Communication-Efficient and Collision-Free Motion Planning of Underwater Vehicles via Integral Reinforcement Learning

Jing Yan, *Senior Member, IEEE*, Wenqiang Cao, Xian Yang, Cailian Chen, *Member, IEEE*, and Xinping Guan, *Fellow, IEEE*

***Abstract*— Motion planning of underwater vehicles is regarded as a promising technique to make up the flexibility deficiency of underwater sensor networks (USNs). Nonetheless, the unique characteristics of underwater channel and environment make it challenging to achieve the above mission. This article is concerned with a communication-efficient and collision-free motion planning issue for underwater vehicles in fading channel and obstacle environment. We first develop a model-based integral reinforcement learning (IRL) estimator to predict the stochastic signal-to-noise ratio (SNR). With the estimated SNR, an integrated optimization problem for the codesign of communication efficiency and motion planning is constructed, in which the underwater vehicle dynamics, communication capacity, collision avoidance, and position control are all considered. In order to tackle this problem, a model-free IRL algorithm is designed to drive underwater vehicles to the desired position points while maximizing the communication capacity and avoiding the collision. It is worth mentioning that, the proposed motion planning solution in this article considers a realistic underwater communication channel, as well as a realistic dynamic model for underwater vehicles. Finally, simulation and experimental results are demonstrated to verify the effectiveness of the proposed approach.**

***Index Terms*— Collision-free, communication-efficient, motion planning, reinforcement learning, underwater vehicle.**

I. Introduction

IN ORDER to understand and explore the ocean, many underwater sensor nodes, including multibeam swath bathymeter, sonar array, and acoustic Doppler current profiler, have been deployed to form the underwater sensor networks (USNs) [1], [2]. The deployment of USNs can increase the space-time cover ability of ocean monitoring; however, USNs lack the necessary flexibility and autonomy, which cannot deal with highly dynamic uncertainties in complex underwater environment. With regard to this, underwater vehicle-assisted USNs have been emerged as a new promising communication platform in future ocean-observation systems, due to the high mobility, controllable maneuver, and on-demand deployment. These appealing advantages have enabled various applications, including intrusion surveillance, data gathering, geographic mapping, petroleum exploration, and transmission of images from remote sites (see [3], [4], [5] and references therein).

In underwater vehicle-assisted USNs, one of the most critical issues is to plan paths for underwater vehicles. For instance, an energy-efficient motion planning strategy was provided in [6] to balance the communication energy consumption and prolong the network lifetime. Yetkin et al. [7] incorporated the environment information into the path planning of underwater vehicles, through which a decision-theoretic-based subsea search algorithm was designed. In [8], the end-to-end data freshness constraint was conducted to determine the paths of underwater vehicles, whose aim was to retrieve the collected data to control center as soon as possible. Followed by this, a heuristic algorithm was provided in [9] to optimize the paths of underwater vehicles, with respect to data quality and underwater coverage efficiency. Wang et al. [10] employed acoustic camera to capture the position and shape of unknown underwater pipelines. These schemes are well developed; however, they do not take the collision avoidance into consideration. As we have seen already, obstacles such as wrecks and plankton inevitably exist in water, while at the same time the collision between vehicles may occur when they work together. The above collision constraint has a strong impact on the motion safety and communication channel of underwater vehicles [11]. Therefore, it is necessary to plan a collision-free path for each underwater vehicle.

To resolve the above problem, Song et al. [12] developed a joint flocking and guidance scheme for underwater vehicles evolving in environments with obstacles. In [13], an artificial potential-based motion planning strategy was conducted, where the software-defined technology was employed to improve the scalability and controllability. A multilayered motion planning scheme was presented in [14] for underwater navigation, wherein a local motion planner was employed to avoid collision with obstacles. Note that the dynamics models

Manuscript received 1 August 2022; revised 11 November 2022; accepted 28 November 2022. This work was supported in part by the National Natural Science Foundation of China under Grant 62222314, Grant 61973263, Grant 61873345, and Grant 62033011; in part by the Youth Talent Program of Hebei under Grant BJ2020031; in part by the Distinguished Young Foundation of Hebei Province under Grant F2022203001; in part by the Central Guidance Local Foundation of Hebei Province under Grant 226Z3201G; and in part by the Three-Three-Three Foundation of Hebei Province under Grant C20221019. *(Corresponding author: Jing Yan.)*

Jing Yan, Wenqiang Cao, and Xian Yang are with the Institute of Electrical Engineering, Yanshan University, Qinhuangdao 066004, China (e-mail: jyan@ysu.edu.cn; cwq@stumail.ysu.edu.cn; xyang@ysu.edu.cn).

Cailian Chen and Xinping Guan are with the School of Electronic Information and Electrical Engineering, Shanghai JiaoTong University, Shanghai 200240, China (e-mail: cailianchen@sjtu.edu.cn; xpguan@sjtu.edu.cn).

Color versions of one or more figures in this article are available at https://doi.org/10.1109/TNNLS.2022.3226776.

Digital Object Identifier 10.1109/TNNLS.2022.3226776

of underwater vehicles in [12], [13], and [14] treat each vehicle as a point mass, such that a second-order nonlinear equation can be conducted to describe the kinematics of underwater vehicles. However, the kinematics model cannot capture the low-level interactions during the implementation of motion planning on actual vehicle. As such, Heshmati-Alamdari et al. [15] jointly considered the kinematics and dynamics models of underwater vehicles, through which a model predictive controller was developed to steer each underwater vehicle to the desired trajectory with collision avoidance. In [16], an adaptive motion controller was provided to achieve finite-time formation control and obstacle avoidance for underwater vehicles. Nevertheless, the above motion planning schemes rely on full or partial knowledge of underwater vehicle dynamic model. Due to the harsh ocean conditions, it is difficult if not impossible to acquire the accurate dynamics model of underwater vehicles. With regard to this, an artificial potential function-based motion planning scheme was developed in [17] to relax the dependence of model parameters for underwater vehicles. Also of relevance, the Astar algorithms were designed in [18] and [19] to steer underwater vehicles to the target points with collision avoidance. Although the artificial potential function and Astar algorithms are simple to implement, they are easy to get stuck at locally optimal value. More recently, Jiang et al. [20] and Kontoudis and Vamvoudakis [21] employed the distributed learning algorithms to reduce the dependency on model parameters and achieve global optimization; however they are not developed in the context of collision-free motion planning for underwater vehicles. Due to the complex dynamics of underwater vehicles, the issue of how to adopt the learning strategy to design a collision-free motion planning scheme without relying on models for underwater vehicles is largely unexplored.

Apart from that, most of the existing motion planning schemes focus on the control techniques and they ignore the influence of underwater acoustic communication channel. To be specific, they assume that the channel quality is approximated by a deterministic disk model. The above assumption is reasonable for terrestrial vehicles; however, it is not valid for underwater vehicles. As has been pointed out in [22] and [23], underwater acoustic communication suffers from stronger shadowing and multipath fading than the terrestrial radio wave communication. Ignoring the shadowing and multipath fading factors may result in the deterioration of communication quality during the motion planning process. Thereby, we need to incorporate the underwater communication quality into the motion planning procedure metric, such that a communication-efficient and collision-free motion planning scheme can be developed to improve the communication capacity via the control feedback of underwater vehicles. The above idea is similar to the codesign of estimation and communication for multiagent systems, e.g., [24], [25]. To this end, we notice that some communication-efficient motion planning schemes have been developed for terrestrial vehicles. For instance, a gradient estimation-based motion controller was developed in [26] to optimize the communication chain. In [27] and [28], two codesign frameworks for aerial vehicle motion planning and communication efficiency were constructed. Yan and Mostofi [29] integrated the probabilistic signal-to-noise ratio (SNR) prediction approach into the router planning of robots. Followed by this, Ali et al. [30] extended the first-order linear kinematic model of vehicles to the second-order, through which a motion-communication cooptimization solution was designed. Note that the dynamics model of vehicles in [26], [27], [28], [29], and [30] is reduced to a first-order or second-order kinematic equation; however, it cannot capture the actual dynamics model of underwater vehicles. In [31], the kinematic and dynamics models of robots were incorporated into the communication-aware motion planning. Nonetheless, the least square estimators are developed in the above literatures to seek the SNR parameters, which are easy to trap in local optimum. To compensate these shortcomings, the distributed learning can offer us with a feasible solution, since it seeks a global optimum solution via online learning and iteration [32]. Our previous works [33], [34] employed the learning algorithm to solve the underwater localization problem. Nevertheless, how to develop a learning-based solution that can jointly solve collision-free motion planning and global-optimum SNR estimation for underwater vehicles is still an unsolved issue.

This article studies a communication-efficient and collision-free motion planning problem for underwater vehicles in fading channel and obstacle environment. A novel two-stage solution is developed, i.e., sensor nodes predict the SNR parameters in the first stage, and underwater vehicles dynamically adjust their positions in the second stage. In such a solution, underwater vehicles behave as mobile communication relaying nodes whose aim is to improve the communication capacity. Main contributions of this article lie in three aspects.

1) *Integrated Optimization Framework for the Codesign of Communication Efficiency and Motion Planning:* We develop an integrated optimization framework, including underwater vehicle dynamics, communication capacity, collision avoidance, and position control. As far as we know, this is the first integrated optimization framework for the communication-efficient and collision-free motion planning of underwater vehicles that involves realistic communication channel and dynamic model. It connects the communication capacity of sensor nodes with the motion planning of underwater vehicles, which can improve the communication capacity.
2) *Model-Based Learning Estimator for Online SNR Prediction:* A model-based integral reinforcement learning (IRL) estimator is designed to predict the SNR of underwater vehicles. It can effectively predict the probabilistic SNR parameters with limited channel knowledge. Compared with least squares or maximum likelihood-based estimators, e.g., [29], [30], [31], the IRL estimator in this article can avoid local minimum. Meanwhile, the shadowing and multipath fading effects are considered in this article, which are ignored by terrestrial vehicles, e.g., [26].
3) *Model-Free Learning Algorithm for Collision-Free Motion Planning:* With the predicted SNR information, a model-free IRL algorithm is developed to steer underwater vehicles to the desired position points while avoiding collision with obstacles and the other vehicles. Different from the motion planning controllers in [12], [13], and [14], the developed motion planning algorithm in this article not only considers the collision

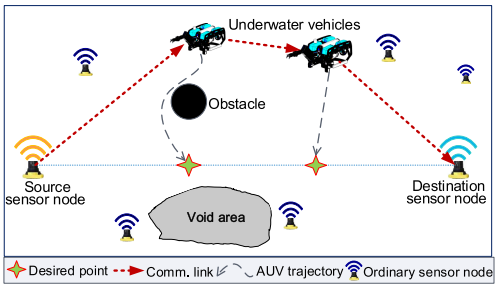


Fig. 1. Communication links from source sensor node to destination sensor node through the motion relays of underwater vehicles.

avoidance, but also takes into account the dynamics model of underwater vehicles. Compared with the solutions in [15] and [16], it relaxes the dependency on vehicle model parameters.

II. System Model and Problem Formulation

We would like to transmit the data information from source sensor node to destination sensor node, as shown in Fig. 1. To this end, the following four types of nodes are provided.

1) *Source Sensor Node:* The position of source sensor node is fixed, and it sends the collected data to the destination sensor node through the relays of underwater vehicles.
2) *Destination Sensor Node:* The position of destination sensor node is also fixed, whose objective is to indirectly receive the data from source sensor node.
3) *Underwater Vehicle:* Underwater vehicles act as mobile communication relaying nodes.
4) *Ordinary Sensor Node:* Ordinary sensor nodes are provided to sense and collect channel measurements to underwater vehicles, which do not undertake the relay task.

On the basis of the above framework, we employ a team of ordinary sensor nodes to predict the stochastic SNR parameters. In view of this, let $\mathbf{p}_s = \{\mathbf{p}_{s,1}, \ldots, \mathbf{p}_{s,N}\}$ be the team position set of ordinary sensor nodes and $\mathbf{p}_{s,i} = [x_{s,i}, y_{s,i}, z_{s,i}]^{\mathrm{T}}$ be the position vector of ordinary sensor node $i \in \mathcal{V}_s = \{1, \ldots, N\}$, where $x_{s,i}$, $y_{s,i}$, and $z_{s,i}$ are the positions on X-, Y-, and Z-axis, respectively. Let $\mathcal{E} = \{R_0, R_1, \ldots, R_{M+1}\}$ be the end-to-end communication chain, where R_0 denotes the source sensor node, R_{M+1} denotes the destination sensor node, and the others are the underwater vehicles. Specifically, $\mathcal{V} = \{R_1, \ldots, R_M\}$ represents the set of underwater vehicles, where each underwater vehicle R_i relays data from its single source neighbor R_{i-1} to its single destination neighbor R_{i+1}. For underwater vehicle R_i, its position vector can be defined as $\mathbf{p}_i = [x_i, y_i, z_i]^{\mathrm{T}}$, while the position vectors of source senor node and destination senor node are defined as $\mathbf{p}_0 = [x_0, y_0, z_0]^{\mathrm{T}}$ and $\mathbf{p}_{M+1} = [x_{M+1}, y_{M+1}, z_{M+1}]^{\mathrm{T}}$, respectively. Besides that, N and M are the total numbers of ordinary sensor node and underwater vehicles.

The inertial reference frame (IRF) and body-fixed reference frame (BRF) are jointly utilized to depict the dynamic model of underwater vehicles. The position and orientation vector for underwater vehicle $R_i \in \mathcal{V}$ in IRF is defined as $\boldsymbol{\eta}_i = [\mathbf{p}_i; \psi_i]$, where ψ_i is the angle on yaw. The linear and angle velocity vector in BRF is $\mathbf{v}_i = [u_i, v_i, w_i, r_i]^{\mathrm{T}}$, where u_i, v_i, and w_i are the linear velocities on surge, sway, and heave, respectively. In addition, r_i is the angle velocity on yaw. From [35], [36], the dynamic model of underwater vehicle R_i is

$$\begin{aligned} \dot{\boldsymbol{\eta}}_i &= \mathbf{J}_i(\boldsymbol{\eta}_i)\mathbf{v}_i \\ \mathbf{M}_i\dot{\mathbf{v}}_i + \mathbf{C}_i(\mathbf{v}_i)\mathbf{v}_i + \mathbf{D}_i(\mathbf{v}_i)\mathbf{v}_i + \mathbf{g}_i(\boldsymbol{\eta}_i) &= \boldsymbol{\tau}_i \end{aligned} \tag{1}$$

where $\mathbf{M}_i \in \mathcal{R}^{4\times4}$, $\mathbf{C}(\mathbf{v}_i) \in \mathcal{R}^{4\times4}$ and $\mathbf{D}(\mathbf{v}_i) \in \mathcal{R}^{4\times4}$ are the inertia, Coriolis-centripetal, and damping matrices, respectively. $\mathbf{J}_i(\boldsymbol{\eta}_i) \in \mathcal{R}^{4\times4}$ is the rotation matrix, $\mathbf{g}_i(\boldsymbol{\eta}_i) \in \mathcal{R}^4$ is the hydrostatic force, and $\boldsymbol{\tau}_i = [\tau_{u_i}, \tau_{v_i}, \tau_{w_i}, \tau_{r_i}]^{\mathrm{T}}$ is the control input, where $\tau_{u_i}, \tau_{v_i}, \tau_{w_i}$, and τ_{r_i} are the control forces on surge, sway, heave, and yaw, respectively.

Define $\mathbf{X}_i = [\boldsymbol{\eta}_i; \mathbf{v}_i]$, and hence, model (1) is rearranged as

$$\dot{\mathbf{X}}_i = \mathbf{A}_i\mathbf{X}_i + \mathbf{B}_i\boldsymbol{\tau}_i + \mathbf{G}_i \tag{2}$$

with

$$\mathbf{A}_i = \begin{bmatrix} \mathbf{0} & \mathbf{J}_i(\boldsymbol{\eta}_i) \\ \mathbf{0} & -\mathbf{M}_i^{-1}(\mathbf{C}_i(\mathbf{v}_i)+\mathbf{D}_i(\mathbf{v}_i)) \end{bmatrix} \tag{3}$$

$$\mathbf{B}_i = \begin{bmatrix} \mathbf{0} \\ \mathbf{M}_i^{-1} \end{bmatrix}, \quad \mathbf{G}_i = \begin{bmatrix} \mathbf{0} \\ -\mathbf{M}_i^{-1}\mathbf{g}_i(\boldsymbol{\eta}_i) \end{bmatrix}. \tag{4}$$

In order to improve the communication capacity, the link capacity $\mathcal{C}_i$ of underwater vehicle R_i is introduced. Referring to [37], one knows the link capacity of an end-to-end communication chain is equal to the capacity of the worst link. Along with this, the link capacity $\mathcal{C}_i$ is defined as

$$\mathcal{C}_i = \min\{c_{i-1,i}, c_{i,i+1}\} \tag{5}$$

where $c_{i-1,i}$ is the link capacity between source neighbor R_{i-1} and vehicle R_i. Meanwhile, $c_{i,i+1}$ is the link capacity between vehicle R_i and destination neighbor R_{i+1}.

