Hindawi
Mathematical Problems in Engineering
Volume 2021, Article ID 6652232, 12 pages
https://doi.org/10.1155/2021/6652232

Research Article

Multiconstrained Gliding Guidance Based on Optimal and Reinforcement Learning Method

Luo Zhe,[1] Li Xinsan ,[1,2] Wang Lixin,[1] and Shen Qiang[1]

[1]*Xi'an High-Tech Institution, Xi'an 710025, China*
[2]*School of Astronautics, Northwest University, Xi'an 710072, China*

Correspondence should be addressed to Li Xinsan; lixinsan_lixinsan@163.com

Received 20 December 2020; Accepted 13 May 2021; Published 27 May 2021

Academic Editor: Defeng He

In order to improve the autonomy of gliding guidance for complex flight missions, this paper proposes a multiconstrained intelligent gliding guidance strategy based on optimal guidance and reinforcement learning (RL). Three-dimensional optimal guidance is introduced to meet the terminal latitude, longitude, altitude, and flight-path-angle constraints. A velocity control strategy through lateral sinusoidal maneuver is proposed, and an analytical terminal velocity prediction method considering maneuvering flight is studied. Aiming at the problem that the maneuvering amplitude in velocity control cannot be determined offline, an intelligent parameter adjustment method based on RL is studied. This method considers parameter determination as a Markov Decision Process (MDP) and designs a state space via terminal speed and an action space with maneuvering amplitude. In addition, it constructs a reward function that integrates terminal velocity error and gliding guidance tasks and uses Q-Learning to achieve the online intelligent adjustment of maneuvering amplitude. The simulation results show that the intelligent gliding guidance method can meet various terminal constraints with high accuracy and can improve the autonomous decision-making ability under complex tasks effectively.

1. Introduction

Hypersonic gliding vehicles have become the focus and hotspot of the aerospace industry due to their long-range flight and large-scale maneuverability. Guidance is one of the core technologies, which requires the control vehicles to complete a given guidance task under the conditions of meeting various process constraints. Gliding guidance faces challenges such as complex flight environment, strong uncertainty, diversified flight missions, multiple processes, and terminal constraints [1]. Therefore, gliding guidance methods need to ensure the accuracy of terminal constraints, the robustness of process deviations, and the adaptability of diverse guidance tasks.

In gliding guidance field, the standard trajectory tracking is the most traditional gliding guidance method. This method can be divided into two parts: first, the standard trajectory design that meets a variety of process constraints and terminal constraints, and second, the guidance command calculation to ensure guidance accuracy and robustness, that is, trajectory tracking [2]. This method has strong reliability and can reduce the online cost of calculation, but the standard trajectory and tracking control parameters need to be redesigned when guidance mission is changed, which limits the adaptability [3]. Predictor-corrector needs to predict the terminal status online and correct the current guidance parameters based on terminal errors [4]. However, the analytical predictor-corrector method requires a lot of simplification of the motion model, and it is difficult to guarantee the guidance accuracy; the numerical predictor-corrector method requires complicated online calculations, which limits the real-time performance [5]. Based on gliding flight characteristics and the two-point boundary value problem, a multiconstraint optimal gliding guidance is derived using the principle of maximum value [6]. However, the terminal velocity control accuracy is greatly affected by the feedback coefficient and usually needs to be adjusted artificially [7, 8].

Artificial intelligence based on machine learning is a hot topic in current research. RL, as an algorithm embodying intelligent decision-making, has been recognized by many scholars and has employed in the field of path planning and parameter determination preliminary [9–11]. Junell et al. [12] present the setup and results of a high-level RL problem for both simulation and real flight tests. The problem provided is that of a quadrotor taking pictures of a disaster site, in which the environment is completely unknown at first and the agent must learn where the sites of interest are and the most efficient way to get there. As for the intercept guidance problem, Gaudet employed RL to learn a homing-phase guidance law that is optimal with respect to the missile's airframe dynamics as well as sensor and actuator noise and delays. However, the state space and action space are not given [13]. Furthermore, for the exo-atmospheric interception of maneuvering targets with line-of-sight angle and their rate of change information only, reinforcement metalearning is employed to optimized policy to adapt to target acceleration, and the policy has superior performance as compared to augmented zero-effort miss guidance with perfect target acceleration knowledge [14, 15]. In [16], RL is utilized to generate reference bank angle commands for directing the aircraft to close proximity of the updraft, and then, the problem of online trajectory generation is reduced to a simple search in a static state-action value table. In addition, the deep RL is effective for aircraft landing guidance. In [17], an environment which can be applied in agent training to solve aircraft landing guidance problem is proposed, and the agent receives the current state of the aircraft and the runway in a sector and outputs an action to guide the aircraft. The Deep Q Network (DQN) algorithm is used to verify the feasibility of the environment. In short, RL has been studied in the area of agent planning and control, but there are no public results in hypersonic vehicle guidance.

Aiming at the key problems of traditional gliding guidance methods and the advantages of RL methods, this paper proposes a multiconstrained intelligent gliding guidance strategy based on optimal guidance, predictor-corrector, and RL. First, the optimal gliding guidance method is used to meet the terminal latitude, longitude, altitude, and flight-path-angle (FPA) constraints. Second, a velocity control strategy based on lateral maneuver is proposed, and the terminal velocity is analyzed and predicted considering the characteristics of gliding flight and lateral maneuver comprehensively. Finally, a framework model of RL is established, and Q-learning is used to adjust the maneuvering amplitude intelligently in velocity control to ensure the terminal velocity control accuracy. This guidance strategy will solve the problem of "dimensional disaster" and guarantee learning efficiency and then realize multiconstrained adaptive guidance.

2. Intelligent Gliding Guidance Problem Formulation

Intelligent gliding guidance requires interaction and perception between the vehicle and the environment to improve the adaptive ability when the flight mission is changed online. The near space where the gliding aircraft is located is extremely complicated, the flight distance is long, and the atmospheric density and the aerodynamic coefficient are greatly deviated, so the use of constant parameters to control the entire gliding flight must have major defects. In addition, the gliding vehicle is faced with diversified flight missions, or even the missions are changed online, so the guidance parameters need to be adjusted online according to the actual flight status and the current mission. The thesis utilizes the optimal control, predictor-corrector, and RL to achieve the guidance goal comprehensively. The following key issues need to be solved when using RL to achieve intelligent parameter adjustment:

(1) Rapid convergence problem of online learning under high dynamic and strong time-varying conditions. The particularity of gliding flight makes it impossible to obtain massive data samples offline, so only the online learning can be performed based on the current flight status and guidance mission. Gliding guidance is a highly dynamic and time-varying centroid control process, which means the efficiency of learning determines the quality of the guidance parameters and the success or failure of the guidance task directly. Therefore, gliding flight must be optimized in guidance strategies and methods to improve learning efficiency.

(2) The "dimensional disaster" problem is brought by the time-continuous guidance and multiple states for RL. Both the flight states, including flight time, position, and velocity, and control variables' angle of attack and the bank angle are time-continuous physical quantities. However, RL is a typical time-discrete optimization process, so it will easily lead to a sudden increase in the dimension of the variable if multidimensional state quantities and control quantities are used to build a discrete RL model. This phenomenon is called "dimensional disaster," which will reduce the learning rate.

