控制与决策
Control and Decision

基于深度强化学习的多潜器编队控制算法设计

闫敬, 徐龙, 曹文强, 杨晛, 罗小元

引用本文:
闫敬,徐龙,曹文强,杨,罗小元. 基于深度强化学习的多潜器编队控制算法设计[J]. 控制与决策, 2023, 38(5): 1457−1463.

在线阅读 View online: https://doi.org/10.13195/j.kzyjc.2022.1424

基于深度强化学习的多潜器编队控制算法设计

闫 敬†, 徐 龙, 曹文强, 杨 晛, 罗小元

(燕山大学 电气工程学院，河北 秦皇岛 066004)

摘 要: 考虑水下未知信道与不确定模型参数,提出基于深度强化学习的多潜器编队控制算法. 首先,提出基于环境采样数据的最小二乘估计器,用于预测在衰落环境下的未知信道; 其次, 根据信道预测估计器得出的信噪比(SNR), 建立通信有效性与编队稳定性的联合优化问题, 并给出基于深度强化学习-深度确定性策略梯度算法(DDPG)的编队控制算法; 最后, 通过仿真与实验结果验证所提出算法的有效性, 参考仿真结果并相比于直接编队控制, 考虑通信有效性的情况下所提出算法提升了 13.5 %的通信性能.

关键词: 信道预测；潜器；编队控制；深度强化学习；联合优化

中图分类号: TP273 **文献标志码:** A

DOI: 10.13195/j.kzyjc.2022.1424

引用格式: 闫敬, 徐龙, 曹文强, 等. 基于深度强化学习的多潜器编队控制算法设计 [J]. 控制与决策, 2023, 38(5): 1457-1463.

Design of formation control algorithm for multiple autonomous underwater vehicles based on deep reinforcement learning

YAN Jing†, XU Long, CAO Wen-qiang, YANG Xian, LUO Xiao-yuan

(School of Electrical Engineering，Yanshan University，Qinhuangdao 066004，China)

Abstract: This paper considers the unknown underwater channel and the uncertain model parameters, and hence, a multiple autonomous underwater vehicles (AUVs) formation control algorithm based on deep reinforcement learning is proposed. Firstly, a least square estimator based on environmental sampling data is developed to predict the unknown channel in fading environment. Then, according to the sigal-noise ratio (SNR) obtained by the channel prediction estimator, the co-optimization problem of communication effectiveness and formation stability is established. Based on this, the formation control algorithm based on the depth deterministic strategy gradient algorithm (DDPG) is designed. Finally, simulation and experimental results verify the effectiveness of the proposed algorithm. According to the simulation results, compared with the direct formation control, the communication performance is improved by 13.5% considering the communication efficiency.

Keywords: channel prediction；AUV；formation control；deep reinforcement learning；co-optimization

0 引 言

近年来, 潜器 (autonomous underwater vehicle, AUV) 在海洋研究领域发挥着重要作用, 例如入侵监视、海上救援和石油勘探等[1-2]. 其中编队控制是多潜器协同控制的重要组成部分, 并在许多应用中具有更高的工作效率和更好的系统稳定性.

在多潜器编队系统设计过程中, 关键问题在于设计合适的编队控制算法, 使得多潜器保持特定编队队形. Liu 等[3] 采用反馈线性化补偿潜器模型的不确定性, 设计了多潜器鲁棒编队控制算法. Gao 等[4] 在潜器编队控制器中加入扰动观测器, 以确保固定时间内潜器编队的稳定性. 此外, Millán 等[5] 提出了一种反馈前馈联合的多潜器编队控制器, Wang 等[6] 设计了一种基于神经网络的自适应编队控制算法. 然而, 上述编队控制器依赖或部分依赖于潜器模型. 为此, 学者们提出了基于强化学习的编队控制算法克服对模型参数的依赖. 例如: Yuan 等[7] 采用径向基函数神经网络识别不确定的潜器动力学模型, 并给出基于强化学习的编队控制算法. Gu 等[8] 提出了无模型的潜器编队控制算法. 然而, 基于强化学习的编队控制算法

收稿日期: 2022-08-07；**录用日期:** 2022-12-30.

基金项目: 国家自然科学基金项目 (62222314, 61973263, 61873345, 62033011)；河北省自然科学基金项目 (2022203001, BJ2020031)；河北省中央引导地方基金项目 (226Z3201G).

责任编委: 杨涛.

†通讯作者. E-mail: jyan@ysu.edu.cn.

在设计过程中需要频繁试凑基函数结构与形式,而且采用单层结构的神经网络,拟合能力较弱. 此外,目前潜器编队控制研究主要聚焦在编队控制方法设计上,并没有考虑水声通信信道对潜器编队控制的影响,即忽略了水声通信通道的影响[9]. 然而,相比于电磁波通信信道,水声通信信道具有更强的阴影衰落和多径衰落,忽略上述因素将导致潜器编队过程中通信效率下降,甚至引起编队任务失效.

基于上述考虑,本文提出基于深度强化学习-深度确定性策略梯度(deterministic diagnostic pattern generation, DDPG)的多潜器编队控制算法,主要贡献如下: 1) 联合优化框架. 不同于其他潜器编队控制研究[9],提出水下信道预测与编队控制协同优化框架,该框架考虑了通信有效性与编队稳定性. 2) 提出无模型多潜器编队控制算法. 考虑通信质量提出基于DDPG的编队控制算法. 相比于文献[7-8],该编队控制算法摆脱了对模型参数的依赖.

