Hindawi
Mathematical Problems in Engineering
Volume 2020, Article ID 1836159, 17 pages
https://doi.org/10.1155/2020/1836159

Research Article

An Algorithm of Reinforcement Learning for Maneuvering Parameter Self-Tuning Applying in Satellite Cluster

Xiao Wang,[1] Peng Shi,[1] Changxuan Wen,[2] and Yushan Zhao[1]

[1]*Beihang University, Beijing 102206, China*
[2]*Beijing Institute of Technology, Beijing, 100081, China*

Correspondence should be addressed to Peng Shi; shipeng@buaa.edu.cn

Received 8 November 2019; Revised 20 March 2020; Accepted 6 April 2020; Published 27 April 2020

Academic Editor: Ibrahim Zeid

Satellite cluster is a type of artificial cluster, which is attracting wide attention at present. Although the traditional empirical parameter method (TEPM) has the potential to deal with the mission of satellite flocking, it is difficult to select the proper parameters. In order to improve the flight effect in the problem of satellite cluster, as well as to make the selection of flight parameters more reasonable, the traditional sensing zones are improved. A 3σ position error ellipsoid and an induction ellipsoid are applied for substituting the traditional repulsing zone and attracting zone, respectively. Besides, we propose an algorithm of reinforcement learning for parameter self-tuning (RLPST), which is based on the actor-critic framework, to automatically learn the suitable flight parameters. To obtain the parameters in the repulsing zone, orientating zone, and attracting zone of each member in the cluster, a three-channel learning framework is designed. The learning process makes the framework finally find the suitable parameters. Numerical experimental results have shown the superiorities compared to the traditional method, which include trajectory deviation and sensing rate or terminal matching rate, as well as the improvement of the flight paths under the learning framework.

1. Introduction

Satellite cluster is a new space architecture emerging after satellite constellation and satellite formation in recent years [1–3]. Different from the satellite formation, which is generally required to design geometric configuration and control desired formation, satellite cluster emphasizes more on the coordination and cooperation among the members in a system. Through the specific technology of satellite cluster, multiple spacecraft with the same or different functions can be connected into an organic whole by a self-organizing network. Thus, the system is expected to have the flexibility to realize one or more tasks [4, 5].

At present, the research of satellite cluster can be generally divided into two types: one is a long-distance loose cluster, which mainly considers the drift and periodic configuration design under long-term condition, and the other is a short-distance cluster, which involves the techniques about cooperative control and collision avoidance. For the long-distance cluster, the boundedness of the system is mainly considered, which means that the influence of various perturbations on the boundary of the cluster will be analysed [6]. From the perspective of fuel optimization, the maneuver sequences will be solved for maintaining the loose flying of the cluster for a long time. Mazal and Gurfil developed a cluster flight control algorithm based on fuel-efficiency for distance-keeping of spacecraft cluster [7]. Based on the relative elements, Wang and Nakasuka applied nonlinear programming for solving the orbit design of fractionated spacecraft [8]. For a long-distance cluster, Dang et al. found the analytic distance bounds for the coplanar relative motion [9]. On the contrary, the techniques of multiagent control, such as graph theory and consistency algorithms, are being continuously studied for the problems of a short-distance cluster. In the field of formation control modelled by second-order dynamics, Ren et al. studied the consensus-based formation control in the absence of centralised leadership [10, 11]. Considering the fixed and

switching topologies, Olfati-Saber and Murray solved the consensus problems for networks of dynamic agents [12]. For the leader-follower consensus problem, Song et al. proposed a pinning control algorithm based on graph theory to handle the condition without a strongly connected interaction graph [13]. Aiming at the multivehicle system of double-integrator dynamic, Qin et al. investigated the consensus strategies to deal with the time-varying reference velocity [14]. However, the complex information link and massive computing are always confusing the algorithms based on graph or consistency theory.

Inspired by biological clusters, humans have constructed a variety of artificial clusters, such as robot clusters and unmanned aerial vehicle clusters. The traditional empirical parameter method (TEPM) has been proved to be effective in many multiagent fields. For describing the motion of flocking particles, Reynolds created a distributed behavioural model [15]. Vicsek et al. suggested a model where the particles were driven with constant velocity and the system was biologically motivated [16]. To reveal the relationship between the individual and the group based on behavioural transitions, Couzin et al. presented a self-organizing model of group formation [17]. Based on these classic models, the algorithms for realizing the specific missions of the cluster were proposed [18], and some attraction/repulsion functions, which are used for achieving swarm aggregation, were designed [19]. With the deep discussion about the interactions between the particles in a swarm, the rule-based control or behaviour-based control was gradually concerned to act in dynamic multiagent systems [20, 21]. Specifically, in the field of aeronautic and aerospace, the behaviour-based path-planning technique for configuring the cluster structures [22], avoiding collision [23], and dealing with needs of aviation swarm convoy [24] were studied, respectively. Due to the wide application prospect of behaviour control methods in biological clusters, more and more space agencies are expecting to introduce the concept of biological clusters into space systems. In this way, it will be possible to make the satellite cluster similar to a biological cluster for completing complex space tasks with simple and cheap spacecraft. Nevertheless, the selection of the behaviour parameters, which is generally decided through the experiences from the scholars, has not been very deeply discussed yet. In order to train the behaviour parameters, one can apply supervised learning if the prior experiences can be obtained. However, such experiences are usually hard to obtain. Therefore, it is a promising direction to find a way to optimize the parameters without the experience data.

In recent years, reinforcement learning has been paid more and more attention in the field of intelligent clusters. Through interacting with the environment, agents in the cluster can optimize their maneuvering strategies under the model-free condition [25,26]. Therefore, the traditional maneuvering strategies, which are based on man-made rules, can be improved. Instead of using the fixed rules, Morihiro et al. proposed a self-organized framework based on reinforcement learning for the flocking agents to conduct group missions [27]. In cooperative multirobot systems, Gu and Yang applied fuzzy policies with policy gradient approach to solve leader-follower problems [28]. Under self-organizing principles derived from natural interactions, Chen et al. solved a swarm pursuit game through a multi-agent reinforcement learning framework [29], and Hung and Givigi presented a Q-learning algorithm, which was applicable to a stochastic environment, for the flocking fixed-wing unmanned aerial vehicles [30]. Therefore, reinforcement learning is a promising method for dealing with the cluster problem of multiple agents. As the most energetic branch of present reinforcement learning, the actor-critic method is suitable for motion problem of continuous agent systems [31].

In this paper, we propose an innovative self-organizing algorithm of reinforcement learning for parameter self-tuning (RLPST), which is based on the actor-critic framework for the path planning of short-distance satellite cluster. The proposed learning algorithm is composed of three channels which can automatically adjust the flight parameters of an agent in the zone of repulsing (ZOR), zone of orientation (ZOO), and zone of attraction (ZOA), respectively. Through iterative learning, the maneuvering strategies of the cluster can be optimized. In this way, the disadvantages under traditional control strategies based on man-made experience parameters for the cluster are broken, and it is expectable to apply the proposed algorithm in a variety of clustering tasks.

The main contributions of this paper are as follows: (1) it is the first time to apply the actor-critic framework into the path planning of satellite cluster. Aiming at two kinds of classical space cluster scenarios, we introduce the reinforcement learning to deal with the relative distance between the members of short-distance satellite cluster, which fills the blank of self-parameter tuning based on the heuristic method in the field of a short-distance satellite cluster. Under the same three flocking principles of Reynolds, we have compared the results under the proposed RLPST and the ones under TEPM and proved the superiority of the proposed algorithm. (2) It is the first time to apply the actor-critic framework to optimize the flight parameters of ZOR, ZOO, and ZOA in a cluster through three channels, respectively, instead of directly optimizing the maneuver of the agent. In this way, it makes full use of the known model information. Besides, the learning difficulty when applying the reinforcement learning to the satellite cluster problem, which has large-scale continuous state and action space, can be reduced.

The structure of this paper is as follows: Section 2 presents the model of satellite cluster and related sensing areas; Section 3 discusses reinforcement learning algorithms for continuous systems; Section 4 applies reinforcement learning to the motion of satellite cluster; Section 5 simulates the proposed algorithm under two classic scenarios, respectively; and Section 6 discusses the simulation results. Finally, Section 7 draws the conclusions.

2. Problem Statement

It is supposed that the subject of the research is a satellite cluster with N members, with a virtual host point, O, which

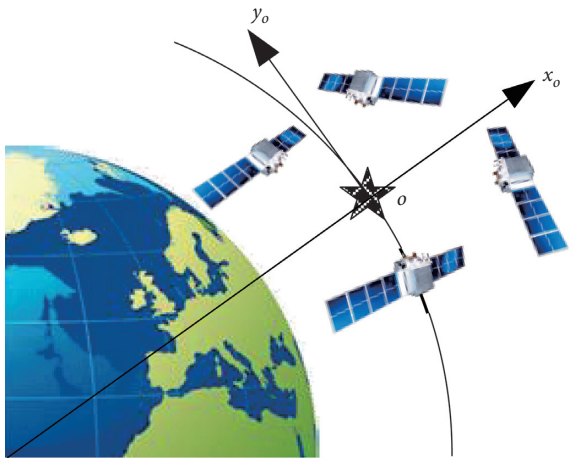

Figure 1: Flocking satellites and the virtual host point.

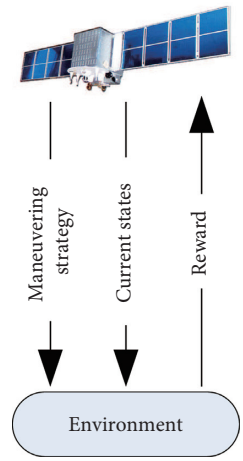

Figure 2: The interaction with the environment of the flocking satellite member.

is near to the center of the cluster in space. Figure 1 draws the flocking satellites and the virtual host point. In Figure 2, it shows that each member in the cluster is an agent, which has the ability of induction, interacting with the environment. By inputting the current states and the maneuvering strategy, the member is able to obtain the reward for preparing the correction of the strategy.

The symbol θ is denoted as the true anomaly, a as the semimajor axis, and e as the eccentricity of the orbit of the virtual host satellite, and then the following equations are obtained [32]:

$$\dot{\theta} = \sqrt{\frac{\mu}{a^3}}\,\frac{(1+e\cos\theta)^2}{(1-e^2)^{3/2}},$$
$$\ddot{\theta} = -\frac{\mu}{a^3}\,\frac{2e\sin\theta(1+e\cos\theta)^3}{(1-e^2)^3}. \tag{1}$$

To facilitate the description of the problem, the following coordinate systems are established: (a) Earth centered inertial ($Ox_iy_iz_i$); (b) orbital coordinate system of the member satellite ($Ox_oy_oz_o$); and (c) orbital coordinate system of the virtual host satellite ($Ox_ry_rz_r$). In $Ox_ry_rz_r$, the position vector $\mathbf{x}_i = [x_i, y_i, z_i]^T$ is denoted. Therefore, according to the two-body motion rule of spacecraft, ignoring the second-order small quantities, the dynamic equation of the ith member in the cluster can be expressed as follows [32]:

$$\left\{\ddot{x}_i - 2\dot{\theta}\dot{y}_i - \ddot{\theta}y_i - \dot{\theta}^2x_i = \frac{\mu x}{r_f^3}\left[2+3\frac{x}{r_f}\right] + f_i^x, \ddot{y}_i + 2\dot{\theta}\dot{x}_i + \ddot{\theta}x_i - \dot{\theta}_i^2 y = -\frac{\mu y}{r_f^3}\left[-1+3\frac{x}{r_f}\right] + f_i^y, \ddot{z}_i = -\frac{\mu z}{r_f^3}\left[-1+3\frac{x}{r_f}\right] + f_i^z,\right. \tag{2}$$

where r_f represents the distance between the mass point of the satellite to the origin and $f_i^j\,(j = x, y, z)$ represents the force in the corresponding channel.

