Hindawi
Mathematical Problems in Engineering
Volume 2021, Article ID 6691254, 9 pages
https://doi.org/10.1155/2021/6691254

Research Article

A Multiagent Reinforcement Learning Solution for Geometric Configuration Optimization in Passive Location Systems

Shengxiang Li,[1] **Haisi Li**,[1] **Ke Ke**,[2] **Ou Li**,[1] **Guangyi Liu**,[1] **Siyuan Ding**,[3] **and Yijie Bai**[1]

[1]*PLA Strategy Support Force Information Engineering University, Zhengzhou, China*
[2]*National Digital Switching System Engineering and Technological R&D Center, Zhengzhou, China*
[3]*Key Laboratory of Experimental Physics and Computational Mathematics, Beijing 100076, China*

Correspondence should be addressed to Shengxiang Li; lishengxiangzz@163.com and Haisi Li; lihaisi2018@126.com

Received 6 February 2021; Revised 20 April 2021; Accepted 23 May 2021; Published 2 June 2021

Academic Editor: Leonardo Acho

Passive location systems receive electromagnetic waves at one or multiple base stations to locate the transmitters, which are widely used in security fields. However, the geometric configurations of stations can greatly affect the positioning precision. In the literature, the geometry of the passive location system is mainly designed based on empirical models. These empirical models, being hard to track the sophisticated electromagnetic environment in the real world, result in suboptimal geometric configurations and low positioning precision. In order to master the characteristics of complicated electromagnetic environments to improve positioning performance, this paper proposes a novel geometry optimization method based on multiagent reinforcement learning. In the proposed method, agents learn to optimize the geometry cooperatively by factorizing team value function into agentwise value functions. To facilitate cooperation and deal with data transmission challenges, a constraint is imposed on the data sent from the central station to vice stations to ensure conciseness and effectiveness of communications. According to the empirical results under direct position determination systems, agents can find better geometric configurations than the existing methods in complicated electromagnetic environments.

1. Introduction

Passive location techniques are used for various scenarios, such as telecommunication pseudobase station discovery, and aviation interference investigation. Traditional passive location algorithms [1] mainly include angle of arrival (AOA) [2], difference of time arrival (TDOA) [3], and frequency difference of arrival (FOA) [4]. These algorithms locate transmitters after estimating the signal parameter called the two-step positioning methods. The direct position determination (DPD) [5, 6] uses observations from all the stations to locate the transmitter without estimating the signal parameters, which outperforms two-step methods in low signal-to-noise ratio (SNR) scenarios [7].

The geometric configurations of stations can significantly affect the positioning precision [8], both in two-step and DPD positioning algorithms. In the literature, some existing studies tried to obtain general principles in geometric configurations from massive experiments [9, 10]. 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 several studies that have employed heuristic methods, such as genetic algorithm (GA) [11] or particle swarm optimization (PSO) [12], to search the optimal geometry. 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. Therefore, it is vital to adjust geometric configuration to fit the sophisticated electromagnetic environment, so as to improve precision in a passive positioning task. This problem is regarded as a sequential decision-making problem in a real-world complex electromagnetic environment, rather than an optimal geometry searching problem based on empirical models.

Reinforcement learning (RL) is a viable and elegant approach to yield an optimal policy for sequential decision-making problems [13]. The tricky electromagnetic environment can be tracked by RL in a trial-and-error paradigm. The nonlinear and parameterized deep neural network (DNN), providing the compact and powerful representation of experiences, can adapt to the complicated electromagnetic spatial distribution accurately. Therefore, this paper addresses the problem of finding optimal geometric configuration in the passive location system through deep reinforcement learning (DRL) [14].

Under the framework of DRL, a station is used as a mobile agent. The terms' station and agent are hereafter used interchangeably which can receive signals and decide where to go. Agents need to optimize the geometric configuration collaboratively to improve the positioning precision, and they can share information via communication channels to facilitate the collaboration. However, the communication traffic matters when the number of agents increases, especially in adverse communication conditions.

This paper proposes an efficient multiagent reinforcement learning algorithm to optimize the geometric configuration for passive location systems. To this end, each station is regarded as a mobile agent with all agents having a collective objective of finding an optimal geometry to improve the positioning precision. To facilitate the collaborations among agents, they are trained based on value function decomposition, which can solve the credit assignment problem among agents implicitly. For a vice station, it needs to obtain information from other stations to improve the evaluation of the situation and promote the quality of decisions on where to go. Meanwhile, it is necessary to reduce the communication traffic due to transmission and processing challenges. A mutual information objective function then is employed to constrain the messages sent to vice stations to ensure the expressiveness and conciseness. The proposed method is evaluated on simulated DPD positioning tasks in a complicated electromagnetic environment. The results demonstrated that the agents can find better geometric configurations than existing methods.

2. Background

This section introduces the relevant background on passive locations (concretely, DPD) and MARL.

2.1. Passive Location with DPD. Consider H transmitters and L stations intercepting the transmitted signal, as shown in Figure 1. Each station is equipped with an antenna array consisting of M elements. The hth transmitter's position is denoted by $\mathbf{p}_h = [x_h, y_h]^\top$. The complex envelopes of the signals observed by the ℓth station are given by the following equation [5]:

$$\mathbf{z}_\ell(t) = \sum_{h=1}^{H} b_{\ell h}\boldsymbol{\alpha}_\ell(\mathbf{p}_h)v_h\left(t - \tau_\ell(\mathbf{p}_h) - t_h^{(0)}\right) + \mathbf{n}_\ell(t), \quad (1)$$

where $0 \le t \le T$, $\mathbf{z}_\ell(t)$ is a complex time dependent $M \times 1$ observation vector, and b_ℓ is an unknown complex scalar representing the channel attenuation between the hth transmitter and the ℓth station. Moreover, $\alpha_\ell(\mathbf{p}_h)$ is the ℓth array response to the signal transmitted from position $\mathbf{p}_h$, and $v_h(t - \tau_\ell(\mathbf{p}_h) - t_h^{(0)})$ is the hth signal waveform transmitted at time $t_h^{(0)}$ and delayed by $\tau_\ell(\mathbf{p}_h)$. The vector $\mathbf{n}_\ell(t)$ represents noise, interference, and multipath effects on the signals.

