Received: Added at production | Revised: Added at production | Accepted: Added at production

DOI: xxx/xxxx

ARTICLE TYPE

Mars Powered Descent Phase Guidance Law Based on Reinforcement Learning for Collision Avoidance †

Yao Zhang[1] | Tianyi Zeng*[2] | Yanning Guo[3] | Guangfu Ma[3]

[1]School of Engineering , University of Southampton, Hampshire, U.K.
[2]Rolls-Royce UTC in Manufacturing and On-wing Technology, University of Nottingham, Nottinghamshire, U.K.
[3]Department of Control Science and Engineering, Harbin Institute of Technology, Harbin, China

Correspondence
*Tianyi Zeng, with Rolls-Royce UTC in Manufacturing and On-wing Technology, University of Nottingham, NG8 1BB, U.K.
Email: tianyi.zeng@nottingham.ac.uk

Summary

This paper proposes a reinforcement learning-based guidance law for Mars powered descent phase, which is an effective online calculation method that handles the nonlinearity caused by the mass variation and avoids collisions. The reinforcement learning method is designed to solve the constrained nonlinear optimization problem by using a critic neural network. Specifically, to cope with the position constraint (i.e. glide-slope constraint) and the thrust force limit constraint, a modified cost function is proposed, and the associated Hamilton-Jacobi-Bellman equation is solved online without using an actor neural network, which significantly reduces the computational burden. The convergence of the critic neural network is proven. Simulation results show the effectiveness of the proposed method.

KEYWORDS:
Collision Avoidance, Powered Descent Phase, Deep Space Exploration, Constraints

1 | INTRODUCTION

As one of the most important space activities, deep space exploration has attracted the attention from various countries all over the world. Particularly, exploring Mars helps scientists learn about momentous shifts in climate that can fundamentally alter planets. The development of earth independence extends human presence beyond low Earth orbit and cislunar space and onto Mars. Missions during this stage of exploration range from 2-3 years with safe return of the crew to Earth taking months. Several missions have been undertaken recently, such as 'Tianwen-1' [1] and 'Perseverance' [2]. Landing robotic spacecraft, and possibly someday humans, on Mars is a technological challenge as of multiple attempted Mars landings by robotic and uncrewed spacecraft, only ten have had successful soft landings. It therefore has been recognized that soft and precision landing on Mars is the solid foundation of Mars exploration [3].

Entry, Descent, and Landing, referred to as EDL, is the shortest and most intense phase of Mars landing missions. Before the powered descent phase, the rover enters the Martian atmosphere during which the guidance is designed to track a pre-optimized trajectory. There are various guidance laws of Mars entry with Low-L/D vehicle or earth reentry [4,5,6]. These guidance laws have high reference values for the Mid-L/D vehicle, and they can be divided into three categories: trajectory planning, predictor-corrector and tracking guidance.

As the last stage of the EDL process, the powered descent phase takes a vital role in ensuring a safe and precise landing on the Martian surface. One of the major issues faced by a Mars landing rover is unforeseen obstacles in this flight trajectory and therefore, a guidance design that takes collision avoidance into account is crucial for a safe landing. This will benefit the scope of exploration on the Martian surface by overcoming limitations in landing caused due to uneven terrains in the flight path. For the powered descent phase guidance, guidance designed for lunar landing or other planetary landing missions contribute

†This work is supported by National Natural Science Foundation of China under Grant 62203219, 61973100, 61876050.

to the design for Mars landing problems[7,8]. Back in the 1970s, Apollo polynomial guidance was proposed for lunar landing[9], inspired by which gravity turn guidance[10] was investigated for Mars landing. Multi-stages guidance[11] and feedback guidance[12] were proposed for the powered descent phase in Mars landing missions. With the development of deep space exploration, the focus has shifted from the simple touch-down on the planetary surface to the optimal or sub-optimal guidance design with high reliability of avoiding collisions on complex terrains.

Various research on the powered descent phase guidance design to avoid collisions has been undertaken. Zero-Effort-Miss/Zero-Effort-Velocity(ZEM/ZEV) guidance was designed to achieve an energy optimal solution[13], based on which a modified ZEM/ZEV guidance was proposed to ensure no collisions between the lander and martian surface occur[14]. By considering a virtual-terminal velocity, a two-phase ZEM/ZEV guidance was designed to avoid the collision[15]. Considering the collision avoidance requirement as a constraint on the lander position (such as glide-slope constraints), model predictive control (MPC) has been utilized to generate the optimal guidance by solving a constrained optimal control problem[16]. MPC was also adopted to cope with the rectangle constraint for some flat landing terrains[17]. Other approaches to avoid collisions are to limit the terminal landing error within a small range so that the lander is kept close to the pre-select landing site subject to unknown disturbances, such as sliding mode control-based guidance design[18] and minimum-landing-error guidance design[19]. Due to the long disturbance between Mars and the Earth, the fuel loaded and consumption has attracted attention in the meantime. Some guidance design focused on minimizing the fuel consumption[20,21,22], which led to a non-convex optimization problem and online calculation requirements.

The challenges of designing optimal guidance for the powered descent phase can be concluded as follows

- the mass variation is significant so the dynamics is considered as a nonlinear model;
- the unknown disturbances including dust storms and wind acting on the lander prevent the lander from landing on a precise site;
- the constraints on both the position and the thrust limit should also be considered, which brings challenges to the online optimization.

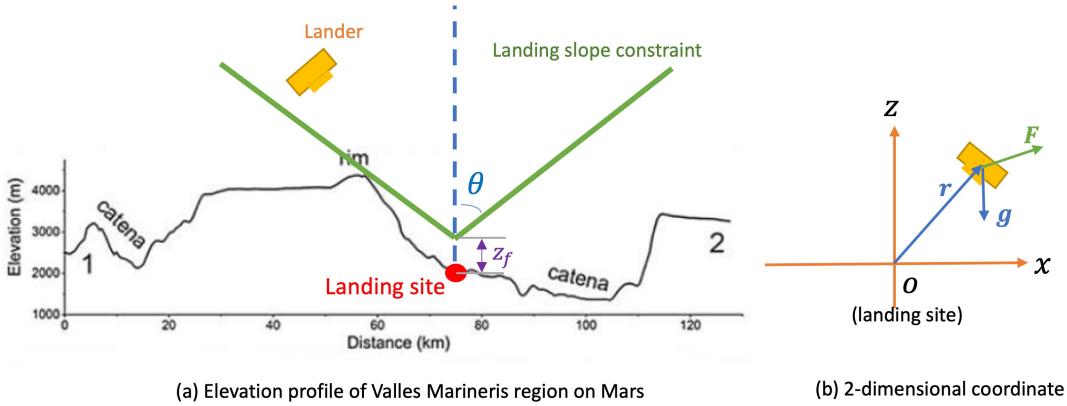
Due to the ability to cope with multiple constraints and the nonlinearity of the system, reinforcement learning, such as approximate dynamic programming[23,24], has great potential to handle the above challenges simultaneously. It has been recognized that intelligent algorithms make contributions to spacecraft guidance and control design[25]. Reinforcement learning-based guidance design for the powered descent phase was proposed in[26], with a calculation algorithm to generate an optimal control signal. However, the unknown disturbance and calculation burden are remained to be further investigated.

In this paper, we focus on developing novel guidance that ensures high reliability in avoiding collisions and strong robustness against unknown disturbances. A reinforcement learning method is designed by using a critic neural network to estimate the optimal cost function and therefore obtain the optimal control input. Meanwhile, the constraints on both the position and the thrust limit are fully considered during the whole landing process. Unlike the widely used reinforcement learning method in which an actor neural network (NN) is also used along with the critic NN, this paper only utilizes a critic NN to estimate the optimal solution whose NN weights are calculated by an adaptive law, which significantly reduces the computational burden and is attractive for the practical online implementation. The proposed guidance can be straightforwardly applied to other planetary landing missions.

The rest of the paper is organized as follows. In Section II, the mathematical model of the lander during the power descent phase is introduced, and physical constraints and control objectives are both analyzed. Section III proposes the reinforcement learning-based guidance for the powered descent phase. The simulation results are shown in Section IV. Finally, Section V concludes the paper.

2 | MATHEMATICAL MODEL AND CONTROL PROBLEM

In this section, the mathematical model of the lander dynamics during the powered descent phase is introduced indicating the nonlinearities caused by the significant mass variation and martian surface disturbances. To achieve a precision and safe landing, several physical constraints during the powered descent phase that should be considered in guidance design are analyzed.



