

Received March 18, 2021, accepted April 14, 2021, date of publication April 26, 2021, date of current version May 3, 2021.

Digital Object Identifier 10.1109/ACCESS.2021.3075475

Finding an Optimal Geometric Configuration for TDOA Location Systems With Reinforcement Learning

SHENGXIANG LI¹, GUANGYI LIU¹, SIYUAN DING², HAISI LI¹, AND OU LI¹

¹PLA Strategy Support Force Information Engineering University, Zhengzhou 450001, China

²Key Laboratory of Experimental Physics and Computational Mathematics, Beijing 100038, China

Corresponding author: Ou Li (zzliou@126.com)


ABSTRACT In TDOA passive location tasks, the geometric configuration can greatly affect the positioning precision due to the complicated characteristics of electromagnetic environment. How to find an appropriate path to a good geometry to locate the transmitter accurately is vital in practical location tasks. This paper proposes a novel geometry optimization method based on deep reinforcement learning. In the proposed method, stations are regarded as mobile agents that can receive wireless signals decide where to go. All agents are controlled by an actor-critic learner, which is trained on the experiences collected from executing the TDOA location task repeatedly. To evaluate the trained agents, a TDOA location simulator environment with complex electromagnetic characteristics is developed. The empirical results show that, the learner mastered useful strategies and navigated to optimal geometric configurations efficiently. A visual depiction of highlights of the learner's behavior in TDOA passive location tasks can be viewed in the video provided in the supplementary material.

INDEX TERMS Geometry optimization, passive location, TDOA, reinforcement learning, actor-critic.

I. INTRODUCTION

Passive location techniques are used for various scenarios, such as telecommunication pseudo base station discovery, aviation interference investigation, etc. Passive location systems based on *time differences of arrivals* (TDOA) is widely used for its simplicity and crypticity. In a TODA location system, a number of spatially separated sensors capture the signals emitted by the transmitter and estimate time differences of arrivals to locate the transmitter [1]–[4].

However, the geometric configurations of stations can significantly affect the positioning precision [5], [6]. In the literature, some existing studies tried to obtain general principles of geometric configurations from massive experiments [6], [7]. And only some rough conclusions have been drawn. For instance, all stations should not line up, or stations should form a triangle to surround the transmitter. There also exist many studies that have employed heuristic methods, such as genetic algorithm (GA) [8]–[11] or particle swarm optimization (PSO) [12]–[14], to search the optimal geometry.

The associate editor coordinating the review of this manuscript and approving it for publication was Tallha Akram .

These methods are based on empirical models in which signals are assumed to propagate ideally. However, in the real world, an electromagnetic environment changes abruptly with the positions of stations due to various factors, such as signal frequency, interference, attenuation, multipath, obstacles, and noises. These factors can hardly be described fully in empirical models, leading to suboptimal geometric configurations and low positioning precision. What more important is that, GA and PSO are hard to realize in the real physical world. Take PSO as an example, one can hard to use thousands of particles (in the real world, they may be UAVs or unmanned cars) to navigate. Even if the physical limitation was satisfied, these heuristic methods suffer from amnesia, which means repeated search for the same problem when faced again. This paper provides a new perspective for geometric optimization, i.e., a sequential decision-making problem which needs to specify a path that navigates to an optimal geometry in a complex electromagnetic environment.

This paper make effort to achieve these goals:

- Find an optimal geometric configuration in less time and with shorter total length of paths;

- As long as the positioning precision is guaranteed, keep away from the transmitter;
- Behaviors are comprehensible.

Reinforcement learning (RL) is a viable and elegant approach to yield an optimal policy for sequential decision-making problems [15]–[17]. In TODA location tasks, the tricky electromagnetic spatial distribution can be tracked by RL in a *trial-and-error* paradigm with the non-linear and parameterized deep neural network (DNN), which provides the compact and powerful representation of experiences. Therefore, this paper address the problem of finding optimal geometric configuration in the TDOA location systems through deep reinforcement learning (DRL) [18]–[22].

Under the framework of DRL, all stations are regarded as mobile agents.¹ These agents capture the radio signal in the air and send it along with other observations to a central learner. The learner makes the decision on where to go based on the information gathered from all the agents, then, learn from the reward given by the environment. The learner is actor-critic style, in which the actor is a multi-dimension Gaussian actor with parameters generated by a neural network. Actions are sampled from these Gaussian distributions in each time step. The critic neural network is updated by minimizing the square of time difference error. Then the critic is used to update the actor neural network according to policy gradient theory. To evaluate the proposed method, a TDOA location task environment with complicated electromagnetic characteristics is developed, in which the multipath effect and interference as well as forbidden regions are considered. The results show that agents can learn stably and exhibit useful strategies to find optimal geometric configurations efficiently.

II. BACKGROUND

This section introduces the relevant background on TDOA passive location and DRL.

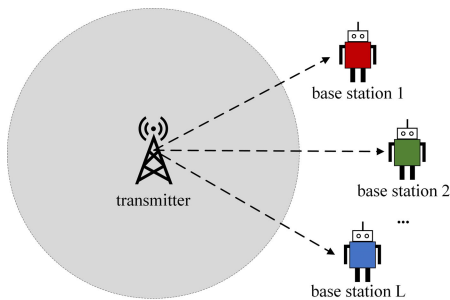


FIGURE 1. The topology of a TDOA location system: In a 2-D plane, L stations receive the emitted signals to estimate the position of the transmitter based on time differences of arrivals.

A. PASSIVE LOCATION WITH TDOA

Consider one transmitter and L stations intercepting the transmitted signal in a 2-D plane, as shown in FIGURE 1. The i -th

¹The terms *station* and *agent* are hereafter used interchangeably.

station's position is denoted by $p_i = (x_i, y_i)^\top$, and the position of the transmitter is denoted by $p_\star = (x_\star, y_\star)^\top$. The sampled observations at station i is denoted by

$$z_i(k) = b_i u(k - \tau_i) + \epsilon_i(k), \quad i = 1, \dots, L, k = 1, \dots, K, \quad (1)$$

where $u(k)$ is the signal radiating from the transmitter, K is the length of signal, b_i is the attenuation, τ_i is the time delay with respect to station i , and $\epsilon_i(k)$ is spatially additive white Gaussian noise. From the signals received by station i , j ($i \neq j$), i.e., z_i and z_j , the time difference of arrival Δ_{ij} can be obtained through correlation function

$$\Delta_{i,j} = \text{Corr}(z_i, z_j), \quad (2)$$

where $\text{Corr}(z_i, z_j)$ refers to the function that calculates the time lag of z_i to z_j . Note that, the resolution of lag time estimation is the sample interval, which determines the resolution of positioning. With positions p_i, p_j, p_\star and time lag $\Delta_{i,j}$, the distance equation is established as follows:

$$\Delta_{i,j} \cdot c = \|p_i - p_\star\|_2 - \|p_j - p_\star\|_2 \quad (3)$$

where $\|\cdot\|_2$ is the square norm, and $c = 3 \times 10^8$ m/s is the speed of light. With L spatially separated stations, there are $\binom{2}{L}$ distance equations can be established:

$$\begin{cases} \Delta_{1,2} \cdot c = \|p_1 - p_\star\|_2 - \|p_2 - p_\star\|_2 \\ \dots \\ \Delta_{L-1,L} \cdot c = \|p_{L-1} - p_\star\|_2 - \|p_L - p_\star\|_2. \end{cases} \quad (4)$$

In 2-D plane, the position of the transmitter is estimated with at least 3 stations. More stations can contribute to more distance equations which can enhance the robustness of estimation with *Least Square* (LS) algorithms.