2.1. Position Error Model of a Cluster Member. The position of a satellite is objective; however, it cannot be accurately known. In different types of missions, there will always be measurement, navigation, control, and other deviations and the influences of perturbation. These factors will cause the real orbit diverged from the nominal orbit of the spacecraft, resulting in trajectory deviation. Taking the Gaussian distribution as an example, the covariance matrix of the spacecraft state distribution is denoted as P, and then, it is

obtained that the real state vector $\mathbf{x}$ is in a hyperellipsoid, which is centered on the nominal state vector $\overline{\mathbf{x}}$ [33, 34]. The sphere of the hyperellipsoid can be expressed as follows:

$$(\mathbf{x}-\overline{\mathbf{x}})^{\mathrm{T}}\mathbf{P}^{-1}(\mathbf{x}-\overline{\mathbf{x}})=l^2. \tag{3}$$

Meanwhile, the probability density function of the relative state error distribution is

$$p(l)=\frac{1}{\sqrt{(2\pi)^n}}\int_0^l \exp\left(-\frac{1}{2}r^2\right)f(r)\mathrm{d}r, \tag{4}$$

where n represents the dimension of the state space and l represents the Markov distance constant. Specifically, when $l=3$, equation (3) represents the 3σ error ellipsoid.

The above equation shows a six-dimensional ellipsoid. The matrix, $\mathbf{A}$, is denoted as

$$\mathbf{A}=\frac{\mathbf{P}^{-1}}{l^2}=\ -l^2=\begin{bmatrix}\mathbf{A}_{rr} & \mathbf{A}_{rv}\\ \mathbf{A}_{vr} & \mathbf{A}_{vv}\end{bmatrix}, \tag{5}$$

where the matrix $\mathbf{A}$ is a real symmetric positive definite matrix. The symbol, $\mathbf{R}$, is denoted as the position component of $\mathbf{x}$; therefore, the position error ellipsoid of the spacecraft can be expressed as follows:

$$\mathbf{R}^T\mathbf{A}_{rr}\mathbf{R}=1. \tag{6}$$

2.2. Sensing Area Division of a Cluster Member. Traditionally, for dealing with problems of cluster, the sensing areas, which are generally known as zone of repulsion (ZOR), zone of orientation (ZOO), and zone of attraction (ZOA), are defined from the inside to the outside as the spherical regions [17]. Figure 3(a) shows these traditional ZOR, ZOO, and ZOA. Under such uniform sensing areas, it will be certainly convenient for describing the problem and designing the cluster control strategy. However, considering the location deviations of spacecraft cluster members and the capability of the attaching sensors, it is necessary to improve the way for dividing sensing areas. Here, we redefine the sensing areas of a satellite member in the cluster, which is illustrated in Figure 3(b).

According to Figures 3(a) and 3(b), we see that, for each member satellite in the cluster, there exists three sensing areas. Compared with the traditional sensing areas, the redefined ones have replaced the ZOR and ZOA part with the 3σ error ellipsoid and the induction ellipsoid, respectively. The details of the redefined sensing areas are expressed below.

2.2.1. ZOR. The 3σ error ellipsoid area of each member is defined as the zone of repulsion. It is assumed that the position deviation obeys the Gaussian distribution. As a result, each member in the cluster has its own position error ellipsoid. If the ellipsoid of a specific member interacts with the one of other individuals, the collision between the two members may occur. Therefore, such an ellipsoid is the ZOR for making repulsive force to avoid the probable collision. The members in the ZOR will make repulsive force on the center member. In this way, it will avoid the individuals in the cluster getting too close from each other.

2.2.2. ZOO. The orientation area is defined as a standard sphere with a specific radius. For a specific member, its ZOO is an ideal zone that neighbours, which are located in such an area, keep suitable distance with this specific member. This member will receive the orientation force from the neighbours in its ZOO, which makes the member tends to align its speed with its neighbours gradually. In this way, the flight process will be smooth.

2.2.3. ZOA. The attracting area is defined as the induction ellipsoid. Traditionally, the attracting area is uniform. However, in the case of spacecraft cluster problem, due to the capability of the sensing elements, the sensing ability may be strong or weak in different directions. Therefore, we use an ellipsoid model to nearly describe the induction area of the member in the cluster.

2.3. Location Criterion of Sensing Zones. For the members located in the sensing areas of the ith satellite, it is important to determine which region these members belong to. In traditional ways, the belonging sensing area is usually determined by the location of the mass center of the member. Different from the traditional method, this paper applies the idea of the Box method [35], which takes the location relation of error ellipsoids as the criterion to judge whether two members in the cluster are repulsive or not. If the error ellipsoids intersect with each other, the repulsion force will be generated between the two members.

In order to detect the position relation between the two error ellipsoids, the algebraic criterion is needed. During this process, it needs to carry out the affine transformation on the two ellipsoids. The process of affine transformation is shown in Figure 4.

Suppose that the S_1 frame, which is centered at the nominal mass point of ith member, is parallel to $Ox_ry_rz_r$. Then, the position error ellipsoid of the ith member is expressed as follows:

$$\mathbf{R}^T\mathbf{A}_{rr}^i\mathbf{R}=1. \tag{7}$$

Denote the symbol $\mathbf{X}=[x,y,z,1]^T$; therefore, the error ellipsoids of the ith member and the jth member can be transformed as

$$\begin{aligned}\mathbf{X}^T\mathbf{B}_i^{S_1}\mathbf{X}&=\mathbf{X}^T\begin{bmatrix}\mathbf{A}_{rr}^i & 0_{3\times1}\\ 0_{1\times3} & -1\end{bmatrix}\mathbf{X}=0,\\ \mathbf{X}^T\mathbf{B}_j^{S_1}\mathbf{X}&=\mathbf{X}^T\left(\mathbf{T}_1^{-1}\right)^T\begin{bmatrix}\mathbf{A}_{rr}^j & 0_{3\times1}\\ 0_{1\times3} & -1\end{bmatrix}\mathbf{T}_1^{-1}\mathbf{X}=0,\end{aligned} \tag{8}$$

where $\mathbf{B}_i^{S_1}$ and $\mathbf{B}_j^{S_1}$ are ellipsoidal quadratic matrices and $\mathbf{T}_1$ is a translation matrix for the jth member.

The details of the affine transformation process are shown in Appendix A. Here, we have the final expressions:

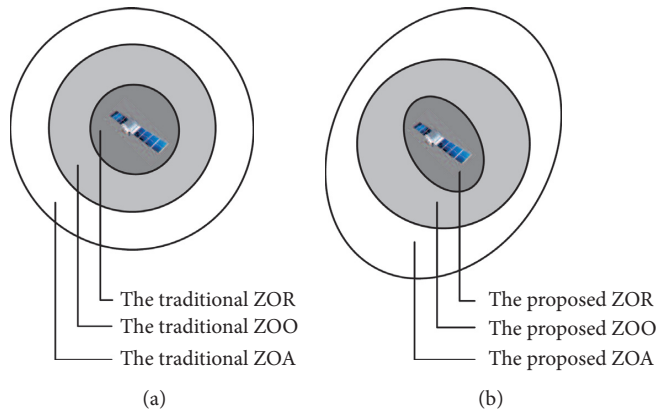

Figure 3: Sensing area division of member satellite in the cluster: (a) traditional sensing areas; (b) proposed sensing areas.

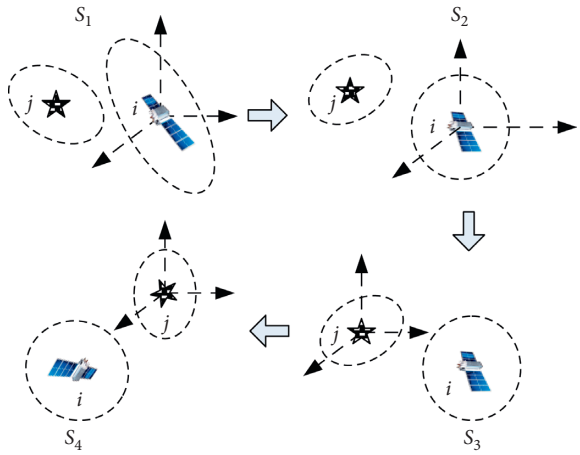

Figure 4: Affine transformation process on the ellipsoids of the ith and the jth member.

$$\mathbf{X}^T \mathbf{B}_i^{S_4} \mathbf{X} = (x - d_1)^2 + (y - d_2)^2 + (z - d_2)^2 - 1 = 0, \quad (9)$$

$$\mathbf{X}^T \mathbf{B}_j^{S_4} \mathbf{X} = \frac{x^2}{a^2} + \frac{y^2}{b^2} + \frac{z^2}{c^2} - 1 = 0, \quad (10)$$

where the frame comes to S_4 and the condition $a \le b \le c$ is satisfied.

Therefore, we can obtain the standard discriminants as equations (9) and (10). The characteristic polynomial is defined as follows:

$$f(\lambda) = \det\left(\lambda \mathbf{B}_j^{S_4} + \mathbf{B}_i^{S_4}\right). \quad (11)$$

The relevant characteristic equation of the above polynomial is

$$f(\lambda) = 0. \quad (12)$$

According to the location judging algebraic criterion [36], the position relation between the ellipsoid of the ith member and the one of the jth members can be detected.

If the characteristic equation has two different real roots, the two ellipsoids are separated. Thus, it will be easy to judge if the jth member locates in the ZOO or ZOA of the ith member or not. Otherwise, the two ellipsoids are not separated, which means that the jth member locates in the ZOR of the i member, and the repulsive force is generated.

2.4. Analysis of the Force Acting in Sensing Areas. For the ith member, when the sensing sets X_r, X_o, and X_a, which relate to the ZOR, ZOO, and ZOA, respectively, are obtained, we can calculate the force directions of the ith member. It is noted that, for each member in a cluster, the member will have its own ZOR, ZOO, and ZOA. Therefore, we only talk about the condition of an arbitrary member in a cluster here. When this arbitrary member is mentioned, it is called "center member" for distinguishing it from the neighbours in its three sensing areas.

2.4.1. Force Direction in ZOR. From the left part in Figure 5, it shows that the individual m is closer to the center member than the individual n under the S_1 frame. According to the traditional repulsive rule, the center member will be repulsed by the individual m more than the individual n. However, for the ellipsoidal ZOR of the center member, the intensity of the repulsive force should be related to the close degree from the individual m or n to the boundary of ZOR. Because the traditional repulsing area is defined as a standard sphere, the boundary is uniform in all directions, which is convenient to calculate the intensity of the repulsive force. Therefore, a special treatment is needed to deal with the nonuniform boundary of the repulsive area, which is shown in the right part in Figure 5. After the affine transformation process, the distance from the center member to the individual m and the one to the individual n is approximately equal. The reason of this situation is because that both of the individual m and n are originally near to the boundary of the repulsing area.