(3) Design of reward function in multiconstrained gliding guidance. The final task of gliding guidance is to meet the terminal latitude, longitude, altitude, velocity, and FPA constraints. The higher the accuracy, the better the guidance performance. Therefore, the reward function must include all terminal constraints in theory. However, the above constraints have mutual influences and couplings, and the magnitudes of the dimensions and actual values are quite different. If there is a defect in reward function design, it will affect the learning efficiency and even lead to the failure of learning and guidance tasks.

Aiming at the abovementioned intelligent gliding guidance problems, the following strategies are proposed.

The intelligent gliding guidance strategy shown in Figure 1 can be described as follows.

Firstly, in the guidance strategy, optimal guidance, predictor-corrector, and RL are combined to achieve the guidance task. Analytical optimal guidance is introduced to meet the constraints of latitude, longitude, altitude, and

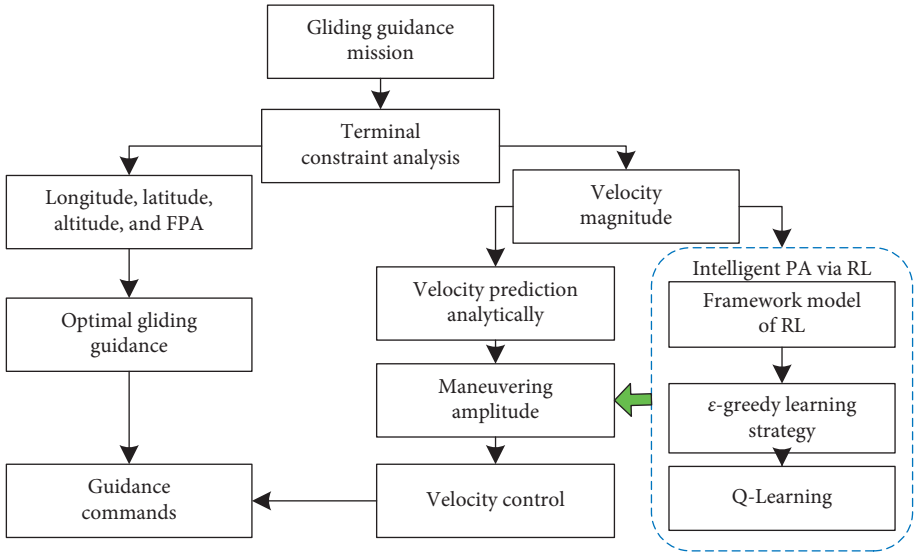

Figure 1: Intelligent gliding guidance strategy (PA: parameter adjustment).

FPA, and lateral sinusoidal maneuver is proposed to control terminal velocity. As for the problems of uncertainty in gliding flight and diversified flight missions, especially the influence of terminal velocity prediction error on guidance performance, RL is further used to adjust the maneuvering amplitude intelligently in velocity control.

Secondly, in the learning method, the online RL is adopted to realize intelligent parameter adjustment. RL does not require sample construction, but only needs to build the state space and action space, design the reward function, and select the action instruction to obtain the maximum reward iteratively. To ensure the online learning efficiency of RL, the determination of the maneuvering amplitude in velocity control is the unique learning task. Therefore, the terminal velocity is taken as the state variable, the maneuvering amplitude is taken as the action variable, and the dimensions of the state space and the action space are reduced through a reasonable discretization process. Furthermore, on the basis of highlighting terminal velocity control, a reward function is designed considering all the guidance constraints.

Finally, the ε-greedy strategy is used for iterative learning, and the most typical Q-Learning method is employed to update the reward value function in RL.

3. Optimal Gliding Guidance and Velocity Control

According to the intelligent guidance strategy, the analytical optimal guidance method is adopted to meet the terminal latitude, longitude, altitude, and FPA constraints, and the maneuvering flight is further increased in the lateral direction. The terminal velocity is predicted analytically considering lateral maneuvering flight.

3.1. Optimal Gliding Guidance. The gliding guidance mission is to generate guidance commands based on current states which can control the glider to satisfy the multiple terminal required constraints. The current flight states are velocity v, FPA θ, and velocity azimuth angle σ which is measured from the north in a clockwise direction, longitude λ, latitude ϕ, and altitude h. As for the gliding guidance, the terminal constraints are

$$x_f = \left(\lambda_f, \phi_f, h_f, \theta_f\right). \tag{1}$$

In the previous research, we have designed the optimal gliding guidance law in longitudinal and lateral directions, in which the quasi-equilibrium glide condition employed to simplify the guidance model and minimum energy consumption is taken as the optimal performance index. The guidance law is the required overload in the two directions:

$$\begin{cases} u_y^* = n_y^*, \\ u_z^* = n_z^*. \end{cases} \tag{2}$$

In longitudinal direction, the guidance goal is to satisfy the terminal altitude and flight-path angle constraints at the given terminal position, and the guidance law is [8]

$$u_y^* = k\left(C_1 L_R - C_2\right) + 1, \tag{3}$$

where the coefficient $k = (g_0/v^2) \approx (g/v^2)$, L_R is the current range, calculated from initial position (λ_0, ϕ_0) and the current location (λ, ϕ), and C_1 and C_2 are coefficients obtained based on optimal control theory [8]:

$$\begin{cases} C_1 = \dfrac{6\left(\left(L_R - L_{Rf}\right)\left(\theta_f + \theta\right) - 2h + 2h_f\right)}{k^2\left(L_R - L_{Rf}\right)^3}, \\ C_2 = \dfrac{2\left(L_R L_{Rf}\left(\theta - \theta_f\right) - L_{Rf}^2\left(2\theta + \theta_f\right) + L_R^2\left(2\theta_f + \theta\right) + 3\left(L_{Rf} + L_R\right)\left(h_f - h\right)\right)}{k^2\left(L_R - L_{Rf}\right)^3}. \end{cases} \tag{4}$$

The total gliding range L_{Rf} can be calculated according to the initial position (λ_0, ϕ_0) and the location of the given target (λ_f, ϕ_f):

$$L_{Rf} = R_e \arccos\left(\sin\,\phi_0 \sin\,\phi_f + \cos\,\phi_0 \cos\,\phi_f \cos\left(\lambda_f - \lambda_0\right)\right). \tag{5}$$

The lateral guidance mission is to eliminate heading error at the terminal flight range L_{Rf}. The heading error is designed as

$$\Delta\sigma = \sigma_{\mathrm{LOS}} - \sigma. \tag{6}$$

In equation (6), σ_{LOS} is the line-of-sight (LOS) angle measured from the north in a clockwise direction. The LOS angle can be computed according to the current position (λ, ϕ) and the target:

$$\tan\,\sigma_{\mathrm{LOS}} = \frac{\sin\left(\lambda_f - \lambda\right)}{\cos\,\phi\,\tan\,\phi_f - \sin\,\phi\,\cos\left(\lambda_f - \lambda\right)}. \tag{7}$$

The lateral optimal guidance law with minimum energy consumption is

$$u_z^* = \frac{\Delta\sigma}{k\left(L_{Rf} - L_R\right)}. \tag{8}$$

With the optimal gliding guidance law shown in equations (3) and (8), the control variables, namely, angle of attack and bank angle can be calculated:

$$\begin{cases} \dfrac{\rho v^2 S_m C_L\left(Ma, \alpha\right)}{2g_0} = \sqrt{n_y^{*2} + n_z^{*2}}, \\ \upsilon = \arctan\left(\dfrac{n_z^*}{n_y^*}\right), \end{cases} \tag{9}$$

where g_0 is the gravity acceleration on the sea level. The first function in equation (9) is inverse interpolation.

