

基于深度强化学习与混合专家模型的水声通感一体波束赋形设计

叶小文, 林恒羿, 周茂炜, 吴怡

(福建师范大学光电与信息工程学院, 福建 福州 350117)

摘要: 针对水声通感一体化系统中通信与感知性能均衡问题, 提出一种兼具高效性与泛化能力的波束赋形方案。首先, 针对水声长传播时延所导致的信道状态信息过期及跨时隙回波干扰问题, 引入深度强化学习框架进行波束赋形决策。其次, 为克服传统深度强化学习方法在适应多样化性能需求方面的局限, 构建基于多任务混合专家模型的神经网络, 以适配不同目标偏好因子下的通感性能权衡。此外, 为刻画当前波束赋形决策对系统性能的长期影响, 设计多步跨任务经验回放机制, 联合利用多任务样本与多步奖励提升学习效率。仿真结果表明, 相较于基准方案, 所提方法能够在不同目标偏好因子下取得更具帕累托效率的解, 并展现出更强的泛化能力。

关键词: 水声网络; 通感一体化; 深度强化学习; 混合专家模型

中图分类号: TN929.5

文献标志码: A

doi: 10.11959/j.issn.1000

Underwater acoustic ISAC beamforming based on deep reinforcement learning with mixture of experts

Ye Xiaowen, Lin Hengyi, Zhou Maowei, Wu Yi

College of Photonic and Electronic Engineering, Fujian Normal University, Fuzhou 350117, China

Abstract: To achieve the trade-off between communication and sensing performance in underwater acoustic integrated sensing and communication systems, an efficient and highly generalizable beamforming scheme was proposed. First, to address the issues of outdated channel state information and cross-time slot echo interference caused by long propagation delays, a deep reinforcement learning framework was introduced for beamforming decision-making. Second, to overcome the limitations of traditional deep reinforcement learning methods in adapting to diverse performance requirements, a neural network based on a multi-task mixture-of-experts model was constructed to adapt to the trade-offs between communication and sensing performance under different objective-preference factors. In addition, to characterize the long-term impact of current beamforming decisions on system performance, a multi-step cross-task experience replay mechanism was designed, which combines multi-task samples and multi-step rewards to improve learning efficiency. Simulation results demonstrated that compared with benchmark schemes, the proposed scheme obtained more Pareto-efficient solutions under different objective-preference factors and exhibited stronger generalization capability.

Key words: underwater acoustic networks, integrated sensing and communication, deep reinforcement learning, mixture of experts

收稿日期: 2026-08-21; 修回日期: 2026-0x-0x

通信作者: 吴怡, wuyi@fjnu.edu.cn

基金项目: 国家自然科学基金资助项目 (No.U25A20398, No.62501157); 福建省青年科技人员育成基金资助项目 (No.2025350410)

Foundation Items: The National Natural Science Foundation of China (No.U25A20398, No.62501157), The Foundation for Cultivated Young Talents of Fujian Province (No.2025350410)

0 引言

水声网络是实现水下设备广泛互联和开展持续海洋观测的关键支撑^[1]。随着水下感知与通信应用不断增多,传感器规模化部署对频谱资源的需求持续增长,而可用水声频谱十分有限,供需矛盾日益突出。通感一体化(integrated sensing and communication, ISAC)将声呐感知与水声通信集成于同一系统,通过共享频谱、波形和硬件资源,为缓解频谱资源紧张提供了新的技术途径^[2]。这种一体化设计不仅能够提高频谱利用效率,还可以降低系统成本与实现复杂度,因而成为未来水声网络的重要发展方向之一^[3]。

目前,研究者围绕水声通感一体化系统的性能优化开展了大量研究^[4-21]。其中,文献[4-15]主要研究水声通感一体化波形设计,相关方案大致可分为单载波和多载波两类。在单载波系统中,文献[4]提出扇区调制线性调频(linear frequency modulation, LFM)信号,通过菲涅尔域到时域映射实现数据的直接解调,同时提高速度估计性能。在此基础上,文献[5]进一步设计了一种混合单载波波形,将用于通信的连续相位调制与用于感知的 LFM 信号相结合。然而,LFM 波形存在较为严重的距离与多普勒耦合问题,且干扰抑制能力有限。为降低多传感器之间的相互干扰,文献[6]提出一种基于广义正弦调频的水声通感一体化波形。文献[7]进一步将广义正弦调频与二进制相移键控调制相结合,实现了较高的距离和速度分辨率。此外,文献[8]采用改进遗传算法优化广义正弦调频的参数,文献[9]则研究了叠加波形以及基于压缩感知的检测干扰消除方案。不同于上述以单载波波形设计为主的研究,近年来研究者开始探索多载波水声通感一体化系统^[10-15]。文献[10]设计了一种新型感知接收机,采用加权马氏距离区分正交时频空反射信号与环境噪声。文献[11]针对下行正交频分复用信号,提出了一种最大似然多普勒频移估计方法。文献[12]的研究结果表明,在水声双扩展信道中,正交啁啾分复用波形的性能优于正交频分复用波形。文献[13]和文献[14]分别基于正交频分复用与正交啁啾分复用信号,设计了面向多传感器协作的通信辅助感知方案。此外,文献[15]通过联合设计发射端波形与接收端信道估计方法,对水声双弥散信道下的传输问题进行了研究。

除波形设计外,研究者还围绕水声通感一体化系统的传输资源分配和自主式水下潜航器轨迹优化开展了相关研究^[16-21]。文献[16]将感知与通信任务分配至不同时隙,并采用演员-评论家算法,根据感知数据优化潜航器的航行轨迹。在此基础上,文献[17]面向多个潜航器共存的场景,设计了一种基于多智能体近端策略优化的覆盖路径规划方案。文献[18]利用先验检测信息调整通信发射功率,以降低数据包碰撞率。文献[19]设计了一种二维分组随机搜索算法,对子载波交织模式和波束赋形进行联合优化。文献[20]提出了一种测距增强方法,以提高目标感知精度。此外,文献[21]研究了一种基于超结构的水声通信与波达方向感知方案。

然而,将上述水声通感一体化方案^[4-21]应用于实际系统时,仍面临以下三方面问题。第一,现有方案普遍忽略了水声网络中的长传播时延。由于电磁波传播速度极快,传播时延在地面通感一体化系统中通常可以忽略;而水中声速仅约为 1500 m/s,传播时延不可忽视。长传播时延会形成水声场景特有的跨时隙干扰:先前发射信号产生的感知回波可能与后续通感一体化信号在通信接收端叠加,从而干扰后续信号的正常接收。此外,在长传播时延的影响下,基于瞬时信道状态信息所做出的波束赋形决策,可能无法匹配信号实际传输时的信道状态。第二,现有研究未充分考虑水声信道复杂的时间相关性。本质上,水声通感波束赋形的联合优化属于长期时序决策问题。传统数学优化方法仅关注系统的瞬时性能,既未能有效利用历史信道信息所蕴含的演化规律,也忽略了当前决策对未来性能的延迟影响,因而难以在全时间维度上保证决策的全局最优性。相比之下,无模型的深度强化学习(deep reinforcement learning, DRL)方法能够通过与环境持续交互来优化长期累积收益,更适用于处理此类时序决策问题。第三,现有基于 DRL 的水声网络研究^[22-24]通常对通信与感知任务进行独立优化,且目标偏好因子固定不变,导致算法在通感性能动态权衡场景下的泛化能力受限。由于 DRL 算法的性能在很大程度上依赖于神经网络是否针对特定偏好因子训练,当实际场景中出现训练阶段未曾见过的目标偏好因子时,神经网络往往需要从头开始训练,显著降低了算法适应新任务的效率。因此,亟需设计一种高效的水声通感一体化优化方案,以解

决上述问题,这也是本文的研究动机。

本文融合多任务深度强化学习与混合专家模型 (mixture-of-experts, MoE), 提出一种具备多目标优化与泛化能力的水声通感一体化波束赋形优化方法, 称为深度多任务混合专家 (deep multi-task MoE, DeepMTE)。与现有研究独立优化水声通信与感知功能不同, 本文将通信与感知波束赋形联合优化建模为多目标马尔可夫决策过程^[25], 旨在满足最大发射功率约束的前提下, 实现通信总速率与感知信噪比之间的有效权衡。为实现多目标优化并提升算法对不同目标偏好因子的泛化能力, 本文将 MoE 架构^[26]嵌入深度确定性策略梯度算法^[27], 将每个目标偏好因子对应的优化过程视为一项独立任务, 并通过共享门控网络与专家网络进行多任务联合训练。此外, 针对传统经验回放机制难以有效评估长传播时延下当前波束赋形决策对后续多个时隙系统性能的影响, 且难以适配多任务联合训练需求的问题, 本文设计了一种多步跨任务经验回放机制, 通过联合利用不同任务的经验样本共同训练 MoE 架构, 并采用多步回报综合评估当前决策对后续多个时隙系统性能的影响。仿真结果表明, 与其他方案相比, DeepMTE 所得解具有更高的帕累托效率与更快的收敛速度, 同时对不同目标偏好因子和天线数量表现出更强的鲁棒性。

本文的主要贡献总结如下。

(1) 构建面向多目标水声通感一体波束赋形优化的联合多任务 DRL 与 MoE 框架。该框架将不同目标偏好因子对应的优化过程视为不同任务, 通过共享门控网络和专家网络开展多任务联合训练, 以实现通信与感知性能之间的有效权衡, 并提升 DeepMTE 在不同目标偏好因子下的泛化能力。

(2) 设计多步跨任务经验回放机制。与传统经验回放^[28]不同, 该机制从不同任务的经验池中提取样本, 对 MoE 架构进行联合训练, 同时引入多步学习方法^[29], 综合当前与未来奖励以评估长传播时延下波束赋形决策对当前及后续系统性能的影响, 从而有效提升 DeepMTE 的学习效率。

符号说明: 向量 (矩阵) 采用粗体小写 (大写) 字母表示, $\mathbb{C}^{N \times K}$ 表示 $N \times K$ 维复矩阵空间, $|\cdot|$ 表示绝对值, $\|\cdot\|_2$ 表示 2 范数, $\lceil \cdot \rceil$ 表示向上取整函数, $\mathbb{E}[\cdot]$ 表示数学期望; $\text{Tr}(\cdot)$ 、 $(\cdot)^T$ 、 $(\cdot)^H$ 分别表示矩阵迹、转置与共轭转置。

1 系统模型

考虑如图 1 所示的水声 ISAC 系统, 其中浮标配备由 N 个阵元构成的均匀线阵, 在服务 U 个单天线水下传感器的同时, 探测一个点状目标。传感器集合记为 $u \in \mathcal{U} = \{1, 2, \dots, U\}$ 。在三维笛卡尔坐标系中, 浮标、目标及传感器 u 的位置坐标分别记为 $[0, 0, 0]$, $\mathbf{g}_{\text{tar}} = [x_{\text{tar}}, y_{\text{tar}}, z_{\text{tar}}]$, $\mathbf{g}_u = [x_u, y_u, z_u]$ 。系统以离散时隙 $t \in \mathcal{T} = \{1, 2, \dots, T\}$ 进行传输, 记 Δt 为一个时隙的持续时间。根据文献^[22], 浮标到目标、浮标到传感器 u 以及目标到传感器 u 的单向传播时延分别表示为 $T_{\text{b,tar}} = \lceil \|\mathbf{g}_{\text{tar}}\|_2 / (c \cdot \Delta t) \rceil$, $T_{\text{b,u}} = \lceil \|\mathbf{g}_u\|_2 / (c \cdot \Delta t) \rceil$, $T_{\text{tar,u}} = \lceil \|\mathbf{g}_{\text{tar}} - \mathbf{g}_u\|_2 / (c \cdot \Delta t) \rceil$, 其中 $c \approx 1500 \text{ m/s}$ 为水声信号的传播速度。