(a) Elevation profile of Valles Marineris region on Mars

(b) 2-dimensional coordinate

FIGURE 1 Terrain data[27] and coordinate: (a) Elevation profile of Valles Marineris region on Mars; (b) Two-dimensional coordinate

2.1 | Mathematical model

Choose the pre-selected landing site as the origin of the coordinate, as shown in Fig. 1. To avoid collisions between the lander and the martian surface, the desired landing site is set to be above the pre-selected landing site, so the desired landing site is denoted by $\boldsymbol{r}^* = [0, z_f]^{\mathrm{T}}$ with z_f as the terminal altitude in the powered descent phase. The terminal altitude is preset to fit the missions, which is approximately within the range from 5 meters to 20 meters above the ground.

During the powered descent phase, the dynamical model of the lander is as follows

$$\begin{cases} \dot{\boldsymbol{r}}(t) &= \boldsymbol{v}(t) \\ \dot{\boldsymbol{v}}(t) &= \boldsymbol{a}(t) + \boldsymbol{g} + \boldsymbol{d}(t) \\ \boldsymbol{F}(t) &= m(t)\boldsymbol{a}(t) \\ \dot{m}(t) &= \frac{\|\boldsymbol{F}(t)\|}{c} \end{cases} \tag{1}$$

where $\boldsymbol{r} = [r_x, r_z]^{\mathrm{T}} \in \mathbb{R}^2$ is the lander position vector with r_x and r_z as components along x-axis and z-axis, respectively; $\boldsymbol{v} = [v_x, v_z]^{\mathrm{T}} \in \mathbb{R}^2$ is the lander velocity vector with v_x and v_z as components along x-axis and z-axis, respectively; $\boldsymbol{a} = [a_x, a_z]^{\mathrm{T}} \in \mathbb{R}^2$ is the thrust acceleration vector with a_x and a_z as components along x-axis and z-axis, respectively; $\boldsymbol{F} = [F_x, F_z]^{\mathrm{T}} \in \mathbb{R}^2$ is the thrust force vector with F_x and F_z as components along x-axis and z-axis, respectively, with $\|\boldsymbol{F}\| = \sqrt{F_x^2 + F_z^2}$, and $\boldsymbol{F}$ is also the control command to be designed in this paper; $\boldsymbol{d} = [d_x, d_z]^{\mathrm{T}} \in \mathbb{R}^2$ is the disturbance acceleration vector caused by wind and dust storms with d_x and d_z as components along x-axis and z-axis, respectively; m is the mass of the lander which is time-varying; c is a positive constant parameter calculated as $c = g_e I_{sp}$ with $g_e = 9.807$ as the gravitational acceleration of the Earth at sea level and I_{sp} as the specific impulse of the thruster; $\boldsymbol{g} = [0, -3.7114]^{\mathrm{T}}$ is the martian gravitational acceleration.

Define the state vector as $\boldsymbol{x} = [x_1, x_2, x_3, x_4]^{\mathrm{T}} = [r_x, r_z, v_x, v_z]^{\mathrm{T}}$, the control input vector as $\boldsymbol{u} = [u_1, u_2]^{\mathrm{T}} = [F_x, F_z]^{\mathrm{T}}$, and the disturbance vector as $\boldsymbol{w} = [0, 0, d_x, d_z + g]^{\mathrm{T}}$, we have the state-space model of (1) as follows

$$\dot{\boldsymbol{x}}(t) = \boldsymbol{A}\boldsymbol{x}(t) + \boldsymbol{B}(t)\boldsymbol{u}(t) + \boldsymbol{w}(t) \tag{2}$$

where

$$\boldsymbol{A} = \begin{bmatrix} \boldsymbol{0}_2 & \boldsymbol{I}_2 \\ \boldsymbol{0}_2 & \boldsymbol{0}_2 \end{bmatrix}, \boldsymbol{B}(t) = \frac{1}{m(t)} \begin{bmatrix} \boldsymbol{0}_2 \\ \boldsymbol{I}_2 \end{bmatrix} \tag{3}$$

and (2) is a perturbed time-varying model. Note that due to the external disturbance $\boldsymbol{w}(t)$ and the time-varying system matrix $\boldsymbol{B}(t)$, the model (2) behaves as a nonlinear system.

2.2 | Physical constraints

To avoid collisions during whole landing mission, the landing glide-slope constraint in the powered descent phase shown in Fig. 1 is considered, where $\theta > 0$ is a constant representing the glide-slope gradient designed based on the specific terrain around

the target landing site. The smaller value of θ is chosen, the stronger constraints on the state is formed. This leads to position constraints as follows

$$|r_x| \leq \tan\theta(r_z - z_f) \tag{4}$$

$$r_z \geq z_f \tag{5}$$

where the correlation of (4) and (5) is ensured by (10).

Due to the thrust limitation, the control input constraint along two axes should be considered as follows

$$|F_x| \leq F_{x,max} \tag{6}$$

$$|F_z| \leq F_{z,max} \tag{7}$$

where $F_{x,max}$ and $F_{z,max}$ are positive constants representing the maximal thrust force that can be provided by the actuators.

2.3 | Control objective

For a fixed fight time denoted by t_f, the control objective is to solve the following constrained optimization problem

$$\min_{\boldsymbol{u}(t)} \|\boldsymbol{r}(t_f)\|^2 \tag{8}$$

subject to

$$\begin{aligned}
&\dot{\boldsymbol{x}}(t) = \boldsymbol{A}\boldsymbol{x}(t) + \boldsymbol{B}(t)\boldsymbol{u}(t) + \boldsymbol{w}(t)\\
&|x_1| \leq \tan\theta(x_2 - z_f), x_2 \geq z_f, |u_1| \leq F_{x,max}, |u_2| \leq F_{z,max} \quad \forall t \in [0, t_f)\\
&\dot{\boldsymbol{r}}(t_f) = 0
\end{aligned}$$

To explicitly cope with state constraints and input constraints, the cost function (8) is reformulated as

$$V(\boldsymbol{x}, t) = \|\boldsymbol{r}(t_f)\|^2 + \int_t^{t_f} \left(X(\boldsymbol{x}(\tau)) + U(\boldsymbol{u}(\tau))\right) d\tau \tag{9}$$

where $X(\boldsymbol{x}(\tau))$ is to cope with the state constraints, and $U(\boldsymbol{u}(\tau))$ is to cope with the input constraints. Design the constraint-handling terms in the following forms

$$X(\boldsymbol{x}(\tau)) = \frac{\sigma_1}{\tan\theta(x_2 - z_f) - |x_1(\tau)|} + \frac{\sigma_2}{x_2(\tau) - z_f} \tag{10}$$

and

$$U(\boldsymbol{u}(\tau)) = 2\int_0^{u_1} F_{x,max}\tanh^{-1}\left(\frac{v_1}{F_{x,max}}\right) R_1 dv_1 + 2\int_0^{u_2} F_{z,max}\tanh^{-1}\left(\frac{v_2}{F_{z,max}}\right) R_2 dv_2 \tag{11}$$

where $\sigma_1 > 0$ and $\sigma_2 > 0$ are arbitrarily small constants, and $\tanh(\bullet)$ is the hyperbolic function, R_1 and R_2 are positive weights, which are to be tuned to achieve a good trade-off between control efforts and state regulation.

To minimize the reformulated cost function (9), the state constraints (4) and (5) are satisfied, because from (10), $\frac{\sigma_1}{\tan\theta(x_2 - z_f) - |x_1(\tau)|} \to +\infty$ holds if $|x_1(\tau)| \to \tan\theta(x_2 - z_f)$, and $\frac{\sigma_2}{x_2(\tau) - z_f} \to +\infty$ holds if $x_2(\tau) \to z_f$.

3 | OPTIMAL GUIDANCE DESIGN BASED ON REINFORCEMENT LEARNING

In this section, the optimal guidance is firstly formulated, and the associated Hamilton-Jacobi-Bellman (HJB) equation is derived and found to be time-varying and nonlinear. Then a reinforcement learning method is proposed to solve the HJB equation by using a critic NN to estimate the optimal cost function. The convergence of the estimation error of NN weights is proven.