For brevity, we use $\alpha_{\ell h}$ and $\tau_{\ell h}$ instead of $\alpha_\ell(\mathbf{p}_h)$ and $\tau_\ell(\mathbf{p}_h)$. The observed signal can be partitioned into K sections with length $(T/K) \gg \max\{\tau_\ell\}$. Taking the Fourier transform of each section, we obtain

$$\begin{aligned}\mathbf{z}_\ell(j,k) &= \sum_{h=1}^{H} b_{\ell h}\boldsymbol{\alpha}_{\ell q}v_h(j,k)e^{-i\omega_j\left[\tau_{\ell h}+t_h^{(0)}\right]} + \mathbf{n}_\ell(j,k) \\ &= \sum_{h=1}^{H} \overline{\boldsymbol{\alpha}}_\ell(j,\mathbf{p}_h,b_{\ell h})\bar{s}_h(j,k) + \mathbf{n}_\ell(j,k), \end{aligned} \quad (2)$$

$$= \mathbf{A}_\ell(j)\bar{\mathbf{v}}(j,k) + \mathbf{n}_\ell(j,k), \quad (3)$$

where $j = 1, \ldots, J$ is the index of Fourier coefficients and $k = 1, \ldots, K$ is the time section index, $i = \sqrt{-1}$.

In (2), we have

$$\begin{aligned}\overline{\boldsymbol{\alpha}}_\ell(j,\mathbf{p}_h,b_{\ell h}) &\triangleq b_{\ell h}\boldsymbol{\alpha}_{\ell h}e^{-i\omega_j\tau_{\ell h}}, \\ \bar{s}_h(j,k) &\triangleq v_h(j,k)e^{-i\omega_j t_h^{(0)}}.\end{aligned} \quad (4)$$

Then, the vector $\overline{\alpha}_\ell(j,\mathbf{p}_h,b_{\ell h})$ concludes all information about the transmitter's position. Furthermore, the phase shift caused by the transmit time $t_h^{(0)}$ is cancelled out when $\bar{s}_h(j,k)$ is used by the DPD method.

In (3), the received signal is presented in matrix notation with

$$\begin{aligned}\mathbf{A}_\ell(j) &\triangleq [\overline{\boldsymbol{\alpha}}_\ell(j,\mathbf{p}_1,b_{\ell 1}),\ldots,\overline{\boldsymbol{\alpha}}_\ell(j,\mathbf{p}_H,b_{\ell H})], \\ \bar{\mathbf{v}}(j,k) &\triangleq [\bar{s}_1(j,k),\ldots,\bar{s}_H(j,k)].\end{aligned} \quad (5)$$

Since the vector $\bar{\mathbf{v}}(j,k)$ is the same at all stations, the observed vectors of all stations can be concatenated together as

$$\mathbf{z}(j,k) = \mathbf{A}(j)\bar{\mathbf{v}}(j,k) + \mathbf{n}(j,k), \quad (6)$$

where

$$\begin{aligned}\mathbf{z}(j,k) &\triangleq \left[\mathbf{z}_1^\top(j,k),\ldots,\mathbf{z}_L^\top(j,k)\right]^\top, \\ \mathbf{A}(j) &\triangleq \left[\mathbf{A}_1^\top(j),\ldots,\mathbf{A}_L^\top(j)\right]^\top, \\ \mathbf{n}(j,k) &\triangleq \left[\mathbf{n}_1^\top(j,k),\ldots,\mathbf{n}_L^\top(j,k)\right]^\top.\end{aligned} \quad (7)$$

Assume the hth column of $\mathbf{A}(j)$ is denoted by $\overline{\alpha}(j,\mathbf{p}_h,\mathbf{b}_h)$, corresponding to the hth emitter, and can be factored as

$$\overline{\boldsymbol{\alpha}}(j,\mathbf{p}_h,\mathbf{b}_h) = \Gamma_h(j)\mathbf{H}\mathbf{b}_h, \quad (8)$$

where $\Gamma_h(j) \triangleq \mathrm{diag}\{\alpha_{1h}^\top e^{-i\omega_j\tau_{1h}},\ldots,\alpha_{Lh}^\top e^{-i\omega_j\tau_{Lh}}\}$ is a diagonal matrix whose elements are the response of the arrays at all stations, $\mathbf{b}_h \triangleq [b_{1h},\ldots,b_{Lh}]^\top$, $\mathbf{H} \triangleq \mathbf{I} \otimes \mathbf{1}_M$, $\mathbf{I}_L$ stands for the identity matrix of sizes $L \times L$, $\mathbf{1}_M$ stands for $M \times 1$ column vector of ones, and $\otimes$ stands for the Kronecker product.

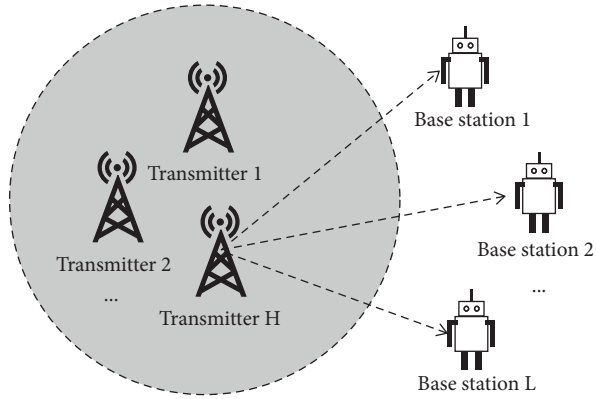

Figure 1: The topology of a passive location system.