1 问题描述

假定潜器可通过浮标与潜标进行自定位[10],为此,潜器四自由度运动模型描述如下:

$$\dot{\boldsymbol{\eta}}_{\bar{m}} = \boldsymbol{J}(\varphi_{\bar{m}})\boldsymbol{V}_{\bar{m}},$$

$$\boldsymbol{M}_{\bar{m}}\dot{\boldsymbol{V}}_{\bar{m}} + \boldsymbol{C}_{\bar{m}}(\boldsymbol{V}_{\bar{m}})\boldsymbol{V}_{\bar{m}} + \boldsymbol{D}_{\bar{m}}(\boldsymbol{V}_{\bar{m}})\boldsymbol{V}_{\bar{m}} = \boldsymbol{\tau}_{\bar{m}}. \quad (1)$$

其中: $\boldsymbol{V}_{\bar{m}} = [u_{\bar{m}}, v_{\bar{m}}, w_{\bar{m}}, r_{\bar{m}}]^{\mathrm{T}}$ 为潜器速度向量, $\bar{m} \in \{1, 2, \ldots, \bar{M}\}$, $\bar{M}$ 为潜器个数, $u_{\bar{m}}$、$v_{\bar{m}}$ 和 $w_{\bar{m}}$ 为潜器在纵荡、横荡、垂荡方向上的线速度, $r_{\bar{m}}$ 为潜器偏航角的速度; $\boldsymbol{\eta}_{\bar{m}} = [x_{\bar{m}}, y_{\bar{m}}, z_{\bar{m}}, \varphi_{\bar{m}}]^{\mathrm{T}}$ 为潜器位置向量, $x_{\bar{m}}$、$y_{\bar{m}}$ 和 $z_{\bar{m}}$ 为潜器位置, $\varphi_{\bar{m}}$ 为潜器偏航角度; $\boldsymbol{M}_{\bar{m}} \in \boldsymbol{R}^{4\times4}$ 为惯性矩阵; $\boldsymbol{C}_{\bar{m}}(\boldsymbol{V}_{\bar{m}}) \in \boldsymbol{R}^{4\times4}$ 为科里奥利和向心耦合矩阵; $\boldsymbol{D}_{\bar{m}}(\boldsymbol{V}_{\bar{m}}) \in \boldsymbol{R}^{4\times4}$ 为阻尼矩阵; $\boldsymbol{\tau}_{\bar{m}} = [\tau_{u_{\bar{m}}}, \tau_{v_{\bar{m}}}, \tau_{w_{\bar{m}}}, \tau_{r_{\bar{m}}}]^{\mathrm{T}}$ 为潜器控制量, $\tau_{u_{\bar{m}}}$、$\tau_{v_{\bar{m}}}$、$\tau_{w_{\bar{m}}}$ 和 $\tau_{r_{\bar{m}}}$ 分别为纵荡、横荡、垂荡和偏航方向上的作用力; $\boldsymbol{J}(\varphi_{\bar{m}})$ 为旋转矩阵.

$\Upsilon_{\mathrm{RX}}(q)$ 为潜器接收到潜标发来的信号强度, $q_b = [x_b, y_b, z_b]^{\mathrm{T}}$ 表示位置,潜器位置为 $q = [x, y, z]^{\mathrm{T}}$. 潜标发送端的信道强度 P_T 经当前信道增益 $g(q)$ 放大,最后在窄带宽信道模型下潜器接收到的信号强度可描述为 $\Upsilon_{\mathrm{RX}}(q) = g(q)P_T + \rho$,其中 ρ 为发送端和接受端的热噪声. 潜器过滤热噪声后接收到的信号强度为 $\Upsilon(q) = \Upsilon_{\mathrm{RX}}(q) - \rho$. 根据文献[11],两点间信噪比模型可表示为

$$\begin{aligned}&\mathrm{SNR}_{\mathrm{dB}}(q, q_b) = \\ &K_{\mathrm{PL}} - 10n_{\mathrm{PL}}\log_{10}(d(q, q_b)) + \bar{\sigma}_{\mathrm{SH}}(q, q_b) + \\ &\bar{\mu}_{\mathrm{MP}}(q, q_b) - 10d(q, q_b)\log_{10}(\alpha(f)) - N_{\mathrm{dB}}. \end{aligned} \quad (2)$$

其中: $d(q, q_b)$ 为潜器与潜标之间的欧氏距离; K_{PL} 和 n_{PL} 为路径损耗参数,取决于环境传播介质的影响; $\bar{\sigma}_{\mathrm{SH}}(q, q_b)$ 为环境中的阴影损耗强度; $\bar{\mu}_{\mathrm{MP}}(q, q_b)$ 为信号传播过程中的多径损耗; $\alpha(f)$ 为水下吸声系数; N_{dB} 为水下信道模型噪声. 上述模型中, N_{dB} 可通过环境先验数据测量得知,根据环境采样数据可以估计路径损耗参数 K_{PL} 和 n_{PL}. 至于阴影衰减参数 $\bar{\sigma}_{\mathrm{SH}}(q, q_b)$ 和多径损耗参数 $\bar{\mu}_{\mathrm{MP}}(q, q_b)$,以往的经验数据表明,它们服从零均值的高斯分布[12]. 针对阴影损耗, Gudmundson[13]用指数衰减的空间相关函数表征阴影衰减分量,假设阴影衰减服从以下高斯分布: $f_{\bar{\sigma}}(x) = \dfrac{1}{\sqrt{2\pi\alpha}}\mathrm{e}^{\frac{x^2}{2\alpha}}$. 环境中任意两点间的阴影衰减空间相关性可表述为

$$\mathrm{cov}(\bar{\sigma}(q_1), \bar{\sigma}(q_2)) = E\{\bar{\sigma}(q_1)\bar{\sigma}(q_2)\} = \alpha\mathrm{e}^{\frac{\|q_1 - q_2\|}{\beta}}.$$

其中: α 为阴影效应的方差, β 反应阴影衰减的空间相关性.

领导者-追随者潜器编队系统需实现以下任务:

$$\begin{cases}\lim\limits_{t\to\infty}\|\boldsymbol{p}_{ld} - \boldsymbol{p}_{ld\_g}\| = 0,\\ \lim\limits_{t\to\infty}\|\boldsymbol{p}_{f} - \boldsymbol{p}_{f\_g}\| = 0,\\ \lim\limits_{t\to\infty}\|\boldsymbol{p}_{ld} - \boldsymbol{r}_{ldf} - \boldsymbol{p}_{f}\| = 0.\end{cases} \quad (3)$$

其中: $\boldsymbol{p}_{ld}$ 为领导者潜器位置; $\boldsymbol{p}_f$ 为其余追随潜器; $\boldsymbol{r}_{ldf}$ 为两者期望相对位置向量; $\boldsymbol{p}_{ld\_g}$ 和 $\boldsymbol{p}_{f\_g}$ 为潜器的目标点位置.

为了实现上述目的,需要解决以下问题: 1) 衰落环境下的水声信道预测. 设计了一种基于环境数据采样的最小二乘估计器,以实现对未知环境下的信道参数估计. 2) 多潜器编队控制. 提出基于深度强化学习的潜器编队控制算法,在通信约束下对潜器进行控制,以提升通信有效性.

