sensors

MDPI

Communication

CACLA-Based Trajectory Tracking Guidance for RLV in Terminal Area Energy Management Phase

Xuejing Lan [1], Zhifeng Tan [1], Tao Zou [1,*] and Wenbiao Xu [2]

[1] School of Mechanical and Electrical Engineering, Guangzhou University, Guangzhou 510006, China; lanxj@gzhu.edu.cn (X.L.); tzf710647@163.com (Z.T.)
[2] Guangdong Province Institute of Metrology, Guangzhou 510450, China; xwb8911@scm.com.cn
* Correspondence: tzou@gzhu.edu.cn

Abstract: This paper focuses on the trajectory tracking guidance problem for the Terminal Area Energy Management (TAEM) phase of the Reusable Launch Vehicle (RLV). Considering the continuous state and action space of this guidance problem, the Continuous Actor–Critic Learning Automata (CACLA) is applied to construct the guidance strategy of RLV. Two three-layer neuron networks are used to model the critic and actor of CACLA, respectively. The weight vectors of the critic are updated by the model-free Temporal Difference (TD) learning algorithm, which is improved by eligibility trace and momentum factor. The weight vectors of the actor are updated based on the sign of TD error, and a Gauss exploration is carried out in the actor. Finally, a Monte Carlo simulation and a comparison simulation are performed to show the effectiveness of the CACLA-based guidance strategy.

Keywords: RLV; guidance; TAEM; CACLA; TD learning

Citation: Lan, X.; Tan, Z.; Zou, T.; Xu, W. CACLA-Based Trajectory Tracking Guidance for RLV in Terminal Area Energy Management Phase. *Sensors* **2021**, *21*, 5062. https://doi.org/10.3390/s21155062

Academic Editors: Charlie Yang and Giancarlo Ferrigno

Received: 6 June 2021
Accepted: 23 July 2021
Published: 26 July 2021

1. Introduction

Advanced Reusable Launch Vehicle (RLV) is a space vehicle that can transport people or payloads into a predetermined orbit and can be reused many times [1,2]. RLV highly integrates and develops aerospace technology and aeronautics technology. It is the inevitable trend of the development of the space transportation systems and has important military and civil values [3]. Therefore, many countries have researched RLV to reduce the cost of future space transportation [4,5].

The main return modes of RLV are parachute vertical descent scheme, thrust reversing vertical landing scheme, and gliding flight horizontal landing scheme. In this paper, RLV uses the gliding flight horizontal landing scheme, which has a long deceleration time, less overload, and a wider re-entry corridor. Because RLV has no engine thrust to re-fly during its return, it is necessary to strictly manage the remaining energy of RLV to ensure its safe horizontal landing. The return process of RLV includes the initial re-entry stage, the Terminal Area Energy Management (TAEM) stage and the automatic landing stage [6–8]. In the initial reentry phase, the atmosphere is thin, and the trajectory control ability of RLV is weak [9–11]. In the automatic landing phase, RLV is very close to the ground, and the adjustable range and remaining time of the trajectory are very limited [12,13]. Therefore, TAEM is the most important phase for the return landing mission [14,15].

The energy change of RLV in TAEM is closely related to the trajectory shape. Thus, the RLV must track the reference trajectory to ensure the safe flight and accurate landing [16,17]. However, the complex flight environment and mission requirements pose challenges to the TAEM guidance system [18–20]. In the research of guidance algorithms, the mature guidance method is to use a small perturbation approximation method or feedback linearization theory to obtain the linear model of RLV. Then, the guidance law is designed based on robust control theory or Linear Quadratic Regulator (LQR) to track the reference trajectory [21–23]. The performance of these methods has a tight relationship with the accuracy of RLV modeling.

At present, the research on such guidance methods is focused on reducing the impact of interference and uncertainty on the system. The guidance algorithm directly oriented to the nonlinear model of RLV includes sliding mode theory, fuzzy theory, and adaptive theory [24,25]. However, these methods still have some difficulties in engineering implementation and need to be further studied. On the other hand, under the influence of a complex flight environment, RLV may deviate from the preset flight trajectory seriously and can not return to the preset flight trajectory. Thus, the online autonomous reconstruction of reference trajectory is an effective way to improve the reliability of RLV [26–28]. Then, the guidance law based on the preset reference trajectory cannot apply to the tracking task of the newly reconstructed reference trajectory. Therefore, this paper intends to study an intelligent guidance technology to achieve the adaptive tracking of the reconstructed reference trajectory.

Reinforcement Learning (RL), as a kind of algorithm in the machine learning field, follows the idea of human learning through environmental feedback, with the aim to solve the guidance problem of RLV in the complex flight environment. To the best of our knowledge, the research on the combination of RL and traditional TAEM guidance is still rare. At present, traditional RL algorithms mainly solve problems with finite discrete action and state space [29–31]. However, many practical problems (such as the guidance problem discussed in this paper) have continuous state and action space, which makes learning a good strategy more complex [32–35]. Therefore, scholars have done much research on RL in the continuous domain [36–39]. The actor–critic algorithm is an effective method to deal with the problem of the "curse of dimension" based on the application of function approximation technology [40–42]. In addition, an improved actor–critic algorithm called Continuous Actor–Critic Learning Automata (CACLA) is developed and performs well in the scene with continuous action and state space, such as LTE-A cellular networks and robotic control tasks [43–45].

To improve the intelligence and adaptiveness of RLV, in this paper, we will use CACLA to construct a trajectory tracking guidance strategy for the TAEM phase of RLV. The Markov Decision Process (MDP) of the guidance problem is modeled based on the trajectory tracking errors and the guidance command increments. The critic and actor of CACLA are modeled by two three-layer neuron networks, respectively. The online weight learning process is realized by an improved model-free Temporal Difference (TD) learning algorithm. Then, the guidance commands of RLV are obtained based on a Gauss exploration in the actor. Compared with the existing research on guidance strategies, the main contributions of this paper are as follows:

(i) An intelligent trajectory tracking guidance strategy is proposed based on CACLA for RLV in terminal area energy management phase.
(ii) The guidance strategy is a data-based guidance method with the ability to learn online, with no need to know the accurate system model.
(iii) The guidance strategy has good adaptability and robustness, and can be used to track the reconstructed reference trajectory.