3.2. Velocity Control and Analysis. The premise of the velocity control is that the terminal velocity can be obtained rapidly and accurately, so an analytical method is used to predict the terminal velocity. The major forces acted on the glider are aerodynamic lift and earth gravity, and the velocity differential is

$$\dot{v} = -\left(\frac{D}{m}\right) - g\,\sin\,\theta. \tag{10}$$

Equation (10) contains all flight states, which means solving equation (10) requires other differential equations, which makes the analytical solution impossible. Therefore, equation (10) needs to be converted reasonably according to the gliding flight characteristics to obtain the terminal velocity analytically. The purpose of velocity control is to satisfy the velocity magnitude constraint at the given range, and the terminal arrival time is free, so equation (10) can be reconstructed based on the range differential:

$$\frac{\mathrm{d}v}{\mathrm{d}L_R} = \frac{\left(\mathrm{d}v/\mathrm{d}t\right)}{\left(\mathrm{d}L_R/\mathrm{d}t\right)} = \frac{-\left(D/m\right) - g\,\sin\,\theta}{v\,\cos\,\theta}. \tag{11}$$

Furthermore, the state parameters in equation (11) are cured using the "averaging method." The time-varying drag acceleration in the remaining range is assumed to be the average value of the current actual acceleration D_c and the terminal acceleration D_f. In the terminal stage, the altitude is relatively lower and the vehicle has sufficient aerodynamic lift to achieve gliding flight, that is, the longitudinal component of the lift is basically equal to the negative gravity:

$$L_f \cos\,\upsilon \approx mg, \tag{12}$$

where L_f is aerodynamic lift in terminal position. For quasi-balanced glide vehicle, the lift-drag ratio $(R_{(L/D)})$ is always large and the variation is small, which means $R_{(L/D)}$ can be considered as a constant in a guidance cycle. As a result, the aerodynamic drag acceleration is

$$D_f = \frac{L_f}{R_{(L/D)}} = \frac{mg}{R_{(L/D)}\cos\,\upsilon}. \tag{13}$$

Then, the average aerodynamic drag can be indirectly expressed as

$$\overline{D} = \frac{D_c + D_f}{2} \approx \frac{D_c}{2} + \frac{mg}{2R_{L/D}\cos\,\upsilon}. \tag{14}$$

FPA θ in equation (11) needs to be converted. Due to the equilibrium gliding characteristic, both the FPA and its differentiation are small. Therefore, FPA θ can be simplified as the average value of the current FPA and terminal FPA constraint:

$$\begin{cases} \cos\,\theta = 1, \\ \tan\,\overline{\theta} \approx \overline{\theta} = \dfrac{\theta + \theta_f}{2}. \end{cases} \tag{15}$$

In order to predict the terminal velocity, the bank angle υ in the drag acceleration needs to be converted as well. The bank angle υ can be obtained based on the second function in equation (9). However, the terminal velocity is uncontrolled, so we need to introduce maneuver further.

Maneuvering flight can be reflected in the longitudinal and lateral, respectively, and the more complex the form of maneuvering trajectory, the greater the hit point prediction error. The surface-symmetrical glider needs to keep quasi-equilibrium gliding flight in longitudinal profile, so only lateral pendulum maneuvering trajectory can be achieved by adding the maneuvering item to the original optimal guidance law shown in equations (3) and (8). Then, the bank angle is computed as

$$v = \arctan\left(\frac{n_z^* + n_{vc}}{n_y^*}\right), \tag{16}$$

where n_{vc} is maneuvering overload used to control terminal velocity. Maneuvering is a damage to the original optimal guidance law, so the design of maneuvering overload needs to minimize its influence on the terminal guidance accuracy. The full cycle of lateral sinusoidal maneuver can increase process energy loss and can make the lateral error produced by maneuver cancel positive and negative. As a result, the maneuvering overload is designed as

$$n_{vc} = A_{vc}\sin\left(2k_m\pi\frac{L_R}{L_{Rf}}\right), \quad k_m \in \mathbf{Z}_+, \tag{17}$$

where A_{vc} is maneuvering amplitude and k_m represents the frequency. In equation (17), $n_{vc}(L_R = L_{Rf}) = 0$, where n_{vc} will be zero at the terminal position to ensure guidance accuracy. In equation (16), when the initial heading error is small, the optimal overload command n_z^* is basically zero, that is, the laterally required overload is mainly the maneuvering overload n_{vc}. Therefore, the bank angle can be calculated from the average lateral overload $\overline{n}_{vc}$ during the gliding flight:

$$\overline{n}_{vc} = \frac{\int_0^{L_{Rf}}\left|A_{vc}\sin\left(2k_m\pi L_R/L_{Rf}\right)\right|\mathrm{d}L_R}{L_{Rf}} = \frac{2A_{vc}}{\pi}. \tag{18}$$

To calculate the bank angle, it needs to simplify the longitudinal optimal overload command and obtain the overload factor in remaining flight analytically. Equilibrium gliding is the main flight characteristics of a gliding vehicle. The altitude changes very gently, that is, the longitudinal overload remains unchanged basically. Therefore, the average overload of the glide flight can be calculated based on the current required overload and terminal overload:

$$\overline{n}_y = \frac{n_y^* + 1}{2}, \tag{19}$$

where "1" is the terminal overload, which means the strict constant altitude flight in terminal point. Combining the average overload in equations (18) and (19), the bank angle can be calculated as

$$\cos\overline{v} = \frac{\pi\left(n_y^* + 1\right)}{\sqrt{16A_{vc}^2 + \pi^2\left(n_y^* + 1\right)^2}}. \tag{20}$$

Using the average drag acceleration in equation (14), the average FPA in equation (15), and the average bank angle in equation (20) replace the current flight status in equation (11), it can transformed into

$$v\frac{\mathrm{d}v}{\mathrm{d}L_R} = -\frac{D_c}{2m} - \frac{g\sqrt{16A_{vc}^2 + \pi^2\left(n_y^* + 1\right)^2}}{2R_{L/D}\pi\left(n_y^* + 1\right)} - g\frac{\theta + \theta_f}{2}. \tag{21}$$

From current states to terminal states, resolve the definite integration for the two sides of equation (21), and we can obtain the predicted terminal velocity:

$$v_{fp}^2 = v^2 - 2\left(\frac{D_c}{2m} + \frac{g\sqrt{16A_{vc}^2 + \pi^2\left(n_y^* + 1\right)^2}}{2R_{(L/D)}\pi\left(n_y^* + 1\right)} + g\frac{\theta + \theta_f}{2}\right)\left(L_{Rf} - L_R\right). \tag{22}$$

It can be known from equation (22) that the larger the maneuvering amplitude A_{vc}, the larger the energy consumption, that is, the smaller the terminal velocity. In theory, by setting different maneuvering amplitudes A_{vc}, different terminal velocities can be obtained. Conversely, the required maneuvering amplitude can also be calculated analytically according to the terminal velocity constraint based on equation (22). However, the predicted terminal velocity must be biased, so calculating amplitude according to equation (22) directly will affect the velocity control accuracy. To this end, the paper adjusts the maneuvering amplitude A_{vc} intelligently to achieve precise control of the terminal velocity. Figure 2 shows the predicted terminal velocity obtained by analytical and numerical methods, respectively. From the simulation results, it can be known that equation (22) can be used to predict the terminal velocity accurately and reflect the relationship between terminal velocity and maneuvering amplitude.