The additive noise vector $\mathbf{n}(j,k)$ is assumed to be a realization of a circularly complex Gaussian process with zero mean. The second-order moments is given by

$$\begin{aligned}\mathbb{E}\left[\mathbf{n}(j,k)\mathbf{n}^H(i,m)\right] &= \mathbf{W}\delta_{k,m}\delta_{j,i},\\ \mathbb{E}\left[\mathbf{n}(j,k)\mathbf{n}^\top(i,m)\right] &= 0.\end{aligned} \tag{9}$$

The covariance matrix $\mathbf{W}$ represents the thermal noise as well as interference. In the case of spatially white noise, $\mathbf{W}$ is a block diagonal matrix given by

$$\mathbf{W} = \operatorname{diag}\left\{\sigma_1^2, \ldots, \sigma_L^2\right\} \otimes \mathbf{1}_M. \tag{10}$$

Assume that signals and noise are uncorrelated so that

$$\lim_{K\longrightarrow\infty} \frac{1}{K}\sum_{k=1}^{K} \overline{\mathbf{v}}(j,k)\mathbf{n}^H(j,k) = 0. \tag{11}$$

A matrix is defined to construct the DPD estimator as follows:

$$\mathbf{U}(j) \triangleq \mathbf{Z}_{\mathbf{vz}}^H(j)\mathbf{Z}_{\mathbf{vv}}^{-1}(j), \tag{12}$$

where $\mathbf{Z}_{\mathbf{sr}}^H(j) \triangleq (1/K)\sum_{k=1}^{K} \overline{\mathbf{v}}(j,k)\mathbf{z}^H(j,k)$ and $\mathbf{Z}_{\mathbf{vv}}^{-1}(j) \triangleq (1/K)\sum_{k=1}^{K} \overline{\mathbf{v}}(j,k)\overline{\mathbf{v}}^H(j,k)$. The matrix $\mathbf{Z}_{\mathbf{vv}}(j)$ becomes diagonal for large K if the signals are uncorrelated. The hth column of $\mathbf{U}(j)$ and its ℓth subvector are denoted by $\mathbf{u}_h(j)$ and $\mathbf{u}_{\ell h}(j)$, respectively. The DPD estimator for general noise covariance is presented as

$$\widehat{\mathbf{p}_h} = \arg\max_{\mathbf{p}_h}\left\{\mathbf{u}_h^H\overline{\mathbf{W}}_h^{-(1/2)}\mathbf{P}_{\overline{\mathbf{A}}_h}\overline{\mathbf{W}}_h^{-(1/2)}\mathbf{u}_h\right\}, \tag{13}$$

where $\mathbf{u}_h \triangleq [\mathbf{u}_h^\top(1), \ldots, \mathbf{u}_h^\top(J)]$, $\overline{\mathbf{A}}_h \triangleq \overline{\mathbf{W}}_h^{-(1/2)}\Gamma_h\mathbf{H}$, $\Gamma_h \triangleq [\Gamma_h^\top(1), \ldots, \Gamma_h^\top(J)]^\top$, $\overline{\mathbf{W}}_h \triangleq \mathbf{G}_{h,h}^{-1} \otimes \mathbf{W}$, $\mathbf{G}_{h,h} \triangleq \operatorname{diag}\{[\mathbf{R}_{\mathbf{vv}}(1)]_{h,h}, \ldots, [\mathbf{R}_{\mathbf{vv}}(J)]_{h,h}\}$, $\mathbf{P}_{\overline{\mathbf{A}}_h} \triangleq \overline{\mathbf{A}}_h\overline{\mathbf{A}}_h^+$, and $\overline{\mathbf{A}}_h^+$ is the pseudoinverse of $\overline{\mathbf{A}}_h$.

In the case of partially white noise with a spectral density matrix defined in (10), the DPD estimator becomes

$$\widehat{\mathbf{p}_h} = \arg\max_{\mathbf{p}_h}\sum_{\ell=1}^{L}\sigma_\ell^{-2}\left|\sum_{j=1}^{J} e^{i\omega_j\tau_\ell(\mathbf{p}_h)}\boldsymbol{\alpha}_{\ell h}^H\mathbf{u}_{\ell h}(j)\right|^2. \tag{14}$$

According to [5], the Cramér–Rao lower bound (CLRB) on the covariance of any unbiased estimator of the position vector with no model errors is

$$\mathrm{CRLB} = \frac{1}{2K}\mathrm{Re}\left\{\left(\boldsymbol{\Omega}^H\boldsymbol{\Omega}\right)^{-1}\right\}, \tag{15}$$

where $\Omega \triangleq \mathbf{P}_{\overline{\mathbf{A}}_h}^{\perp}\widetilde{\mathbf{D}}_h[\mathbf{I}_2 \otimes \mathbf{b}_h]$, $\widetilde{\mathbf{D}}_h \triangleq \overline{\mathbf{W}}_h^{-(1/2)}[(\partial\Gamma_h/\partial x_h)(\partial\Gamma_h/\partial y_h)]\mathbf{H}$, and $\mathbf{P}_{\overline{\mathbf{A}}_h}^{\perp} \triangleq = \mathbf{I} - \mathbf{P}_{\overline{\mathbf{A}}_h}$.

The CRLB obtained from (15), determined by received signals $\mathbf{z}$ and locations of stations $\mathbf{p}$, is utilized as the reward function. The CRLB plays a major role in developing the passive location agents through MARL.