B. DEEP REINFORCEMENT LEARNING

In reinforcement learning, an agent interacts with the environment for a given goal, which is modeled as a *Markov Decision Process* (MDP): $\langle \mathcal{S}, \mathcal{A}, \mathcal{P}, r, \gamma \rangle$. At time t , it observes state $s_t \in \mathcal{S}$ with \mathcal{S} denoting the state space, takes action $a_t \in \mathcal{A}$ with \mathcal{A} representing the action space, receives a reward $r_t \in \mathbb{R}$, and moves to the new state $s_{t+1} \in \mathcal{S}$ with probability $p(s_{t+1}|s_t, a_t) \in \mathcal{P}$. The agent aims to learn a policy that maximizes the cumulative sum of rewards [17]

$$J(\pi) = \mathbb{E} \left[\sum_{t=0}^{\infty} \gamma^t r_t \mid \pi \right], \quad (5)$$

where $\gamma \in [0, 1]$ is the discount factor that determines the importance of future rewards. The state-value function, which starts from state s and follows policy π , is denoted by

$$V(s) = \mathbb{E} \left[\sum_{t=0}^{\infty} \gamma^t r_t \mid s_t = s; \pi \right]. \quad (6)$$

The action-value function, that starts from state s , takes action a and follows policy π , is denoted by

$$Q(s, a) = \mathbb{E} \left[\sum_{t=0}^{\infty} \gamma^t r_t \mid s_t = s, a_t = a; \pi \right]. \quad (7)$$

And it is obvious that $V(s) = \int_a Q(s, a) da$. Denote the probability density, from state s to state s' after t steps following policy π , as $p(s|s', t, \pi)$, then, the discounted state distribution is $\rho(s') \triangleq \int_S \sum_{t=1}^{\infty} \gamma^{t-1} p_1(s) p(s|s', t, \pi) ds$. The *objective function* in reinforcement learning is to find the optimal policy π that maximizes the expected expectation long-term return,

$$J(\pi) = \int_S \rho(s) \int_A \pi(s, a) r(s, a) da ds = \mathbb{E}_{s \sim \rho, a \sim \pi} [r(s, a)] \quad (8)$$

where $\mathbb{E}_{s \sim \rho}[\cdot]$ denotes the expected value with respect to discounted state distribution $\rho(\cdot)$, following the policy π .

Tabular RL fails to handle huge or continuous state and action spaces. To this end, deep learning is leveraged to approximate the states and actions, which contributes to a new powerful technique, namely, DRL. DRL has made remarkable achievements in the fields of Chess [23], video games [24], and physical control tasks [20]. In DRL, the value function and policy are parameterized as $Q_\omega(s, a)$ and $\pi_\theta(s, a)$, respectively. According to *Policy Gradient* (PG) theory [18], the objective function can be expressed by

$$J(\pi_\theta) = \int_S \int_A \rho(s) \cdot \pi_\theta(a, s) \cdot A_\omega(s, a) da ds, \quad (9)$$

where $A_\omega(s, a) = Q_\omega(s, a) - V_\omega(s, a)$ is the *advantage function*. We can obtain the on-policy gradient by differentiating the performance function and applying an approximation

$$\nabla_\theta J(\pi_\theta) \approx \int_S \int_A \rho(s) \nabla_\theta \pi_\theta(a, s) A_\omega(s, a) da ds \quad (10)$$

$$= \int_S \rho(s) \int_A \nabla_\theta \pi_\theta(a, s) A_\omega(s, a) da ds \quad (11)$$

$$= \mathbb{E}_{s \sim \rho, a \sim \pi} [\nabla_\theta \log \pi_\theta(s, a) A_\omega(s, a)], \quad (12)$$

where $\nabla_\theta \log \pi_\theta(s, a)$ is the *score function*. In (10), the approximation drops a term that depends on the action-value gradient $\nabla_\theta A_\omega(s, a)$ and [25] argues that this is a good approximation since it can preserve the set of local optima to which gradient ascent converges.

III. RL-BASED GEOMETRY OPTIMIZATION

This section presents a RL-based geometric configuration optimization method for passive location systems.

A. MODEL FRAMEWORK

In this paper, a TDOA location system is considered with L mobile stations (e.g., UAVs equipped with positioning devices), i.e., L TDOA agents. Each agent transfers the intercepted signals to a central processing agent where the emitter's position is estimated. Agents have no knowledge

of the emitter and the electromagnetic environment. Due to the influence of multipath and noises, the signals received by different agents may vary. To adapt to the complicated electromagnetic spatial distribution accurately, a DRL based method, with positioning error being the reward function, is considered. The key elements in the MARL scheme are defined as follows.

States. States consists of geometric information of stations, features of received signals, and positioning information. At time step t , the signal emitted by the transmitter is intercepted by agent $1, \dots, L$, denoted by z_1^t, \dots, z_L^t . The features of the raw signal with high dimension are extracted by function $f(\cdot)$. The state is defined by $s_t = (p_1^t, \dots, p_L^t, f(z_1^t), \dots, f(z_L^t), \hat{p}_0, \hat{\sigma}_{\text{est}})$, \hat{p}_0 and $\hat{\sigma}_{\text{est}}$ are the center and variance of estimations of the transmitter in the same time step.

Actions. Actions represent the decisions regarding where to receive signals at next step. Let $a_i^t = (\Delta x_i^t, \Delta y_i^t)$ denote the action of agent i at time step t , where Δx_i^t and Δy_i^t are movements on the x -axis and y -axis, respectively. Then the joint action of all agents is denoted as $a_t = (\Delta x_1^t, \Delta y_1^t, \dots, \Delta x_L^t, \Delta y_L^t)$.

Rewards. This paper aims to develop agents that can properly adjust the geometry automatically to improve the positioning precision. To this end, we assess agents' behavior by positioning errors. Two types of positioning errors are widely used:

- CRLB is an effective index of the precision of a passive location system. Let the background position of the transmitter be $p_\star = (x_\star, y_\star)^T$. Then, the CRLB is a function of state s and the background position p_\star .
- The statistic errors such as the *mean error* (ME) and the *root mean square error* (RMSE) are popular errors. These errors need more estimation at each time step to calculate the statistic performance of positioning under the current geometry configuration.

The CRLB is popular for it can evaluate the positioning error without estimating the transmitter's position from received signals. But it depends on noise power spectral, which varies when stations are placed in different positions. More importantly, the CRLB is a lower bound of positioning error instead of error itself, and we can not always ignore the gap. The RMSE reflects the positioning performance more directly, independent of passive location algorithms. Hence, we utilize RMSE instead of CRLB to shape the reward relating to positioning error:

$$R_{\text{RMSE}}(p, p_\star) = -10 \log_{10} \left(1 + \sqrt{\frac{1}{N_{\text{est}}} \sum_{k=1}^{N_{\text{est}}} \|\hat{p}_\star^k - p_\star\|^2} \right), \quad (13)$$

where \hat{p}_\star^k is the k -th estimation of p_\star , and N_{est} denotes the estimation times for each geometric configuration.