Note that the link capacity is a function of bandwidth and communication quality SNR [37]. Hence, the SNR between source neighbor R_{i-1} and vehicle R_i is expressed as

$$\begin{aligned} &\mathrm{SNR}_{\mathrm{dB}}(\mathbf{p}_{i-1}, \mathbf{p}_i) \\ &\quad = K_{\mathrm{dB}} - 10n_{\mathrm{PL}}\log_{10}(l_{i-1,i}) - 10l_{i-1,i}\log_{10}(\alpha(f)) \\ &\qquad - N_{\mathrm{dB}}^0 + \sigma_{\mathrm{SH}}(\mathbf{p}_{i-1}, \mathbf{p}_i) + \mu_{\mathrm{MP}}(\mathbf{p}_{i-1}, \mathbf{p}_i) \end{aligned} \tag{6}$$

where $10\log_{10}\alpha(f) = ((0.11f^2)/(1+f^2)) + (44f^2/(4100+f^2)) + 2.75\times10^{-4}f^2 + 0.003$. In addition, K_{dB} denotes the average energy consumption of transmitting 1 bit data in dB, n_{PL} denotes the spreading coefficient, $l_{i-1,i} = \|\mathbf{p}_{i-1} - \mathbf{p}_i\|$ denotes the relative distance between source neighbor R_{i-1} and underwater vehicle R_i, N_{dB}^0 denotes the noise power spectral density in dB, and $\alpha(f)$ is the acoustic absorption with frequency f. Moreover, $\sigma_{\mathrm{SH}}(\mathbf{p}_{i-1}, \mathbf{p}_i)$ and $\mu_{\mathrm{MP}}(\mathbf{p}_{i-1}, \mathbf{p}_i)$ represent the location-related stochastic parameters, which reflect the effects of shadowing and multipath fading, respectively.

Remark 1: Different from the simplified SNR models in [38] and [39], the shadow fading parameter σ_{SH} and the multipath parameter μ_{MP} are both considered in this article, which can well capture the realistic underwater environment. An example of the shadowing and multipath fading is shown in Fig. 2.

From (6) and noting with the Shannon–Hartley theorem [26], the link capacity $c_{i-1,i}$ which provides the theoretical upper bound can be obtained as

$$c_{i-1,i} = B\log_2(1 + S_{i-1,i}) \tag{7}$$

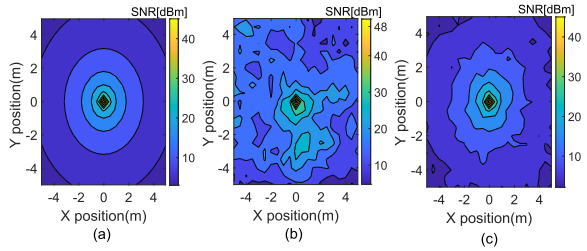


Fig. 2. (a) Only the path loss is considered. (b) Path loss and shadowing are considered. (c) Path loss and multipath fading are considered.

with

$$S_{i-1,i} = 10^{\frac{\mathrm{SNR_{dB}}(\mathbf{p}_{i-1},\mathbf{p}_i)}{10}} \tag{8}$$

where B denotes the communication bandwidth. Similarly, the detailed expression of $c_{i,i+1}$ can also be acquired.

Assumption 1: The obstacles in underwater environment can be covered by convex cylinders. The mth $(m = 1, 2, \ldots)$ obstacle in the environment is denoted as $B_m(O_m, \rho_m)$, where O_m is its center and $\rho_m(\rho_m > 0)$ is its radius.

Definition 1 (Obstacle Set): For underwater vehicle R_i at time t, the detected obstacle set can be defined as the subset $\Omega_i^t \subset \{B_1(O_1, \rho_1), \ldots, B_m(O_m, \rho_m), \ldots\}$ in the detected range of underwater vehicle $R_i \in \mathcal{V}$.

Assumption 2: The obstacles are sparsely exist in the underwater environment, and hence, the impact of obstacles on communication channel is ignored. Of note, this assumption has been made in some existing works, e.g., [29], [40].

Assumption 3: The offline map of the target area is to be known by the sonar installed on underwater vehicle. Meanwhile, underwater vehicle is equipped with camera, which is capable of online detecting obstacles within a certain range.

Accordingly, the following two problems are formulated.

Problem 1 (SNR Prediction in Fading Channel): The shadow fading and multipath parameters in SNR model cannot be obtained previously. In view of this, we attempt to design a model-based IRL estimator to capture the unknown channel parameters. This problem can be reduced to the estimation of K_{dB}, n_{PL}, σ_{SH} and μ_{MP} with the limited channel measurements from ordinary sensor nodes $i \in \{1, \ldots, N\}$.

Problem 2 (Collision-Free Motion Planning): It is impossible to acquire the accurate dynamic model of underwater vehicles. Meanwhile, obstacles increase the difficulty of motion planning of underwater vehicles. In view of this, we aim to employ IRL to develop a model-free and collision-free motion planning algorithm. This problem is reduced to maximize $\mathcal{C}_i$ while guaranteeing $\|\mathbf{p}_i - O_m\| > \rho_m$ and $\|\mathbf{p}_i - \mathbf{p}_j\| > 0$.

III. Main Results

We first design an IRL-based estimator to capture the unknown shadowing and multipath parameters. Along with this, the IRL is adopted to develop a model-free and collision-free motion planning algorithm for underwater vehicles. Finally, the theoretical analysis for our solution is presented.

A. IRL-Based Estimator for Online SNR Prediction

Initially, underwater vehicle $R_i \in \mathcal{V}$ at location $\mathbf{p}_i$ broadcasts an initiator message to its neighboring ordinary sensor nodes. Then, underwater vehicle R_i switches into the listening mode. For any ordinary sensor node $j \in \mathcal{N}_i$, it senses the SNR of underwater vehicle R_i, denoted by $\mathrm{SNR_{dB}}(\mathbf{p}_i, \mathbf{p}_{s,j})$ where $\mathcal{N}_i$ is the neighboring ordinary sensor set of underwater vehicle R_i. After that, ordinary sensor node $j \in \mathcal{N}_i$ replies its position and SNR measurement to underwater vehicle R_i. Repeating the above procedure, the collected messages on underwater vehicle $R_i \in \mathcal{V}$ can be expressed as

$$\left\{\mathbf{p}_{s,j},\ \mathrm{SNR_{dB}}(\mathbf{p}_i, \mathbf{p}_{s,j})\right\}_{j\in\mathcal{N}_i}. \tag{9}$$

For clear of expression, the ordinary sensor nodes in set $\mathcal{N}_i$ are labeled as $1_i, 2_i, \ldots, |\mathcal{N}_i|_i$. We stack the above SNR measurements into a vector $\mathbf{Y}_{\mathrm{dB}}^{R_i} = [\mathrm{SNR_{dB}}(\mathbf{p}_i, \mathbf{p}_{s,1_i}), \ldots, \mathrm{SNR_{dB}}(\mathbf{p}_i, \mathbf{p}_{s,|\mathcal{N}_i|_i})]^{\mathrm{T}}$. Noting with (6), one has

$$\mathbf{Y}_{\mathrm{dB}}^{R_i} = \mathbf{H}_{R_i}\boldsymbol{\theta}_{R_i} - \boldsymbol{\varepsilon}_{R_i} + \boldsymbol{\sigma}_{R_i} + \boldsymbol{\mu}_{R_i} \tag{10}$$

with

$$\mathbf{H}_{R_i} = \begin{bmatrix} 1 & -10\log_{10}\left(\left\|\mathbf{p}_{s,1_i} - \mathbf{p}_i\right\|\right) \\ \vdots & \vdots \\ 1 & -10\log_{10}\left(\left\|\mathbf{p}_{s,|\mathcal{N}_i|_i} - \mathbf{p}_i\right\|\right) \end{bmatrix}$$

$$\boldsymbol{\varepsilon}_{R_i} = \begin{bmatrix} 1 & \left\|\mathbf{p}_{s,1_i} - \mathbf{p}_i\right\| \\ \vdots & \vdots \\ 1 & \left\|\mathbf{p}_{s,|\mathcal{N}_i|_i} - \mathbf{p}_i\right\| \end{bmatrix} \begin{bmatrix} N_{\mathrm{dB}}^0 \\ 10\log_{10}(\alpha(f)) \end{bmatrix}$$

where $\boldsymbol{\theta}_{R_i} = [K_{\mathrm{dB},i},\ n_{\mathrm{PL},i}]^{\mathrm{T}}$, $\boldsymbol{\sigma}_{R_i} = [\sigma_{\mathrm{SH}}(\mathbf{p}_i, \mathbf{p}_{s,1_i}), \ldots, \sigma_{\mathrm{SH}}(\mathbf{p}_i, \mathbf{p}_{s,|\mathcal{N}_i|_i})]^{\mathrm{T}}$ and $\boldsymbol{\mu}_{R_i} = [\mu_{\mathrm{MP}}(\mathbf{p}_i, \mathbf{p}_{s,1_i}), \ldots, \mu_{\mathrm{MP}}(\mathbf{p}_i, \mathbf{p}_{s,|\mathcal{N}_i|_i})]^{\mathrm{T}}$. It is worth mentioning that N_{dB}^0 and f can be acquired by the priori knowledge.

In (6), $\boldsymbol{\theta}_{R_i}$ is a deterministic vector, representing the offset and slope of path loss, while $\boldsymbol{\sigma}_{R_i}$ and $\boldsymbol{\mu}_{R_i}$ are location-related random shadowing and multipath fading parameter vectors, respectively. We can estimate $\boldsymbol{\theta}_{R_i}$ by the available measurements; however, one cannot estimate $\boldsymbol{\sigma}_{R_i}$ and $\boldsymbol{\mu}_{R_i}$ due to their randomness. For that reason, we estimate the statistical characteristics of $\boldsymbol{\sigma}_{R_i}$ and $\boldsymbol{\mu}_{R_i}$, rather than the real-time values of $\boldsymbol{\sigma}_{R_i}$ and $\boldsymbol{\mu}_{R_i}$. It is assumed that $\boldsymbol{\sigma}_{R_i}$ is captured by a zero-mean Gaussian noise with an exponential spatial correlation. Similar to the assumption in [29] and [41], $\boldsymbol{\mu}_{R_i}$ is captured by lognormal distribution without the spatial correlation. Accordingly, we employ the spatial correlation to predict the SNR parameters. Then, the covariance matrices of $\boldsymbol{\sigma}_{R_i}$ and $\boldsymbol{\mu}_{R_i}$ can be expressed as

$$\boldsymbol{\Upsilon}_{R_i} = \xi_{R_i}^2 \begin{bmatrix} 1 & \cdots & \exp\left\{\dfrac{d_{1_i,|\mathcal{N}_i|_i}}{-\varphi_{R_i}}\right\} \\ \exp\left\{\dfrac{d_{2_i,1_i}}{-\varphi_{R_i}}\right\} & \cdots & \exp\left\{\dfrac{d_{2_i,|\mathcal{N}_i|_i}}{-\varphi_{R_i}}\right\} \\ \vdots & \ddots & \vdots \\ \exp\left\{\dfrac{d_{|\mathcal{N}_i|_i,1_i}}{-\varphi_{R_i}}\right\} & \cdots & 1 \end{bmatrix}$$

$$\boldsymbol{\Psi}_{R_i} = \rho_{R_i}^2 \mathbf{I}_{|\mathcal{N}_i|_i} \tag{11}$$

where $\xi_{R_i}^2$ is the power of shadowing, φ_{R_i} is the parameter controlling the spatial correlation, $\rho_{R_i}^2$ is the multipath fading power, and $d_{j_i,\vec{j}_i}$ is the distance between ordinary sensor nodes $j_i \in \{1_i, \ldots, |\mathcal{N}_i|_i\}$, and $\vec{j}_i \in \{1_i, \ldots, |\mathcal{N}_i|_i\}/\{j_i\}$.

Since $\boldsymbol{\sigma}_{R_i}$ and $\boldsymbol{\mu}_{R_i}$ are independent, one can define the variance of $\mathbf{Y}_{\mathrm{dB}}^{R_i}$ as $\boldsymbol{\Omega}_{R_i} = \boldsymbol{\Upsilon}_{\boldsymbol{R_i}} + \boldsymbol{\Psi}_{R_i}$. Hence, the estimations of $\boldsymbol{\theta}_{R_i}$ and $\boldsymbol{\Omega}_{R_i}$ can be acquired by maximizing the following maximum likelihood function, i.e.,

$$f\left(\mathbf{Y}_{\mathrm{dB}}^{R_i}\middle|\boldsymbol{\theta}_{R_i}, \boldsymbol{\Omega}_{R_i}\right) = \frac{1}{\sqrt{(2\pi)^{|\mathcal{N}_i|_i}\left|\boldsymbol{\Omega}_{R_i}\right|}}\exp\left\{-\frac{1}{2}\left(\mathbf{Y}_{\mathrm{dB}}^{R_i} - \mathbf{H}_{R_i}\boldsymbol{\theta}_{R_i} + \boldsymbol{\varepsilon}_{R_i}\right)^{\mathrm{T}}\boldsymbol{\Omega}_{R_i}^{-1} \times \left(\mathbf{Y}_{\mathrm{dB}}^{R_i} - \mathbf{H}_{R_i}\boldsymbol{\theta}_{R_i} + \boldsymbol{\varepsilon}_{R_i}\right)\right\} \quad (12)$$

and hence, for $\boldsymbol{\Delta}_{R_i} = \mathbf{Y}_{\mathrm{dB}}^{R_i} - \mathbf{H}_{R_i}\boldsymbol{\theta}_{R_i} + \boldsymbol{\varepsilon}_{R_i}$, it yields

$$\ln f\left(\mathbf{Y}_{\mathrm{dB}}^{R_i}\middle|\boldsymbol{\theta}_{R_i}, \boldsymbol{\Omega}_{R_i}\right) = -\frac{1}{2}\boldsymbol{\Delta}_{R_i}^{\mathrm{T}}\boldsymbol{\Omega}_{R_i}^{-1}\boldsymbol{\Delta}_{R_i} - \frac{1}{2}\ln\left|\boldsymbol{\Omega}_{R_i}\right| - \frac{|\mathcal{N}_i|_i}{2}\ln(2\pi). \quad (13)$$

In our previous work [42], a separate design strategy was developed, where $\boldsymbol{\theta}_{R_i}$ was estimated in Phase I, $\xi_{R_i}^2$, $\rho_{R_i}^2$, and φ_{R_i} were estimated in Phase II, and the iteration was conducted in Phase III. The above separate design has high computational complexity, while its estimation accuracy is sensitive to the measurement noise. To cover these deficiencies, this article jointly estimates $\boldsymbol{\theta}_{R_i}$, $\xi_{R_i}^2$, $\rho_{R_i}^2$, and φ_{R_i}. To this end, we differentiate (13) with respect to variable $\boldsymbol{\theta}_{R_i}$ and variance $\boldsymbol{\Omega}_{R_i}$, and hence, one can further have

$$\frac{\partial \ln f\left(\mathbf{Y}_{\mathrm{dB}}^{R_i}\middle|\boldsymbol{\theta}_{R_i}, \boldsymbol{\Omega}_{R_i}\right)}{\partial \boldsymbol{\theta}_{R_i}} = -\mathbf{H}_{R_i}^{\mathrm{T}}\boldsymbol{\Omega}_{R_i}^{-1}\boldsymbol{\Delta}_{R_i} \quad (14)$$

$$\frac{\partial \ln f\left(\mathbf{Y}_{\mathrm{dB}}^{R_i}\middle|\boldsymbol{\theta}_{R_i}, \boldsymbol{\Omega}_{R_i}\right)}{\partial \boldsymbol{\Omega}_{R_i}} = \frac{\boldsymbol{\Omega}_{R_i}^{-1}\boldsymbol{\Delta}_{R_i}\boldsymbol{\Delta}_{R_i}^{\mathrm{T}}\boldsymbol{\Omega}_{R_i}^{-1} - \boldsymbol{\Omega}_{R_i}^{-1}}{2}. \quad (15)$$

From (14) and (15), the optimization of $\boldsymbol{\theta}_{R_i}$ and $\boldsymbol{\Omega}_{R_i}$ is acquired by solving $\boldsymbol{\Delta}_{R_i} = \mathrm{argmin}_{\boldsymbol{\Delta}_{R_i}} ||((\partial \ln f(\mathbf{Y}_{\mathrm{dB}}^{R_i}|\boldsymbol{\theta}_{R_i}, \boldsymbol{\Omega}_{R_i}))/(\partial\boldsymbol{\theta}_{R_i}))||$ and $((\partial \ln f(\mathbf{Y}_{\mathrm{dB}}^{R_i}|\boldsymbol{\theta}_{R_i}, \boldsymbol{\Omega}_{R_i}))/(\partial\boldsymbol{\Omega}_{R_i})) = 0$, which are equal to $\boldsymbol{\Delta}_{R_i} = \boldsymbol{\varkappa}_{R_i}$ and $\boldsymbol{\Omega}_{R_i} = \boldsymbol{\Delta}_{R_i}\boldsymbol{\Delta}_{R_i}^{\mathrm{T}}$. Of note, $\boldsymbol{\varkappa}_{R_i}$ is the sum vector of multipath and shadow fading. Based on this and with the definition in (11), one rearranges $\boldsymbol{\Omega}_{R_i}$ as

$$\begin{bmatrix} \xi_{R_i}^2 + \rho_{R_i}^2 & \cdots & \xi_{R_i}^2\exp\left\{\frac{d_{1_i,|\mathcal{N}_i|_i}}{-\varphi_{R_i}}\right\} \\ \xi_{R_i}^2\exp\left\{\frac{d_{2_i,1_i}}{-\varphi_{R_i}}\right\} & \cdots & \xi_{R_i}^2\exp\left\{\frac{d_{2_i,|\mathcal{N}_i|_i}}{-\varphi_{R_i}}\right\} \\ \vdots & \ddots & \vdots \\ \xi_{R_i}^2\exp\left\{\frac{d_{|\mathcal{N}_i|_i,1_i}}{-\varphi_{R_i}}\right\} & \cdots & \xi_{R_i}^2 + \rho_{R_i}^2 \end{bmatrix} = \begin{bmatrix} ([\boldsymbol{\Delta}_{R_i}]_{1_i})^2 & \cdots & [\boldsymbol{\Delta}_{R_i}]_{1_i}[\boldsymbol{\Delta}_{R_i}]_{|\mathcal{N}_i|_i} \\ [\boldsymbol{\Delta}_{R_i}]_{2_i}[\boldsymbol{\Delta}_{R_i}]_{1_i} & \cdots & [\boldsymbol{\Delta}_{R_i}]_{2_i}[\boldsymbol{\Delta}_{R_i}]_{|\mathcal{N}_i|_i} \\ \vdots & \ddots & \vdots \\ [\boldsymbol{\Delta}_{R_i}]_{|\mathcal{N}_i|_i}[\boldsymbol{\Delta}_{R_i}]_{1_i} & \cdots & ([\boldsymbol{\Delta}_{R_i}]_{|\mathcal{N}_i|_i})^2 \end{bmatrix} \quad (16)$$

where $[\boldsymbol{\Delta}_{R_i}]_j$ is the jth element of $\boldsymbol{\Delta}_{R_i}$.