4. Framework of Reinforcement Learning

This paper employs artificial intelligence (AI) algorithm to modify the maneuvering amplitude. Machine learning is an important research content in the field of AI, which includes supervised learning, semisupervised learning, and RL [10]. Among them, RL can describe and solve the problem that the agent learns strategies to maximize rewards or achieve specific goals in the process of interacting with the environment and has been used in the control field of robots and agents [12].

4.1. Description of Reinforcement Learning. RL is a tentative evaluation process. The learning process is shown in Figure 3. The agent selects actions to the environment and receives reinforcement signals (reward or punishment). According to the reinforcement signal and the current state of the environment, the agent then chooses the next action, in which the principle of choice is to increase the probability of positive reinforcement (award). The action chosen affects not only the immediate reinforcement value but also the state of the environment at the next moment and the final reinforcement value [12].

A common model of RL is the standard MDP, which consists of five elements $(\mathbf{S}, \mathbf{A}, P_{sa}, \gamma, R)$:

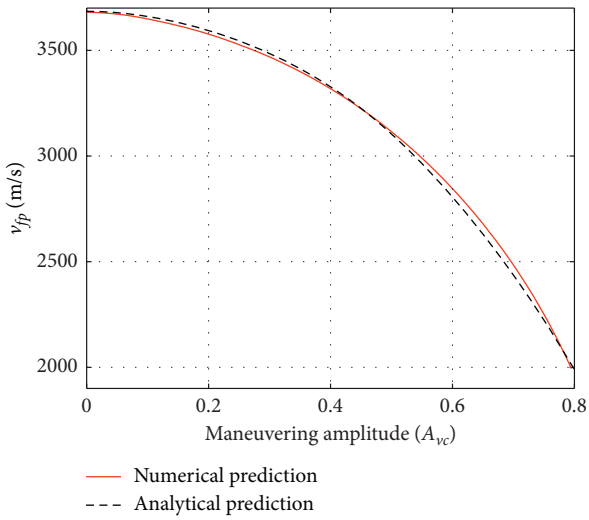

Figure 2: Relationship between predicted velocities and maneuvering amplitude.

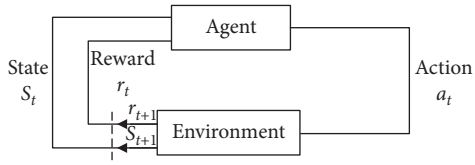

Figure 3: Framework of RL.

(1) S represents a state space, which represents a data set composed of parameter information describing the movement state of the agent, such as the position and velocity of the aircraft.

(2) A represents an action space, which represents a discrete data set composed of specific actions or operation commands that can directly affect the movement state of the agent.

(3) P_{sa} is the state transition probability, which indicates the probability distribution of transition to other states after the action $a \in \mathbf{A}$ in the current state, and can be determined by the object model usually. According to whether P_{sa} is known, RL is divided into two categories, model based and model free.

(4) $\gamma \in [0, 1)$ is a discount factor, which reflects that the future states are worthy of affecting the current rewards. The larger the discount factor, the more attention by decision makers to long-term benefits. Conversely, a smaller value means that decision makers pay more attention to current interests.

(5) R is a reward function, which represents the reward obtained by transferring the state s_t at a certain moment to the next state s_{t+1} with the help of action a_t, which needs to be quantified during the learning process.

4.2. Description of Q-Learning. RL requires iterative calculations, including typical value iterations and strategy iterations. However, these two methods are difficult to solve the problem of high-dimensional or time-continuous learning. In addition, they both rely on known environmental models, that is, both are model-based learning algorithms. Q-Learning is a model-free iterative learning method that can solve complex high- dimensional learning problems.

Q-Learning is a typical representative of Temporal-Difference (TD), which uses the sampled and estimated values of the state to update the current state value function. A simplest TD algorithm, namely, TD (0), is [13]

$$V(s_t) \longleftarrow V(s_t) + \alpha[R_{t+1} + \gamma V(s_{t+1}) - V(s_t)]. \quad (23)$$

In Q-Learning, $Q^{\pi}(s, a)$ is a value function of state behavior, which represents the specific value of the value function corresponding to the current state and action under the current policy π. If the state space is m-dimensional and the action space is n-dimensional, then $Q^{\pi}(s, a)$ is a $m \times n$-dimensional table, so it can be called a Q-table. The update algorithm of the Q-Learning is [14]

$$Q(s, a) \longleftarrow Q(s, a) + \alpha\left[R + \gamma \max_{a} Q(s', a) - Q(s, a)\right]. \quad (24)$$

The specific operation steps of Q-Learning algorithm are as follows [13, 14]:

(1) Initialize $Q(s, a)$ table arbitrarily.

(2) Initialize states $\boldsymbol{S}$ (for each episode).

(3) Repeat the following operations:

  (1) Choose action a using policy (e.g., å-greedy) derived from current Q
  (2) Perform the current action a and obtain quantized reward R and next state s'
  (3) Update Q-table, $Q(s, a) \longleftarrow Q(s, a) + \alpha[R + \gamma \max_{a} Q(s', a) - Q(s, a)]$
  (4) Update state $s \longleftarrow s'$
  (5) Finish the learning of current episode if s is terminal

(4) Repeat step (3) via the updated Q-table, until learning times are satisfied.

5. Intelligent Modification of Feedback Coefficient

In the terminal velocity control based on lateral maneuver, the intelligent adjustment of the maneuvering amplitude is an effective means to eliminate process deviations and velocity prediction errors and to respond to a variety of flight missions. Because the terminal velocity is only related to the current and future maneuvering amplitudes and not to the past information, the determination of the maneuvering amplitudes is consistent with the MDP process. According to the requirements of RL and Q-Learning, it is necessary to build an intelligent parameter adjustment model and design the state and action space and the reward function according

to the actual guidance task. The intelligent parameter adjustment logic based on RL is shown in Figure 4.

5.1. Learning Policy Design. The ε-greedy policy is used to determine the maneuvering amplitude online to meet the terminal velocity constraint. In the first parameter adjustment, the Q-table is initialized with a random method to explore more states and actions. In the subsequent adjustments, the initial Q-table is inherited from the previous adjustments to velocity up the iterative convergence rate. In addition, in order to give full play to the Q-learning algorithm's exploration and optimization capabilities, ε can be selected to be larger in the early stages of learning to explore more states and actions and gradually reduced in the later stages to make gliding guidance make correct actions to ensure terminal guidance accuracy based on obtained experience [18]. The discount factor is $\gamma = 0.9$, which means that Q-learning pays more attention to long- term rewards. The process of intelligent parameter adjustment based on Q-learning is shown in Figure 5.

