踪方法,比非认知类算法具有更高的鲁棒性。文献[13]探讨了多音正弦调频自适应发射波形在主动认知声呐系统中的应用。通过调整谐波权重,能够针对新场景和环境进行微调。文献[14]提出了一种认知声呐波形设计方法,使用遗传算法对最大信噪比准则函数进行优化,获取发射线性调频信号的最优起始频率。近年来,强化学习(reinforcement learning, RL)的快速发展[15-18]使认知声呐系统的实时性获得了显著提升的可能,文献[19]提出了一种基于 Q 学习的目标回波分辨方法,能够实时调整波形参数,解决低速目标在浅海环境中受到回声杂波干扰的问题。然而,现有的认知声呐研究缺乏多参数耦合的高维决策机制且信道特性适配性不足。

本文提出一种基于多智能体强化学习(multi-agent reinforcement learning, MARL)的主动声呐发射波形与声源级联合优化方法。引入奖励塑形手段与双深度 Q 网络(double deep Q-Network, DDQN)网络提高智能体的决策能力。结合南海实测声速与 Bellhop 软件对所提方法的性能进行了验证与评价。

1 问题建模与系统分析

考虑如图 1 所示的水下主动声呐探测场景,由于水下环境中的反射和折射现象,声波沿着多条路径到达目标再返回接收器,多径效应引起信号到达延迟、相位和幅度差异。设定去程信道为 $h_1(t)$,回程信道为 $h_2(t)$,目标响应函数为 $h_s(t)$,考虑多径信道传播的水下目标的回波模型为:

$$r(t) = h_1(t) \otimes s(t) \otimes h_s(t) \otimes h_2(t) + n(t) \tag{1}$$

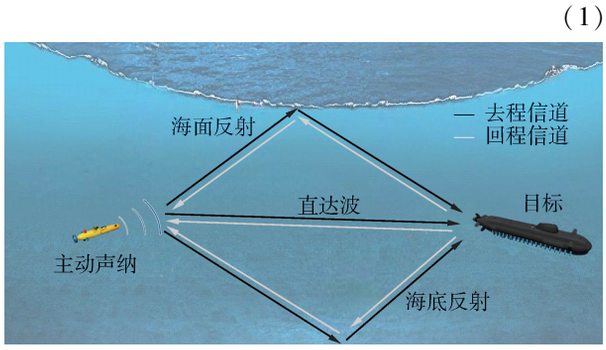


图 1 主动声呐探测场景

Fig. 1 Active sonar detection scene

令 $h(t) = h_1(t) \otimes h_s(t) \otimes h_2(t)$ 表示总传播信道。则式(1)可简化为:

$$r(t) = s(t) \otimes h(t) + n(t) \tag{2}$$

式中:$s(t)$ 为发射信号;$n(t)$ 为背景干扰;$h(t) = \sum_{k=1}^{K} \alpha_k \delta(t - \tau_k)$;$K$ 为回波多径的数量;α_k 为每条路径的衰减因子;τ_k 为每条路径的双程传播时延。

水声信道的多径结构会导致不同频点的信号经过信道后产生不同程度的衰减,即频率选择特性[20-21]。表现为信道的频谱能量分布起伏较大,在一些频段形成波峰,可以较好地传输发射信号的能量,在一些频段形成波谷,对信号能量传输造成较大抑制。频率选择性衰落源于不同路径信号在频域上的相位干涉。在频域中,不同频率分量因路径差异发生相干叠加,导致某些频率被增强,另一些被削弱。

假设总传播信道 $h(t)$ 的频域响应可表示为各路径贡献的复数和:

$$H(f) = \sum_{k=1}^{K} \alpha_k \exp(-\mathrm{j}2\pi f \tau_k) \tag{3}$$

$H(f)$ 的幅度响应为:

$$|H(f)| = \sqrt{\sum_{k=1}^{N} \alpha_k^2(f) + 2\sum_{i<j} \alpha_i(f)\alpha_j(f)\cos(2\pi f(\tau_j - \tau_i))} \tag{4}$$

相位响应为:

$$\angle H(f) = \arctan\left(\frac{-\sum_{k=1}^{K} \alpha_k(f)\sin(2\pi f\tau_k)}{\sum_{k=1}^{K} \alpha_k(f)\cos(2\pi f\tau_k)}\right) \tag{5}$$

式中∠表示复数的相位角。

从式(4)可以看出信道的幅频响应具有明显的干涉结构。以一个简单特例进一步分析,设存在 2 条路径(直射路径 $k=1$、反射路径 $k=2$),则信道的幅度响应为:

$$|H(f)| = \sqrt{\alpha_1^2 + \alpha_2^2 + 2\alpha_1\alpha_2\cos(2\pi f(\tau_2 - \tau_1))} \tag{6}$$

幅度响应呈现周期波动,干涉周期为 $\Delta f = \frac{1}{\tau_2 - \tau_1}$。当 $f = \frac{n}{\tau_2 - \tau_1}$($n$ 为整数),$|H(f)| = \alpha_1 + \alpha_2$。此时,$f$ 对应幅度响应的峰值,多径分量产生同相叠加效应。当 $f = \frac{n + 0.5}{\tau_2 - \tau_1}$,$|H(f)| = |\alpha_1 - \alpha_2|$。此时,$f$ 对应幅度响应的谷底,多径分量出现反相抵消。