2. Problem Formulation

2.1. Dynamics of RLV

The RLV to be studied in this paper uses the gliding flight horizontal landing scheme. There is no engine thrust during the return process, and the gliding maneuvering return depends on the aerodynamic force generated by the movement of RLV in the atmosphere. This kind of return mode can make a proper orbit manoeuvre by controlling the direction of lift, thus creating good conditions for horizontal landing on the runway of the landing site. In this paper, it is assumed that the earth is flat and non-rotating in TAEM. Thus, the dynamic model of RLV is established as follows:

$$\dot{v} = -\frac{D}{m} - g\sin\gamma, \tag{1}$$

$$\dot{\gamma} = \frac{L\cos\sigma}{mv} - \frac{g\cos\gamma}{v}, \tag{2}$$

$$\dot{\chi} = \frac{L\sin\sigma}{mv\cos\gamma}, \tag{3}$$

$$\dot{h} = v\sin\gamma, \tag{4}$$

$$\dot{x} = v\cos\gamma\cos\chi, \tag{5}$$

$$\dot{y} = v\cos\gamma\sin\chi, \tag{6}$$

where the states are calculated based on a landing coordinate system. v, h, and γ represent the velocity, altitude, and flight path angle of RLV, respectively. x and y represent the longitudinal and lateral positions of RLV in the landing coordinate system. χ is the heading angle relative to the runway centerline. m represents the mass of vehicle, and g represents the gravitational acceleration. The bank angle of RLV is denoted by σ. In addition, L and D denote the aerodynamic lift force and drag force, respectively, as follows:

$$L = qSC_L, \tag{7}$$

$$D = qSC_D, \tag{8}$$

where S is the reference area of vehicle, and q is the dynamic pressure. C_L (C_D) denotes the aerodynamic lift coefficient (drag coefficient), which can be determined by the angle-of-attack α and the Mach number M with a two-dimensional table look-up.

The total state vector of RLV is concluded as

$$X = [v, \gamma, \chi, h, x, y]^T. \tag{9}$$

The nominal reference trajectory can be obtained by off-line trajectory planning or online trajectory planning algorithms. The tracking error vector is defined according to the current state vector X and the state vector of reference trajectory X_r:

$$\Delta X = X - X_r. \tag{10}$$

At the end of the TAEM, the RLV must achieve the desired Approach and Landing Interface (ALI) states X_{ALI} to ensure the safety of the automatic landing phase. In view of the trajectory tracking guidance method, the TAEM terminal states can meet the ALI constraints by accurate tracking the reference trajectory with the tracking error satisfies

$$|\Delta X| \leq \varepsilon_X, \tag{11}$$

where ε_X is the boundary of trajectory tracking error. Although there is no strict restriction on the terminal flight path angle, in order to ensure the high-precision tracking of the reference trajectory, the flight path angle still needs to be considered in the guidance strategy.

2.2. Markov Decision Processes

The trajectory tracking problem of RLV should first be modeled as a Markov Decision Process (MDP) to enable the work of CACLA . The state of MDP s is defined by the trajectory tracking errors:

$$s = \Delta X, \tag{12}$$

The action a is defined based on the guidance command increments:

$$a = [\Delta\alpha, \Delta\sigma]^T. \tag{13}$$

where $\Delta\alpha = \alpha - \alpha_r$ is the increment of current angle of attack α relative to the reference one α_r. $\Delta\sigma = \sigma - \sigma_r$ is the increment of current bank angle σ relative to the reference bank angle σ_r.

The immediate reward r is defined by

$$r = s^T M s + a^T G a, \tag{14}$$

where $M^T = M > 0$ and $G^T = G > 0$ are square matrices. The smaller the error between the current state and the reference state, the smaller the immediate reward. $V(s)$ is the state value function represents the expected accumulative total rewards from the state s. The task of trajectory tracking guidance is to solve an optimal control strategy to minimize the state value function $V(s)$.

3. CACLA-Based Guidance Strategy

3.1. CACLA Algorithm for Trajectory Tracking

CACLA is a reinforcement learning algorithm that can be effectively implemented in continuous state and action space. In this paper, we will use CACLA to learn a guidance law for RLV to track the reference trajectory. There are two modules in CACLA, named "critic" and "actor". In this work, two three-layer neuron networks (NN) are used to realize the function approximation for the critic and actor, respectively. In the critic, there are 6 input layer neurons, k hidden layer neurons, and 1 output layer neuron. In the actor, there are 6 input layer neurons, q hidden layer neurons, and 2 output layer neurons. Then, the state value function is approximated as

$$V(s) = \theta_2 \cdot \phi(\theta_1 s), \tag{15}$$

where $\theta_1 \in \mathbb{R}^{k\times 6}$ and $\theta_2 \in \mathbb{R}^{1\times k}$ are the weight vectors of NN in critic. $\phi(\cdot)$ is the basic function defined as

$$\phi(z) = \frac{1-\exp(-z)}{1+\exp(-z)}. \tag{16}$$

Moreover, the action function is approximated as

$$A(s) = \psi_2 \cdot \phi(\psi_1 s), \tag{17}$$

where $\psi_1 \in \mathbb{R}^{q\times 6}$ and $\psi_2 \in \mathbb{R}^{2\times q}$ are the weight vectors of NN in actor. To enable the exploration in CACLA, a Gaussian distribution policy $P(s,a)$ centered on $A(s)$ is defined:

$$P(s,a) = \frac{1}{\sqrt{2\pi}\mu}\exp(-\frac{(a-A(s))^2}{2\mu^2}), \tag{18}$$

where π and μ are constant parameters. Thus, the action a is achieved according to this Gauss exploration.

From the definition in (13), we can further obtain the actual guidance commands α and σ based on the known reference commands:

$$[\alpha, \sigma]^T = a + [\alpha_r, \sigma_r]^T, \tag{19}$$

When the guidance commands α and σ are applied to the dynamic model of RLV, the next total state vector of RLV can be obtained. Compared to the states and commands of the given reference trajectory, the next state of MDP $s(t+1)$ and the immediate reward $r(t)$ can be obtained. Then, the model free TD learning algorithm is used to update the weight vectors of critic.

$$\theta_1(t+1) = \theta_1(t) + \Delta\theta_1(t+1), \tag{20}$$

$$\theta_2(t+1) = \theta_2(t) + \Delta\theta_2(t+1), \tag{21}$$

$$\Delta\theta_1(t+1) = \eta\Delta\theta_1(t) + \varsigma(1-\eta)e_1(t+1)\delta(t), \tag{22}$$

$$\Delta\theta_2(t+1) = \eta\Delta\theta_2(t) + \varsigma(1-\eta)e_2(t+1)\delta(t), \tag{23}$$

where η is the momentum factor, and ς is the learning rate. $e_1 \in \mathbb{R}^{k\times 6}$ and $e_2 \in \mathbb{R}^{1\times k}$ are eligibility traces updated as follows:

$$e_1(t+1) = \lambda\tau e_1(t) + \nabla_{\theta_1} V_t(s(t)), \tag{24}$$

$$e_2(t+1) = \lambda\tau e_2(t) + \nabla_{\theta_2} V_t(s(t)), \tag{25}$$

where τ is the discount factor, and λ is a trace decay parameter. $\delta(t)$ is the TD error defined as

$$\delta(t) = r(t) + \tau V_t(s(t+1)) - V_t(s(t)). \tag{26}$$

If the TD error $\delta(t) > 0$, the weight vector of actor will be updated by

$$\psi_1(t+1) = \psi_1(t) + \beta(a(t) - A_t(s(t))\nabla_{\psi_1} A_t(s(t)), \tag{27}$$

$$\psi_2(t+1) = \psi_2(t) + \beta(a(t) - A_t(s(t))\nabla_{\psi_2} A_t(s(t)), \tag{28}$$

where β is a learning rate. $a(t)$ is the explored action, and $A_t(s(t))$ is the output of the actor.