Based on (14)–(16), we can easily obtain

$$\mathbf{Y}_{\mathrm{dB}}^{R_i} - \mathbf{H}_{R_i}\boldsymbol{\theta}_{R_i} + \boldsymbol{\varepsilon}_{R_i} = \boldsymbol{\varkappa}_{R_i} \quad (17)$$

$$\xi_{R_i}^2 + \rho_{R_i}^2 = \frac{1}{|\mathcal{N}_i|_i}\sum_{j_i=1_i}^{|\mathcal{N}_i|_i}([\boldsymbol{\Delta}]_{j_i})^2 \quad (18)$$

$$\xi_{R_i}^2\exp\left\{\frac{d_{j_i,\vec{j}_i}}{-\varphi_{R_i}}\right\} = [\boldsymbol{\Delta}]_{j_i}[\boldsymbol{\Delta}]_{\vec{j}_i}. \quad (19)$$

We define $\boldsymbol{\chi}_{R_i} = [\boldsymbol{\theta}_{R_i}; \ln\xi_{R_i}^2; 1/\varphi_{R_i}; \rho_{R_i}^2]$, while the estimations of $\boldsymbol{\theta}_{R_i}$, $\xi_{R_i}^2$, φ_{R_i} and $\rho_{R_i}^2$ are denoted by $\hat{\boldsymbol{\theta}}_{R_i}$, $\hat{\xi}_{R_i}^2$, $\hat{\varphi}_{R_i}$ and $\hat{\rho}_{R_i}^2$, respectively. Based on the above results, the optimization of $\boldsymbol{\theta}_{R_i}$, $\xi_{R_i}^2$, φ_{R_i} and $\rho_{R_i}^2$ is conducted as

$$\hat{\boldsymbol{\chi}}_{R_i}^* = \mathrm{argmin}\ \underbrace{\left\{\lambda_1\left(\frac{1}{|\mathcal{N}_i|_i}\boldsymbol{\Delta}_{R_i}^{\mathrm{T}}\boldsymbol{\Delta}_{R_i} - \hat{\xi}_{R_i}^2 - \hat{\rho}_{R_i}^2\right)^2\right.}_{\text{Part 1}} + \underbrace{\sum_{(j_i,\vec{j}_i)\in\mathbf{G}}\lambda_2\left[\ln\hat{\xi}_{R_i}^2 - \frac{d_{j_i,\vec{j}_i}}{\hat{\varphi}_{R_i}} - \ln([\boldsymbol{\Delta}]_{j_i}[\boldsymbol{\Delta}]_{\vec{j}_i})\right]^2}_{\text{Part 2}} + \underbrace{\left.\left(\mathbf{Y}_{\mathrm{dB}}^{R_i} - \mathbf{H}_{R_i}\boldsymbol{\theta}_{R_i} + \boldsymbol{\varepsilon}_{R_i}\right)^{\mathrm{T}}\mathbf{Q}_1\left(\mathbf{Y}_{\mathrm{dB}}^{R_i} - \mathbf{H}_{R_i}\boldsymbol{\theta}_{R_i} + \boldsymbol{\varepsilon}_{R_i}\right)\right\}}_{\text{Part 3}} \quad (20)$$

where $\mathbf{G} = \{(1_i, 2_i), \ldots, (j_i, \vec{j}_i), \ldots, (|\mathcal{N}_i|_i, |\mathcal{N}_i|_i - 1)\}$ denotes the set of ordinary sensor nodes pairs for j_i and $\vec{j}_i$. In addition, $\lambda_1 > 0$ and $\lambda_2 > 0$ are the tuning indexes of shadowing and multipath cost function terms, respectively. Moreover, $\mathbf{Q}_1$ a positive definite matrix.

Let $\hat{\boldsymbol{\chi}}_{R_i}$ be the estimation of $\boldsymbol{\chi}_{R_i}$, and $\mathbf{u} \in \mathcal{R}^4$ be the increment input vector of $\hat{\boldsymbol{\chi}}_{R_i}$. Hence, the estimation procedure of $\boldsymbol{\chi}_{R_i}$ can be described as

$$\dot{\hat{\boldsymbol{\chi}}}_{R_i} = \mathbf{u}. \quad (21)$$

A model-based IRL estimator is developed to seek $\hat{\boldsymbol{\chi}}_{R_i}^*$, whose basic idea is to minimize the integral temporal difference error [43], [44]. Then, the cost function is defined as

$$g_1(\hat{\boldsymbol{\chi}}_{R_i}, \mathbf{u}) = \lambda_1\left(\frac{1}{|\mathcal{N}_i|_i}\boldsymbol{\Delta}_{R_i}^{\mathrm{T}}\boldsymbol{\Delta}_{R_i} - \hat{\xi}_{R_i}^2 - \hat{\rho}_{R_i}^2\right)^2 + \sum_{(j_i,\vec{j}_i)\in\mathbf{G}}\lambda_2\left[\ln\hat{\xi}_{R_i}^2 - \frac{d_{j_i,\vec{j}_i}}{\hat{\varphi}_{R_i}} - \ln([\boldsymbol{\Delta}]_{j_i}[\boldsymbol{\Delta}]_{\vec{j}_i})\right]^2 + \boldsymbol{\Delta}_{R_i}^{\mathrm{T}}\mathbf{Q}_1\boldsymbol{\Delta}_{R_i} + \mathbf{u}^{\mathrm{T}}\mathbf{R}_1\mathbf{u} \quad (22)$$

where $\mathbf{R}_1$ is a positive definite matrix.

From (22), the value function for the estimation of $\boldsymbol{\chi}_{R_i}$ is

$$V_1(\hat{\boldsymbol{\chi}}_{R_i}(t)) = \int_t^{t+T} g_1(\hat{\boldsymbol{\chi}}_{R_i}, \mathbf{u})\mathrm{d}\tau + V_1(\hat{\boldsymbol{\chi}}_{R_i}(t+T)) \quad (23)$$

and hence, the optimal value of $\boldsymbol{\chi}_{R_i}$ is to select $\mathbf{u}$, such that an optimal update policy of $\mathbf{u}$ can be obtained, i.e.,

$$\mathbf{u}^* = \underset{\mathbf{u}}{\mathrm{argmin}}\left\{\int_t^{t+T} g_1(\hat{\boldsymbol{\chi}}_{R_i}, \mathbf{u})\mathrm{d}\tau + V_1(\hat{\boldsymbol{\chi}}_{R_i}(t+T))\right\}. \quad (24)$$

In the following, the IRL strategy includes two steps, i.e., policy evaluation and policy improvement. In policy evaluation, $V_1(\hat{\boldsymbol{\chi}}_{R_i}(t))$ is evaluated by using (23), given the current update policy. In policy improvement, the optimal update policy is selected until the convergence is reached for the iteration procedure. The above steps are detailed as follows.

1) *Initialization:* Initially, the policy and value function are set as $\mathbf{u}^{(0)}(0)=\mathbf{0}$ and $V^{(0)}(\hat{\boldsymbol{\chi}}_{R_i}(0))=0$.
2) *Policy Evaluation:* For each iteration s, one calculates the following value function:

$$V_1^{(s)}(\hat{\boldsymbol{\chi}}_{R_i}(t)) = \int_t^{t+T} g_1(\hat{\boldsymbol{\chi}}_{R_i},\mathbf{u}^{(s)})\mathrm{d}\tau + V_1^{(s)}(\hat{\boldsymbol{\chi}}_{R_i}(t+T)). \quad (25)$$

3) *Policy Improvement:* Find an updated control policy $\mathbf{u}_1^{(s+1)}$ through the following rule:

$$\mathbf{u}^{(s+1)} = \underset{\mathbf{u}^{(s)}}{\operatorname{argmin}}\left\{\int_t^{t+T} g_1(\hat{\boldsymbol{\chi}}_{R_i},\mathbf{u}^{(s)})\mathrm{d}\tau + V_1^{(s)}(\hat{\boldsymbol{\chi}}_{R_i}(t+T))\right\} \quad (26)$$

which yields $\mathbf{u}^{(s+1)} = -(1/2)\mathbf{R}_1^{-1}((\partial V_1^{(s)}(\hat{\boldsymbol{\chi}}_{R_i}(t)))/(\partial\hat{\boldsymbol{\chi}}_{R_i}))$.

In order to smoothly approximate the value function $V_1^{(s)}(\hat{\boldsymbol{\chi}}_{R_i}(t))$, a critic network is introduced as

$$V_1^{(s)}(\hat{\boldsymbol{\chi}}_{R_i}(t)) = \mathbf{W}_1^{(s)^{\mathrm{T}}}\boldsymbol{\phi}_1(\hat{\boldsymbol{\chi}}_{R_i}(t)) \quad (27)$$

where $\boldsymbol{\phi}_1(\hat{\boldsymbol{\chi}}_{R_i}(t))$ is basis function for weight vector $\mathbf{W}_1$.

Based on (27), one rearranges (25) as

$$\mathbf{W}_1^{(s)^{\mathrm{T}}}\boldsymbol{\phi}_1(\hat{\boldsymbol{\chi}}_{R_i}(t)) = \int_t^{t+T} g_1(\hat{\boldsymbol{\chi}}_{R_i},\mathbf{u}^{(s)})\mathrm{d}\tau + \mathbf{W}_1^{(s)^{\mathrm{T}}}\boldsymbol{\phi}_1(\hat{\boldsymbol{\chi}}_{R_i}(t+T)) \quad (28)$$

and its residual error can be expressed as

$$e_1(t) = \mathbf{W}_1^{(s)^{\mathrm{T}}}\left(\boldsymbol{\phi}_1(\hat{\boldsymbol{\chi}}_{R_i}(t)) - \boldsymbol{\phi}_1(\hat{\boldsymbol{\chi}}_{R_i}(t+T))\right) - \int_t^{t+T} g_1(\hat{\boldsymbol{\chi}}_{R_i},\mathbf{u}^{(s)})\mathrm{d}\tau. \quad (29)$$

The purpose of weight updating is to minimize the overall residual error, i.e., min $\int_0^\infty e_1(\tau)\mathrm{d}\tau$. Hence, the recursive least square method is used to update the weight $\mathbf{W}_1$. Let $\varrho_1^{(s)} = \int_t^{t+T} g_1(\hat{\boldsymbol{\chi}}_{R_i},\mathbf{u}^{(s)})\mathrm{d}\tau$ denote the value of the cost function for a period of time under control policy $\mathbf{u}^{(s)}$. Along with this, the following law is adopted to update $\mathbf{W}_1^{(s)}$, i.e.,

$$\mathbf{W}_1^{(s)} = \mathbf{W}_1^{(s-1)} + \frac{\mathbf{P}_{s-1}\boldsymbol{\phi}_1(t)\left(\varrho_1^{(s)} - \boldsymbol{\phi}_1(t)^{\mathrm{T}}\mathbf{W}_1^{(s-1)}\right)}{\kappa_1 + \boldsymbol{\phi}_1(t)^{\mathrm{T}}\mathbf{P}_{s-1}\boldsymbol{\phi}_1(t)} \quad (30)$$

with variance matrix $\mathbf{P}_s$ to adjust the update speed, i.e.,

$$\mathbf{P}_s = \mathbf{P}_{s-1} - \frac{\mathbf{P}_{s-1}\boldsymbol{\phi}_1(t)\boldsymbol{\phi}_1(t)^{\mathrm{T}}\mathbf{P}_{s-1}}{\kappa_1 + \boldsymbol{\phi}_1(t)^{\mathrm{T}}\mathbf{P}_{s-1}\boldsymbol{\phi}_1(t)} \quad (31)$$

where κ_1 is the forgetting factor in the weight update process. In addition, $\boldsymbol{\phi}_1(t) = \boldsymbol{\phi}_1(\hat{\boldsymbol{\chi}}_{R_i}(t)) - \boldsymbol{\phi}_1(\hat{\boldsymbol{\chi}}_{R_i}(t+T))$ is the excitation function, whose role is to drive the weight to accomplish convergence. Moreover, $||\mathbf{W}_1^{(s)} - \mathbf{W}_1^{(s-1)}|| < \epsilon_1$ and $||\hat{\boldsymbol{\chi}}_{R_i}(t+T) - \hat{\boldsymbol{\chi}}_{R_i}(t)|| < \epsilon_2$ denote the termination conditions for the above two interaction functions, where ϵ_1 and ϵ_2 are small positive decimals.

When the above iteration procedure is ended, the optimal SNR parameters $\boldsymbol{\theta}_{R_i}^*$, $\xi_{R_i}^{*2}$, $\varphi_{R_i}^*$ and $\rho_{R_i}^{*2}$ can be obtained.

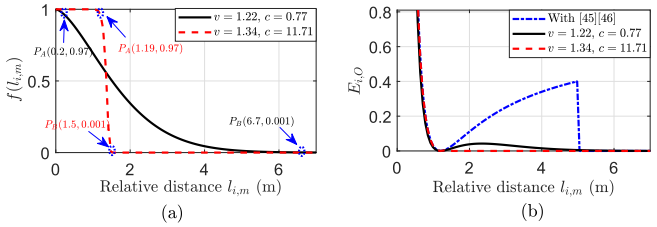


Fig. 3. Example of obstacle avoidance function where $\rho_{om} = 1.2$ m. (a) Steepness of $f(l_{i,m})$. (b) Value of $E_{i,O}$.

Accordingly, the predicted mean of SNR between source neighbor R_{i-1} and vehicle R_i is expressed as

$$\overrightarrow{\mathrm{SNR}}_{\mathrm{dB}}(\mathbf{p}_{i-1},\mathbf{p}_i) = \mathbf{h}_{R_i}\boldsymbol{\theta}_{R_i}^* - \varepsilon_{R_i} + \boldsymbol{\Xi}_{R_i}^{\mathrm{T}}\boldsymbol{\Omega}_{R_i}^{-1}\left(\mathbf{Y}_{\mathrm{dB}}^{R_i} - \mathbf{H}_{R_i}\boldsymbol{\theta}_{R_i}^* + \boldsymbol{\varepsilon}_{R_i}\right) \quad (32)$$

with

$$\mathbf{h}_{R_i} = [1, -10\log_{10}(\|\mathbf{p}_i - \mathbf{p}_{i-1}\|)]$$
$$\boldsymbol{\Xi}_{R_i} = \xi_{R_i}^{*2}\left[\mathrm{e}^{-\frac{\|\mathbf{p}_i-\mathbf{p}_{s,1_i}\|}{\varphi_{R_i}^*}}, \ldots, \mathrm{e}^{-\frac{\|\mathbf{p}_i-\mathbf{p}_{s,|\mathcal{N}_i|_i}\|}{\varphi_{R_i}^*}}\right]^{\mathrm{T}}$$
$$\varepsilon_{R_i} = N_{\mathrm{dB}}^0 + 10\log_{10}(\alpha(f)).$$

B. Model-Free and Collision-Free Motion Planning Algorithm

The following cost function is defined for underwater vehicle R_i to maximize its channel capacity $\mathcal{C}_i$, i.e.,

$$E_{i,\mathrm{L}}(\mathbf{p}_i) = \frac{1}{\mathcal{C}_i^2(\mathbf{p}_i)}. \quad (33)$$

Then, the obstacle avoidance function for underwater vehicle R_i can be defined as

$$E_{i,\mathrm{O}}(\mathbf{p}_i) = \sum_{m\in\Omega_i^t} f(l_{i,m})\left(\frac{1}{l_{i,m}} - \frac{1}{\rho_{om}}\right)^2 \quad (34)$$

with

$$f(l_{i,m}) = \exp\left(-\frac{1}{2}\left(\frac{l_{i,m}^2}{\nu^2}\right)^c\right) \quad (35)$$

where $l_{i,m} = \rho_m + \rho_{i,m}$, $\rho_{i,m}$ is the minimum distance from underwater vehicle R_i to obstacle m, ρ_{om} is the radius of the mth obstacle's influence boundary cylinder, c is the steepness of repulsive function, and ν is the repulsive range.