Therefore, through the matrix, $\mathbf{T}_{tr}$, which represents the transformation matrix from the S_1 frame to the S_2 frame, the direction of repulsive force of the ith member in the cluster is expressed as

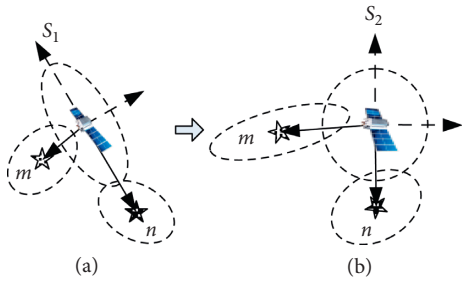

Figure 5: Affine transformation process in repulsing area.

$$p_r^i = \mathrm{T}_{\mathrm{tr}}^{-1} \sum_{j \in X_r} \left(\left(\frac{1}{\left\| \tilde{r}_{ij} \right\|} - \frac{1}{d_r} \right) \tilde{r}_{ij} \right) \Big/ \left\| \sum_{j \in X_r} \left(\left(\frac{1}{\tilde{r}_{ij}} - \frac{1}{d_r} \right) \tilde{r}_{ij} \right) \right\|, \tag{13}$$

where $\tilde{r}_{ij}$ is the relative position vector in S_2 frame and d_r is the boundary of the repulsing area, which is expressed as follows:

$$d_r = 1.2 \max_{j \in X_r} \left\| \tilde{r}_{ij} \right\|. \tag{14}$$

It is noticed that the constant coefficient 1.2 is used to avoid ambiguity caused by zero denominator.

2.4.2. Force Direction in ZOO. Due to the specific definition of ZOO, which is a standard sphere with radius d_m, the direction of orientation force can be expressed as

$$p_o^i = \frac{\mathbf{v}_i + \sum_{j \in X_o} \mathbf{v}_j}{\left\| \mathbf{v}_i + \sum_{j \in X_m} \mathbf{v}_j \right\|}, \tag{15}$$

where $\mathbf{v}_i$ and $\mathbf{v}_j$ represent the velocity of the ith member and the jth member in $Ox_r y_r z_r$, respectively.

2.4.3. Force Direction in ZOA. Compared with the traditional ZOA, the proposed ZOA is set as the induction area of the center member, which means that the sensing ability is not uniform for the center member. In addition, we need to guarantee that the intensity of attractive force should be zero at the boundary of ZOO. As a result, the corresponding boundary of attracting area of each member in the cluster needs to be calculated.

The direction of the attractive force is expressed as

$$p_a^i = \sum_{j \in X_a} \frac{\left(\left(1/d_a - r_{ij} \right) - \left(1/d_a - d_m \right) \right) \mathrm{r}_{ij}}{\left\| \left(\left(1/d_a - r_{ij} \right) - \left(1/d_a - d_m \right) \right) \mathrm{r}_{ij} \right\|}, \tag{16}$$

where d_a is the boundary of the attracting area, which can be defined as follows:

$$d_a = 1.2 \max_{j \in X_a} \left\| r_{ij} \right\|. \tag{17}$$

It is noted that, in Figure 6, we see that the individual m and the individual n are located at the boundary of ZOA. The center member is expected to judge which is the farthest neighbour in its ZOA. Then, the relative position vector from the center to that neighbour will be used to generate d_a.

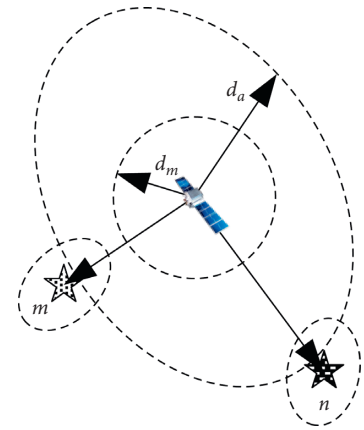

Figure 6: Affine transformation process in attracting area.

When the virtual host satellite is in a circle orbit or near-circle orbit, the conditions, $\dot{\theta} = n$ and $\ddot{\theta} = 0$, are satisfied. To denote the symbol as $X = \left[x \;\; y \;\; z \;\; \dot{x} \;\; \dot{y} \;\; \dot{z} \right]^T$, the dynamic model, which is expressed in equation (2), can be rewritten as follows:

$$\begin{aligned} \dot{X} &= \begin{bmatrix} 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 \\ 3n^2 & 0 & 0 & 0 & 2n & 0 \\ 0 & 0 & 0 & -2n & 0 & 0 \\ 0 & 0 & -n^2 & 0 & 0 & 0 \end{bmatrix} \begin{bmatrix} x \\ y \\ z \\ \dot{x} \\ \dot{y} \\ \dot{z} \end{bmatrix} + \begin{bmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} f_x \\ f_y \\ f_z \end{bmatrix} \\ &= BX + Ca. \end{aligned} \tag{18}$$

To denote the symbols, $c_{ri} \in [-c_{ri}^{\min}, c_{ri}^{\max}]a$, $c_{oi} \in [-c_{oi}^{\min}, c_{oi}^{\max}]$, and $c_{ai} \in [-c_{ai}^{\min}, c_{ai}^{\max}]$ as the flight parameters related to ZOR, ZOO, and ZOA respectively, the motion controller of the ith member in the spacecraft cluster is designed as

$$a_i = -BX - \nu + c_{ri} p_r^i + c_{oi} p_o^i + c_{ai} p_a^i, \tag{19}$$

where $\nu = \left[\dot{x} \;\; \dot{y} \;\; \dot{z} \right]^{\mathrm{T}}$ and $i = 1, 2, \ldots, N$. It is noticed that the term $-BX - \nu$ is added to make the agent move stably during the control gap.

Based on equations (13) to (16), the force directions of the ith member, p_r^i, p_o^i, and p_a^i, can be calculated. Therefore, for the controller shown in equation (19), the key is to find the corresponding parameters c_{ri}, c_{oi}, and c_{ai}. The effect of cluster flight will be largely determined by these parameters.

In traditional ways, the parameters are selected according to the experimental results or the expert experiences, which is known as TEPM. Nevertheless, considering the intelligent development of spacecraft and the raising labour cost, the satellite cluster needs to have a certain automatic capability to adjust the parameters in the future. To achieve this goal, an innovative algorithm of RLPST is proposed, which applies the reinforcement learning framework and is expected to make the flight parameters self-tuned along with multiple learning times.

3. Continuous Reinforcement Learning

3.1. The Fuzzy Inference System. In order to apply the reinforcement learning into the space cluster, which has continuous dynamic systems, it is reasonable to find a way to not only avoid the curse of dimensionality but also have the clear physical meaning. Therefore, a zero-order Takagi–Sugeno (T-S) fuzzy system is employed as the approximator. It is assumed that the fuzzy system has L rules and n input variables. The fuzzy inference rule is

$$\text{Rule } l\text{: IF } s_1 \text{ is } F_1^l, \ldots, \text{ and } s_n \text{ is } F_n^l \text{ then } z_l = \phi_l, \tag{20}$$

where $s_i\,(i = 1, \ldots, n)$ represents the ith input of the fuzzy system, F_i^l represents the fuzzy set of the ith input variable, z_l represents the output of the lth rule, and ϕ_l represents the output parameter.

With the h membership functions of each s_i, the output of the fuzzy system is expressed as

$$Z(\mathbf{s}) = \frac{\sum_{l=1}^{L}\left[\left(\prod_{i=1}^{n}\mu^{F_i^l}(s_i)\right)\phi^l\right]}{\sum_{l=1}^{L}\left(\prod_{i=1}^{n}\mu^{F_i^l}(s_i)\right)} = \sum_{l=1}^{L}\Psi_l(\mathbf{s})\phi_l, \tag{21}$$

where $\mathbf{s} = [s_1, \ldots, s_n]^{\mathrm{T}}$ is the state vector and $\mu^{F_i^l}$ is the membership function of s_i under the lth rule. In addition, the expression of $\Psi_l(\mathbf{s})$ is as follows:

$$\Psi_l(\mathbf{s}) = \frac{\prod_{i=1}^{n}\mu^{F_i^l}(s_i)}{\sum_{l=1}^{L}\left(\prod_{i=1}^{n}\mu^{F_i^l}(s_i)\right)} = \frac{\omega_l(\mathbf{s})}{\sum_{l=1}^{L}\omega_l(\mathbf{s})}. \tag{22}$$

3.2. The Actor-Critic Learning Algorithm. Reinforcement learning is a type of algorithm that interacts with the environment. The agent optimizes its behaviour through the rewards obtained from the environment for maximizing the total benefits. In the Markov process, the value function of reinforcement learning can be expressed as

$$V_t(\mathbf{s}_t) = E\left\{\sum_{i=t}^{\infty}\gamma^{i-t}R_i\right\}, \tag{23}$$

where $\gamma \in [0, 1)$ is the discount factor and R_i is the immediate reward which is obtained from the environment.

In order to solve Markov decision problem in continuous action space, a type of reinforcement learning algorithm called adaptive heuristic critic (AHC) has been widely studied and applied. In the AHC algorithm, the value function and the policy function are approximated, respectively. In this way, the learning structure is called the actor-critic framework. In such a learning algorithm, the critic part is used to estimate the value function, while the actor part is used to generate the action. To generalize the state space and the action space, the critic part and the actor part are both composed of T-S systems. To apply the temporal difference (TD) learning method, we need two critic parts for estimating the current value function $V_t(\mathbf{s}_t)$ and the next value function $V_t(\mathbf{s}_{t+1})$. The temporal difference can be expressed as follows:

$$\Delta_t = R_t + \gamma V_t(\mathbf{s}_{t+1}) - V_t(\mathbf{s}_t). \tag{24}$$

E is denoted as the variance of the difference signal, which is shown as

$$E = \frac{1}{2}\Delta_t^2, \tag{25}$$

and the adaptive update rule of the parameters in the critic is expressed as

$$\phi^C(t+1) = \phi^C(t) - \alpha\frac{\partial E}{\partial \phi^C}, \tag{26}$$

where α is the learning rate of the critic.

Furthermore, according to the gradient descent method, it is shown that

$$\frac{\partial E}{\partial \phi^C} = \Delta_t\left[-\frac{\partial V_t(\mathbf{s}_t)}{\partial \phi^C}\right]. \tag{27}$$

To sum up, we have

$$\phi^C(t+1) = \phi^C(t) - \alpha\left[R_t + \gamma V_t(\mathbf{s}_{t+1}) - V_t(\mathbf{s}_t)\right]\left[-\frac{\partial V_t(\mathbf{s}_t)}{\partial \phi^C}\right], \tag{28}$$

$$pc\frac{\partial V_t(\mathbf{s}_t)}{\partial \phi^C} = \left[\Psi_1(\mathbf{s}_t), \Psi_2(\mathbf{s}_t), \ldots, \Psi_L(\mathbf{s}_t)\right]. \tag{29}$$

Combining with equation (22), equation (28) can be solved.

The adaptive update rule for the critic part is shown as above. As for the actor part, the adaptive update rule of the output parameter, ϕ^A, is expressed as

$$\phi^A(t+1) = \phi^A(t) + \beta\Delta_t\frac{\partial u_t}{\partial \phi^A}, \tag{30}$$

where β is the learning rate of the actor. The partial derivative of u_t is expressed as follows:

$$\frac{\partial u_t}{\partial \phi^A} = \left[\Psi_1(\mathbf{s}_t), \Psi_2(\mathbf{s}_t), \ldots, \Psi_L(\mathbf{s}_t)\right]. \tag{31}$$

4. Algorithm of Reinforcement Learning for Parameter Self-Tuning in Satellite Cluster

The proposed learning framework in this paper is single-looped, which can be divided into three channels of repulsing area (r), orientating area (o), and attracting area (a), respectively. The input of the fuzzy system is single, which is defined as the proportion of the total number of the sensing members in every sensing area. The ith member is taken as an example, and its inputs for fuzzy systems are expressed as

$$\begin{aligned} s_a &= \frac{n_a}{N}, \\ s_o &= \frac{n_o}{N}, \\ s_r &= \frac{n_r}{N}, \end{aligned} \tag{32}$$

where n_a, n_o, and n_r represent the number of sensing neighbours in the repulsing area, orientating area, and attracting area, respectively. Besides, the symbol N represents the total number of all sensing neighbours of the ith agent. Therefore, s_a, s_o, and s_r represent the proportions of sensing members in the corresponding areas.