2 主要结果

2.1 衰落环境下的水声信道预测

根据环境采样数据对信道参数估计,其中潜标位置 q_b 已知,采样数据收集过程如下: 首先给出 n 个SNR采样点 $[q_1, \ldots, q_n]$; 其次潜器在采样位置完成SNR数据测量; 最后得到SNR测量向量,记作 $\boldsymbol{Y}_{\mathrm{dB}} = [y_1, \ldots, y_n]^{\mathrm{T}}$. 给出如下信道预测模型:

$$\boldsymbol{Y}_{\mathrm{dB}} = \boldsymbol{H}\theta + \sigma_{\mathrm{SH}} + \mu_{\mathrm{MP}} - \boldsymbol{\varepsilon}. \quad (4)$$

其中

$$\boldsymbol{H} = \begin{bmatrix}1 & -10\log_{10}(d(q_1, q_b))\\ \vdots & \vdots\\ 1 & -10\log_{10}(d(q_n, q_b))\end{bmatrix}, \quad (5)$$

$$\boldsymbol{\varepsilon}=\begin{bmatrix}1 & d(q_1,q_b)\\ \vdots & \vdots\\ 1 & d(q_n,q_b)\end{bmatrix}\begin{bmatrix}N_{\rm dB}\\ 10\log_{10}(\alpha(f))\end{bmatrix}. \tag{6}$$

这里:$\theta=[K_{\rm PL},n_{\rm PL}]^{\rm T}$、$\sigma_{\rm SH}=[\bar{\sigma}_{\rm SH}(q_1,q_b),\dots,\bar{\sigma}_{\rm SH}(q_n,q_b)]^{\rm T}$,$\mu_{\rm MP}=[\bar{\mu}_{\rm MP}(q_1,q_b),\dots,\bar{\mu}_{\rm MP}(q_n,q_b)]^{\rm T}$;$\theta$、$\sigma$和$\mu$分别为采样点的路径损耗参数、阴影和多径衰落分量. 前文提到多径变量和阴影变量是服从零均值的高斯分布,SNR测量向量服从均值为$\boldsymbol{H}\theta+\varepsilon$、方差为$\boldsymbol{R}+\mu^2\boldsymbol{I}_n$的高斯分布,其中$\boldsymbol{I}_n$为$n\times n$的单位矩阵. 此处的方差表示阴影效应空间相关性的协方差矩阵$\boldsymbol{R}$与相互独立多径效应方差对角矩阵$\mu^2\boldsymbol{I}_n$之和. 基于此,$\boldsymbol{Y}$的概率密度函数描述为

$$\begin{aligned}&f(\boldsymbol{Y}|\theta,\alpha,\beta,\mu^2)=\\&\frac{1}{\sqrt{(2\pi)^n|\boldsymbol{R}+\mu^2\boldsymbol{I}_n|}}\exp\Big\{-\frac{1}{2}(\boldsymbol{Y}-\boldsymbol{H}\theta-\varepsilon)^{\rm T}\times\\&(\boldsymbol{R}+\mu^2\boldsymbol{I}_n)^{-1}(\boldsymbol{Y}-\boldsymbol{H}\theta-\varepsilon)\Big\},\end{aligned} \tag{7}$$

其中

$$\boldsymbol{R}=\alpha\begin{bmatrix}1 & \exp\left(\frac{d(q_1,q_2)}{-\beta}\right) & \dots & \exp\left(\frac{d(q_1,q_n)}{-\beta}\right)\\ \exp\left(\frac{d(q_2,q_1)}{-\beta}\right) & 1 & \dots & \exp\left(\frac{d(q_2,q_n)}{-\beta}\right)\\ \vdots & \vdots & \ddots & \vdots\\ \exp\left(\frac{d(q_n,q_1)}{-\beta}\right) & \exp\left(\frac{d(q_n,q_2)}{-\beta}\right) & \dots & 1\end{bmatrix}.$$

用$\chi=\alpha+\mu^2$表示阴影和多径效应造成的随机损耗之和. 由于阴影损耗和多径损耗服从均值为零的高斯分布,最小二乘估计器表示为

$$\hat{\theta}_{\rm LS}=(\boldsymbol{H}^{\rm T}\boldsymbol{H})^{-1}\boldsymbol{H}^{\rm T}(\boldsymbol{Y}_{\rm dB}+\boldsymbol{\varepsilon}), \tag{8}$$

$$\begin{aligned}&\hat{\chi}_{{\rm LS}|\theta=\hat{\theta}_{\rm LS}}=\\&\frac{1}{n}(\boldsymbol{Y}_{\rm dB}+\boldsymbol{\varepsilon}-\boldsymbol{H}\hat{\theta}_{\rm LS})^{\rm T}(\boldsymbol{Y}_{\rm dB}+\varepsilon-\boldsymbol{H}\hat{\theta}_{\rm LS})=\\&\frac{1}{n}(\boldsymbol{Y}_{\rm dB}+\varepsilon)^{\rm T}(\boldsymbol{I}_n-\boldsymbol{H}(\boldsymbol{H}^{\rm T}\boldsymbol{H})^{-1}\boldsymbol{H}^{\rm T})(\boldsymbol{Y}_{\rm dB}+\varepsilon).\end{aligned} \tag{9}$$

采样点处阴影和多径效应信号强度可表示为

$$\begin{aligned}\boldsymbol{Y}_{\rm LS}&=\boldsymbol{Y}_{\rm dB}+\varepsilon-\boldsymbol{H}\hat{\theta}_{\rm LS}=\\&(\boldsymbol{I}_n-\boldsymbol{H}(\boldsymbol{H}^{\rm T}\boldsymbol{H})^{-1}\boldsymbol{H}^{\rm T})(\boldsymbol{Y}_{\rm dB}+\boldsymbol{\varepsilon}).\end{aligned} \tag{10}$$

对阴影衰减空间相关性参数(α,β)进行数值估计,信道参数估计过程如下:

$$\begin{aligned}&[\hat{\alpha}_{\rm LS};\hat{\beta}_{\rm LS}]=\\&\arg\min_{\alpha,\beta}\sum_{i,j}w_{ij}\ln\Big(\frac{\alpha\exp\Big\{-\frac{1}{\beta}\|q_i-q_j\|\Big\}}{(\boldsymbol{Y}_{\rm LS}\boldsymbol{Y}_{\rm LS})_{i,j}}\Big).\end{aligned} \tag{11}$$