It can be seen that the actor update process is performed only when the TD error is positive. Therefore, the actor of CACLA is updated based on the sign of TD error, not on the value of TD error as other actor–critic learning methods do. Moreover, another difference from most other actor–critic learning methods is that CACLA directly update the actor by the error in the action space, not the error in the policy space.

3.2. Application of Guidance Strategy

Due to the high cost of RLV, it is necessary to train the critic and actor offline before the guidance strategy is implemented in RLV. The flowchart of the off-line training of CACLA is shown in Figure 1. The updating procedure of critic and actor is performed at each step of TAEM, and is continued until RLV reaches the terminal ALI. At the end of TAEM if RLV has not met the ALI constraints, the TAEM guidance process will be performed again from the start of TAEM with adjusted initial parameters or structure of NN. For example, the latest updated weight vectors of the critic and the actor can be used as the initial values to achieve better guidance accuracy. When the off-line training of critic and actor achieves the required guidance accuracy, the weight vectors of critic and actor can be saved and be used in practical guidance missions.

The online learning of CACLA is effective to ensure the adaptive tracking of the reference trajectory, which may be reconstructed online to improve the reliability of RLV. On the other hand, even if the reference trajectory is not reconstructed, the online learning of CACLA is also necessary and helpful to improve the intelligence level of the guidance system to cope with the impact of the complex environment. Figure 2 shows the framework of the CACLA-based guidance system. The online learning procedure of guidance strategy is the same as the offline training. The weight vectors of critic are updated by (20)–(23). The weight vectors of actor are updated by (27) and (28), but only when the TD error is positive. Based on the output of the actor and the reference commands, the actual guidance commands are obtained by (19) and applied to the dynamic model of RLV. In order to simulate the actual flight TAEM environment of RLV, uncertainties and disturbances are added to the flight process. In this paper, it is assumed that the guidance system can accurately know the states of RLV through sensors or observers. Therefore, the research on state perception error or observation error will not be discussed in detail.

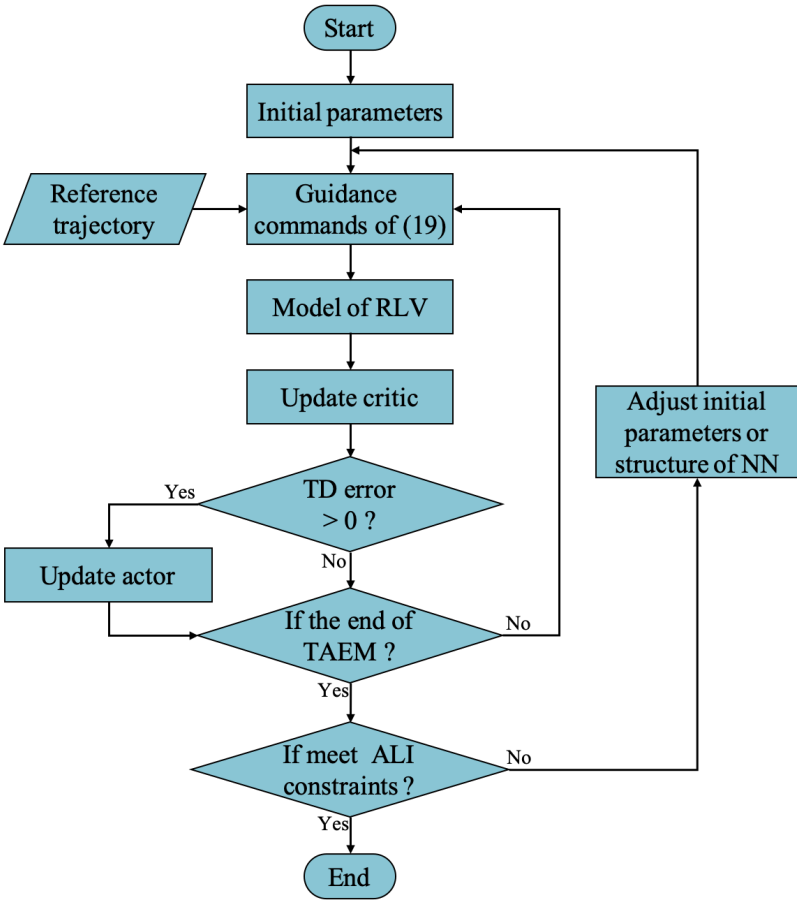

Figure 1. The flowchart of the off-line training of CACLA.

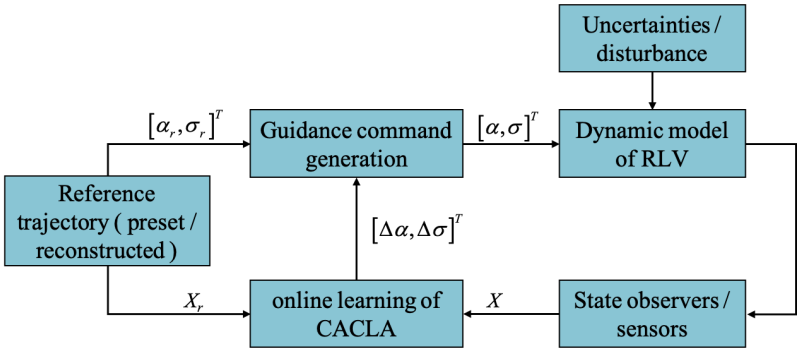

Figure 2. The framework of guidance system.