5.2. State Space Construction. State space is an indispensable element in RL, which is a data set that reflects the state of the flight process or the terminal state and must include all possible state parameter values. The thesis uses the intelligent parameter adjustment method to meet the terminal velocity constraint, so the state space can be designed as a data set composed of terminal velocity. Gliding guidance is a time-continuous centroid control problem, and its terminal velocity must also be time-continuous. Therefore, when using discrete RL for intelligent parameter adjustment, it is necessary to discretize the terminal velocity, that is, the state space is a data set composed of the discretized terminal velocity.

According to the previous research [8], the terminal velocity range is set to [2000, 4000] m/s. We discretize the terminal velocity into an equally spaced state space with 51 discrete points and an interval of 40 m/s. Then, the state space $\boldsymbol{S}$ is $\boldsymbol{S}$ = [2000, 2040, 2080, 2120, 2160, 2200, 2240, 2280, 2320, 2360, 2400, 2440, 2480, 2520, 2560, 2600, 2640, 2680, 2720, 2760, 2800, 2840, 2880, 2920, 2960, 3000, 3040, 3080, 3120, 3160, 3200, 3240, 3280, 3320, 3360, 3400, 3440, 3480, 3520, 3560, 3600, 3640, 3680, 3720, 3760, 3800, 3840, 3880, 3920, 3960, 4000].

5.3. Action Space Construction. According to the definition of action space in RL, "action" needs to influence all the above "state." There are many factors that affect the state, that is, the terminal velocity, including the current flight states and maneuver such as up, down, left, and right. Overly complex action space will increase the search difficulty and then affect learning efficiency. Aiming at this problem, the "action" is designed as the maneuvering amplitude that can directly affect the gliding flight and terminal velocity, that is, the action space is a data set of the maneuvering amplitude. In the previous research, the optimal guidance and maneuvering deceleration methods were used to generate guidance commands, in which the maneuvering amplitude

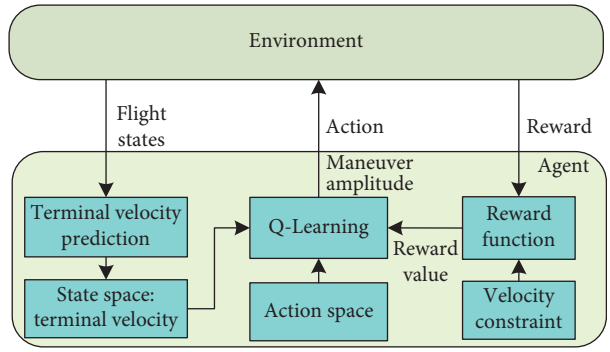

Figure 4: Block diagram of intelligent parameter adjustment using RL.

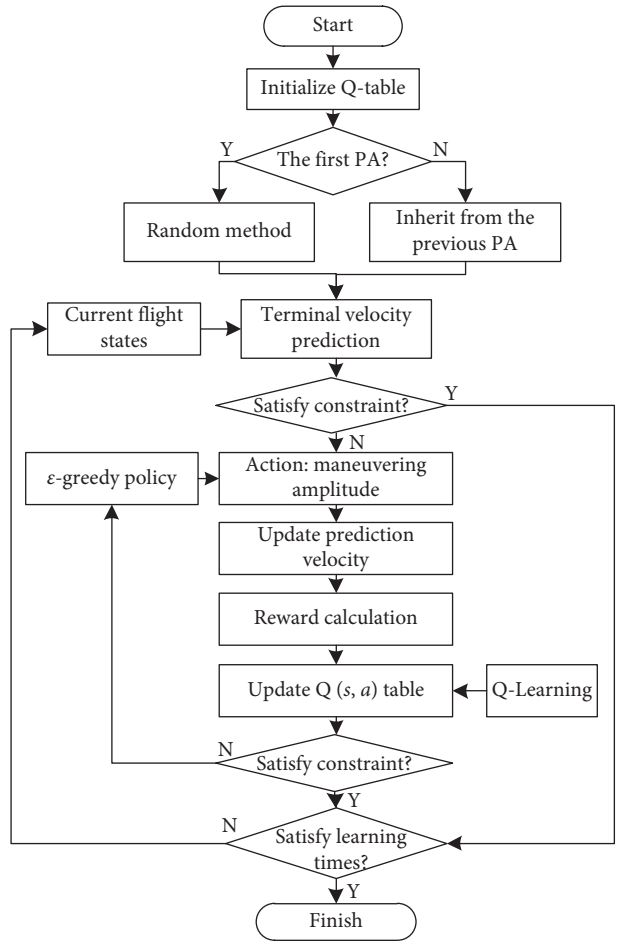

Figure 5: Flowchart of intelligent PA using Q-learning.

was set 0–0.75 to make terminal velocity varying within the range of 2000–4000 m/s [7, 8]. Therefore, the paper appropriately expands the range to [0, 0.8] and then designs an action space composed of 30 discrete points: $\boldsymbol{A}$ = [0, 0.1, 0.2, 0.3, 0.35, 0.40, 0.45, 0.5, 0.52 0.54, 0.56, 0.58, 0.6, 0.62. 0.64, 0.66, 0.68, 0.7, 0.71, 0.72, 0.73, 0.74, 0.75, 0.76, 0.77, 0.78, 0.785, 0.79, 0.795, 0.8].

5.4. Reward Function Design. As the core of RL, quantitative reward function is used to judge the performance of action. In the gliding guidance studied by this paper, the RL method

is used to modify the maneuvering amplitude online to meet the velocity constraints with high accuracy. Therefore, we design the reward function according to the terminal velocity constraint v_f^* and predicted value v_{fp}:

$$R = \begin{cases} -\dfrac{\left|v_{fp} - v_f^*\right|}{100}, & \text{if } \left|v_{fp} - v_f^*\right| < 200, \\ -2 * \dfrac{\left|v_{fp} - v_f^*\right|}{100}, & \text{if } \left(v_{fp} - v_f^*\right) < -200, \\ -1.5 * \dfrac{\left|v_{fp} - v_f^*\right|}{100}, & \text{if } \left(v_{fp} - v_f^*\right) > 200. \end{cases} \tag{25}$$

The physical meaning of the reward function in equation (25) is as follows. When the difference between the predicted velocity and the required velocity is less than 200 m/s, the reward value is a negative absolute value of the velocity difference. When the predicted velocity is much smaller than the required velocity, meaning the excessive energy loss will lead to the flight mission failure, the most severe "punishment" should be given. When the predicted velocity is much higher than the required constraint, excessively fast velocity will cause the process constraints such as dynamic pressure and overload to be exceeded, then the relatively "punishment" will be given. The purpose of the reward equation (25) is to control the predicted velocity and the required velocity to be equal so that the maximum reward value is "zero."

