Reinforcement Learning-based Adaptive Trajectory Planning for AUVs in Under-ice Environments

Chaofeng Wang*, Li Wei*, Zhaohui Wang*, Min Song†, and Nina Mahmoudian‡

*Dept. of Electrical and Computer Engineering, Michigan Technological University, MI 49931, USA
†Dept. of Electrical and Computer Engineering, Stevens Institute of Technology, NJ 07030, USA
‡Dept. of Mechanical Engineering-Engineering Mechanics, Michigan Technological University, MI 49931, USA

***Abstract*—This work studies online learning-based trajectory planning for multiple autonomous underwater vehicles (AUVs) to estimate a water parameter field of interest in the under-ice environment. A centralized system is considered, where several fixed access points (APs) on the ice layer are introduced as gateways for communications between the AUVs and a remote data fusion center (FC). We model the water parameter field of interest as a Gaussian process (GP) with unknown hyper-parameters. The AUV trajectories for sampling are determined on an epoch-by-epoch basis. At the end of each epoch, the APs relay the observed field samples from all the AUVs to the FC which computes the posterior distribution of the field based on the Gaussian process regression (GPR) and estimates the field hyper-parameters. The optimal trajectories of all the AUVs in the next epoch are determined to minimize a long-term cost that is defined based on the field uncertainty reduction and the AUV mobility cost, subject to the kinematics constraint, the communication range constraint and the sensing area constraint. We formulate the adaptive trajectory planning problem as a Markov decision process (MDP). A reinforcement learning (RL)-based online learning method is designed to determine the optimal AUV trajectories in a constrained continuous space. Simulation results show that the proposed learning-based trajectory planning algorithm has performance similar to a benchmark method that assumes perfect knowledge of the field hyper-parameters.**

I. Introduction

Autonomous underwater vehicles (AUVs) are emerging as attractive platforms for remote underwater exploration and monitoring. Given the high cost of AUVs and their deployments, the AUV trajectories need to be carefully designed to collect the "best" data over scalar or vector fields that vary on a range of spatial and temporal scales [1]–[4]. Compared to the open water scenario, research on the under-ice AUV trajectory planning has been very limited, with existing work mainly focused on a single AUV [5]–[8], and the AUV trajectory is typically pre-programmed with decision autonomy to handle malfunctions and external events [6], [9].

This work studies the adaptive trajectory planning of multiple AUVs in the under-ice environment for estimateion of a water parameter field of interest. Particularly, we consider a centralized system as illustrated in Fig. 1, where the fixed access points (APs) on the ice layer serve as gateways for communications between the AUVs and a remote data fusion center (FC). The AUV trajectories are determined by the FC on a time epoch-by-epoch basis based on the samples collected in the past epochs.

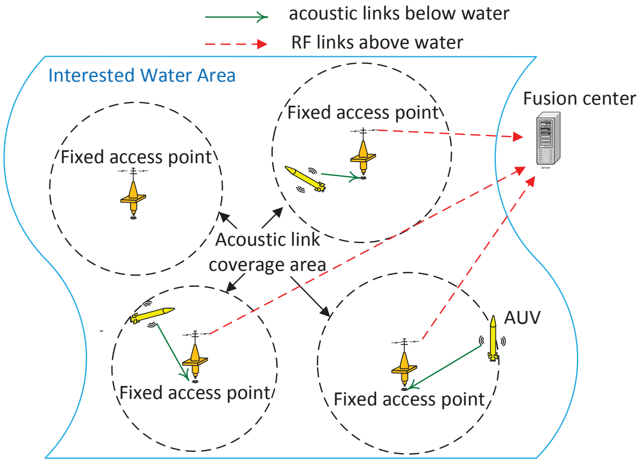


Fig. 1. An illustration of a system layout with 3 AUVs and 4 APs.

In this work, the water parameter field of interest is modeled as a Gaussian process (GP) with unknown hyper-parameters [10]. At the end of each epoch, the APs relay the field samples collected by the AUVs to the FC where the field hyper-parameters are estimated via the maximum likelihood method [10], and the posterior field distribution and the field uncertainty are computed via the Gaussian process regression (GPR) [11]. The AUV trajectories in the next epoch will then be determined based on the current system state including the current positions of all the AUVs and the field knowledge, with an aim of minimizing a long-term system cost that is defined based on the field uncertainty reduction and the AUV mobility cost. The AUV trajectories are expected to satisfy several practical constraints, including the kinematics constraint, the constraint on the communication range, and the constraint of being within the area of interest.

We formulate the adaptive trajectory planning problem as a Markov decision process (MDP) [12] with a constrained continuous action space. A reinforcement learning (RL)-based method is designed for online learning of the optimal action, i.e., the trajectories of all the AUVs, which satisfies the constraints. The knowledge for determining the optimal trajectories in each epoch is first obtained by transferring the historical knowledge used to determine the trajectories in the previous epoch and then is further adjusted based on the newly collected system cost. The proposed RL-based trajectory planning algorithm is validated using simulated 2-

978-1-5386-4814-8/18/$31.00 ©2018 IEEE

dimensional (2D) fields. The simulation results show that the proposed algorithm achieves performance similar to a benchmark method that assumes perfect knowledge of the field hyper-parameters.