基于该特性本文采用多载波梳状谱波形作为声呐发射信号,当信道环境发生变化时,声呐发出不同频谱分布的梳状谱信号来适应信道的幅频响应。梳

状谱波形由 N 个等间隔子脉冲组成,每个子脉冲为单频连续波(continuous wave, CW)。发射信号为:

$$s(t)=\sum_{i=1}^{N}A_i\cos(2\pi f_i t+\phi_i) \tag{7}$$

式中:A_i 为子脉冲幅度;f_i 为子脉冲频点;ϕ_i 为初始相位。假设每个子脉冲频点 f_i 对应的信道幅度为 $|H(f_i)|$,波形优化的目的是通过调梳状谱子脉冲的频谱分布,使得信道适配度(channel fitness, CF)尽可能大。CF 用以衡量发射信号能量与信道选择特性的匹配程度,其定义为:

$$C_F=\sum_{n}^{N}\frac{|H(f_i)|^2}{N\cdot\max(|H(f)|^2)} \tag{8}$$

除了频率选择特性,水声信道的幅度衰减特性反映了发射信号在海洋信道传播过程中的能量损失。固定的发射声源级(source level, SL)难以精准补偿动态变化的传播损失,使得接收端回波信噪比波动较大,导致探测虚警严重或发射能耗较高。因此,需要一种动态的 SL 调整策略,通过感知回波信噪比对 SL 进行调控,能够使声呐系统实现对接收回波能量水平的稳定控制。无论目标在何种距离,声呐都能在一个理想的接收水平上持续探测目标。确保声呐在复杂的海洋环境中,可以高效地获取目标信息,同时减少不必要的声呐能耗。

因此,本文的优化目标为联合调整声呐梳状谱发射波形频点分布 $\{f_1, f_2, \cdots, f_N\}$ 与声源级 SL,从而最大化信道适配度 CF 的同时最小化声源级 SL,即:

$$\max_{\{f_i\},\mathrm{SL}}\mathrm{E}[C_F]-\lambda\mathrm{E}[S_L],$$

$$\text{s.t.}\quad S_{\mathrm{NR}}=10\lg\left(\frac{\sum_{i=1}^{N}|H(f_i)|^2A_i^2}{P_{\mathrm{noise}}}\right)\geqslant S_{\mathrm{NR}}^{\mathrm{th}} \tag{9}$$

式中:$\mathrm{E}[\cdot]$ 为数学期望算子;S_{NR} 为优化后的回波信噪比;$S_{\mathrm{NR}}^{\mathrm{th}}$ 为设置的回波信噪比期望阈值,表示希望声呐探测系统接收端维持的探测信噪比;P_{noise} 为噪声功率。$|H(f_i)|^2$ 的大小与 CF 的大小呈线性正相关,A_i 的大小与 SL 的大小成正相关。

2　基于多智能体协作强化学习的发射参数联合优化

为求解式(9)所示的优化问题,本文利用多智能体深度强化学习建立声呐发射参数联合优化机制,通过智能体与环境的交互不断优化声呐发射的波形和声源级,使声呐在复杂海洋条件下实现稳健与高效的目标探测效果。

2.1　强化学习基本概念

强化学习是一种使智能体通过与环境的交互学习策略的机器学习方法[16-17],在时间步骤 t,智能体观察其周围的环境,并获得状态作为 $s^t\in S$,然后根据一定的策略 $\Psi(a\mid s)$ 执行动作 $a^t\in A$,即代表在状态 s^t 下采取动作 a^t 的概率。环境受到所执行的动作的影响,并相应地转移到下一个状态 s^{t+1}。奖励 $r^t=R(s^t,a^t)$ 用作评估动作 a^t 的作用效应,为智能体提供下一步的策略调整依据。智能体在时间步 t 时与环境的这种交互形成一种体验,该体验可以由一个元组(s^t,a^t,r^t,s^{t+1}) 描述,该过程又被称之为马尔科夫决策过程(Markov decision process, MDP)。MDP 的目标是寻找最优的策略使得系统能够最大化奖励。

2.2　多智能体强化学习框架设计

对于本文需要求解的主动声呐多发射参数联合优化问题,若使用单智能体强化学习(single-agent reinforcement learning, SARL)优化多个参数,会产生严重的维度爆炸问题。单个智能体必须同时处理所有参数的组合,参数空间呈指数级增长,会导致收敛速度慢,学习效率低,且容易陷入局部最优。为了解决这个问题,本文设计了如图 2 所示的多智能体强化学习框架,通过将优化任务分解波形优化和声源级调控,并分配给不同的智能体,减少每个智能体需要处理的状态和动作空间,缓解了维度爆炸,提高了优化效率。

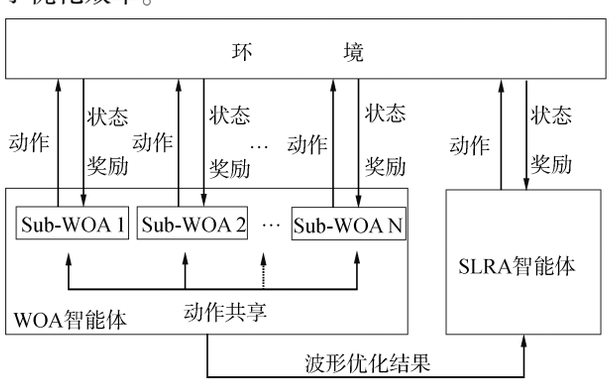


图 2　多智能体协作强化学习框架

Fig. 2　Multi-agent cooperative reinforcement learning framework

该框架包括 2 个主要的智能体组件。

1)波形优化智能体(waveform optimization agent,WOA):由 N 个子智能体(sub-WOA)组成,数量与梳状谱子脉冲数一致,每个智能体控制一个频点 f_i 的选择,根据当前频点的信道增益 $H(f_i)$ 选择一个动作优化目标检测性能。

每个 Sub-WOA 的状态空间为:

$$o_i^t=[f_i^t,|H(f_i^t)|^2] \tag{10}$$

式中:f_i^t 和 $|H(f_i^t)|^2$ 分别为当前时间步第 i 个 Sub-WOA 所选频点和对应的信道频谱的能量;f_i^t 为历史 T 步的频点选择状态。

每个 Sub-WOA 的动作空间 A_w 由子脉冲的不同频点构成,定义为 $A_w = \{f_1, f_2, \cdots, f_M\}$,$M$ 为动作空间的长度。