式(11)中$i\neq j$,取其对数可得

$$\Big[\ln(\hat{\alpha}_{\rm LS});\frac{1}{\hat{\beta}_{\rm LS}}\Big]=(\widetilde{\boldsymbol{M}}^{\rm T}\boldsymbol{W}\widetilde{\boldsymbol{M}})^{-1}\widetilde{\boldsymbol{M}}^{\rm T}\boldsymbol{W}\boldsymbol{b}. \tag{12}$$

其中

$$\tilde{\boldsymbol{M}}=\begin{bmatrix}1 & -d_1\\ \vdots & \vdots\\ 1 & -d_{\frac{n(n-1)}{2}}\end{bmatrix},\ \boldsymbol{b}=\begin{bmatrix}\ln((\boldsymbol{Y}_{\rm LS}\boldsymbol{Y}_{\rm LS})_{1,2})\\ \vdots\\ \ln((\boldsymbol{Y}_{\rm LS}\boldsymbol{Y}_{\rm LS})_{n-1,n})\end{bmatrix},$$

$\boldsymbol{W}=[w_{ij}]\in\boldsymbol{R}^{n\times n}$为权重矩阵,权重$w_{ij}$取残差平方的倒数. 式(12)中,$0<(\boldsymbol{Y}_{\rm LS}\boldsymbol{Y}_{\rm LS})_{i,j}<\hat{\chi}_{{\rm LS}|\theta=\hat{\theta}_{\rm LS}},\hat{\beta}_{\rm LS}>0$. 根据以上结果,针对多径衰减参数,选取相同位置的采样点数据,将多径衰减分量提取出来,即$\hat{\mu}^2_{\rm LS}=\hat{\chi}_{{\rm LS}|\theta=\hat{\theta}_{\rm LS}}-\hat{\alpha}_{\rm LS}$. 基于参数估计结果,环境中信噪比均值和方差结果如下:

$$\begin{aligned}&\bar{\Upsilon}_{\rm dB}(q)=\boldsymbol{h}(q)\theta-\varepsilon_{q,q_b}+\phi^{\rm T}(q)\boldsymbol{R}^{-1}(\boldsymbol{Y}_{\rm dB}-\boldsymbol{H}\theta+\boldsymbol{\varepsilon}),\\&\boldsymbol{V}^2_{\rm dB}(q)=\mu^2+\alpha-\phi^{\rm T}(q)\boldsymbol{R}^{-1}\phi(q).\end{aligned} \tag{13}$$

其中:$\boldsymbol{h}(q)=[1,-10\log_{10}(d(q,q_b))]$,$\phi=\alpha[{\rm e}^{-\frac{\|q-q_1\|}{\beta}},\dots,{\rm e}^{-\frac{\|q-q_n\|}{\beta}}]^{\rm T}$,结合式(6)可知$\varepsilon_{q,q_b}=N_{\rm dB}+d(q,q_b)10\log_{10}(\alpha(f))$表示水下噪声.

2.2　信道感知下的多潜器编队控制

不失一般性,用3个潜器作为本文的控制对象,其中3个潜器分别为领导者潜器、跟随者潜器1和跟随者潜器2,其位置分别用$\boldsymbol{p}_{ld}$、$\boldsymbol{p}_{f1}$和$\boldsymbol{p}_{f2}$表示.

记观测状态为$\boldsymbol{s}_{ld}=[x_{ld},y_{ld},z_{ld},\varphi_{ld},u_{ld},v_{ld},w_{ld},r_{ld}]$,$\boldsymbol{s}_{f1}=[x_{f1},y_{f1},z_{f1},\varphi_{f1},u_{f1},v_{f1},w_{f1},r_{f1}]$,$\boldsymbol{s}_{f2}=[x_{f2},y_{f2},z_{f2},\varphi_{f2},u_{f2},v_{f2},w_{f2},r_{f2}]$,$\boldsymbol{s}=[\boldsymbol{s}_{ld},\boldsymbol{s}_{f1},\boldsymbol{s}_{f2}]$为整体观测. 观测控制输出为$\boldsymbol{a}=[\tau_{uld},\tau_{vld},\tau_{wld},\tau_{rld},\tau_{ufl1},\tau_{vfl1},\tau_{wfl1},\tau_{rfl1},\tau_{ufl2},\tau_{vfl2},\tau_{wfl2},\tau_{rfl2}]$. 基于此,编队学习奖励函数为

$$r_{\rm sum}=r_{ld}+r_f+r_{eg}+r_{ep\_{\rm dot}}+r_{\rm snr}. \tag{14}$$

其中:$r_{ld}=c_{ld}e_p$为潜器当前位置与目标点的距离误差项,$e_p=\|\boldsymbol{p}_{ld}-\boldsymbol{p}_{ld_g}\|+\|\boldsymbol{p}_{f1}-\boldsymbol{p}_{f1_g}\|+\|\boldsymbol{p}_{f2}-\boldsymbol{p}_{f2_g}\|$为潜器当前位置与目标点的距离误差项;$r_f=c_fe_f$为跟随者潜器当前位置与编队期望位置的距离误差项,$e_f=\|\boldsymbol{p}_{ld}-\boldsymbol{r}_{ldf}-\boldsymbol{p}_f\|$;$r_{eg}=\boldsymbol{a}\boldsymbol{C}\boldsymbol{a}^{\rm T}$为能耗惩罚项,对潜器能量进行约束;$r_{ep_{\rm dot}}=c_{ep_{\rm dot}}\dot{e}_p$为目标点距离变化率奖励,促进算法收敛;$r_{\rm snr}$为编队中所有潜器当前位置预测到的信噪比奖励,该信噪比由式(13)进行计算,具体表达为

$$r_{\rm snr}=c_{\rm snr}(\bar{\Upsilon}_{\rm dB}(\boldsymbol{p}_{ld})+\bar{\Upsilon}_{\rm dB}(\boldsymbol{p}_{f1})+\bar{\Upsilon}_{\rm dB}(\boldsymbol{p}_{f2})), \tag{15}$$

c_{ld}、c_{fl}、c_{eg}、$c_{ep_{\rm dot}}$为小于零的常数,$c_{\rm snr}$为大于零的常数,$\boldsymbol{C}$为正定矩阵.