The main contributions of this work are in the following.

- The developed algorithm is non-myopic and for multiple AUVs, while most existing works on non-myopic planning consider only a single vehicle [13]–[15].
- This work performs the online learning of the field hyper-parameters, while many existing works assume known *a priori* of the field knowledge [2], [16], [17].
- The developed algorithm considers a continuous action space, while many existing works consider either a discrete action space or a finite number of pre-determined trajectory patterns [14], [15], [18].

The rest of the paper is organized as follows. The system model is presented in Section II. The RL-based adaptive trajectory planning algorithm is developed in Section III. Evaluation of the proposed algorithm is included in Section IV. Conclusions are drawn in Section V.

II. System Model and Problem Formulation

In this section, we describe the system in details and build a mathematical model for the field estimation. The trajectory planning for multiple AUVs is then formulated as an optimization problem under constraints.

A. System Description

The system under consideration consists of multiple AUVs, several fixed APs and a remote FC. Denote the set of the AUVs as $\mathcal{M} = \{1, 2, ..., |\mathcal{M}|\}$. The AUVs are equipped with sensors and acoustic communication devices. They take field measurements at different sampling locations as navigate along their trajectories. A total number of $N_{\rm AP}$ APs are placed at fixed locations which collect data from all the AUVs via acoustic links. The APs send the observation data and location information of all the AUVs to a data FC via high data rate radio links where the FC performs further data processing. An illustration of the system layout with 3 AUVs and 4 APs is shown in Fig. 1. The underwater area of interest can be described by a continuous location set $\mathcal{X}_{\rm area} \subset \mathbb{R}^{\rm D}$ with $\rm D = 2$ or $\rm D = 3$. The field can be described as $f(\mathbf{x})$, where $\mathbf{x} \in \mathcal{X}_{\rm area}$ represents a location in the area of interest.

The system operates on an epoch-by-epoch basis. The proposed trajectory planning mechanism for AUVs in each epoch is described as in Fig. 2. The planned trajectory of each AUV in the ℓth epoch consists of K waypoints in K time slots and is determined at the end of the $(\ell-1)$th epoch, i.e., $\tilde{\mathbf{y}}_i(\ell) := [\mathbf{y}_{i1}(\ell); \mathbf{y}_{i2}(\ell); \cdots; \mathbf{y}_{iK}(\ell)]$. Each AUV takes field measurements around the waypoints, and after reach the last waypoint in the current epoch, it transmits the observed data and the corresponding sampling locations to the nearest AP via acoustic links in water. The APs then relay all the information to the FC via radio links above water. The FC estimates the field based on all the observation data, estimates the field knowledge, determines the trajectories $\{\tilde{\mathbf{y}}_i(\ell+1), i \in \mathcal{M}\}$ for all the AUVs in the next epoch, and transmits via APs the planned trajectories to all the AUVs. At the end of the ℓth epoch, all the AUVs receive their planned trajectories in the next epoch.

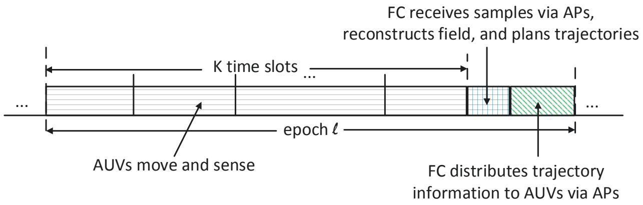

FC receives samples via APs,
reconstructs field, and plans trajectories
K time slots
...
...
...
epoch ℓ
AUVs move and sense
FC distributes trajectory
information to AUVs via APs

Fig. 2. Epoch structure for water parameter field estimation using AUVs.

B. Constraints on Sampling Trajectories

The planned trajectories must satisfy practical constraints. In this work, we consider three constraints related to kinematics, the communication range, and the sensing area.

1) Kinematics Constraint: Due to the limited travel speed of an AUV, in each epoch, the distance between any two consecutive waypoints for each AUV is constrained as

$$\sqrt{||\mathbf{y}_{ij}(\ell) - \mathbf{y}_{i,j+1}(\ell)||^2} \leq \kappa_{\rm up}, \forall i \in \mathcal{M} \tag{1}$$

with $1 \leq j \leq K-1$, and

$$\sqrt{||\mathbf{y}_{iK}(\ell) - \mathbf{y}_{i1}(\ell+1)||^2} \leq \kappa_{\rm up}, \forall i \in \mathcal{M} \tag{2}$$

where $\kappa_{\rm up}$ is the maximal distance that an AUV can travel within one time slot.

2) Communication Range Constraint: Since the field samples of each AUV must be sent to an AP in the last time slot in each epoch, we must ensure that in the Kth time slot of each epoch, each AUV must be within the communication range of at least one of the $N_{\rm AP}$ APs, i.e.,

$$\sqrt{||\mathbf{y}_{iK}(\ell) - \mathbf{y}_{\rm AP}^{(j)}||^2} < \kappa_{\rm comm}, \ \exists j \in \mathcal{I}_{\rm AP}, \ \forall i \in \mathcal{M} \tag{3}$$

where $\mathcal{I}_{\rm AP} := \{1, 2, \cdots, N_{\rm AP}\}$ is the AP index set, $\mathbf{y}_{\rm AP}^{(j)}$ is the location of the jth AP, and $\kappa_{\rm comm}$ is the communication range that ensures error-free transmission between an AP and an AUV.