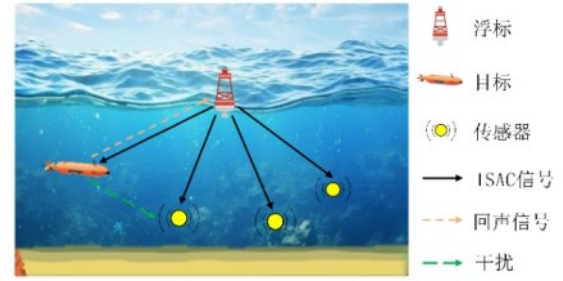


图1 水声 ISAC 系统

(1) 水声信道模型。设 $\mathbf{h}_{\text{b,tar}}(t) \in \mathbb{C}^{N \times 1}$, $\mathbf{h}_{\text{b,u}}(t) \in \mathbb{C}^{N \times 1}$ 与 $h_{\text{tar,u}}(t)$ 分别表示浮标到目标、浮标到传感器 u 以及目标到传感器的信道函数。各信道在一个时隙内保持不变, 但可随不同时隙发生变化。一般而言, 水声信道同时受大尺度效应 (如路径损耗) 与小尺度效应 (如多径衰落) 的影响^[3]。对于距离 d 和频率 f , 与距离和频率相关的路径损耗可表示为 $A(d, f) = A_0 d^k (a(f))^d$, 其中 A_0 为归一化常数, k 为扩展系数。根据索普公式^[31], 当 f 以 kHz 为单位时, 以 dB/km 为单位的吸收系数 $a(f)$ 表示如下: $10 \lg a(f) = 0.11 f^2 / (1 + f^2) + 44 f^2 / (4100 + f^2) + 2.75 \times 10^{-4} f^2 + 0.003$ 。文献^[31]指出, 浅海声传播主要存在三条路径: 直射路径、海面反射路径与海底反射路径, 对应的传播距离分别记为 d_{LoS} 、 $d_{\text{NLoS,s}}$ 、 $d_{\text{NLoS,b}}$ 。其中, 直达径属于视距路径, 而海面反射路径和海底反射路径则视为非视距路径。

本文对所有信道采用莱斯信道模型^[31]。例如,

在时隙 t , 信道 $\mathbf{h}_{b,\text{tar}}(t)$ 可表示为:

$$\begin{aligned} \mathbf{h}_{b,\text{tar}}(t) = & \frac{1}{\sqrt{A(d_{\text{LoS},s}f)}} \sqrt{\frac{\omega}{1+\omega}} \mathbf{h}_{b,\text{tar}}^{\text{LoS}}(t) + \\ & \frac{1}{\sqrt{A(d_{\text{NLoS},s}f)}} \sqrt{\frac{1}{2(1+\omega)}} \mathbf{h}_{b,\text{tar}}^{\text{NLoS},s}(t) e^{-j2\pi f \Delta\tau_s} + \\ & \frac{1}{\sqrt{A(d_{\text{NLoS},b}f)}} \sqrt{\frac{1}{2(1+\omega)}} \mathbf{h}_{b,\text{tar}}^{\text{NLoS},b}(t) e^{-j2\pi f \Delta\tau_b} \quad (1) \end{aligned}$$

式中, ω 为莱斯因子; $\Delta\tau_s = (d_{\text{NLoS},s} - d_{\text{LoS}})/c$ 和 $\Delta\tau_b = (d_{\text{NLoS},b} - d_{\text{LoS}})/c$ 分别表示海面反射路径和海底反射路径相对于直射路径的相位时延差^[31]。在公式(1)中, $\mathbf{h}_{b,\text{tar}}^{\text{LoS}}(t)$ 表示为视距分量, 其定义为 $[1, e^{j\pi \sin \psi(t)}, \dots, e^{j\pi(N-1) \sin \psi(t)}]^T$, 其中 $\psi(t)$ 为发射角。非视距信道分量 $\mathbf{h}_{b,\text{tar}}^{\text{NLoS},s}(t)$ 与 $\mathbf{h}_{b,\text{tar}}^{\text{NLoS},b}(t)$ 的演化过程采用一阶高斯-马尔可夫过程建模^[2], 即:

$$\begin{cases} \mathbf{h}_{b,\text{tar}}^{\text{NLoS},s}(t) = \rho \mathbf{h}_{b,\text{tar}}^{\text{NLoS},s}(t-1) + \sqrt{1-\rho^2} \hat{\mathbf{h}}_{b,\text{tar}}^{\text{NLoS},s}(t) \\ \mathbf{h}_{b,\text{tar}}^{\text{NLoS},b}(t) = \rho \mathbf{h}_{b,\text{tar}}^{\text{NLoS},b}(t-1) + \sqrt{1-\rho^2} \hat{\mathbf{h}}_{b,\text{tar}}^{\text{NLoS},b}(t) \end{cases} \quad (2)$$

其中, $\rho \in [0, 1]$ 表示时间相关性系数; $\hat{\mathbf{h}}_{b,\text{tar}}^{\text{NLoS},s}(t)$ 、 $\mathbf{h}_{b,\text{tar}}^{\text{NLoS},s}(t)$ 与 $\mathbf{h}_{b,\text{tar}}^{\text{NLoS},s}(t-1)$ 相互独立且同分布, 向量中各元素均服从零均值、单位方差的复高斯分布。信道 $\mathbf{h}_{b,u}(t)$ 和 $h_{\text{tar},u}(t)$ 采用相同形式的信道模型。为便于后续阐述, 定义:

$$\begin{cases} \mathbf{H}_{b,\text{tar}}(t) = \mathbf{h}_{b,\text{tar}}(t) \mathbf{h}_{b,\text{tar}}^T(t - T_{b,\text{tar}}) \in \mathbb{C}^{N \times N} \\ \mathbf{H}_{b,u}(t) = [\mathbf{h}_{b,1}(t), \dots, \mathbf{h}_{b,U}(t)] \in \mathbb{C}^{N \times U} \\ \mathbf{h}_{\text{tar},u}(t) = [h_{\text{tar},1}(t), \dots, h_{\text{tar},U}(t)]^T \in \mathbb{C}^{U \times 1} \end{cases} \quad (3)$$

(2) 环境噪声模型。水声环境噪声主要来源于四类: 湍流噪声(N_{tu})、航运噪声(N_{sh})、风浪噪声(N_{wi})以及热噪声(N_{th})。根据文献^[31], 各类环境噪声的经验功率谱密度可表示为频率 f 的函数。其中 f 的单位为 kHz, 功率谱密度的单位为 dB re $\mu\text{Pa}^2/\text{Hz}$:

$$\begin{aligned} 10\lg N_{\text{tu}}(f) &= 17 - 30\lg f / 10\lg N_{\text{sh}}(f) = 40 + 20(s - 0.5) + 26\lg f - 60\lg(f + 0.03) \\ 10\lg N_{\text{wi}}(f) &= 50 + 7.5w^{1/2} + 20\lg f - 40\lg(f + 0.4) \\ 10\lg N_{\text{th}}(f) &= -15 + 20\lg f \end{aligned} \quad (4)$$

其中, $s \in [0, 1]$ 为航运因子, w 为风速, 单位为 m/

s。因此, 环境噪声总功率为 $N_{\text{total}} = \int_{f_c - B/2}^{f_c + B/2} N_{\text{tu}}(f) +$

$$N_{\text{sh}}(f) + N_{\text{wi}}(f) + N_{\text{th}}(f) df.$$

(3) 信号模型。浮标的发射信号为通信符号与感知信号的加权叠加, 表达式为:

$$\mathbf{x}(t) = \mathbf{W}_c(t) \mathbf{v}_c(t) + \mathbf{W}_s(t) \mathbf{v}_s(t) = \mathbf{W}(t) \mathbf{v}(t) \quad (5)$$

其中, $\mathbf{v}_c(t)$ 为通信符号, 满足 $\mathbb{E}\{\mathbf{v}_c(t) \mathbf{v}_c^H(t)\} = \mathbf{I}_U$; $\mathbf{v}_s(t) \in \mathbb{C}^{N \times 1}$ 为感知信号, 满足 $\mathbb{E}\{\mathbf{v}_s(t) \mathbf{v}_s^H(t)\} = \mathbf{I}_N$; 两类信号统计独立且互不相关。 $\mathbf{W}_c(t) \in \mathbb{C}^{N \times U}$ 和 $\mathbf{W}_s(t) \in \mathbb{C}^{N \times N}$ 分别为通信与感知波束赋形矩阵。定义组合矩阵 $\mathbf{W}(t) \triangleq [\mathbf{W}_c(t), \mathbf{W}_s(t)]$, 以及 $\mathbf{v}(t) \triangleq [\mathbf{v}_c^T(t), \mathbf{v}_s^T(t)]^T$ 。假设时隙时长 Δ_t 足够长, 使得单次发射的通感一体化信号的所有多径分量均能在同一时隙内到达接收端^[13]。由于回波传播路径较长, 传播时延不可忽略, 回波可能延迟至后续时隙才到达。因此, 先前发射信号的回波可能与当前或后续时隙接收的一体化信号叠加, 形成跨时隙干扰。

(4) 感知模型。如图2所示, 浮标发射的感知信号经目标反射后返回浮标。浮标在时隙 t 接收到的目标回波信号可表示为 $\mathbf{y}_{\text{tar}}(t) = \mathbf{H}_{b,\text{tar}}(t) \mathbf{x}(t - 2T_{b,\text{tar}}) + \mathbf{n}_b(t)$, 其中 $\mathbf{n}_b(t) \sim \mathcal{CN}(0, N_{\text{total}} \mathbf{I})$ 为环境噪声^[31]。由此, 时隙 t 的目标感知信噪比可表示为:

$$\Gamma_{\text{tar}}(t) = \frac{\text{Tr}(\mathbf{W}^H(t - 2T_{b,\text{tar}}) \mathbf{H}_{b,\text{tar}}^H(t) \mathbf{H}_{b,\text{tar}}(t) \mathbf{W}(t - 2T_{b,\text{tar}}))}{N_{\text{total}}} \quad (6)$$

(5) 通信模型。如图2所示, 由于传播时延不可忽略, 先前发射的通感一体化信号所产生的回波可能在后续时隙到达传感器, 从而干扰当前信号的接收。因此, 传感器 u 在时隙 t 的接收信号表示为:

$$\begin{aligned} y_u(t) &= \mathbf{h}_{b,u}^T(t) \mathbf{x}(t - T_{b,u}) + \\ & h_{\text{tar},u}(t) \mathbf{h}_{b,u}^T(t - T_{\text{tar},u}) \mathbf{x}(t - T_{b,\text{tar}} - T_{\text{tar},u}) + n_u(t) \end{aligned} \quad (7)$$

其中, $\mathbf{x}(t - T_{b,\text{tar}} - T_{\text{tar},u})$ 为浮标在时隙 $t - T_{b,\text{tar}} - T_{\text{tar},u}$ 发射、并于时隙 $t - T_{\text{tar},u}$ 经目标反射后到达传感器 u 的通感一体化信号; $\mathbf{n}_u(t) \sim \mathcal{CN}(0, N_{\text{total}})$ 为环境噪声。因此, 传感器 u 在时隙 t 的通信信干噪比为:

$$\text{SINR}_u(t) = \frac{\left| \mathbf{h}_{b,u}^T(t) \mathbf{w}_u(t - T_{b,u}) \right|^2}{\sum_{\substack{j=1 \\ j \neq u}}^{N+U} \left| \mathbf{h}_{b,u}^T(t) \mathbf{w}_j(t - T_{b,u}) \right|^2 + P_e + N_{\text{total}}} \quad (8)$$