Remark 2: In (35), a large c causes a steep shape for obstacle avoidance, while a large ν causes a wide repulsive range. In view of this, a point $P_{\mathrm{A}} = [l_{\mathrm{left}}, \kappa_1]$ that tends to ρ_{om} from left is selected to capture the steepness, and a point $P_{\mathrm{B}} = [l_{\mathrm{right}}, \kappa_2]$ that stay off ρ_{om} from right is selected to capture the repulsive range. Of note, $0 < \kappa_2 < \kappa_1 < 1$. Then, c and ν are selected as $c = ((\ln(\log_{\kappa_1}\kappa_2))/(2\ln l_{\mathrm{right}} - \ln l_{\mathrm{left}}))$ and $\nu = (l_{\mathrm{left}}/(\exp(((\ln(-2\ln\kappa_1))/2c))))$. An example of the above selection result is shown in Fig. 3(a).

Most of the existing works (e.g., [13], [45]) set $f(l_{i,m})$ as 1 if $l_{i,m} \le \nu$, and $f(l_{i,m}) = 0$ otherwise. The above design leads to the discontinuity of repulsive potential field, which causes the system dithering. To avoid this shortage,

a smooth coefficient $f(l_{i,m})$ is introduced to eliminate the impact of avoidance function $E_{i,\mathrm{O}}(\mathbf{p}_i)$ beyond the safety distance. Clearly, $E_{i,\mathrm{O}}(\mathbf{p}_i)=0$ is equivalent to $l_{i,m}=\rho_{om}$, which means underwater vehicle R_i moves on the influence boundaries of the obstacles. If $l_{i,m}>\nu$, one regards that underwater vehicle R_i has already escaped the influences of obstacles, and hence, the value of $f(l_{i,m})$ is very small. If $l_{i,m}\leq\nu$, one regards that underwater vehicle R_i enters into the influences of obstacles, and hence, the value of $f(l_{i,m})$ is increased with the decreasing of $l_{i,m}$. The effect of $f(l_{i,m})$ on the obstacle avoidance function is depicted by Fig. 3(b).

Underwater vehicle R_i also requires to avoid collision with its neighboring vehicle $j\in\mathcal{N}_i^*$, where $\mathcal{N}_i^*$ is its neighboring set. Similar to (34), the internal collision avoidance function for underwater vehicle R_i can be denoted by

$$E_{i,j}(\mathbf{p}_i)=\sum_{j\in\mathcal{N}_i^*}\exp\left(-\frac{1}{2}\left(\frac{l_{i,j}^2}{\nu_1^2}\right)^{c_1}\right)\frac{1}{l_{i,j}^2} \tag{36}$$

where $l_{i,j}=\|\mathbf{p}_i-\mathbf{p}_j\|$ is the relative distance between underwater vehicle R_i and neighboring vehicle $j\in\mathcal{N}_i^*$. In addition, ν_1 and c_1 are the repulsive range and steepness for the internal collision avoidance, respectively.

With (2), (33), (34), and (36), the total cost function for the motion planning of underwater vehicle R_i is

$$\begin{aligned}&\bar{g}_i(\hat{\mathbf{p}}_i,\tau_i)\\&=\beta_1E_{i,\mathrm{L}}(\hat{\mathbf{p}}_i)+\beta_2E_{i,\mathrm{O}}(\hat{\mathbf{p}}_i)+\beta_3E_{i,j}(\hat{\mathbf{p}}_i)+\boldsymbol{\tau}_i^{\mathrm{T}}\bar{\mathbf{R}}_i\boldsymbol{\tau}_i\\&=\frac{\beta_1}{\mathcal{C}_i^2(\hat{\mathbf{p}}_i)}+\beta_2\sum_{m\in\Omega_i^t}f(\hat{l}_{i,m})\left(\frac{1}{\hat{l}_{i,m}}-\frac{1}{\rho_{om}}\right)^2\\&\quad+\beta_3\sum_{j\in\mathcal{N}_i^*}\exp\left(-\frac{1}{2}\left(\frac{l_{i,j}^2}{\nu_1^2}\right)^{c_1}\right)\frac{1}{l_{i,j}^2}+\boldsymbol{\tau}_i^{\mathrm{T}}\bar{\mathbf{R}}_i\boldsymbol{\tau}_i\end{aligned} \tag{37}$$

where $\hat{\mathbf{p}}_i$ and $\hat{l}_{i,m}$ denote the estimations of $\mathbf{p}_i$ and $l_{i,m}$, respectively. $\bar{\mathbf{R}}_i$ is a positive definite matrix. Besides that, β_1, β_2, and β_3 are positive constants, whose role is to balance the communication efficiency and collision avoidance.

Based on (37), the value function for control input $\boldsymbol{\tau}_i$ is

$$\bar{V}_i(\hat{\mathbf{X}}_i(t))=\int_t^{t+T}\bar{g}_i(\hat{\mathbf{p}}_i,\tau_i)\mathrm{d}\tau+\bar{V}_i(\hat{\mathbf{X}}_i(t+T)) \tag{38}$$

where $\hat{\mathbf{X}}_i$ denotes the estimated value of $\mathbf{X}_i$.

Hence, one can construct the following optimal problem:

$$\boldsymbol{\tau}_i^*(t)=\underset{\boldsymbol{\tau}_i}{\operatorname{argmin}}\int_t^{t+T}\bar{g}_i(\hat{\mathbf{p}}_i,\tau_i)\mathrm{d}\tau+\bar{V}_i(\hat{\mathbf{X}}_i(t+T)). \tag{39}$$

Meanwhile, dynamic model (2) can be rearranged as

$$\dot{\mathbf{X}}_i=\mathbf{A}_i\mathbf{X}_i+\mathbf{B}_i\boldsymbol{\tau}_i^{(s)}+\mathbf{B}_i\left(\boldsymbol{\tau}_i-\boldsymbol{\tau}_i^{(s)}\right)+\mathbf{G}_i \tag{40}$$

where $\boldsymbol{\tau}_i^{(s)}$ is the updated policy in the sth iteration and $\boldsymbol{\tau}_i$ is an admissible policy for the learning procedure.

Combining (38) with (40), the derivative of the value function in the sth iteration can be calculated as

$$\begin{aligned}\dot{\bar{V}}_i^{(s)}(\hat{\mathbf{X}}_i)&=\nabla\bar{V}_{\mathbf{X}_i}^{(s)\mathrm{T}}\left(\mathbf{A}_i\hat{\mathbf{X}}_i+\mathbf{B}_i\boldsymbol{\tau}_i^{(s)}+\mathbf{G}_i\right)\\&\quad-2\left(-\frac{1}{2}\bar{\mathbf{R}}_i^{-1}\mathbf{B}_i^{\mathrm{T}}\nabla\bar{V}_{\mathbf{X}_i}^{(s)}\right)^{\mathrm{T}}\bar{\mathbf{R}}_i\left(\boldsymbol{\tau}_i-\boldsymbol{\tau}_i^{(s)}\right)\\&=-\bar{g}_i\left(\hat{\mathbf{p}}_i,\boldsymbol{\tau}_i^{(s)}\right)-2\boldsymbol{\tau}_i^{(s+1)\mathrm{T}}\bar{\mathbf{R}}_i\left(\boldsymbol{\tau}_i-\boldsymbol{\tau}_i^{(s)}\right)\end{aligned} \tag{41}$$

where $\bar{\mathbf{R}}_i=\mathrm{diag}(r_{i,1},r_{i,2},r_{i,3},r_{i,4})$ is a positive definite matrix, and $\nabla\bar{V}_{\mathbf{X}_i}^{(s)}=\partial\bar{V}^{(s)}(\mathbf{X}_i)/\partial\mathbf{X}_i$. The model parameter $\mathbf{B}_i$ can be eliminated through the model-based control policy $-(1/2)\bar{\mathbf{R}}_i^{-1}\mathbf{B}_i^{\mathrm{T}}\nabla\bar{V}_{\mathbf{X}_i}^{(s)}$ which is similar to the one in Section III-A. Hence, the desired policy $\boldsymbol{\tau}_i^{(s)}$ can be obtained by solving (41).

In the following, a model-free policy iteration algorithm is employed to seek the optimal update policy $\boldsymbol{\tau}_i^*$.

1) *Initialization:* Initially, the policy and value function are set as $\boldsymbol{\tau}_i^{(0)}(0)=\mathbf{0}$ and $\bar{V}_i^{(0)}(\hat{\mathbf{X}}_i(0))=0$, respectively.
2) *Policy Evaluation:* For each iteration s, calculate the following value function obtained by integrating (41):

$$\begin{aligned}&\bar{V}_i^{(s)}(\hat{\mathbf{X}}_i(t+T))-\bar{V}_i^{(s)}(\hat{\mathbf{X}}_i(t))\\&\quad=-\int_t^{t+T}2\boldsymbol{\tau}_i^{(s+1)\mathrm{T}}\bar{\mathbf{R}}_i\left(\boldsymbol{\tau}_i-\boldsymbol{\tau}_i^{(s)}\right)\mathrm{d}\tau\\&\qquad-\int_t^{t+T}\bar{g}_i\left(\hat{\mathbf{p}}_i,\boldsymbol{\tau}_i^{(s)}\right)\mathrm{d}\tau.\end{aligned} \tag{42}$$

3) *Policy Improvement:* Find an updated control policy $\boldsymbol{\tau}_i^{(s+1)}$ through the following rule:

$$\boldsymbol{\tau}_i^{(s+1)}=\underset{\boldsymbol{\tau}_i^{(s)}}{\operatorname{argmin}}\int_t^{t+T}e_2\left(\boldsymbol{\tau}_i^{(s)}\right)\mathrm{d}\tau \tag{43}$$

with

$$\begin{aligned}e_2(\boldsymbol{\tau}_i^{(s)})&=\bar{V}_i^{(s)}(\hat{\mathbf{X}}_i(t+T))-\bar{V}_i^{(s)}(\hat{\mathbf{X}}_i(t))\\&\quad+\int_t^{t+T}2\boldsymbol{\tau}_i^{(s+1)\mathrm{T}}\bar{\mathbf{R}}_i\left(\boldsymbol{\tau}_i-\boldsymbol{\tau}_i^{(s)}\right)\mathrm{d}\tau\\&\quad+\int_t^{t+T}\bar{g}_i\left(\hat{\mathbf{p}}_i,\boldsymbol{\tau}_i^{(s)}\right)\mathrm{d}\tau.\end{aligned} \tag{44}$$

In the above process, the analytical forms of $\bar{V}_i^{(s)}$ and $\boldsymbol{\tau}_i^{(s)}$ are unknown previously. In order to smoothly approximate the value function $\bar{V}_i^{(s)}$ and the desired policy $\boldsymbol{\tau}_i^{(s)}$, the critic network and actor network are introduced as

$$\bar{V}_i^{(s)}(\hat{\mathbf{X}}_i)=\bar{\mathbf{W}}_i^{(s)\mathrm{T}}\boldsymbol{\phi}_2^i(\hat{\mathbf{X}}_i) \tag{45}$$

$$\boldsymbol{\tau}_i^{(s+1)}=\left[\boldsymbol{\tau}_{i,1}^{(s+1)};\,\boldsymbol{\tau}_{i,2}^{(s+1)};\,\boldsymbol{\tau}_{i,3}^{(s+1)};\,\boldsymbol{\tau}_{i,4}^{(s+1)}\right] \tag{46}$$

where $\boldsymbol{\phi}_2^i$ is the basis function vector for the weight vector $\bar{\mathbf{W}}_i$. In addition, $\boldsymbol{\tau}_{i,\vec{i}}^{(s+1)}=\tilde{\mathbf{W}}_{i,\vec{i}}^{(s+1)^{\mathrm{T}}}\boldsymbol{\phi}_{3,\vec{i}}^i(\hat{\mathbf{X}}_i)$, where $\boldsymbol{\phi}_{3,\vec{i}}^i$ is the basis function vector for the weight vector $\tilde{\mathbf{W}}_{i,\vec{i}}$. Of note, $\tilde{\mathbf{W}}_{i,\vec{i}}$ is the $\vec{i}$th policy weight for $\vec{i}\in\{1,\ldots,4\}$.

Noting with (45) and (46), one can deduce the following result from (44), i.e.,

$$\begin{aligned}e_2\left(\boldsymbol{\tau}_i^{(s)}\right)&=\bar{\mathbf{W}}_i^{(s)^{\mathrm{T}}}\left(\boldsymbol{\phi}_2^i(\hat{\mathbf{X}}_i(t+T))-\boldsymbol{\phi}_2^i(\hat{\mathbf{X}}_i(t))\right)\\&\quad+2\sum_{\vec{i}=1}^{4}r_{i,\vec{i}}\int_t^{t+T}\tilde{\mathbf{W}}_{i,\vec{i}}^{(s+1)^{\mathrm{T}}}\boldsymbol{\phi}_{3,\vec{i}}^i(\hat{\mathbf{X}}_i(\tau))\bar{\tau}_{i,\vec{i}}(t)\mathrm{d}\tau\\&\quad+\int_t^{t+T}\bar{g}_i\left(\hat{\mathbf{p}}_i,\boldsymbol{\tau}_i^{(s)}\right)\mathrm{d}\tau\end{aligned} \tag{47}$$

where $\bar{\tau}_{i,\vec{i}}(t)=\tau_{i,\vec{i}}-\tau_{i,\vec{i}}^{(s)}$. Specifically, $\tau_{i,\vec{i}}$ denotes the $\vec{i}$th element of $\boldsymbol{\tau}_i$, and $\tau_{i,\vec{i}}^{(s)}$ denotes the $\vec{i}$th element of $\boldsymbol{\tau}_i^{(s)}$.