Another important aspect we need consider is the total path length agents cover till reaching the optimal geometric

configuration. The path reward at time step t' is represented as follows:

$$R_{\text{path}} = \sum_{t=1}^{t'} \sum_{i=1}^L \sqrt{|\Delta x_i^t|^2 + |\Delta y_i^t|^2}. \quad (14)$$

The global reward consists of positioning error and total path length:

$$R = R_{\text{RMSE}} + \zeta R_{\text{path}}, \quad (15)$$

where $\zeta > 0$ is the coefficient that determines the relative importance of these two rewards.

B. LEARN TO OPTIMIZE THE GEOMETRY

This section presents an actor-critic algorithm for geometric configuration optimization in TDOA location tasks. The overall architecture of the proposed method is illustrated in FIGURE 2. The actor takes in the state s made up of information from all the agents and yields the actions to tell them where to go in the next time step. The critic use both state s and the action a given by the actor to evaluate the decision of the actor, i.e., calculate $Q(s, a)$. Both the actor and critic are approximated with neural networks and parameterized by θ and ω , respectively.

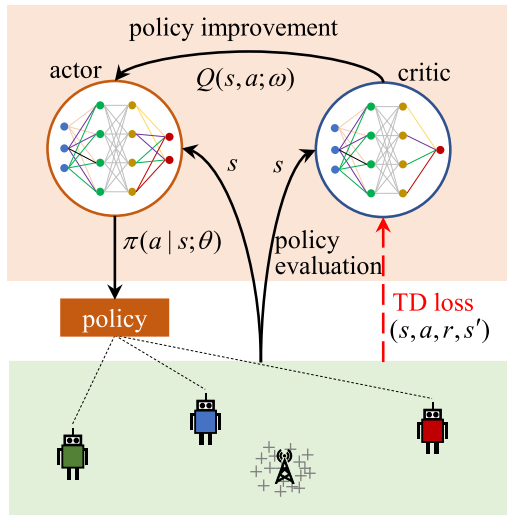


FIGURE 2. Schematics of the proposed method: An actor-critic learner that learns positioning strategies from executing passive location tasks repeatedly.

At the critic step, the agent takes in the state s and action a to estimate the reward-to-go, i.e., $Q(s, a; \omega)$. A well trained agent can estimate $Q(s, a; \omega)$ accurately, but at the beginning, the agent has only rough estimation and the gap is defined as *time difference* (TD) error:

$$\delta_t = R_t + \gamma Q(s_{t+1}, a_{t+1}; \omega) - Q(s_t, a_t; \omega). \quad (16)$$

The critic neural network is trained via minimizing the square of the TD error through *Stochastic Gradient Descent* (SGD):

$$\mathcal{J}_\omega = (R_t + \gamma Q(s_{t+1}, a_{t+1}; \omega) - Q(s_t, a_t; \omega))^2, \quad (17)$$

$$\nabla_\omega \mathcal{J}_\omega = -\delta_t \nabla_\omega Q(s_t, a_t; \omega), \quad (18)$$

$$\omega \leftarrow \omega + \alpha \delta_t \nabla_\omega Q(s_t, a_t; \omega), \quad (19)$$

where α is the stepsize for parameters update. Then the actor network is trained through *Policy Gradient Ascent* (PGA):

$$\theta \leftarrow \theta + \eta \nabla_\theta \pi_\theta(s, a) A(s, a; \omega), \quad (20)$$

where η is the stepsize for actor update and $A(s, a; \omega) = R_t + \gamma Q(s_{t+1}, a_{t+1}; \omega) - Q(s_t, a_t; \omega)$ is the advantage function.

Algorithm 1 Geometric Configuration Optimization for TDOA Location Systems With Actor-Critic

- 1: Initialize the TDOA passive location system with target transmitter emitting signals;
- 2: Initialize the critic and actor with parameters ω , θ , respectively;
- 3: Initialize the iteration counter $t \leftarrow 0$.
- 4: **repeat**
- 5: **for** $i = 1 : L$ **do**
- 6: Intercept the signals z_i^t and sent it to the central station;
- 7: **end for**
- 8: The central station update the state $s_t = (p_1^t, \dots, p_L^t, f(z_1^t), \dots, f(z_L^t), \hat{p}_0, \hat{\sigma}_{\text{est}})$;
- 9: Sample an action from the policy $a_t \sim \pi_\theta(\cdot | s_t)$;
- 10: Execute a_t for the passive location task and calculate receive the reward R_t ;
- 11: Update the parameters of value networks: $\delta_t = R_t + \gamma Q(s_{t+1}, a_{t+1}; \omega) - Q(s_t, a_t; \omega)$
 $\omega \leftarrow \omega + \alpha \delta_t \nabla_\omega Q(s_t, a_t; \omega)$
- 12: Update the parameters of policy network: $\theta \leftarrow \theta + \eta \nabla_\theta \pi_\theta(s, a) A(s, a; \omega)$
- 13: Update the counter $t \leftarrow t + 1$;
- 14: **until** The task is completed or reaching the maximum of counter.

The details of the method is summarized in Algorithm 1.

IV. EXPERIMENTS

In this section, a TDOA location task environment with complicated electromagnetic characteristics is developed, based on which the proposed method is evaluated.

A. THE TDOA LOCATION TASK ENVIRONMENT

In the experiment, the simulator's geographical coverage is a circular region with radius being 4km, as shown in FIGURE 3. The transmitter is at the center of the map and equipped with an isotropically radiating antenna. The channel attenuation is a function of the receiver's position p : $b(p) \propto \lambda_s / (4\pi d)$, i.e., the *free space path loss*, where λ_s is the wave length of signal and d is the distance to the transmitter. It is obvious that when the agents get close to the transmitter, the SNR increases and the positioning error declines. The RMSE is expected to be invariant under the same geometric configuration but to change mildly as agents move slowly, which is conducive to training the agents. According to the reward R_{RMSE} defined in (13), we designed an experiment

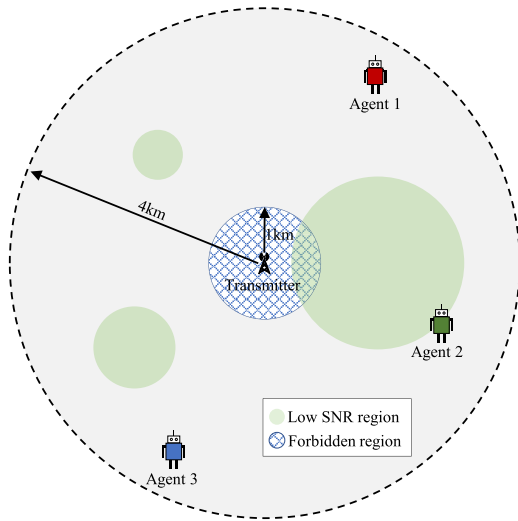


FIGURE 3. The TDOA location tasks environment. The transmitter is plotted as a black square in the center, surrounded by an 1-km forbidden region. There are several low SNR regions plotted in light green, where noises become stronger when agents enter it. Three agents are plotted as red, green and blue dots, respectively.

to explore the influence of estimation times ($N_{est} = 10, 30, 50, 80, 100$) on the RMSE of positioning. FIGURE 4 shows the RMSE curves with different N_{est} as agents get close to the transmitter. It can be seen that, $N_{est} = 100$ is a good choice when trade off between the asymptotic property and computation complexity of reward function.