2.2. Multiagent Reinforcement Learning. In reinforcement learning, an agent interacts with the environment for a given goal. 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 reward $r_t \in \mathbb{R}$, and moves to the new state $s_t \in \mathcal{S}$. The agent aims to learn a policy that maximizes the long-term reward. The action-value function, which starts from state s, takes action a, and follows policy π, is denoted by $Q^\pi(s,a)$ [13]:

$$Q^\pi(s,a) = \mathbb{E}\left[\sum_{t=0}^{\infty}\gamma^t r_t | s_t = s, a_t = a\right], \tag{16}$$

where $\gamma \in [0,1]$ is the discount factor that determines the importance of future rewards.

In the multiagent reinforcement learning (MARL) [15, 16], agents (robots, UAVs, sensors, etc.) interact with a shared environment to complete the given tasks. Basically, agents are the learnable units that want to learn policies in order to maximize the long-term reward through interactions with the environment. Most MARL problems are classified as NP-hard problems [17] for the sophisticated environments and the combinatorial nature of the problem.

In a cooperative MARL problem, agents must jointly optimize an accumulative scalar team reward over time. The centralized RL approach can be employed to solve the cooperation problem, i.e., all state observations are merged together and the problem is reduced to a single agent problem with a combinatorial action space. Whereas, according to Peter [18], the naive centralized RL methods fail to find the global optimum even if it is possible to solve the problems with such huge state and action spaces. The challenge lies in the fact that some of the agents may become

lazy and unable to learn and cooperate as they should. This may cause the whole system to face a failure. They addressed these problems by training individual agents with a value decomposition network (VDN) architecture. The agents learn to decompose the team value function into agentwise value functions as follows:

$$Q^{\text{tot}}(\rho, a) = \sum_i Q^i\left(\rho^i, a^i; \omega^i\right), \tag{17}$$

where ρ and a represent the observation-action history and joint action, respectively, and ω^i is the value function parameters of agent i. VDN aims to learn an optimal linear value decomposition from the team reward signal, by backpropagating the total Q gradient through deep neural networks representing the summation of individual value functions. The VDN solves the credit assignment among agents implicitly without any specific reward for individual agents. Rashid [19] regarded the cooperative MARL problem as the VDN does, but added a constraint on the objective:

$$\frac{\partial Q^{\text{tot}}}{\partial Q^i} \geq 0, \quad \forall i, \tag{18}$$

which makes the weights of the mixing network positive and ensures monotonic improvement.

3. MARL-Based Geometry Optimization

This section proposes a MARL-based geometric configuration optimization method for passive location systems.

3.1. Model Framework. In this paper, a DPD location system is considered with L mobile stations (e.g., UAVs equipped with positioning equipment), i.e., L DPD 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 vary. To adapt to the complicated electromagnetic spatial distribution accurately, a MARL-based method, with positioning error being the reward function, is considered. The key elements in the MARL scheme are defined as follows:

(i) *States*. At each time step t, agent i intercepts signals, o_t^i, emitted by the transmitter. The total messages it receives from other agents are denoted by m_t^{in}. The state of agent i is denoted by $s_t^i = (o_t^i, m_t^{in}, \mathbf{p}_t^i) \in \mathcal{S}^i$, where $\mathbf{p}_t^i$ is the position of agent i at time t, defined in Section 3.1. Then, the global state is represented as $s = (s^1, \ldots, s^L) \in \mathcal{S}$, where $\mathcal{S} = \prod_{i=1}^L \mathcal{S}^i$.

(ii) *Actions*. Actions represent the decisions regarding where to receive signals at next step. Let $a^i = (\varphi^i, d^i) \in \mathcal{A}^i$ denote the action of agent i, where φ^i and d^i represent its moving direction and distance, respectively. And, the joint action of all agents is denoted as $a = (a^1, \ldots, a^L) \in \mathcal{A}$, $\mathcal{A} = \prod_{i=1}^L \mathcal{A}^i$.

(iii) *Rewards*. This paper aims to develop agents that can properly adjust the geometry automatically to improve the positioning precision. To this end, we evaluate agents' behavior by positioning errors. Two types of positioning errors are considered:

(a) CRLB is an effective index for evaluating the precision of a passive location system. Let the background position of the transmitter be $\mathbf{p}_\star = (x_\star, y_\star)$. Then, the CRLB is a function of state s and the background position $\mathbf{p}_\star$, i.e., $\text{CRLB}(\mathbf{p}_\star, s)$, according to (15).

(b) The statistic error is a popular class of position errors, such as the mean error (ME) and the mean square error (MSE).

Among the errors listed above, only CRLB can assess the geometry without estimating the target position, which reduces considerable amounts of time and computing in training. Therefore, CRLB is used as the reward function in training the DPD agents.

3.2. Learn to Optimize the Geometry. This section presents an efficient multiagent actor-critic algorithm for geometric configuration optimization in passive location tasks. The overall architecture of the proposed method is illustrated in Figure 2. It is developed based on two main considerations: (i) factorizing the global value function into individual value functions with local observations for better collaboration and (ii) utilizing information constraints to facilitate communications and optimize the messages to tackle the transmission challenges.