4. Simulation Results

In this section, a Monte Carlo simulation and a comparison simulation are performed to evaluate the effectiveness of the proposed intelligent guidance strategy. In the TAEM phase, the velocity and altitude are initialized as $v = 900$ m/s and $h = 28$ km. The lateral ground track position is initialized as $x = -10$ km and $y = -50$ km. The flight path angle is initialized as $\gamma = -8$ deg, and the initial heading angle is set towards the ALI. The desired ALI conditions are defined as $v < 180$ m/s, $h = 3 \pm 0.1$ km, $x = -21 \pm 0.3$ km, $y = 0 \pm 0.1$ km, and $\chi = 0 \pm 5$ deg. Although there is no strict restriction on the terminal flight path angle, to ensure the high precision tracking of the reference trajectory, the flight path angle still needs to be considered in the guidance strategy. Thus, the boundaries of trajectory tracking errors for each state of RLV are set as $\varepsilon_v = 100$ m/s, $\varepsilon_\gamma = 5$ deg, $\varepsilon_\chi = 5$ deg, $\varepsilon_h = 0.1$ km, $\varepsilon_x = 0.3$ km, and $\varepsilon_y = 0.1$ km. In addition, the guidance commands are subject to the bank-angle rate limit of 10 deg/s and the angle-of-attack rate

limit of 10 deg/s in the simulation. The parameters of CACLA are set as $\eta = 0.2$, $\varsigma = 0.4$, $\tau = 0.2$, $\lambda = 0.1$, $\beta = 0.1$, $k = 10$, and $q = 8$. The initial $\Delta\theta_1(0)$, $\Delta\theta_2(0)$, $e_1(0)$, and $e_2(0)$ are set as zero.

4.1. Monte Carlo Simulation

In the actual flight process, the deviations of the aerodynamic model and atmospheric density model inevitably exist due to modeling uncertainties or unknown disturbances. In order to evaluate the performance of the CACLA-based guidance strategy, a Monte Carlo simulation is performed with a variety of aerodynamic coefficient deviations and atmospheric density deviations that are subject to a Gaussian distribution given in Table 1. The reference trajectory is planned by the trajectory planning algorithm of [27] in an ideal environment.

Table 1. Model deviations for Monte Carlo simulations.

Parameter	Mean	Three Standard Deviations
Atmospheric density	0.0	15%
Aerodynamic lift coefficient	0.0	15%
Aerodynamic drag coefficient	0.0	15%

The RLV states of the reference trajectory and the 100 guidance trajectories are shown together in Figures 3–6, where the dashed red line represents the reference trajectory, and the solid black lines represent the guidance trajectories. Figure 3 shows the three-dimensional TAEM trajectories of RLV in Monte Carlo simulation. The three views of the TAEM trajectories and the velocity profiles with respect to altitude in Monte Carlo simulation are depicted in Figure 4. Because the trajectory propagation simulation is terminated at the desired altitude $h = 3$ km, the RLV can meet the terminal altitude constraints at ALI. The terminal velocities are all less than 180 m/s, meeting the requirements. The terminal errors of longitudinal position are within 0.3 km, and the terminal errors of lateral position are within 0.1 km. The flight path angle and heading angle profiles with respect to time in Monte Carlo simulation are shown in Figure 5. There is no strict requirement for terminal flight path angle in TAEM, thus angle errors are allowable. The terminal errors of heading angle are within 0.5 deg. Therefore, all the TAEM guidance trajectories of RLV meet the requirements of tracking accuracy. In addition, the angle of attack and bank angle profiles with respect to time in Monte Carlo simulation are illustrated in Figure 6. It can be seen that the guidance commands have been adjusted online to cope with the uncertainties and disturbances in the actual flight environment. The detailed terminal conditions of the 100 guidance trajectories are presented in Table 2, meeting all the terminal constraints. These Monte Carlo simulation results validate the effectiveness of the proposed intelligent guidance strategy.

Table 2. TAEM terminal conditions in Monte Carlo simulation.

Conditions	Maximum	Minimum	Mean	Variance	Desired Value
V_f (m/s)	178.8269	159.3518	167.9724	4.654	<180
x_f (km)	-20.7174	-21.2804	-20.9724	0.15224	-21 ± 0.3
y_f (km)	0.0934	-0.0862	0.0043	0.0479	0 ± 0.1
χ_f (deg)	0.1533	-0.0237	0.0638	0.0365	0 ± 5

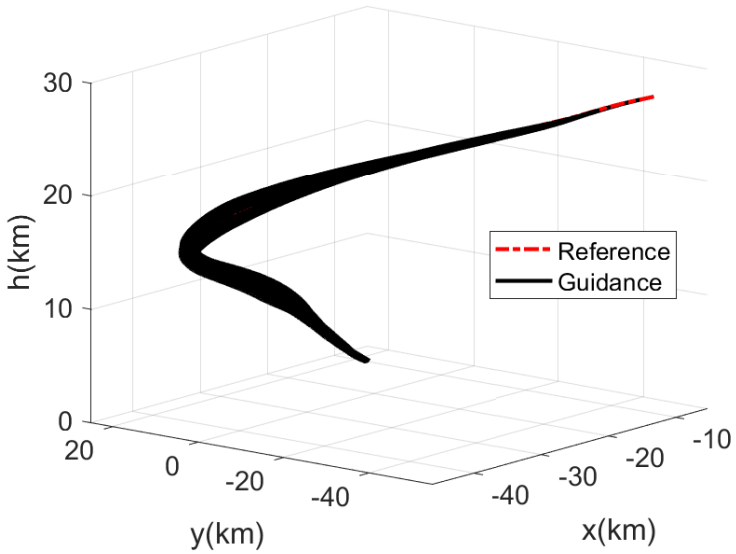

Figure 3. Three-dimensional TAEM trajectories of RLV in Monte Carlo simulation.

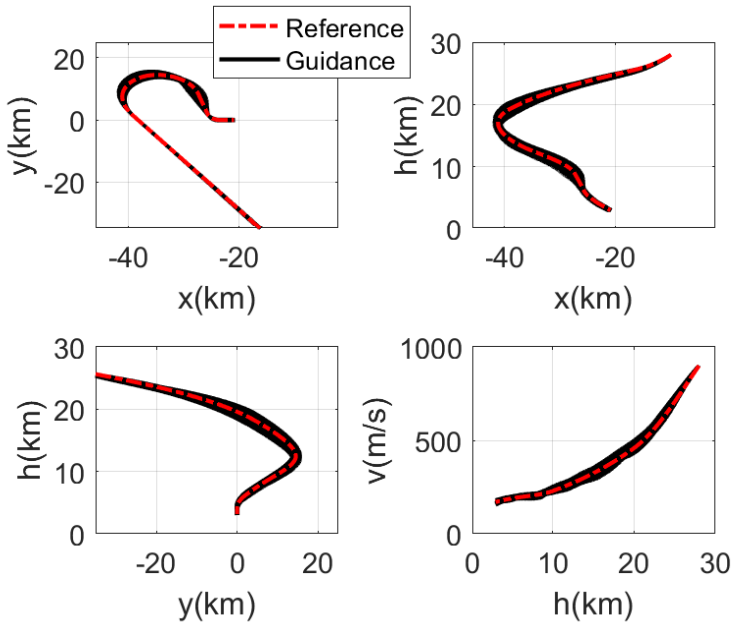

Figure 4. Three views of the TAEM trajectories and the velocity profiles with respect to altitude in Monte Carlo simulation.