The noise and interference, as well as the multipath effect, are all considered in this environment. The background noise is modeled by the spatially white noise defined in (1). To include the interference, as well as the multipath effect, some low regions, highlighted in green in FIGURE 3, is placed in the simulator, where the noises are stronger than other areas. The centers of these low SNR regions are $(2000, 0)$, $(-2000, -2000)$, $(-1500, 1700)$, and the corresponding radii are 1500, 600, 400. When agents are in these

regions, the amplitude of noises becomes greater than that in free space loss region. The radius of a low SNR region is denoted by u_{SNR} , and the distance of an agent to the center of the low SNR region is d . Then, the amplitude of noise is amplified K_{SNR} times: $K_{SNR} = (u_{SNR} - d)/u_{SNR} \times 9 + 1$, $d \leq u_{SNR}$.

Furthermore, in the real world, we can not get too close to the transmitter, therefore, there is a forbidden area (the radius is $u_{TR} = 1000$) around the transmitter, shown as the gray shadow region in FIGURE 3. Also, agents should not go too further. The boarder of the simulator is drawn as black dashed circle. The distance from the boarder of the simulator to the center of the transmitter is denoted by u_{BO} , and $u_{BO} = 4000$. Stations being in the forbidden region or on the outside of the simulator leads to punishment with additional reward:

$$R_P = \begin{cases} d - u_{TR} + R_0, & d \leq u_{TR}. \\ u_{BO} - d + R_0, & d \geq u_{BO}. \\ 0, & \text{otherwise.} \end{cases} \quad (21)$$

where R_0 is the basic punitive reward in the experiment, and $R_0 = -1000$.

B. SETUP

In the experiment, one central station and two vice stations are used to perform the task of cooperatively optimizing the geometric configuration in an area consisting of free propagation regions, low SNR regions, and forbidden regions. At the beginning of a task, all the stations are initialized with random positions between the forbidden circle and the boarder of the simulator. At each time step, stations receive and transfer the signals to the central station, which makes decisions about movement on the x -axis and y -axis. The central station is an actor-critic style RL learner. The actor neural network takes in a 12-dimension state s_t and gives a 6-dimension action a_t . It has three hidden layers and each hidden layer has 256 neural units, which is activated

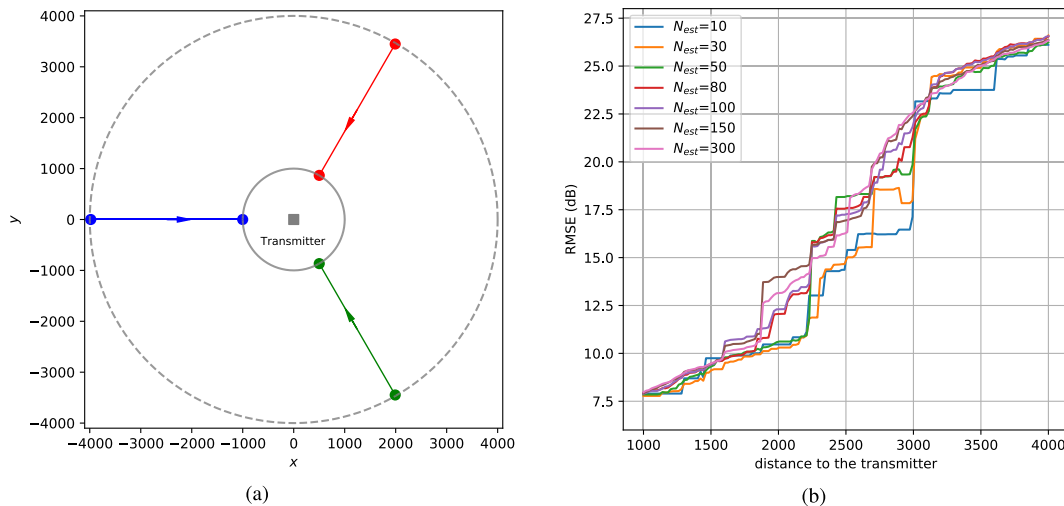


FIGURE 4. Positioning RMSE curves when get close to the transmitter. (a): The trajectory of agents. (b) Positioning RMSE curves with different estimation times $N_{est} = [10, 30, 50, 80, 100, 150, 300]$.

with ReLU function. The actor is designed as a stochastic Gaussian actor which gives a set of parameters $(\mu, \sigma) = [(\mu_1^x, \sigma_1^x), (\mu_1^y, \sigma_1^y), \dots, (\mu_3^x, \sigma_3^x), (\mu_3^y, \sigma_3^y)]$. (μ, σ) determines 6 Gaussian distributions that generate the action for the stations. The critic neural network is similar to the actor but its input and output are (s_t, a_t) and $Q(s_t, a_t)$. In each time steps, all the stations carry the location task together for $N_{est} = 100$ times, and calculate the reward $R_{RMSE}(p, p_*)$. The task is completed when $RMSE(p, p_*) \leq 10$, and agents obtain an immediate reward of 1000. The maximum of time steps taken in one location task is 100. The details of parameters for setting up the experiment is summarized in Table 1.

C. RESULTS AND ANALYSIS

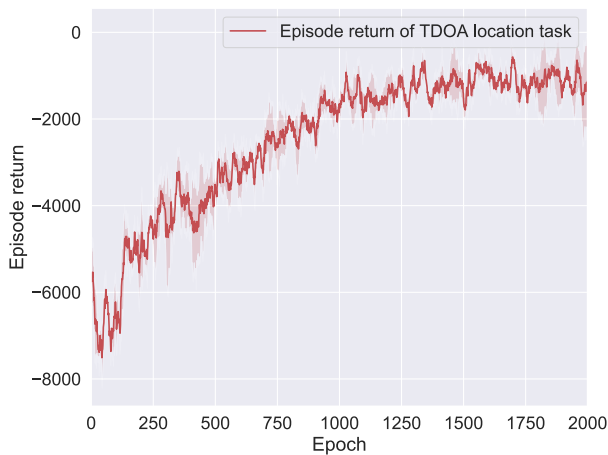
1) TRAINING CURVES

We trained the agents for 2000 epochs, 1.6 million interactions with the TDOA location environment proposed in Section IV-A, over 5 random seeds that initialize the actor and critic. The training curves are presented in FIGURE 5. In the training process, there are some indicators that reflect how well agents master the skills that we expect them to

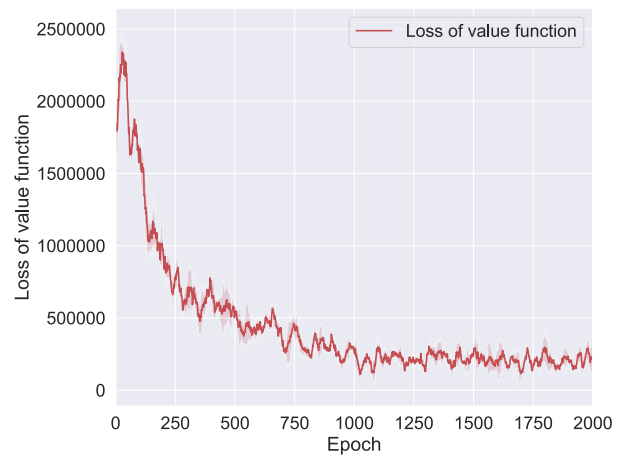
TABLE 1. The details for setting up the TDOA location task with reinforcement learning.