为此,编队航迹规划的整体动作奖励函数为

$$Q(s_t,a_t)=\sum_{i=t}^{t_f}\gamma^{i-t}r_{\rm sum}(s_i,a_i), \tag{16}$$

其中$0<\gamma<1$为衰减系数.

根据式(16),建立贝尔曼方程

$$Q(s_t,a_t)=r_{\text{sum}}(s_t,a_t)+\gamma Q(s_{t+1},a_{t+1}). \tag{17}$$

进而优化问题转化为

$$a_t^*=\arg\max\{Q(s_t,a_t)\}, \tag{18}$$

其中a_t^*为多潜器编队最优策略.

本文采用DDPG算法设计多潜器编队控制算法. 考虑均值为零的OU噪声值Γ,OU过程中的控制输出为$\boldsymbol{a}=\pi(s|\varpi^\pi)+\Gamma$. 具体而言,基于DDPG的多潜器编队控制算法包含4个网络,分别是:当前策略网络$\pi(s|\varpi^\pi)$、目标策略网络$\pi'(s|\varpi^{\pi'})$、当前价值网络$Q(s,a|\varpi^Q)$和目标价值网络$Q'(s,a|\varpi^{Q'})$,其中当前价值网络$Q(s,a|\varpi^Q)$用于拟合整体动作奖励函数$Q(s_t,a_t)$. 为了实现上述目的,首先从经验池中随机采样N条观测数据,然后结合目标价值网络$Q'(s,a|\varpi^{Q'})$,采用梯度下降的方式对当前价值网络$Q(s,a|\varpi^Q)$权重更新,具体方式如下:

$$\varpi_{t+1}^Q=\varpi_t^Q-\ell\nabla_{\varpi^Q}\text{Loss}(\varpi^Q). \tag{19}$$

其中

$$\text{Loss}(\varpi^Q)=\frac{1}{N}\sum\left(y_t-Q(s_t,a_t|\varpi^Q)\right)^2, \tag{20}$$

$$\nabla_{\varpi^Q}\text{Loss}(\varpi^Q)=-\frac{2}{N}\sum\left((y_t-Q(s_t,a_t|\varpi^Q))\frac{\partial Q(s_t,a_t|\varpi^Q)}{\partial\varpi^Q}\right), \tag{21}$$

$$y_t=r_{\text{sum}}(s_t,a_t)+\gamma Q'(s_{t+1},a_{t+1}|\varpi^{Q'}). \tag{22}$$

其中:ϖ_t^Q为t时刻当前价值网络的权重;ℓ为梯度下降的学习率,且$0<\ell<1$.

在当前价值网络$Q(s,a|\varpi^Q)$更新优化的基础上,需要最大化当前价值网络$Q(s,a|\varpi^Q)$的输出值. 为实现上述目的,采用梯度上升法对当前策略网络$\pi(s|\varpi^\pi)$权重进行更新,具体过程如下:

$$\varpi_{t+1}^\pi=\varpi_t^\pi+\bar{\lambda}\nabla_{\varpi^\pi}\boldsymbol{J}(\varpi^\pi). \tag{23}$$

其中

$$\nabla_{\varpi^\pi}\boldsymbol{J}(\varpi^\pi)=\frac{1}{N}\nabla_a Q(s,a|\varpi^Q)\nabla_{\varpi^\pi}\pi(s|\varpi^\pi); \tag{24}$$

$\bar{\lambda}$为梯度上升的学习率且$0<\bar{\lambda}<1$;根据当前策略网络和价值网络的权重更新结果,采用软更新的方式更新权重,即

$$\begin{aligned}\varpi^{\pi'}&=\rho\varpi^\pi+(1-\rho)\varpi^{\pi'},\\ \varpi^{Q'}&=\rho\varpi^Q+(1-\rho)\varpi^{Q'},\end{aligned} \tag{25}$$

ρ为小于1的常数,以调整权重更新速度.

2.3 性能分析

为证明算法收敛性,给出以下引理[14-16].

引理1 随机过程

$$\Delta_{t+1}(s)=(1-o_t(s))\Delta_t(s)+o_t(s)F_t(s).$$

其中:$0\leqslant o_t(s)\leqslant 1,\sum_t o_t(s)=\infty,\sum_t o_t^2(s)<\infty$. 若满足以下条件,则$\{\Delta_t\}$收敛到零:

$$\|E[F_t(s_t,a_t)]\|_\infty\leqslant\gamma\|\Delta_t(s_t,a_t)\|_\infty,\ \gamma<1; \tag{26}$$

$$\text{Var}[F_t(s_t,a_t)]\leqslant K(1+\|\Delta_t(s_t,a_t)\|_\infty^2),\ K>0. \tag{27}$$

引理2 网络结构中,给定经验数据足够多时,若隐藏单元数量足够大,则可以在训练误差线性收敛到零的情况下获得全局最优.

定理1 给定潜器和环境的交互信息,在满足单步奖励函数r_{sum}有界、控制器决策速率远大于环境瞬时变化速率,以及引理1和引理2假设条件的情况下,DDPG算法的动作价值函数逼近器$Q(s,a|\varpi^Q)$逼近于最优奖励累计和.

证明 在式(17)贝尔曼方程基础上,考虑DDPG最优网络输出,可以得到最优动作价值函数$Q^*(s_t,a_t)$,贝尔曼最优方程可以表示为

$$\begin{aligned}&Q(s_{t+1},a_{t+1})=\\&Q(s_t,a_t)+o_t(s)[r_{\text{sum}}(s_t,a_t)+\\&\gamma\max(Q(s_{t+1},a_{t+1}))-Q(s_t,a_t)].\end{aligned} \tag{28}$$

注意到,式(28)可以重新整理为

$$\begin{aligned}&Q(s_{t+1},a_{t+1})=\\&(1-o_t(s))Q(s_t,a_t)+o_t(s)[r_{\text{sum}}(s_t,a_t)+\\&\gamma\max(Q(s_{t+1},a_{t+1}))].\end{aligned} \tag{29}$$

将式(29)两边同时减去$Q^*(s_t,a_t)$,可得

$$\begin{aligned}&\Delta_{t+1}(s_t,a_t)=\\&(1-o_t(s))\Delta_t(s_t,a_t)+o_t(s)[r_{\text{sum}}(s_{t+1},a_{t+1})+\\&\gamma\max Q(s_{t+1},a_{t+1})-Q^*(s_t,a_t)],\end{aligned} \tag{30}$$

其中$\Delta_{t+1}(s_t,a_t)=Q(s_{t+1},a_{t+1})-Q^*(s_t,a_t)$. 同样的,$\Delta_t(s_t,a_t)=Q(s_t,a_t)-Q^*(s_t,a_t)$.