2)声源级调控智能体(sound source-level control agent,SLRA):该智能体根据实时回波信噪比调节发射的声源级从而保证系统的信噪比稳定性。

SLRA 的状态空间为全局状态空间:

$$o_{\mathrm{SLRA}}^t = [S_{\mathrm{NR}}^t, S_L^t] \tag{11}$$

式中:S_{NR}^t 为当前时间步的回波信噪比;S_L^t 为当前时间步的发射声源级。

SLRA 的动作空间 A_s 定义为 $A_s = \{p_1, p_2, \cdots, p_k\}$,该空间由不同大小的声源级构成。

在学习过程中,WOA 内部的各个子智能体可以获取其他子智能体的频点选择动作信息从而调整自身频点选择动作,防止频点聚集。SLRA 可以根据 WOA 的频点选择信息获取实时回波信噪比来调整对应声源级。

2.3 奖励塑形

对于 WOA 来说,通常需要使每个子智能体尽量寻找到最佳信道增益频点,才能保证叠加之后的梳状谱信号整体探测效果提升,这种情况下仅将信道适配增益作为子智能体的奖励函数即可,但若仅以信道增益最大化为目标,这种“贪婪策略”会形成囚徒困境,各 Sub-WOA 会独立抢占当前信道增益峰值频点,导致频点聚集,使得梳状谱的带宽过窄影响接收端匹配滤波的距离分辨力。因此,此任务并不属于完全合作型的多智能体任务,而是每个智能体都选择自己的最优动作,同时还要考虑其他智能体的频点信息,防止出现频点重叠。每个子智能体在获取了各个子智能的频点选择信息之后依靠有奖有惩的奖惩函数使其自发的产生频点冲突惩罚塑形下的合作策略。

第 i 个 Sub-WOA 子智能体奖励为:

$$r_i = \alpha \cdot \frac{|H(f_i)|^2}{\max(|H(f)|^2)} - \beta \cdot \sum_{j \neq i} \mathrm{e}^{-|f_i - f_j|/\Delta B} \tag{12}$$

式中:等式右边第 1 项为信道适配奖励,衡量频点与信道主瓣的对齐程度;第 2 项为频点冲突惩罚,为了防止子智能体选择同一个频点导致梳状谱波形设计失效;α 和 β 为权重系数。

式(11)的奖励函数设计已经较为合理,然而,由于信道频谱响应具有多峰特性,某些频点的信道增益可能在局部峰值附近变化剧烈,导致信道适配奖励出现突变。例如,当智能体调整频点时,即使微小的频偏也可能导致信道增益从峰值急剧下降到谷值,从而使得奖励函数剧烈波动,不利于智能体学习到稳定的策略。为了解决这个问题,本文设计了一种奖励平滑塑形方法,使得即使频点调整导致信道增益变化较大时,奖励函数的变化也较为平缓,从而引导智能体更稳定地探索和优化。

具体而言,对每个频点 f,计算其邻域 $[f - L\Delta f, f + L\Delta f]$ 内的高斯加权平均增益:

$$\tilde{H}(f_i) = \frac{1}{C} \sum_{l=-L}^{L} H(f + l\Delta f) \cdot \mathrm{e}^{-\frac{(l\Delta f)^2}{2\sigma^2}} \tag{13}$$

式中:σ 为高斯核标准差(控制平滑强度);$C = \sum_{l=-L}^{L} \mathrm{e}^{-\frac{(l\Delta f)^2}{2\sigma^2}}$ 为归一化因子;Δf 为频点分辨率;L 为窗口半宽。

至此,本文重构第 i 个 Sub-WOA 的奖励函数为:

$$\tilde{r}_i = \alpha \cdot \frac{|\tilde{H}(f_i)|^2}{\max(|\tilde{H}(f)|^2)} - \beta \cdot \sum_{j \neq i} \mathrm{e}^{-|f_i - f_j|/\Delta B} \tag{14}$$

对于 SLRA 来说,WOA 的频点选择直接影响 CF,而 CF 的优劣决定了回波信噪比的潜在水平。SLRA 的目标是通过调节声源级 SL 维持 SNR 稳定,同时最小化能耗。若 SLRA 不了解 WOA 的频点分布,则无法准确评估当前波形的信道适配效果,导致声源级的欠补偿或过补偿。因此 SLRA 的奖励函数为:

$$r_{\mathrm{SLRA}} = -\left|S_{\mathrm{NR}}^{\mathrm{new}} - S_{\mathrm{NR}}^{\mathrm{th}} + S_L^{\mathrm{old}} - S_L^{\mathrm{act}}\right| \tag{15}$$

式中:$S_{\mathrm{NR}}^{\mathrm{new}}$ 为根据 WOA 的频点选择信息计算得到的实时回波信噪比;S_L^{old} 为初始声源级;S_L^{act} 为智能体选择的动作对应的优化声源级。

2.4 基于 DDQN 的智能体学习网络

在强化学习中,智能体通过与环境交互学习如何在状态 s 下选择动作 a,以最大化其长期累积奖励。Q learning[22] 是一种常用的值迭代方法,通过学习状态-动作价值函数(Q 值函数)指导智能体的动作选择。Q 值函数 $Q(s,a)$ 表示在状态 s 下执行动作 a 后,智能体所能获得的预期累积奖励:

$$Q(s,a) = \mathrm{E}[r + \gamma \max_{a'} Q(s',a')] \tag{16}$$

式中:r 为即时奖励;γ 为折扣因子,分别表示当前奖励和未来奖励的重要性;s' 为更新后的状态;a' 为处于状态 s' 时,Q 值最大对应的动作。

在 Q learning 中,Q 值表随着每次交互不断更新,但对于高维状态空间,不可以直接存储和更新 Q 值表。为了解决这一问题,DQN[23] 使用神经网络 $Q(s,a;\theta)$ 逼近 Q 值函数,参数 θ 为网络权重,S 为神经网络输入的当前状态,输出的是所有可能动作的 Q 值。