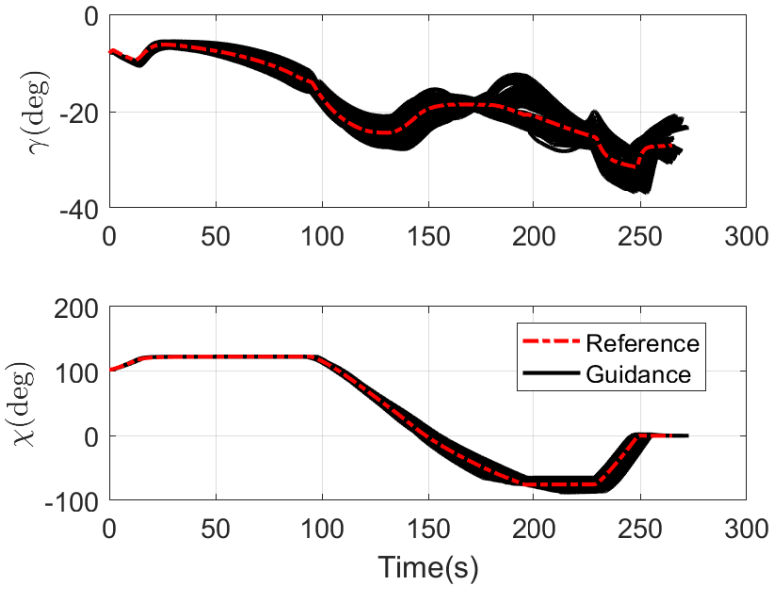

Figure 5. Flight path angle and heading angle profiles with respect to time in Monte Carlo simulation.

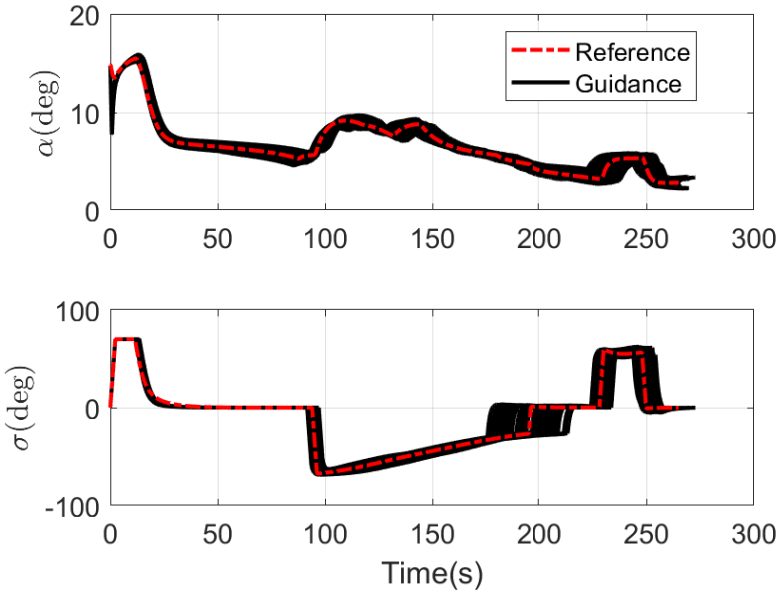

Figure 6. Angle of attack and bank angle profiles with respect to time in Monte Carlo simulation.

4.2. Comparison Simulation

Under the influence of complex flight environment, RLV may deviate from the preset flight trajectory seriously and cannot return to the preset flight trajectory. To demonstrate the adaptability of the CACLA-based guidance strategy, deviations in initial conditions are applied to the RLV, where $v = 950$ m/s, $h = 28.5$ km, $x = -5$ km, $y = -30$ km. Then, a new reference trajectory is reconstructed by the trajectory planning algorithm in [27]. The parameters of CACLA are not changed, and a PID guidance law based on the preset reference trajectory is designed for comparison. In the guidance environment, the random aerodynamic coefficient deviation and atmospheric density deviation given in Table 1 are performed.

The RLV states of the reconstructed reference trajectory, the guidance trajectory by using CACLA-based guidance law and the guidance trajectory by using PID guidance law are shown together in Figures 7–10. Figure 7 shows the three-dimensional TAEM trajectories of RLV in comparison simulation. The three views of the TAEM trajectories and the velocity profiles with respect to altitude in comparison simulation are depicted

in Figure 8. Because the trajectory propagation simulation is terminated at the desired altitude $h = 3$ km, the RLV can meet the terminal altitude constraints at ALI. The terminal velocities are less than 180 m/s, meeting the requirements. The terminal errors of lateral position are within 0.1 km. The terminal error of longitudinal position of CACLA guidance trajectory is within 0.3 km, meeting the requirements. However, the terminal longitudinal position of PID guidance trajectory is -21.5007, which does not meet the terminal error requirements. The flight path angle and heading angle profiles with respect to time in comparison simulation are shown in Figure 9. Although there is no strict requirement for terminal flight path angle in TAEM, the terminal flight path angle of CACLA guidance trajectory is closer to that of reconstructed reference trajectory than that of PID guidance trajectory. The terminal errors of heading angle are within 0.5 deg. In addition, the angle of attack and bank angle profiles with respect to time in comparison simulation are illustrated in Figure 10. The detailed terminal conditions of the reconstructed reference trajectory, the CACLA guidance trajectory and the PID guidance trajectory are presented in Table 3. It can be seen that the PID guidance law is inappropriate in the tracking task of the newly reconstructed reference trajectory. However, the CACLA-based guidance law can meet all the terminal constraints. Therefore, this comparison simulation results illustrate the advantages of the proposed CACLA-based guidance strategy.

Table 3. TAEM terminal conditions in comparison simulation.

