Leveraging Curriculum Reinforcement Learning for Rocket Powered Landing Guidance and Control

Han Yuan
Beijing Institute of Astronautical Systems Engineering
Beijing, China
yuan-h15@tsinghua.org.cn

Yongzhi Zhao
Beijing Institute of Astronautical Systems Engineering
Beijing, China
zhaoyzh_01@139.com

Yu Mou
Beijing Institute of Astronautical Systems Engineering
Beijing, China
emu1982@sina.com

Xiaojun Wang
China Academy of Launch Vehicle Technology
Beijing, China
wangxj99@139.com

2021 China Automation Congress (CAC) | 978-1-6654-2647-3/21/$31.00 ©2021 IEEE | DOI: 10.1109/CAC53003.2021.9728176

***Abstract*—Future missions to recover the first stage of rocket require advanced guidance and control algorithms which can overcome stochastic disturbances, so as to achieve pinpoint landing. Recent studies have shown promising results of reinforcement learning (RL) based powered landing guidance and control method. However, many of them suffer from poor sample efficiency due to the sparse and complex reward, that trades off different goals in terms of attitude stability, terminal velocity, terminal attitude, terminal position and fuel usage. In order to address the above-mentioned problem, we propose a novel curriculum RL method to learn a fuel optimal pinpoint landing policy by maximizing the complex reward in this paper. The basic idea is to start with an easier task, and then increase the difficulty level by adding more goals. In this way, each task provides a good initial policy for the next task, making it easier to train the policy. The plane rigid dynamic model of rocket powered landing is adopted. The simulation result shows that the proposed RL method, which combined curriculum learning and Twin Delayed Deep Deterministic Policy Gradient (TD3) algorithm, can learn a better policy than the TD3 algorithm without curriculum under the same number of samples.**

***Keywords*—Reusable rocket, Powered landing guidance and control, Reinforcement learning, Optimal control, Curriculum learning**

I. INTRODUCTION

Recently, as rockets like Falcon-9 and New shepherd successfully reused the first stage and reduced the cost of launch by vertical landing, the technology of powered landing guidance has gained much attention. Researches about powered landing have been carried out since the 1960s for the Moon [1]-[2] and Mars [3]-[4] exploration mission. In the preceding missions, the precision of landing position is kilometer level, and the aerodynamics of powered landing guidance model is neglected. However, in the case of rocket vertical landing, the precision of landing position should be meter-level and the aerodynamics should not be neglected. With the developments of hardware and algorithm, the numerical trajectory planning based guidance and control approaches are proposed to improve the fuel efficiency and ability to cope with disturbances and initial states dispersions, these approaches are called "computational guidance and control" (CG&C) [5]. The objective function of CG&C is typically referred to as fuel-optimal problem [6]-[10] and energy-optimal problem [11]. Both of the two types of objective functions have their weaknesses. If the objective function is energy-optimal, the propellant usage will be significantly increased [12]. If the objective function is fuel-optimal, the thrust solution will have a Bang-Bang structure, and it is not robust under disturbances when the value of thrust is the maximum before touchdown [13]-[14].

Different from traditional CG&C approaches which adopt the deterministic model to generate control commands, many model-free reinforcement learning approaches adopt the stochastic model and generates control commands by maximizing the expectation of the objective function, in this way model uncertainty is considered. With the development of artificial intelligence, reinforcement learning is widely used in guidance and control problems in aerospace [15]-[23]. The reinforcement learning algorithms used in those studies, generally include deterministic policy gradient methods [24]-[26] and stochastic policy gradient methods [27]-[29]. Among the studies about powered landing guidance and control [20]-[23], some of them adopt a plane particle dynamic model, and the magnitude of uncertainty is small [20]-[21]. However, when a more complex dynamic model is adopted [22]-[23], we need to trade-off between more goals, which makes the policy hard to converge. In order to speed up training, they introduce shaping reward, however, since the shaping reward changes the objective, they fail to find a fuel-optimal policy. Therefore, it is important to improve both of sample efficiency and performance for reinforcement learning based powered landing guidance and control.

An effective way to learn faster is called curriculum reinforcement learning, which learn the complex task (target task) via a series of subtasks (source tasks) from simple to complex, the knowledge learned from previous task, include policy and value functions, are transferred to the next task. The curriculum learning was initially proposed to improve the sample efficiency for supervised learning [30], then the idea of curriculum learning was extended to reinforcement learning, and achieves higher sample efficiency [31]-[33]. In recent years, the curriculum reinforcement learning approach was applied to many types of applications, such as four-legged robot locomotion [34], autonomous driving [35], unmanned aerial vehicles control [36], all of them have achieved better sample efficiency and performance. However, it remains a challenge to construct a good curriculum for a specific task that benefits the learning process.