其中, $\mathbf{w}_u(t)$ 为矩阵 $\mathbf{W}(t)$ 的第 u 列向量; $P_e = \sum_{j=1}^{N+U} \left| h_{\text{tar},u}(t) \mathbf{h}_{b,u}^T(t - T_{\text{tar},u}) \mathbf{w}_j(t - T_{b,\text{tar}} - T_{\text{tar},u}) \right|^2$ 表示目标回波引起的跨时隙干扰。由此, 传感器 u 在时隙 t 的通信速率为: $R_u(t) = \log_2(1 + \text{SINR}_u(t))$ 。

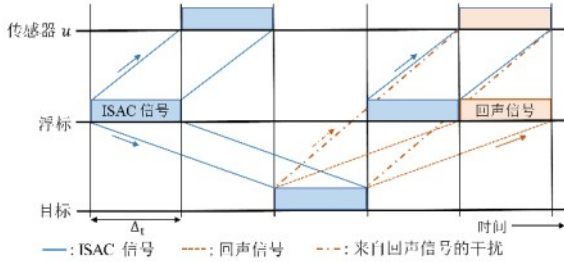


图2 浮标、目标和传感器 u 之间的传输模型

本文通过对通信波束赋形矩阵 $\mathbf{W}_c(t)$ 与感知波束赋形矩阵 $\mathbf{W}_s(t)$ 进行联合优化, 实现两个优化目标: (i) 最大化通信总速率; (ii) 最大化感知信噪比。对应的优化问题数学表述如下:

$$\begin{aligned} \max_{\mathbf{w}_c(t), \mathbf{w}_s(t)} & \left(\sum_{t=1}^T \sum_{u=1}^U R_u(t + T_{b,u}), \sum_{t=1}^T \Gamma_{\text{tar}}(t + 2T_{b,\text{tar}}) \right) \\ \text{s.t. } C_1: & \text{Tr}(\mathbf{W}_c^H(t) \mathbf{W}_c(t)) + \text{Tr}(\mathbf{W}_s^H(t) \mathbf{W}_s(t)) \\ & \leq P_{\max}, \forall t \in \mathcal{T} \end{aligned} \quad (9)$$

式中, $P_{\max}$ 为浮标节点的最大发射功率, 约束 C_1 用于将浮标总发射功率限制在最大功率 $P_{\max}$ 以内。

上述优化问题具有以下三个特征。第一, 问题具有非凸性。通信总速率和感知信噪比与通信波束赋形矩阵 $\mathbf{W}_c(t)$ 和感知波束赋形矩阵 $\mathbf{W}_s(t)$ 呈非线性耦合关系。此外, 由公式(2)可知, 水声信道具有时间相关性, 因此问题(9)本质上是一个长期时序优化问题。第二, 浮标只能基于过时的信道状态信息进行决策。通常情况下, 浮标在时隙 t 仅能根据当前可观测的信道矩阵 $\mathbf{H}_{b,u}(t)$ 和 $\mathbf{H}_{b,\text{tar}}(t)$ 来设计赋形矩阵 $\mathbf{W}_c(t)$ 和 $\mathbf{W}_s(t)$ 。然而, 受水声长传播时延和信道时变性的影响, 传感器 u 实际接收信号时

所经历的有效信道为 $\mathbf{h}_{b,u}(t + T_{b,u})$, 而浮标接收目标回波时的有效信道为 $\mathbf{H}_{b,\text{tar}}(t + 2T_{b,\text{tar}})$ 。因此, 基于当前信道状态生成的波束赋形矩阵难以与信号实际传输时的信道状态实现准确匹配。第三, 通信与感知优化目标之间存在固有权衡。在总发射功率约束下, 为通信波束赋形矩阵 $\mathbf{W}_c(t)$ 分配更多功率虽有助于提高通信总速率, 但会相应减少感知波束赋形矩阵 $\mathbf{W}_s(t)$ 的可用功率, 从而降低感知信噪比。因此, 问题(9)是一个多目标优化问题。

2 DeepMTE算法

传统基于数学优化的方法依赖于实时信道状态信息, 在该场景下难以获得有效解。此外, 由于水声信道具有时间相关性, 传统优化方法既难以利用历史信道信息, 也未考虑当前决策对后续系统性能的长期影响。若浮标能够学习信道随时间的演化规律以及不同时隙间的时序相关性, 则有望确定合适的 $\mathbf{W}_c(t)$ 和 $\mathbf{W}_s(t)$ 。为此, 本文首先将多目标波束赋形优化问题建模为多目标马尔可夫决策过程, 引入 DRL 算法, 通过智能体与环境的持续交互来优化长期累积收益, 从而求解此类时序决策问题。

然而, 传统深度强化学习方法通常面向固定的单一优化目标, 在系统优化目标偏好发生变化时泛化能力有限。因此, 本文提出一种基于多任务 DRL 与 MoE 模型的波束赋形优化方法 DeepMTE, 并设计多步跨任务经验回放机制, 旨在最大发射功率约束下实现通信与感知性能的权衡, 同时提升算法在不同目标偏好因子下的泛化能力。

2.1 多目标马尔可夫决策过程

在本文建模中, 目标、水下传感器以及无线信道共同构成环境, 搭载 DeepMTE 算法的浮标被视为智能体。形式上, 多目标马尔可夫决策过程包含 5 个要素: 状态 $\mathbf{S}(t)$ 、动作 $\mathbf{A}(t)$ 、状态转移概率 $p(\mathbf{S}(t+1)|\mathbf{S}(t), \mathbf{A}(t))$ 、奖励向量 $\mathbf{r}(t+1)$ 以及用于权衡通信和感知性能的目标偏好因子 $b \in [0, 1]$ 。与单目标马尔可夫决策过程不同, 多目标马尔可夫决策过程的奖励以向量形式表示为 $\mathbf{r}(t+1) = [r_c(t+1), r_s(t+1)]^T$, 其中每个分量分别对应一个独立优化目标^[25]。通过设置不同的目标偏好因子, 多目标深度强化学习可将奖励向量 $\mathbf{r}$ 聚合为单一标量奖励, 从而生成能够协调多目标之间权衡关系的决策策略。下文将针对多目标水声通

感波束赋形优化问题, 详细完成多目标马尔可夫决策过程的建模。

(1) 系统状态需要包含足够的信息, 以便智能体做出合理的通感一体化波束赋形决策。智能体基于估计的信道状态信息 $\mathbf{H}_{b,u}(t)$ 和 $\mathbf{H}_{b,tar}(t)$ 来设计 $\mathbf{W}_c(t)$ 和 $\mathbf{W}_s(t)$, 以尽可能匹配信号实际传输时对应的信道状态 $\mathbf{h}_{b,u}(t + T_{b,u})$ 与 $\mathbf{H}_{b,tar}(t + 2T_{b,tar})$ 。虽然信道估计值无法精确反映未来真实信道, 但能够为波束赋形设计提供有效参考。因此, 时隙 t 的可观测状态 $\mathbf{S}(t)$ 定义为: $\mathbf{S}(t) = [\mathbf{H}_{b,u}(t), \mathbf{H}_{b,tar}(t)]$ 。

(2) 智能体动作与系统优化变量相关联。根据式(9), 智能体在时隙 t 的动作表示为 $\mathbf{A}(t) = [\mathbf{W}_c(t), \mathbf{W}_s(t)] \in \mathbb{C}^{N \times (U+N)}$, 所有分量均为待优化连续变量。2.3 节将设计一种基于缩放动作选取的策略, 详细阐述如何利用神经网络生成通信与感知波束赋形矩阵。

(3) 状态转移概率记作 $p(\mathbf{S}(t+1)|\mathbf{S}(t), \mathbf{A}(t))$, 表示智能体在状态 $\mathbf{S}(t)$ 下执行动作 $\mathbf{A}(t)$ 后, 系统转移至状态 $\mathbf{S}(t+1)$ 的概率^[25]。由于智能体在时隙 t 无法提前获取未来真实信道 $\mathbf{h}_{b,u}(t + T_{b,u})$ 与 $\mathbf{H}_{b,tar}(t + 2T_{b,tar})$, 故其对环境仅具备部分可观测性, 因此问题(9)属于部分可观测马尔可夫决策过程。

(4) 奖励函数 $\mathbf{r}(t+1)$ 用于评价智能体在状态 $\mathbf{S}(t)$ 下执行动作 $\mathbf{A}(t)$ 的效果。本文旨在使智能体学习有效的通感联合波束赋形策略, 同时最大化通信总速率与感知信噪比。鉴于通信与感知目标之间存在固有权衡, 我们定义奖励向量 $\mathbf{r}(t+1)$ 来评估动作 $\mathbf{A}(t)$ 的性能, 即:

$$\begin{aligned} \mathbf{r}(t+1) &= [r_c(t+1), r_s(t+1)]^T \\ &= \left[\sum_{u=1}^U R_u(t + T_{b,u}), \Gamma_{tar}(t + 2T_{b,tar}) \right]^T \end{aligned} \quad (10)$$

其中, 奖励分量 $r_c(t+1)$ 与 $r_s(t+1)$ 分别评价动作 $\mathbf{A}(t)$ 在通信和感知两方面带来的性能; 由于 $R_u(t + T_{b,u})$ 与 $\Gamma_{tar}(t + 2T_{b,tar})$ 处于同一数量级^[32], 因此在计算中仅保留纯数值, 省略其量纲。

(5) 目标偏好因子 b 。在任意时隙, 给定状态 $\mathbf{S}(t)$, 智能体依据策略 $\pi(\mathbf{A}(t)|\mathbf{S}(t))$ 选取动作 $\mathbf{A}(t)$ 。为求解最优波束赋形策略 π^* , 智能体最大化累积奖励, 表达式如下:

$$\begin{aligned} G(t) &= \sum_{l=t}^{\infty} \gamma^{l-t} \mathbf{b}^T \mathbf{r}(l+1) \\ &= \sum_{l=t}^{\infty} \gamma^{l-t} [b, 1-b] \mathbf{r}(l+1) \end{aligned} \quad (11)$$

式中, γ 为折扣因子, $\mathbf{b}$ 为目标偏好向量。状态-动作对 $(\mathbf{S}(t), \mathbf{A}(t))$ 对应的动作价值函数 (即 Q 值) 由策略 π 与状态转移概率共同决定, 定义为:

$$Q_{\pi}(\mathbf{S}(t), \mathbf{A}(t)) = \mathbb{E}_{\pi}[G(t)|\mathbf{S}(t), \mathbf{A}(t)] \quad (12)$$

至此, 波束赋形优化问题已被建模为多目标马尔可夫决策过程。然而, 由于智能体在时隙 t 无法预先获知未来信道模型 $\mathbf{h}_{b,u}(t + T_{b,u})$ 与 $\mathbf{H}_{b,tar}(t + 2T_{b,tar})$, 状态转移概率 $p(\mathbf{S}|\mathbf{S}, \mathbf{A})$ 难以获得。这一难点促使本文采用无模型 DRL 方案进行求解。