The critic part and the actor part are composed of fuzzy systems. The inference rule is shown as

$$R_l\big|_q : \text{IF } s_q \text{ is } A_q^l \text{ THEN } Z_l\big|_q = \phi_l^C\big|_q, \tag{33}$$

where $\{\cdot\}|_q$ represents the variable in the q channel ($q = \{a, o, r\}$). As mentioned in Section 3.1, it is supposed that the input has h membership functions and the fuzzy system has L rules in total. It is noticed that, because the input is single, it meets that $L = h$. Therefore, according to the membership degree of the input, the output can be calculated:

$$\Psi_l\big|_q = \frac{\mu^{A_q^l}}{\sum_{l=1}^{L}\left(\mu^{A_q^l}\right)}, \tag{34}$$

$$V_q = \sum_{l=1}^{L}\left(\Psi_l\,\big|_q\right)\cdot\left(\phi_l^C\,\big|_q\right), \quad q = r, o, a. \tag{35}$$

The fuzzy inference process of the actor part is similar to that of the critic part and the difference lies in the consequent parameter to each membership degree:

$$c_q = \sum_{l=1}^{h}\left(\Psi_l\,\big|_q\right)\cdot\left(\phi_l^A\,\big|_q\right), \quad q = r, o, a. \tag{36}$$

The three sensing areas have different proportions of sensing members; therefore, the designed reward function, $R_t|_q$, is expressed as

$$R_t = \begin{cases} R_t\,\big|_r = 10, R_t\,\big|_o = 0, R_t\,\big|_a = 0, & s_r > 0, \\ R_t\,\big|_a = 10, R_t\,\big|_o = 0, R_t\,\big|_r = 0, & s_r = 0, s_a > \varepsilon, \\ R_t\,\big|_o = 10, R_t\,\big|_r = 0, R_t\,\big|_a = 0, & \text{others.} \end{cases} \tag{37}$$

From the structure of the reward function, it is shown that if the proportion of the sensing member, s_r, is positive, the system will receive a positive reward, which will stimulate the system to enhance the coefficient of force in the repulsing sensing area. Except that the condition where s_r is positive, the rewards of other sensing areas will be decided according to the states of s_a. When s_a is larger than ε, which is a positive separator, the reward about ZOA is positive; otherwise, the reward about ZOO is positive. It is mentioned that the rewards of the three areas cannot be calculated simultaneously. Otherwise, it may make the parameters enhanced simultaneously, which may make the learning invalidate. Therefore, the calculation process of the reward needs to be prioritized according to specific tasks. The whole diagram of learning logic is illustrated in Figure 7.

In Figure 7, there exists two critic parts and one actor part in each channel. The two critic parts are applied to estimate the value of current time, $V(t)$, and the value of next time, $V(t+1)$. According to $s_a(t)$, $s_o(t)$, and $s_r(t)$, the parameters, c_{ai}, c_{oi}, and c_{ri} are calculated. Bring these parameters into the motion controller, which is designed in equation (19), $s_r(t+1)$, $s_o(t+1)$, and $s_a(t+1)$ are obtained. Besides, the immediate reward, R_t, is also acquired from the environment. According to R_t, $V(t)$, and $V(t+1)$, the time difference, Δ_t, is calculated. The output parameters of the critic part and the actor part can be adjusted according to Δ_t.

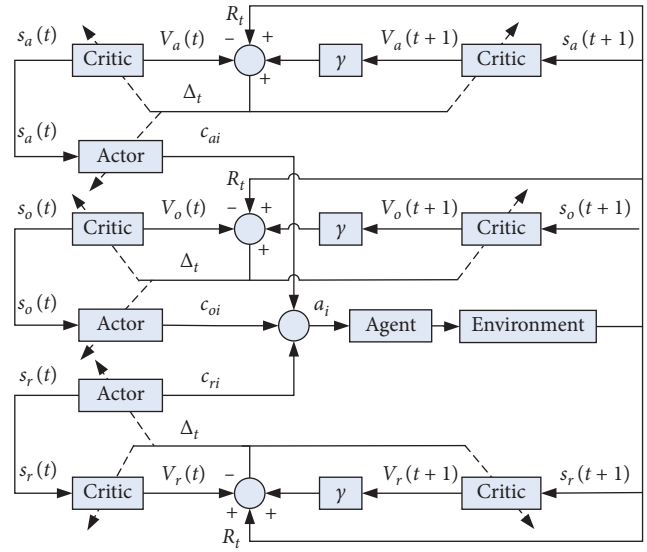

Figure 7: The diagram of learning logic.

To sum up, the learning algorithm of RLPST is shown as Algorithm 1.

5. Simulation

A cluster with four satellite members, which are numbered from No. 1 to No. 4, is selected as the numerical experimental object. It is supposed that the reference orbit is a circular orbit with the radius of 10^4 km. The symbols, $\mathbf{x}_{10}$, $\mathbf{x}_{20}$, $\mathbf{x}_{30}$ and $\mathbf{x}_{40}$, are denoted as the initial states of the satellites in the cluster from No. 1 to No. 4. The first three items of these vectors represent the relative position in m, while the last three items represent the relative velocity in m/s. In this numerical experiment, for each cluster member, the quadratic matrix of the position error ellipsoid is set as $\mathbf{A}$ and that of the induction ellipsoid is set as $\mathbf{M}$. Based on the reference in [33, 34, 37], the values of $\mathbf{A}$ and $\mathbf{M}$ are set as follows:

$$\begin{aligned} \mathbf{A} &= \begin{bmatrix} 215.41 & -84.43 & 56.29 \\ -84.43 & 312.91 & -97.50 \\ 56.29 & -97.50 & 231.66 \end{bmatrix}, \\ \mathbf{M} &= \begin{bmatrix} 3.11\mathrm{e}^{-7} & 2.19\mathrm{e}^{-8} & -4.65\mathrm{e}^{-9} \\ 2.18\mathrm{e}^{-8} & 2.78\mathrm{e}^{-7} & 3.63\mathrm{e}^{-8} \\ -4.65\mathrm{e}^{-9} & 3.63\mathrm{e}^{-8} & 2.92\mathrm{e}^{-7} \end{bmatrix}. \end{aligned} \tag{38}$$

Two classic scenarios, which include the scenario of adding members into the cluster and the scenario of members following a flight path, are considered, respectively. It is noticed that when we talk about adding members,

(1) **for all** cluster members **do**
(2) **for all** channel **do**
(3) Initialize the membership functions
(4) Initialize $V_q = 0$, $\phi_l^C|_q = 0$, $\phi_l^A|_q = 0$, for $l = 1, \ldots, L$;
(5) **end for**
(6) **end for**
(7) **for** each episode **do**
(8) **for all** cluster members **do**
(9) Initialize states of the cluster member
(10) **for all** Time step **do**
(11) Calculate the 3σ position error ellipsoid according to equation (7)
(12) Maintain all sensing neighbours of the cluster member
(13) Obtain the sensing sets, X_r, X_o, and X_a, respectively, based on the results of equation (11)-equation (12)
(14) Calculate the force direct p_i^r, p_i^o, and p_i^a according to equation (13), equation (15), and equation (16), respectively
(15) **for all** channel **do**
(16) Calculate the output of the actor c_q through equation (36)
(17) Calculate the output of the critic $V_q(t)$ from equation (35)
(18) Interact with the environment
(19) Obtain the reward R_t, and the output of the critic $V_q(t+1)$
(20) Calculate the time difference Δ_t from equation (24)
(21) Update $\phi_l^C|_q$ and $\phi_l^A|_q = 0$ according to equation (26) and equation (30), respectively
(22) **end for**
(23) **end for**
(24) **end for**
(25) **end for**

Algorithm 1: RLPST.

the combination of No. 1 and No. 2 is set as the original cluster, while No. 3 and No. 4 are the ones who want to add into. Besides, when the members are following a flight path, No. 1 is set as the leader satellite, which holds a desired path, and others are the ones who need to follow the path of No. 1. With the limits of $c_{ri} \in [-20, 20]$, $c_{oi} \in [-20, 20]$, and $c_{ai} \in [-20, 20]$, the experimental results are shown below.

5.1. Scenario of Adding Members into the Cluster. In this scenario, satellites No. 1 and No. 2, which are considered as the members of original cluster, are flying together with a proper distance. Satellites No. 3 and No. 4 aim to merge into the cluster from the different directions, respectively. The goal of the mission is to make the four members get into a new cluster with a proper distance from each other at the terminal time. A proper distance means that there is no neighbour located in the ZOR of each member in the cluster, and it makes the neighbours located in the ZOO as much as possible. The initial states of the cluster are listed in Table 1.

For representing the smoothness of the flight paths, the signal σ is defined to express the deviation degree from the whole flight path to the center baseline:

$$\sigma = \sqrt{\sum_{i=3}^{N} \int_{t=0}^{T} \left(\boldsymbol{\rho}_i(t) - \boldsymbol{\rho}_{\text{ref}}(t)\right)^2}, \tag{39}$$

where $\boldsymbol{\rho}_i$ and $\boldsymbol{\rho}_{\text{ref}}$ represent the position vector of the ith member and the corresponding position vector on the center baseline, respectively.

For the mission of adding members into an original cluster, it is appropriate to judge the terminal matching degree of new adding members. Therefore, the signal η_m, which is called the terminal matching rate, is defined to represent the degree of terminal status in ZOO:

$$\eta_m = \frac{\sum_{i=3}^{N} \text{Num}_i^m}{\sum_{i=3}^{N} (N-2)}, \tag{40}$$

where Num_i^m represents the number of neighbours in the ZOO of the ith member.

In order to express the effectiveness of improving the flight paths under the proposed RLPST, the signal Cost is defined to represent the quality of distances among new adding members, which is shown as

$$\begin{aligned} \text{Cost} = &\sum_{i=3, j \in X_r}^{N} \int_{t=0}^{T} \left(m_r \frac{1}{\left\| r_n(t) \right\|} \right) \mathrm{d}t \\ &+ \sum_{i=3, j \in X_a}^{N} \int_{t=0}^{T} \left(m_a \left\| r_n(t) \right\| \right) \mathrm{d}t, \end{aligned} \tag{41}$$

where m_r and m_a represent the corresponding coefficients of sets X_r and X_a, respectively.

k is denoted as the empirical parameter to substitute the value of c_r, c_o, and c_a in TEPM; then, the experimental results under the simulation time T with 1000 s are shown below.

From Figures 8(a) and 8(b), the trajectories of TEPM with $k = 3$ and $k = 6.5$ are illustrated, respectively. In Figure 8(a), it is seen that the terminal positions of the four members are relatively far away from each other, which does

Table 1: Initial states of the members in the cluster.