In this paper, a novel curriculum reinforcement learning based approach is proposed to learn a better policy with less samples. Since the main difficulties lie in trading off between different goals, we propose a two-stage curriculum RL method to achieve different goals gradually. To be more specific, we consider five different goals, including the attitude stability, the terminal velocity, the terminal attitude, the terminal position and the fuel usage. Details of the two-stage curriculum RL method are as follows: In the first stage, the task is to learn to keep the attitude stable, and to land on the ground with a small touchdown velocity and attitude error. In the final stage, the task is to learn to achieve all goals, especially focusing on the position error and fuel usage.

The policy network and value network are shared for all those tasks, which are trained in a sequential manner. As a result, the policy learned by the first task, can provide a good initial policy for the final task, making it easier to train the policy. The TD3 algorithm (Twin Delayed Deep Deterministic Policy Gradient) [26] is adopted in our method. Simulation results show that, compared to reinforcement learning without curriculum (standard RL), the proposed curriculum reinforcement learning (curriculum RL) approach improves both of sample efficiency and performance.

The rest of this paper is organized as follows: In section II, the models of rocket powered landing are established. In section III, the novel curriculum RL approach is proposed. In section IV, we present the simulation for the proposed approach. And in Section V, the whole study is concluded.

II. PROBLEM FORMULATION

A. Dynamics Equation

The plane rigid dynamic model is adopted to formulate the landing dynamic in the target coordinate system, while the rocket is controlled by thrust magnitude and thrust direction. The geometry of rocket landing is shown in Fig. 1, and the equations of motion are shown in (1).

$$\begin{cases} \dfrac{\mathrm{d}x}{\mathrm{d}t} = v\cos\gamma \\ \dfrac{\mathrm{d}y}{\mathrm{d}t} = v\sin\gamma \\ \dfrac{\mathrm{d}\theta}{\mathrm{d}t} = \omega \\ \dfrac{\mathrm{d}v}{\mathrm{d}t} = -\dfrac{D+T\cos(\alpha+\varepsilon)}{m} - g\sin\gamma \\ \dfrac{\mathrm{d}\gamma}{\mathrm{d}t} = \dfrac{L-T\sin(\alpha+\varepsilon)}{mv} - \dfrac{g}{v}\cos\gamma \\ \dfrac{\mathrm{d}\omega}{\mathrm{d}t} = \dfrac{M_{\mathrm{A}} - Tx_{\mathrm{T}}\sin\varepsilon}{I} \\ \dfrac{\mathrm{d}m}{\mathrm{d}t} = -\dfrac{T}{v_{\mathrm{ex}}} \end{cases} \tag{1}$$

Where x and y are the range and altitude, respectively. θ is the angle of attitude. v is the velocity. γ is the flight path angle. ω is the angular velocity. m is the mass of the rocket. T is the thrust magnitude, ε is the angle of thrust direction, v_{ex} is the effective exhaust velocity of rocket engines, x_{T} is the distance between mass center and engine. D is the drag force, L is the lift force, M_{A} is the aerodynamic moment. The aerodynamics are expressed as (2), the lift force is regarded as a trigonometric function. α is the AoA, it is calculated by $\alpha = \theta - \gamma$. I is the inertia, it is assumed to be $I = ml_I^2$. g is the gravitational acceleration, the spherical earth model is adopted, so g is assumed to be $g = g_0 R_{\mathrm{E}}^2/(R_{\mathrm{E}}+y)^2$, where $g_0 = 9.82\,\mathrm{m/s^2}$, $R_{\mathrm{E}} = 6370\,\mathrm{km}$.

$$\begin{cases} D = C_{\mathrm{D}} S_{\mathrm{ref}} q \\ L = C_{\mathrm{L}} S_{\mathrm{ref}} q \sin\alpha \\ M_{\mathrm{A}} = -L x_{\mathrm{A}} \\ q = \dfrac{1}{2}\rho v^2 \end{cases} \tag{2}$$

Where C_{D} and C_{L} are aerodynamic coefficients, S_{ref} is the reference area of the rocket, x_{A} is the distance between mass center and aerodynamic center, and $x_{\mathrm{A}} > 0$ means it is aerodynamically stable, q is the dynamic pressure, ρ is the atmospheric density and is expressed as $\rho = \rho_0 e^{-y/h_{\mathrm{MCP}}}$, where $\rho_0 = 1.225\,\mathrm{kg/m^3}$, $h_{\mathrm{MCP}} = 7.11\,\mathrm{km}$.