Parameters	description
Environment Scope	radius=4km
Transmitters and stations	1 transmitter, 3 stations (2000, 0, 1500)
Low SNR regions (x, y, radius)	(-2000, 2000, 600) (-1500, 1700, 400)
Actor neural network	(12, 256, 256, 256, 6)
Critic neural network	(18, 256, 256, 256, 1)
Learning rate	actor:3e-4; critic:1e-3
Buffer size	8e5
Max episode length	100
Position estimation times	100
Minimum RMSE for task accomplishment	10
Reward for task accomplishment	1000
Reward for being at forbidden region	-1000

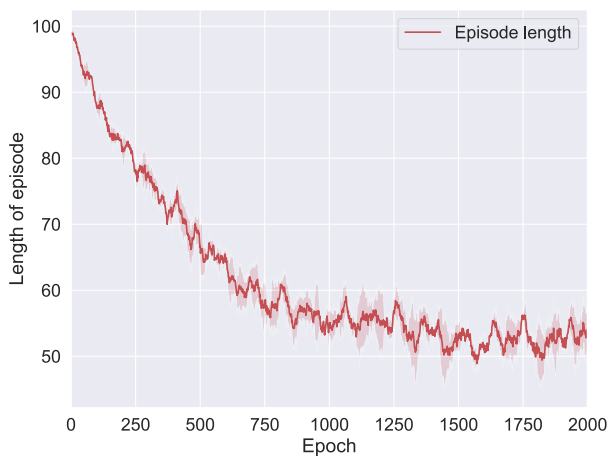
grasp. In a TDOA location task, agents need to find an optimal geometric to reduce the positioning error. We can comprehend the learning process intuitively with: **episode return**, **loss of value function**, **episode length**, and **policy entropy**. Episode return, mainly associated with the positioning error, is a key indicator. From FIGURE 5-(a), we can see that,



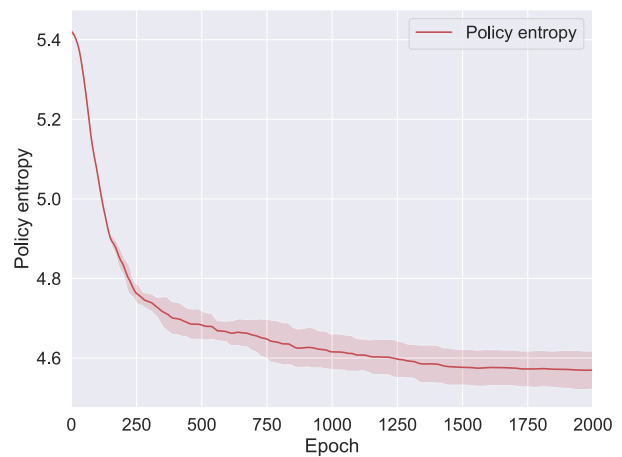
(a)



(b)



(c)



(d)

FIGURE 5. Training curves of the after 2000 epochs of training: (a) Averaged episode reward. (b) Loss of value function. (c) Episode length. (d) Policy entropy.

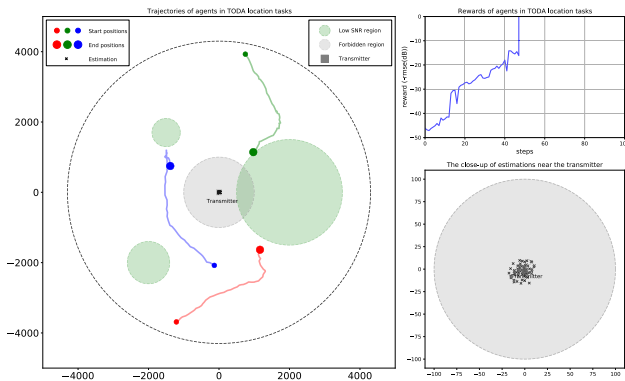


FIGURE 6. The learned agents executing TDOA location task: The left shows the trajectories of all the agents from start positions to the final geometry, the top right is the immediate reward in each step, and the bottom left plots the close-up of estimation near the transmitter.

the episode return increases gradually with training going on, which means that the cumulative positioning error of an episode declines. FIGURE 5-(b) shows the decrease of value function loss, which indicates the critic can estimate the return more accurately. FIGURE 5-(c) illustrates that the total steps of an episode drops with the training process, which suggests actions become more effective. In FIGURE 5-(d), the action entropy declines, which reveals that agents are more confident and have less exploration.

2) THE LEARNED AGENTS

To demonstrate the skills that agents obtained in TDOA location tasks, we visualized the process by recording the trajectories of agents from different initial positions to the optimal geometries. We also plotted the immediate reward in each step, which mainly consists of RMSE of positioning, to understand the decisions made by agents. The estimations near the transmitter are scattered as a close-up to demonstrate the positioning error more clearly. FIGURE 6 shows the trajectory, immediate reward, and the close-up of estimations over a TDOA location task. With more steps are taken, the immediate reward increase gradually, and the final geometric configuration is a triangle with the transmitter at the inner of it, which is consistent with the knowledge in TDOA passive location fields [5], [11]. To test the learned agents comprehensively, we specified four different initial positions which are more difficult for humans to find a path to the optimal geometries. As shown in FIGURE 7, agents can find elegant paths to achieve the given positioning precision in all difficult scenarios. FIGURE 8 shows 12 trajectories with random initial positions, the learned agents can find the optimal geometries efficiently across all the random scenarios. Agents mastered effective strategies to accomplish the location tasks as follows:

- **Take a detour.** Avoid the forbidden and low SNR regions to gain a safer and more effective path.