在上述多目标马尔可夫决策过程框架下, 智能体的决策动作 $\mathbf{A}(t) = [\mathbf{W}_c(t), \mathbf{W}_s(t)]$ 处于连续动作空间。深度确定性策略梯度算法^[27]等传统基于策略的深度强化学习算法虽能有效处理连续控制任务, 但仅局限于单目标优化, 难以泛化至不同的优化目标偏好。为克服上述缺陷, 本文提出一种嵌入 MoE 模型的新型多任务深度强化学习框架 DeepMTE, 用于求解所建立的多目标马尔可夫决策过程问题。该算法的核心思路是: 将每一组目标偏好因子对应的优化问题单独作为一项任务, 并通过 MoE 架构对全部任务进行联合训练^[26]。借助该机制, DeepMTE 能够针对训练过程中未曾出现的目标偏好因子, 快速求解出性能优异的波束赋形方案。

具体地, 训练过程共设定 M 个任务, 任务索引 $m \in \mathcal{M} = \{1, 2, \dots, M\}$ 对应偏好向量为 $\mathbf{b}_m$ 的优化问题。任务 m 对应的 DRL 状态、动作、奖励分别记作 $\mathbf{S}_m(t)$ 、 $\mathbf{A}_m(t)$ 、 $\mathbf{r}_m(t+1)$ 。图 3 展示了 DeepMTE 完整框架, 主要包含三个组成部分: (i) 用于生成通信波束 $\mathbf{W}_{c,m}(t)$ 和感知波束 $\mathbf{W}_{s,m}(t)$ 的多任务深度确定性策略梯度算法框架; (ii) 用于对奖励向量 $\mathbf{r}(t+1)$ 进行处理的目标偏好向量模块; (iii) 用于提升跨不同任务泛化能力的 MoE 架构。

2.2 多任务 MoE 架构

DeepMTE 算法使用两类神经网络, 分别命名为 MoE 动作网络与 MoE 评价网络。MoE 动作网络包含两个子网络: 由参数 Θ_a 定义的 MoE 主动作网络和由参数 Θ_a^* 定义的 MoE 目标动作网络。MoE 评

价网络由参数为 Θ_c 的MoE主评价网络和参数为 Θ_a 的MoE目标评价网络组成。如图3所示,所有神经网络均采用相同的结构,即包含1个共享门控网络、 L 个专家网络和 M 个输出层,但各自承担不同功能。

MoE主动作网络基于状态 $\mathbf{S}_m(t)$ 和目标偏好向量 $\mathbf{b}_m$ 生成临时通感一体化波束赋形决策 $\pi_{a,m}(\mathbf{S}_m(t), \mathbf{b}_m; \Theta_a)$ 。在第2.3节中,一种新的动作选择策略将用于确定最终实际执行的波束赋形决策 $\mathbf{A}_m(t)$ 。MoE主评价网络用于计算动作价值,即Q值 $Q(\mathbf{S}_m(t), \mathbf{A}_m(t), \mathbf{b}_m; \Theta_c)$,并通过贝尔曼方程对主动作网络的性能进行评估^[33]。MoE目标动作网络与MoE目标评价网络分别作为对应主网络的缓慢更新副本,在训练阶段用于计算目标动作价值 $Q(\mathbf{S}_m(t), \mathbf{A}_m(t), \mathbf{b}_m; \Theta_c)$,以指导 Θ_a 和 Θ_c 的更新。

在该神经网络框架中,共享门控网络根据输入数据确定分配给各专家网络的权重,各专家网络从 M 项任务对应的系统状态 $\mathbf{S}_m(t)$ 中学习不同的特征,输出层聚合所有专家网络的知识,生成临时波束赋形决策。下文以MoE主动作网络为例,对其内部组成模块展开详细说明。

(1) 共享门控网络。共享门控网络由输入层、全连接层和Softmax层组成^[25]。给定目标偏好向量 $\mathbf{b}_m \in \mathbb{R}^{2 \times 1}$,输入层将其送入全连接层进行处理。全连接层输出一个 L 维向量,表达式为: $\mathbf{c}_{g,f} = \Theta_{a,g}^T \mathbf{b}_m$,其中, $\Theta_{a,g} \in \mathbb{R}^{2 \times L}$ 为全连接层参数。向量

$\mathbf{c}_{g,f}$ 的第 l 个元素 $c_{g,f}^l$ 表示分配给专家网络 l 的原始权重,其中 $l \in \mathcal{L} = \{1, 2, \dots, L\}$ 。为保证所有专家权重之和等于1,通过Softmax层对 $\mathbf{c}_{g,f}$ 进行归一化,得到门控输出 $\mathbf{c}_{g,s} = \text{Softmax}(\mathbf{c}_{g,f})$,其第 l 个元素计算如下:

$$c_{g,s}^l = \frac{\exp(c_{g,f}^l)}{\sum_{i=1}^L \exp(c_{g,f}^i)} \quad (13)$$

通过在所有任务间共用同一门控网络,DeepMTE能够对训练过程中未曾出现的目标偏好因子保持泛化能力。

(2) 专家网络。将每个目标偏好因子对应的优化问题视为一项独立任务,原多目标优化问题(9)可转化为多任务学习问题。通常,单个专家网络(例如单纯的多任务网络)会因参数之间的相互干扰而导致性能下降^[26]。因此,DeepMTE采用多个专家网络分别学习输入状态的不同特征。例如,部分专家网络侧重于学习能够提升下行通信总速率的通信策略,另一些专家网络则侧重于学习能够提高感知信噪比的感知策略。此外,通过共享门控网络,DeepMTE能够根据具体任务动态组合相关专家,从而实现自适应且具有针对性的输出。所有专家网络均采用相同的结构,由输入层、门控循环单元层^[25]和全连接层组成。给定输入状态 $\mathbf{S}_m(t)$,输入层将其送入门控循环单元层,由门控循环单元层从时序数据中提取时间依赖关系,随后将提取的特

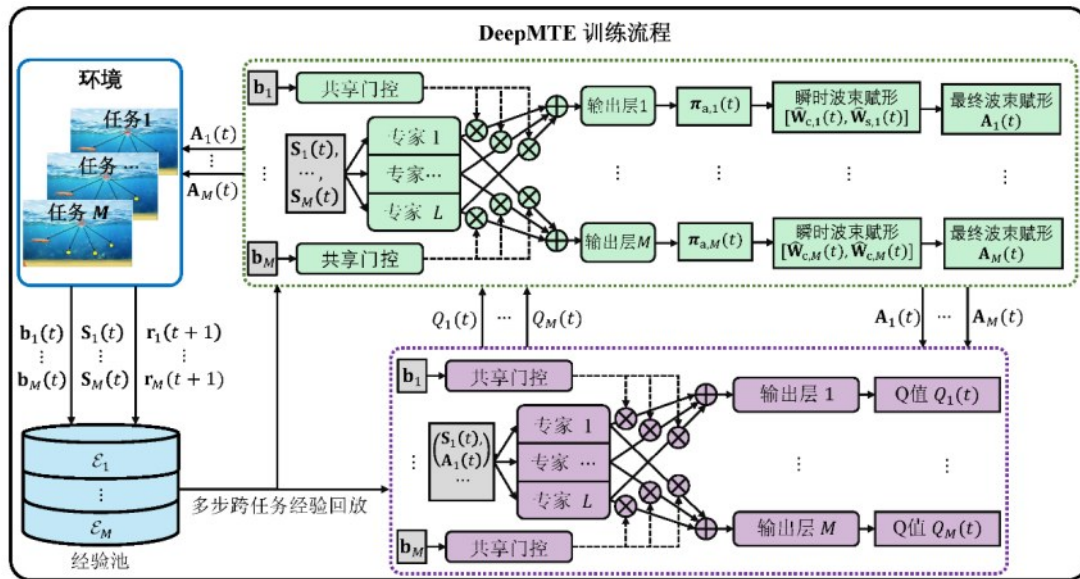


图3 DeepMTE 框架

征送入全连接层做进一步处理。由于神经网络无法直接处理复数矩阵,本文首先提取 $\mathbf{S}_m(t)$ 中每个元素的实部和虚部,将其展平后拼接为实值向量 $\hat{\mathbf{S}}_m(t) \in \mathbb{R}^{2(NU+N^2) \times 1}$,随后输入专家网络进行处理。设 $\Theta_{a,c,l} \in \mathbb{R}^{2(NU+N^2) \times N_c}$ 为第 l 个专家网络的参数,其中 N_c 为全连接层的神经元数量。全连接层的输出表达式为:

$$\mathbf{c}_{c,l} = \Theta_{a,c,l}^T \cdot \hat{\mathbf{S}}_m(t) \quad (14)$$

(3) 输出层。结合共享门控网络与所有专家网络的输出,任务 $m \in \mathcal{M}$ 对应的输出层输入可表示为 $\mathbf{c}_{o,l} = \sum_{l=1}^L \mathbf{c}_{g,s}^l \mathbf{c}_{c,l}$ 。通常,通信波束赋形矩阵 $\mathbf{W}_c \in \mathbb{C}^{N \times U}$ 和感知波束赋形矩阵 $\mathbf{W}_s \in \mathbb{C}^{N \times N}$ 无法直接生成,因为神经网络无法输出复数值。因此,任务 m 的输出层首先生成实值波束赋形决策 $\pi_{a,m}$:

$$\pi_{a,m} = \Theta_{a,o,m}^T \mathbf{c}_{o,l} \quad (15)$$

式中, $\Theta_{a,o,m} \in \mathbb{R}^{N_c \times 2(NU+N^2)}$ 为任务 m 对应输出层的参数。为简化表述,令 $\Theta_{a,c} = [\Theta_{a,c,1}, \Theta_{a,c,2}, \dots, \Theta_{a,c,L}]$ 。MoE主动动作网络的总参数 Θ_a 由共享门控网络、全部专家网络以及所有输出层的参数共同构成,即: $\Theta_a = \{\Theta_{a,g}, \Theta_{a,c}, \Theta_{a,o,1}, \dots, \Theta_{a,o,M}\}$ 。

2.3 缩放动作选取策略

根据式(9),搭载DeepMTE的浮标通过优化通信与感知波束赋形矩阵,在满足最大发射功率约束的条件下,最大化通信总速率与感知信噪比。为严格满足功率约束,本文采用缩放动作选取策略^[34]。

在时隙 t ,对于 $\forall m \in \mathcal{M}$,MoE主动动作网络基于状态 $\mathbf{S}_m(t)$ 与目标偏好向量 $\mathbf{b}_m$ 输出临时通感波束赋形决策 $\pi_{a,m}(\mathbf{S}_m(t), \mathbf{b}_m; \Theta_{a,g}, \Theta_{a,c}, \Theta_{a,o,m})$ 。由于神经网络无法直接输出复数矩阵,本文将该向量 $\pi_{a,m}(\mathbf{S}_m(t), \mathbf{b}_m; \Theta_{a,g}, \Theta_{a,c}, \Theta_{a,o,m})$ 划分为两段实值分量:第一段维度为 $2NU$,用于编码通信波束赋形信息;第二段维度为 $2N^2$,用于编码感知波束赋形信息。进一步地,通信波束赋形矩阵的构造规则:取第一段中的前 NU 个元素作为复数幅值,后 NU 个元素作为复数相位,并通过逐元素组合构成复数矩阵;感知波束赋形矩阵则基于第二段采用相同的构造规则。由此,可分别得到瞬时通信波束赋形矩阵 $\tilde{\mathbf{W}}_{c,m}(t) \in \mathbb{C}^{N \times U}$ 和瞬时感知波束赋形矩阵 $\tilde{\mathbf{W}}_{s,m}(t) \in \mathbb{C}^{N \times N}$ 。