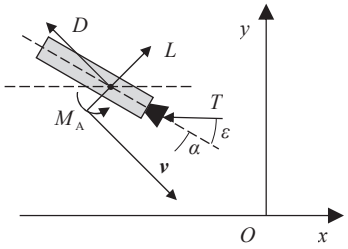


Fig. 1. Geometry of the rocket landing

Thrust magnitude and thrust direction are bounded as (3). The attitude should be stable during flight, we adopt angle and angular velocity criterion to estimate it, the attitude control is regarded as a failure when $|\omega| > \omega_{\max}$ or $\sin\theta > 0$.

$$\begin{cases} T_{\min} \le T \le T_{\max} \\ -\varepsilon_{\max} \le \varepsilon \le \varepsilon_{\max} \end{cases} \tag{3}$$

The rocket needs to land on a target position, with a small touchdown velocity and upright attitude, so the terminal constraints are (4). The objective of the two types of landing problem is to minimize the fuel usage, as shown in (5).

$$\begin{cases} |x(t_{\mathrm{f}})| \le x_{\mathrm{limit}} \\ v(t_{\mathrm{f}})^2 + 2g_0 y(t_{\mathrm{f}}) \le E_{\mathrm{limit}}, \\ |\theta(t_{\mathrm{f}}) + 90^\circ| \le \theta_{\mathrm{limit}}, \\ |v(t_{\mathrm{f}})\cos\gamma(t_{\mathrm{f}})| \le v_{x,\mathrm{limit}}, \\ m(t_{\mathrm{f}}) \ge m_{\mathrm{dry}} \end{cases} \tag{4}$$

$$\min J = -m(t_{\mathrm{f}}) \tag{5}$$

B. Markov Decision Process for Rocket Powered Landing

The rocket powered landing is formalized as a Markov Decision Process (MDP) in this subsection. The MDP is a type of discrete-time optimal control process, and consists of states, actions, transitions between states and reward function. In the MDP, the next state only depends on the current state and action, and does not depend on the history of states and actions. In this paper, the transitions are calculated by integrating the (1). For the i th step, the action is defined as $\boldsymbol{a}_i = [a_{1,i}, a_{2,i}]$, the of the elements in $\boldsymbol{a}_i$ are shown in (6), the state is defined as (7).

$$\begin{cases} a_{1,i} = (2T_i - T_{\min} - T_{\max})/(T_{\max} - T_{\min}) \\ a_{2,i} = \varepsilon_i/\varepsilon_{\max} \end{cases} \tag{6}$$

$$\boldsymbol{s}_i = [x_i, y_i, \theta_i, v_i, \gamma_i, \omega_i, m_i, a_{1,i-1}, \alpha_{2,i-1}] \tag{7}$$

Here we define the condition for terminating an episode. For the case of successful landing, the episode is terminated by $y \le 0$. Because $\mathrm{d}y/\mathrm{d}t$ is usually smaller than zero during flight with a good powered landing policy [6]-[10], we add a terminate condition, $\mathrm{d}y/\mathrm{d}t > 0$. We also terminate the episode when the attitude is unstable. As a result, the condition is

$$y \le 0 \text{ or } \sin\gamma > 0 \text{ or } |\omega| > \omega_{\max} \text{ or } \sin\theta > 0 \tag{8}$$

A good landing policy should trade off multiple goals. By normalizing the constraints shown in (4), we get the dimensionless variables in (9). Specifically, e_x, e_v, e_θ and e_{vx} are used to denote the relative error of terminal position, mechanical energy, attitude and horizontal velocity, respectively. The e_0 denotes the acceptable error, we suppose $e_0 = 0.5$. And we define a dimensionless variable r_{fuel} for calculating the reward for fuel usage in (10) by normalizing the objective in (5). The reward for terminal states should be negatively correlated with e_x, e_v, e_θ, e_{vx}, and positively correlated with r_{fuel}. We will introduce the details of reward function in the next section.