Let $\bar{\varrho}_i^{(s)}=\int_t^{t+T}\bar{g}_i(\hat{\mathbf{p}}_i,\boldsymbol{\tau}_i^{(s)})\mathrm{d}\tau$ denote the value of the cost function for a period of time T under the given control

Algorithm 1 IRL-Based Motion Planning Controller

Input: $\boldsymbol{\chi}_{R_i}(0)$, $\mathbf{X}_i(0)$, $\boldsymbol{\tau}_i(0)$
Output: The optimal control policy $\boldsymbol{\tau}_i^*$

```
while ||1/C_i(t+T) − 1/C_i(t)|| < ε_4 do
    Collect SNR measurement data Y_dB^{R_i} and p_{s,j}
    s ← 0;
    for t = 0 : T_{max,1} do
        Calculate (22), (26) and ϱ_1^{(s)} in turn
        Collect data χ̂_{R_i}(t), χ̂_{R_i}(t+T), u_1^{(s)} for (30)
        if The termination conditions are satisfied then
            break;
        s ← s + 1;
    s ← 0;
    Obtain χ̂*_{R_i}, and construct the link capacity C_i
    for t = 0 : T_{max,2} do
        Determine (37), (47) and ϱ̄_i^{(s)} in trun
        Collect data X̂_i(t), X̂_i(t+T), τ_i^{(s)} for (48)
        if ||W̄_i^{(s)} − W̄_i^{(s−1)}|| < ε_3 then
            break;
        s ← s + 1;
```

policy $\boldsymbol{\tau}_i^{(s)}$. From (47), the update weight vector can be updated by the following iteration procedure, i.e.,

$$\bar{\mathbf{W}}_i^{(s)} = \bar{\mathbf{W}}_i^{(s-1)} + \frac{\bar{\mathbf{P}}_{i,s-1}\boldsymbol{\phi}(t)\left(\bar{\varrho}_i^{(s)} - \boldsymbol{\phi}(t)^{\mathrm{T}}\bar{\mathbf{W}}_i^{(s-1)}\right)}{\bar{\kappa} + \boldsymbol{\phi}(t)^{\mathrm{T}}\bar{\mathbf{P}}_{i,s-1}\boldsymbol{\phi}(t)} \tag{48}$$

with variance matrix $\bar{\mathbf{P}}_i$ to adjust the update speed, i.e.,

$$\bar{\mathbf{P}}_{i,s} = \bar{\mathbf{P}}_{i,s-1} - \frac{\bar{\mathbf{P}}_{i,s-1}\boldsymbol{\phi}(t)\boldsymbol{\phi}(t)^{\mathrm{T}}\bar{\mathbf{P}}_{i,s-1}}{\bar{\kappa} + \boldsymbol{\phi}(t)\bar{\mathbf{P}}_{i,s-1}\boldsymbol{\phi}(t)} \tag{49}$$

where $\bar{\kappa}$ is the forgetting factor. In addition, $\boldsymbol{\phi}(t) = [\boldsymbol{\phi}_2^i(\hat{\mathbf{X}}_i(t)) - \boldsymbol{\phi}_2^i(\hat{\mathbf{X}}_i(t+T)); \ 2r_{i,1}\int_t^{t+T}\boldsymbol{\phi}_{3,1}^i(\hat{\mathbf{X}}_i)\bar{\tau}_{i,1}\mathrm{d}\tau; \ 2r_{i,2}\int_t^{t+T}\boldsymbol{\phi}_{3,2}^i(\hat{\mathbf{X}}_i)\bar{\tau}_{i,2}\mathrm{d}\tau; \ 2r_{i,3}\int_t^{t+T}\boldsymbol{\phi}_{3,3}^i(\hat{\mathbf{X}}_i)\bar{\tau}_{i,3}\mathrm{d}\tau; \ 2r_{i,4}\times\int_t^{t+T}\boldsymbol{\phi}_{3,4}^i(\hat{\mathbf{X}}_i)\bar{\tau}_{i,4}\mathrm{d}\tau]$, which continuously stimulates the weight to converge to the appropriate value. Besides that, $||\bar{\mathbf{W}}_i^{(s)} - \bar{\mathbf{W}}_i^{(s-1)}|| < \epsilon_3$ and $||(1/(\mathcal{C}_i(t+T))) - (1/(\mathcal{C}_i(t)))|| < \epsilon_4$ are end conditions, where ϵ_3 and ϵ_4 are positive decimals.

Based on the above iteration procedure, the optimization control input $\boldsymbol{\tau}_i^*$ can be obtained, as depicted by Algorithm 1, where $T_{\max,1}$ and $T_{\max,2}$ represent the maximum times for SNR prediction and motion planning, respectively.

Remark 3: In Section III-A, a model-based IRL estimator is adopted, since the kinematic model (21) that is employed can be accurately known by underwater vehicles. By contrast, the dynamic model (2) is adopted in Section III-B. Due to the harsh ocean environment, it is difficult to acquire the accurate dynamics model of underwater vehicles, e.g., $\mathbf{A}_i$, $\mathbf{B}_i$, and $\mathbf{G}_i$. In view of this, a model-based IRL estimator is developed in Section III-B, even if it is complicated in implementation.

C. Preformation Analysis

For SNR prediction, the optimal policy is given in (26), whose convergence is presented as follows.

Theorem 1: Given an initial admissible policy $\mathbf{u}^{(0)}(0)$, the policy iteration (26) can make $\mathbf{u}^{(s+1)}$ converge to the optimal policy $\mathbf{u}^*$, such that $\hat{\boldsymbol{\chi}}_{R_i}^*$ can also be obtained.

Proof: The proof is given in Appendix A. ■

For motion planning control, the model-free IRL is adopted to find the optimal control strategy $\boldsymbol{\tau}_i^*$, as provided by (39). With regard to this, the following convergence analysis is presented for the optimal control strategy $\boldsymbol{\tau}_i^*$.

Theorem 2: Given an initial admissible policy $\boldsymbol{\tau}_i^{(0)}(0)$, the policy iterations (42) and (43) can make $\boldsymbol{\tau}_i^{(s+1)}$ converge to the optimal policy $\boldsymbol{\tau}_i^*$.

Proof: The proof is given in Appendix B. ■

The underwater acoustic communication in this article is divided into the following two parts: 1) data collection of SNR measurements (see Section III-A) and 2) collision avoidance between different vehicles (see Section III-B). Then, the communication complexity is studied by counting the transmitted and received scalars for each node, as similar to [34] and [46].

Step 1 (Complexity in Part 1): Recall that underwater vehicle $R_i \in \mathcal{V}$ broadcasts an initiator message to its neighboring ordinary sensor nodes, through which underwater vehicle $R_i \in \mathcal{V}$ receives the replies from ordinary sensor node $j \in \mathcal{N}_i$, i.e., $\{\mathbf{p}_{s,j}, \mathrm{SNR}_{\mathrm{dB}}(\mathbf{p}_i, \mathbf{p}_{s,j})\}_{j\in\mathcal{N}_i}$. Based on this, underwater vehicle $R_i \in \mathcal{V}$ transmits 1 scalars and receives $|\mathcal{N}_i|_i$ scalars during the data collection procedure. Along with this, any ordinary sensor node $j \in \mathcal{N}_i$ receives the initiator message from underwater vehicle R_i, and then it replies its position and SNR measurement to underwater vehicle R_i. Correspondingly, ordinary sensor node $j \in \mathcal{N}_i$ transmits four scalars and receives one scalars during the data collection procedure.

Step 2 (Complexity in Part 2): During the internal collision avoidance procedure, underwater vehicle $R_i \in \mathcal{V}$ transmits its position information to its neighboring vehicles, and meanwhile it receives the position information from neighboring vehicles. Thus, the transmitted and received scalars for underwater vehicle $R_i \in \mathcal{V}$ are given as four and four, respectively. In addition, the ordinary sensor nodes do not implement communication task in this part, so the transmitted and received scalars for ordinary sensor node $j \in \mathcal{N}_i$ are all zeros.

The collision avoidance analysis is presented as follows.

Corollary 1: Given cost function (37) and value function (38), underwater vehicle R_i never collides with obstacles or neighbor R_j, if the following condition is satisfied, i.e.,

$$\begin{aligned}\bar{g}_i\left(\hat{\mathbf{p}}_i, \tau_i^{(0)}\right)\big|_{t=0} &= \beta_1 E_{i,\mathrm{L}}(\hat{\mathbf{p}}_i)|_{t=0} + \boldsymbol{\tau}_i^{(0)^{\mathrm{T}}}\bar{\mathbf{R}}_i\boldsymbol{\tau}_i^{(0)}\big|_{t=0} \\ &< \min\left\{\beta_2 E_{i,\mathrm{O}}^{\max}, \beta_3 E_{i,j}^{\max}\right\}\end{aligned} \tag{50}$$

where $\beta_2 E_{i,\mathrm{O}}^{\max} = \sum_{m\in\Omega_i^l}((\sqrt{\beta_2}/\hat{l}_{i,m}) - (\sqrt{\beta_2}/\rho_{om}))^2$ and $\beta_3 E_{i,j}^{\max} = \sum_{j\in\mathcal{N}_i^*}(\beta_3/l_{i,j}^2)$.

Proof: The proof is given in Appendix C. ■

IV. Simulation and Experimental Studies

A. Simulation Results

In this section, simulation results are presented to verify the effectiveness. Specifically, the positions of source sensor node and destination sensor node are set as $[-30, -30, -8]^{\mathrm{T}}$ and $[30, 30, -8]^{\mathrm{T}}$, respectively. The initial position and orientation vectors of underwater vehicles R_1 and R_2 are set as

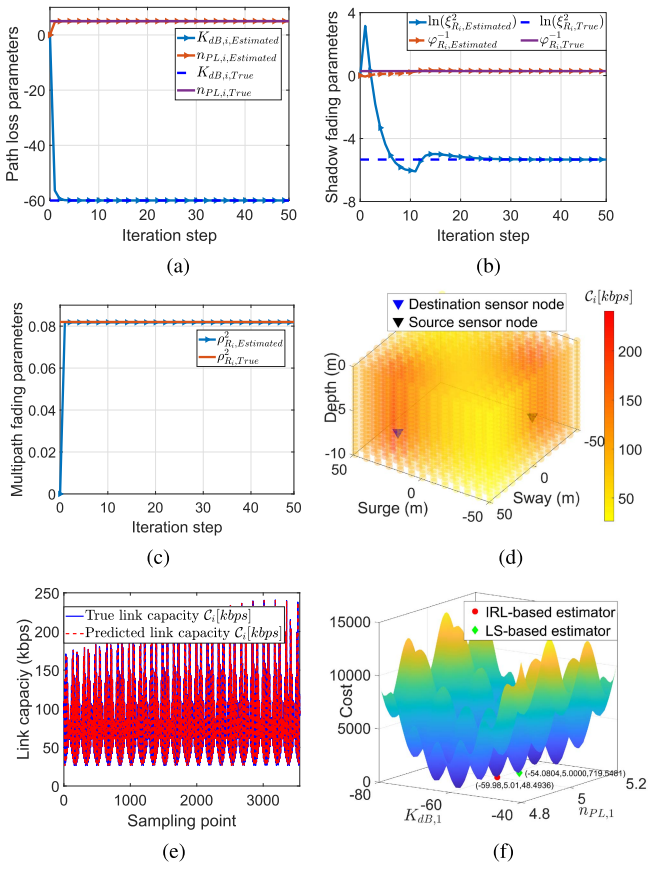


Fig. 4. Simulation results for the RL-based online SNR prediction. (a) Estimation of path loss. (b) Estimation of shadowing. (c) Estimation of multipath fading. (d) Predicted link capacity. (e) Link capacity with noise. (f) Comparison with [29].

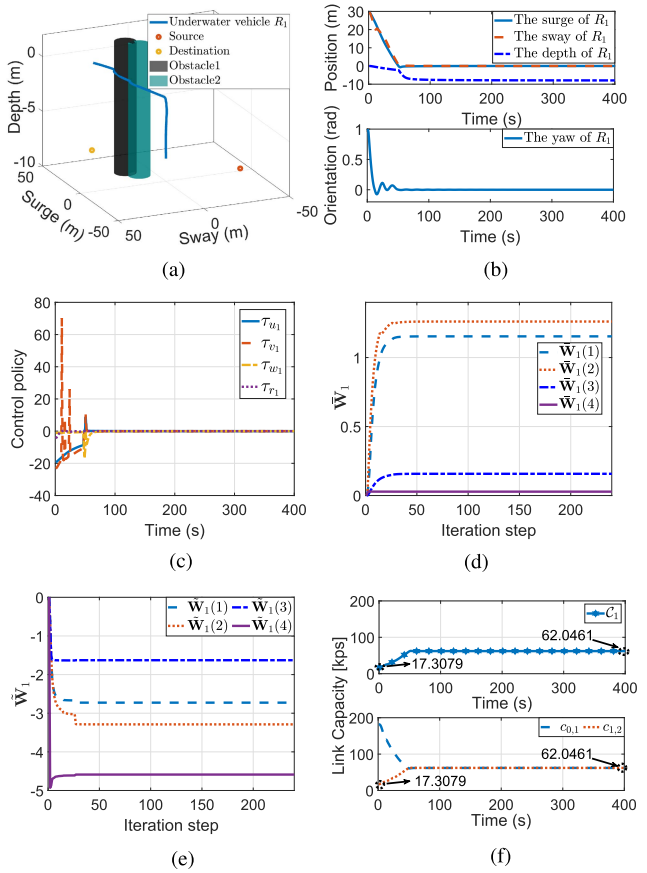


Fig. 5. Simulation results for motion planning of a single vehicle. (a) Trajectory of vehicle R_1. (b) Position and orientation. (c) Optimal control policy. (d) Learned weight vector $\tilde{\mathbf{W}}_1$. (e) Learned weight vector $\tilde{\mathbf{W}}_1$. (f) Link capacity.

$\boldsymbol{\eta}_1 = [29.5, 29.5, 0, 1]^{\mathrm{T}}$ and $\boldsymbol{\eta}_2 = [29.5, 29.5, 0, 2]^{\mathrm{T}}$, respectively. In addition, the parameters of steepness and repulsive range are given as $c = 10$, $\upsilon = 10$, $\nu_1 = 1$, $c_1 = 10$.

1) IRL-Based Estimator for Online SNR Prediction: We first verify the effectiveness of the SNR estimator as proposed in Section III-A, where the actual channel parameters are set as $\boldsymbol{\theta}_{R_i} = [-60, 5]^{\mathrm{T}}$, $\xi_{R_i} = 0.07$, $\varphi_{R_i} = 3.51$ and $\rho_{R_i} = 0.2864$. Besides that, $\mathbf{Q}_1 = \mathrm{diag}([5, 5, 5, 5])$, $\lambda_1 = 0.5$, $\lambda_2 = 0.01$, and $T = 0.1$. Accordingly, the SNR estimator is adopted, and hence, the estimated path loss, shadowing, and multipath fading parameters are shown in Fig. 4(a)–(c), respectively. On the basis of this, the link capacity in underwater area can be shown in Fig. 4(d). The estimated parameters converge to the true values, which verify the effectiveness of the proposed SNR estimator in this article.

In [29], the least square estimator is adopted to estimate the SNR parameters, but the least square estimator can make the parameters fall into local optimum. To show the above phenomenon, we assume that the SNR measurement process is polluted by external noise, where the external noise is set as $\digamma_{\#} = 5000\cos^2(0.25K_{\mathrm{dB},1} + 5.3) + 2000\cos^2(0.5n_{\mathrm{PL},1} + 0.3)$. Take Part 3 as an example, the optimization problem can be updated as $\hat{\boldsymbol{\chi}}^*_{R_i} = \mathrm{argmin}\{\mathrm{Part\ 3} + \digamma_{\#}\}$. With the IRL-based estimator in this article, the predicted link capacity by using 3500 different sampling points is shown in Fig. 4(e). Meanwhile, the cost comparison by using the least square estimator (e.g., [29]) and the IRL-based estimator in this article is provided by Fig. 4(f). We find that the parameters estimation result of IRL-based estimator proposed in this article is closer to the true value than that of the least squares estimator.

2) Motion Planning of a Single Underwater Vehicle: With the predicted SNR information, we consider a simple motion planning scenario, i.e., a single underwater vehicle is deployed to relay the data from source sensor node to destination sensor node. Along with this, the value and policy basis functions of underwater vehicle R_1 can be expressed as $\boldsymbol{\phi}^1_2 = [e^2_{1,u_1}, e^2_{2,v_1}, e^2_{1,w_1}, e^2_{1,r_1}]^{\mathrm{T}}$, $\boldsymbol{\phi}^1_{4,1} = [2e_{1,u_1}]^{\mathrm{T}}$, $\boldsymbol{\phi}^1_{4,2} = [2e_{1,v_1}]^{\mathrm{T}}$, $\boldsymbol{\phi}^1_{4,3} = [2e_{1,w_1}]^{\mathrm{T}}$, and $\boldsymbol{\phi}^1_{4,4} = [2e_{1,r_1}]^{\mathrm{T}}$, respectively. Of note, e_{1,u_1},e_{1,v_1}, e_{1,w_1} and e_{1,r_1} denote the error components of underwater vehicle R_1 on surge, sway, depth, and yaw, respectively. Meanwhile, $\beta_1 = 100$, $\beta_2 = 1$, $\rho_{o1} = 6$, $\rho_{o2} = 6$, and $\bar{\mathbf{R}}_1 = \mathrm{diag}([0.1, 0.08, 1, 1.2])$. Thereby, the trajectory of underwater vehicle R_1 is presented in Fig. 5(a), whose position and orientation are shown in Fig. 5(b). Correspondingly, the optimal policy of underwater vehicle R_1 is shown in Fig. 5(c), where the learned weights are presented in Fig. 5(d) and (e). Based on this, the link capacity of underwater vehicle R_1 and the segmented link capacity $c_{0,1}$ and $c_{1,2}$ are shown in Fig. 5(f). Clearly, the collision avoidance can be guaranteed. Meanwhile, at the beginning, the link capacity of the underwater vehicle is very poor. Through the motion planning procedure, the link capacity gradually increases, wherein $c_{0,1}$ and $c_{1,2}$ become