3.2.1. Value Decomposition. As shown in Figure 2, the global value function is factorized into linear combination of individual value functions as follows:

$$Q_{\text{tot}}(\rho, a; \omega) = \sum_i Q^i\left(\rho^i, a^i; \omega^i\right), \tag{19}$$

where $\rho = (\rho^1, \ldots, \rho^N)$. And, $\rho^i = ((s_1^i, a_1^i), \ldots, (s_t^i, a_t^i))$ is the history of local observations, actions, and messages received. Local value functions are parameterized by $\omega^V = ((\omega^1)^\top, \ldots, (\omega^N)^\top)^\top$. The policy of each station maps the history of observations and actions to the next action: $\pi_{\theta^i}(\rho^i, a^i)$. The joint policy for the location system is denoted by $\pi_\theta(\rho, a) = \prod_i \pi_{\theta^i}(\rho^i, a^i)$. Both actor and critic of each agent utilize the gated recurrent unit (GRU) [20] to process the input of observation history. GRU is a special kind of recurrent neural network that has the ability to capture the long range connections of states. The mixing network and individual value functions are trained in an end-to-end manner by minimizing the TD loss as follows:

$$\mathcal{L}_{\text{TD}}\left(\omega^V\right) = \left[r + \gamma Q_{\text{tot}}\left(\rho', a'; \omega^V\right) - Q_{\text{tot}}\left(\rho, a; \omega^V\right)\right]^2. \tag{20}$$

3.2.2. Information Constraint. For the central station, it must collect observations from all the stations to estimate the transmitter's position. Nevertheless, a vice station j just

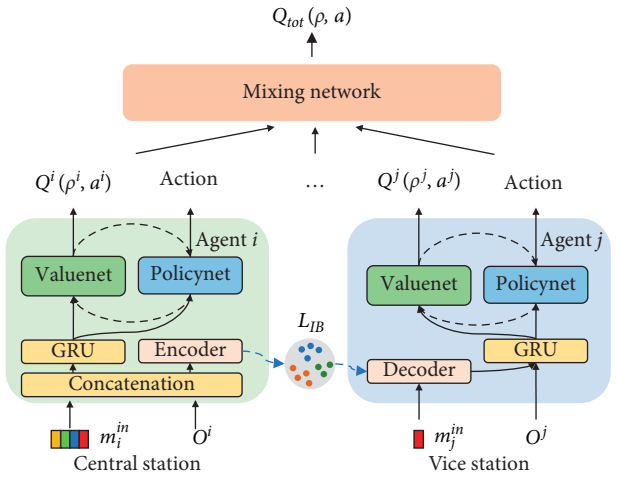

FIGURE 2: Schematics of the proposed method. The global value function is decomposed into the agentwise value functions.

needs the data that can help to make better decisions. Therefore, the central station must learn how to send messages as short as possible but enough for vice stations to act better. A natural solution is to add information constraints. In practice, to improve the effectiveness of messages sent to vice stations, it is necessary to maximize the mutual information of messages and station's actions. Let c be the index of the central station; then, mutual information is defined by

$$I(m_c, a^v) = \int p(m_c, a^v) \log \frac{p(m_c, a^v)}{p(m_c) p(a^v)} \mathrm{d}m_c \mathrm{d}a^v, \quad (21)$$

where m_c represents the message sent from central station to a vice station and $m_j^{in} = m_c\,(j \neq c)$ and a^v is the joint action of all vice stations.

However, if this is the only objective, agents could always ensure a maximally informative representation by taking the identity encoding of raw data ($m_c = \rho$), which contradicts the transmission reduction goal. To increase the conciseness of messages, the complexity of the messages is limited by the constraint $I(m_c, \rho) \leq H_0$. It is then possible to learn an encoding m_c, which is maximally expressive about a^v in addition to being maximally compressive about ρ. With Lagrange multiplier β, the information bottleneck (IB) is defined as follows:

$$\mathscr{L}_{\mathrm{IB}} = -I\left(m_c, a^v; \omega^{\mathrm{IB}}\right) + \beta I\left(m_c, \rho; \omega^{\mathrm{IB}}\right), \quad (22)$$

where ω^{IB} represents the parameters of the encoder and the decoder network.

The value networks are then trained together with the encoder and the decoder by minimizing an overall objective:

$$\mathscr{L}(\omega) = \mathscr{L}_{\mathrm{TD}}\left(\omega^V\right) + \lambda \mathscr{L}_{\mathrm{IB}}\left(\omega^{\mathrm{IB}}\right), \quad (23)$$

where ω consists of ω^V and ω^{IB} and λ is the weight that trades off between these two subobjectives.

The policy gradient [21] of station i is defined as

$$g^i = \nabla \pi_{\theta^i}\left(\rho^i, a^i\right) A\left(\rho^i, a^i\right), \quad (24)$$

where

$$A\left(\rho^i, a^i\right) = Q\left(\rho^i, a^i\right) - \sum_{a^{i,\prime} \in \mathscr{A}^i} \pi\left(\rho^i, a^{i,\prime}\right) Q\left(\rho^i, a^{i,\prime}\right). \quad (25)$$

The policy of station i is optimized through the gradient ascend:

$$\theta^i \leftarrow \theta^i + \eta g^i = \theta^i + \eta \nabla \pi_{\theta^i}\left(\rho^i, a^i\right) A\left(\rho^i, a^i\right), \quad (26)$$

where θ^i refers to the parameters of station i's policy and $\eta > 0$ shows the step size. The details of the training process are shown in Algorithm 1.

4. Experiments

In this section, we develop a simulated electromagnetic environment for passive location tasks, based on which the agents are evaluated.

4.1. Environment. In the experiment, the simulator's geographical coverage is 10 km×10 km, as shown in Figure 3. The transmitter is located in the center of the map and is equipped with an isotopically radiating antenna. The signal model, defined by (1), is employed with some modifications. The channel attenuation is a function of the receiver's position $\mathbf{p}$: $b_\ell(\mathbf{p}) \propto \lambda_s/(4\pi\ d)$, which follows the free space path loss. The noise and interference, as well as the multipath effect, are all compassed in the noise $\mathbf{n}_\ell$, which is modeled by the spatially white noise in (10). There are some low regions, highlighted in green in Figure 3, where the noises are stronger than other areas. It should be noted that due to these low SNR regions, the contours of SNR turn into irregular concentric rings. Furthermore, in the real world, it is also impossible to approach too close to the transmitter; therefore, there is a forbidden 1 km area around the transmitter in the simulator.