6. Simulations and Analysis of Guidance Performance

Taking CAV-H as the simulation object [19], numerical simulation is used to test the performance of the intelligent gliding guidance method. The initial parameters: velocity is 6500 m/s, FPA is 0°, velocity azimuth equals to the LOS azimuth, longitude is 0°, latitude is 0°, and altitude is 65 km. Terminal parameters: longitude is 95°, latitude is 10°, altitude is 30 km, FPA is 0°, and velocity is 2600 m/s. In the additional lateral maneuvering (17), the maneuvering frequency is $k_m = 3$, which means a three-cycle sinusoidal maneuver is added to control the terminal velocity. For the parameters of RL, the learning period is 800 km, and the parameters are adjusted only within the first 8000 km range of the gliding flight. In ε-greedy, the first learning $\varepsilon_1 = 0.3$, the second learning $\varepsilon_2 = 0.2$, the third learning $\varepsilon_1 = 0.1$, and the subsequent learning is all zero. The terminal velocity error range in Q-learning is 40 m/s, that is, when the difference between the predicted velocity and the required velocity is less than this value, the terminal velocity constraint is considered to be satisfied.

6.1. Nominal Performance Test. According to the parameter settings of the paper, a total of ten complete maneuvering amplitude adjustments are performed during the gliding flight. Figure 6 shows the number of convergence steps and the cumulative reward value of the first and seventh parameter adjustments. It can be known from the results that the more number of convergence steps, the smaller cumulative return value. In addition, each parameter adjustment can converge after a certain number of shocks, the correction steps of the deceleration maneuvering amplitude will within two times, and the cumulative reward value will also close to a stable maximum "zero" after convergence. Comparing the effects of the two parameter adjustments, it can be seen that the convergence rate of the seventh parameter adjustment is about three times that of the first. The main reason is that, for the first time, the Q-table is initialized with random method, and the parameter ε setting is large. During the learning process, it is necessary to experience multiple "attempts" to obtain a lot of experience, resulting in a decrease in the convergence rate. Subsequent parameter adjustment inherits the Q-table obtained from the previous time, which already contains the value function information that can meet the terminal velocity constraint. As a result, it is more experienced and the convergence rate can be greatly improved.

Guidance commands are generated using the optimal gliding guidance law and the maneuvering amplitude determined by RL, and the main simulation results are shown in Figure 7. It can be seen from the simulation results that the intelligent gliding guidance method can control the vehicle to meet the terminal latitude and longitude, altitude, velocity, and FPA constraints. The terminal position error is about 12m, the altitude error is 0.6 m, the FPA error is -0.011°, and the velocity magnitude error is 20 m/s. In the initial stage of gliding, despite the large longitudinal overload command, the high flight altitude and small atmospheric density cause the vehicle to fail to achieve a balanced gliding, resulting in the synchronous jumps of altitudes and velocities. As the altitude continues to decrease and the atmospheric density continues to increase, the vehicle has sufficient lift to achieve gliding flight. In the lateral direction, optimal guidance and velocity control lead to a three-cycle sinusoidal maneuver for lateral overload and bank angle, to eliminate heading errors and meet terminal velocity constraints. Because the cycle of RL is 800 km, there is a discontinuous step change in the maneuvering amplitude, and the decreasing maneuvering amplitude is beneficial to reduce the influence of maneuvering flight on the guidance accuracy. In addition, the analytical prediction value of the terminal velocity is close to the velocity constraint value, and the variation range is small, which lays the foundation for high-precision control of the velocity.

6.2. Adaptive Performance Test. In order to further verify the adaptability of the intelligent gliding guidance method, keeping the terminal latitude, longitude, altitude, and FPA constraints unchanged, set different velocity magnitude constraints for testing. The simulation results are shown in Table 1 and Figure 8. It can be seen that the intelligent gliding guidance method can adjust the maneuvering

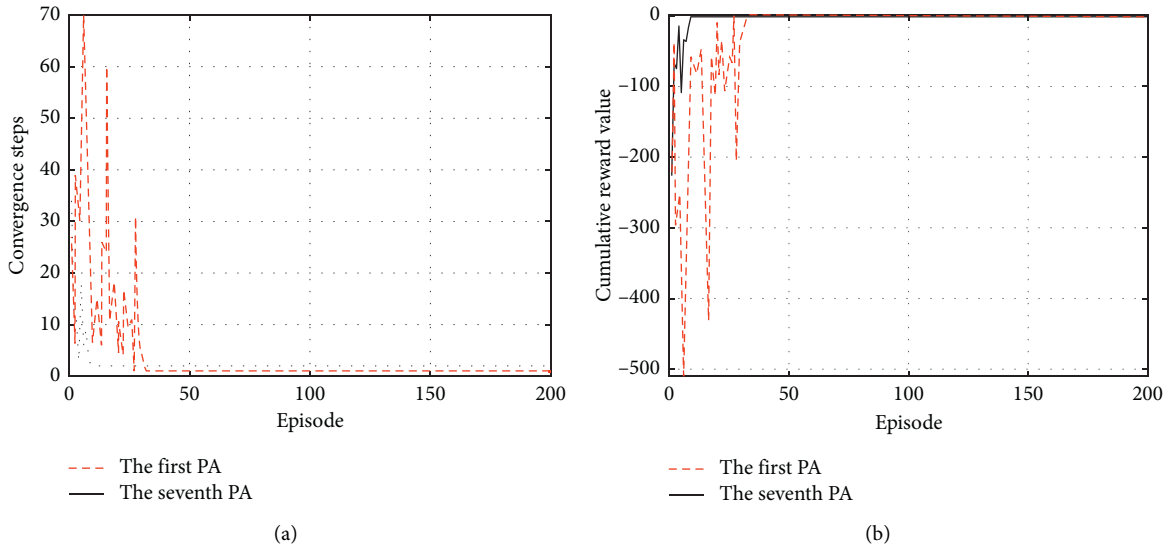

Figure 6: Main results for two times of PA. (a) Number of convergence steps. (b) Cumulative reward value.