传统的 DQN 在 Q 值更新时,容易因为过高估计最大 Q 值(即 $\max Q(s',a')$)而产生偏差,导致不稳定的学习。DDQN[24] 通过引入 2 个独立的 Q 值

网络来缓解这一问题,一个用于选择动作,另一个用于估计目标 Q 值。图 3 所示为 DDQN 的网络结构,DDQN 更新为:

$$y = r + \gamma Q(s', \mathrm{argmax}_{a'} Q(s', a'; \theta); \theta^-) \quad (17)$$

式中 θ 是当前 Q 值网络的参数,用于选择下一个状态 s' 的最优动作 a'; θ^- 是目标网络的参数,用于估计该动作的Q 值。

损失函数定义为:

$$L(\theta) = \mathrm{E}[(y - A(s', a'; \theta))^2] \quad (18)$$

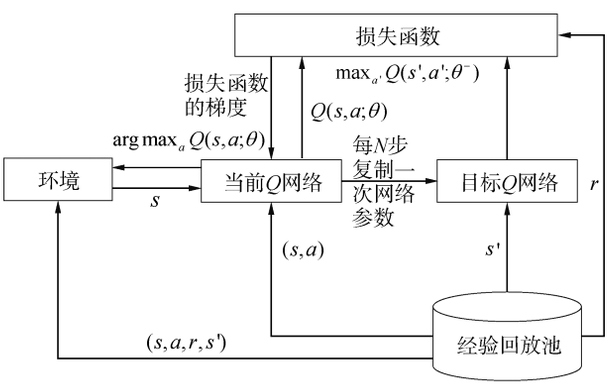


图 3　DDQN 网络结构

Fig. 3　DDQN network architecture

本文将 DDQN 作为本文所提方法中智能体的学习网络,通过使用不同的网络分别进行动作选择和 Q 值估计,DDQN 有效减轻了单一网络高估 Q 值的偏差问题,提升了 Q 值估计的准确性和学习稳定性。

2.5　多智能体强化学习算法流程

为了便于简洁表示,用 $i = 1, 2, \cdots, N, N+1$ 为所有智能体进行编号,其中前 N 个智能体代表 WOA 的子智能体,第 $N+1$ 个智能体代表 SLRA,本文所提算法的实现流程为:

1)初始化各智能体(包括所有 Sub-WOA 和 SLRA)的 DDQN 网络参数 θ、经验回放池 D_i、学习率 α、折扣因子 γ 和探索率 ε;

2)在每个训练轮次开始时,初始化全局环境状态;

3)对于轮次内的当前时间步 t,各智能体根据其当前状态 s_i^t 选择动作 a_i^t;所有智能体执行联合动作 $[a_1^t, a_2^t, \cdots, a_N^t, a_{N+1}^t]$;

4)环境据此转移至新状态并产生反馈。各智能体基于其状态 s_i^t、动作 a_i^t、新状态 s_i^{t+1} 及环境反馈计算奖励 r_i^t,并将经验元组 $(s_i^t, a_i^t, r_i^t, s_i^{t+1})$ 存入对应经验池 D_i。

5)各智能体从 D_i 中采样 mini-batch 样本,依据式(16)和(17)计算当前 Q 值和损失,并通过梯度下降更新参数 θ。每间隔 U 步,将当前网络参数 θ 复制至目标网络参数 θ^-。

6)训练过程中逐步衰减探索率 ε 以降低策略随机性。

2.6　计算复杂度分析

为了分析所提算法的计算复杂度,假设算法包含 N 个 Sub-WOA 和 1 个 SLRA,每个 Sub-WOA 的 DDQN 网络均有一个输入层(维度为 H_{in})、2 个隐藏层(维度分别为 H_1 和 H_2)和一个输出层(维度为 H_{out}),SLRA 的网络具有同样的结构,对应的参数分别为 H'_{in}、H'_1、H'_2 和 H'_{out}。

每个 Sub-WOA 的网络前向传播复杂度为:

$$\Gamma_{\mathrm{WOA}} = O(H_{\mathrm{in}}H_1 + H_1H_2 + H_2H_{\mathrm{out}}) \quad (19)$$

SLRA 的网络前向传播复杂度为:

$$\Gamma_{\mathrm{SLRA}} = O(H'_{\mathrm{in}}H'_1 + H'_1H'_2 + H'_2H'_{\mathrm{out}}) \quad (20)$$

每个时间步所有智能体与环境交互复杂度 Γ_1 为:

$$\Gamma_1 = O((N\Gamma_{\mathrm{WOA}} + \Gamma_{\mathrm{SLRA}}) \quad (21)$$

智能体从缓冲区采样大小为 B 的 mini-batch 并计算损失,经验回放更新复杂度 Γ_2 为:

$$\Gamma_2 = O(B(N\Gamma_{\mathrm{WOA}} + \Gamma_{\mathrm{SLRA}})) \quad (22)$$

使用梯度下降更新 DDQN 参数,网络参数更新的复杂度 Γ_3 由网络总参数量决定:

$$\Gamma_3 = O((N\Gamma_{\mathrm{WOA}} + \Gamma_{\mathrm{SLRA}}) \quad (23)$$

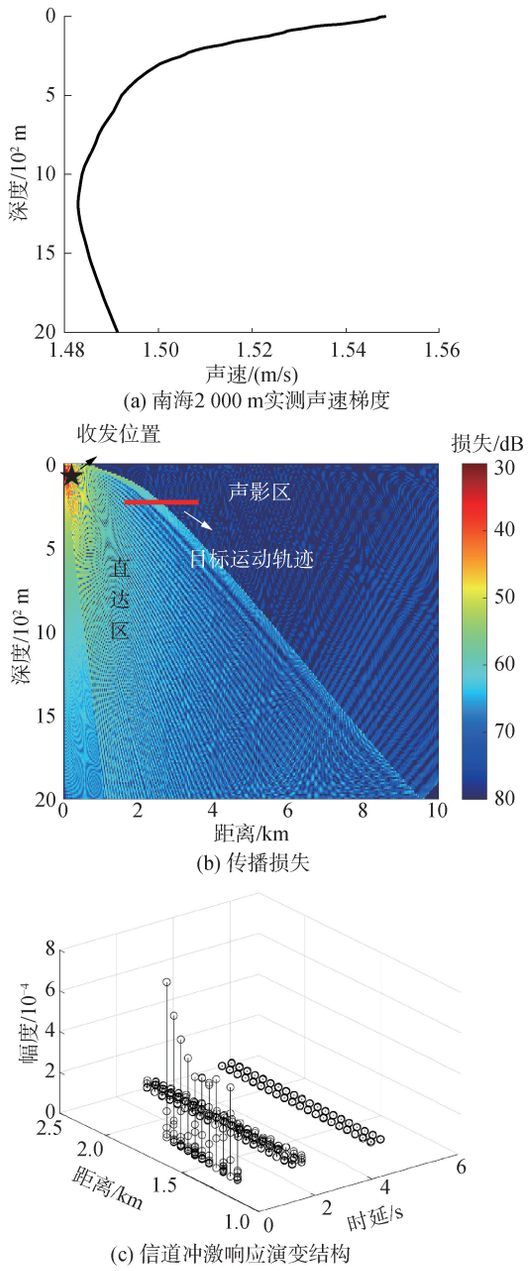
因此经过 E 个训练回合,T 个时间步的训练阶段总复杂度为 $O(E \times T \times (\Gamma_1 + \Gamma_2 + \Gamma_3))$。执行阶段在线决策仅需前向传播,计算复杂度为 $O(N \times \Gamma_{\mathrm{WOA}} + \Gamma_{\mathrm{SLRA}})$。

3　数值验证

基于南海实测声速梯度数据与 Bellhop 软件重构得到的典型深海信道环境对所提算法的性能进行验证。

3.1　基于 Bellhop 的水声信道重构

图 4(a)给出了 2 000 m 深度的南海实测声速梯度,将其输入 bellhop,设置中心工作频率为 6 kHz,收发深度为 50 m,目标深度为 300 m,海深 2 000 m,海底底质为泥沙,海底衰减系数和密度分别为 0.8 dB/λ 和 1.8 g/cm^3,得到了如图 4(b)所示的传播损失。可以看到声传播模式呈明显的分层特性,左下为直达声区,右上为声影区。目标在直达声区内运动时,传播损失较小,当运动到声影区时,传播损失急剧增大。为了模拟水下探测时,目标位置从直达声区变换到声影区的过程,设定了一组目标水平距离点位,设定目标深度为 300 m,距离点位水平分布范围为 1.8~3.8 km,步进为 100 m,得到的信道冲激响应演变结构如图 4(c)所示。可以看到第 1~11 帧信道位于直达声区,存在较强的直达声线,而第 12~20 帧信道位于声影区,不存在直达声线。



(a) 南海2 000 m实测声速梯度

(b) 传播损失

(c) 信道冲激响应演变结构

图 4 实测深海声速剖面及 Bellhop 仿真结果

Fig. 4 Measured deep-sea sound velocity profile and Bellhop simulation results

由于 Bellhop 仿真信道为一发一收单程信道，为了模拟收发合置声呐主动探测的双程信道，根据互易性原理，将该信道数据与自身进行卷积得到总传播信道，用于本文所提算法验证。声呐初始发射声源级设置为 180 dB，目标点位对应的传播损失从 Bellhop 结果中提取得到，声呐工作频段为 2～8 kHz，采样率 32 kHz，发射信号波形为梳状谱波形，子脉冲为 CW 脉冲，个数设置为 5，海洋环境背景噪声级设置为 55 dB，目标强度设置为 9 dB，接收为 16 元直线阵，回波信噪比的联合优化期望阈值设置为 0 dB。

3.2 训练参数设置与训练过程

本文所提算法在训练时设置折扣因子为 0.5，全连接层大小为[256, 256, 32]，探索因子从 1 逐渐缩减至 0.001，权重系数 α 和 β 分别设置为 1 和 0.2，批处理样本数量为 256，最大训练轮次为 600，每轮次循环次数为 100，目标网络更新频率为 0.1，经验回放缓冲区最大长度为 1×10^6，激活函数选用 Tanh 函数。

本文使用平均奖励评价训练效果，平均奖励由数据训练过程中所有智能体获得奖励的总和取平均得到。图 5 显示了不同学习率下的平均奖励收敛曲线，可以看出当学习率过小时，智能体的学习过程会变得非常缓慢，需要较长时间才能收敛，随着学习率的增大，收敛速度逐渐加快，但当学习率过大时，智能体可能会跳过潜在的优秀策略使得最终的学习效果不如预期，因此最终本文将算法的学习因子设置为 1×10^{-3}。

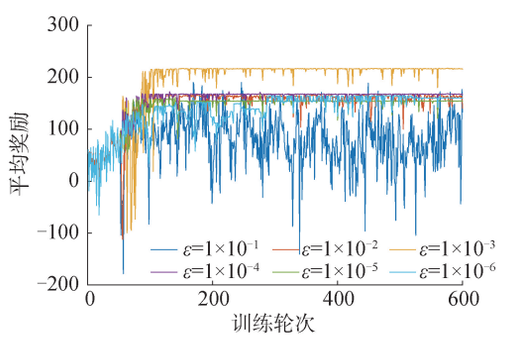


图 5 不同学习率下的平均奖励对比

Fig. 5 Comparison of average rewards under different learning rates

3.3 多参数联合优化结果

图 6 展示了在 20 帧动态信道条件下，基于所提方法的发射波形优化与固定发射波形的对比结果，每帧对应不同的信道状态。可以看出，优化波形的子脉冲频点分布随帧序号显著变化，优先选择信道增益高的“波峰”频段，避开衰减严重的“波谷”频段。固定波形的子脉冲始终均匀分布于 3、4、5、6、7 kHz，缺乏环境反馈，未适配信道特性，无法规避信道衰减。通过奖励函数中的频点冲突惩罚项，WOA 智能体自主避免相邻子脉冲频点过近，保障波形带宽与距离分辨力。