3) Sensing Area Constraint: We assume that all the AUVs should stay within the area of interest, i.e.,

$$\mathbf{y}_{ij}(\ell) \in \mathcal{X}_{\rm area}, \forall i \in \mathcal{M}, i \geq 0, 0 \leq j \leq K, \ell \geq 0. \tag{4}$$

C. Modeling Real Trajectories of AUVs

Denote $\mathcal{Y}(\ell) := \{\tilde{\mathbf{y}}_1(\ell), \tilde{\mathbf{y}}_2(\ell), \cdots, \tilde{\mathbf{y}}_{|\mathcal{M}|}(\ell)\}$ as the planned trajectories consisting of waypoints for all the AUVs in the ℓth epoch. Due to the complex underwater environment, the AUVs may not arrive at each planned waypoint exactly. We model the true sampling location of the ith AUV in the kth time slot within the ℓth epoch as

$$\mathbf{x}_{ik}(\ell) = \mathbf{y}_{ik}(\ell) + \mathbf{e}_{ik}(\ell), \tag{5}$$

where $\mathbf{e}_{ik}(\ell) \in \mathbb{R}^{\rm D}$ is a noise vector which describes the location inaccuracy, and is assumed following a uniform distribution $\mathcal{U}(-\epsilon, \epsilon)$ [19] with $\epsilon \ll \kappa_{\rm comm}$ and $\epsilon \ll \kappa_{\rm up}$ being the navigation error.

The exact sampling locations of the ith AUV in the ℓth epoch are described by $\tilde{\mathbf{x}}_i(\ell) = [\mathbf{x}_{i1}(\ell); \mathbf{x}_{i2}(\ell); \cdots ; \mathbf{x}_{iK}(\ell)]$. Denote $\mathcal{X}_{\rm samp}(\ell) := \left\{\tilde{\mathbf{x}}_1(\ell), \tilde{\mathbf{x}}_2(\ell), ..., \tilde{\mathbf{x}}_{|M|}(\ell)\right\}$ as the sampling locations of all the AUVs in the ℓth epoch, $\mathcal{Z}(\ell)$ as all the sampling location from epoch 0 to epoch ℓ, and $\tilde{\mathbf{p}}(\ell) := [\mathbf{x}_{1K}(\ell-1); \mathbf{x}_{2K}(\ell-1); \cdots ; \mathbf{x}_{|\mathcal{M}|K}(\ell-1)]$ as the locations of all the AUVs at the beginning of the ℓth epoch.

D. Gaussian Process Regression for Field Estimation

In this work, we model the field of interest as a GP, and employ the GPR for field estimation. We first obtain a discrete set of target points $\mathcal{X}$ by discretizing the area $\mathcal{X}_{\rm area}$. We intend to minimize the field uncertainty over the target points rather than the whole area of interest to reduce the computational complexity. The set $\mathcal{X}$ can be selected based on application requirements or to balance the field estimation accuracy and the computational complexity. We assume that the total number of elements in $\mathcal{X}$ is $N_{\mathcal{X}}$. The field of interest is then modeled as a GP with zero mean,

$$f(\mathbf{x}) \sim \mathcal{GP}(0, \mathcal{K}(\mathbf{x}, \mathbf{x}')), \forall \mathbf{x}, \mathbf{x}' \in \mathcal{X}_{\rm area} \tag{6}$$

where $\mathcal{K}(\mathbf{x}, \mathbf{x}')$ is the value of covariance function at locations $\mathbf{x}$ and $\mathbf{x}'$ which describes the spatial correlation between locations $\mathbf{x}$ and $\mathbf{x}'$.

There are various types of covariance functions that can be employed [10]. In this work, we consider the squared exponential covariance function,

$$\mathcal{K}(\mathbf{x}, \mathbf{x}') = \sigma_f^2 \exp\left\{-(\mathbf{x}-\mathbf{x}')^{\rm T} \mathbf{\Lambda}^{-2} (\mathbf{x}-\mathbf{x}')\right\}, \tag{7}$$

where $\mathbf{\Lambda} = \mathrm{diag}([d_1, \cdots, d_{\rm D}])$ with $\rm D = 2$ or 3 being the dimension of the water area and d_i being the distance scale that determines the spatial correlation of two locations, and σ_f^2 is the signal variance.

In the ℓth epoch, a set of field observations can be obtained,

$$\boldsymbol{\psi}(\ell) = f\left(\mathcal{X}_{\rm samp}(\ell)\right) + \mathbf{n}(\ell), \tag{8}$$

where $f(\mathcal{X}_{\rm samp}(\ell))$ are the field values at the locations in $\mathcal{X}_{\rm samp}(\ell)$, and $\mathbf{n}(\ell)$ is the observation noise with each of its elements assumed following a Gaussian distribution $\mathcal{N}(0, \sigma_n^2)$.