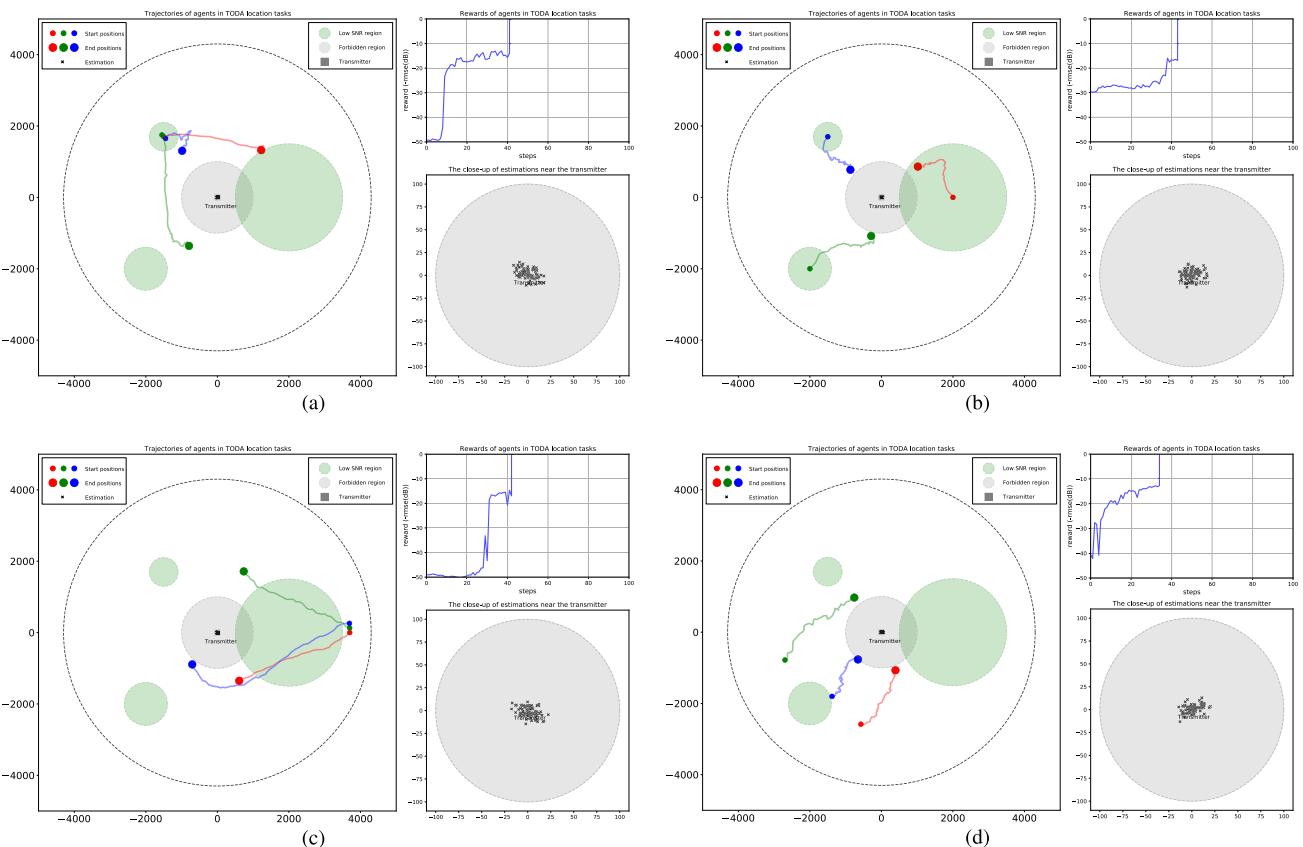


FIGURE 7. Trajectories of learned agents in TDOA location tasks with specified initial geometric configurations.

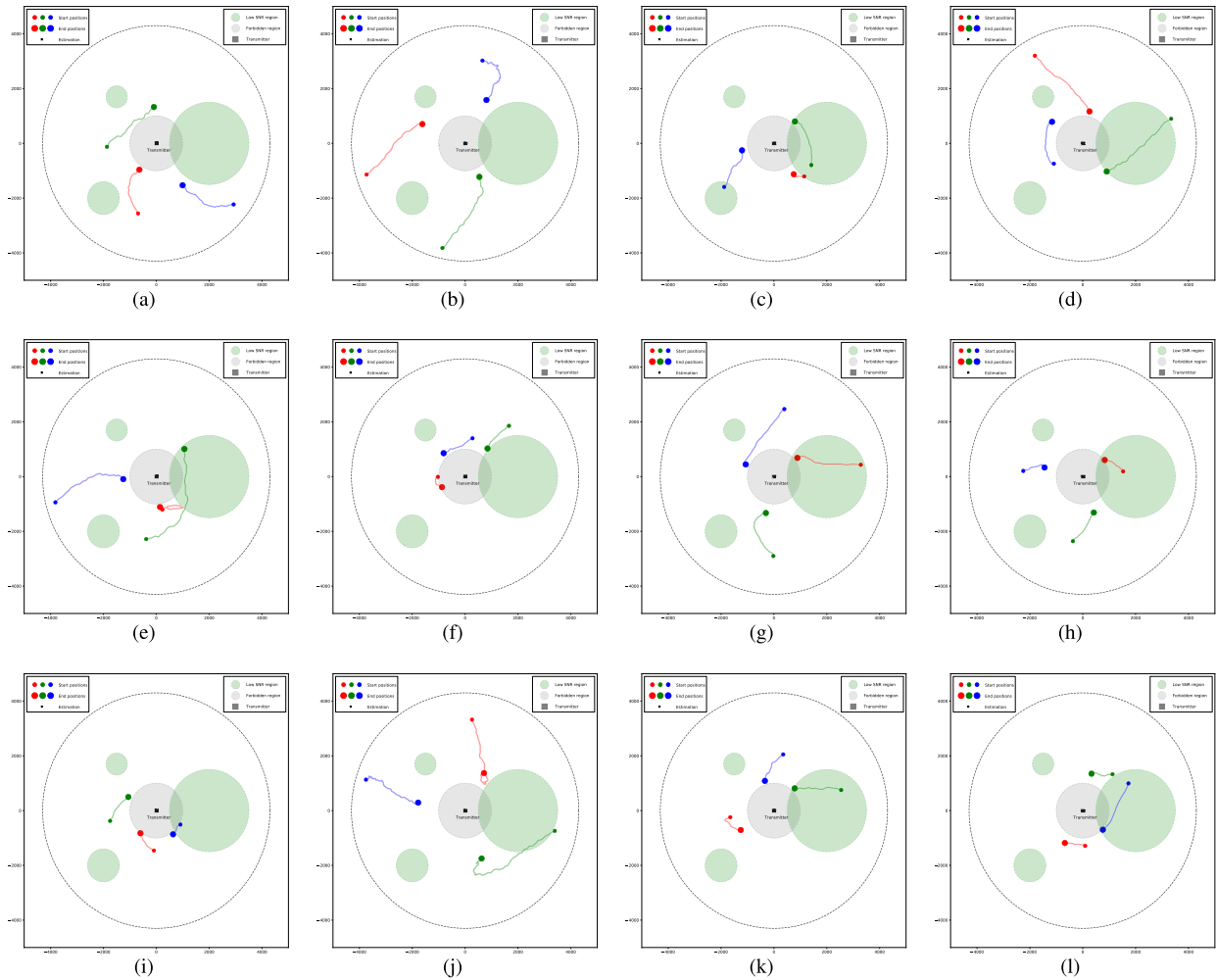


FIGURE 8. Trajectories of finding the optimal geometric configuration in TDOA location tasks when agents are initialized with random positions.

- **Sacrifice the immediate reward.** For the sake of maximizing the long-term return, agents may sacrifice the immediate reward, i.e., crossing low SNR regions to shorter the path to optimal geometry.
- **Geometry first, distance later.** Agents mastered a useful strategy that adjusts the geometry to encircle the transmitter firstly, then shrink the encirclement.