3.1 | Optimal Guidance Design for Constrained Landing Problem

The Hamiltonian function for (9) is in the following form

$$H(\boldsymbol{x},\boldsymbol{u},V,t)=\frac{\partial V}{\partial \boldsymbol{x}^{\mathrm{T}}}(\boldsymbol{A}\boldsymbol{x}+\boldsymbol{B}\boldsymbol{u}+\boldsymbol{w})+\frac{\sigma_1}{\tan\theta(x_2-z_f)-|x_1|}+\frac{\sigma_2}{x_2-z_f}+U(\boldsymbol{u}) \tag{12}$$

Define $\boldsymbol{u}^*=[u_1^*,u_2^*]^{\mathrm{T}}$ as the optimal control input. From (9), we have the following optimal cost function

$$V^*(\boldsymbol{x},t)=\min_{\boldsymbol{u}}\|\boldsymbol{r}(t_f)\|^2+\min_{\boldsymbol{u}}\int_t^{t_f}(X+U(\boldsymbol{u}^*))\,d\tau \tag{13}$$

The optimal control input $\boldsymbol{u}^*$ satisfies the HJB equation as follows

$$\begin{aligned}-\frac{\partial V^*}{\partial t}&=\min_u H(\boldsymbol{x},\boldsymbol{u}^*,V^*,t)\\&=\frac{\partial V^*}{\partial \boldsymbol{x}^{\mathrm{T}}}(\boldsymbol{A}\boldsymbol{x}+\boldsymbol{B}\boldsymbol{u}^*+\boldsymbol{w})+X+2\int_0^{u_1^*}F_{x,max}\tanh^{-1}\left(\frac{v_1}{F_{x,max}}\right)R_1dv_1+2\int_0^{u_2^*}F_{z,max}\tanh^{-1}\left(\frac{v_2}{F_{z,max}}\right)R_2dv_2\end{aligned} \tag{14}$$

The optimal control input $\boldsymbol{u}^*=[u_1^*,u_2^*]^{\mathrm{T}}$ can be obtained[28] by solving $\frac{\partial H(\boldsymbol{x},\boldsymbol{u}^*,V^*)}{\partial \boldsymbol{u}^*}=0$, which yields

$$\left(\frac{\partial V^*}{\partial \boldsymbol{x}^{\mathrm{T}}}B\right)^{\mathrm{T}}+\begin{bmatrix}2R_1F_{x,max}\tanh^{-1}(u_1/F_{x,max})\\2R_2F_{z,max}\tanh^{-1}(u_2/F_{z,max})\end{bmatrix}=\begin{bmatrix}0\\0\end{bmatrix}$$

which gives

$$B^{\mathrm{T}}\frac{\partial V^*}{\partial \boldsymbol{x}}+\begin{bmatrix}2R_1F_{x,max}&0\\0&2R_2F_{z,max}\end{bmatrix}\tanh^{-1}\begin{bmatrix}u_1/F_{x,max}\\u_2/F_{z,max}\end{bmatrix}=\begin{bmatrix}0\\0\end{bmatrix}$$

Therefore, we have

$$\begin{aligned}\boldsymbol{u}^*&=-\begin{bmatrix}F_{x,max}&0\\0&F_{z,max}\end{bmatrix}\tanh\left(\begin{bmatrix}\frac{1}{2R_1F_{x,max}}&0\\0&\frac{1}{2R_2F_{z,max}}\end{bmatrix}\boldsymbol{B}^{\mathrm{T}}\frac{\partial V^*}{\partial \boldsymbol{x}}\right)\\&=-\begin{bmatrix}F_{x,max}&0\\0&F_{z,max}\end{bmatrix}\tanh(\boldsymbol{\phi})\end{aligned} \tag{15}$$

where $\boldsymbol{\phi}=\begin{bmatrix}\frac{1}{2R_1F_{x,max}}&0\\0&\frac{1}{2R_2F_{z,max}}\end{bmatrix}\boldsymbol{B}^{\mathrm{T}}\frac{\partial V^*}{\partial \boldsymbol{x}}$, and for a vector $\boldsymbol{\phi}=[\phi_1,\phi_2]^{\mathrm{T}}$, the function $\tanh(\boldsymbol{\phi})$ is defined as $\tanh(\boldsymbol{\phi})=[\tanh(\phi_1),\tanh(\phi_2)]^{\mathrm{T}}$.

From (15), it is clear that due to the property of hyperbolic function, the input constraints (6) and (7) are satisfied for all $t>0$.

Substituting (15) into (11) gives

$$\begin{aligned}U(\boldsymbol{u}^*)=&2\int_0^{u_1^*}F_{x,max}\tanh^{-1}\left(\frac{v_1}{F_{x,max}}\right)R_1dv_1+2\int_0^{u_2^*}F_{z,max}\tanh^{-1}\left(\frac{v_2}{F_{z,max}}\right)R_2dv_2\\=&2F_{x,max}\tanh^{-1}(u_1^*/F_{x,max})R_1u_1^*+F_{x,max}^2R_1\ln\left(1-(u_1^*/F_{x,max})^2\right)\\&+2F_{z,max}\tanh^{-1}(u_2^*/F_{z,max})R_2u_2^*+F_{z,max}^2R_2\ln\left(1-(u_2^*/F_{z,max})^2\right)\\=&\left[\tanh(\phi_1)\ \tanh(\phi_2)\right]\begin{bmatrix}F_{x,max}&0\\0&F_{z,max}\end{bmatrix}\boldsymbol{B}^{\mathrm{T}}\frac{\partial V^*}{\partial \boldsymbol{x}}+\left[\ln(1-\tanh^2(\phi_1))\ \ln(1-\tanh^2(\phi_2))\right]\begin{bmatrix}F_{x,max}^2R_1\\F_{z,max}^2R_2\end{bmatrix}\end{aligned} \tag{16}$$

To rewrite the HJB equation (14) as a function of optimal cost function V^* only, we substitute (16) into (14), which gives

$$-\frac{\partial V^*}{\partial t}=\frac{\partial V^*}{\partial \boldsymbol{x}^{\mathrm{T}}}(\boldsymbol{A}\boldsymbol{x}+\boldsymbol{w})+\frac{\sigma_1}{\tan\theta(x_2-z_f)-|x_1(\tau)|}+\frac{\sigma_2}{x_2(\tau)-z_f}+\left[\ln(1-\tanh^2(\phi_1))\ \ln(1-\tanh^2(\phi_2))\right]\begin{bmatrix}F_{x,max}^2R_1\\F_{z,max}^2R_2\end{bmatrix} \tag{17}$$

It is obvious that the optimal cost V^* can be obtained by solving HJB equation (17), and then the optimal control input can be obtained by (15). But it is not possible to get its analytical solution, since the HJB equation is time-varying and nonlinear. In the next subsection, a reinforcement learning method is utilized to solve the HJB equation online, where the optimal cost function $V^*(\boldsymbol{x},t)$ is estimated and approximated by a critic NN.

3.2 | Reinforcement Learning for Optimal Guidance Implementation

The critic NN is designed in the following time-varying regressor form

$$V^*(\boldsymbol{x},t)=\boldsymbol{W}^{\mathrm{T}}\boldsymbol{\psi}(\boldsymbol{x},t_{go})+\epsilon \tag{18}$$

where $t_{go}=t_f-t$ is the time-to-go, $\boldsymbol{W}\in\mathbb{R}^l$ denotes a constant NN weight, $\boldsymbol{\psi}(\boldsymbol{x},t_{go})=[\psi_1(\boldsymbol{x},t_{go}),\cdots,\psi_l(\boldsymbol{x},t_{go})]^{\mathrm{T}}\in\mathbb{R}^l$ is the regressor with l is the number of neurons, and $\epsilon\in\mathbb{R}$ is the residual NN error.

The partial derivative of (18) with respect to $\boldsymbol{x}$ [29] is

$$\frac{\partial V^*(\boldsymbol{x},t)}{\partial \boldsymbol{x}}=\frac{\partial \boldsymbol{\psi}^{\mathrm{T}}(\boldsymbol{x},t_{go})}{\partial \boldsymbol{x}}\boldsymbol{W}+\frac{\partial \epsilon}{\partial \boldsymbol{x}} \tag{19}$$

The partial derivative of (18) with respect to t [29] is

$$\frac{\partial V^*(\boldsymbol{x},t)}{\partial t}=\frac{\partial \boldsymbol{\psi}^{\mathrm{T}}(\boldsymbol{x},t_{go})}{\partial t}\boldsymbol{W}+\frac{\partial \epsilon}{\partial t} \tag{20}$$

Assumption 1. The NN weight $\boldsymbol{W}$ is bounded by $\|\boldsymbol{W}\|\le W_n$, the regressor $\boldsymbol{\psi}$ is bounded by $\|\boldsymbol{\psi}\|\le\psi_n$, and the partial derivatives of $\boldsymbol{\psi}$ with respect to $\boldsymbol{x}$ and t are respectively bounded by $\|\frac{\partial V^*(\boldsymbol{x},t)}{\partial \boldsymbol{x}^{\mathrm{T}}}\|\le\psi_x$ and $\|\frac{\partial V^*(\boldsymbol{x},t)}{\partial t}\|\le\psi_t$, where W_n, ψ_n, ψ_x, and ψ_t are positive constants.