为进一步提升DeepMTE的探索能力,本文在

生成动作时添加截断高斯噪声以构造平滑动作,表示如下: $[\hat{\mathbf{W}}_{c,m}(t), \hat{\mathbf{W}}_{s,m}(t)] \leftarrow [\tilde{\mathbf{W}}_{c,m}(t), \tilde{\mathbf{W}}_{s,m}(t)] + \epsilon_a$,式中, $\epsilon_a \in \mathbb{C}^{N \times (U+N)}$ 的每个元素独立服从截断高斯分布 $\text{clip}(\mathcal{N}(0, \sigma_a^2), -c_a, c_a)$, σ_a^2 为噪声方差, c_a 为截断阈值。当智能体学习到有效策略后,不再需要过度探索^[25],因此引入衰减系数 $\iota \in (0,1)$,使噪声方差随时隙逐步递减。设初始方差为 $\sigma_{a,\text{init}}^2$,则时隙 t 对应的噪声方差为 $\sigma_a^2(t) = \sigma_{a,\text{init}}^2 \iota^t$ 。为满足给定的最大传输功率约束,定义缩放系数为 $\zeta = P_{\max} / (\text{Tr}(\hat{\mathbf{W}}_{c,m}^H(t) \hat{\mathbf{W}}_{c,m}(t)) + \text{Tr}(\hat{\mathbf{W}}_{s,m}^H(t) \hat{\mathbf{W}}_{s,m}(t)))$,最终实际执行的波束赋形矩阵由下式确定:

$$\mathbf{A}_m(t) = \begin{cases} [\sqrt{\zeta} \hat{\mathbf{W}}_{c,m}(t), \sqrt{\zeta} \hat{\mathbf{W}}_{s,m}(t)] & \text{若 } \zeta < 1 \\ [\hat{\mathbf{W}}_{c,m}(t), \hat{\mathbf{W}}_{s,m}(t)] & \text{其他} \end{cases} \quad (16)$$

2.4 多步跨任务经验回放机制

由前述SINR表达式可知,DeepMTE做出的波束赋形决策 $\mathbf{A}(t) = [\mathbf{W}_c(t), \mathbf{W}_s(t)]$ 会影响未来时隙 $t + T_{b,\text{tar}} + T_{\text{tar},u}$ 各传感器的信干噪比。也就是说,时隙 t 执行的动作 $\mathbf{A}(t)$ 会影响后续多个时隙的奖励。

然而传统经验回放机制^[28]仅利用单步即时奖励训练深度神经网络,导致DeepMTE难以学习到有效的波束赋形策略。此外,这类传统机制在单任务场景下效果良好,但无法直接适配DeepMTE框架下的多任务学习场景。为解决该问题,本文为DeepMTE设计一种全新的多步跨任务经验回放机制。其核心思路为:(i)采用多步学习公式计算损失函数;(ii)为每个任务单独设置经验池,同时利用多个任务的样本联合训练MoE架构。具体过程如下。

(1) 经验存储。为存储并复用历史交互经验,为每个任务 $m \in \mathcal{M}$ 分配一个容量为 E 的先进先出经验池 $\mathcal{E}_m$ 。在每个时隙中,智能体计算出任务 m 对应的奖励 $\mathbf{r}_m(t+1)$ 后,以五元组 $(\mathbf{S}_m(t), \mathbf{A}_m(t), \mathbf{r}_m(t+1), \mathbf{S}_m(t+1), \mathbf{b}_m)$ 的形式采集单条经验 $e_m(t)$,并将其存入经验池 $\mathcal{E}_m$ 。由于水下传播时延较长,智能体在时隙 t 所执行动作对应的奖励向量 $\mathbf{r}(t+1)$,需等到时隙 $t + T_{b,u}$ 获得通信速率 $R_u(t + T_{b,u})$,以及时隙 $t + 2T_{b,\text{tar}}$ 获得感知信噪比 $\Gamma_{\text{tar}}(t + 2T_{b,\text{tar}})$ 之后,才能完整计算得出。因此,对应的经验样本无法即时生成并用于网络训练,这与无线场景中奖励几乎可实时获得、经验可立即采

集的特点形成显著差异。

(2) 多步学习。为评估当前波束赋形决策对后续多个时隙系统性能的长期影响, 本文在 Deep-MTE 算法中引入多步学习方法。其核心思想是结合即时奖励 $\mathbf{r}_m(t+1)$ 与后续连续多步奖励 $\{\mathbf{r}_m(t+2), \mathbf{r}_m(t+3), \dots, \mathbf{r}_m(t+K)\}$ 共同评价动作 $\mathbf{A}_m(t)$ 的质量, 其中 K 为多步学习步数。具体而言, 每一轮训练从经验池 $\mathcal{E}_m$ 中采样 N_{train} 条经验构成小批量样本 $\mathcal{B}_m$ 。对每条采样经验 $e_m(i)$, 从 $\mathcal{E}_m$ 中调取其后续 $K-1$ 条经验 $\{e_m(i+1), e_m(i+2), \dots, e_m(i+K-1)\}$, 并依据式 (17) 计算多步目标值 $z_m(i+K)$; 其中 $\mathbf{A}_m(i+K)$ 由 MoE 目标动作网络根据状态 $\mathbf{S}_m(i+K)$ 输出。因此, 在任务 m 下, 用于训练 MoE 主评价网络的损失函数计算如下:

$$L_m(\Theta_{c,g}, \Theta_{c,e}, \Theta_{c,o,m}) = \frac{1}{N_{\text{train}}} \sum_{e_m(i) \in \mathcal{B}_m} (z_m(i+K) - Q(\mathbf{S}_m(i), \mathbf{A}_m(i), \mathbf{b}_m; \Theta_{c,g}, \Theta_{c,e}, \Theta_{c,o,m}))^2 \quad (18)$$

式中, $\Theta_{c,g}$ 、 $\Theta_{c,e}$ 、 $\Theta_{c,o,m}$ 分别为 MoE 主评价网络共享门控网络、专家网络以及任务 m 专属输出层的参数; $\Theta_{c,g}^-$ 、 $\Theta_{c,e}^-$ 、 $\Theta_{c,o,m}^-$ 分别表示 MoE 目标评价网络中对应部分的参数。同理, 任务 m 下 MoE 主动作

$$z_m(i+K) = \begin{cases} \sum_{k=1}^K \gamma^{k-1} \mathbf{b}_m^T \mathbf{r}_m(i+k) + \gamma^K Q(\mathbf{S}_m(i+K), \mathbf{A}_m(i+K), \mathbf{b}_m; \Theta_{c,g}^-, \Theta_{c,e}^-, \Theta_{c,o,m}^-) & \text{若 } i \leq T-K+1 \\ \sum_{k=1}^{T-i+1} \gamma^{k-1} \mathbf{b}_m^T \mathbf{r}_m(i+k) & \text{其他} \end{cases} \quad (17)$$

此外, 采用随机梯度下降算法^[25]对 MoE 主评价网络参数 $\{\Theta_{c,g}, \Theta_{c,e}, \Theta_{c,o,1}, \Theta_{c,o,2}, \dots, \Theta_{c,o,M}\}$ 执行更新:

$$\Theta_{c,i} \leftarrow \Theta_{c,i} - \alpha_c \nabla_{\Theta_{c,i}} L(\Theta_{c,g}, \Theta_{c,e}, \Theta_{c,o,1}, \Theta_{c,o,2}, \dots, \Theta_{c,o,M}), \quad (22)$$

其中, $i \in \{g, e, \{o, 1\}, \{o, 2\}, \dots, \{o, M\}\}$, α_c 为 MoE 主评价网络的学习率。同理, MoE 主动作网络参数 $\{\Theta_{a,g}, \Theta_{a,e}, \Theta_{a,o,1}, \Theta_{a,o,2}, \dots, \Theta_{a,o,M}\}$ 更新规则如下:

$$\Theta_{a,i} \leftarrow \Theta_{a,i} - \alpha_a \nabla_{\Theta_{a,i}} L(\Theta_{a,g}, \Theta_{a,e}, \Theta_{a,o,1}, \Theta_{a,o,2}, \dots, \Theta_{a,o,M}), \quad (23)$$

式中, α_a 为 MoE 主动作网络的学习率。由此可见, 共享门控网络与全部专家网络作为底层共享模块, 利用从所有任务中采集的经验进行联合训练; 每个

网络的训练损失函数为:

$$L_m(\Theta_{a,g}, \Theta_{a,e}, \Theta_{a,o,m}) = \frac{1}{N_{\text{train}}} \sum_{e_m(i) \in \mathcal{B}_m} (\nabla_{\mathbf{A}_m(i)} Q(\mathbf{S}_m(t), \mathbf{A}_m(t), \mathbf{b}_m; \Theta_{c,g}, \Theta_{c,e}, \Theta_{c,o,m}) \times \nabla_{(\Theta_{a,g}, \Theta_{a,e}, \Theta_{a,o,m})} \pi_{a,m}(\mathbf{S}_m(t), \mathbf{b}_m; \Theta_{a,g}, \Theta_{a,e}, \Theta_{a,o,m})) \quad (19)$$

(3) 跨任务损失函数。在 DeepMTE 框架中, 各专家网络都需要在不同目标偏好向量 $\mathbf{b}$ 下保持有效的特征学习能力。若仅使用有限的偏好向量 $\mathbf{b}$ 进行训练, 会限制专家网络实现专业化。类似地, 共享门控网络负责建立偏好向量 $\mathbf{b}$ 到各专家权重的映射关系; 若仅基于少量目标偏好向量对应的经验样本进行训练, 门控网络只能学习到有限偏好下的专家权重组合, 难以泛化至新的目标偏好向量。

为解决该问题, 在神经网络训练阶段, 从 M 个任务的经验池中分别采样 N_{train} 条经验进行联合损失计算。MoE 主评价网络的总体损失函数为:

$$L(\Theta_{c,g}, \Theta_{c,e}, \Theta_{c,o,1}, \Theta_{c,o,2}, \dots, \Theta_{c,o,M}) = \frac{1}{M} \sum_{m=1}^M L_m(\Theta_{c,g}, \Theta_{c,e}, \Theta_{c,o,m}) \quad (20)$$

对应地, MoE 主动作网络的总体损失函数为:

$$L(\Theta_{a,g}, \Theta_{a,e}, \Theta_{a,o,1}, \Theta_{a,o,2}, \dots, \Theta_{a,o,M}) = \frac{1}{M} \sum_{m=1}^M L_m(\Theta_{a,g}, \Theta_{a,e}, \Theta_{a,o,m}) \quad (21)$$

输出层仅使用对应单一任务的数据完成训练, 保证各任务均可输出独立且精准的预测结果。

(4) 延迟策略更新。在经典深度确定性策略梯度算法^[27]中, MoE 主动作网络与 MoE 主评价网络通常执行同步更新, 但 MoE 主评价网络的估计误差容易导致算法收敛至次优解。为克服这一缺陷, DeepMTE 采用解耦式更新策略^[33]: MoE 主评价网络每更新 F_c 次, 才对 MoE 主动作网络执行一次更新。同时, 降低 MoE 目标动作网络与 MoE 目标评价网络的更新频率以提升训练稳定性, 通常使其与 MoE 主动作网络的更新频率保持一致。具体通过软更新方法^[33]对两个目标网络参数进行迭代修正:

$$\begin{cases} \Theta_{c,i}^- \leftarrow \chi_c \Theta_{c,i} + (1 - \chi_c) \Theta_{c,i}^- \\ \Theta_{a,i}^- \leftarrow \chi_a \Theta_{a,i} + (1 - \chi_a) \Theta_{a,i}^- \end{cases} \quad (24)$$