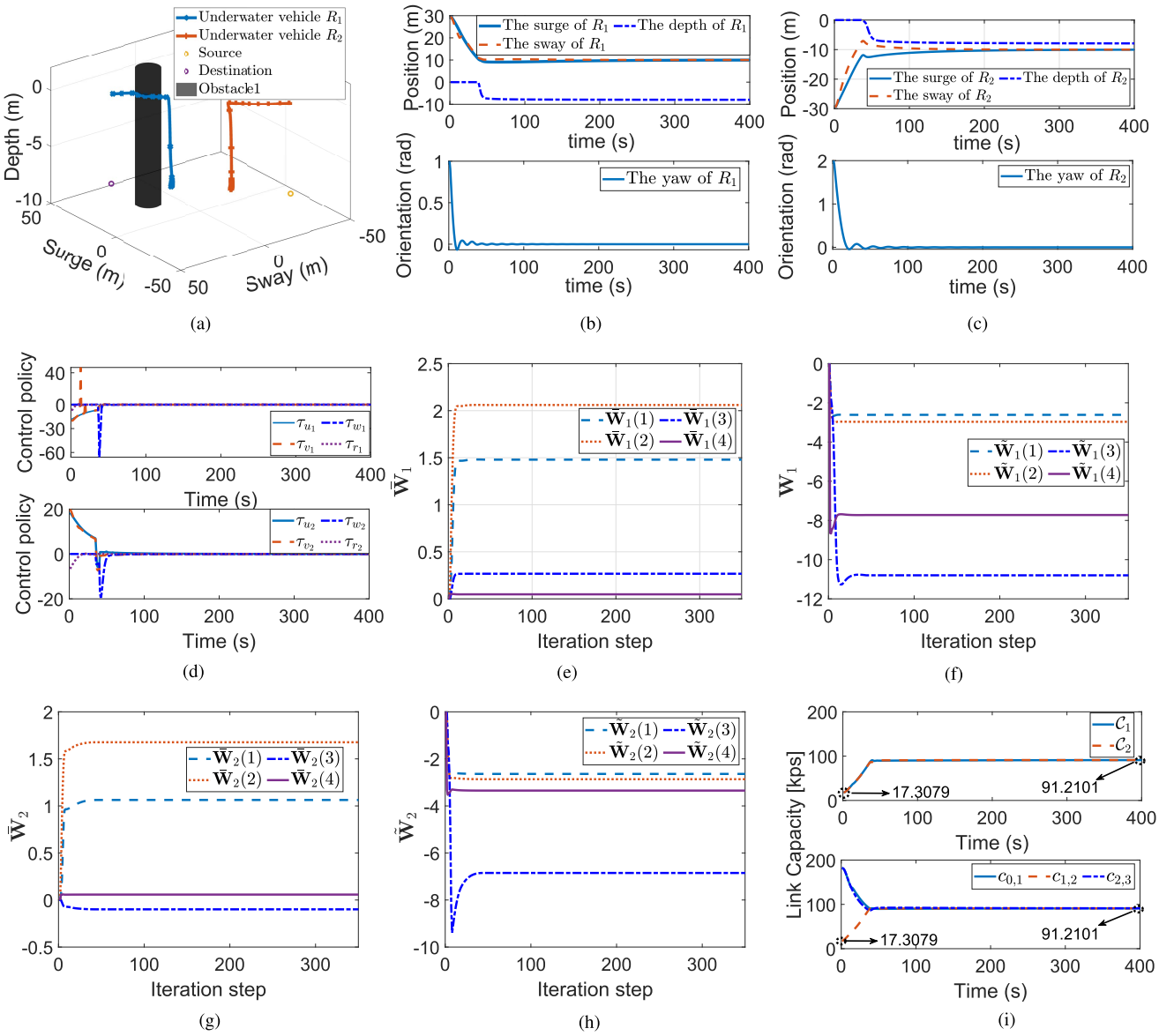


Fig. 6. Simulation results for the motion planning of multiple underwater vehicles. (a) Trajectories of two underwater vehicles. (b) Position and orientation of vehicle R_1. (c) Position and orientation of vehicle R_2. (d) Optimal control policies. (e) Learned weight vector $\tilde{\mathbf{W}}_1$. (f) Learned weight vector $\tilde{\mathbf{W}}_1$. (g) Learned weight vector $\tilde{\mathbf{W}}_2$. (h) Learned weight vector $\tilde{\mathbf{W}}_2$. (i) Link capacity for $\mathcal{E} = \{R_0, R_1, R_2, R_3\}$.

equal and reach stability after $t = 50$ s. Once $\mathcal{C}_1$ is maximized, underwater vehicle R_1 can hover at $(0, 0, -8)$. Overall, the link capacity is increased by 258.48% than the initial value, i.e., from 17.3079 to 62.0461.

3) Motion Planning of Multiunderwater Vehicle: Next, we consider a general motion planning scenario, i.e., two underwater vehicles are deployed to relay the data from source sensor node to destination sensor node. To this end, the value and policy basis functions of underwater vehicles are defined as the same in Section IV-A2. In addition, $\beta_1 = 100$, $\beta_2 = 1$ and $\bar{\mathbf{R}}_1 = \bar{\mathbf{R}}_2 = \text{diag}([0.1, 0.08, 1, 1.2])$. Accordingly, the trajectories of underwater vehicles R_1 and R_2 are shown in Fig. 6(a), whose position and orientation are presented in Fig. 6(b) and (c). Correspondingly, the optimal policies of underwater vehicles R_1 and R_2 are shown in Fig. 6(d), where the learned weights are presented in Fig. 6(e)–(h). Clearly, the collision avoidance is also be guaranteed, while the learned weights can converge to the optimal values. Based on this, the link capacities of the two underwater vehicles and the segmented link capacity $c_{0,1}$, $c_{1,2}$, and $c_{2,3}$ are shown in Fig. 6(i). From Fig. 6(i), we know the link capacity of the networks is gradually increased by the motion-planning process, where $c_{0,1}$, $c_{1,2}$, and $c_{2,3}$ become equal after $t = 38$ s and reach stability after $t = 50$ s. Once $\mathcal{C}_1$ and $\mathcal{C}_1$ are maximized, underwater vehicles R_1 and R_2 hover at $(10, 10, -8)$ and $(-10, -10, -8)$, respectively. Overall, the link capacity is increased by 426.99%, i.e., from 17.3079 to 91.2101. These results demonstrate the meaning and necessary of our communication-efficiency motion planning solution.

4) Comparison With the Other Motion Planning Solutions: Note that a Lyapunov guidance vector field (LGVF)-based motion planning algorithm was provided in [26], where the

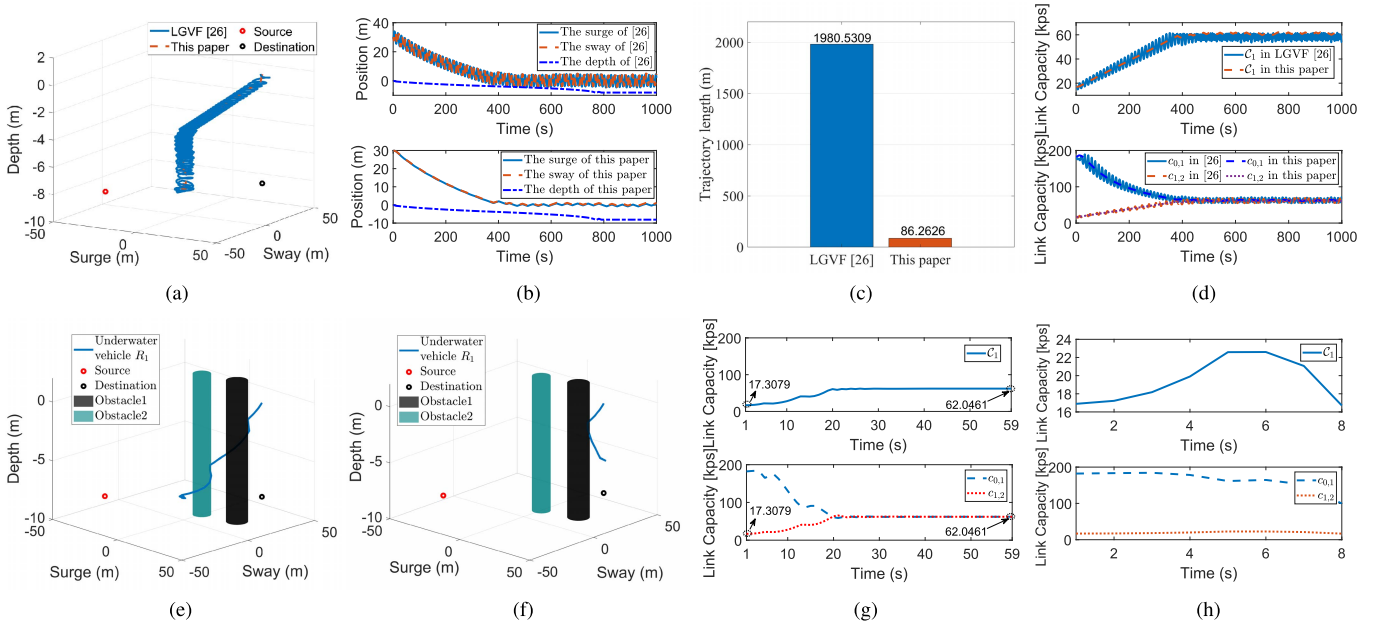


Fig. 7. Comparison for the IRL-based motion planning solution with the other existing solutions. (a) Comparison with LGVF [26]. (b) Position of R_1. (c) Trajectory length of two algorithms. (d) Link capacity. (e) Motion trajectory in Case 1. (f) Motion trajectory in Case 2. (g) Link capacity in Case 1. (h) Link capacity in Case 2.

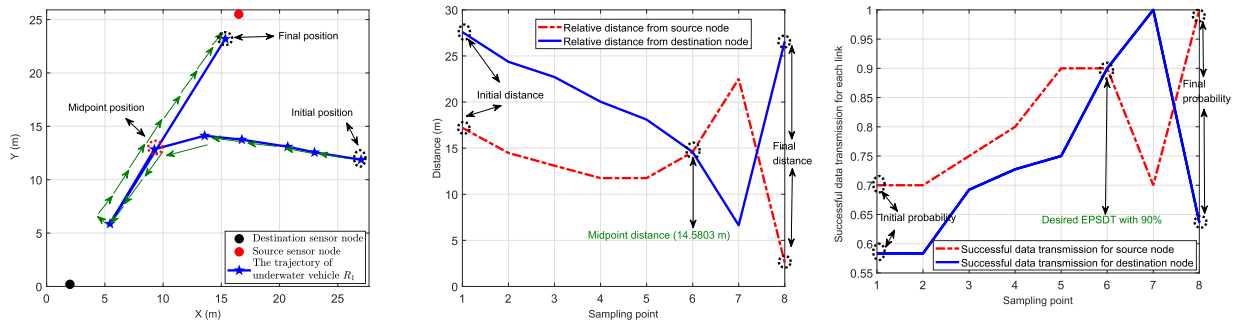
spiral forward was performed by vehicle. Clearly, the spiral forward can increase the path length of vehicle, which may reduce the lifetime of vehicle. By ignoring the underwater obstacle, the LGVF-based motion planning algorithm is adopted here, and hence, the trajectories of the underwater vehicle R_1 by using the above two algorithms are shown in Fig. 7(a). The positions on surge, sway, and depth are given in Fig. 7(b). The comparison of the path length required by the two algorithms is shown in Fig. 7(c). Meanwhile, the link capacities of the two algorithms are shown in Fig. 7(d). From Fig. 7(a)–(d), we can see that the link capacities by using the above two algorithms can both be improved; however, the path length required in this article is less than the one in [26] since the spiral forward is not required in this article.

Another important characteristic of our solution is the independence of system model, i.e., it is not necessary to know the nominal value of the underwater vehicle in advance. This characteristic is of great practical significance to ocean monitoring because it is difficult to obtain the nominal value in harsh underwater environment. With respect to this, the model-based learning controller (e.g., [33]) is adopted by the underwater vehicle. Then, the following two cases are considered: 1) the model matrix $\mathbf{M}_1$ can be accurately obtained and 2) the model matrix $\mathbf{M}_1$ cannot be accurately obtained due to environment noise and model uncertainty. By employing model-based learning approach, the motion trajectories of underwater vehicle R_1 in Case 1 and Case 2 are shown in Fig. 7(e) and (f), respectively. Correspondingly, the link capacities in Case 1 and Case 2 are presented in Fig. 7(g) and (h), respectively. Clearly, the underwater vehicle can achieve the motion planning task when the model information is accurate, and meanwhile, the link capacity is significantly improved. However, when the model information is inaccurate, the motion planning task of underwater vehicle cannot be well achieved, which can result in the failure of capacity improvement. The above results demonstrate that the model-free motion planning solution developed in this article is meaningful and necessary for underwater vehicles.

Fig. 8. Experiment deployment, where an underwater vehicle, a source sensor node, and a destination sensor node is included.

B. Experimental Results

This section presents the experimental results. As depicted in Fig. 5(a), the desired relay position for a single underwater vehicle is on the midpoint between the source sensor node and the destination sensor node. With regard to this, three nodes including an underwater vehicle, a source sensor node, and a destination sensor node are considered, as shown in Fig. 8. The work frequency band of the wireless communication system is within 21–27 kHz, and it adopts the orthogonal



(a) (b) (c)

Fig. 9. Experimental results for deployment of underwater vehicle. (a) Motion trajectory of underwater vehicle. (b) Relative distances with source and destination sensor nodes. (c) Probability of successful data transmission for each link.

frequency division multiple access (OFDM) mode. In order to overcome the multipath and Doppler effects, the cyclic prefix is extended and the frequency interval is enlarged. Due to this, our communication system has a stable communication rate of 300 b/s and a maximum communication rate of 2000 b/s. Different from the terrestrial environment, the SNR $S_{i-1,i}$ is affected by path loss, shadow fading, multipath fading, and various external underwater disturbance, especially in shallow water near the shore. In the experiment, with the change of distance, the $S_{i-1,i}$ value is about 0.15, so the communication rate is 300 b/s. For underwater vehicle, the BlueROV from Blue Robotics is adopted, which features six thrusters, a flight controller, a wireless communication unit, and a Raspberry Pi.

In the following, the underwater vehicle patrols in different positions to relay the data of source sensor node to the destination sensor node, whose motion trajectory is shown in Fig. 9(a). For clear description, the relative distance between underwater vehicle and source (or destination) sensor node is provided in Fig. 9(b). Correspondingly, the end-to-end probability of successful data transmission (EPSDT) is shown in Fig. 9(c), which is defined as the successful data transmission of the worst link. We find that the successful data transmission for each communication link is increased with the increase of relative distance. Meanwhile, the successful data transmission can reach to the same value (i.e., 90%) when the relative distances for the two links are the same (i.e., 14.5803 m). Based on the definition of EPSDT, we can know that the EPSDT can reach to the maximum when the underwater vehicle is on the midpoint between the source sensor node and the destination sensor node. It is clear that these results are consistent with the simulation results. Similarly, the results for multiple underwater vehicles can also be obtained, and this part is omitted here due to page limitation.

V. Conclusion

This article gives a communication-efficient and collision-free motion planning solution for underwater vehicles. By adopting the model-based IRL, an online SNR estimator is designed to capture the unknown shadowing and multipath parameters, such that the SNR in unvisited positions can be predicted by underwater vehicles. With the predicted channel information, a model-free IRL motion algorithm is conducted to drive underwater vehicles to the desired position points while maximizing the communication capacity and avoiding the collision. Finally, simulation and experimental results are both presented to verify the effectiveness.

In the future, we will employ the distributed learning approach to resolve the codesign problem of underwater detection, communication, and control. Meanwhile, how to verify the results in ocean environment is also our future work.

Appendix A
Proof of Theorem 1

Given an initial admissible policy $\mathbf{u}^{(0)}(0)$ with the system trajectory of $\dot{\hat{\boldsymbol{\chi}}}_{R_i} = \mathbf{u}^{(s+1)}$, the task of this proof is to prove $V_1^*(\hat{\boldsymbol{\chi}}_{R_i}, \mathbf{u}^*) \le V_1^{(s+1)}(\hat{\boldsymbol{\chi}}_{R_i}, \mathbf{u}^{(s+1)}) \le V_1^{(s)}(\hat{\boldsymbol{\chi}}_{R_i}, \mathbf{u}^{(s)})$ where $V_1^*(\hat{\boldsymbol{\chi}}_{R_i}, \mathbf{u}^*) = \min_{\mathbf{u}} \int_0^\infty g_1(\hat{\boldsymbol{\chi}}_{R_i}, \mathbf{u}) \mathrm{d}\tau$.