令$F_t(s_t,a_t)=r_{\text{sum}}(s_t,a_t)+\gamma\max Q(s_{t+1},a_{t+1})-Q^*(s_t,a_t)$,则$F_t(s_t,a_t)$的期望计算如下:

$$\begin{aligned}&E(F_t(s_t,a_t))=\\&E(r_{\text{sum}}(s_t,a_t)+\gamma\max Q(s_{t+1},a_{t+1})-Q^*(s_t,a_t))=\\&\sum\wp_a(s_t,s_{t+1})(r_{\text{sum}}(s_t,a_t)+\\&\gamma\max Q(s_{t+1},a_{t+1})-Q^*(s_t,a_t))=\\&(\boldsymbol{G}Q_t)(s,a)-(\boldsymbol{G}Q^*)(s_t,a_t).\end{aligned} \tag{31}$$

其中: $\wp_a(s_t, s_{t+1})$ 为输出动作的状态转移概率, $\boldsymbol{G}$ 为 $Q(s_t, a_t)$ 的不动点算子. 由于此不动点算子超范数收敛[14], 式(31)的期望值可化简为

$$\|E[F_t(s_t, a_t)]\|_\infty \leqslant \gamma \|\Delta_t(s_t, a_t)\|_\infty, \tag{32}$$

由式(32)可知, 所提出的编队控制算法满足引理 1 中的条件 1, 即式(26).

其次, 通过计算 $F_t(s_t, a_t)$ 的方差证明所提出控制算法可以满足引理 1 的第 2 个前提条件, 即式(27). 基于此思路, $F_t(s_t, a_t)$ 方差计算为

$$\begin{aligned}&\mathrm{Var}[F_t(s_t, a_t)] = [(F_t(s_t, a_t) - \|E[F_t(s_t, a_t)]\|)^2] = \\ &\mathrm{Var}[r_{\mathrm{sum}}(s_t, a_t) + \gamma \max Q(s_{t+1}, a_{t+1})].\end{aligned} \tag{33}$$

由于 $r_{\mathrm{sum}}(s_t, a_t)$ 有界, 式(33)可整理为

$$\mathrm{Var}[F_t(s_t, a_t)] \leqslant K(1 + \|\Delta_t(s_t, a_t)\|_\infty^2). \tag{34}$$

根据引理 1, 式(32)中 $\Delta_t(s_t, a_t)$ 依概率收敛到零, $Q(s_t, a_t)$ 收敛到最优值. 同时, 本文潜器与环境交互数据的经验池数据量以及网络隐藏神经元的个数满足引理 2 的条件, 因此本文 DDPG 控制算法($o_t(s) = 1$)中的价值网络 $Q(s, a|\varpi^Q)$ 可以保证拟合动作价值 $Q^*(s_t, a_t)$, 并收敛于全局最优, 此时潜器可收敛于期望位置. □

3　仿真及实验研究

设定潜器初始位置为 $\boldsymbol{p}_{ld} = [10, 20, -15]^{\mathrm{T}}$, $\boldsymbol{p}_{f1} = [7, 22, -15]^{\mathrm{T}}$, $\boldsymbol{p}_{f2} = [13, 14, -8]^{\mathrm{T}}$, 其目标点位置分别为 $\boldsymbol{p}_{ld\_g} = [30, 2, -5]^{\mathrm{T}}$, $\boldsymbol{p}_{f1\_g} = [28.59, 2, -6.414]^{\mathrm{T}}$, $\boldsymbol{p}_{f2\_g} = [30, 0.59, -6.414]^{\mathrm{T}}$, 移动感知的潜标位置为 $\boldsymbol{p}_{\mathrm{station}} = [10, 8, -21]^{\mathrm{T}}$.

3.1　水声信道预测仿真

设置真实的信道参数值分别为 $\theta = [-43, 3]^{\mathrm{T}}$, $\alpha = 1.2$ 和 $\mu = 1.15$, 基于此信道参数, 采用所提出信道参数估计算法, 得到估计参数值见表 1. 由表 1 可知, 估计参数与真实值十分接近, 验证了信道估计器的有效性. 完成信道估计后, 范围内的信道预测结果如图 1 所示.

表 1　信道参数估计值

变量名称	变量估计值	变量名称	变量估计值
$\hat{\theta}_{\mathrm{LS}}$	[−43.133 5; 3.048 2]	$\hat{\beta}_{\mathrm{LS}}$	18.029 3
$\hat{\alpha}_{\mathrm{LS}}$	1.001 7	$\hat{\mu}^2_{\mathrm{LS}}$	2.324 4

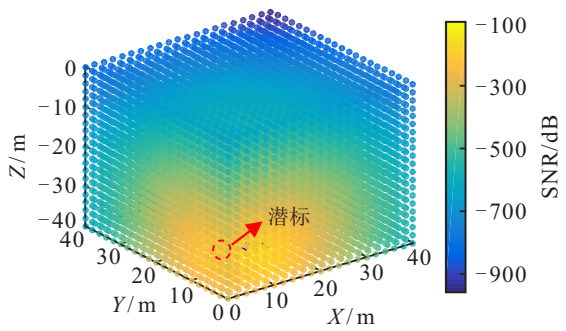


图 1　水声信道分布情况

3.2　多潜器编队控制仿真

设置 3 个潜器的期望相对队形为直角三角形, 使用 256 × 256 的全连接层网络, 动作网络和目标网络学习率分别为 0.000 5、0.001, 随机采样数量 $N = 256$, 采用 $1e^6$ 容量的经验池存放数据训练网络. 经过 2 500 次回合的训练, 编队效果如图 2 所示. 可以看出, 潜器可收敛到期望位置. 图 3 为 DDPG 控制算法在训练过程中的奖励变化趋势.