其中, $i \in \{g, e, \{o, 1\}, \dots, \{o, M\}\}$; $\chi_a \in [0, 1]$ 和 $\chi_c \in [0, 1]$ 分别为 MoE 目标动作网络和 MoE 目标评价网络的软更新系数。

2.5 算法执行流程

DeepMTE 采用离线训练-在线执行的部署架构, 其在实际系统中的具体实施步骤如下。

(1) 离线训练。为明确多任务离线训练所覆盖的任务范围, 定义目标偏好因子 b 的取值空间为连续区间 $[0, 1]$ 。由于该参数空间为连续无限空间, 有限次离线训练无法实现对所有可能取值的穷举遍历。为此, 在离线训练阶段, 我们从 $[0, 1]$ 中随机采样 M 个目标偏好因子, 构成训练任务集合, 每个取值对应一项独立的多目标波束赋形任务。为验证模型对未见目标偏好的泛化能力, 在线测试阶段所采用的目标偏好均不包含在离线训练任务集合中。在此阶段, 智能体在基于实测数据校准的数学模型所搭建的仿真环境中进行交互, 以保证仿真环境的真实性。该仿真系统集成了多径传播、环境噪声以及信号响应等模型。DeepMTE 的核心步骤如下: (i) 初始化 M 个并行仿真环境; (ii) 根据第 m 个环境的状态 $\mathbf{S}_m(t)$ 与目标偏好向量 $\mathbf{b}_m$, 生成临时波束赋形决策 $\pi_{a,m}$; (iii) 通过缩放动作选择策略确定最终波束赋形决策 $\mathbf{A}_m(t)$, 并与对应仿真环境交互; (iv) 根据式(10)计算奖励 $\mathbf{r}_m(t+1)$, 获取下一时刻状态 $\mathbf{S}_m(t+1)$; (v) 构造经验样本 $e_m(t) = (\mathbf{S}_m(t), \mathbf{A}_m(t), \mathbf{r}_m(t+1), \mathbf{S}_m(t+1))$, 并存入对应任务的经验池 $\mathcal{E}_m$; (vi) 从 $\mathcal{E}_m$ 中随机采样 N_{train} 条经验构建小批量样本 $\mathcal{B}_m$; (vii) 计算每条采样经验 $e_m(i)$ 的目标值; (viii) 分别依据式(22)和式(23)更新 MoE 主评价网络与 MoE 主动作网络; (ix) 通过式(24)对 MoE 目标动作网络和 MoE 目标评价网络的参数执行软更新。当神经网络收敛后, DeepMTE 模型可部署至实际系统中在线执行。

(2) 在线执行。在部署阶段, 智能体可能面临训练阶段未曾出现的目标偏好向量 $\mathbf{b}$ 。对于这些未知偏好, DeepMTE 借助 MoE 架构的泛化能力, 将离线阶段从多个训练任务中学习到的跨任务知识迁移至新任务, 同时结合少量在线微调, 实现对未见偏好的快速适配。因此, 本文方案的泛化性并不依赖于对偏好参数空间的穷举覆盖, 而是依托在有限离线任务中习得的共享知识结构, 有效支撑模型在面对新偏好时的快速响应与适应能力。然而, 尽管

经过充分训练的多任务 MoE 架构, 尤其是共享门控网络与专家网络, 对集合 $\{\mathbf{b}_m | m \in \mathcal{M}\}$ 具备较强的泛化能力, 但每个输出层最初仅针对某一特定任务 m 进行优化。为适配全新的偏好向量 $\mathbf{b}$, 移除所有预训练输出层, 并新增一个专属于该新任务的输出层。由于输出层仅包含少量参数^[32], 部署后的 DeepMTE 可以快速完成微调, 仅需少量额外训练即可获得适用于新任务的有效波束赋形策略。

3 仿真分析

本节基于 Python 3.6 仿真平台对 DeepMTE 算法进行性能测试。所研究的系统包含 1 个位于坐标 (0,0,0) 米处的浮标、1 个位于坐标 (800,600,300) 米处的目标以及 $U=3$ 个随机分布在 $[600,960] \times [300,660] \times [100,300]$ 米空间范围内的传感器。关键参数设置如下: 莱斯因子 $\omega = 6$, 时间相关系数 $\rho = 0.9$, 最大发射功率 $P_{\text{max}} = 200\text{dB re } 1\mu\text{Pa } 1\text{m}$, 载波频率 $f = 3\text{kHz}$ 。传播衰减系数 $k = 1.5$, 航运噪声系数 $s = 0.5$, 风速 $w = 10\text{m/s}$, 海水密度 1020kg/m^3 , 水深 300 米。天线默认数量 $N = 6$ 。DeepMTE 算法的超参数汇总于表 1。多任务训练设置任务数量 $M = 20$, 每个任务的目标偏好因子 b 在区间 $[0, 1]$ 内随机采样得到。若无特殊说明, 在线执行阶段所用的 b 设定为 0.6。

本次仿真对比以下几种方案。(i)DeepMTE: 本文所提基于 MoE 的多任务 DRL 的多目标水声 ISAC 方案; (ii)DeepMTE-w: 该版本移除多任务 MoE 架构, 智能体从零开始训练, 同时将多步跨任务经验回放简化为多步经验回放, 用于验证多任务 MoE 架构的有效性; (iii)DeepMTE-c: 该方案采用传统经验回放机制^[27]替换本文设计的多步跨任务经验回放机制, 用于验证其有效性; (iv)DeepMTE-o: DeepMTE 的简化版本, 去掉在线微调阶段, 用以验证在线微调环节的作用; (v)DeepMTE-m: 该方案采用传统多任务架构^[35]替换本文设计的 MoE 架构, 用于验证其有效性; (vi)AC: 采用演员-评论家 (actor-critic, AC) 算法^[25]求解优化问题 (9), 智能体仅使用单步即时经验完成训练, 不使用历史经验; (vii)FPM: 采用分式规划与主化-最小化联合优化方法^[30]求解优化问题 (9), 该方案依靠实时信道状态信息完成波束赋形决策; (viii)SCA: 采用逐次凸近似方法^[36]求解优化问题 (9),

该方案依靠实时信道状态信息完成波束赋形决策；(ix)**预测型波束赋形**：采用深度学习^[37]求解优化问题(9)，该方案依赖先验的波束赋形作为训练标签。

表 1 算法参数

仿真参数	参数值
专家网络数量 L	5
网络学习率 α_a	$1e-4$
网络学习率 α_c	$2e-4$
经验回放缓存容量 E	$1e3$
回报折扣因子 γ	0.95
多步时序差分步长 K	15
单网络神经元数量	128
训练小批量样本规模 N_{train}	64
网络软更新系数 χ_a	0.001
网络软更新系数 χ_c	0.001
初始探索方差 $\sigma_{a,\text{init}}^2$	0.9
探索率衰减系数 l	0.999

本文采用通信总速率与感知信噪比作为性能指标。在每个时隙中，通信总速率和感知信噪比均按照短期平均值计算，对应表达式分别为 $\sum_{\tau=t-N_r+1}^t \sum_{u=1}^U R_u(\tau + T_{b,u})/N_r$ 、 $\sum_{\tau=t-N_r+1}^t \Gamma_{\text{tar}}(\tau + 2T_{b,\text{tar}})/N_r$ ，其中 N_r 设置为 2000。为保证可靠性，所有仿真均重复 20 次，以获取平均结果。

图 4(a)和图 4(b)分别展示了不同方案实现的通信总速率与感知信噪比。可以观察到，由于水声信道的快速时变特性，所有曲线均呈现出显著波动。换言之，在不同时隙下，同一方案对应的波束赋形策略可能存在差异^[38]。

(1) DeepMTE 与 DeepMTE-w 对比。在所有对比方案中，DeepMTE 与 DeepMTE-w 均实现了最高的通信总速率与感知信噪比。与 DeepMTE-w 相比，DeepMTE 的起始性能更优，初始总速率高出约 18.82%，初始感知信噪比高出 35.52%。这一提升源于 DeepMTE-w 训练初期经验样本有限，必须从零开始学习；而 DeepMTE 采用预训练的多任务 MoE 架构，即使在预训练任务与当前部署任务不完全一致的情况下，仍能实现高效热启动，并快速适配新任务。此外，与 DeepMTE-w 相比，Deep-

MTE 收敛所需的时隙数减少了 68.81%。这是因为随着试错交互的累积，智能体收集到足够多的经验来训练神经网络，使得 DeepMTE-w 的性能逐步逼近 DeepMTE。上述结果表明，MoE 架构显著提升了 DRL 算法的初始化质量和收敛速度。不过，在交互量充足、算法最终收敛后，该架构并未带来进一步性能增益。

(2) DeepMTE 与 DeepMTE-c 对比。尽管 DeepMTE-c 同样采用多任务混合专家模型完成离线训练与在线执行，但其使用的是传统经验回放机制^[27]。该机制本质上为单任务场景设计，无法促进不同任务间的泛化。因此，与 DeepMTE 相比，DeepMTE-c 收敛所需的时隙数多出约 30%。此外，传统机制仅依赖单步即时奖励来训练神经网络，导致智能体难以有效学习水声长传播时延对系统性能的长远影响。相比之下，DeepMTE 凭借其多步跨任务回放机制，能够更准确地评估这一影响。这一优势使 DeepMTE 相较于 DeepMTE-c 取得了显著的性能提升，总速率和感知信噪比分别提高约 6.08% 和 5.39%，充分验证了所提经验回放机制的有效性。

(3) DeepMTE 与 FPM 对比。如图 4 所示，FPM 作为一种典型的优化方法，性能远低于 DeepMTE，其总速率损失达 21.28%，感知信噪比下降 14.88%。这一性能差距源于 FPM 依赖实时信道状态信息进行波束赋形设计。具体而言，FPM 在每个时隙均基于当前信道状态信息计算通信与感知波束矩阵，但由于水声长传播时延和信道快速时变特性，这些波束在后续传输中会失效。相反，DeepMTE 借助神经网络的推理能力与强化学习的适应能力，能够学习具备前瞻性的波束赋形决策，即便无法获取准确的未来信道状态信息，决策依然保持有效。

(4) DeepMTE 与 SCA 对比。由图 4 可知，SCA 实现的性能略低于 FPM。这是由于在水声长传播时延场景下，目标函数涉及多时间步耦合，FPM 框架凭借其构造代理函数的能力，能够将历史变量自然纳入迭代优化过程，而 SCA 仅依赖当前时刻的梯度信息进行近似，难以有效处理跨时隙耦合项。尽管如此，FPM 与 SCA 的通信和感知性能均远低于 DeepMTE。这一显著差距源于传统优化方法本质上依赖于实时信道状态信息进行波束赋

形设计：FPM 和 SCA 在每个时隙均基于当前 CSI 计算波束赋形向量，但由于水声传播时延较长且信道快速时变，计算出的波束矩阵在后续实际传输时已严重过期失效。相比之下，DeepMTE 即便无法获取准确的未来 CSI，其决策仍保持有效。以上结果表明，在本文所设定的过期 CSI、长传播时延以及相应水声信道模型条件下，DeepMTE 相较于基于数学优化的 FPM 与 SCA 方法展现出明显的性能优势。