To this end, we set $\mathbf{u}^{(s)}$ as an admissible policy. Based on this, we take the derivative of $V_1^{(s)}(\hat{\boldsymbol{\chi}}_{R_i})$ along $\dot{\hat{\boldsymbol{\chi}}}_{R_i} = \mathbf{u}^{(s+1)}$, through which one has

$$\dot{V}_1^{(s)}(\hat{\boldsymbol{\chi}}_{R_i}, \mathbf{u}^{(s+1)}) = \left(\frac{\partial V_1^{(s)}(\hat{\boldsymbol{\chi}}_{R_i})}{\partial \hat{\boldsymbol{\chi}}_{R_i}}\right)^{\mathrm{T}} \mathbf{u}^{(s+1)}. \tag{51}$$

According to the definition of $V_1^*(\hat{\boldsymbol{\chi}}_{R_i}, \mathbf{0})$, the Hamilton–Jacobi–Bellman (HJB) equation becomes

$$\left(\frac{\partial V_1^{(s)}(\hat{\boldsymbol{\chi}}_{R_i})}{\partial \hat{\boldsymbol{\chi}}_{R_i}}\right)^{\mathrm{T}} \mathbf{u}^{(s)} + g_1(\hat{\boldsymbol{\chi}}_{R_i}, \mathbf{u}^{(s)}) = 0. \tag{52}$$

Combining (51) with (52), we can get

$$\dot{V}_1^{(s)}(\hat{\boldsymbol{\chi}}_{R_i}, \mathbf{u}^{(s+1)}) = \left(\frac{\partial V_1^{(s)}(\hat{\boldsymbol{\chi}}_{R_i})}{\partial \hat{\boldsymbol{\chi}}_{R_i}}\right)^{\mathrm{T}} \left(\mathbf{u}^{(s+1)} - \mathbf{u}^{(s)}\right) - g_1(\hat{\boldsymbol{\chi}}_{R_i}, \mathbf{u}^{(s)}). \tag{53}$$

With (26), we have $((\partial V_1^{(s)}(\hat{\boldsymbol{\chi}}_{R_i}))/(\partial \hat{\boldsymbol{\chi}}_{R_i}))^{\mathrm{T}} = -2(\mathbf{u}^{(s+1)})^{\mathrm{T}} \mathbf{R}_1$. Based on this, one can rearrange (53) as

$$\begin{aligned}
&\dot{V}_1^{(s)}(\hat{\boldsymbol{\chi}}_{R_i}, \mathbf{u}^{(s+1)}) \\
&\quad = -2\mathbf{u}^{(s+1)\mathrm{T}} \mathbf{R}_1 \left(\mathbf{u}^{(s+1)} - \mathbf{u}^{(s)}\right) - g_1(\hat{\boldsymbol{\chi}}_{R_i}, \mathbf{u}^{(s)}) \\
&\quad = -\mathbf{u}^{(s+1)\mathrm{T}} \mathbf{R}_1 \mathbf{u}^{(s+1)} - \lambda_1 \left(\frac{1}{|\mathcal{N}_i|_i} \boldsymbol{\Delta}_{R_i}^{\mathrm{T}} \boldsymbol{\Delta}_{R_i} - \hat{\xi}_{R_i}^2 - \hat{\rho}_{R_i}^2\right)^2 \\
&\qquad - \left(\mathbf{u}^{(s+1)} - \mathbf{u}^{(s)}\right)^{\mathrm{T}} \mathbf{R}_1 \left(\mathbf{u}^{(s+1)} - \mathbf{u}^{(s)}\right)
\end{aligned}$$

$$
\begin{aligned}
&- \sum_{(j_i,\vec{j}_i)\in\mathbf{G}} \lambda_2\left[\ln\hat{\xi}_{R_i}^2 - \frac{d_{j_i,\vec{j}_i}}{\hat{\varphi}_{R_i}} - \ln([\mathbf{\Delta}]_{j_i}[\mathbf{\Delta}]_{\vec{j}_i})\right]^2 \\
&- \mathbf{\Delta}_{R_i}^{\mathrm{T}}\mathbf{Q}_1\mathbf{\Delta}_{R_i} \\
&= -\big(\mathbf{u}^{(s+1)} - \mathbf{u}^{(s)}\big)^{\mathrm{T}}\mathbf{R}_1\big(\mathbf{u}^{(s+1)} - \mathbf{u}^{(s)}\big) \\
&- g_1(\hat{\boldsymbol{\chi}}_{R_i}, \mathbf{u}^{(s+1)}). \qquad (54)
\end{aligned}
$$

Clearly, $\dot{V}_1^{(s)}(\hat{\boldsymbol{\chi}}_{R_i}, \mathbf{u}^{(s+1)}) \le 0$ is always satisfied. Note that $V_1^{(s)}(\hat{\boldsymbol{\chi}}_{R_i}, \mathbf{u}^{(s+1)})$ is positive definite, continuous, and differentiable. In view of this, $V_1^{(s)}(\hat{\boldsymbol{\chi}}_{R_i}, \mathbf{u}^{(s+1)})$ can be regarded as the Lyapunov function of $\mathbf{u}^{(s+1)}$, through which one knows $\mathbf{u}^{(s)}$ and $\mathbf{u}^{(s+1)}$ are both the admissible policies.

Based on the above conclusion, we require to prove that $V_1^{(s+1)}(\hat{\boldsymbol{\chi}}_{R_i}, \mathbf{u}^{(s+1)}) - V_1^{(s)}(\hat{\boldsymbol{\chi}}_{R_i}, \mathbf{u}^{(s)}) \le 0$ in the trajectory along $\dot{\hat{\boldsymbol{\chi}}}_{R_i} = \mathbf{u}^{(s+1)}$. Combining (51) with (52), we have

$$
\begin{aligned}
&V_1^{(s)}(\hat{\boldsymbol{\chi}}_{R_i}, \mathbf{u}^{(s)}) - V_1^{(s+1)}(\hat{\boldsymbol{\chi}}_{R_i}, \mathbf{u}^{(s+1)}) \\
&= \int_0^\infty g_1(\hat{\boldsymbol{\chi}}_{R_i}, \mathbf{u}^{(s)})\mathrm{d}\tau - \int_0^\infty g_1(\hat{\boldsymbol{\chi}}_{R_i}, \mathbf{u}^{(s+1)})\mathrm{d}\tau \\
&= \int_0^\infty \left(\frac{\partial V_1^{(s+1)}(\hat{\boldsymbol{\chi}}_{R_i})}{\partial \hat{\boldsymbol{\chi}}_{R_i}}\right)^{\mathrm{T}} \mathbf{u}^{(s+1)}\mathrm{d}\tau \\
&\quad - \int_0^\infty \left(\frac{\partial V_1^{(s)}(\hat{\boldsymbol{\chi}}_{R_i})}{\partial \hat{\boldsymbol{\chi}}_{R_i}}\right)^{\mathrm{T}} \mathbf{u}^{(s+1)}\mathrm{d}\tau \\
&= \int_0^\infty \Big(\dot{V}_1^{(s+1)}(\hat{\boldsymbol{\chi}}_{R_i}, \mathbf{u}^{(s+1)}) - \dot{V}_1^{(s)}(\hat{\boldsymbol{\chi}}_{R_i}, \mathbf{u}^{(s+1)})\Big)\mathrm{d}\tau \\
&= \int_0^\infty \Big(-g_1(\hat{\boldsymbol{\chi}}_{R_i}, \mathbf{u}^{(s+1)}) - \dot{V}_1^{(s)}(\hat{\boldsymbol{\chi}}_{R_i}, \mathbf{u}^{(s+1)})\Big)\mathrm{d}\tau \\
&= \int_0^\infty \big(\mathbf{u}^{(s+1)} - \mathbf{u}^{(s)}\big)^{\mathrm{T}}\mathbf{R}_1\big(\mathbf{u}^{(s+1)} - \mathbf{u}^{(s)}\big)\mathrm{d}\tau. \qquad (55)
\end{aligned}
$$

Therefore, one can easily obtain $V_1^{(s+1)}(\hat{\boldsymbol{\chi}}_{R_i}, \mathbf{u}^{(s+1)}) \le V_1^{(s)}(\hat{\boldsymbol{\chi}}_{R_i}, \mathbf{u}^{(s)})$. Noting with the definition of $V_1^*(\hat{\boldsymbol{\chi}}_{R_i}, \mathbf{u}^*)$, one can further have $V_1^*(\hat{\boldsymbol{\chi}}_{R_i}, \mathbf{u}^*) \le V_1^{(s+1)}(\hat{\boldsymbol{\chi}}_{R_i}, \mathbf{u}^{(s+1)}) \le V_1^{(s)}(\hat{\boldsymbol{\chi}}_{R_i}, \mathbf{u}^{(s)})$, which means the optimal value of $\hat{\boldsymbol{\chi}}_{R_i}$ (i.e., $\hat{\boldsymbol{\chi}}_{R_i}^*$) can also be obtained. That completes the proof.

Appendix B
Proof of Theorem 2

Given an initial admissible policy $\boldsymbol{\tau}_i^{(0)}(0)$ with system trajectory of $\dot{\bar{\mathbf{X}}}_i = \mathbf{A}_i\mathbf{X}_i + \mathbf{B}_i\boldsymbol{\tau}_i^{(s)} + \mathbf{G}_i$, the task of this proof is to prove the solution to the model-free Bellman equation is the same as the model-based Bellman equation.

First, the form of model-based Bellman equation is established. Differentiating the value function (38), we have the following Bellman equation, i.e.,

$$
\begin{aligned}
\bar{H}\big(\bar{V}_i^{(s)}, \boldsymbol{\tau}_i^{(s)}\big) &= \nabla\bar{V}_{\mathbf{X}_i}^{(s)\mathrm{T}}\big(\mathbf{A}_i\hat{\mathbf{X}}_i + \mathbf{B}_i\boldsymbol{\tau}_i^{(s)} + \mathbf{G}_i\big) + \bar{g}_i\big(\hat{\mathbf{p}}_i, \boldsymbol{\tau}_i^{(s)}\big) \\
&= 0. \qquad (56)
\end{aligned}
$$

By the stationarity condition $\partial H(\bar{V}_i^{(s)}, \boldsymbol{\tau}_i^{(s)})/\partial\boldsymbol{\tau}_i^{(s)} = 0$, we can have the following optimal control, i.e.,

$$
\boldsymbol{\tau}_i^{(s+1)} = -\frac{1}{2}\bar{\mathbf{R}}_i^{-1}\mathbf{B}_i^{\mathrm{T}}\nabla\bar{V}_{\mathbf{X}_i}^{(s)}. \qquad (57)
$$

From (56) and (57), one obtains

$$
\begin{aligned}
\bar{H}\big(\bar{V}_i^{(s)}, \boldsymbol{\tau}_i^{(s)}\big) &= \nabla\bar{V}_{\mathbf{X}_i}^{(s)\mathrm{T}}\big(\mathbf{A}_i\hat{\mathbf{X}}_i + \mathbf{B}_i\boldsymbol{\tau}_i^{(s)} + \mathbf{G}_i\big) + \boldsymbol{\tau}_i^{(s)\mathrm{T}}\bar{\mathbf{R}}_i\boldsymbol{\tau}_i^{(s)} \\
&+ \frac{\beta_1}{\mathcal{C}_i^2(\hat{\mathbf{p}}_i)} + \beta_2\sum_{m\in\Omega_i'} f(\hat{l}_{i,m})\left(\frac{1}{\hat{l}_{i,m}} - \frac{1}{\rho_{om}}\right)^2 \\
&+ \beta_3\sum_{j\in\mathcal{N}_i^*}\exp\left(-\frac{1}{2}\left(\frac{l_{i,j}^2}{\nu_1^2}\right)^{c_1}\right)\frac{1}{l_{i,j}^2} \\
&= 0. \qquad (58)
\end{aligned}
$$

On the other hand, the form of model-free Bellman equation can also be established. From (41), we have

$$
\begin{aligned}
H\big(\bar{V}_i, \boldsymbol{\tau}_i^{(s)}\big) &= \nabla\bar{V}_{\mathbf{X}_i}^{(s)\mathrm{T}}\big(\mathbf{A}_i\hat{\mathbf{X}}_i + \mathbf{B}_i\boldsymbol{\tau}_i^{(s)} + \mathbf{B}_i\big(\boldsymbol{\tau}_i - \boldsymbol{\tau}_i^{(s)}\big) + \mathbf{G}_i\big) \\
&+ \bar{g}_i\big(\hat{\mathbf{p}}_i, \boldsymbol{\tau}_i^{(s)}\big) + 2\boldsymbol{\tau}_i^{(s+1)\mathrm{T}}\bar{\mathbf{R}}_i\big(\boldsymbol{\tau}_i - \boldsymbol{\tau}_i^{(s)}\big) \\
&= 0. \qquad (59)
\end{aligned}
$$

Substituting (57) into (59), we can have the same Bellman equation as (56). Therefore, the optimal solution $\boldsymbol{\tau}_i^*$ to the Bellman function in model-free equation is the same as that of the Bellman function in model-based equation.

Note that the convergence of the solution to model-based Bellman equation [i.e., (57)] has been proven in Theorem 1. Based on this, one knows the policy iterations (42) and (43) can make $\boldsymbol{\tau}_i^{(s+1)}$ converge to the optimal policy $\boldsymbol{\tau}_i^*$. That completes the proof.

Appendix C
Proof of Corollary 1

The proof includes two parts. First, we prove that $\bar{V}_i^{(s)}(\hat{\mathbf{X}}_i(t), \boldsymbol{\tau}_i^{(s)})$ is decreasing function with policy $\boldsymbol{\tau}_i^{(s+1)} = (-1/2)\bar{\mathbf{R}}_i^{-1}\mathbf{B}_i^{\mathrm{T}}\nabla\bar{V}_{\mathbf{X}_i}^{(s)}$. Taking the derivative of $\bar{V}_i(\hat{\mathbf{X}}_i(t), \boldsymbol{\tau}_i^{(s+1)})$ along $\dot{\hat{\mathbf{X}}}_i = \mathbf{A}_i\hat{\mathbf{X}}_i + \mathbf{B}_i\boldsymbol{\tau}_i^{(s+1)} + \mathbf{G}_i$, we have

$$
\dot{\bar{V}}_i^{(s)}\big(\hat{\mathbf{X}}_i(t), \boldsymbol{\tau}_i^{(s+1)}\big) = \nabla\bar{V}_{\hat{\mathbf{X}}_i}^{(s)}\big(\mathbf{A}_i\hat{\mathbf{X}}_i + \mathbf{B}_i\boldsymbol{\tau}_i^{(s+1)} + \mathbf{G}_i\big). \qquad (60)
$$

According to (40), we have

$$
\nabla\bar{V}_{\mathbf{X}_i}^{(s)\mathrm{T}}\mathbf{G}_i = -\nabla\bar{V}_{\mathbf{X}_i}^{(s)\mathrm{T}}\big(\mathbf{A}_i\hat{\mathbf{X}}_i + \mathbf{B}_i\boldsymbol{\tau}_i^{(s)}\big) - \bar{g}_i\big(\hat{\mathbf{p}}_i, \boldsymbol{\tau}_i^{(s)}\big). \qquad (61)
$$

Combining $\boldsymbol{\tau}_i^{(s+1)} = -(1/2)\bar{\mathbf{R}}_i^{-1}\mathbf{B}_i^{\mathrm{T}}\nabla\bar{V}_{\mathbf{X}_i}^{(s)}$ with (60) and (61), one can further have

$$
\begin{aligned}
&\dot{\bar{V}}_i^{(s)}\big(\hat{\mathbf{X}}_i(t), \boldsymbol{\tau}_i^{(s+1)}\big) \\
&= -2\boldsymbol{\tau}_i^{(s+1)\mathrm{T}}\bar{\mathbf{R}}_i\big(\boldsymbol{\tau}_i^{(s+1)} - \boldsymbol{\tau}_i^{(s)}\big) - \bar{g}_i\big(\hat{\mathbf{p}}_i, \boldsymbol{\tau}_i^{(s)}\big) \\
&= -\boldsymbol{\tau}_i^{(s+1)\mathrm{T}}\bar{\mathbf{R}}_i\boldsymbol{\tau}_i^{(s+1)} - \big(\boldsymbol{\tau}_i^{(s+1)} - \boldsymbol{\tau}_i^{(s)}\big)^{\mathrm{T}}\bar{\mathbf{R}}_i\big(\boldsymbol{\tau}_i^{(s+1)} - \boldsymbol{\tau}_i^{(s)}\big) \\
&\quad - \beta_2 E_{i,\mathrm{O}}(\hat{\mathbf{p}}_i) - \beta_3 E_{i,j}(\hat{\mathbf{p}}_i) \le 0 \qquad (62)
\end{aligned}
$$

which means $\bar{V}_i^{(s)}(\hat{\mathbf{X}}_i(t), \boldsymbol{\tau}_i^{(s+1)}) \le \bar{V}_i^{(s)}(\hat{\mathbf{X}}_i(t), \boldsymbol{\tau}_i^{(s)})$. Similarly, the equivalence between the solution of IRL and the optimal policy solution $\boldsymbol{\tau}_i^{(s+1)}$ can be proved by Theorem 2.