Considering (19), the optimal control input $\boldsymbol{u}^*$ given by (15) can be rewritten as

$$\boldsymbol{u}^*=-\begin{bmatrix}F_{x,max}&0\\0&F_{z,max}\end{bmatrix}\tanh\left(\begin{bmatrix}\frac{1}{2R_1F_{x,max}}&0\\0&\frac{1}{2R_2F_{z,max}}\end{bmatrix}\boldsymbol{B}^{\mathrm{T}}\left(\frac{\partial \boldsymbol{\psi}^{\mathrm{T}}(\boldsymbol{x},t_{go})}{\partial \boldsymbol{x}}\boldsymbol{W}+\frac{\partial \epsilon}{\partial \boldsymbol{x}}\right)\right) \tag{21}$$

Since the NN weight $\boldsymbol{W}$ is an unknown vector, so we need to approximate the optimal solution by estimating the NN weight vector.

Define the estimate of the critical NN $V^*(\boldsymbol{x},t)$ as $\hat{V}(\boldsymbol{x},t)$, and define the estimate of the weight $\boldsymbol{W}$ as $\hat{\boldsymbol{W}}$. So we have the following estimated NN

$$\hat{V}(\boldsymbol{x},t)=\hat{\boldsymbol{W}}^{\mathrm{T}}\boldsymbol{\psi}(\boldsymbol{x},t_{go}) \tag{22}$$

which gives the control input as follows by substituting (22) into (21)

$$\boldsymbol{u}=-\begin{bmatrix}F_{x,max}&0\\0&F_{z,max}\end{bmatrix}\tanh\left(\begin{bmatrix}\frac{1}{2R_1F_{x,max}}&0\\0&\frac{1}{2R_2F_{z,max}}\end{bmatrix}\boldsymbol{B}^{\mathrm{T}}\frac{\partial \boldsymbol{\psi}^{\mathrm{T}}(\boldsymbol{x},t_{go})}{\partial \boldsymbol{x}}\hat{\boldsymbol{W}}\right) \tag{23}$$

Substituting (19) and (20) into the HJB equation (17) yeilds

$$\begin{aligned}-\boldsymbol{W}^{\mathrm{T}}\frac{\partial \boldsymbol{\psi}(t,t_{go})}{\partial t}=&\boldsymbol{W}^{\mathrm{T}}\frac{\partial \boldsymbol{\psi}(\boldsymbol{x},t_{go})}{\partial \boldsymbol{x}}(\boldsymbol{A}\boldsymbol{x}+\boldsymbol{w})+\frac{\sigma_1}{\tan\theta(x_2-z_f)-|x_1(\tau)|}+\frac{\sigma_2}{x_2(\tau)-z_f}\\&+\left[\ln(1-\tanh^2(\phi_1))\ \ln(1-\tanh^2(\phi_2))\right]\begin{bmatrix}F_{x,max}^2R_1\\F_{z,max}^2R_2\end{bmatrix}+\epsilon_H\end{aligned} \tag{24}$$

where $\epsilon_H=\frac{\partial \epsilon}{\partial \boldsymbol{x}^{\mathrm{T}}}(\boldsymbol{A}\boldsymbol{x}+\boldsymbol{w})+\frac{\partial \epsilon}{\partial t}$ is the lumped residual error, which is also bounded.

For simplification purposes, we design two intermediate variables as

$$\Xi:=\frac{\partial \boldsymbol{\psi}(t,t_{go})}{\partial t}+\frac{\partial \boldsymbol{\psi}(\boldsymbol{x},t_{go})}{\partial \boldsymbol{x}}(\boldsymbol{A}\boldsymbol{x}+\boldsymbol{w}) \tag{25}$$

and

$$\Theta:=\frac{\sigma_1}{\tan\theta(x_2-z_f)-|x_1(\tau)|}+\frac{\sigma_2}{x_2(\tau)-z_f}+\left[\ln(1-\tanh^2(\phi_1))\ \ln(1-\tanh^2(\phi_2))\right]\begin{bmatrix}F_{x,max}^2R_1\\F_{z,max}^2R_2\end{bmatrix} \tag{26}$$

so (24) can be rewritten as

$$\Theta=-\boldsymbol{W}^{\mathrm{T}}\Xi-\epsilon_H \tag{27}$$

Construct a new vector as

$$\Omega=\eta_1\hat{\boldsymbol{W}}+\eta_2 \tag{28}$$

where η_1 and η_2 are designed as adaptive parameters with the following adaptive laws

$$\begin{cases}\dot{\eta}_1=&-c_l\eta_1+\Xi\Xi^{\mathrm{T}}\\\dot{\eta}_2=&-c_l\eta_2+\Xi\Theta\end{cases} \tag{29}$$

with $\eta_1(0) = \eta_2(0) = 0$ and $c_l > 0$ to be sufficiently large to make η_1 and η_2 bounded.

Lemma 1. (see[30]) The condition $\lambda_{min}(\boldsymbol{\eta}_1) > \kappa > 0$ holds for a positive constant κ provided that the regressor Ξ is persistently excited (PE).

Define the estimation error of the NN weight as $\tilde{\boldsymbol{W}} = \boldsymbol{W}^* - \hat{\boldsymbol{W}}$. Following the existing result in[31], we can rewrite (28) as

$$\Omega = -\eta_1 \tilde{\boldsymbol{W}} + C \tag{30}$$

where $C = -\int_0^t e^{-c_l(t-r)} \epsilon_H(r) \Xi(r) dr$ is bounded by $\|C\| \leq C_N$ with $C_N > 0$ as a constant.

The critic NN weight is calculated by

$$\dot{\hat{\boldsymbol{W}}} = -\boldsymbol{M}\Omega \tag{31}$$

where $\boldsymbol{M}$ is a positive definite constant matrix.

Theorem 1. Consider the time-varying nonlinear dynamic system (2) and the cost function with multiple constraints (13), the optimal thrust force (i.e. control input) (23), and adaptive law (31), the estimate error of weight $\tilde{\boldsymbol{W}}$ converges to a small region around the zero, and the corresponding control input (23) approximates the optimal control input (21). The estimate error of weight $\tilde{\boldsymbol{W}}$ converges to zero when the residual NN error and its derivatives are zero, i.e. $\epsilon = \partial\epsilon/\partial \boldsymbol{x} = \partial\epsilon/\partial t = 0$.

Proof. A Lyapunov candidate is selected as

$$V = \frac{1}{2} \tilde{\boldsymbol{W}}^{\mathrm{T}} \boldsymbol{M}^{-1} \tilde{\boldsymbol{W}} \tag{32}$$

and its first time derivative is

$$\begin{aligned} \dot{V} =& \tilde{\boldsymbol{W}}^{\mathrm{T}} \boldsymbol{M}^{-1} \dot{\tilde{\boldsymbol{W}}} \\ =& -\tilde{\boldsymbol{W}}^{\mathrm{T}} \boldsymbol{\eta}_1 \tilde{\boldsymbol{W}} + \tilde{\boldsymbol{W}}^{\mathrm{T}} C \\ \leq& -\kappa \|\tilde{\boldsymbol{W}}\|^2 + \|\tilde{\boldsymbol{W}}\| C_N \\ \leq& -\left(\kappa - \frac{1}{2\mu}\right) \|\tilde{\boldsymbol{W}}\|^2 + \frac{\mu}{2} C_N^2 \\ \leq& -K_1 V + K_2 \end{aligned} \tag{33}$$

where $K_1 = 2(\kappa - 1/2\mu)/\lambda(M^{-1})$ and $K_2 = \mu C_N^2/2$ are positive constants for selected μ satisfying $\mu > 1/2\kappa$.

Therefore, it is obvious that the NN weight estimation error $\tilde{\boldsymbol{W}}$ is bounded and converges to the region around zero that can be calculated by $\|\tilde{\boldsymbol{W}}\| \leq \sqrt{2K_2/K_1\lambda_{min}(M^{-1})}$. It can be found that a greater value of NN residual error bound C_N leads to a large size of convergence region. The size of this region is also determined by the NN residual error bound C_N, the adaptive learning gain $\boldsymbol{M}$, and the excitation level κ.