Denote $\mathbf{\Psi}(\ell) = \{\boldsymbol{\psi}(\ell')\}_{\ell'=0}^{\ell}$ as available field observations. Denote $\mathbf{C}(\mathcal{A}, \mathcal{B})$ as a matrix whose the (i, j)th element is calculated as $\mathcal{K}(\mathbf{x}_i, \mathbf{x}_j)$, with $\mathbf{x}_i \in \mathcal{A}$ and $\mathbf{x}_j \in \mathcal{B}$. The posterior distribution of the field in the ℓth epoch over the target point set $\mathcal{X}$ can be obtained as

$$f(\mathcal{X}) \sim \mathcal{N}(\boldsymbol{\mu}_\ell, \mathbf{\Sigma}_\ell), \tag{9}$$

with

$$\boldsymbol{\mu}_\ell = \mathbf{C}(\mathcal{X}, \mathcal{Z}(\ell))\mathbf{C}_{\mathcal{Z}}^{-1}\mathbf{\Psi}(\ell), \tag{10}$$

$$\mathbf{\Sigma}_\ell = \mathbf{C}(\mathcal{X}, \mathcal{X}) - \mathbf{C}(\mathcal{X}, \mathcal{Z}(\ell))\mathbf{C}_{\mathcal{Z}}^{-1}\mathbf{C}(\mathcal{Z}(\ell), \mathcal{X}), \tag{11}$$

and $\mathbf{C}_{\mathcal{Z}} = \mathbf{C}(\mathcal{Z}(\ell), \mathcal{Z}(\ell)) + \sigma_n^2\mathbf{I}$, according to [10].

Based on the available observation $\mathbf{\Psi}(\ell)$ at the end of the ℓth epoch, the field hyper-parameters $\boldsymbol{\theta}_{\rm hyper} := \{\sigma_f^2, \mathbf{\Lambda}\}$ can be estimate by maximizing the log marginal likelihood [10],

$$\hat{\boldsymbol{\theta}}_{\rm hyper} = \max_{\boldsymbol{\theta}_{\rm hyper}} \left\{-\frac{1}{2}\mathbf{\Psi}(\ell)^{\rm T}\mathbf{C}_{\mathcal{Z}}^{-1}\mathbf{\Psi}(\ell) - \frac{1}{2}\log|\mathbf{C}_{\mathcal{Z}}|\right\}. \tag{12}$$

The hyper-parameters fully characterize the field spatial correlation, which are unknown *a priori* and estimated on the fly. The optimization problem (12) can be solved using a quasi-Newton method, i.e., the L-BFGS-B method [20].

E. Problem Formulation for Optimal Trajectory Planning

The field uncertainty can be obtained based on the field posterior distribution which is updated through the GPR. Specifically, we denote $\mathbf{u}_\ell := \mathrm{diag}(\mathbf{\Sigma}_{\ell-1})$, to describe the uncertainty of all the target points in $\mathcal{X}$ based on the observations up to the $(\ell-1)$th epoch.

Denote $\mathbf{s}(\ell) = \{\tilde{\mathbf{p}}(\ell), \mathbf{u}_\ell\}$ as the system state at the beginning of the ℓth epoch. Denote $\mathbf{a}(\ell)$ as the action in the ℓth epoch which consists of the planned waypoints for all the AUVs in the ℓth epoch.

The desired trajectories for all the AUVs in the ℓth epoch can be determined to minimize the expected total discounted cost,

$$\min_{\{\mathbf{a}(\ell)\}_{\ell=0}^{\infty}} \mathbb{E}\left\{\sum_{\ell=0}^{\infty} \gamma^\ell C(\mathbf{s}(\ell), \mathbf{a}(\ell))\right\}, \tag{13}$$

where $\gamma \in (0, 1]$ is a discount factor, and $C(\mathbf{s}(\ell), \mathbf{a}(\ell))$ is an application-dependent cost function. In this work, the cost function considers the field uncertainty reduction, the AUV mobility cost based on the planned trajectories, and the constraints from (1) to (4). Next we present the formulation of the cost function used in this work.

1) Cost Function: Denote the current state $\mathbf{s} = \{\tilde{\mathbf{p}}, \mathbf{u}\}$ and the planned trajectories as $\mathbf{a}$. Denote the next state $\mathbf{s}' = \{\tilde{\mathbf{p}}', \mathbf{u}'\}$. The costs, reward, and penalties induced by action $\mathbf{a}$ under the current state $\mathbf{s}$ and the next state $\mathbf{s}'$ are as follows.