State	Value
$\mathbf{x}_{10}$	$[0; 100; 0; 2.678 \times 10^{-2}; -4.715 \times 10^{-5}; 0]^T$
$\mathbf{x}_{20}$	$[0; -100; 0; -2.678 \times 10^{-2}; -5.608 \times 10^{-3}; 0]^T$
$\mathbf{x}_{30}$	$[0; 1000; 0; 2.680 \times 10^{-2}; -4.715 \times 10^{-5}; 0]^T$
$\mathbf{x}_{40}$	$[-1000; -1000; 0; -2.678 \times 10^{-2}; -5.608 \times 10^{-3}; 0]^T$

not satisfy the requirement of the mission. This is because the empirical parameter is selected too small. On the contrary, from Figure 8(b), due to the large empirical parameter, it shows the nonsmooth trajectories of satellites No. 3 and No. 4. Although the requirement of terminal positions of the four satellites is guaranteed, the flying process will waste unnecessary fuels for the nonsmooth flight paths. Therefore, it is seen that we will easily get confused for the selection of empirical parameter under the TEPM. Whether the parameter is selected too large or too small, the flying effect cannot meet the goal of mission.

To compare with the results of TEPM, we set the discount factor γ as 0.8, the reward separator ε as 0.5, the learning rate of the critic α as 10^{-7}, the learning rate of the actor β as 10^{-8}, the coefficients in equation (41) m_r as 1000, and m_a as 10^{-3}. Thus, the results under the proposed RLPST are shown in Figures 9 and 10, where the results of members adding with different learning times are illustrated, respectively. From Figure 9(a), it is seen that the original cluster (which includes satellite No. 1 and satellite No. 2) keeps normally flying, and satellites No. 3 and No. 4 are far from the original cluster at the beginning. In addition, satellite No. 4 is moving nearly in the same direction as the original cluster, while satellite No. 3 is the opposite. Attracted by the original cluster, satellites No. 3 and No. 4 begin to move, where satellite No. 3 appears the tendency to change the moving direction, and satellite No. 4 keeps moving forward. In Figure 9(b), it can be seen that, with the increase in learning times, the direction changing is fixed, and the satellites No. 3 and No. 4 have basically determined the same flight direction as the original cluster. However, they have not integrated with each other yet and they are still attracted by the original cluster continuously. Figure 10 shows the finished training results after 55 times of learning. As the flight progresses, satellites No. 3 and No. 4 finally have merged into the original cluster to form a new cluster, and the mission of adding members is completed.

Figure 11 shows the trajectory deviations under the TEPM and the proposed RLPST, respectively. It is seen that the deviation under the TEPM is relatively large when the empirical parameter is set too large or too small. The result is reasonable because when the empirical parameter is too small, the whole flying condition cannot meet the terminal requirement of the mission, and when the empirical parameter is too large, the flight paths are nonsmooth, which may cause large trajectory deviation as well. When the empirical parameter is set to be an acceptable value, the trajectory deviation will meet the low point in the figure. However, compared with the proposed RLPST, the deviation under the RLPST is obviously lower than that under the TEPM, which means that the proposed RLPST has more smooth flight path which is a benefit for saving fuels and avoiding complex maneuvering strategies.

The terminal matching rate represents the final states of the cluster, and the ideal value is equal to one, which means that the adding member keeps a moderate distance with not only the original cluster members but also other adding members. From Figure 12, it is seen that the rate is different with different empirical parameters, which means that the matching rate cannot be guaranteed optimal under the TEPM. On the contrary, the solid line represents the rate under the RLPST, and it is clear that the rate is equal to one when the learning process is finished.

From Figure 13, the variation in the cost line along with learning times is illustrated. It is clearly seen that the cost is generally decreased with the increasing learning times. The figure indicates that the learning process has reduced the cost effectively, which means that the total flying condition is improved gradually during the process.

5.2. Scenario of Members following a Flight Path. In this scenario, satellite No. 1 is a leader, which has the desired flight path. Satellites No. 2 to No. 4 are expected to follow the path of the leader. The task requires that satellites No. 2, No. 3, and No. 4 can trace the leader effectively. The initial states of cluster members are shown in Table 2.

Similar to Section 5.1, to express the deviation degree from the whole flight path to the center baseline, the following definition is executed:

$$\sigma = \sqrt{\sum_{i=2}^{N} \int_{t=0}^{T} \left(\boldsymbol{\rho}_i(t) - \boldsymbol{\rho}_{\text{ref}}(t)\right)^2}, \tag{42}$$

where $\boldsymbol{\rho}_i$ and $\boldsymbol{\rho}_{\text{ref}}$ represent the position vector of the ith member and the corresponding position vector on the center baseline, respectively.

Besides, the signal *Cost* is also defined to represent the quality of distances among new adding members, which is shown as follows:

$$\begin{aligned} \text{Cost} = &\sum_{i=2, j\in X_r}^{N} \int_{t=0}^{T} \left(m_r \frac{1}{\left\| r_n(t) \right\|} \right) \mathrm{d}t \\ &+ \sum_{i=2, j\in X_a}^{N} \int_{t=0}^{T} \left(m_a \left\| r_n(t) \right\| \right) \mathrm{d}t. \end{aligned} \tag{43}$$

In addition, in the scenario of members following a leader, it will be reasonable to care about how many neighbours can each member sense. The more neighbours that a member can sense, the more information can the member obtain, which will be benefit for planning the flight paths. Therefore, the symbol η_s is defined as the sensing rate for representing the degree of sensing ability of the members in the cluster:

$$\eta_s = \frac{\sum_{i=2}^{N} \text{Num}_i^s}{\sum_{i=2}^{N} (N-1)}, \tag{44}$$

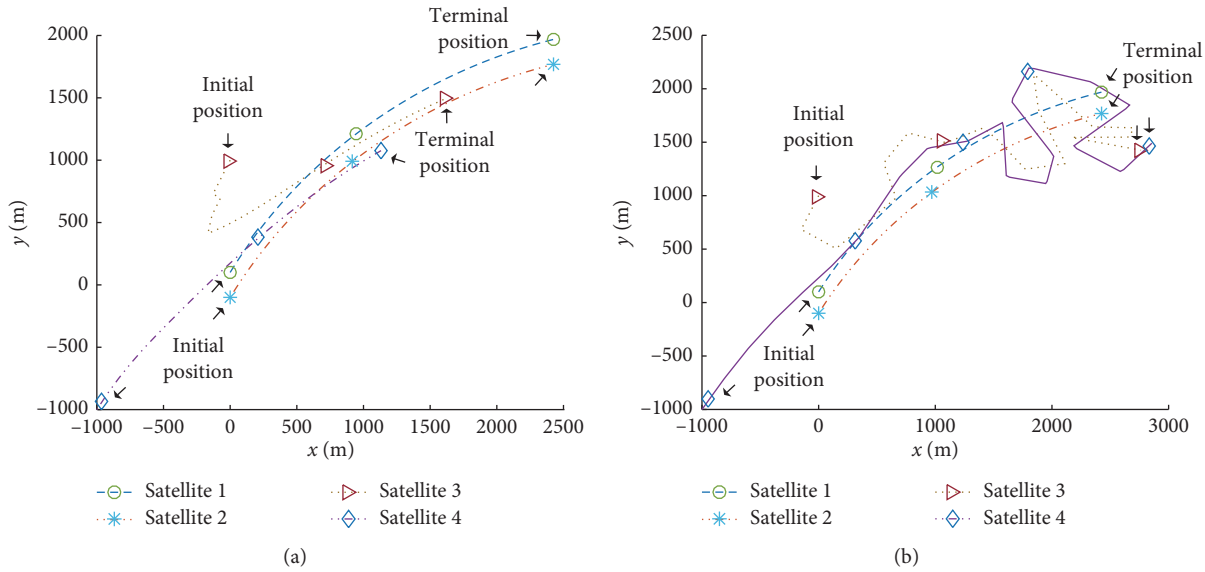

Figure 8: Trajectories of TEPM with different empirical parameters in the scenario of adding members: (a) $k = 3$; (b) $k = 6$.

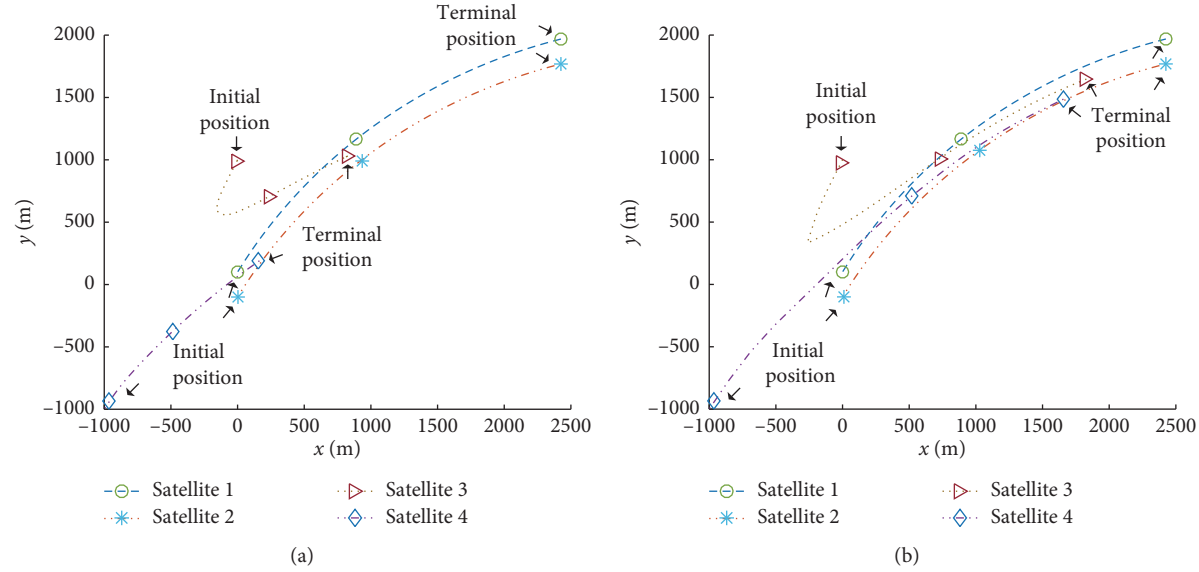

Figure 9: Results of the training different times in the scenario of members adding: (a) 2 times; (b) 30 times.

where Num_i^s represents the number of neighbours that the ith member can sense.

We set the discount factor γ as 0.8, the reward separator ε as 0.2, the learning rate of the critic α as 5×10^{-4}, the learning rate of the actor β as 10^{-4}, the coefficients m_r as 1000, and m_a as 10^{-3}. Thus, the experimental results under the simulation time T with 1000 s are shown below.