Conditions	Reference	CACLA Guidance	PID Guidance	Desired Value
V_f (m/s)	169.9194	171.6748	168.3108	<180
x_f (km)	-20.8309	-20.8800	-21.5007	-21 ± 0.3
y_f (km)	0.0148	-0.0863	0.0712	0 ± 0.1
χ_f (deg)	0.0075	0.0828	-0.0414	0 ± 5

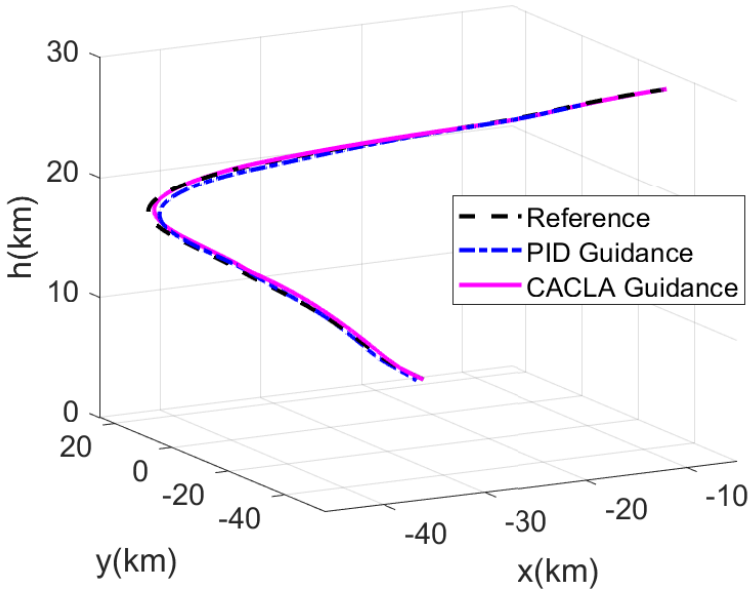

Figure 7. Three-dimensional TAEM trajectories of RLV in comparison simulation.

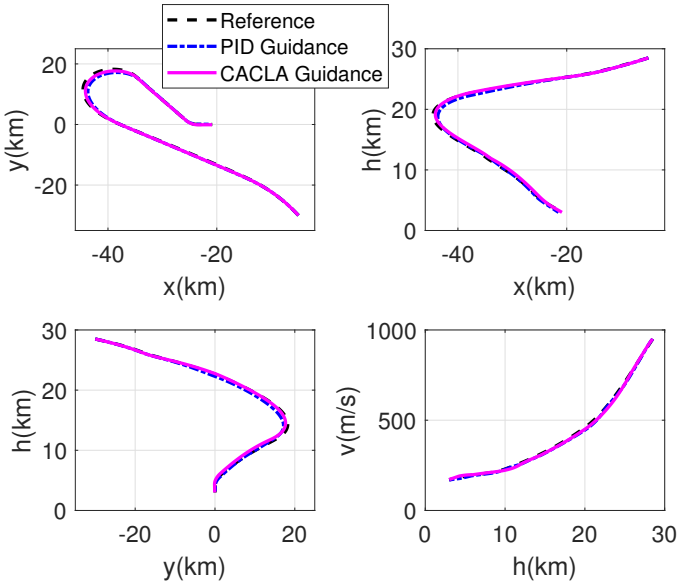

Figure 8. Three views of the TAEM trajectories and the velocity profiles with respect to altitude in comparison simulation.

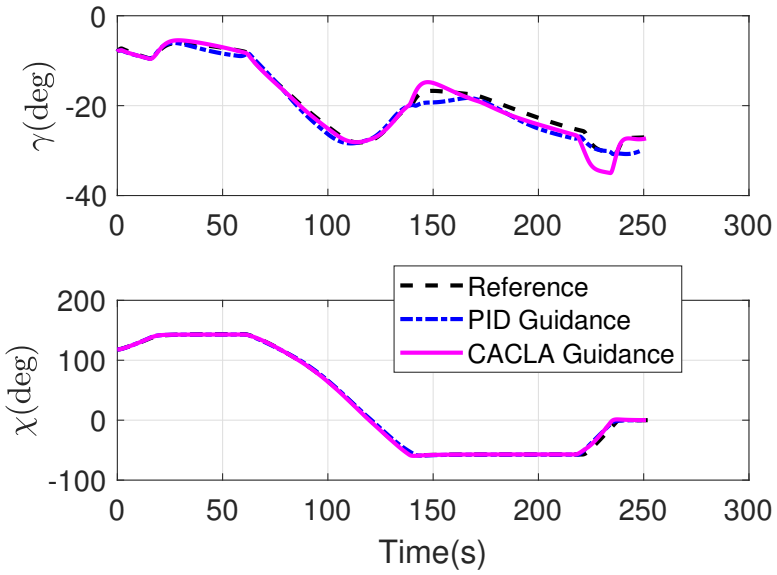

Figure 9. Flight path angle and heading angle profiles with respect to time in comparison simulation.

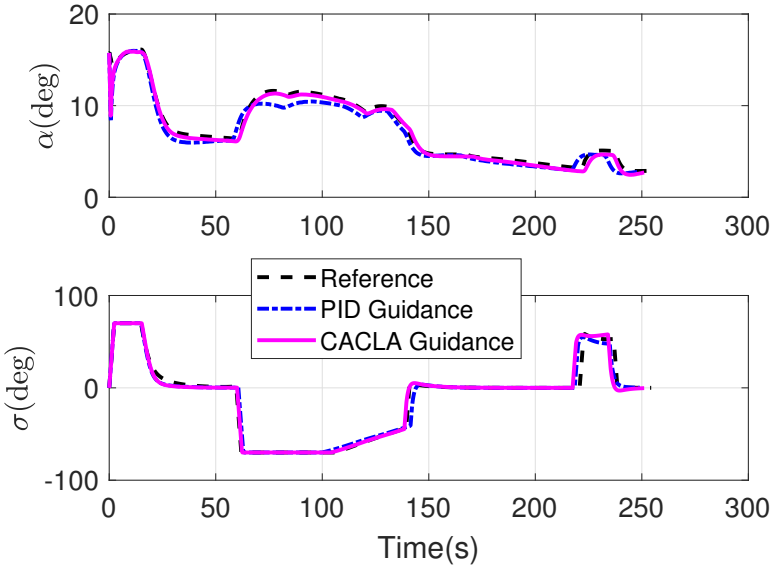

Figure 10. Angle of attack and bank angle profiles with respect to time in comparison simulation.