3) COMPARISON STUDY

Though there are significant differences between the method proposed in this paper and heuristic methods, such as PSO [12], [13]. We tried to solve geometry optimization problem with PSO and made more comprehensive comparison with our RL based methods. According to [12], the standard PSO can be described as:

$$\begin{cases} v_{id}(t+1) = wv_{id}(t) + c_1\xi_1(p_{id}(t) - x_{id}(t)) \\ \quad + c_2\xi_2(p_{gd}(t) - x_{id}(t)) \\ x_{id}(t+1) = x_{id}(t) + v_{id}(t+1), \end{cases} \quad (22)$$

where v_{id} and x_{id} are velocity and position of particles, w is the inertial weight, c_1, c_2 are acceleration coefficients, and

ξ_1, ξ_2 are random variables of uniform distribution $U(0, 1)$. The fitness function is vital to PSO algorithms for it provides the heuristic information for particles' evolution. The fitness function in the experiment is defined by:

$$f_{\text{pso}} = -10 \log_{10} \left(1 + \sqrt{\frac{1}{N_{\text{est}}} \sum_{k=1}^{N_{\text{est}}} \|\hat{p}_{\star}^k - \bar{p}_{\star}\|^2} \right), \quad (23)$$

where \bar{p}_{\star} is the center of estimations. Note that, in (23), the fitness function is obtained from the estimations instead of RMSE that requires the background location of the transmitter. In the experiment, $c_1 = c_2 = 2, w = 0.4$, the number of particles is 100, and the start position of an agent is sampled from normal distribution with standard deviation being 200, which is the same as the maximum moving distance for RL agents in each time step. We tested the PSO agents in difficulty scenarios, and FIGURE 9 shows the results. in FIGURE 9, (a)-(d) are snapshots of particles and their best ones at generation 1, 40, 70, 100 with a specified start positions the same as FIGURE 7-(a). The final geometry found by PSO after 100 iteration is not optimal because the

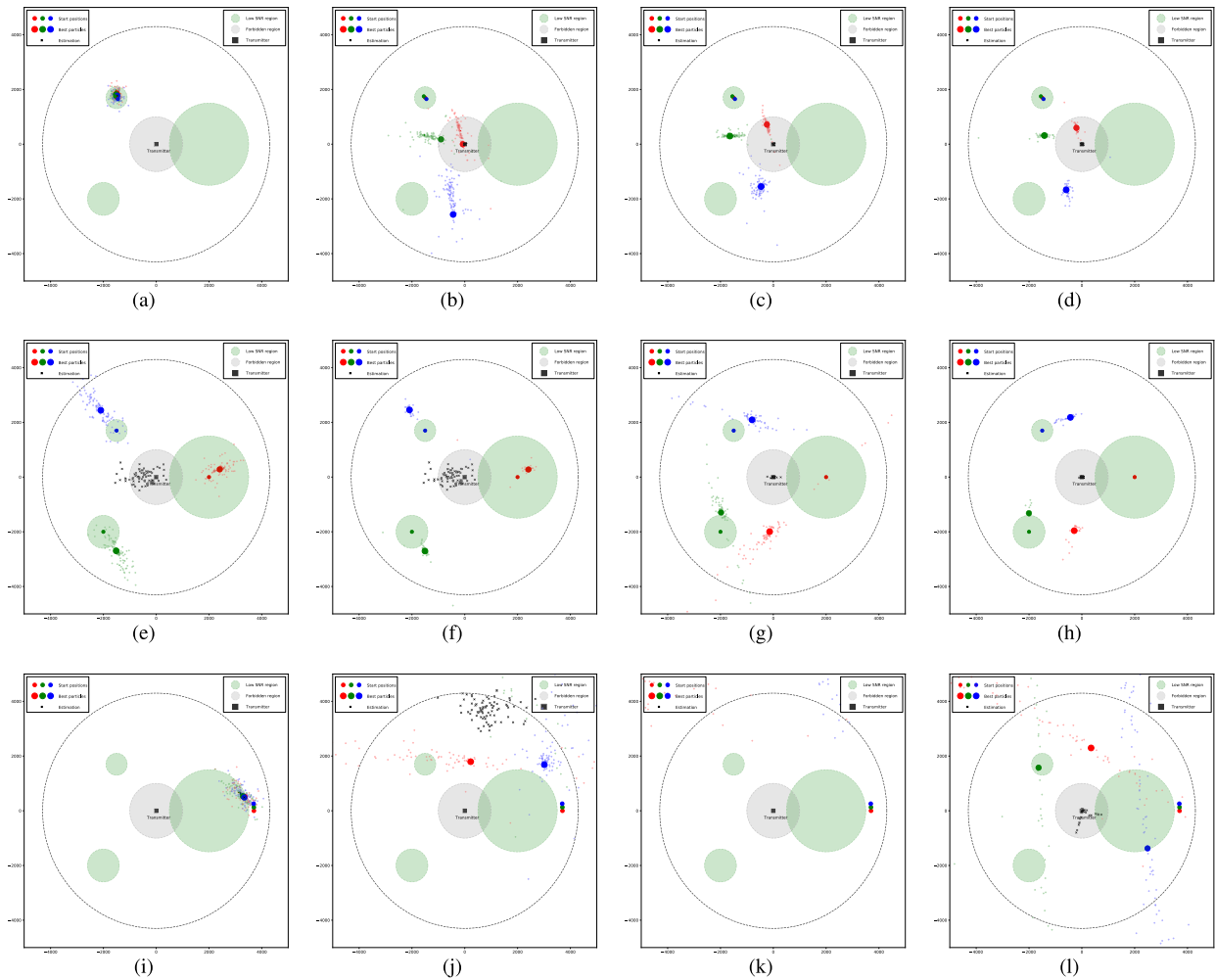


FIGURE 9. The results of comparison experiment that applies PSO to optimize the geometry of TDOA passive location systems. Each row consists of four snapshots in a TDOA location task with specified initial positions.

red agent is in the forbidden region. (e)-(l) are snapshots with other start positions, from which we can see that, the best geometry in the iteration changes abruptly in some cases. In FIGURE 9-(k) most particles are out of the scope of the simulator, which increases the difficulty of agents tracking the path.

From the comparison study presented above, it can be seen that, the RL based method proposed in this paper has obvious advantages in finding a path to optimal geometries in TDOA passive location tasks.

V. CONCLUSION

This paper analyzed the geometry optimization problem of TDOA passive location systems in a complex electromagnetic environment and proposed a reinforcement learning based method to address it in a *try-and-error* fashion. In the method, stations are regarded as mobile agents that can learn to decide where to go. All agents are controlled by an actor-critic reinforcement learner. A TDOA location simulator with complicated electromagnetic is developed to evaluate our method. The empirical results show that, the learned agents exhibited effective strategies that enable them to find

good geometric configurations efficiently. Although TDOA is used in the proposed method, it can be replaced by any other passive location algorithm (e.g., DPD or AOA) to enhance the algorithm flexibility in various location scenarios. In the future, we will address the passive location problem from the perspective of multi-agent reinforcement learning.