$$\begin{cases} e_x = \max\left(\left|x(t_{\mathrm{f}})\right| / x_{\mathrm{limit}}, e_0\right) \\ e_v = \max\left(\dfrac{v(t_{\mathrm{f}})^2 + 2g_0 y(t_{\mathrm{f}})}{E_{\mathrm{limit}}}, e_0\right) \\ e_\theta = \max\left(\left|\cos\theta(t_{\mathrm{f}})\right| / \sin\theta_{\mathrm{limit}}, e_0\right) \\ e_{vx} = \max\left(\left|v(t_{\mathrm{f}})\cos\gamma(t_{\mathrm{f}})\right| / v_{x,\mathrm{limit}}, e_0\right) \end{cases} \tag{9}$$

$$r_{\text{fuel}} = \left(m(t_{\mathrm{f}}) - m_{\mathrm{dry}}\right) / m_0 \tag{10}$$

Although existing studies show that a small value of discount factor could benefit the sample efficiency [20]-[23], such as 0.9, 0.95 and 0.99. But if the discount factor is smaller than 1, the discounted episode return will have an explicit relationship with flight time, which might affect the fuel usage. For example, when the reward for terminal states is negative, and discount factor is smaller than 1, the policy will be inclined to fly more steps to decay the terminal reward [22]. Therefore, we use 1 as the discount factor. And the control period and length of time step is 0.5s.

III. Learning Methodology

A. Policy Optimization Algorithm

In this paper, the TD3 algorithm is adopted to learn the policy. The details of TD3 algorithm are illustrated in [26], here we only introduce the adopted hyper-parameters. The fully connected feed-forward neural networks (as known as multi-layer-perception) are adopted for the policy and value function. The networks have 4 hidden layers and 1 output layer. More details for networks are shown in Table 1.

TABLE I. Policy and Value Functions Network Architecture

Layer	Policy Network		Value Network	
	size	*Activation*	*size*	*Activation*
Hidden 1	64	tanh	64	tanh
Hidden 2~4	64	LeaklyRelu	64	LeaklyRelu
Output	2	tanh	1	Linear

The hyper-parameters are as follows: learning rate for value network and policy network are 2×10^{-5} and 2×10^{-7}, respectively. Batch size is 16,384. Replay buffer size is 10^5. Target smooth factor is 0.005. Both the exploration noise and target policy noise are $N(0, 0.04)$.

B. Target task

Since the details of curriculum design are depended on the specific task, we introduce the details of dynamic in this subsection. We use the Falcon-9 rocket as prototype, and its parameters are listed in TABLE II. Notably, $x_{\mathrm{A}} < 0$, means the unstable aerodynamic model is adopted. The range of initial states and magnitude of wind and thrust disturbance are listed in TABLE III. , they follow uniform distribution.

TABLE II. Parameters of Rocket Landing Task

	Parameter	Value	parameter	Value
Rocket parameters	m_0	35t	$T_{\max}$	854kN
	m_{dry}	28t	$T_{\min}$	342kN
	S_{ref}	10.52m^2	v_{ex}	2770m/s
	x_{A}	-3m	$\varepsilon_{\max}$	10°
	C_{D}	1.5	x_{T}	10m
	C_{L}	1.8	l_l	12m
	$\omega_{\max}$	10°/s		
limitations and constraints	x_{limit}	5m	E_{limit}	9m^2/s^2
	$v_{x,\mathrm{limit}}$	1m/s	θ_{limit}	10°

TABLE III. Initial States and Disturbances

Parameter	Value range	parameter	Value range
x_0	[-700, -500]m	v_0	[220, 240]m/s
y_0	[1970, 2030]m	γ_0	[-66, -64] °
v_w	[-3, 3] m/s	δT	[-8.54, 8.54] kN

The wind is in the x direction, v_w denotes the wind velocity, and it is constant in an episode. δT denotes the thrust disturbance, it is random in each step.

We only give nonzero reward for terminal states, and the terminal reward is shown in (11). The first and second terms are used to encourage landing with small velocity error. The third term is used to encourage landing with small attitude error. The fourth term is used to encourage landing with small position error. The fifth term is used to encourage landing with low fuel usage. And the sixth term, $r_{P\theta}$, is a punishment term for failing to control attitude, when the episode is terminated by $|\omega| > \omega_{\max}$ or $\sin\theta > 0$, $r_{P\theta} = 50\times0.99^i$, otherwise $r_{P\theta}$ is zero. The key design is that $r_{P\theta}$ should decrease with the increase of flight time to encourage success control of attitude.