5. Conclusions

This paper proposed an intelligent trajectory tracking guidance strategy for the TAEM phase of RLV. A reinforcement learning algorithm CACLA is applied to construct the guidance strategy of RLV, which has continuous state and action space. Two three-layer neuron networks are used to realize the function approximation for critic and actor, respectively. Then, an improved model-free TD learning algorithm is used in the weight updating process. A Gauss exploration is carried out to obtain the guidance commands of RLV. Finally, the Monte Carlo simulation and the comparison simulation have performed to show that the proposed guidance strategy can achieve the high-precision tracking of the TAEM reference trajectory with all ALI conditions satisfied. In addition, the CACLA-based guidance strategy is universal, and thus can be used not only in TAEM, but also in the initial re-entry phase and the automatic landing phase.

Author Contributions: Data curation, Z.T. and W.X.; Methodology, T.Z.; Software, X.L.; Validation, X.L., Z.T. and T.Z.; Writing—Original draft, X.L.; Writing—Review and editing, T.Z. and W.X. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported by National Natural Science Foundation of China (Grant no. 61803111) and Guangzhou Science and Technology Project (Grant no. 202102010403).

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Joshi, A.; Sivan, K. Reentry Guidance for Generic RLV Using Optimal Perturbations and Error Weights. In Proceedings of the AIAA Guidance, Navigation, and Control Conference and Exhibit, Honolulu, HI, USA, 18–21 August 2012; American Institute of Aeronautics and Astronautics: Reston, VA, USA, 2012; pp. 1979–2014.
2. Liang, Z.; Hongbo, Z.; Wei, Z. A three-dimensional predictor–corrector entry guidance based on reduced-order motion equations. *Aerosp. Sci. Technol.* **2018**, *73*, 223–231.
3. Hilton, S.; Sabatini, R.; Gardi, A.; Ogawa, H.; Teofilatto, P. Space traffic management: Towards safe and unsegregated space transport operations. *Prog. Aerosp. Sci.* **2019**, *105*, 98–125.
4. Tomatis, C.; Bouaziz, L.; Franck, T.; Kauffmann, J. RLV candidates for European Future Launchers Preparatory Programme. *Acta Astronaut.* **2009**, *65*, 40–46.
5. Hanson, J. A Plan for Advanced Guidance and Control Technology for 2nd Generation Reusable Launch Vehicles. In Proceedings of the AIAA Guidance, Navigation, and Control Conference and Exhibit, Honolulu, HI, USA, 18–21 August 2012; American Institute of Aeronautics and Astronautics: Reston, VA, USA, 2012; pp. h1979–h1989.
6. He, R.; Liu, L.; Tang, G.; Bao, W. Entry trajectory generation without reversal of bank angle. *Aerosp. Sci. Technol.* **2017**, *71*, 627–635.
7. Mao, Q.; Dou, L.; Zong, Q.; Ding, Z. Attitude controller design for reusable launch vehicles during reentry phase via compound adaptive fuzzy H-infinity control. *Aerosp. Sci. Technol.* **2017**, *72*, 36–48.

8. Zang, L.; Lin, D.; Chen, S.; Wang, H.; Ji, Y. An on-line guidance algorithm for high L/D hypersonic reentry vehicles. *Aerosp. Sci. Technol.* **2019**, *89*, 150–162.
9. Wei, X.; Lan, X.; Liu, L.; Wang, Y. Rapid trajectory planning of a reusable launch vehicle for airdrop with geographic constraints. *Int. J. Adv. Robot. Syst.* **2019**, *16*, 1–14.
10. Zhou, H.; Wang, X.; Cui, N. A Novel Reentry Trajectory Generation Method Using Improved Particle Swarm Optimization. *IEEE Trans. Veh. Technol.* **2019**, *68*, 3212–3223.
11. Wang, X.; Guo, J.; Tang, S.; Qi, S.; Wang, Z. Entry trajectory planning with terminal full states constraints and multiple geographic constraints. *Aerosp. Sci. Technol.* **2019**, *84*, 620–631.
12. Li, M.; Hu, J. An approach and landing guidance design for reusable launch vehicle based on adaptive predictor–corrector technique. *Aerosp. Sci. Technol.* **2018**, *75*, 13–23.
13. Hameed, A.S.; Bindu, D.G.R. A Novel Flare Maneuver Guidance for Approach and Landing Phase of a Reusable Launch Vehicle. In *Advances in Science and Engineering Technology, Proceedings of the International Conferences (ASET), Dubai, United Arab Emirates, 26 March–11 April 2019;* IEEE: Piscataway, NJ, USA, 2019; pp. 1–6.
14. Ridder, S.D.; Mooij, E. Optimal Longitudinal Trajectories for Reusable Space Vehicles in the Terminal Area. *J. Spacecr. Rocket.* **2011**, *48*, 642–653.
15. Corraro, F.; Morani, G.; Nebula, F.; Cuciniello, G.; Palumbo, R. GN&C Technology Innovations for TAEM: USV DTFT2 Mission Results. In Proceedings of the 17th AIAA International Space Planes and Hypersonic Systems and Technologies Conference, San Francisco, CA, USA, 11–14 April 2012; American Institute of Aeronautics and Astronautics: Reston, VA, USA, 2012; pp. 2004–2013.
16. Horneman, K.; Kluever, C. Terminal Area Energy Management Trajectory Planning for an Unpowered Reusable Launch Vehicle. In Proceedings of the AIAA Atmospheric Flight Mechanics Conference and Exhibit, Minneapolis, MN, USA, 13–16 August 2012; American Institute of Aeronautics and Astronautics: Reston, VA, USA, 2012; pp. 36–38.
17. Mayanna, A.; Grimm, W.; Well, K. Adaptive Guidance for Terminal Area Energy Management (TAEM)of Reentry Vehicles. In Proceedings of the AIAA Guidance, Navigation, and Control Conference and Exhibit, Honolulu, HI, USA, 18–21 August 2012; American Institute of Aeronautics and Astronautics: Reston, VA, USA, 2012; pp. 1–11.
18. Kluever, C.A. Terminal Guidance for an Unpowered Reusable Launch Vehicle with Bank Constraints. *J. Guid. Control. Dyn.* **2007**, *30*, 162–168.
19. Fonseca, J.; Dilão, R. Dynamic guidance of orbiter gliders: Alignment, final approach, and landing. *CEAS Space J.* **2019**, *11*, 123–145.
20. Pengfei, F.; Fan, W.; Yonghua, F.; Jie, Y. In-flight Longitudinal Guidance for RLV in TAEM Phase. In Proceedings of the 2018 IEEE 4th International Conference on Control Science and Systems Engineering (ICCSSE), Wuhan, China, 24–26 August 2019; pp. 296–303.
21. Baek, J.H.; Lee, D.W.; Kim, J.H.; Cho, K.R.; Yang, J.S. Trajectory optimization and the control of a re-entry vehicle in TAEM phase. *J. Mech. Sci. Technol.* **2008**, *22*, 1099–1110.
22. Cazaurang, F.; Falcoz, A.; Morio, V.; Vernis, P. Robust terminal area energy management guidance using flatness approach. *IET Control Theory Appl.* **2010**, *4*, 472–486.
23. Zheng, B.; Liang, Z.; Li, Q.; Ren, Z. Trajectory tracking for RLV terminal area energy management phase based on LQR. In Proceedings of the 2014 IEEE Chinese Guidance, Navigation and Control Conference, Yantai, China, 8–10 August 2015; pp. 2520–2524.
24. Grantham, K. Adaptive Critic Neural Network Based Terminal Area Energy Management/Entry Guidance. In Proceedings of the 41st Aerospace Sciences Meeting and Exhibit, Reno, NV, USA, 6–9 January 2012; American Institute of Aeronautics and Astronautics: Reston, VA, USA, 2012; pp. 4–6.
25. Mu, L.; Yu, X.; Wang, B.; Zhang, Y.; Wang, X.; Li, P. 3D gliding guidance for an unpowered RLV in the TAEM phase. In Proceedings of the 2018 33rd Youth Academic Annual Conference of Chinese Association of Automation (YAC), Nanjing, China, 18–20 May 2018; pp. 409–414.
26. Kluever, C.; Horneman, K.; Schierman, J. Rapid Terminal-Trajectory Planner for an Unpowered Reusable Launch Vehicle. In Proceedings of the AIAA Guidance, Navigation, and Control Conference, Chicago, IL, USA, 10–13 August 2009.
27. Lan, X.J.; Liu, L.; Wang, Y.J. Online trajectory planning and guidance for reusable launch vehicles in the terminal area. *Acta Astronaut.* **2016**, *118*, 237–245.
28. Lan, X.; Xu, W.; Wang, Y. 3D Profile Reconstruction and Guidance for the Terminal Area Energy Management Phase of an Unpowered RLV with Aerosurface Failure. *J. Aerosp. Eng.* **2020**, *33*, 04020003.
29. Busoniu, L.; Ernst, D.; De Schutter, B.; Babuska, R. Approximate reinforcement learning—An overview. In Proceedings of the IEEE Symposium on Adaptive Dynamic Programming and Reinforcement Learning (ADPRL), Paris, France, 11–15 April 2011; pp. 1–8.
30. Wei, Q.; Liu, D.; Shi, G. A Novel Dual Iterative Q-Learning Method for Optimal Battery Management in Smart Residential Environments. *IEEE Trans. Ind. Electron.* **2014**, *62*, 2509–2518.
31. Tampuu, A.; Matiisen, T.; Kodelja, D.; Kuzovkin, I.; Korjus, K.; Aru, J.; Aru, J.; Vicente, R. Multiagent Cooperation and Competition with Deep Reinforcement Learning. *arXiv* **2015**, arXiv:1511.08779v1.
32. Yang, C.; Chen, C.; Wang, N.; Ju, Z.; Wang, M. Biologically Inspired Motion Modeling and Neural Control for Robot Learning From Demonstrations. *IEEE Trans. Cogn. Dev. Syst.* **2018**, 1, doi:10.1109/TCDS.2018.2866477.