- *Uncertainty reduction reward:* The sampling reward to reduce the field uncertainty by performing the action $\mathbf{a}$ at the system state $\mathbf{s}$ is defined as

$$R(\mathbf{s}, \mathbf{a}) := \frac{\alpha_R}{N_{\mathcal{X}}}\left(||\mathbf{u}||_1 - ||\mathbf{u}'||_1\right), \tag{14}$$

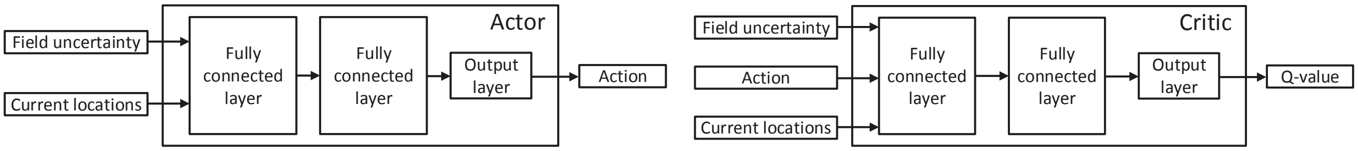
  where α_R is a weighting factor, and $||\mathbf{u}'||_1$ is the summation of all the elements in $\mathbf{u}'$ which describes the total estimation error of the field. We intend to minimize the field uncertainty over the target set.
- *Trajectory cost:* The mobility cost is defined as

$$C_{\rm T}(\mathbf{a}) := \alpha_L L(\mathbf{a}) + \alpha_A A(\mathbf{a}), \tag{15}$$

  where $L(\mathbf{a})$ is the total distance of the planned trajectories based on $\mathbf{a}$, $A(\mathbf{a})$ is the total angle that the AUVs travel along the planned trajectories based on $\mathbf{a}$, and α_L and $\alpha_{\rm A}$ are weighting factors. Less energy will be consumed if an AUV travels less distance and makes less turns.
- *Trajectory constraint penalty:* We define a penalty term for the case if the planned trajectories do not satisfy constraints (3) to (4). The penalty is defined as

$$C_{\rm P}(\mathbf{a}) := \alpha_{\rm p1} I_1 + \alpha_{\rm p2} I_2, \tag{16}$$

  where $\alpha_{\rm p1}$ and $\alpha_{\rm p2}$ are positive values, and I_1 and I_2 are indication functions for constraints (3) and (4),



(a) Actor network

(b) Critic network

Fig. 3. Illustration of the forward structures of the actor network and the critic network.

respectively, which equal 1 if the constraints are not satisfied and 0 otherwise.

Hence, the cost function in (13) used in this work can be described as

$$C(\mathbf{s},\mathbf{a}) = -R(\mathbf{s},\mathbf{a}) + C_{\mathrm{T}}(\mathbf{a}) + C_{\mathrm{P}}(\mathbf{a}). \tag{17}$$

III. Reinforcement Learning-based Adaptive Trajectory Planning

The proposed optimization problem in (13) can be taken as an MDP when the field hyper-parameters are known *a priori*. In this section, we employ an actor-critic-based algorithm, namely, the deep deterministic policy gradient (DDPG) algorithm [21], to solve the proposed MDP.

A. DDPG Basics and Design

In an actor-critic method, the actor learns how to generate the optimal action while the critic learns how to provide action evaluation which helps the actor to improve its action generation strategy. In the DDPG algorithm, an actor is represented by a neural network which takes the system state $\mathbf{s}$ as the input and takes the optimal action $\mathbf{a}$ under the system state $\mathbf{s}$ as the output. A critic is represented by another neural network which takes the system state $\mathbf{s}$ and the action $\mathbf{a}$ as the inputs and takes the Q-value function $Q(\mathbf{s},\mathbf{a})$ as the output, which indicates the expected cost after taking action $\mathbf{a}$ under system state $\mathbf{s}$. In the learning process, the actor network is leveraged to provide the action $\mathbf{a}$ to be executed under the state $\mathbf{s}$. After performing the action $\mathbf{a}$, the corresponding cost $C(\mathbf{s},\mathbf{a})$ can be obtained. Based on the obtained cost, the weights of the critic network are adjusted to better approximate the Q-value function $Q(\mathbf{s},\mathbf{a})$. Then, the weights of the actor are adjusted using the policy gradient method such that the action obtained by the actor could result in lower expected cost. For more details about the DDPG method, please refer to [21].

The structural design of the actor and the critic in this work is shown in Fig. 3. For the actor, as illustrated in Fig. 3(a), the field uncertainty and the current locations of all the AUVs go through two fully connected layers with rectified linear units (ReLUs) as the activation functions. The output layer takes the summation of the outputs of the second fully connected layer and uses a bounded tanh activation function to ensure that the action satisfies the kinematics constraints (1) and (2). For the critic, as shown in Fig. 3(b), the field uncertainty, and the current locations and actions of all the AUVs go through two fully connected layers with ReLUs as the activation functions. The output layer of the critic is the summation of the outputs of the second fully connected layer. In each training iteration, the parameters of the actor and the critic networks are updated based on one iteration of the backpropagation algorithm [22].

For the constraints (3) and (4), we modify the DDPG algorithm by using two experience replay buffers. Experience replay is a technique used to train the actor and the critic with system transition samples drawn from a buffer which consists of historical transitions from the previous experience. The two buffers consist of transitions whose actions satisfy the constraints (3) and (4) and otherwise, respectively. To learn the optimal actions which satisfy the constraints, we should ensure that the actor and the critic learn sufficient samples from both buffers in the training process. In this way, the actor will generate actions which have less cost and satisfy the constraints (3) and (4) while the critic could evaluate the actions and states without a bias.