4.2. Setup. Consider one central station and $L-1$ vice stations with the task of cooperatively optimizing the geometric configuration in an area consisting of free propagation regions, low SNR regions, and forbidden regions. At each time step t, stations observe the environment to obtain the state s_t and make decisions about moving in direction φ on distance d, e.g., a_t. While moving, stations shut off the positioning and communication devices until arriving the next positions. If the time step t reaches the maximum of $t_{\max}$, the location task ends. Figure 4 demonstrates the process of executing a passive location task in training and execution mode in different branches. With geometry formed by stations at each time step t, the reward is given by the theoretic error bound, CRLB:

$$r_t(s_t, a_t) = -\mathrm{CRLB}(\mathbf{z}, \mathbf{p}, \mathbf{p}_\star), \quad (27)$$

where $\mathbf{z}$ is the received signals and $\mathbf{p} = (\mathbf{p}^1, \ldots, \mathbf{p}^L)$ refers to the positions of all L stations. Also, the root mean square error (RMSE) is calculated to describe the positioning error more intuitively:

(1) Initialize the DPD passive location system with target transmitter emitting signals, specify the number of stations L and the central station c;
(2) Initialize neural network parameters ω, θ
(3) Initialize the iteration counter $t \longleftarrow 0$.
(4) repeat
(5) for $i = 1: L, i \neq c$ do
(6) Intercept the signals o_t^i;
(7) Send the state (s_t^i, a_t^i) to the central station;
(8) end for
(9) The central station intercepts signals o_t^c and send m_c to vice stations;
(10) Update the parameters of value networks:
$\omega \leftarrow \omega + \eta \nabla_\omega \mathscr{L}(\omega)$;
(11) for all i do
(12) Update the parameters of policy network:
$\theta^i \leftarrow \theta^i + \eta \nabla \pi_{\theta^i}(\rho^i, a^i) A(\rho^i, a^i)$;
(13) end for
(14) Update the counter $t \leftarrow t + 1$;
(15) until the task is completed or reaching the maximum of counter

Algorithm 1: Geometric configuration optimization with multiagent reinforcement learning.

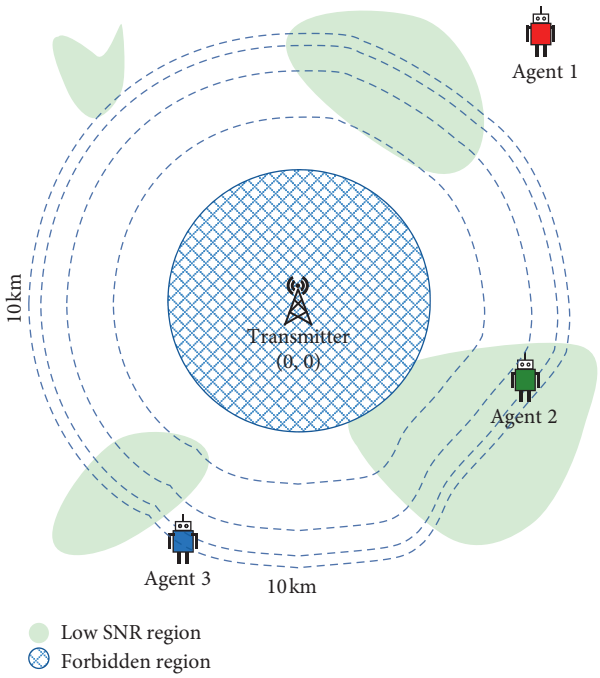

Figure 3: The passive location task environment. There are one transmitter and three mobile stations. Noises become strong in the low SNR regions. Agents are randomly distributed near the edge of the square at the beginning of a task.

$$\mathrm{RMSE}(\mathbf{p}, \mathbf{p}_\star) = \sqrt{\frac{1}{N_{\mathrm{est}}} \sum_{k=1}^{N_{\mathrm{est}}} \left\| \widehat{\mathbf{p}}_\star^k - \mathbf{p}_\star \right\|^2}, \tag{28}$$

where $\widehat{\mathbf{p}}_\star^k$ is the kth estimation of $\mathbf{p}_\star$ and N_{est} denotes the estimation times for each geometric configuration.

4.3. Results and Analysis. The agents are trained in the passive positioning task mentioned above by setting the maximum time step to 100. For the sake of comparison, a basic version of the proposed method is also evaluated. In that version, the central station sends nothing except for the reward (naive DPD agents).

The top segment of Figure 5 shows the learning curve in terms of the averaged reward for DPD agents with communications versus naive DPD agents. DPD agents with communications converge to a much higher return than the naive DPD agents, which indicates that, with messages sent by the central station, vice stations are able to estimate the value function more accurately. In other words, communications are essential to geometry optimization in DPD location tasks. The bottom segment of Figure 5 illustrates the information bottleneck loss $\mathscr{L}_{\mathrm{IB}}$ against the training epochs. $\mathscr{L}_{\mathrm{IB}}$ declines quickly through training. The proposed agents can achieve a higher position precision with lighter communication overhead.

To show the learned decomposition of value functions, Figure 6 demonstrates the error curve, normalized value functions, and the agents' situations when learned agents perform a certain DPD positioning task. According to the top segment of Figure 6, both CRLB and RMSE decline with more steps taken by agents. Furthermore, the RMSE of positioning converges to the CRLB with respect to optimization steps. It means that agents can find geometric configurations where estimation error becomes closer and closer to the CRLB, which is the best achievable output for passive location systems. The middle and bottom segments of Figure 6 show that when the agents are in the low SNR area, their value functions decrease and the positioning error increases, which is consistent with our experiences.