From Figures 14(a) and 14(b), the trajectories under TEPM with $k = 6$ and $k = 22$ are illustrated, respectively. In Figure 14(a), it is seen that satellites No. 2 to No. 4 fly aside from satellite No. 1 in the latter half of flight. The reason why their flight paths deviate from the leader is because that the empirical parameter is set too small that the members cannot sense the leader. It indicates that, in TEPM, if the empirical parameter is too small, some unexpected situations will occur which may result in the failure of the mission. In Figure 14(b), it is seen that the flight paths of the members are nonsmooth, which is not a good flight condition for following the leader. Certainly, this is because of the large value of the selected empirical parameter. Figures 14(a) and 14(b) have further explained the difficulties for selecting empirical parameter. The unsuitable selections may cause bad flying results or even lead to the failure of the mission. Besides, compared with the results from the Figures 14(a) and 14(b), the final training effect of members flight following under RLPST is illustrated in Figure 15. It is seen that satellite No. 1 is the leader in the cluster and satellites No. 2, No. 3 and No. 4 are scattered from each other at initial

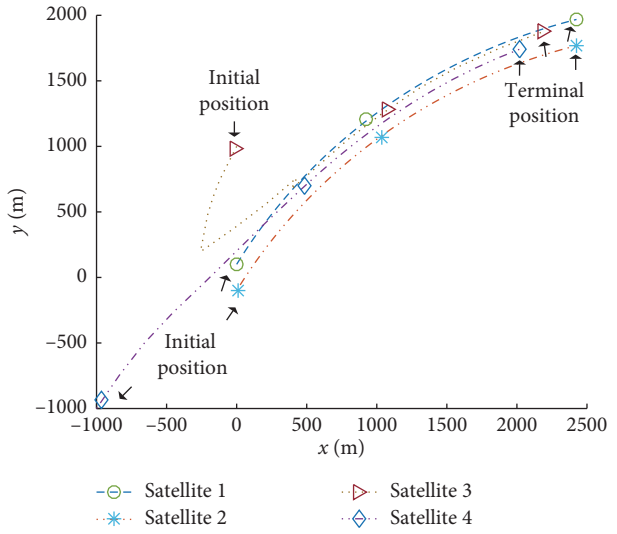

Figure 10: Results of the final training in the scenario of members adding.

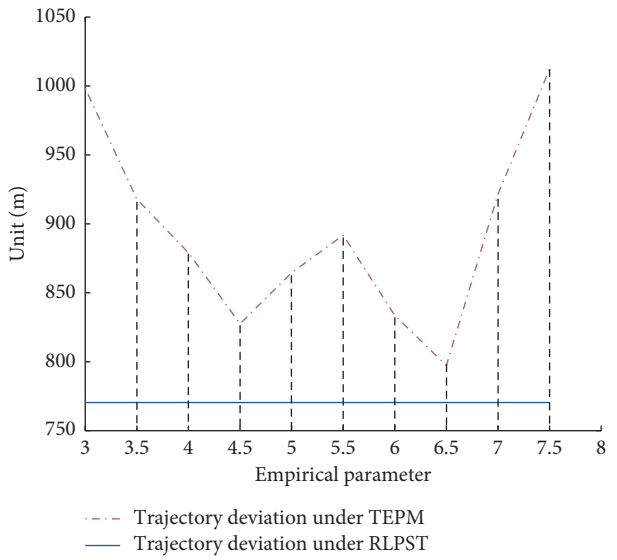

Figure 11: Curves of trajectory deviations under TEPM and RLPST in the scenario of adding members.

condition. When the training is finished, satellites No. 2 to No. 4 have the ability to follow the leader successfully and keep a smooth flight path.

From Figure 16, it is seen that when the empirical parameter is set from 12 to 16, the sensing rate is equal to one. In such a condition, all members in the cluster can sense other members during the whole flight. However, when the empirical parameter is too small or too large, the sensing rate will not be guaranteed to be one, which means that flight effect may be badly influenced. On the contrary, the solid line represents the sensing rate under the proposed RLPST, which is guaranteed to be one when the learning process is finished. The figure shows the superiority of the proposed RLPST because of the assurance of the optimal sensing rate.

Figure 17 shows the trajectory deviations under the TEPM and the proposed RLPST, respectively. Similar to the

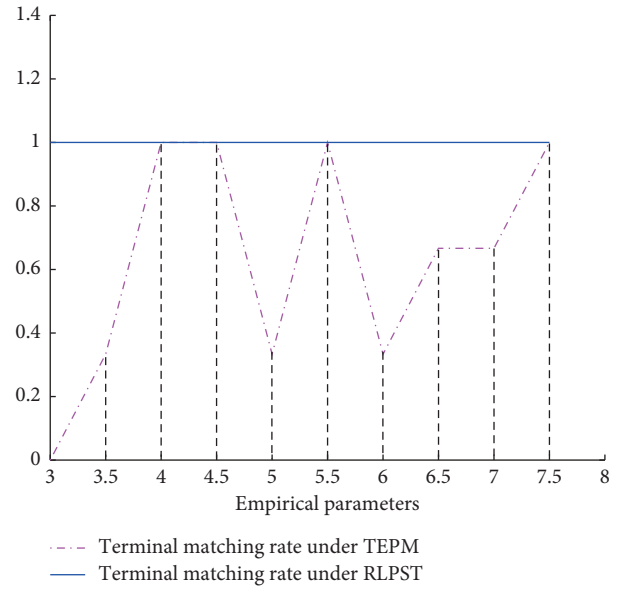

Figure 12: Curves of terminal matching rate under TEPM and RLPST in the scenario of adding members.

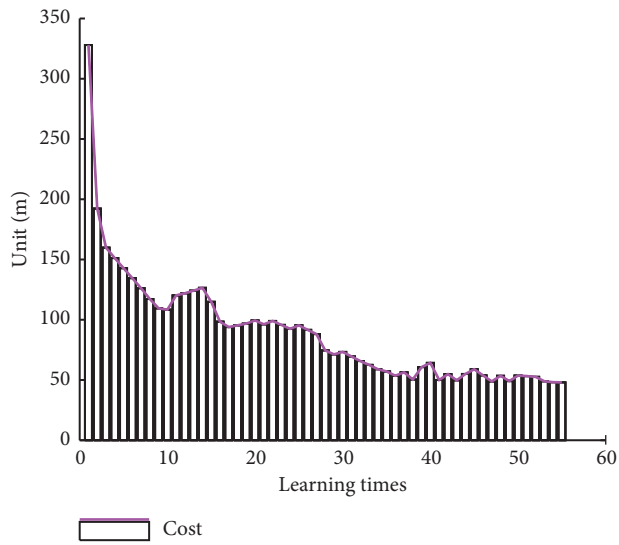

Figure 13: Curve of cost line along with learning times in the scenario of adding members.

Table 2: Initial states of the members in the cluster.

State	Value
$\mathbf{x}_{10}$	$[0; 0; 0; 0; 0; 0]^T$
$\mathbf{x}_{20}$	$[500; 500; 0; -2.678 \times 10^{-2}; -5.608 \times 10^{-3}; 0]^T$
$\mathbf{x}_{30}$	$[-500; 500; 0; 2.680 \times 10^{-2}; -4.715 \times 10^{-5}; 0]^T$
$\mathbf{x}_{40}$	$[-500; -500; 0; -2.678 \times 10^{-2}; -5.608 \times 10^{-3}; 0]^T$

condition shown in Figure 11, when the empirical parameter is too small or too large, the deviation is obviously high because of the badly flight results. In the figure, if the parameter is chosen as about 12, it has the lowest value of deviation. However, compared with the value under RLPST,

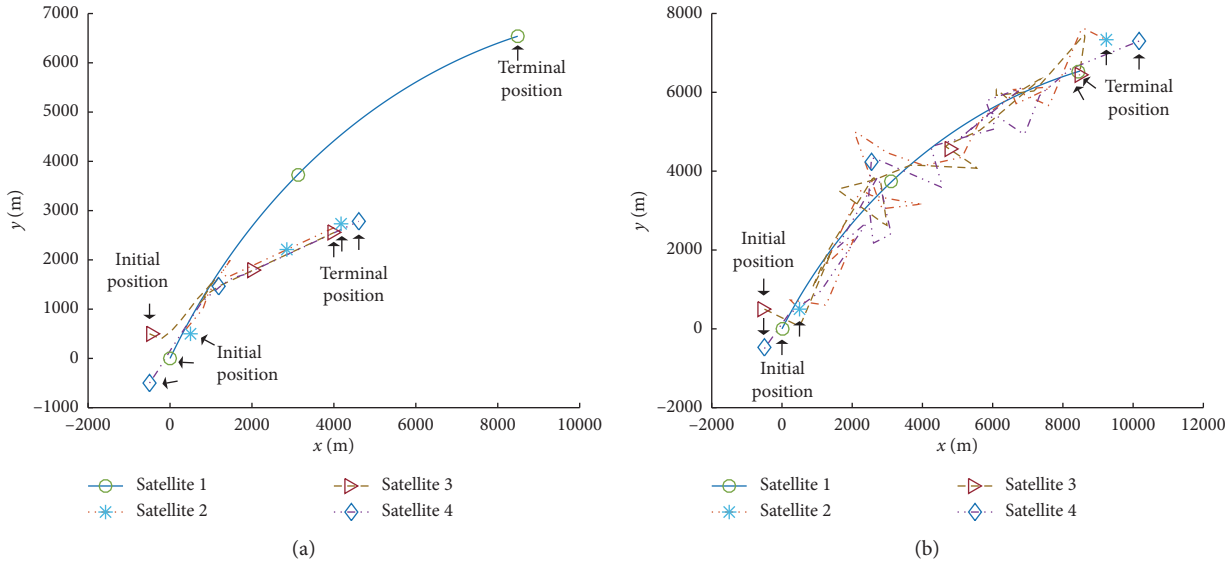

Figure 14: Trajectories of TEPM with different empirical parameters in the scenario of members following a flight path: (a) $k = 6$; (b) $k = 22$.

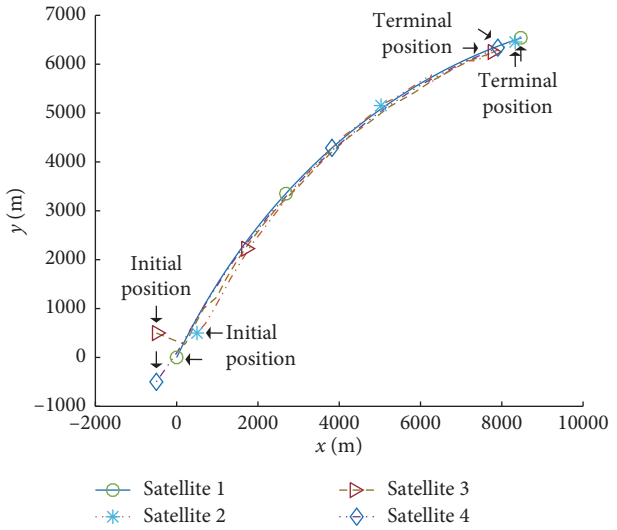

Figure 15: Results of final training in the scenario of members following a flight path.

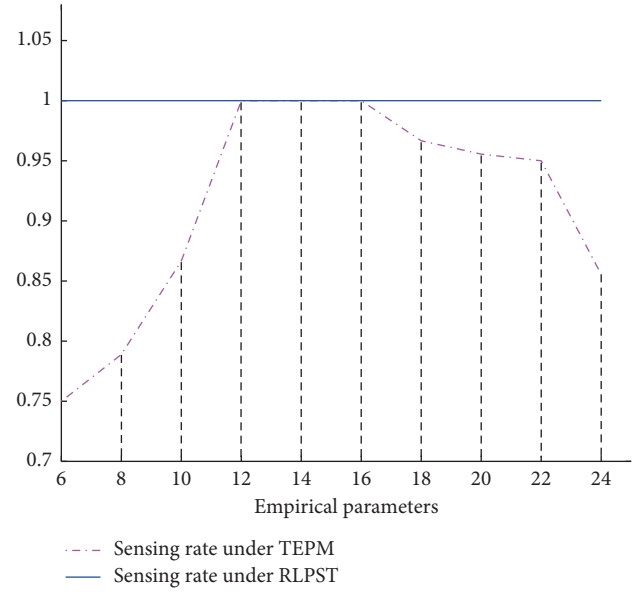

Figure 16: Curves of sensing rate under TEPM and RLPST in the scenario of members following a flight path.

it indicates that the value under RLPST is still smaller than the lowest value under TEPM. Therefore, the RLPST method has the ability to meet the more lower deviation within the safety flight range, which makes the flight path more smooth.