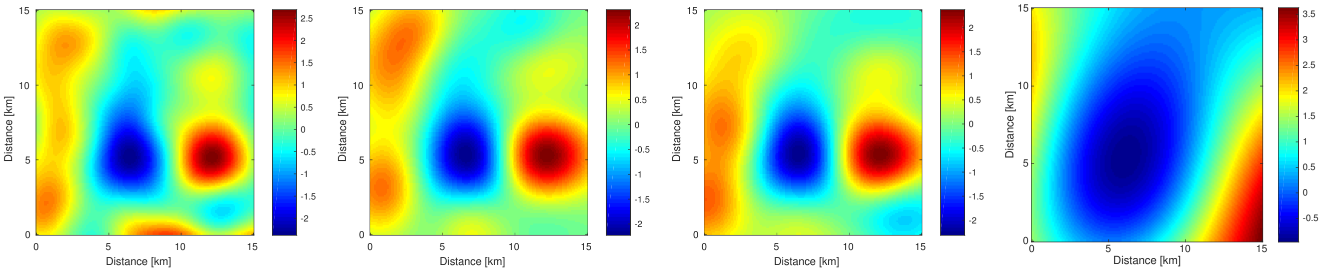
If the field hyper-parameters are known *a priori*, the above modified DDPG algorithm can be used to learn the optimal actions offline, and the corresponding performance serves as an upper bound for the proposed online learning strategy.

B. Online Learning for Trajectories Planning with Unknown Field Hyper-parameters

In practice, the perfect knowledge of the field hyper-parameters is often unavailable. It is generally the case that those hyper-parameters should be estimated online during the sampling process. We propose an online learning algorithm which incorporates the modified DDPG algorithm to determine the optimal trajectories of all the AUVs in each epoch, where the field hyper-parameters are online estimated. Specifically, at the end of each epoch, the unknown field hyper-parameters in the covariance function (7) can be estimated by solving the optimization problem in (12) based on all the current and historical observations. After obtaining the estimated hyper-parameters, the learned weights of the actor and critic networks in the previous epoch are transferred directly to the current epoch. The modified DDPG algorithm is then applied based on the available knowledge of the actor and the critic, as well as the estimated field hyper-parameters, to learn the optimal trajectories for the future epoch. In this way, the optimal trajectories for each epoch can be learned online based on the online estimated field hyper-parameters.

IV. Algorithm Evaluation

We consider an under-ice field of interest in a 2D 15 km × 15 km square area, and the target set $\mathcal{X}$ consists of 16



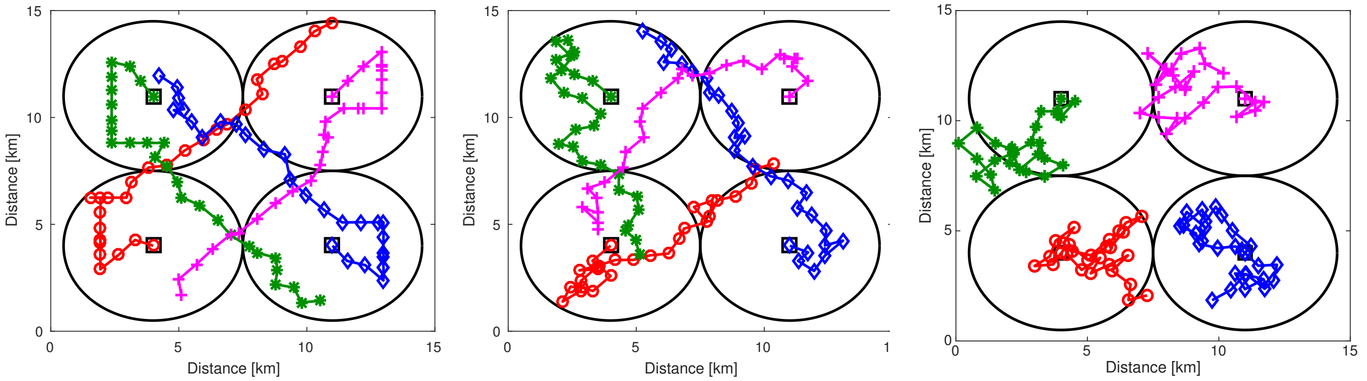
(a) True simulated field

(b) Estimated field, Scheme 1

(c) Estimated field, Scheme 2

(d) Estimated field, Scheme 3

Fig. 4. The true simulated field and the estimated fields obtained by three schemes.



(a) Scheme 1

(b) Scheme 2

(c) Scheme 3

Fig. 5. Trajectories of 4 AUVs obtained by three schemes, where the black squares and the black circles indicate the positions of 4 APs and the communication ranges of the APs, respectively. The black circles are also the initial deployment locations of the 4 AUVs.

× 16 grid points where the latitude and longitude distance between any two consecutive locations is 1 km. The 2D water parameter field is generated based on the circulant embedding method [23] with the field hyper-parameters as $\sigma_f^2 = 1$ and $\mathbf{\Lambda} = \mathrm{diag}([0.3,\ 0.3])$.