33. Yang, C.; Chen, C.; He, W.; Cui, R.; Li, Z. Robot Learning System Based on Adaptive Neural Control and Dynamic Movement Primitives. *IEEE Trans. Neural Netw. Learn. Syst.* **2019**, *30*, 777–787.
34. Wang, N.; Chen, C.; Yang, C. A Robot Learning Framework based on Adaptive Admittance Control and Generalizable Motion Modeling with Neural Network Controller. *Neurocomputing* **2019**, *390*, doi:10.1016/j.neucom.2019.04.100.
35. Zhao, Z.; Liu, Z. Finite-Time Convergence Disturbance Rejection Control for a Flexible Timoshenko Manipulator. *IEEE/CAA J. Autom. Sin.* **2021**, *8*, 161–172.
36. Al-Talabi, A.A.; Schwartz, H.M. Kalman fuzzy actor-critic learning automaton algorithm for the pursuit-evasion differential game. In Proceedings of the IEEE International Conference on Fuzzy Systems, Vancouver, BC, Canada, 24–26 July 2016; pp. 1015–1022.
37. Gerken, A.; Spranger, M. Continuous Value Iteration (CVI) Reinforcement Learning and Imaginary Experience Replay (IER) for learning multi-goal, continuous action and state space controllers. *arXiv* **2019**, arXiv:1908.10255v1.
38. Zimmer, M.; Weng, P. Exploiting the Sign of the Advantage Function to Learn Deterministic Policies in Continuous Domains. *arXiv* **2019**, arXiv:1906.04556v2.
39. Lan, X.; Liu, Y.; Zhao, Z. Cooperative control for swarming systems based on reinforcement learning in unknown dynamic environment. *Neurocomputing* **2020**, *410*, doi:10.1016/j.neucom.2020.06.038.
40. Leuenberger, G.; Wiering, M.A. Actor-Critic Reinforcement Learning with Neural Networks in Continuous Games. In Proceedings of the International Conference on Agents and Artificial Intelligence (ICAART), Funchal, Portugal, 16–18 January 2018.
41. Jiang, X.; Yang, J.; Tan, X.; Xi, H. Observation-based Optimization for POMDPs with Continuous State, Observation, and Action Spaces. *IEEE Trans. Autom. Control.* **2018**, 1–8, doi:10.1109/TAC.2018.2861910.
42. Lan, X.; Liu, L.; Wang, Y. ADP-Based Intelligent Decentralized Control for Multi-Agent Systems Moving in Obstacle Environment. *IEEE Access* **2019**, *7*, 59624–59630.
43. Van Hasselt, H. Reinforcement Learning in Continuous State and Action Spaces. In *Reinforcement Learning*; Springer: Berlin/Heidelberg, Germany, 2012; pp. 207–251.
44. Comsa, I.S.; Aydin, M.; Zhang, S.; Kuonen, P.; Wagen, J.F.; Yao, L. Scheduling policies based on dynamic throughput and fairness tradeoff control in LTE-A networks. In Proceedings of the 39th Annual IEEE Conference on Local Computer Networks, Edmonton, AB, Canada, 8–11 Septmber 2014; pp. 418–421.
45. Hafez, M.B.; Weber, C.; Wermter, S. Curiosity-driven exploration enhances motor skills of continuous actor-critic learner. In Proceedings of the 7th Joint IEEE International Conferences on Development and Learning and Epigenetic Robotics (ICDL-Epirob), Lisbon, Portugal, 18–21 September 2017; pp. 39–46.