Figure 7 demonstrates the final geometric configuration found by the proposed agents as well as that optimized by the GA [1]. According to the geometry yielded by the GA, there is a station in the low SNR region, which is a suboptimal geometry. In other words, the GA optimizes the geometry on the empirical model, which cannot identify the low SNR regions in the simulator. By contrast, the trained agents can

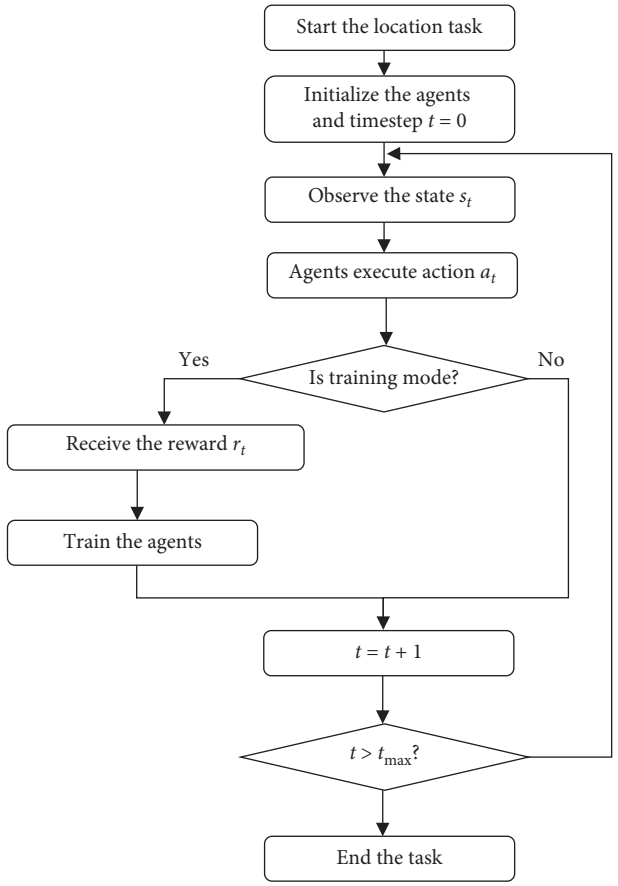

Figure 4: The flowchart of finding optimal geometry in based on MARL algorithm.

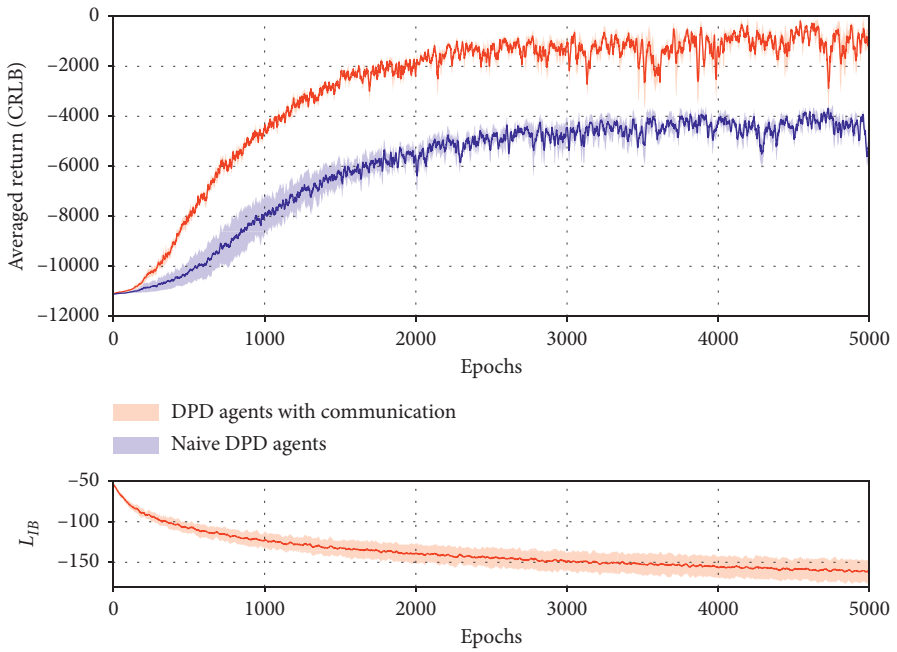

Figure 5: Learning Curves. Top: the averaged return (CRLB) of the naive and communication version of the agents. Bottom: the information bottleneck loss $\mathscr{L}_{\mathrm{IB}}$ curve over five random seeds.

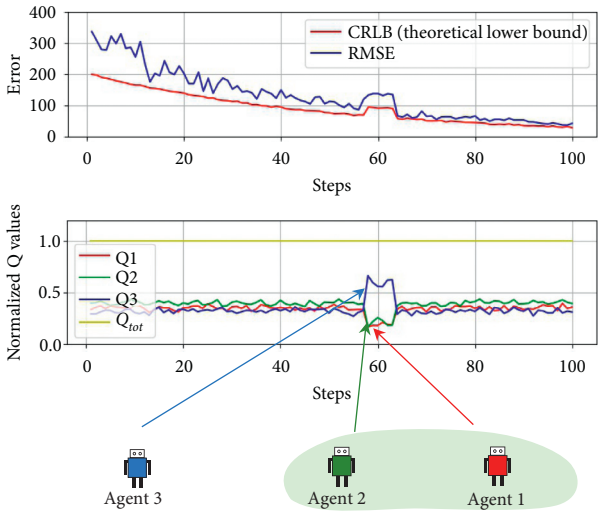

Figure 6: Learned agents executing the passive location task. Top: the CRLB and RMSE decline with more steps taken by agents. Middle and bottom: the agentwise factorized value functions change when agents are in different situations.