Furthermore, the corresponding control input (23) also converges to a small region around the optimal input (21).

The optimal control input is the solution of minimization of the cost function (9), in which the system state is forced to converge. Therefore, when the optimal control input is achieved, the convergence of the state is guaranteed.

Consider a special case where $\epsilon = \partial\epsilon/\partial \boldsymbol{x} = \partial\epsilon/\partial t = 0$ holds, which means $C = 0$ in (30) holds. Hence, we have

$$\dot{V} = -\tilde{\boldsymbol{W}}^{\mathrm{T}} \boldsymbol{\eta}_1 \tilde{\boldsymbol{W}} < -\kappa \|\tilde{\boldsymbol{W}}\|^2 \leq -\frac{2\kappa}{\lambda_{max}(M^{-1})} V \tag{34}$$

Therefore, in this special case, the estimation error of NN weight exponentially converges to zero, meanwhile, the proposed control input (23) approximates the optimal solution (21). This completes the proof. □

4 | SIMULATION RESULTS

In this section, the effectiveness of the proposed guidance is verified by simulation results. The specific impulse is $I_{sp} = 360$, and the initial mass of the lander is $m(0) = 56000\text{kg}$[1]. The fuel on board is 16000kg. The whole flight time is 50 s. The thrust limit constraints are $F_{x,max} = 2 \times 10^6$ N and $F_{z,max} = 0.5 \times 10^6$ N. The final touch-down height is selected as $z_f = 0.5$ m. The glide-slope constraint shown in Fig.1 is set as $\theta = \frac{\pi}{3}$ rad. The sampling time is 0.1 s. The weight parameters in cost function are set to be $R_1 = 0.5$ and $R_2 = 0.8$. Following the existing literatures[18], we model the external disturbance as sinusoidal functions. The external disturbance is modelled as $d_x = d_z = 0.5\sin(0.2t) + \mathcal{N}$ m/s^2. We have injected zero-mean white Gaussian noise $\mathcal{N} \sim (0, 0.1)\text{m/s}^2$ in the external disturbance to match the realistic scenario.

The regressor in critic NN is chosen as

$$\boldsymbol{\psi}(\boldsymbol{x}, t_{go}) = \left[x_1 \frac{t_{go}}{t_f}, x_2 x_3^2 \frac{t_{go}^3}{t_f^4}, x_4 \frac{t_{go}^4}{t_f^3}, \sin(x_2 x_4) \frac{t_{go}^2}{t_f^2}, x_3 \frac{t_{go}}{t_f^2}, x_4 \frac{t_{go}}{t_f}, \sin(x_1)\sin(x_4) x_2 x_4 x_3 t_{go}, x_2 \frac{t_{go}^2}{t_f^2}\right]^{\mathrm{T}}$$

with the initial estimate as $\hat{\boldsymbol{W}}(0) = [0, 0, 0, 0, 0, 0, 0, 0]^{\mathrm{T}}$.

The simulation of the proposed guidance is run on a PC with Intel(R) Core i5 CPU @ 2.70Ghz, 8.00 GB memory, 64-bit OS. PC running time is 4 minutes. The convergence trajectories of NN weight are shown in Fig. 2, in which it can be found that the estimate of the 8 NN weights converge within 13 s, and this demonstrates the effectiveness of the proposed adaption method for ensuring the convergence of NN estimate as proposed in Theorem 1.

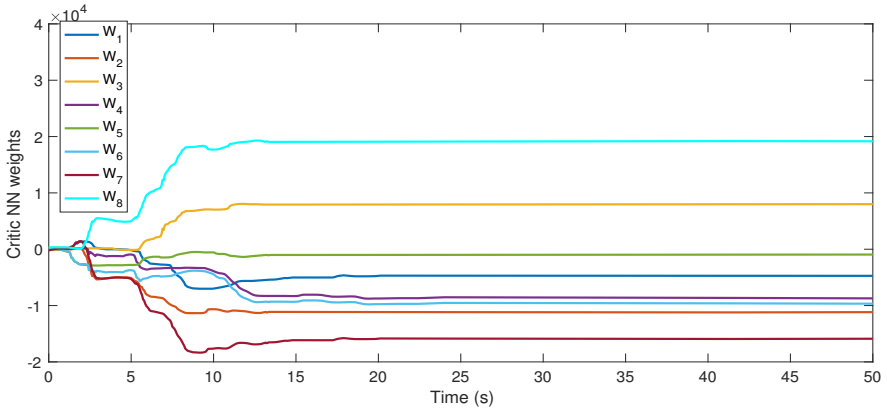


FIGURE 2 Critic NN weights convergence trajectories

4.1 | Case 1: Multiple initial positions with variation along x-axis

The initial position along the z-axis is chosen as $r_z(0) = 1500$ m, and a range of the initial positions along the x-axis are chosen as $r_x(0) = -1800 : 200 : 2000$ m. The initial velocity is chosen as $\boldsymbol{v}(0) = [100, -75]^{\mathrm{T}}$ m/s.

The landing trajectories are shown in Fig. 3, in which it can be seen that with different initial positions, the lander achieves precision landings in all cases and the glide-slope constraint is satisfied even when the initial position is close to the bound with an initial velocity towards the bound (see the light blue line on the far right). From the velocity trajectories in Fig. 4, it can be seen that the terminal velocity along the x-axis is less than 3 m/s, which can be classified as a soft landing. The thrust force is demonstrated in Fig. 5, in which it can be found that the input limitation is satisfied all the time, and particularly, at the beginning when the lander is relatively far away from the landing site, the input constraint is active (see the period from 0 s to 4.5 s), which also verifies the effectiveness of the proposed guidance in handling input constraints. The mass variation of the lander during the powered descent phase is shown in Fig. 6, in which it can be found that the maximum fuel consumption is 4356 kg, which is 7.76% of the original weight and 27.3% of the initial fuel weight.

4.2 | Case 2: Multiple initial positions with variation along z-axis

The initial position along the x-axis is chosen as $r_x(0) = -2000$ m, and a range of the initial positions along the z-axis are chosen as $r_z(0) = 1520 : 20 : 1900$ m. The initial velocity is chosen as $\boldsymbol{v}(0) = [100, -75]^{\mathrm{T}}$ m/s.

The landing trajectories are shown in Fig. 7, in which it can be seen that with different initial positions along the z-axis (i.e. the altitude), the lander achieves precision landings and the glide-slope constraint is satisfied in all cases. From the velocity trajectories in Fig. 8, it can be seen that the terminal velocity along the z-axis converges to a small region around zero. The thrust force demonstrated in Fig. 9 shows that the input constraint is active (see the period from 0 s to 3.1 s) and no constraints violation happened subject to the nonlinear dynamic model and unknown disturbances. The mass variation of the lander during