图 7 显示了所提算法在 20 帧信道数据中相比于固定发射策略获得的信噪比优化增益，其为波形优化前后的回波信噪比之差。由于固定发射策略在每一帧中对信道的适配程度动态变化，因此优化结果呈现明显的波动趋势，例如信噪比优化增益达到 8.44 dB 的第 6 帧数据其固定发射波形的频点对应信道衰减较为严重，因此其优化空间较大，如图 8(a)所示，图中归一化幅度表示对信道频谱进行归一化处理。而低提升帧(如第 10 帧仅提升 1.2 dB，

其对应的信道直达途径存在单根能量较突出的声线,导致干涉效应不明显)则是因为固定策略的发射频点对应的信道增益已经较高,使得优化空间被物理性压缩,如图 8(b)所示。整体来看,所提算法使所有帧的发射波形频谱分布尽可能与信道选择特性相适配,均实现了正向优化,平均信噪比优化增益达到 5.24 dB,使得下一步的声源级优化可以达到更低的功耗,实现更高效的发射资源管理。

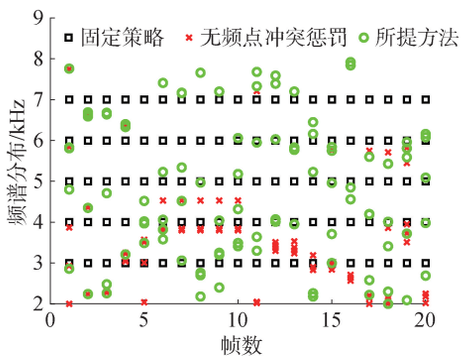


图 6　波形优化结果

Fig. 6　Waveform optimization results

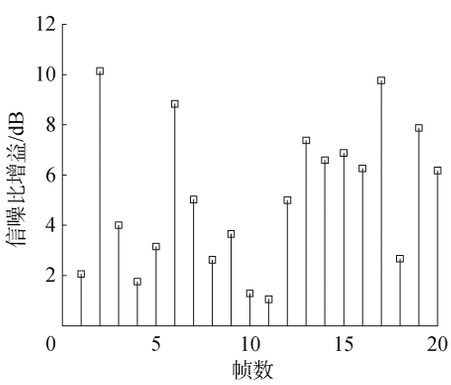


图 7　波形优化后的信噪比增益

Fig. 7　SNR gain after waveform optimization

图 9 显示了联合优化后发射声源级的调控结果,可以看出,优化后的发射声源级不再固定不变,而是更加灵活。在目标位于近距离直达声区阶段,由于传播损失较低,智能体感知到固定发射策略对应的回波信噪比超过设定阈值,适应性地降低发射声源级以节省发射能量,平均每帧声源级降低 32.5 dB;当目标进入声影区时,传播损失急剧增大,智能体感知到固定发射策略对应的回波信噪比低于设定阈值,随之提高发射声源级使得声呐在远距离恶劣状态下也能进行稳健探测。除此之外,与仅优化发射声源级相比,联合发射波形的优化使得每一帧的发射声源级进一步降低,使声呐系统的能耗得到了进一步缩减。

3.4　性能对比分析

为了定性验证所提算法对多发射策略联合优化的性能,本文采用 2 种基准方案进行对比:1)基于单智能体 DQN 算法的方案(single-agent deep Q-network, SADQN),仅使用一个智能体同时调控所有发射参数;2)基于多智能体 DQN 算法的方案(multi-agent deep Q-network, MADQN),采用与本文一致的多智能体协作框架,但不对奖励函数和网络结构进行改进。所有方法在上述每帧信道条件下进行了 20 次试验,统计平均其结果。其中,梳状谱波形优化性能由信道适配度 CF 衡量,其范围为 0~1,CF 越大,代表梳状谱波形与信道的频率选择适配性最高。声源级优化性能使用优化后回波信噪比的均值和标准差来衡量,均值越接近设定信噪比,标准差越小,代表声源级优化性能越好。