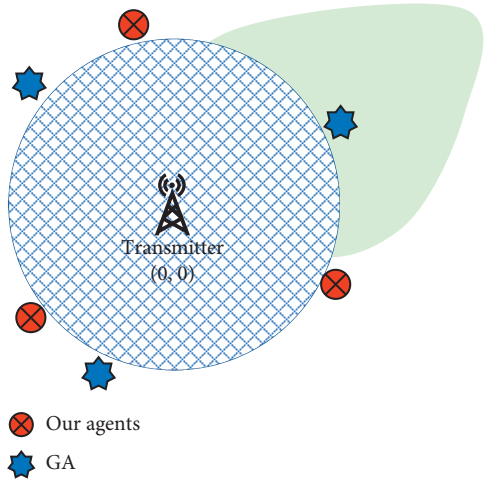

Figure 7: Optimized geometric configuration compared to the existing methods. Red: geometry found by our trained agents. Blue: geometry found by PSO, with one station in the low SNR region.

avoid low SNR regions and find the optimal geometry successfully.

5. Conclusions

This paper analyzed the geometry optimization problem of passive location systems in a complex electromagnetic environment and proposed a MARL method to address it in a try-and-error fashion. In the method, by factorizing the global value function into the agentwise value functions, agents can learn to optimize the geometric configuration cooperatively. Moreover, by adding the mutual information constraints, the communication traffic from the central station to vice stations can be greatly reduced while effectiveness is ensured. A simulator with a sophisticated electromagnetic environment for passive location task is also developed, the results on which showed that the agents could find better geometric configurations than existing methods.

This paper should be seen as a first attempt at learning geometric configuration optimization through MARL in a passive location task. Although DPD is used in the proposed method, it can be replaced by any other passive location algorithm (e.g., TDOA or AOA) to enhance the algorithm flexibility in various location scenarios.

Data Availability

The data used to support the findings of the study are available from Shengxiang Li (lishengxiangzz@163.com) upon request.

Conflicts of Interest

The authors declare that they have no conflicts of interest.

References

[1] J. Don, "Statistical theory of passive location systems"" *IEEE Transactions on Aerospace and Electronic Systems*, vol. 2, pp. 183–198, 1984.

[2] S. O. Al-Jazzar and Y. Jaradat, "AOA-based drone localization using wireless sensordoublets," *Physical Communication*, vol. 42, 2020.

[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 Transactions on Signal Processing*, vol. 55, no. 2, pp. 684–696, 2007.

[4] F. Ma, F. Guo, and L. Yang, "Low-complexity TDOA and FDOA localization: a compromise between two-step and DPD methods," *Digital Signal Processing*, vol. 96, Article ID 102600, 2020.

[5] A. Amar and A. J. Weiss, "Direct position determination in the presence of model errors-known waveforms," *Digital Signal Processing*, vol. 16, no. 1, pp. 52–83, 2006.

[6] J. Weiss, "Direct position determination of narrowband radio transmitters"" in *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, vol. 2, Quebec, Canada, May 2004.

[7] T. Tirer and A. J. Weiss, "High resolution direct position determination of radio frequency sources," *IEEE Signal Processing Letters*, vol. 23, no. 2, pp. 192–196, 2016.

[8] Krzysztof Bronk and J. Stefanski, "Bad geometry effect in the TDOA systems," *Polish Journal of Environmental Studies*, vol. 16, no. Jan, pp. 11–13, . 2007.

[9] I. Martin-Escalona and F. Barcelo-Arroyo, "Impact of geometry on the accuracy of the passive-TDOA algorithm," in *Proceedings of the IEEE International Symposium on Personal, Indoor and Mobile Radio Communications, PIMRC*, pp. 1–6, Cannes, France, October 2008.

[10] B. Sun, "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.

[11] Bo Wang and L. Xue, "Station arrangement stretegy of TDOA location system based on genetic algorithm," *Systems Engineering and Electronics*, vol. 31, pp. 2125–2128, 2009.

[12] G. Zhou et al., "Analysis of the influence of base station layout on location accuracy based on TDOA," *Command Control and Simulation*, vol. 39, pp. 119–126, 2017.
[13] R. S. Sutton and A. G. Barto, *Reinforcement Learning: An Introduction*, MIT Press, Cambridge, MA, USA, 2018.
[14] Y. Fenjiro and H. Benbrahim, "Deep reinforcement learning overview of the state of the art," *Journal of Automation, Mobile Robotics and Intelligent Systems*, vol. 12, no. 3, pp. 20–39, 2018.
[15] L. Busoniu, R. Babuska, and B. De Schutter, "Multi-agent reinforcement learning: a survey," in *Proceedings of 9th International Conference on Control, Automation, Robotics and Vision*, Singapore, December 2006.
[16] K. Tuyls and G. Weiss, "Multiagent learning: basics, challenges, and prospects," *AI Magazine*, vol. 33, no. 3, pp. 41–52, 2012.
[17] D. S. Bernstein, R. Givan, N. Immerman, and S. Zilberstein, "The complexity of decentralized control of markov decision processes," *Mathematics of Operations Research*, vol. 27, no. 4, pp. 819–840, 2002.
[18] S. Peter, "Value-decomposition networks for cooperative multi-agent learning based on team reward," in *Proceedings of the 2018 International Joint Conference on Autonomous Agents and Multiagent Systems*, vol. 3, pp. 2085–2087, Richland, SC, USA, 2018.
[19] T. Rashid, "QMIX: Monotonic value function factorisation for deep multi-agent reinforcement Learning," in *Proceedings of 35th International Conference on Machine Learning*, vol. 10, pp. 6846–6859, Stockholm, Sweden, July 2018.
[20] K. Cho, "Learning phrase representations using RNN encoder-decoder for statistical machine translation," *Computer Science*, 2014, https://arxiv.org/abs/1406.1078.
[21] R. S. Sutton, "Policy gradient methods for reinforcement learning with function approximation," *Advances in Neural Information Processing Systems*, vol. 12, pp. 1057–1063, 2000.