The duration of one time slot is 1,000 seconds (16.7 minutes), and one epoch consists of 3 time slots, leading to an epoch duration of 50 minutes. We consider a total of 9 epochs in the sampling process, which yields a deployment time duration of 7.5 hours in total. The simulated system consists of 4 AUVs and 4 APs. The 4 APs are located at $(4\ \text{km}, 4\ \text{km})$, $(4\ \text{km}, 11\ \text{km})$, $(11\ \text{km}, 4\ \text{km})$, and $(11\ \text{km}, 11\ \text{km})$, respectively. Those four locations are also the initial deployment sites of the 4 AUVs. The maximal navigation error is $\epsilon = 5$ m. The maximal speed of each AUV is 1 m/s, and the maximal distance of an AUV can travel within one time slot is therefore $\kappa_{\text{up}} = 1$ km. The communication range for underwater acoustic links between an AUV and an AP is $\kappa_{\text{comm}} = 3.5$ km. The discounted factor is $\gamma = 0.99$. The weights in the total cost function (17) are $\alpha_R = -10$, $\alpha_L = 1 \times 10^{-3}$, $\alpha_A = 5 \times 10^{-2}$, $\alpha_{p1} = 2$, and $\alpha_{p1} = 4$.

We evaluate the field estimation performance of three schemes.

- *Scheme 1:* The clairvoyant method which determines the sampling trajectories through the offline modified DDPG algorithm based on the perfect knowledge of the field hyper-parameters;
- *Scheme 2:* The proposed online RL algorithm which determines the sampling trajectories epoch-by-epoch through the modified DDPG algorithm where the field hyper-parameters are online estimated in each epoch based on the collected samples;
- *Scheme 3:* All the AUVs sample the water parameter field via random walk. Here, we present the simulation result that is selected among 10,000 Monte Carlo runs which yields the minimal total cost.

We take the normalized mean square error (NMSE) as a performance metric for the field estimation, which describes the normalized difference between the true field and the estimated field,

$$\text{NMSE} := \frac{\int_{\mathcal{X}_{\text{area}}} ||f(\mathbf{x}) - \hat{f}(\mathbf{x})||^2 d\mathbf{x}}{\int_{\mathcal{X}_{\text{area}}} ||f(\mathbf{x})||^2 d\mathbf{x}}, \tag{18}$$

where f is the true field and $\hat{f}$ is the estimated field based on the mean of the GPR.

The three schemes are first examined from the perspectives of the total traveled distance, the total traveled angle and the NMSE, as shown in Table I. Scheme 1 achieves the least

TABLE I
PERFORMANCE COMPARISON OF THREE SCHEMES.

	Scheme 1	Scheme 2	Scheme 3
Total traveled distance [km]	74.4	77.9	78.1
Total traveled angle [rad]	76.6	117.4	131.5
Normalized mean square error	0.11	0.15	1.07

total traveled distance and the least total traveled angle, while Scheme 2 has a similar total traveled distance but greater total traveled angle. The performance gap is due to the fact that Scheme 2 estimates the field hyper-parameters and determines the actions online. The total traveled distance and the total traveled angle obtained by Scheme 3 are similar to those of Scheme 2. However, the NMSEs obtained by Schemes 1 and 2 are significantly smaller than that of Scheme 3, where a marginal difference of the NMSEs obtained by Schemes 1 and 2 can be observed.

The simulated true field and the estimated fields by the three schemes are presented in Fig. 4. One can see that Schemes 1 and 2 can capture important features of the true field, and the estimated field by Scheme 3 is significantly different from the true field. The planned trajectories obtained by the three schemes are shown in Fig. 5. To explore the area with high uncertainty, the trajectories determined by Scheme 1 spread out more than those of Schemes 2 and 3, which results in the largest sensed area. The sensed area based on the trajectories obtained by Scheme 2 at the early epochs is small due to the inaccurate field hyper-parameter estimation based on limited field samples at the early stage. With more field samples collected, the trajectory pattern obtained by Scheme 2 is similar to the pattern obtained by Scheme 1 which tends to explore the area with high uncertainty.

V. CONCLUSIONS

This work investigated the adaptive trajectory planning of multiple AUVs for the water parameter field estimation in the under-ice environment. An online learning-based trajectory planning algorithm was proposed to adaptively determine the trajectories of AUVs. The field of interest was modeled as a GP with unknown hyper-parameters. The field hyper-parameters and the field posterior distribution were estimated online based on the collected samples. The adaptive trajectory planning problem was formulated as an MDP with a goal of minimizing a long-term cost that is defined based on the field uncertainty reduction and the AUV mobility cost, subject to the kinematics constraint, the communication range constraint and the sensing area constraint. A RL-based method was designed to solve the above MDP with a constrained continuous action space. The simulation results showed that the proposed RL-based adaptive trajectory planning algorithm achieved the performance close to a benchmark method that assumes perfect knowledge of the field hyper-parameters.

ACKNOWLEDGEMENT

This work was supported in part by the National Science Foundation under grants CNS-1551067 and ECCS-1651135 (CAREER), and by the ICC Paul William Seed Grant of Michigan Technological University.

REFERENCES

[1] N. E. Leonard, D. A. Paley, F. Lekien, R. Sepulchre, D. M. Fratantoni, and R. E. Davis, "Collective motion, sensor networks, and ocean sampling," *Proc. of the IEEE*, vol. 95, no. 1, pp. 48–74, Jan. 2007.