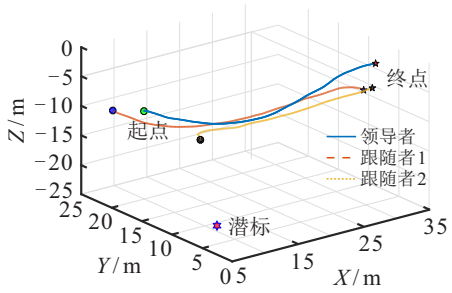


图 2　场景 1-编队控制轨迹

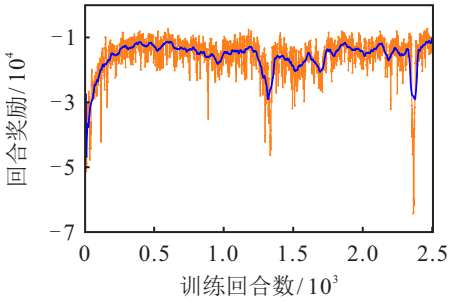


图 3　DDPG 训练奖励收敛过程

考虑以下 3 个场景: 场景 0, 在编队控制器中考虑 SNR 较大权重的情况; 场景 1, 考虑较小权重 SNR; 场景 2, 不考虑 SNR 信息, 如文献[17]方法. 场景 2 编队轨迹如图 4 所示, 场景 1 与场景 2 编队过程中 SNR 均值比较见图 5. 场景 1 比场景 2 在通信性能上提升了 13.5 %, 虽然场景 2 中的编队径直驶向目标点, 但是信道质量难以保证, 实际情况中往往直接导致潜器通信失效, 不能实现期望的编队效果, 因此所提出协同框架更具实际意义. 对比场景 0 的编队控制效果见图 6. 场景 0 考虑了较大权重的 SNR, 潜器在后续的编队控制中向潜标靠近, 导致编队控制效果失败. 根据以上仿真分析可知, 由于潜器所处环境不同, 编队和通信的权重需要根据实际情况进行权衡, 本文主要考虑了通信对编队控制的影响, 未来的工作中将考虑如何合理权衡编队与通信感知问题.

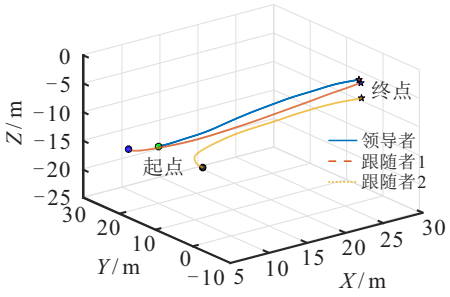


图 4 场景 2-编队控制轨迹

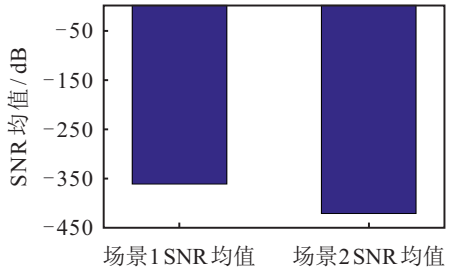


图 5 场景 1 与场景 2 的信道质量比较

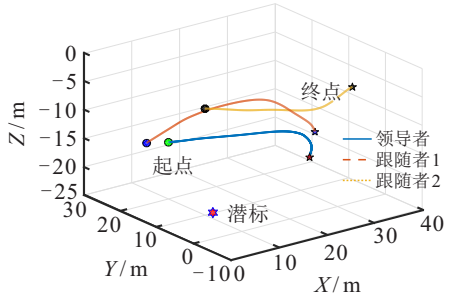


图 6 场景 0-编队控制轨迹

3.3 多潜器编队实验

本节给出编队控制算法实验结果. 由于实验条件有限,仅验证机器人的控制算法部分. 实验平台硬件条件如图 7 所示. 潜器位置部分由 UWB (ultra wide band) 系统获得,姿态信息由潜器的陀螺仪获得,同时岸边的控制中心负责计算控制输入,通过匿名无线传输模块传送给潜器并驱动其运动,上位机负责实时记录实验数据. 为了方便实验,给定领导者和两个跟随者的初始位置分别为 $\boldsymbol{p}_{ld} = [2.206, 1.333, 0]^{\mathrm{T}}, \boldsymbol{p}_{f1} = [2.19, 1.356, 0]^{\mathrm{T}}, \boldsymbol{p}_{f2} = [3.91, 7.272, 0]^{\mathrm{T}}$; 目标点为 $\boldsymbol{p}_{ld\_g} = [10, 10, 0]^{\mathrm{T}}, \boldsymbol{p}_{f1\_g} = [9.239, 10.707, 0]^{\mathrm{T}}, \boldsymbol{p}_{f2\_g} = [10.707, 10.707, 0]^{\mathrm{T}}$.

图 7 硬件系统

在给定初始点和目标点的情况下,多潜器轨迹如图 8 所示. 与目标位置的距离误差如图 9 所示. 参考图 8 和图 9,在水流以及其他实验误差的影响下,跟随者 2 在 37 s 时已经接近了目标点,但领导者和跟随者 1 距离目标点还有一定距离,由于编队队形的约束,跟随者 2 调节自身位置向领导者靠近,跟随者 2 在 37 ~ 60 s 有远离目标位置的趋势. 到 60 s 时,领导者和跟随者 1 已经接近到目标点,此时跟随者 2 受到编队队形和目标位置的约束,开始减小距离误差靠近目标点. 同时可以看出在 60 ~ 100 s 时间段内,领导者和跟随者 1 也受到跟随者 2 的影响,有一定趋势远离目标点,最终在目标位置和编队约束下逐渐形成编队队形. 实验过程中,3 个潜器的控制过程不仅受到环境的扰动,同时受到编队队形的约束,因此潜器与目标点的距离误差是一个动态调节的过程. 忽略实验环境以及设备的测量误差,在容许的误差范围内,实验结果表明所设计的控制算法可以用于多潜器编队控制. 实验视频参见如下链接: https://v.youku.com/v_show/id_XNTg5MTM3OTk4NA.