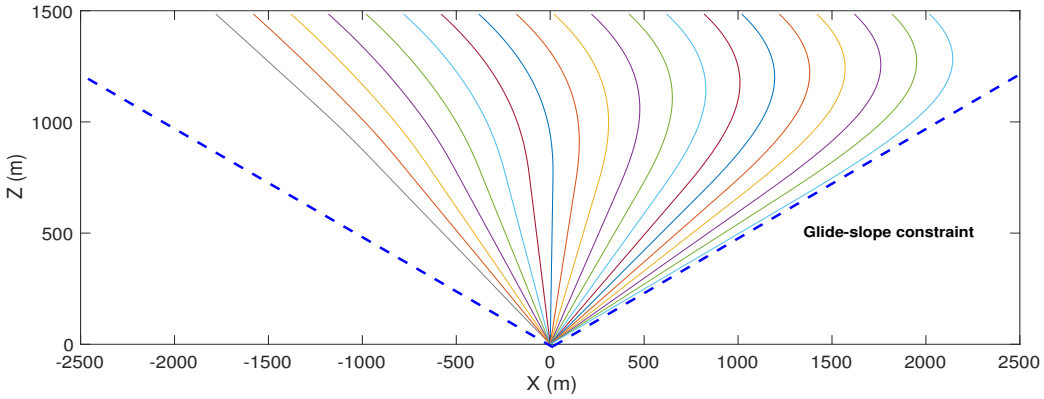


FIGURE 3 Landing trajectories

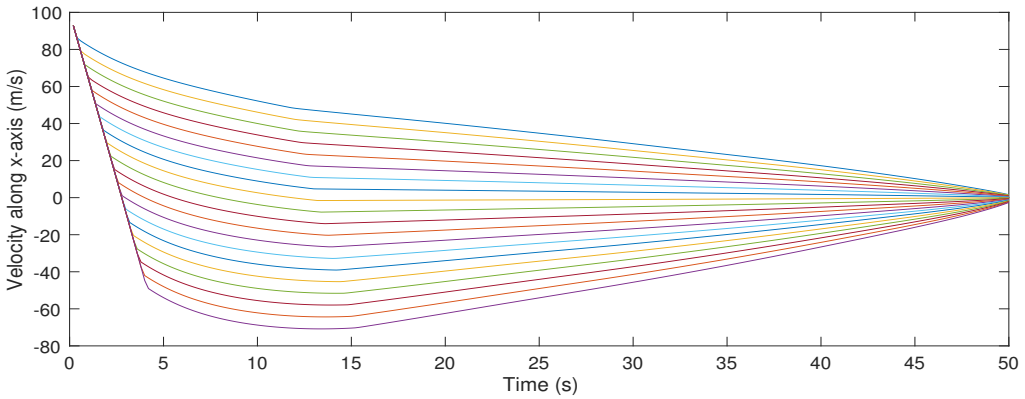


FIGURE 4 Velocities along x-axis vs time

the powered descent phase is shown in Fig. 10, in which it can be found that the maximum fuel consumption is 5857 kg, which is 10.45% of the original weight.

4.3 | Case 3: Multiple initial velocities with variation along x-axis

The initial velocity along the z-axis is chosen as $v_z(0) = -75$ m/s, and a range of the initial velocities along the x-axis are chosen as $v_x(0) = 0 : 10 : 190$ m/s. The initial position is chosen as $\boldsymbol{r}(0) = [1000, 1500]^{\mathrm{T}}$ m.

The landing trajectories are shown in Fig. 11, which verified that with different initial positions along the z-axis (i.e. the altitude), the lander achieves precision landings and the glide-slope constraint is satisfied in all cases. Therefore, the effectiveness of the proposed guidance in coping with state constraints is verified. From the velocity trajectories in Fig. 12, it can be seen that the terminal velocity along the z-axis converges to a small region around zero. The thrust force demonstrated in Fig. 13 shows that the input constraint is active (see the period from 0 s to 3.7 s) and no constraint violation happens subject to the nonlinear dynamic model and unknown disturbances. The mass variation of the lander during the powered descent phase is shown in Fig. 14, in which it can be found that the maximum fuel consumption is 6693 kg, which is 11.98% of the original weight.