(5) DeepMTE 与 AC 对比。与 FPM 以及 SCA 不同，AC 方案采用 DRL 算法进行波束赋形，在通信总速率与感知信噪比上分别实现了约 7.28% 与 4.23% 的增益。但 AC 方案存在两个固有缺陷：其一，它从所有可能动作的概率分布中随机采样动作，可能得到次优决策；其二，它依赖单步经验训练深度神经网络，导致参数更新存在强时序相关性，进而降低学习效率^[25]。相比之下，DeepMTE 生成确定性的波束赋形决策，并利用多步历史经验进行训练。通过克服 AC 方案的随机性与短视性，DeepMTE 实现了显著的性能提升，通信总速率增益达 18.41%，感知信噪比增益达 12.71%。

(6) DeepMTE 与 DeepMTE-o 对比。DeepMTE-o 虽然基于充分训练的多任务 MoE 架构构建，但省略了 DeepMTE 核心的在线微调步骤。由于无法适配实时环境，其性能显著下降，与 DeepMTE 相比，通信总速率降低 9.74%，感知信噪比降低 8.14%。不过，多任务 MoE 架构的固有优势，即共享门控网络与专家网络带来的泛化能力，使其即便面临训练数据分布偏移，仍能保持相对稳定的性能。与 AC 相比，DeepMTE-o 仍保持优势，其总速率增益为 6.88%，感知信噪比增益为 3.54%。上述结果验证了多任务 MoE 架构在复杂多目标水声 ISAC 系统中具备出色的泛化能力。

(7) DeepMTE 与 DeepMTE-m 对比。结果表明，DeepMTE-m 的整体性能略逊于 DeepMTE，其通信总速率和感知信噪比分别下降了约 0.3bps/Hz 和 0.14dB。这一性能差距主要源于：DeepMTE-m 将所有任务的信息压缩在同一网络中进行训练，任务间不可避免存在相互干扰，难以根据不同目标偏好因子提供差异化的优化输出；而 DeepMTE 通过多个专家网络的并行学习与门控网络的动态调度，能够有效解耦不同任务间的冲突，在保持通信与感

知性能平衡的同时，实现更优的综合性能。

(8) DeepMTE 与预测型波束赋形对比。如图 4 所示，预测型波束赋形方案取得了接近 DeepMTE 的通信总速率和感知信噪比。然而，预测型波束赋形方案存在两个关键局限：其一，训练过程需要以先验最优波束赋形矩阵作为标签，而这些标签通常依赖于特定的目标偏好因子，一旦系统通信与感知的权衡偏好发生改变，神经网络需要重新训练，缺乏泛化能力；其二，实际水声环境具有时变和非平稳特性，当训练环境与部署环境不一致时，预先训练好的预测模型将难以适应。相较之下，DeepMTE 通过引入 MoE 架构，能够对不同目标偏好因子具备较强的泛化性；同时，其基于强化学习的在线交互机制能够持续收集环境经验样本，从而实现对水声环境变化的自适应调整。因此，DeepMTE 在实际部署中具有更高的灵活性和鲁棒性^[34]。

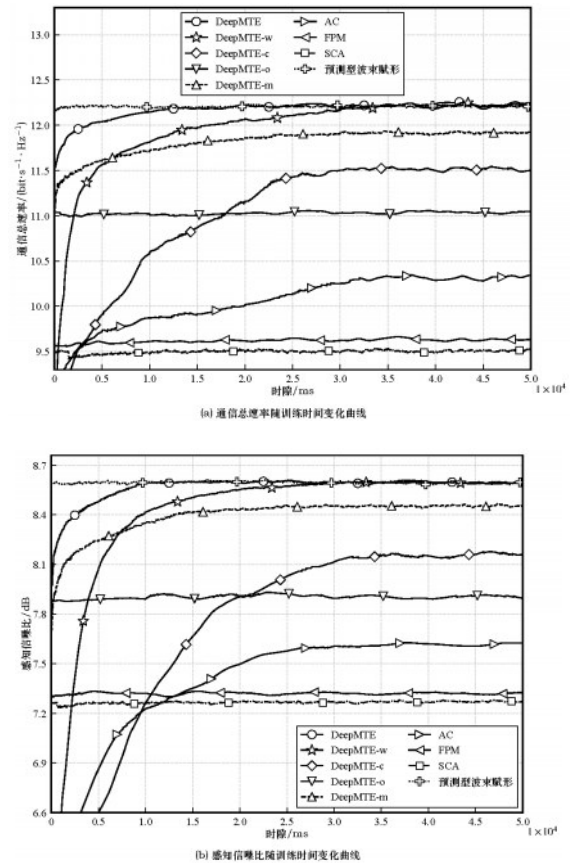


图4 各算法的通信总速率与感知信噪比随时隙变化情况

其次，为验证 DeepMTE 算法对不同目标偏好因子的鲁棒性，因子 b 以 0.2 为步长，由 0.2 递增至 0.8。从图 5(a)和图 5(b)可以看出，随着 b 增大，各

算法的通信总速率均上升,而感知信噪比则降低。该趋势源于通信与感知性能之间的内在权衡: b 越大,优化目标中通信速率的权重越高,感知性能的优化则相应减弱。因此,所有算法在 b 增大时更偏向提升通信速率。此外,随着 b 增大,各方案之间的通信总速率差距不断扩大,而感知信噪比差距则持续缩小。但无论在何种 b 取值下,DeepMTE在系统性能上均明显优于对比算法。具体而言,相较于DeepMTE-c、DeepMTE-o、DeepMTE-m、AC、FPM和SCA算法,所提DeepMTE方案的通信总速率增益分别超过3.84%、10.00%、1.83%、18.41%、24.07%和26.10%;感知信噪比增益分别超过4.41%、5.92%、1.10%、8.91%、12.08%和12.94%。上述结果表明,DeepMTE在目标偏好因子发生变化时具备更强的鲁棒性。更为重要的是,通过调整 b 的取值,能够有效捕捉水声通信性能与感知性能之间的权衡关系,这与现有地面ISAC系统中基于帕累托优化得到的通感性能折中规律相一致^[39]。

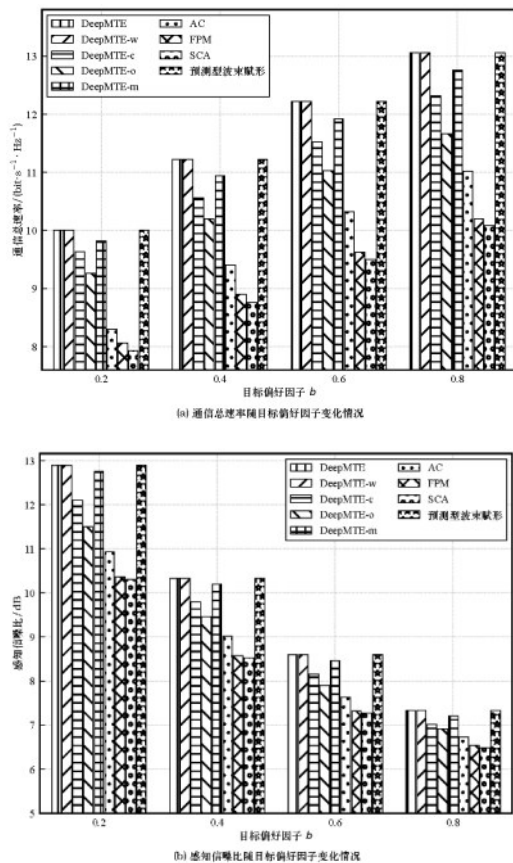


图5 不同目标偏好因子下的通信总速率与感知信噪比性能

为研究DeepMTE在不同天线数量下的鲁棒性,令天线数 N 以2为步长,由6递增至12。图6(a)和图6(b)分别展示了各算法的通信总速率与感知信噪比。正如预期,随着 N 增大,所有方案的总速率与感知信噪比均单调提升。这是因为较大的 N 能够为水声ISAC信号传输提供更多的空间自由度。得益于更智能、更高效的联合优化策略,DeepMTE在通信与感知性能上始终优于其他方案。在不同 N 取值下,DeepMTE的增益如下:相较于DeepMTE-c,总速率和感知信噪比分别提升6.08%和4.35%以上;相较于DeepMTE-o,分别提升9.71%和7.14%以上;相较于DeepMTE-m,分别提升2.14%和1.54%以上;相较于AC,分别提升16.19%和12.63%以上;相较于FPM,分别提升20.94%和15.63%以上;相较于SCA,分别提升22.10%和16.25%以上。

图6同时表明,随着 N 增大,DeepMTE与其他方案在通信和感知两方面的性能差距均逐步扩大。例如,当 N 从6增加到12时,DeepMTE相较于DeepMTE-c的通信总速率提升量从0.70bps/Hz升至0.97bps/Hz,感知信噪比提升量从0.44dB升至0.56dB。上述结果表明,在不同天线数量下,DeepMTE相比其他方案展现出更强的鲁棒性。

图7对比了DeepMTE与单步经验回放方法在不同传播时延下的性能,涵盖近零时延(近似无线信道)、中等时延和较长时延三种典型场景。横坐标四元组表示浮标到目标及三个传感器之间的传播时延。仿真表明,随着传播时延增大,传感器及目标到浮标的距离增加,路径损耗加剧,CSI时效性恶化,两种方法的系统性能均下降。在近零时延下,两者性能相当,多步奖励未带来明显增益,表明即时奖励已足够有效。随着时延增加,单步方法性能明显退化,因其仅依赖当前时隙奖励,难以捕捉长期影响;而DeepMTE借助多步奖励机制,能有效关联当前动作与未来多个时隙的通信与感知性能,在长时延下仍保持较优性能。当时延进一步增大,回声干扰跨越更多时隙,时序关系更复杂,DeepMTE的优势愈发突出。这充分证明了多步机制的有效性主要来源于对水声长传播时延效应的适配,而非一般意义上多步TD学习的通用性能提升。

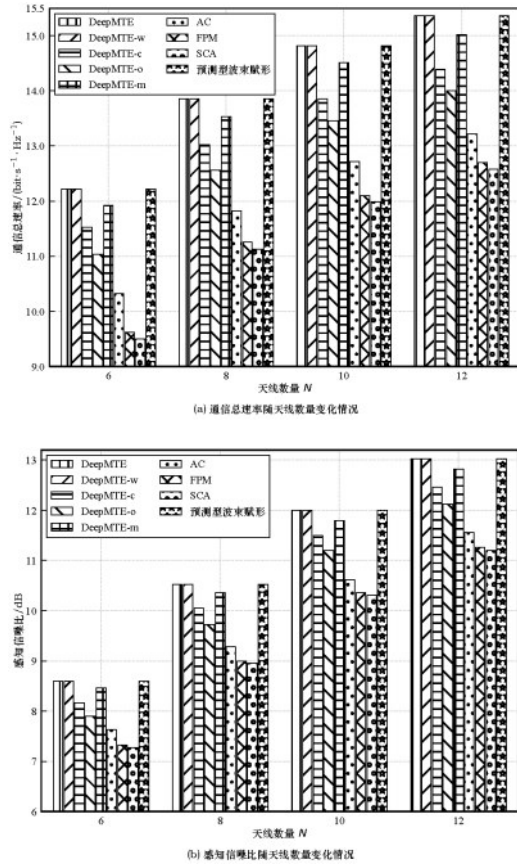


图6 不同天线数量下的通信总速率与感知信噪比性能

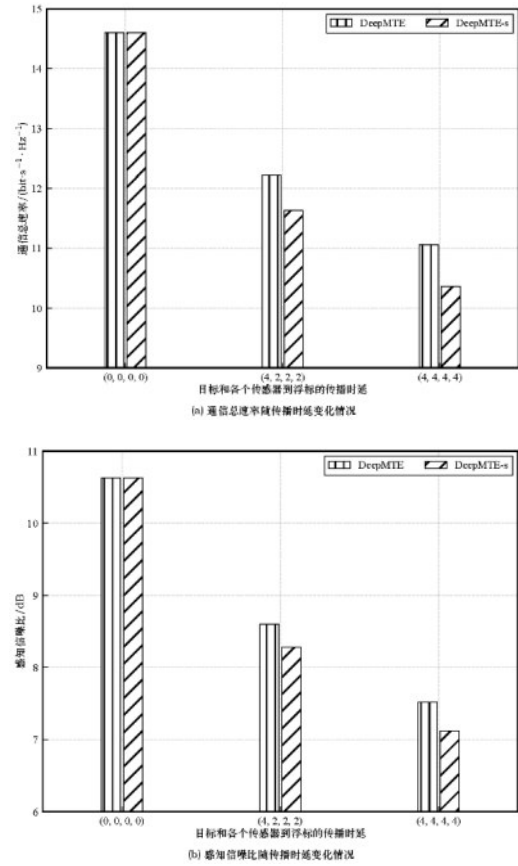


图7 不同传播时延下的通信总速率与感知信噪比性能

4 结束语

本文提出了一种面向水声 ISAC 系统的新型波束赋形优化方案 DeepMTE，该方案引入目标偏好因子，以动态权衡通信与感知性能。DeepMTE 融合深度强化学习与 MoE 架构以提升算法在不同偏好因子下的泛化能力。此外，针对长传播时延下当前波束赋形决策对后续多个时隙系统性能的长期影响，设计了多步跨任务经验回放机制。仿真结果证实，与各基线方案相比，DeepMTE 能够得到帕累托效率更高的解，且在不同天线配置下均展现出更强的鲁棒性。同时，本文验证了 DeepMTE 对未见过的目标偏好因子具有良好的泛化能力。