Next, we prove that the collision never occurs between underwater vehicle R_i and obstacles or neighbor R_j if (50) are satisfied. Assume that at $t = t^*$, underwater vehicle R_i collides with obstacles or neighbor R_j, then cost function becomes

$$
\begin{aligned}
\bar{g}_i\big(\hat{\mathbf{p}}_i, \tau_i^{(s)}\big)\big|_{t=t^*} &= \beta_1 E_{i,\mathrm{L}}(\hat{\mathbf{p}}_i)|_{t=t^*} + \boldsymbol{\tau}_i^{(s)\mathrm{T}}\bar{\mathbf{R}}_i\boldsymbol{\tau}_i^{(s)}\big|_{t=t^*} \\
&+ \max\big\{\beta_2 E_{i,\mathrm{O}}^{\max}, \beta_3 E_{i,j}^{\max}\big\}. \qquad (63)
\end{aligned}
$$

Then, we can conclude that

$$
\bar{g}_i\big(\hat{\mathbf{p}}_i, \tau_i^{(s)}\big)\big|_{t=t^*} > \min\big\{\beta_2 E_{i,\mathrm{O}}^{\max}, \beta_3 E_{i,j}^{\max}\big\}. \qquad (64)
$$

From (62), (63), and (64), we have

$$\bar{g}_i\left(\hat{\mathbf{p}}_i, \tau_i^{(s+1)}\right)\Big|_{t=t^*} < \bar{g}_i\left(\hat{\mathbf{p}}_i, \tau_i^{(s)}\right)\Big|_{t=t^*} < \bar{g}_i\left(\hat{\mathbf{p}}_i, \tau_i^{(0)}\right)\Big|_{t=0} < \min\left\{\beta_2 E_{i,\mathrm{O}}^{\max}, \beta_3 E_{i,j}^{\max}\right\} \quad (65)$$

which means (65) has contradiction with (64). Therefore, underwater vehicle R_i never collides with obstacles or neighbor R_j if (50) is satisfied. That completes the proof.

REFERENCES

[1] Z. Peng, D. Wang, and J. Wang, "Data-driven adaptive disturbance observers for model-free trajectory tracking control of maritime autonomous surface ships," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 32, no. 12, pp. 5584–5594, Dec. 2021.

[2] X. Geng and Y. R. Zheng, "Exploiting propagation delay in underwater acoustic communication networks via deep reinforcement learning," *IEEE Trans. Neural Netw. Learn. Syst.*, early access, May 17, 2022, doi: 10.1109/TNNLS.2022.3170050.

[3] A. Vasilijevic, D. Nad, F. Mandic, N. Miskovic, and Z. Vukic, "Coordinated navigation of surface and underwater marine robotic vehicles for ocean sampling and environmental monitoring," *IEEE/ASME Trans. Mechatronics*, vol. 22, no. 3, pp. 1174–1184, Jun. 2017.

[4] I. Jawhar, N. Mohamed, J. Al-Jaroodi, and Z. Sheng, "An architecture for using autonomous underwater vehicles in wireless sensor networks for underwater pipeline monitoring," *IEEE Trans. Ind. Informat.*, vol. 15, no. 3, pp. 1329–1340, Mar. 2019.

[5] R. Su, D. Zhang, C. Li, Z. Gong, R. Venkatesan, and F. Jiang, "Localization and data collection in AUV-aided underwater sensor networks: Challenges and opportunities," *IEEE Netw.*, vol. 33, no. 6, pp. 86–93, Nov. 2019.

[6] Y. Jing, Y. Xian, X. Luo, and C. Chen, "Energy-efficient data collection over AUV-assisted underwater acoustic sensor network," *IEEE Syst. J.*, vol. 12, no. 4, pp. 3519–3530, Dec. 2018.

[7] H. Yetkin, C. Lutz, and D. Stilwell, "A decision-theoretic approach to acquire environmental information for improved subsea search performance," *Ocean Eng.*, vol. 204, no. 1, pp. 1–12, May 2020.

[8] M. T. R. Khan, Y. Z. Jembre, S. H. Ahmed, J. Seo, and D. Kim, "Data freshness based AUV path planning for UWSN in the internet of underwater things," in *Proc. IEEE Global Commun. Conf. (GLOBECOM)*, Dec. 2019, pp. 1–6.

[9] V. Yordanova and B. Gips, "Coverage path planning with track spacing adaptation for autonomous underwater vehicles," *IEEE Robot. Autom. Lett.*, vol. 5, no. 3, pp. 4774–4780, Jul. 2020.

[10] Y. Wang et al., "Acoustic camera-based pose graph SLAM for dense 3-D mapping in underwater Environments," *IEEE J. Ocean Eng.*, vol. 46, no. 3, pp. 829–847, Jul. 2021.

[11] M. Boban, T. T. V. Vinhoza, M. Ferreira, J. Barros, and O. K. Tonguz, "Impact of vehicles as obstacles in vehicular ad hoc networks," *IEEE J. Sel. Areas Commun.*, vol. 29, no. 1, pp. 15–28, Jan. 2011.

[12] Z. Song, D. Lipinski, and K. Mohseni, "Multi-vehicle cooperation and nearly fuel-optimal flock guidance in strong background flows," *Ocean Eng.*, vol. 141, pp. 388–404, Sep. 2017.

[13] C. Lin, G. Han, J. Du, Y. Bi, L. Shu, and K. Fan, "A path planning scheme for AUV flock-based internet-of-underwater-things systems to enable transparent and smart ocean," *IEEE Internet Things J.*, vol. 7, no. 10, pp. 9760–9772, Oct. 2020.

[14] S. Mahmoudzadeh, D. M. W. Powers, and A. Atyabi, "UUV's hierarchical DE-based motion planning in a semi dynamic underwater wireless sensor network," *IEEE Trans. Cybern.*, vol. 49, no. 8, pp. 2992–3005, Aug. 2019.

[15] S. Heshmati-Alamdari, A. Nikou, and D. V. Dimarogonas, "Robust trajectory tracking control for underactuated autonomous underwater vehicles in uncertain environments," *IEEE Trans. Autom. Sci. Eng.*, vol. 18, no. 3, pp. 1288–1301, Jul. 2021.

[16] Y. Shou, B. Xu, H. Lu, A. Zhang, and T. Mei, "Finite-time formation control and obstacle avoidance of multi-agent system with application," *Int. J. Robust Nonlinear Control*, vol. 32, no. 5, pp. 2883–2901, Mar. 2022.

[17] B. K. Sahu and B. Subudhi, "Flocking control of multiple AUVs based on fuzzy potential functions," *IEEE Trans. Fuzzy Syst.*, vol. 26, no. 5, pp. 2539–2551, Oct. 2018.

[18] J. Wang, Z. Wu, M. Tan, and J. Yu, "3-D path planning with multiple motions for a gliding robotic dolphin," *IEEE Trans. Syst., Man, Cybern. Syst.*, vol. 51, no. 5, pp. 2904–2915, May 2021.

[19] X. Yu, W.-N. Chen, T. Gu, H. Yuan, H. Zhang, and J. Zhang, "ACO-A*: Ant colony optimization plus A* for 3-D traveling in environments with dense obstacles," *IEEE Trans. Evol. Comput.*, vol. 23, no. 4, pp. 617–631, Aug. 2019.

[20] L. Jiang, H. Huang, and Z. Ding, "Path planning for intelligent robots based on deep Q-learning with experience replay and heuristic knowledge," *IEEE/CAA J. Autom. Sinica*, vol. 7, no. 4, pp. 1179–1189, Jul. 2020.

[21] G. P. Kontoudis and K. G. Vamvoudakis, "Kinodynamic motion planning with continuous-time Q-learning: An online, model-free, and safe navigation framework," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 30, no. 12, pp. 3803–3817, Dec. 2019.

[22] A. Stefanov and M. Stojanovic, "Design and performance analysis of underwater acoustic networks," *IEEE J. Sel. Areas Commun.*, vol. 29, no. 10, pp. 2012–2021, Dec. 2011.

[23] Y. He, G. Han, Z. Tang, M. Martinez-Garcia, and Y. Peng, "State prediction-based data collection algorithm in underwater acoustic sensor networks," *IEEE Trans. Wireless Commun.*, vol. 21, no. 4, pp. 2830–2842, Apr. 2022.

[24] X. Cao, J. Zhang, and H. V. Poor, "Joint energy procurement and demand response towards optimal deployment of renewables," *IEEE J. Sel. Topics Signal Process.*, vol. 12, no. 4, pp. 657–672, Aug. 2018.

[25] X. Cao et al., "Joint estimation of clock skew and offset in pairwise broadcast synchronization mechanism," *IEEE Trans. Commun.*, vol. 61, no. 6, pp. 2508–2521, Jun. 2013.

[26] C. Dixon and E. W. Frew, "Optimizing cascaded chains of unmanned aircraft acting as communication relays," *IEEE J. Sel. Areas Commun.*, vol. 30, no. 5, pp. 883–898, Jun. 2012.

[27] Y. Guo, C. You, C. Yin, and R. Zhang, "UAV trajectory and communication co-design: Flexible path discretization and path compression," *IEEE J. Sel. Areas Commun.*, vol. 39, no. 11, pp. 3506–3523, Nov. 2021.

[28] Q. Wu, Y. Zeng, and R. Zhang, "Joint trajectory and communication design for multi-UAV enabled wireless networks," *IEEE Trans. Wireless Commun.*, vol. 17, no. 3, pp. 2109–2121, Mar. 2018.

[29] Y. Yan and Y. Mostofi, "Robotic router formation in realistic communication environments," *IEEE Trans. Robot.*, vol. 28, no. 4, pp. 810–827, Aug. 2012.

[30] U. Ali, H. Cai, Y. Mostofi, and Y. Wardi, "Motion-communication co-optimization with cooperative load transfer in mobile robotics: An optimal control perspective," *IEEE Trans. Control Netw. Syst.*, vol. 6, no. 2, pp. 621–632, Jun. 2019.

[31] D. B. Licea, M. Bonilla, M. Ghogho, S. Lasaulce, and V. S. Varma, "Communication-aware energy efficient trajectory planning with limited channel knowledge," *IEEE Trans. Robot.*, vol. 36, no. 2, pp. 431–442, Apr. 2020.

[32] M. Chen et al., "Distributed learning in wireless networks: Recent progress and future challenges," *IEEE J. Sel. Areas Commun.*, vol. 39, no. 12, pp. 3579–3605, Dec. 2021.

[33] J. Yan, X. Li, X. Yang, X. Luo, C. Hua, and X. Guan, "Integrated localization and tracking for AUV with model uncertainties via scalable sampling-based reinforcement learning approach," *IEEE Trans. Syst., Man, Cybern. Syst.*, vol. 52, no. 11, pp. 6952–6967, Nov. 2022.

[34] J. Yan, Y. Meng, X. Yang, X. Luo, and X. Guan, "Privacy-preserving localization for underwater sensor networks via deep reinforcement learning," *IEEE Trans. Inf. Forensics Security*, vol. 16, pp. 1880–1895, 2021.

[35] R. Cui, L. Chen, C. Yang, and M. Chen, "Extended state observer-based integral sliding mode control for an underwater robot with unknown disturbances and uncertain nonlinearities," *IEEE Trans. Ind. Electron.*, vol. 64, no. 8, pp. 6785–6795, Aug. 2017.

[36] R. Cui, C. Yang, Y. Li, and S. Sharma, "Adaptive neural network control of AUVs with control input nonlinearities using reinforcement learning," *IEEE Trans. Syst., Man, Cybern., Syst.*, vol. 47, no. 6, pp. 1019–1029, Jun. 2017.

[37] T. Rappaport, *Wireless Communications, Principles and Practice.* Upper Saddle River, NJ, USA: Prentice-Hall, 2002.

[38] Y. Zhang, Z. Zhang, L. Chen, and X. Wang, "Reinforcement learning-based opportunistic routing protocol for underwater acoustic sensor networks," *IEEE Trans. Veh. Technol.*, vol. 70, no. 3, pp. 2756–2770, Mar. 2021.

[39] D. E. Lucani, M. Medard, and M. Stojanovic, "Underwater acoustic networks: Channel models and network coding based lower bound to transmission power for multicast," *IEEE J. Sel. Areas Commun.*, vol. 26, no. 9, pp. 1708–1719, Dec. 2008.

[40] W. Hurst, H. Cai, and Y. Mostofi, "Communication-aware RRT: Path planning for robotic communication operation in obstacle environments," in *Proc. IEEE Int. Conf. Commun. (ICC)*, Jun. 2021, pp. 1–6.
[41] A. Goldsmith, *Wireless Communication*. Cambridge, U.K.: Cambridge Univ. Press, 2005.
[42] W. Cao, J. Yan, X. Yang, X. Luo, and X. Guan, "Communication-aware formation control of AUVs with model uncertainty and fading channel via integral reinforcement learning," *IEEE/CAA J. Autom. Sinica*, vol. 10, no. 1, pp. 159–176, Jan. 2023.
[43] M. Liu, Y. Wan, F. L. Lewis, and V. G. Lopez, "Adaptive optimal control for stochastic multiplayer differential games using on-policy and off-policy reinforcement learning," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 31, no. 12, pp. 5522–5533, Dec. 2020.
[44] C. He, Y. Wan, Y. Gu, and F. L. Lewis, "Integral reinforcement learning-based multi-robot minimum time-energy path planning subject to collision avoidance and unknown environmental disturbances," *IEEE Control Syst. Lett.*, vol. 5, no. 3, pp. 983–988, Jul. 2021.
[45] J. Zhang, J. Sha, G. Han, J. Liu, and Y. Qian, "A cooperative-control-based underwater target escorting mechanism with multiple autonomous underwater vehicles for underwater Internet of Things," *IEEE Internet Things J.*, vol. 8, no. 6, pp. 4403–4416, Mar. 2021.
[46] X. Shi and J. Wu, "To hide private position information in localization using time difference of arrival," *IEEE Trans. Signal Process.*, vol. 66, no. 18, pp. 4946–4956, Sep. 2018.

Jing Yan (Senior Member, IEEE) received the B.Eng. degree in automation from Henan University, Kaifeng, China, in 2008, and the Ph.D. degree in control theory and control engineering from Yanshan University, Qinhuangdao, China, in 2014.

From January 2016 to September 2016, he held a post-doctoral position with the University of North Texas, Denton, TX, USA. From October 2016 to January 2017, he was a Research Associate with the University of Texas at Arlington, Arlington, TX, USA. He is currently a Full Professor with Yanshan University. He has authored over 60 referred international journals and conference papers. Meanwhile, he has published two books, and he is also the inventor of 17 patents. His research interests include underwater acoustic sensor networks, networked teleoperation systems, and cyberphysical systems.

Dr. Yan received the Excellent Youth Project for NSF of China in 2022, the Distinguished Youth Project for NSF of Hebei Province in 2022, and the Excellence Paper Award from the National Doctoral Academic Forum of System Control and Information Processing in 2012. He currently serves as the Career Associate Editor for the IEEE/CAA JOURNAL OF AUTOMATICA SINICA and an Associate Editor for *Wireless Networks* (Springer) and *IET Control Theory and Applications*.

Wenqiang Cao received the B.S. degree in automation from the Shandong University of Technology, Zibo, China, in 2020. He is currently pursuing the Ph.D. degree in control engineering with Yanshan University, Qinhuangdao, China.

His research interests include distributed formation control, reinforcement learning, and robot control.

Xian Yang received the B.S. degree in automation and the Ph.D. degree in control theory and control engineering from Yanshan University, Qinhuangdao, China, in 2010 and 2016, respectively.

She is currently an Associate Professor with Yanshan University. Her research interests include networked teleoperation systems, underwater cyber-physical systems, and nonlinear control.

Cailian Chen (Member, IEEE) received the B.Eng. and M.Eng. degrees in automatic control from Yanshan University, Qinhuangdao, China, in 2000 and 2002, respectively, and the Ph.D. degree in control and systems from the City University of Hong Kong, Hong Kong, in 2006.

In 2008, she joined as an Associate Professor with the Department of Automation, Shanghai Jiao Tong University, Shanghai, China, where she is currently a Full Professor. She has authored and/or coauthored two research monographs and over 100 referred international journals and conference papers. She is the inventor of more than 20 patents. Her research interests include wireless sensor and actuator network and application in industrial automation, vehicular networks and application in intelligent transportation systems, and estimation and control for multiagent systems.

Prof. Chen received the IEEE TRANSACTIONS ON FUZZY SYSTEMS Outstanding Paper Award in 2008. She was one of the First Prize Winners of Natural Science Award from the Ministry of Education of China in 2006 and 2016, respectively. She was honored as Changjiang Young Scholar by the Ministry of Education of China in 2015 and the Excellent Young Researcher by NSF of China in 2016.

Xinping Guan (Fellow, IEEE) was a Professor and Dean of Electrical Engineering at Yanshan University, Qinhuangdao, China. He is currently a Chair Professor with Shanghai Jiao Tong University, Shanghai, China, where he is the Deputy Director of University Research Management Office and the Director of the Key Laboratory of Systems Control and Information Processing, Ministry of Education of China. He is the Leader of the prestigious Innovative Research Team of the National Natural Science Foundation of China (NSFC). His current research interests include industrial cyber-physical systems, wireless networking and applications in smart city and smart factory, and underwater sensor networks.

Dr. Guan is an Executive Committee Member of the Chinese Automation Association Council and the Chinese Artificial Intelligence Association Council. He received the First Prize of Natural Science Award from the Ministry of Education of China in 2006 and 2016, respectively, and the Second Prize of the National Natural Science Award of China in 2008. He was a recipient of the IEEE TRANSACTIONS ON FUZZY SYSTEMS Outstanding Paper Award in 2008. He is a National Outstanding Youth honored by NSF of China, Changjiang Scholar by the Ministry of Education of China, and State-level Scholar of New Century Bai Qianwan Talent Program of China.