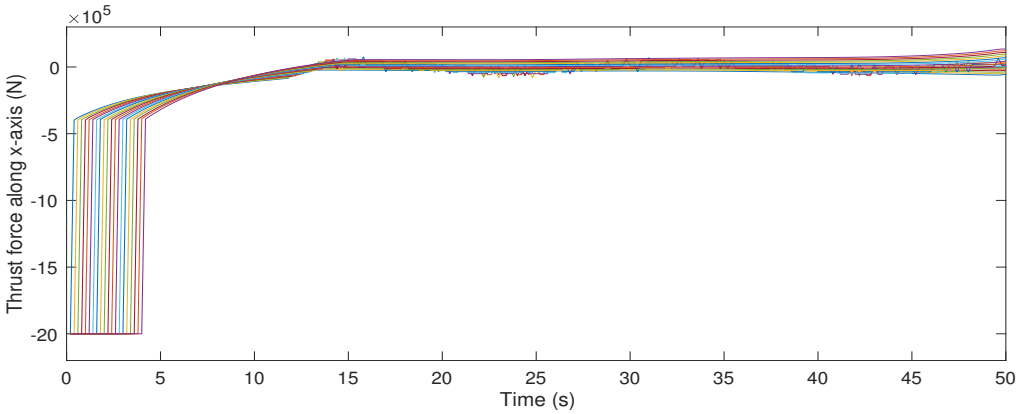


FIGURE 5 Thrust forces along x-axis vs time

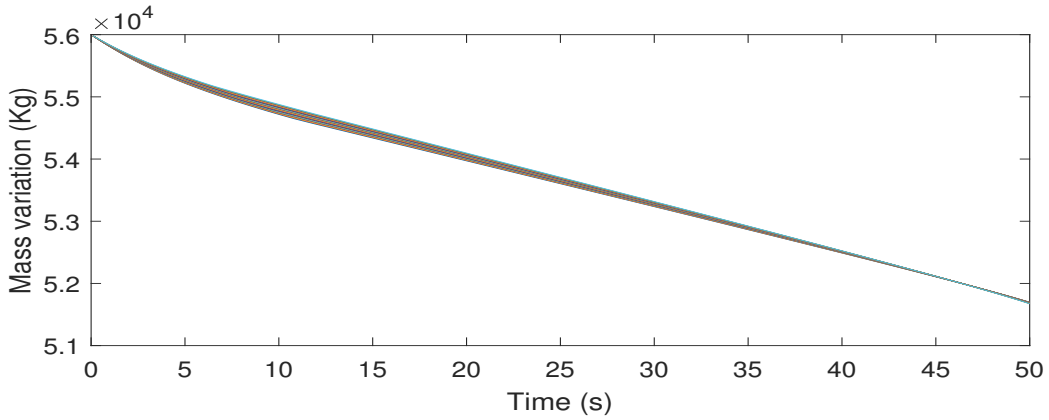


FIGURE 6 Mass variation vs time

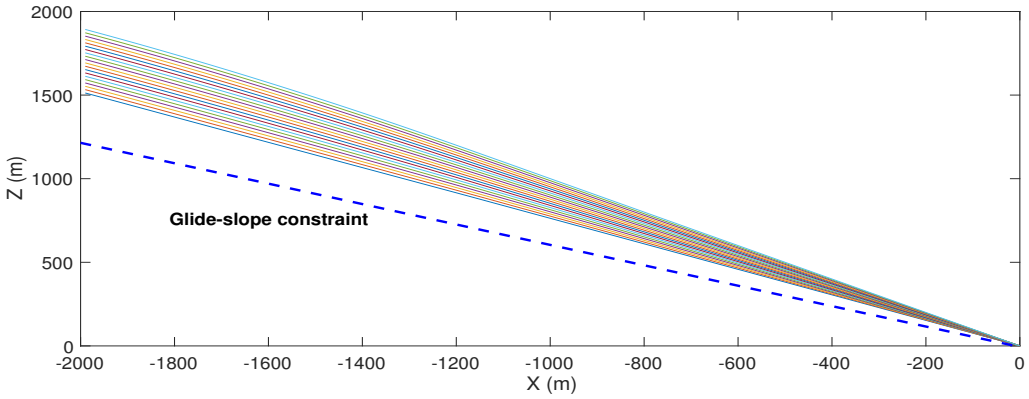


FIGURE 7 Landing trajectories

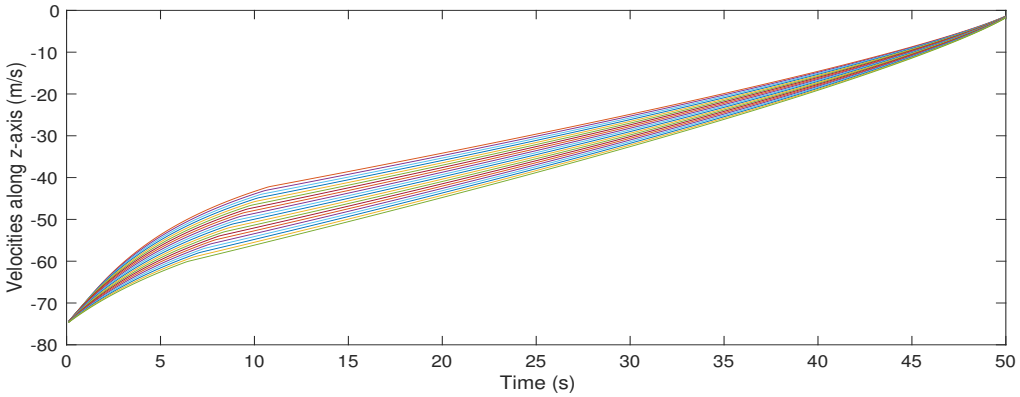


FIGURE 8 Velocities along z-axis vs time

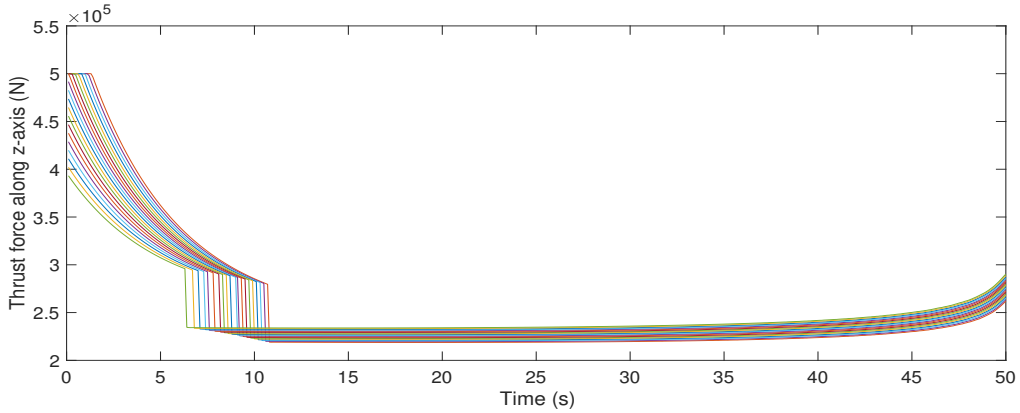


FIGURE 9 Thrust forces along z-axis vs time

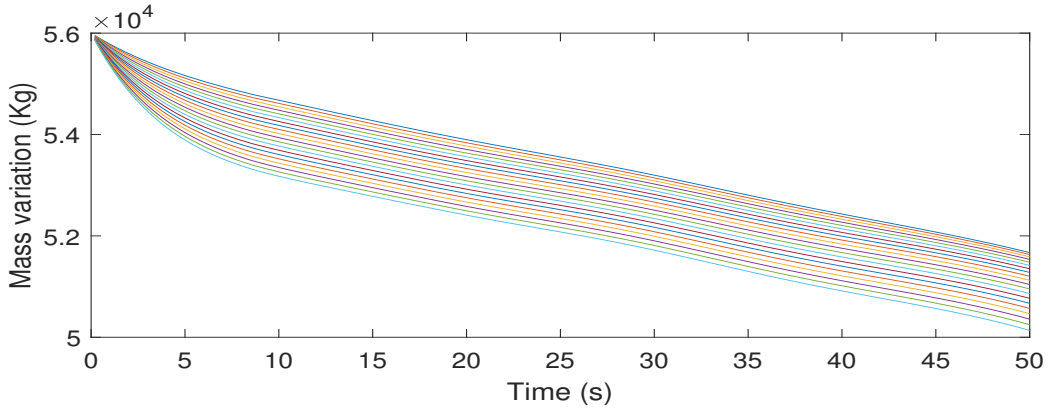


FIGURE 10 Mass variation vs time

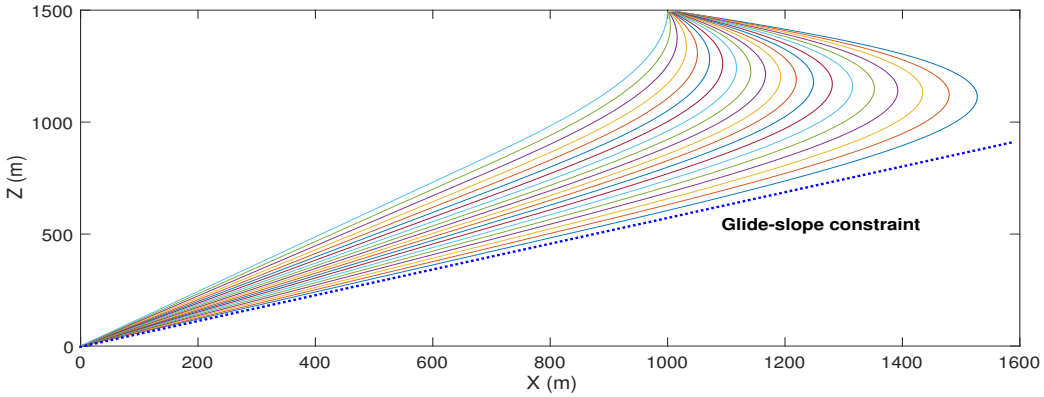


FIGURE 11 Landing trajectories

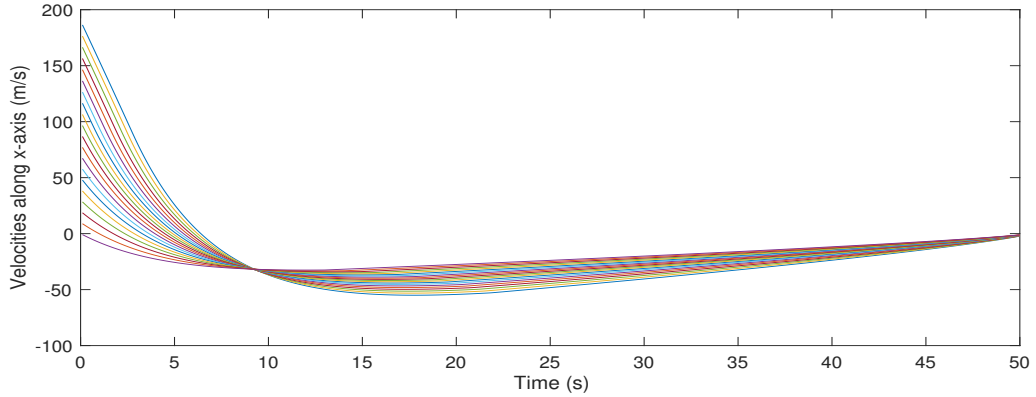


FIGURE 12 Velocities along x-axis vs time

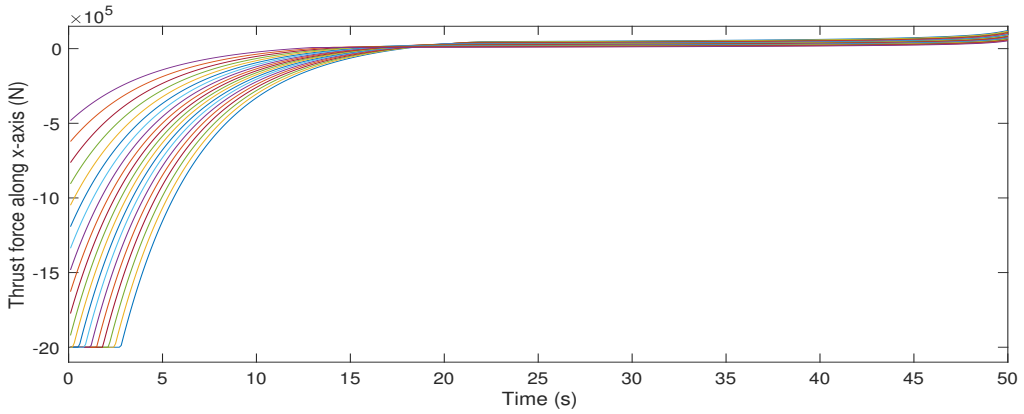


FIGURE 13 Thrust forces along x-axis vs time

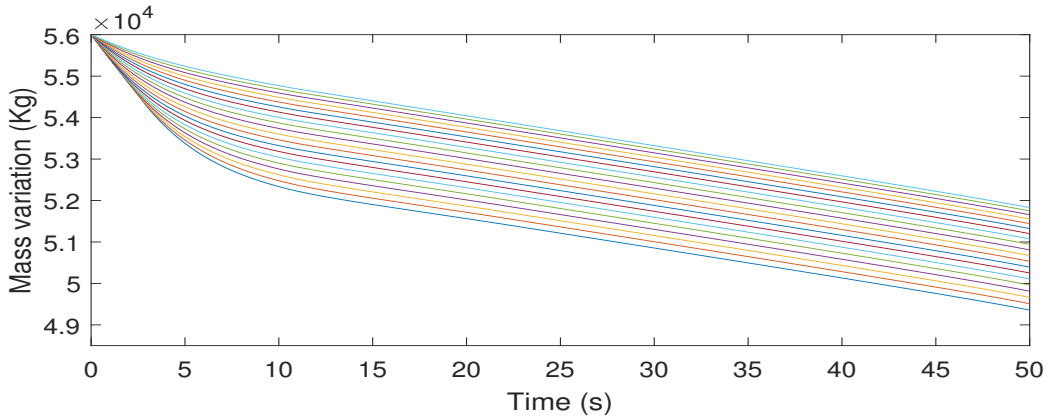


FIGURE 14 Mass variation vs time

5 | CONCLUSIONS

This paper proposed a reinforcement learning-based guidance law design method, in which the nonlinearity of the dynamic model and multiple constraints during the powered descent phase was handled by utilizing a modified cost function. A critic NN was designed to estimate the optimal cost function online, and its convergence has been proven. Simulation results demonstrated that the proposed method was effective with both input constraints and state constraints to be satisfied during the whole landing process. Future work will focus on the reinforcement learning-based guidance law design for 6 degree-of-freedom dynamical models during the powered descent phase.