[2] N. Yilmaz, C. Evangelinos, P. Lermusiaux, and N. Patrikalakis, "Path planning of autonomous underwater vehicles for adaptive sampling using mixed integer linear programming," *IEEE J. Ocean. Eng.*, vol. 33, no. 4, pp. 522 – 537, Oct. 2008.

[3] D. Zhu, H. Huang, and S. X. Yang, "Dynamic task assignment and path planning of multi-AUV system based on an improved self-organizing map and velocity synthesis method in three-dimensional underwater workspace," *IEEE Trans. on Cybernetics*, vol. 43, no. 2, pp. 504–514, Apr. 2013.

[4] K. Szwaykowska and F. Zhang, "Trend and bounds for error growth in controlled Lagrangian particle tracking," *IEEE J. Ocean. Eng.*, vol. 39, no. 1, pp. 10–25, Jan. 2014.

[5] A. Kukulya, A. Plueddemann, T. Austin, R. Stokey, M. Purcell, B. Allen, R. Littlefield, L. Freitag, P. Koski, E. Gallimore, J. Kemp, K. Newhall, and J. Pietro, "Under-ice operations with a REMUS-100 AUV in the Arctic," in *IEEE/OES Autonomous Underwater Vehicles*, Monterey CA, 2010.

[6] A. J. Plueddemann, A. L. Kukulya, R. Stokey, and L. Freitag, "Autonomous underwater vehicle operations beneath coastal sea ice," *IEEE/ASME Trans. on Mechatronics*, vol. 17, no. 1, pp. 54–64, Feb. 2012.

[7] S. E. Webster, C. M. Lee, and J. I. Gobat, "Preliminary results in under-ice acoustic navigation for seagliders in Davis Strait," in *Proc. of MTS/IEEE OCEANS Conf.*, St. John's, Sept. 2014.

[8] J. Ferguson, "Adapting AUVs for use in under-ice scientific missions," in *Proc. of MTS/IEEE OCEANS Conf.*, Sept. 2008.

[9] P. Norgre and R. Skjetne, "Using autonomous underwater vehicles as sensor platforms for ice-monitoring," *Modeling, Identification and Control*, vol. 35, no. 4, pp. 269–277, Nov. 2014.

[10] C. E. Rasmussen, "Gaussian processes in machine learning," *Advanced lectures on machine learning*, pp. 63–71, 2004.

[11] C. K. Williams and C. E. Rasmussen, "Gaussian processes for regression," in *Advances in neural information processing systems*, 1996, pp. 514–520.

[12] R. Bellman, "A Markovian decision process," *Journal of Mathematics and Mechanics*, pp. 679–684, 1957.

[13] R. Martinez-Cantin, N. Freitas, E. Brochu, J. Castellanos, and A. Doucet, "A Bayesian exploration-exploitation approach for optimal online sensing and planning with a visually guided mobile robot," *Automous Robots*, vol. 27, no. 2, pp. 93–103, 2009.

[14] A. Singh, A. Krause, and W. Kaiser, "Nonmyopic adaptive informative path planning for multiple robots," in *Proc. of IJCAI*, Pasadena, CA, Jul. 2009.

[15] P. Morere, R. Marchant, and F. Ramos, "Sequential Bayesian optimization as a POMDP for environment monitoring with UAVs," in *Proc. Proc. of Int'l. Conf. on Robotics and Automation*, Singapore, Jun. 2017.

[16] Y. Xu, J. Choi, and S. Oh, "Mobile sensor network navigation using gaussian processes with truncated observations," *IEEE Trans. on Robotics*, vol. 27, no. 6, pp. 1118–1131, Dec. 2011.

[17] A. Marino, G. Antonelli, A. Aguiar, A. Pascoal, and S. Chiaverini, "A dencentralized strategy for multirobot sampling/patrolling: Theory and experiments," *IEEE Trans. on Control Systems Technology*, vol. 23, no. 1, pp. 313–322, Jan. 2015.

[18] L. Nguyen, S. Kodagoda, R. Ranasinghe, and G. Dissanayake, "Information-driven adaptive sampling strategy for mobile robotic wireless sensor network," *IEEE Trans. on Control Systems Technology*, vol. 24, no. 1, pp. 372–379, Jan. 2016.

[19] S. P. Mertikas, "Error distributions and accuracy measures in navigation: An overview," *Tech. Rep.*, 1985.

[20] R. Byrd, P. Lu, J. Nocedal, and C. Zhu, "A limited memory algorithm for bound constrained optimization," *SIAM Journal on Scientific Computing*, vol. 16, no. 5, pp. 1190–1208, 1995.

[21] T. P. Lillicrap, J. J. Hunt, A. Pritzel, N. Heess, T. Erez, Y. Tassa, D. Silver, and D. Wierstra, "Continuous control with deep reinforcement learning," in *Proc. of the International Conference on Learning Representations (ICLR)*, 2016.

[22] C. M. Bishop, *Pattern Recognition and Machine Learning*, 6th ed. Springer-Verlag New York, 2006.

[23] D. P. Kroese and Z. I. Botev, *Spatial Process Simulation*. Cham: Springer International Publishing, 2015, pp. 369–404. [Online]. Available: https://doi.org/10.1007/978-3-319-10064-7_12