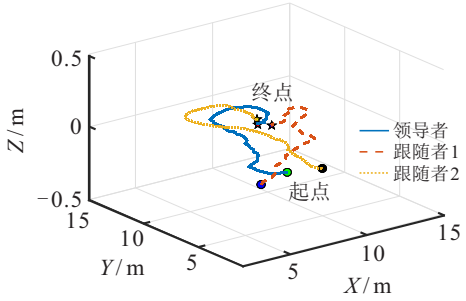
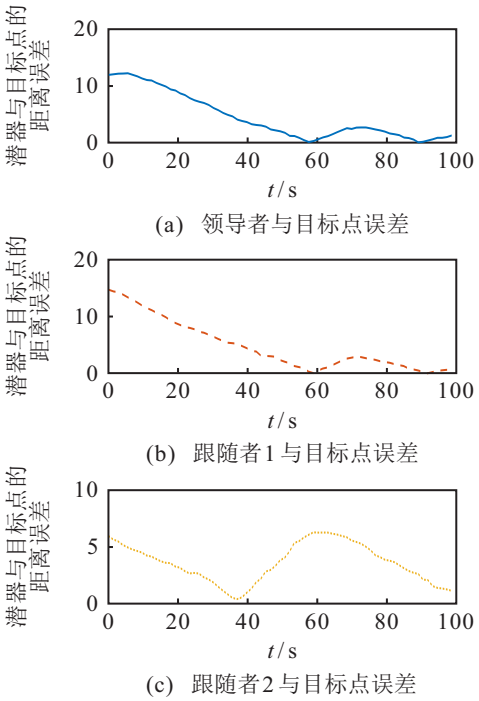


图 8 实验环境中 AUVs 的轨迹



(a) 领导者与目标点误差

(b) 跟随者 1 与目标点误差

(c) 跟随者 2 与目标点误差

图 9 AUVs 与目标点的距离误差

4 结 论

本文设计了基于深度强化学习的多潜器编队控制算法,同时考虑了水声通信质量对编队控制的影响. 所提出的水下信道预测和编队控制协同优化框架可实现通信有效性与编队稳定性,并通过仿真实验验证了所提出算法的有效性. 未来将研究信道预测下最优编队构型问题,同时将所提出控制算法用于真实的海洋环境中.

参考文献(References)

[1] 陈铭治, 朱大奇. FMM与改进GBNN模型相结合的多AUV实时围捕算法[J]. 控制与决策, 2020, 35(12): 2845-2854.
(Chen M Z, Zhu D Q. Multi-AUV real-time hunting control based on FMM and improved GBNN model[J]. Control and Decision, 2020, 35(12): 2845-2854.)

[2] 陈子印, 王宏健, 边信黔, 等. 基于反馈增益的AUV稳定神经网络反步变深控制[J]. 控制与决策, 2013, 28(3): 407-412.
(Chen Z Y, Wang H J, Bian X Q, et al. Stable neural network backstepping for diving control of AUV based on feedback gain[J]. Control and Decision, 2013, 28(3): 407-412.)

[3] Liu H, Wang Y H, Lewis F L. Robust distributed formation controller design for a group of unmanned underwater vehicles[J]. IEEE Transactions on Systems, Man, and Cybernetics: Systems, 2021, 51(2): 1215-1223.

[4] Gao Z Y, Guo G. Fixed-time sliding mode formation control of AUVs based on a disturbance observer[J]. IEEE/CAA Journal of Automatica Sinica, 2020, 7(2): 539-545.

[5] Millán P, Orihuela L, Jurado I, et al. Formation control of autonomous underwater vehicles subject to communication delays[J]. IEEE Transactions on Control Systems Technology, 2014, 22(2): 770-777.

[6] Wang J, Wang C, Wei Y, et al. Filter-backstepping based neural adaptive formation control of leader-following multiple AUVs in three dimensional space[J]. Ocean Engineering, 2020, 201: 107150.

[7] Yuan C Z, Licht S, He H B. Formation learning control of multiple autonomous underwater vehicles with heterogeneous nonlinear uncertain dynamics[J]. IEEE Transactions on Cybernetics, 2018, 48(10): 2920-2934.

[8] Gu N, Wang D, Peng Z H, et al. Model-free containment control of underactuated surface vessels under switching topologies based on guiding vector fields and data-driven neural predictors[J]. IEEE Transactions on Cybernetics, 2022, 52(10): 10843-10854.

[9] Wang C C, Cai W Y, Lu J, et al. Design, modeling, control, and experiments for multiple AUVs formation[J]. IEEE Transactions on Automation Science and Engineering, 2022, 19(4): 2776-2787.

[10] Yan J, Ban H J, Luo X Y, et al. Joint localization and tracking design for AUV with asynchronous clocks and state disturbances[J]. IEEE Transactions on Vehicular Technology, 2019, 68(5): 4707-4720.

[11] Zhang Y, Zhang Z M, Chen L, et al. Reinforcement learning-based opportunistic routing protocol for underwater acoustic sensor networks[J]. IEEE Transactions on Vehicular Technology, 2021, 70(3): 2756-2770.

[12] Cotton S L, Scanlon W G. Higher order statistics for lognormal small-scale fading in mobile radio channels[J]. IEEE Antennas and Wireless Propagation Letters, 2007, 6: 540-543.

[13] Gudmundson M. Correlation model for shadow fading in mobile radio systems[J]. Electronics Letters, 1991, 27(23): 2145.

[14] Melo F S, Ribeiro M I. *Q*-learning with linear function approximation[C]. International Conference on Computational Learning Theory. Berlin, 2007: 308-322.

[15] Jaakkola T, Jordan M I, Singh S P. On the convergence of stochastic iterative dynamic programming algorithms[J]. Neural Computation, 1994, 6(6): 1185-1201.

[16] Du S S, Zhai X, Poczos B, et al. Gradient descent provably optimizes over-parameterized neural networks[C]. IEEE International Conference on Learning Representations. New Orleans, 2019: 1-19.

[17] Zhao Y J, Ma Y, Hu S L. USV formation and path-following control via deep reinforcement learning with random braking[J]. IEEE Transactions on Neural Networks and Learning Systems, 2021, 32(12): 5468-5478.

作者简介

闫敬(1985−), 男, 教授, 博士生导师, 从事水下网络系统感知、组网与控制等研究, Email: jyan@ysu.edu.cn;

徐龙(1994−), 男, 硕士生, 从事深度强化学习下潜器编队控制系统的研究, E-mail: xl@stumail.ysu.edu.cn;

曹文强(1998−), 男, 博士生, 从事潜器组网与编队控制的研究, E-mail: cwq@stumail.ysu.edu.cn;

杨晛(1988−), 女, 副教授, 博士生导师, 从事多潜器协同控制等研究, E-mail: xyang@ysu.edu.cn;

罗小元(1976−), 男, 教授, 博士生导师, 从事多智能体协同控制与智能电网等研究, E-mail: xyluo@ysu.edu.cn.

(责任编辑：郑晓蕾)