References

1. Jiang X, Li S. Enabling technologies for Chinese Mars lander guidance system. *Acta Astronautica* 2017; 133: 375-386.
2. Brugarolas P, Chen A, Johnson A, et al. Concept for On-Board Safe Landing Target Selection and Landing for the Mars 2020 Mission. *Lpi Contributions* 2014; 1795: 8029.
3. Cui P, Gao X, Zhu S, Shao W. Visual navigation using edge curve matching for pinpoint planetary landing. *ACTA ASTRONAUTICA* 2018; 146(MAY): 171-180.
4. Kluever C. Entry guidance performance for Mars precision landing. *Journal of guidance, control, and dynamics* 2008; 31(6): 1537–1544.
5. Gong Y, Guo Y, Ma G, Guo M. Mars entry guidance for mid-lift-to-drag ratio vehicle with control constraints. *Aerospace Science and Technology* 2020; 107: 106361.
6. Long J, Gao A, Cui P, Liu Y. Mars atmospheric entry guidance for optimal terminal altitude. *Acta Astronautica* 2019; 155: 274–286.
7. Liu J, Zeng X, Li C, et al. Landing site selection and overview of China's lunar landing missions. *Space Science Reviews* 2021; 217(1): 1–25.
8. Cui P, Gao X, Zhu S, Shao W. Visual navigation using edge curve matching for pinpoint planetary landing. *Acta Astronautica* 2018; 146: 171–180.
9. Klumpp AR. Apollo lunar descent guidance. *Automatica* 1974; 10(2): 133-146.
10. McInnes , Colin R. Nonlinear transformation methods for gravity-turn descent. *Journal of Guidance Control Dynamics* 1995; 19(1): 247-248.
11. Ma L, Wang K, Xu Z, Shao Z, Song Z, Biegler LT. Multi-point powered descent guidance based on optimal sensitivity. *Aerospace Science and Technology* 2019.
12. Zhang B, Tang S, Pan B. Multi-constrained suboptimal powered descent guidance for lunar pinpoint soft landing. *Aerospace Science and Technology* 2016; 48: 203-213.
13. Guo Y, Hawkins M, Wie B. Waypoint-Optimized Zero-Effort-Miss/Zero-Effort-Velocity Feedback Guidance for Mars Landing. *Journal of Guidance, Control, and Dynamics* 2013; 36(3): 799-809.
14. Zhang Y, Guo Y, Ma G, Zeng T. Collision avoidance ZEM/ZEV optimal feedback guidance for powered descent phase of landing on Mars. *Advances in Space Research* 2017; 59(6): 1514-1525.
15. Wang P, Guo Y, Ma G, Wie B. Two-Phase Zero-Effort-Miss/Zero-Effort-Velocity Guidance for Mars Landing. *Journal of Guidance, Control, and Dynamics* 2021; 44(1): 75-87.
16. Wang Z, Li G, Jiang H, Chen Q, Zhang H. Collision-free navigation of autonomous vehicles using convex quadratic programming-based model predictive control. *IEEE/ASME Transactions on Mechatronics* 2018; 23(3): 1103-1113.

17. Wang T, Guo Y, Zhang Y, Ma G, Liang Z. Model predictive control guidance for constrained mars pinpoint landing. In: 2016 2nd International Conference on Control Science and Systems Engineering (ICCSSE). IEEE. ; 2016: 201–206.

18. Zhang Y, Vepa R, Li G, Zeng T. Mars powered descent phase guidance design based on fixed-time stabilization technique. *IEEE Transactions on Aerospace and Electronic Systems* 2018; 55(4): 2001–2011.

19. Blackmore L, Scharf DP. Minimum-Landing-Error Powered-Descent Guidance for Mars Landing Using Convex Optimization. *Journal of Guidance, Control, and Dynamics* 2010; 33(4): 1161-1171.

20. Acikmese B, Ploen SR. Convex Programming Approach to Powered Descent Guidance for Mars Landing. *Journal of Guidance, Control, and Dynamics* 2007; 30(5): 1353-1366.

21. Ren GF, Gao A, Cui PY, Luan EJ. A Rapid Power Descent Phase Trajectory Optimization Method with Minimum Fuel Consumption for Mars Pinpoint Landing. *Yuhang Xuebao/Journal of Astronautics* 2014; 35(12): 1350-1358.

22. Bai C, Guo J, Zheng H. Minimum-Fuel Powered Descent Guidance for Mars Landing. *2018 9th International Conference on Mechanical and Aerospace Engineering (ICMAE)* 2018.

23. He W, Gao H, Zhou C, Yang C, Li Z. Reinforcement learning control of a flexible two-link manipulator: an experimental investigation. *IEEE Transactions on Systems, Man, and Cybernetics: Systems* 2020; 51(12): 7326–7336.

24. Li C, Ding J, Lewis FL, Chai T. A novel adaptive dynamic programming based on tracking error for nonlinear discrete-time systems. *Automatica* 2021; 129: 109687.

25. Izzo D, Mrtens M, Pan B. A survey on artificial intelligence trends in spacecraft guidance dynamics and control. *Astrodynamics* 2019; 3(4): 287-299.

26. Gaudet B, Linares R, Furfaro R. Integrated guidance and control for pinpoint mars landing using reinforcement learning. In: AAS/AIAA Astrodynamics Specialist Conference. ; 2018: 1–20.

27. Ha B, Sv A. Craters in the vicinity of Valles Marineris region, Mars: Chronological implications to the graben and pits activities - ScienceDirect. *Icarus*; 343.

28. Lewis FL, Vrabie DL, Syrmos VL. *Reinforcement learning and optimal adaptive control*. Optimal Control, Third Edition . 2015.

29. Na J, Li G, Wang B, Herrmann G, Zhan S. Robust optimal control of wave energy converters based on adaptive dynamic programming. *IEEE Transactions on Sustainable Energy* 2018; 10(2): 961–970.

30. Na J, Mahyuddin MN, Herrmann G, Ren X, Barber P. Robust adaptive finite-time parameter estimation and control for robotic systems. *International Journal of Robust and Nonlinear Control* 2015; 25(16): 3045–3071.

31. Na J, Herrmann G. Online adaptive approximate optimal tracking control with simplified dual approximation structure for continuous-time unknown nonlinear systems. *IEEE/CAA Journal of Automatica Sinica* 2014; 1(4): 412–422.

AUTHOR BIOGRAPHY

Yao Zhang received her Ph. D. Harbin Institute of Technology in 2018. She is currently a Assistant Professor/Lecturer in Maritime Engineering at University of Southampton. Her research interest covers sliding mode control, model predictive control and applications on aerospace, marine engineering and renewable energy. She is also an Associate Editor of the IET Generation, Transmission and Distribution. She is the local event coordinator in IEEE UK and Ireland Control Systems Committee and an IEEE WIE Ambassador in UK and Ireland.

Tianyi Zeng received his B.Eng degree in control science and engineering from Harbin Institute of Technology, Harbin, China, in 2015, and the Ph.D. degree in control science and engineering from Beijing Institute of Technology, Beijing, China, in 2020. He is currently a research fellow at the Rolls-Royce University Technology Centre (UTC) in Manufacturing and On-Wing

Technology, University of Nottingham, UK. His research interest includes nonlinear system control, plant/controller co-design, robotic control.

Yanning Guo received the M.S. and Ph.D. degrees in control science and engineering from the Harbin Institute of Technology, in 2008 and 2012, respectively. He is currently a Professor with the Department of Control Science and Engineering, Harbin Institute of Technology, and currently teaches and performs research in the fields of deep space exploration, satellite attitude control.

Guangfu Ma received the M.S. and Ph.D. degrees in control science and engineering from the Harbin Institute of Technology, in 1987 and 1993, respectively. He is currently a Professor with the Department of Control Science and Engineering, Harbin Institute of Technology, where he became an Associate Professor in 1992, a professor in 1997, and currently teaches and performs research in the fields of deep space exploration, satellite attitude control, and nonlinear control.