参考文献:

- [1] Yan J, Guan X, Yang X, et al. A survey on integration design of localization, communication and control for underwater acoustic sensor networks[J]. IEEE Internet of Things Journal, 2025, 12(6): 6300-6324.
- [2] Jehangir A, Ashraf S M, Khalil R A, et al. ISAC-enabled underwater IoT network localization: overcoming asynchrony, mobility, and stratification issues[J]. IEEE Open Journal of the Communications Society, 2024, 5: 3277-3288.
- [3] Wang J, Wang Q, Tao Q, et al. Integrated sonar and communication: a survey[J]. IEEE Communications Surveys & Tutorials, 2026, 28: 4127-4155.
- [4] Bao J, Wang J, Tao Q, et al. Fresnel domain sector modulation LFM signal for underwater integrated communication and detection[J]. IEEE Internet of Things Journal, 2025, 12(3): 3017-3027.
- [5] Lv T, Tian Q, Wang Y, et al. Design of an underwater acoustic waveform and integrated system for communication and detection[J]. Journal of Marine Science and Engineering, 2025, 13(1): 114.
- [6] Men W, Du J, Wang J, et al. Design of ISAC waveform and multiple-access interference suppression receiver for underwater acoustic sensor networks[C]//Proceedings of the 2024 IEEE Global Communications Conference (GLOBECOM). Piscataway: IEEE Press, 2024: 3461-3466.
- [7] Niu Q, Zhang Q, Shi W. Integrated waveform design scheme based on underwater detection and communication[C]//Proceedings of the 2022 International Conference on Signal Processing, Communications and

- Computing (ICSPCC). Piscataway: IEEE Press, 2022: 1-5.
- [8] Liu L, Cui X, Li J, et al. Joint underwater detection and communication with GSFM: integrated waveform design and GA-based parameter optimization[J]. *IEEE Internet of Things Journal*, 2026, 13(7): 14486-14498.
 - [9] Wang Y, Wang J, Zhou M, et al. Integration communication and detection scheme based on superimposed waveform[J]. *IEEE Transactions on Communications*, 2025, 73(7): 5155-5167.
 - [10] Wang J, Lian J, Zhu G. A weighted Mahalanobis distance target sensing strategy for underwater integrated sensing and communication systems[J]. *IEEE Wireless Communications Letters*, 2024, 13(1): 79-83.
 - [11] Wang J, Cui Y, Zhou M, et al. Doppler shift estimation for downlink OFDM signals in underwater acoustic integrated sensing and communication networks[C]//*Proceedings of the 2024 IEEE International Conference on Signal Processing (ICSP)*. Piscataway: IEEE Press, 2024: 125-129.
 - [12] Wang Y, Shi Z, Ma X, et al. A joint sonar-communication system based on multicarrier waveforms[J]. *IEEE Signal Processing Letters*, 2022, 29: 777-781.
 - [13] Men W, Wang J, Dong B, et al. OFDM-based underwater integrated sensing and communication: receiver design for doubly spread acoustic channels[J]. *IEEE Transactions on Communications*, 2025, 73(10): 9372-9387.
 - [14] Wang J, Wang Q, Tao Q, et al. Dictionary-theory-based orthogonal chirp division multiplexing signal for integrated sonar and communication system[J]. *IEEE Transactions on Vehicular Technology*, 2025, 74(5): 7005-7019.
 - [15] Niu Q, Shi W, Jing L, et al. Joint optimization of underwater acoustic ISUDC waveform design and sparse channel estimation algorithms[J]. *IEEE Internet of Things Journal*, 2025, 12(20): 43756-43771.
 - [16] Hu T, Zhuo X, Tang L, et al. Multi-AUV collaborative data collection and trajectory planning in integrated sensing and communication for underwater acoustic networks[C]//*Proceedings of the 2024 IEEE Vehicular Technology Conference (VTC)*. Piscataway: IEEE Press, 2024: 1-6.
 - [17] Wu Y, Zhuo X, Tang L, et al. Coverage path planning for AUVs cooperative environment detection in integrated underwater acoustic communication and detection networks[C]//*Proceedings of the 2024 IEEE Vehicular Technology Conference (VTC)*. Piscataway: IEEE Press, 2024: 1-6.
 - [18] Guo Y, Zhuo X, Tang L, et al. Neighbor discovery with directional transmission in integrated sensing and communication for underwater acoustic networks[C]//*Proceedings of the 2023 IEEE/CIC International Conference on Communications in China Workshops (ICCC Workshops)*. Piscataway: IEEE Press, 2023: 1-6.
 - [19] Men W, Zhao L, Guan Y L, et al. Efficient resource allocation for multi-user and multi-target MIMO-OFDM underwater ISAC[C]//*Proceedings of the 2025 IEEE Global Communications Conference (GLOBECOM)*. Piscataway: IEEE Press, 2025: 1-6.
 - [20] Liang S, Mao W, Zhao D, et al. Ranging enhancement for underwater collision avoidance sonar via ISAC waveform and adaptive sensing acceleration[J]. *IEEE Transactions on Instrumentation and Measurement*, 2026. DOI: 10.1109/TIM.2026.3655998.
 - [21] Li J, Tong F, Jiang W, et al. Metastructure-enabled underwater acoustic integrated sensing and communication[J]. *IEEE Wireless Communications Letters*, 2024, 13(12): 3335-3339.
 - [22] Ye X, Yu Y, Fu L. Deep reinforcement learning based MAC protocol for underwater acoustic networks[J]. *IEEE Transactions on Mobile Computing*, 2022, 21(5): 1625-1638.
 - [23] Tang R, Zhang R, Xu Y, et al. Resource allocation for underwater acoustic sensor networks with partial spectrum sharing: when optimization meets deep reinforcement learning[J]. *IEEE Transactions on Network and Service Management*, 2025, 22(3): 2715-2730.
 - [24] He Y, Han G, Zhu S, et al. Trust management based on attention-weighted federated deep reinforcement learning for underwater acoustic sensor networks[J]. *IEEE Transactions on Mobile Computing*, 2026. DOI: 10.1109/TMC.2026.3651511.
 - [25] Sutton R S, Barto A G. Reinforcement learning: an introduction[M]. Cambridge: MIT Press, 2018.
 - [26] Ma J, Zhao Z, Yi X, et al. Modeling task relationships in multi-task learning with multi-gate mixture-of-experts[C]//*Proceedings of the 2018 ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD)*. New York: ACM Press, 2018: 1930-1939.
 - [27] Lillicrap T P, Hunt J J, Pritzel A, et al. Continuous control with deep reinforcement learning[C]//*Proceedings of the 2016 International Conference on Learning Representations (ICLR)*. San Juan, 2016.
 - [28] Lin L J. Self-improving reactive agents based on reinforcement learning, planning and teaching[J]. *Machine Learning*, 1992, 8(3): 293-321.
 - [29] Mnih V, Badia A P, Mirza M, et al. Asynchronous methods for deep reinforcement learning[C]//*Proceedings of the 2016 International Conference on Machine Learning (ICML)*. New York: PMLR, 2016: 1928-1937.
 - [30] Liu R, Li M, Liu Q, et al. SNR/CRB-constrained joint beamforming and reflection designs for RIS-ISAC systems[J]. *IEEE Transactions on Wireless Communications*, 2024, 23(7): 7456-7470.
 - [31] Wu W, Gao X, Sun C, et al. Shallow underwater acoustic massive MIMO communications[J]. *IEEE Transactions on Signal Processing*, 2021, 69: 1124-1139.
 - [32] Ye X, Song X, Wu Y, et al. Intelligent omni-surface-aided multi-objective ISAC: a meta hybrid deep reinforcement learning approach [J]. *IEEE Transactions on Mobile Computing*, 2026, 25(3): 3760-3775.
 - [33] Fujimoto S, Hoof H, Meger D. Addressing function approximation error in actor-critic methods[C]//*Proceedings of the 2018 International Conference on Machine Learning (ICML)*. Stockholm: PMLR, 2018: 1587-1596.
 - [34] Ye X, Mao Y, Yu X, et al. Integrated sensing and communications for low-altitude economy: a deep reinforcement learning approach[J]. *IEEE Transactions on Wireless Communications*, 2026, 25: 351-367.
 - [35] Aref M A, Jayaweera S K. Spectrum-agile cognitive radios using multi-task transfer deep reinforcement learning[J]. *IEEE Transactions on Wireless Communications*, 2021, 20(10): 6729-6742.
 - [36] Fang T, Ma M, Juntti M, et al. Optimal ISAC beamforming structure and efficient algorithms for sum rate and CRLB balancing[J]. *IEEE Transactions on Signal Processing*, 2026, 74: 2400-2416.
 - [37] Gao Q, Zhong R, Zou Y, et al. Deep learning-based beamforming optimization for ISAC systems: a low-complexity and transferable framework[J]. *IEEE Transactions on Wireless Communications*, 2026, 25: 8754-8768.
 - [38] Ye X, Song X, Wu Y, et al. Integrated sensing and communication for underwater acoustic networks based on deep reinforcement learning[J]. *IEEE Transactions on Mobile Computing*, 2026, 25: 14482-14497.

- [39] Zhao B, Cui Q, Ni W, et al. Pareto optimization for performance trade-off in multi-functional RIS-assisted ISAC systems[J]. IEEE Transactions on Vehicular Technology, 2025, 74(11): 18267-18272.



叶小文 (1996-), 男, 博士, 福建师范大学副教授, 主要研究方向为智能无线通信、水声通信组网、通感一体化等。



林恒羿 (2002-), 男, 硕士, 福建师范大学学生, 主要研究方向为无线通信、强化学习等。



周茂伟 (2004-), 男, 硕士, 福建师范大学学生, 主要研究方向为水声通信、强化学习等。



吴怡 (1970-), 女, 博士, 福建师范大学教授、博士生导师, 主要研究方向为海上/水下通信、信道接入、车联网、智能通信等。