From Figure 18, the variation in cost along with the learning times is illustrated. Similar to the curves drawn in Figure 13, the cost is generally decreased with the increase in learning times. When the learning process is finished, the near lowest cost for the mission is found. The figure indicates that the flight condition is gradually improved during the learning process, and the cost of the flight can be effectively reduced through the proposed RLPST.

6. Discussion

In Section 5.1, we simulate the scenario of members adding under the TEPM and the proposed RLPST, respectively. The results show that it is difficult to select proper empirical parameters under TEPM. In addition, the trajectory deviations and terminal matching rates under TEPM and RLPST are compared. The trajectory deviation under RLPST is lower than that under TEPM. On the contrary, the value of the terminal matching rate under RLPST is guaranteed to equal to one, while that under TEPM cannot be. Therefore, the superiorities of the proposed RLPST are obviously proved. The variation in the cost along with the learning times shows the flight paths can be gradually improved through the learning framework.

In Section 5.2, the scenario of members flight path following under the TEPM and the proposed RLPST are simulated, respectively. Apart from the superiorities of

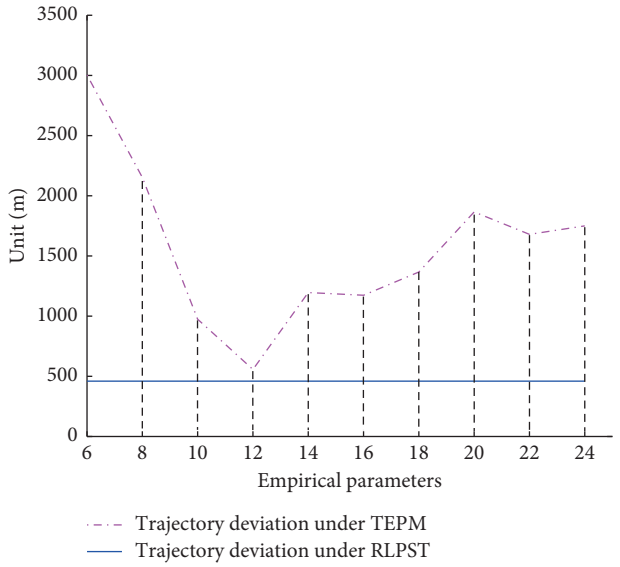

Figure 17: Curves of trajectory deviations under TEPM and RLPST in the scenario of members following a flight path.

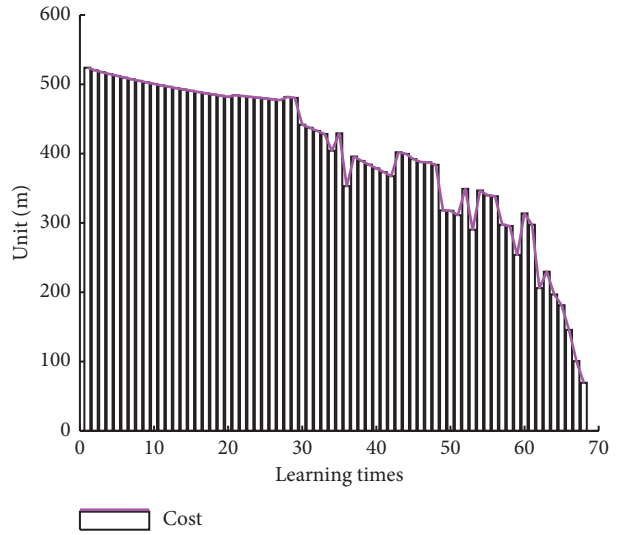

Figure 18: Curve of cost line along with learning times in the scenario of members following a flight path.

Table 3: Time cost and iteration times for the simulated scenarios.

Scenario	Time cost (unit: s)	Iteration times
Adding members into a satellite cluster	131.96	55
Members following a flight path	69.44	68

RLPST in the aspects of low trajectory deviation and decreasing cost, the sensing rate under RLPST shows the advantage compared with that under the TEPM, which means that RLPST method makes the cluster member be able to obtain more information from neighbours for completing the mission. A steep downward trend in Figure 18 is due to the selection of learning rates. At the final part of learning process, the performance of the cluster becomes sensitive to the learning rates, which is still a challenging problem.

The time cost and iteration times for the two simulated scenarios are listed in Table 3.

By comparing with the similar studies of [29, 30], it is seen that the time costs and iteration times of the two simulated scenarios are acceptable, which means the proposed RLPST can improve the flight path within a reasonable payment.

7. Conclusion

Due to the difficulties of parameter selection under TEPM for satellite cluster flying, a type of parameter-self-tuning method based on the actor-critic algorithm is proposed for handling the problem. Considering the specific condition of satellite cluster, the three sensing zones are redefined and the method for determining the belonging zones of sensing members of each cluster member is presented. To tune the flight parameter in each sensing zone, the fuzzy inference systems are employed to compose the actor and critic parts. With the proper design of reward function, a three-channel learning framework of parameter self-tuning for satellite cluster is designed. Compared with the TEPM, the proposed RLPST algorithm shows the superiorities. The results of simulation experiments indicate that the proposed RLPST has the lower trajectory deviation and guarantees the better terminal matching rate for scenario of members adding as well as the better sensing rate for scenario of members flight path following than the TEPM. Besides, the numerical experimental results also have shown the decrease in the cost along with the learning times in the two scenarios, which proves that the proposed RLPST has the ability to gradually improve the flight paths of the satellite cluster under the learning framework.

Appendix

A. The Affine Transformation Process

Recall the S_1 frame, which is centered at the nominal mass point of ith unit, is parallel to $Ox_r y_r z_r$. The position error ellipsoid of the ith unit is expressed as

$$\mathbf{R}^T \mathbf{A}_{rr}^i \mathbf{R} = 1. \tag{A.1}$$

Denote the symbol $\mathbf{X} = [x, y, z, 1]^T$; therefore, the error ellipsoid can be transformed as

$$\mathbf{X}^T \mathbf{B}_i^{S_1} \mathbf{X} = \mathbf{X}^T \begin{bmatrix} \mathbf{A}_{rr}^i & 0_{3\times 1} \\ 0_{1\times 3} & -1 \end{bmatrix} \mathbf{X} = 0, \tag{A.2}$$

where $\mathbf{B}_i^{S_1}$ is the ellipsoidal quadratic matrix. It is assumed that the jth unit is in the sensing area of the ith unit with the relative distance $\mathbf{r}_{ij} = \mathbf{r}_j - \mathbf{r}_i = [x_{ij}, y_{ij}, z_{ij}]^T$, and the position error ellipsoid of the jth unit in S_1 frame can be expressed as

$$\mathbf{X}^T \mathbf{B}_j^{S_1} \mathbf{X} = \mathbf{X}^T \left(\mathbf{T}_1^{-1}\right)^T \begin{bmatrix} \mathbf{A}_{rr}^j & 0_{3\times 1} \\ 0_{1\times 3} & -1 \end{bmatrix} \mathbf{T}_1^{-1} \mathbf{X} = 0, \tag{A.3}$$

where $\mathbf{T}_1$ is the translation matrix, which is expressed as

$$\mathbf{T}_1 = \begin{bmatrix} 1 & 0 & 0 & x_{ij} \\ 0 & 1 & 0 & y_{ij} \\ 0 & 0 & 1 & z_{ij} \\ 0 & 0 & 0 & 1 \end{bmatrix}. \tag{A.4}$$

Because $\mathbf{A}_{rr}^{i}$ is the positive definite symmetric, there exists an orthogonal matrix $\mathbf{Q}$ satisfying the following condition:

$$\mathbf{A}_{rr}^{i} = \mathbf{Q}^{\mathrm{T}}\mathbf{\Lambda}\mathbf{Q} = \begin{bmatrix} Q_{11} & Q_{21} & Q_{31} \\ Q_{12} & Q_{22} & Q_{32} \\ Q_{13} & Q_{23} & Q_{33} \end{bmatrix} \begin{bmatrix} 1/a^2 & 0 & 0 \\ 0 & 1/b^2 & 0 \\ 0 & 0 & 1/c^2 \end{bmatrix} \cdot \begin{bmatrix} Q_{11} & Q_{12} & Q_{13} \\ Q_{21} & Q_{22} & Q_{23} \\ Q_{31} & Q_{32} & Q_{33} \end{bmatrix}, \tag{A.5}$$

where $1/a$, $1/b$, and $1/c$ are the matrix eigenvalues of $\mathbf{A}_{rr}$. Besides, the following condition is satisfied:

$$a \le b \le c. \tag{A.6}$$

Therefore, we have

$$\mathbf{Q}\mathbf{A}_{rr}^{i}\mathbf{Q}^{\mathrm{T}} = \mathbf{\Lambda}. \tag{A.7}$$

Thus, the transformation matrix, $\mathbf{T}_2$, which is applied to align the axes, can be obtained:

$$\mathbf{T}_2 = \begin{bmatrix} Q_{11} & Q_{21} & Q_{31} & 0 \\ Q_{12} & Q_{22} & Q_{32} & 0 \\ Q_{13} & Q_{23} & Q_{33} & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix}. \tag{A.8}$$

Besides, the transformation matrix $\mathbf{T}_3$ is defined as follows:

$$\mathbf{T}_3 = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & b/ba & 0 & 0 \\ 0 & 0 & c/ca & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix}. \tag{A.9}$$

The S_2 frame is defined as the coordinate system for axis alignment of the ith unit. Therefore, the error ellipsoids of the ith unit and the jth unit in S_2 frame can be expressed as

$$\begin{aligned} \mathbf{X}^{\mathrm{T}}\mathbf{B}_i^{S_2}\mathbf{X} &= \mathbf{X}^{\mathrm{T}}\left(\mathbf{T}_3\right)^{\mathrm{T}}\left(\mathbf{T}_2^{-1}\right)^{\mathrm{T}}\mathbf{B}_i^{S_1}\mathbf{T}_2^{-1}\mathbf{T}_3\mathbf{X} = 0, \\ \mathbf{X}^{\mathrm{T}}\mathbf{B}_j^{S_2}\mathbf{X} &= \mathbf{X}^{\mathrm{T}}\left(\mathbf{T}_3\right)^{\mathrm{T}}\left(\mathbf{T}_2^{-1}\right)^{\mathrm{T}}\mathbf{B}_j^{S_1}\mathbf{T}_2^{-1}\mathbf{T}_3\mathbf{X} = 0. \end{aligned} \tag{A.10}$$

After the transformation process, the distance vector from the jth unit to the ith unit is expressed as

$$\begin{aligned} \widetilde{r}_{ij} &= \begin{bmatrix} 1 & 0 & 0 \\ 0 & b/ba & 0 \\ 0 & 0 & c/ca \end{bmatrix}^{-1} \begin{bmatrix} Q_{11} & Q_{21} & Q_{31} \\ Q_{12} & Q_{22} & Q_{32} \\ Q_{13} & Q_{23} & Q_{33} \end{bmatrix} \mathbf{r}_{ij} = \mathbf{T}, \mathbf{r}_{ij} \\ &= \left[\widetilde{x}_{ij}, \widetilde{y}_{ij}, \widetilde{z}_{ij}\right]^{\mathrm{T}}, \end{aligned} \tag{A.11}$$

where $\mathbf{T}_{\mathrm{tr}}$ is the matrix for distance transformation. For satisfying the condition which is applicable for the location judging algebraic criterion, the origin of the frame needs to be translated to the nominal mass point of the jth unit. The translation matrix is denoted as $\mathbf{T}_4$, which is expressed as follows:

$$\mathbf{T}_4 = \begin{bmatrix} 1 & 0 & 0 & -\widetilde{x}_{ij} \\ 0 & 1 & 0 & -\widetilde{y}_{ij} \\ 0 & 0 & 1 & -\widetilde{z}_{ij} \\ 0 & 0 & 0 & 1 \end{bmatrix}. \tag{A.12}$$