REFERENCES

- [1] D. Torrieri, "Statistical theory of passive location systems," *IEEE Trans. Aerosp. Electron. Syst.*, vol. AES-20, no. 2, pp. 183–198, Mar. 1984.
- [2] Y. T. Chan and K. C. Ho, "A simple and efficient estimator for hyperbolic location," *IEEE Trans. Signal Process.*, vol. 42, no. 8, pp. 1905–1915, Aug. 1994.
- [3] K. C. Ho, X. Lu, and L. Kovavisaruch, "Source localization using TDOA and FDOA measurements in the presence of receiver location errors: Analysis and solution," *IEEE Trans. Signal Process.*, vol. 55, no. 2, pp. 684–696, Feb. 2007.
- [4] K. C. Ho, L. Kovavisaruch, and H. Parikh, "Source localization using TDOA with erroneous receiver positions," in *Proc. IEEE Int. Symp. Circuits Syst.*, vol. 3, May 2004, pp. III–453.
- [5] K. Bronk and J. Stefanski, "Bad geometry effect in the TDOA systems," *Polish J. Environ. Stud.*, vol. 16, pp. 11–13, Jan. 2007.
- [6] I. Martin-Escalona and F. Barcelo-Arroyo, "Impact of geometry on the accuracy of the passive-TDOA algorithm," in *Proc. IEEE 19th Int. Symp. Pers., Indoor Mobile Radio Commun.*, Sep. 2008, pp. 1–6.

- [7] B. G. Sun, Q. Miao, J. L. Song, and J. L. Bai, "Analysis of the influence of station placement on the position precision of passive area positioning system based on TDOA," *Fire Control Command Control*, vol. 36, pp. 129–132, 2011.
- [8] D. E. Goldberg, *Genetic Algorithms in Search, Optimization, and Machine Learning*. Reading, MA, USA: Addison-Wesley, 1989.
- [9] K. De Jong, *An Analysis of the Behavior of a Class of Genetic Algorithms*, M.S. thesis, Univ. Michigan, Ann Arbor, MI, USA, 1975.
- [10] J. H. Holland, *Adaptation in Natural and Artificial Systems: An Introductory Analysis with Applications to Biology, Control, and Artificial Intelligence*. Cambridge, MA, USA: MIT Press, 1992.
- [11] B. Wang and L. Xue, "Station arrangement strategy of TDOA location system based on genetic algorithm," *Syst. Eng. Electron.*, vol. 31, no. 9, pp. 2125–2128, 2009.
- [12] Y. Shi and R. C. Eberhart, "A modified particle swarm optimization," in *Proc. IEEE Congr. Evol. Comput.*, 1999, pp. 69–73.
- [13] J. Kennedy and R. C. Eberhart, *Swarm Intelligence*. San Mateo, CA, USA: Morgan Kaufmann, 2001.
- [14] G. Zhou, L. Yang, Z. Liu, and Y. Peng, "Analysis of the influence of base station layout on location accuracy based on TDOA," *Command Control Simul.*, vol. 39, no. 6, pp. 119–126, 2017.
- [15] M. L. Littman, "Value-function reinforcement learning in Markov games," *Cogn. Syst. Res.*, vol. 2, no. 1, pp. 55–66, Apr. 2001.
- [16] L. P. Kaelbling, M. L. Littman, and A. W. Moore, "Reinforcement learning: A survey," *J. Artif. Intell. Res.*, vol. 4, no. 1, pp. 237–285, Jan. 1996.
- [17] R. S. Sutton and A. G. Barto, *Reinforcement Learning: An Introduction*. Cambridge, MA, USA: MIT Press, 2018.
- [18] R. S. Sutton, D. A. McAllester, S. P. Singh, and Y. Mansour, "Policy gradient methods for reinforcement learning with function approximation," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 12, 2000, pp. 1057–1063.
- [19] J. Schulman, S. Levine, P. Moritz, M. Jordan, and P. Abbeel, "Trust region policy optimization," in *Proc. 32th Int. Conf. Mach. Learn.*, 2015, pp. 1889–1897.
- [20] 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. 4th Int. Conf. Learn. Represent.*, 2016.
- [21] D. Silver, A. Huang, C. J. Maddison, A. Guez, L. Sifre, G. van den Driessche, J. Schrittwieser, I. Antonoglou, V. Panneershelvam, M. Lanctot, S. Dieleman, D. Grewe, J. Nham, N. Kalchbrenner, I. Sutskever, T. Lillicrap, M. Leach, K. Kavukcuoglu, T. Graepel, and D. Hassabis, "Mastering the game of go with deep neural networks and tree search," *Nature*, vol. 529, no. 7587, pp. 484–489, Jan. 2016.
- [22] Y. Fenjira and H. Benbrahim, "Deep reinforcement learning overview of the state of the art," *J. Autom., Mobile Robot. Intell. Syst.*, vol. 12, no. 3, pp. 20–39, Dec. 2018.
- [23] D. Silver, J. Schrittwieser, K. Simonyan, I. Antonoglou, A. Huang, A. Guez, T. Hubert, L. Baker, M. Lai, A. Bolton, Y. Chen, T. Lillicrap, F. Hui, L. Sifre, G. van den Driessche, T. Graepel, and D. Hassabis, "Mastering the game of go without human knowledge," *Nature*, vol. 550, no. 7676, pp. 354–359, Oct. 2017.
- [24] V. Mnih, K. Kavukcuoglu, D. Silver, A. A. Rusu, J. Veness, M. G. Bellemare, A. Graves, M. Riedmiller, A. K. Fidjeland, G. Ostrovski, S. Petersen, C. Beattie, A. Sadik, I. Antonoglou, H. King, D. Kumaran, D. Wierstra, S. Legg, and D. Hassabis, "Human-level control through deep reinforcement learning," *Nature*, vol. 518, no. 7540, pp. 529–533, 2015.
- [25] T. Degris and R. S. Sutton, "Off-policy actor-critic," in *Proc. 29th Int. Conf. Mach. Learn.*, 2012, pp. 457–464.



SHENGXIANG LI received the B.S. and M.S. degrees from PLA Strategy Support Force Information Engineering University, Zhengzhou, China, in 2005 and 2018, respectively, where he is currently pursuing the Ph.D. degree. His research interests include wireless communications and reinforcement learning.



GUANGYI LIU received the B.S. degree in communication engineering from PLA Information Engineering University, Zhengzhou, China, in 2005, and the M.S. and Ph.D. degrees in communication engineering from the National Digital Switching System Engineering and Technological Research and Development Center, Zhengzhou, in 2009 and 2015, respectively. His research interests include wireless communications and signal analysis.



SIYUAN DING received the B.S. degree in communication engineering from Shanghai University, Shanghai, China, in 2015, and the M.S. degree in electronic and communication engineering from PLA Strategy Support Force Information Engineering University, Zhengzhou, China, in 2018. Her research interests include machine learning and data mining.



H AISI LI received the B.S. degree in electronic science and technology from Peking University, Beijing, China, in 2018. She is currently pursuing the M.S. degree with PLA Strategy Support Force Information Engineering University, Zhengzhou, China. Her research interest includes radiation source localization.



OU LI received the Ph.D. degree from the National Digital Switching System Engineering and Technological Research and Development Center (NDSC), Zhengzhou, China, in 2001. He is currently a Professor with NDSC. His primary research interests include wireless communication technology, wireless sensor networks, cognitive radio networks, MIMO, and spectrum sensing.

...