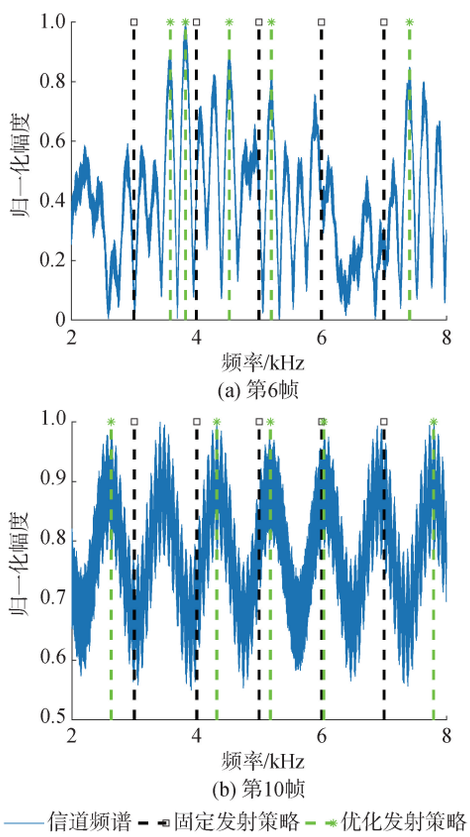


图 8　波形优化细节

Fig. 8　Waveform optimization details

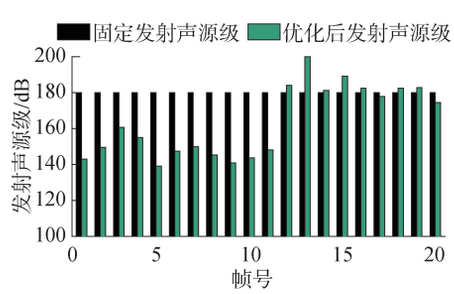


图 9　发射声源级优化结果

Fig. 9　Optimization results of emission sound source level

图 10 展示了 4 种算法(固定策略、SADQN、MADQN 算法及所提算法)在 20 帧信道条件下的

CF 对比结果。固定策略的信道适配度随帧数的增加波动程度较大,这是因为其完全依赖预设发射波形,无法响应信道变化。SADQN 算法由于使用单个智能体集中调节所有发射参数,动作空间维度过大,智能体难以学习至收敛,因此其 CF 的优化效果波动较大,在某些帧的 CF 甚至低于固定策略。MADQN 算法相比于前 2 种策略,优化效果有了明显提升,CF 均提高到了 0.7 以上,但由于信道频谱函数的复杂多峰效应,使其容易陷入局部最优,导致 CF 的提升有限。经过所提方法处理之后,相比于多智能体 DQN 算法,信道适配度得到了进一步的提升,达到了 0.8 以上,相较于 MADQN 算法平均提升了 12%,提升效果在声影区信道(第 12~20 帧)体现得更为明显,这是因为声影区的信道频谱分布中的高频噪声更加明显,此时奖励平滑塑形的作用更大,有效改善了此场景下智能体容易陷入局部最优的问题。

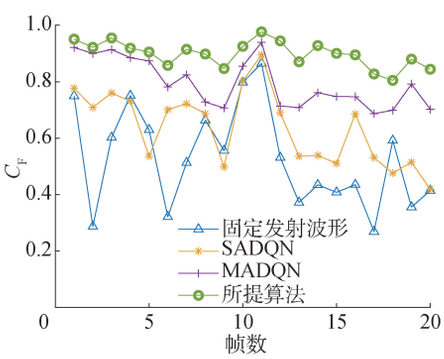


图 10 不同方法的信道适配度对比

Fig. 10 Comparison of CF of different methods

图 11 显示了所提算法与对比算法优化发射声源级后回波信噪比的均值和标准差的对比,图 11 中可以看出,SADQN 算法因动作空间维度爆炸问题(需同时优化多个频点分布与声源级),难以收敛,导致优化误差最大。MADQN 算法的表现较 SADQN 明显提升,但其调控后的回波信噪比仍然存在一定程度的波动。这是由于 WOA 与 SLRA 是协同优化,SLRA 的探索性能会受 WOA 的影响。在 MADQN 算法中,由于 WOA 在寻找最优信道增益频点时,容易出现难以收敛或陷入局部最优的问题,导致下一步 SLRA 的探索效果也受到影响,表现为 MADQN 算法优化后的回波信噪比统计具有较大的标准差。由于固定发射策略对声源级不产生调节作用,因此在图中并未展示。而所提算法调控后的回波信噪比基本保持在设定的 0 dB 附近,且相比于对比算法其波动程度较小,统计优化结果具有更小的标准差(小于 0.41 dB),表现出较好的稳定性,能够有效适应环境变化,确保探测回波的质量。

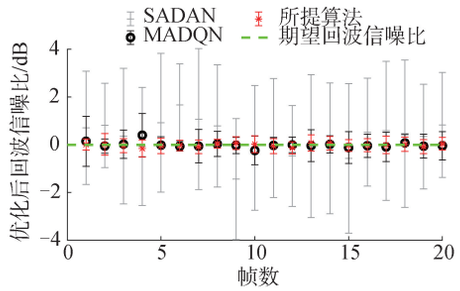


图 11 不同方法的声源级调控效果对比

Fig. 11 Comparison of sound source level control effects of different methods

4 结论

1)针对主动声呐在复杂水声环境中面临的探测性能波动与能量利用效率低等问题,本文将多智能体强化学习引入主动声呐多发射参数优化问题。在保证接收端稳健检测的约束下,最大化梳状谱发射波形与水声信道频率选择特性的适配度,同时最小化发射声源级,节省发射功耗,为复杂水声环境下的声呐智能化探测提供一种新范式。

2)添加高斯平滑引导与频点冲突惩罚等奖励塑形机制,使智能体兼顾波形带宽与信道适配性。使用 DDQN 作为学习网络,提高了智能体的决策稳定性。

3)在深海近程水声信道条件下,所提算法通过自适应调整梳状谱频点分布,使信道适配度提升至 0.8 以上,较基线方法平均提升了至少 12%。通过联合优化策略,结合声源级动态调控,使得回波信噪稳定在设定阈值,波动标准差小于 0.41 dB,明显优于基线方法,显著提升了探测稳定性。

本文仍存在一定局限性:算法验证依赖于信道模型(Bellhop)的准确性,实际复杂海洋环境中的信道预测误差可能影响算法性能;多智能体决策速度有限,可能影响实时性应用。未来工作将结合迁移学习与网络轻量化技术,进一步提升算法在真实动态环境中的实时优化能力。

参考文献:

[1] GUO Qijia, XIE Kean, YE Weibin, et al. A sparse Bayesian learning method for moving target detection and reconstruction[J]. IEEE transactions on instrumentation and measurement, 2025, 74: 4505413.

[2] 兰朝凤,郑智韦,陈欢. 基于复杂声传播环境的水下作战效能评估[J]. 哈尔滨工程大学学报,2025,46(1):166-172.
LAN Chaofeng, ZHENG Zhiwei, CHEN Huan. Method for assessing the combat effectiveness of underwater unmanned clusters based on a complex acoustic propagation environment[J]. Journal of Harbin Engineering University, 2025, 46(1): 166-172.