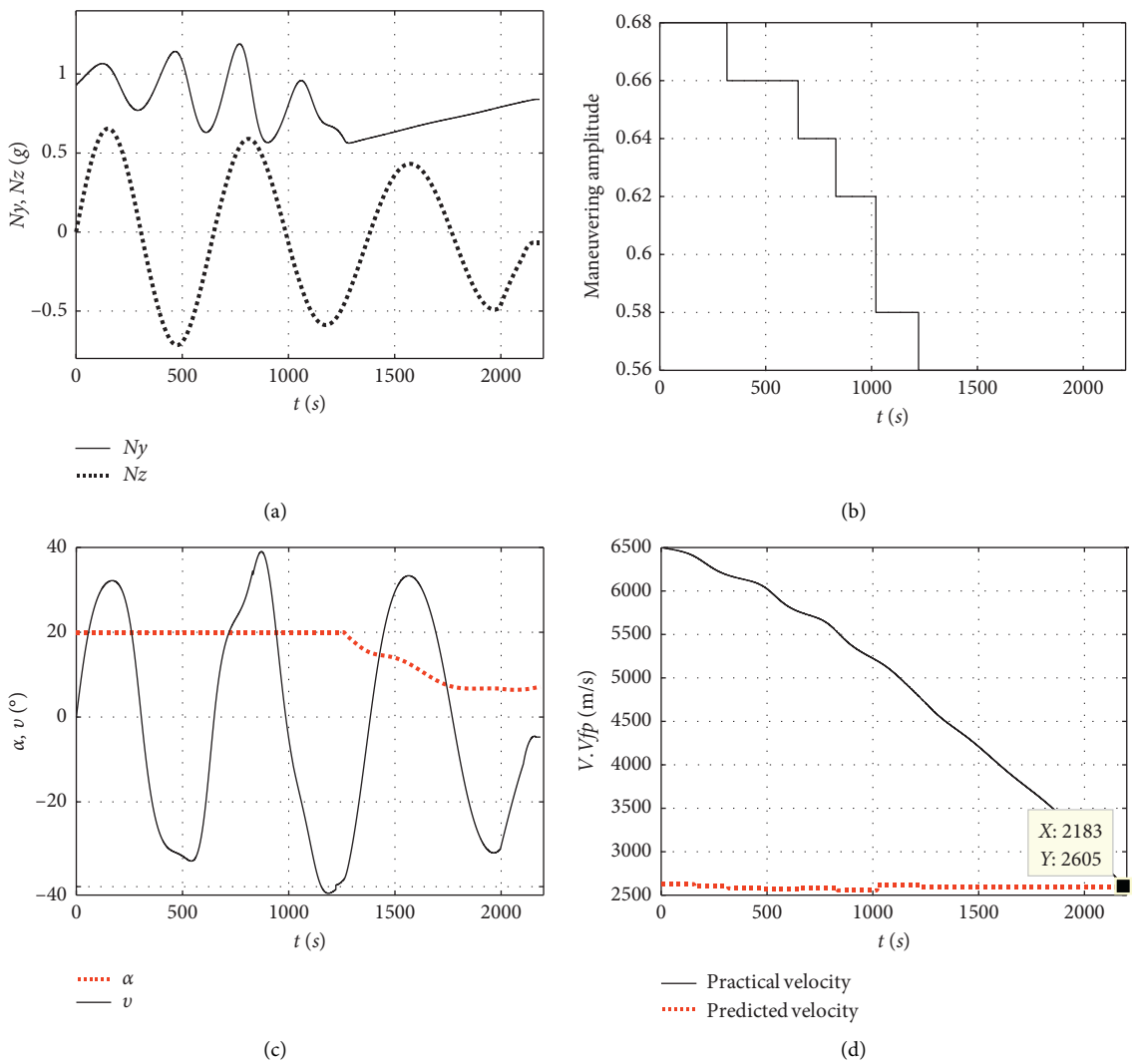

Figure 7: Continued.

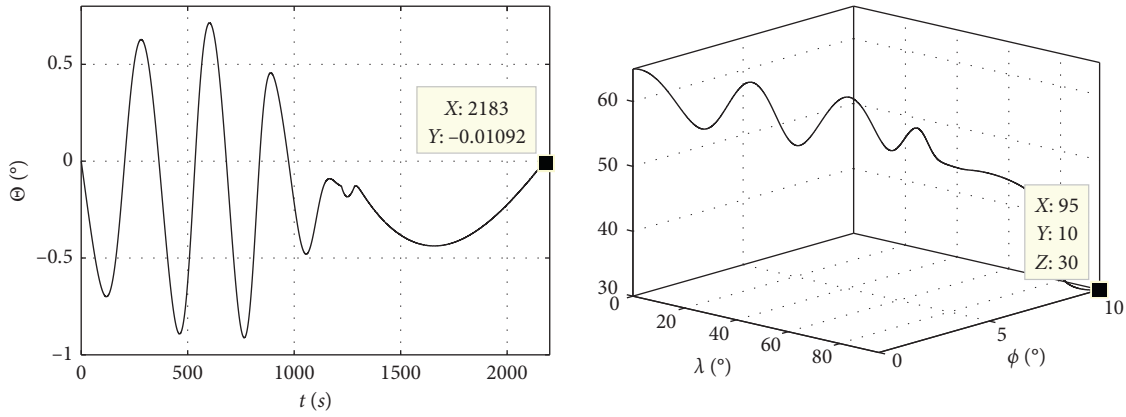

Figure 7: Main ballistic curves of intelligent gliding guidance. (a) Overload commands. (b) Maneuvering amplitude. (c) Control variables. (d) Velocity and terminal predicted velocity. (e) Flight-path angle. (f) Longitude-latitude-altitude.

Table 1: Terminal results under different velocity constraints.

Velocity constraint (m/s)	Position error (m)	Altitude error (m)	FPA (°)	Terminal velocity (m/s)
3200	8.154	2.012	−0.007	3201.013
3000	9.384	1.328	−0.008	3003.108
2800	11.519	1.425	−0.009	2771.499
2600	12.364	0.821	−0.011	2600.766
2400	13.586	0.946	−0.014	2370.207

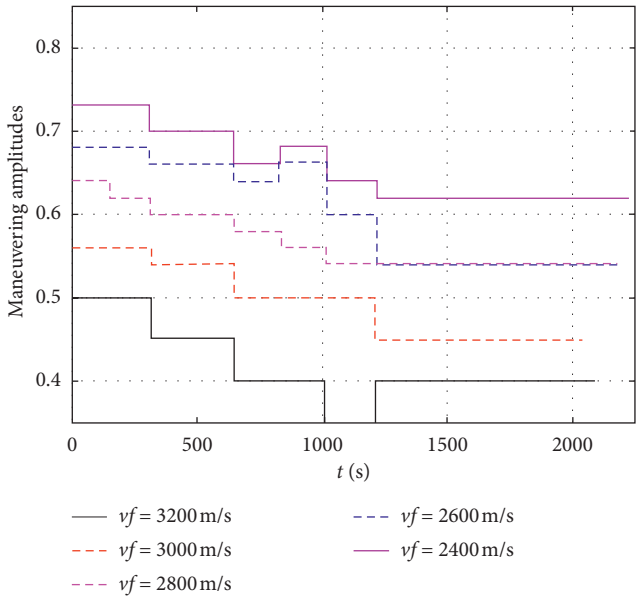

Figure 8: Maneuvering amplitudes under different velocity constraints.

amplitude adaptively according to the terminal constraint value and can still meet different terminal velocity constraints under the premise of ensuring the position, altitude, and FPA constraints. As the terminal velocity decreases, the maneuvering amplitude increases continuously, and the severe maneuvering flight causes the terminal position and FPA errors to increase. As the vehicle approaching the target, the terminal velocity prediction accuracy is improved continuously, and the maneuvering amplitude is continuously optimized and kept constant in the late flight. Comparing the simulation results in Figures 7 and 8, it can be seen that when the terminal velocity constraint is 2600 m/s, different terminal guidance accuracies appear. The major reason is that the random method is employed to initialize Q-table in the first adjustment, which leads to the learning results are different, but they are all beyond the error range of the terminal velocity.