After the translation process, the frame comes to S_3, where the error ellipsoids of the ith unit and the jth unit are expressed as follows:

$$\begin{aligned} \mathbf{X}^{\mathrm{T}}\mathbf{B}_i^{S_3}\mathbf{X} &= \mathbf{X}^{\mathrm{T}}\left(\mathbf{T}_4^{-1}\right)^{\mathrm{T}}\mathbf{B}_i^{S_2}\mathbf{T}_4^{-1}\mathbf{X} = 0, \\ \mathbf{X}^{\mathrm{T}}\mathbf{B}_j^{S_3}\mathbf{X} &= \mathbf{X}^{\mathrm{T}}\left(\mathbf{T}_4^{-1}\right)^{\mathrm{T}}\mathbf{B}_j^{S_2}\mathbf{T}_4^{-1}\mathbf{X} = 0. \end{aligned} \tag{A.13}$$

For aligning the axes, the rotation matrix is denoted as $\mathbf{T}_5$. Thus, we have

$$\begin{aligned} \mathbf{X}^{\mathrm{T}}\mathbf{B}_i^{S_4}\mathbf{X} &= \mathbf{X}^{\mathrm{T}}\left(\mathbf{T}_5^{-1}\right)^{\mathrm{T}}\mathbf{B}_i^{S_3}\mathbf{T}_5^{-1}\mathbf{X} = (x-d_1)^2 + (y-d_2)^2 \\ &\quad + (z-d_2)^2 - 1 = 0, \\ \mathbf{X}^{\mathrm{T}}\mathbf{B}_j^{S_4}\mathbf{X} &= \mathbf{X}^{\mathrm{T}}\left(\mathbf{T}_5^{-1}\right)^{\mathrm{T}}\mathbf{B}_j^{S_3}\mathbf{T}_5^{-1}\mathbf{X} = \frac{x^2}{a^2} + \frac{y^2}{b^2} + \frac{z^2}{c^2} - 1 = 0, \end{aligned} \tag{A.14}$$

where the frame comes to S_4.

Data Availability

The data, such as number of cluster members: N, simulation time: T, initial state of No. 1 satellite: $\mathbf{x}_{10}$, initial state of No. 2 satellite: $\mathbf{x}_{20}$, initial state of No. 3 satellite: $\mathbf{x}_{30}$, initial state of No. 4 satellite: $\mathbf{x}_{40}$, discount factor: γ, separate factor: ε, learning rate: α, learning rate: β, factor for calculating *Cost*: m_r, factor for calculating *Cost*: m_a, reference orbit, matrix of error ellipsoid: $\mathbf{A}$, matrix of induction area: $\mathbf{M}$, deviation degree: σ, matching rate: η_m, sensing rate: η_s, degree of flight quality: *Cost*, time cost, iteration times, trajectories under TEPM, and trajectories under RLPST, used to support the findings of this study are included within the article.

Conflicts of Interest

The authors declare that they have no conflicts of interest.

Acknowledgments

This work was supported by the National Natural Science Foundation of China (11572019) and Shanghai Academy of Spaceflight Technology (SAST2019084).

References

[1] V. Gazi and B. Fidan, "Coordination and control of multi-agent dynamic systems: models and approaches," in *Swarm Robotics*, pp. 71–102, Springer, Berlin, Germany, 2007.

[2] O. Brown and E. Paul, "The value proposition for fractionated space architectures," *Sciences*, vol. 4, no. 5, p. 23, 2006.

[3] O. Brown and E. Paul, "Application of value-centric design to space architectures: the case of fractionated spacecraft," in *Proceedings of the AIAA SPACE 2008 Conference & Exposition*, San Diego, CA, USA, September 2008.

[4] C. Mathieu and A. Weigel, "Assessing the fractionated spacecraft concept," in *Proceedings of the AIAA SPACE 2008 Conference & Exposition*, San Diego, CA, USA, September 2006.

[5] M. Gregory O'Neill and A. L. Weigel, "Assessing fractionated spacecraft value propositions for earth imaging space missions," *Journal of Spacecraft and Rockets*, vol. 48, no. 6, pp. 974–986, 2011.

[6] K. Kholshevnikov and N. Vassiliev, "On the distance function between two keplerian elliptic orbits," *Celestial Mechanics & Dynamical Astronomy*, vol. 75, pp. 10–83, 1999.

[7] L. Mazal and P. Gurfil, "Closed-loop distance-keeping for long-term satellite cluster flight," *Acta Astronautica*, vol. 94, no. 1, pp. 73–82, 2014.

[8] J. Wang and S. Nakasuka, "Cluster flight orbit design method for fractionated spacecraft," *Aircraft Engineering & Aerospace Technology: An International Journal*, vol. 84, no. 5, pp. 330–343, 2012.

[9] Z. Dang, Z. Wang, and Y. Zhang, "Bounds on maximal and minimal distances for coplanar satellite relative motion under given initial conditions," *Aerospace Science & Technology*, vol. 46, pp. 204–209, 2015.

[10] W. Ren, R. Beard, and E. Atkins, "A survey of consensus problems in multiagent coordination," in *Proceedings of the American Control Conference*, pp. 1859–1864, Portland, OR, USA, June 2005.

[11] W. Ren, "Consensus strategies for cooperative control of vehicle formations," *Control Theory & Applications, IET*, vol. 1, no. 2, pp. 505–512, 04 2007.

[12] R. Olfati-Saber and R. Murray, "Consensus problems in networks of agents with switching topology and time-delays," *IEEE Transactions on Automatic Control*, vol. 49, no. 9, pp. 1520–1533, 09 2004.

[13] Q. Song, J. Cao, and W. Yu, "Second-order leader-following consensus of nonlinear multi-agent systems via pinning control," *Systems & Control Letters*, vol. 59, no. 9, pp. 553–562, 2010.

[14] J. Qin, W. X. Zheng, and H. Gao, "Consensus of multiple second-order vehicles with a time-varying reference signal under directed topology," *Automatica*, vol. 47, pp. 1983–1991, 2011.

[15] C. W. Reynolds, "Flocks, herds and schools: a distributed behavioral model," *ACM SIGGRAPH Computer Graphics*, vol. 21, no. 4, pp. 25–34, 1987.

[16] T. Vicsek, A. Czirok, E. Ben-Jacob, I. Cohen, and O. Sochet, "Novel type of phase transition in a system of self-driven particles," *Physical Review Letters*, vol. 75, no. 6, p. 12, 1995.

[17] I. D. Couzin, J. Krause, R. James, G. D. Ruxton, and N. R. Franks, "Collective memory and spatial sorting in animal groups," *Journal of Theoretical Biology*, vol. 218, no. 1, pp. 1–11, 2002.

[18] O.-S. Reza, "Flocking for multi-agent dynamic systems: algorithms and theory," *IEEE Transactions on Automatic Control*, vol. 51, no. 3, pp. 401–420, 2006.

[19] V. Gazi and K. M. Passino, "A class of attraction/repulsion functions for stable swarm aggregations," in *Proceedings of the 41st IEEE Conference on Decision and Control*, pp. 2842–284, Las Vegas, NV, USA, December 2002.

[20] X. Lei, M. Liu, W. Li, and P. Yang, "Distributed motion control algorithm for fission behavior of flocks," in *Proceedings of the 2012 IEEE International Conference on Robotics and Biomimetics (ROBIO)*, pp. 1621–1626, Guangzhou, China, December 2012.

[21] Z. Chen, H. Liao, and T. Chu, "Clustering in multi-agent swarms via medium-range interaction," *EPL (Europhysics Letters)*, vol. 96, Article ID 40015, 11 pages, 2011.

[22] D. Izzo and L. Pettazzi, "Autonomous and distributed motion planning for satellite swarm," *Journal of Guidance Control and Dynamics*, vol. 30, 2007.

[23] S. Nag and L. Summerer, "Behaviour based, autonomous and distributed scatter manoeuvres for satellite swarms," *Acta Astronautica*, vol. 82, no. 1, pp. 95–109, 2013.

[24] X. Liang, Q. Sun, Z. Yin, and Y. Wang, "A study of aviation swarm convoy and transportation mission," *Lecture Notes in Computer Science*, vol. 7929, pp. 368–375, Springer, Berlin, Germany, 2013.

[25] R. S. Sutton, A. G. Barto, and R. J. Williams, "Reinforcement learning is direct adaptive optimal control," *IEEE Control Systems*, vol. 12, no. 2, pp. 19–22, 1992.

[26] A. G. Barto, "Reinforcement learning," *A Bradford Book*, MIT Press, vol. 15, no. 7, pp. 665–685, Cambridge, MA, USA, 1998.

[27] K. Morihiro, T. Isokawa, H. Nishimura, and N. Matsui, "Emergence of flocking behavior based on reinforcement learning," in *Proceedings of the International Conference on Knowledge-Based Intelligent Information & Engineering Systems*, Bournemouth, UK, October 2006.

[28] D. Gu and E. Yang, "Fuzzy policy reinforcement learning in cooperative multi-robot systems," *Journal of Intelligent and Robotic Systems*, vol. 48, no. 1, pp. 7–22, 2007.

[29] C. S. Chen, Y. Hou, and Y. Ong, "A conceptual modeling of flocking-regulated multi-agent reinforcement learning," in *Proceedings of the International Joint Conference on Neural Networks*, pp. 5256–5262, Vancouver, Canada, July 2016.

[30] S.-M. Hung and S. Givigi, "A q-learning approach to flocking with UAVs in a stochastic environment," *IEEE Transactions on Cybernetics*, vol. 47, no. 1, 2016.

[31] M. D. Awheda and H. M. Schwartz, "A residual gradient fuzzy reinforcement learning algorithm for differential games," *International Journal of Fuzzy Systems*, vol. 19, no. 4, pp. 1058–1076, Aug. 2017.

[32] W. H. Clohessy and R. S. Wiltshire, "Terminal guidance system for satellite rendezvous," *Journal of the Aerospace Sciences*, vol. 27, no. 9, pp. 653–658, 1960.

[33] C. Wen, Y. Zhao, and P. Shi, "Precise determination of reachable domain for spacecraft with single impulse," *Journal*

of Guidance, Control, and Dynamics, vol. 37, pp. 1767–1779, 2014.

[34] X. Wang, C. Wen, Y. Zhao, and P. Shi, "Collision possibility detection in the safe corridor of a tumbling target," *Journal of Harbin Institute of Technology*, vol. 4, 2018.

[35] S. Ueda, T. Kasai, and H. Uematsu, "HTV rendezvous technique and GN&C design evaluation based on 1st flight on-orbit operation result," in *Proceedings of the AIAA/AAS Astrodynamics Specialist Conference*, Toronto, Canada, August 2010.

[36] W. Wang, J. Wang, and M.-S. Kim, "An algebraic condition for the separation of two ellipsoids," *Computer Aided Geometric Design*, vol. 18, pp. 531–539, 2001.

[37] H. Shi, Y. Zhao, and P. Shi, "Analysis of trajectory deviation for spacecraft relative motion in close-range," *Journal of Beijing University of Aeronautics and Astronautics*, vol. 43, no. 3, pp. 636–644, 2017.