[3] 佟文涛,葛威,殷敬伟,等. 水声单载波通信中的块稀疏

均衡器[J]. 声学学报,2025,50(2):511-524.
TONG Wentao, GE Wei, YIN Jingwei, et al. Block-wise sparse equalizer for underwater acoustic single-carrier communication[J]. Acta acustica,2025,50(2):511-524.
[4] 梁国龙,张博宇,齐滨,等. 无源声呐水下多目标融合跟踪方法[J]. 声学学报,2024,49(3):501-512.
LIANG Guolong, ZHANG Boyu, QI Bin, et al. Underwater multitarget fusion tracking method for passive sonar[J]. Acta acustica, 2024, 49(3): 501-51.
[5] ZHANG Yi, VENKATESAN R, DOBRE O A, et al. Efficient estimation and prediction for sparse time-varying underwater acoustic channels[J]. IEEE journal of oceanic engineering, 2020, 45(3): 1112-1125.
[6] 郑巧宁, 郑浩赐, 李茂林, 等. 采用改进支持向量机的浅海水声信道小样本估计[J]. 哈尔滨工程大学学报, 2025, 46(3): 390-400.
ZHENG Qiaoning, ZHENG Haoci, LI Maolin, et al. Shallow water acoustic channel small sample estimation using enhanced support vector machines[J]. Journal of Harbin Engineering University, 2025, 46(3): 390-400.
[7] 李昊鑫,肖长诗,元海文,等. 特征降维与融合的水声目标识别方法[J]. 哈尔滨工程大学学报,2025,46(1):102-110.
LI Haoxin,XIAO Changshi,YUAN Haiwei,et al. Underwater acoustic target recognition method based on feature dimension reduction and fusion[J]. Journal of Harbin Engineering University,2025,46(1):102-110.
[8] HAYKIN S. Cognitive radar: a way of the future[J]. IEEE signal processing magazine, 2006, 23(1): 30-40.
[9] LI Xiaohua, LI Yaan, CUI Lin, et al. Research of new concept sonar-cognitive sonar[J]. Journal of marine science and application, 2011, 10(4): 502-509.
[10] LI Xiaohua, LI Yaan, LIN Guancheng, et al. Research of the principle of cognitive sonar and beamforming simulation analysis[C]//2011 IEEE International Conference on Signal Processing, Communications and Computing (ICSPCC). Piscataway, NJ, 2011: 1-5.
[11] CLAUSSEN T, NGUYEN V D. Real-time cognitive sonar system with target-optimized adaptive signal processing through multi-layer data fusion[C]//2015 IEEE International Conference on Multisensor Fusion and Integration for Intelligent Systems (MFI). Piscataway, NJ, 2015: 357-361.
[12] LU Shuping, CHEN Yang, CHEN Fangxiang, et al. Cognitive continuous tracking algorithm for centralized multistatic sonar systems[C]//2021 OES China Ocean Acoustics (COA). Piscataway, NJ, 2021: 1021-1026.
[13] HAGUE D A. Adaptive transmit waveform design for active cognitive sonar using multi-tone sinusoidal frequency modulation[J]. The journal of the acoustical society of America, 2022, 151(4): A101-A101.
[14] PAKDEL A O, AMIRI H, RAZZAZI F. Enhanced target detection using a new cognitive sonar waveform design in shallow water[J]. Applied acoustics, 2023, 205: 109270.
[15] ARULKUMARAN K, DEISENROTH M P, BRUNDAGE M, et al. Deep reinforcement learning: a brief survey[J]. IEEE signal processing magazine, 2017, 34(6): 26-38.
[16] WANG Qiang, ZHAN Zhongli. Reinforcement learning model, algorithms and its application[C]//2011 International Conference on Mechatronic Science, Electric Engineering and Computer (MEC). Piscataway, NJ, 2011: 1143-1146.
[17] KOBER J, BAGNELL J A, PETERS J. Reinforcement learning in robotics: a survey[J]. The international journal of robotics research, 2013, 32(11): 1238-1274.
[18] LEE J, NIYATO D, GUAN Yong liang, et al. Learning to schedule joint radar-communication with deep multi-agent reinforcement learning[J]. IEEE transactions on vehicular technology, 2022, 71(1): 406-422.
[19] FU Yubin, MA Xiaochuan, FENG Chao, et al. Model-based optimal action selection for Dyna-Q reverberation suppression cognitive sonar[J]. EURASIP journal on advances in signal processing, 2023, 2023(1): 116.
[20] WISNIEWSKA D M, JOHNSON M, BEEDHOLM K, et al. Acoustic gaze adjustments during active target selection in echolocating porpoises[J]. Journal of experimental biology, 2012, 215(pt 24): 4358-4373.
[21] KAM C, KOMPELLA S, NGUYEN G D, et al. Frequency selection and relay placement for energy efficiency in underwater acoustic networks[J]. IEEE Journal of oceanic engineering, 2013, 39(2): 331-342.
[22] WATKINS C J C H, DAYAN P. Q-learning[J]. Machine learning, 1992, 8: 279-299.
[23] VAN H H, GUEZ A, SILVER D. Deep reinforcement learning with double Q-learning[C]//Proceedings of the AAAI Conference on Artificial Intelligence. Phoenix, 2016: 431-442.
[24] XI Lei, YU Lu, XU Yanchun, et al. A novel multi-agent DDQN-AD method-based distributed strategy for automatic generation control of integrated energy systems[J]. IEEE transactions on sustainable energy, 2020, 11(4): 2417-2426.

本文引用格式:

生雪莉,穆梦飞,毕耀,等. 多智能体强化学习驱动的主动声呐发射参数联合优化[J]. 哈尔滨工程大学学报,2025,46(8):1557-1565.
SHENG Xueli, MU Mengfei, Bi YAO,et al. Joint optimization of transmit parameters of active sonar driven by multiagent reinforcement learning[J]. Journal of Harbin Engineering University,2025,46(8):1557-1565.