Set the same terminal constraints, use equation (22) to the computer dynamic amplitude analytically (G1), and compare with the intelligent gliding guidance method (G2), and the simulation results are shown in Table 2. It can be known from the simulation results that the terminal velocity error generated by G1 is always larger than 50 m/s, and the terminal velocity error of G2 does not exceed the error range value of 30 m/s. The terminal velocity error of the intelligent method is mainly affected by the error range in Q-learning and prediction error, which means that the terminal velocity control error can be controlled manually. However, due to the influence of the interval between the elements in the action and state space and the consideration of online calculation efficiency, the error range cannot be too small, so as to avoid the failure of online learning.

Table 2: Terminal velocities under actions of G1 and G2.

Velocity constraint (m/s)	G1: terminal velocity (m/s)	G2: terminal velocity (m/s)
3200	3280.637	3211.624
3000	3064.381	2996.674
2800	2745.628	2794.348
2600	2534.329	2621.943
2400	2315.061	2419.815

7. Conclusions

This paper studies an intelligent gliding guidance method based on optimal control, predictor-corrector, and RL. Firstly, the optimal gliding guidance method is introduced to satisfy the constraints of latitude and longitude, altitude, and FPA. Secondly, aiming at the terminal velocity constraint, the velocity control strategy based on lateral maneuver is proposed, and the terminal velocity is predicted considering the gliding flight characteristics and lateral maneuver. Finally, it constructs the frame model of RL, state space, action space, and reward function, Q-learning is used to adjust the maneuvering amplitude intelligently to ensure the accuracy of terminal velocity control. Compared with the traditional standard trajectory tracking and predictor-corrector guidance methods, the main advantages of intelligent gliding guidance are as follows:

(1) Intelligent guidance does not depend on the standard trajectory and can obtain guidance commands in real time according to the current flight states and guidance mission. It has great flexibility and can complete different guidance tasks without adjusting the guidance parameters manually.

(2) RL does not need to establish the sample database and has high calculation efficiency, so it is suitable for the flight control that cannot obtain a large amount of actual flight data and has a high demand for real-time performance.

(3) Both the optimal gliding guidance and terminal velocity prediction use the analytical form to achieve the guidance target, Q-learning has high calculation efficiency and can inherit the excellent experience of previous study. Therefore, the calculation amount of this strategy is small and easy to realize in engineering.

Data Availability

The data used to support the findings of this study are available from the corresponding author upon request.

Conflicts of Interest

The authors declare that they have no conflicts of interest.

Acknowledgments

This research was funded by National Science Basic Research Program of Shaanxi, China (Grant no. 2020JQ-491).

References

[1] L. Zang, D. Lin, S. Chen, H. Wang, and Y. Ji, "An on-line guidance algorithm for high L/D hypersonic reentry vehicles," *Aerospace Science and Technology*, vol. 89, pp. 150–162, 2019.

[2] Y.-L. Zhang, K.-J. Chen, L.-H. Liu, G.-J. Tang, and W.-M. Bao, "Entry trajectory planning based on three-dimensional acceleration profile guidance," *Aerospace Science and Technology*, vol. 48, pp. 131–139, 2016.

[3] H. P. Lahanier and L. Serre, "Trajectory and guidance scheme design for free flight test of hypersonic vehicle," in *Proceedings of the AIAA Guidance, Navigation, and Control Conference, AIAA 2017-2197*, Xiamen, China, March 2017.

[4] A. Joshi, K. Sivan, and S. S. Amma, "Predictor-corrector reentry guidance algorithm with path constraints for atmospheric entry vehicles," *Journal of Guidance, Control, and Dynamics*, vol. 30, no. 5, pp. 1307–1318, 2007.

[5] P. Lu, "Predictor-corrector entry guidance for low-lifting vehicles," *Journal of Guidance, Control, and Dynamics*, vol. 31, no. 4, pp. 1067–1075, 2008.

[6] I. Rusnak, H. Weiss, and G. Hexner, "Optimal guidance laws with prescribed degree of stability," *Aerospace Science and Technology*, vol. 99, no. 5, p. 105780, 2020.

[7] J. Zhu, L. Liu, G. Tang, and W. Bao, "Highly constrained optimal gliding guidance," *Proceedings of the Institution of Mechanical Engineers, Part G: Journal of Aerospace Engineering*, vol. 229, no. 12, pp. 2321–2335, 2015.

[8] J. Zhu and S. Zhang, "Adaptive optimal gliding guidance independent of QEGC," *Aerospace Science and Technology*, vol. 71, pp. 373–381, 2017.

[9] B. Prashant, K. Faruk, and S. Navdeep, "Reinforcement learning based obstacle avoidance for autonomous underwater vehicle," *Journal of Marine Science and Application*, vol. 18, no. 2, pp. 228–238, 2019.

[10] X. Yan, J. Zhu, M. Kuang, and X. Wang, "Aerodynamic shape optimization using a novel optimizer based on machine learning techniques," *Aerospace Science and Technology*, vol. 86, pp. 826–835, 2019.

[11] Y.-H. Wu, Z.-C. Yu, C.-Y. Li, M.-J. He, B. Hua, and Z.-M. Chen, "Reinforcement learning in dual-arm trajectory planning for a free-floating space robot," *Aerospace Science and Technology*, vol. 98, no. 1, p. 105657, 2020.

[12] J. Junell, E. J. Kampeny, C. D. Visser, and Q. Chu, "Reinforcement learning applied to a quadrotor guidance law in autonomous flight," in *Proceedings of the AIAA Guidance, Navigation, and Control Conference AIAA 2015-1990*, Kissimmee, FL, USA, January 2015.

[13] B. Gaudeta and R. Furfaroa, "Missile homing-phase guidance law design using reinforcement learning," in *Proceedings of the AIAA Guidance, Navigation, and Control Conference AIAA 2012-4470*, Minneapolis, Min, USA, August 2012.

[14] B. Gaudeta, R. Furfaroa, and R. Linares, "Reinforcement meta-learning for angle-only intercept guidance of

maneuvering targets," in *Proceedings of the AIAA Scitech 2020 Forum AIAA 2020-0609*, Orlando, FL, USA, January 2020.

[15] B. Gaudet, R. Furfaro, and R. Linares, "Reinforcement learning for angle-only intercept guidance of maneuvering targets," *Aerospace Science and Technology*, vol. 99, p. 105746, 2020.

[16] T. Woodbury, C. Dunny, and J. Valasek, "Autonomous soaring using reinforcement learning for trajectory generation," in *Proceedings of the 52nd Aerospace Sciences Meeting*, National Harbor, MD, USA, January 2014.

[17] Z. Wang, H. Li, H. Wu, F. Shen, and R. Lu, "Design of agent training environment for aircraft landing guidance based on deep reinforcement learning," in *Proceedings of the 2018 11th International Symposium on Computational Intelligence and Design*, pp. 76–79, Hangzhou, China, December 2018.

[18] J. Yang, X. You, G. Wu, M. M. Hassan, A. Almogren, and J. Guna, "Application of reinforcement learning in UAV cluster task scheduling," *Future Generation Computer Systems*, vol. 95, no. 11, pp. 140–148, 2019.

[19] T. H. Phillips, *A Common Aero Vehicle (CAV) Model, Description, and Employ-ment Guide*, Schafer Corporation for AFRL and AFSPC, Arlington, VA, USA, 2003.