It will be shown later that it is hard to find an acceptable policy by standard RL, but a good policy could be found by curriculum RL. The "acceptable policy" means that, in the trajectory controlled by this policy, all terminal constraints are satisfied. The "good policy" means that, not only all terminal constraints are satisfied, but also the fuel usage is near to the open-loop optimal solution.

$$\begin{aligned} r_{\mathrm{f}} = {} & 40\left(1 + \frac{1}{\sqrt{e_v + e_{vx}}}\right) - 2\sqrt{e_v + e_{vx}} + 20\max\left(1 - e_\theta, 0\right) \\ & + 100\max\left(1 - e_x/60, 0\right) + 140 r_{\text{fuel}} - r_{P\theta} \end{aligned} \tag{11}$$

C. Two-stage Curriculum Reinforcement Learning

We construct a two-stage curriculum RL method as follows:

In the first stage, the rocket learns to keep the attitude stable, and to minimize the terminal states error, e_v, e_θ and e_{vx}. In this stage, because the terminal position constraint is not considered, we use large initial states disperses to avoid extrapolation in the next stage, the range of initial states are listed in TABLE IV. We adopt nonzero reward for non-terminal step in this stage, the non-terminal reward for step i is shown in (12). The purpose of this step reward function is to punish for large angular velocity. The terminal reward is shown in (13). Compared with the terminal reward shown in (11), the terms of position error and fuel usage are removed in this stage, and the first term will decay with flight time. The purpose of this form of terminal reward is to encourage the rocket to keep attitude stable firstly, and then to minimize terminal velocity and attitude error. Furthermore, this reward system could encourage short flight time, and the fuel optimal trajectory often have the feature of short flight time [12]. Notice that $50\times(1-0.99)=0.5$, when the attitude is stable and terminal velocity error is large, which mean $(\omega/\omega_{\max})^2$ and $1/\sqrt{e_v+e_{vx}}$ are close to zero, the episode return is independent with time. Later, when the velocity error is small, this episode return will increase with the decrease of flight time. If the average e_v is less than 1 and the episode is not terminated by $|\omega|>\omega_{\max}$ or $\sin\theta>0$ in the last 100 episodes, the training process of the first task will stop. The policy and value networks learned in this task will be the initial policy and value networks for the next task.

$$r_{1,i}=0.5\times 0.99^i-0.3\times(\omega/\omega_{\max})^2 \tag{12}$$

$$r_{1,\mathrm{f}}=50\times 0.99^i\times\left(1+1/\sqrt{e_v+e_{vx}}\right)-2\sqrt{e_v+e_{vx}}+20\max(1-e_\theta,0)-r_{P\theta} \tag{13}$$

TABLE IV. INITIAL STATES IN TASK 1

Parameter	Value range	parameter	Value range
x_0	[-1100, -100]m	v_0	[220, 240]m/s
y_0	[1970, 2030]m	γ_0	[-70, -60] °

Based on the policy learned from the first stage, the rocket is able to keep attitude stable and land with small touchdown velocity and attitude error. In the second stage, we add two additional goals, which are small position error and low fuel usage. In this stage, the reward system is as same as in the target task, which means that, the step reward, shown in (12), is removed, and the terminal reward function is (11). The key of this stage is that the coefficient of r_{fuel} in (11) should be fine-tuned to find a good policy and good sample efficiency. If reward for fuel usage is too large, the rocket could not find a good policy. If reward for fuel usage is too small, the sample efficiency is low.

Based on the two-stage curriculum, the rocket could find a good policy, as demonstrated in the simulation part.

IV. SIMULATION RESULTS

A. Sample Efficiency

To demonstrate the effectiveness of the proposed curriculum reinforcement learning (curriculum RL) method on the training process, we compare the learning curves of curriculum RL and standard reinforcement learning (standard RL). For standard RL, the policy and value networks are randomly initialized, and directly trained in the target task. For curriculum RL, the policy and value networks are randomly initialized for the first stage, and the learned policy and value networks are used for initialization in the final stage (as same as the target task). The network architectures and hyper-parameters for the two reinforcement learning approaches are same.

The learning curves are shown in Fig. 2. Specifically, we are interested in the following variables: episode return, the failure rate of attitude control, and some evaluation metrics for terminal states. The mean value of those variables per 100 episodes is used to plot the learning curves. The failure of attitude control means that the episode is terminated by $|\omega|>\omega_{\max}$ or $\sin\theta>0$. The evaluation metrics for terminal states include mechanical energy, attitude angle error, position error and mass. Those variables are calculated by $\sqrt{v(t_\mathrm{f})^2+2g_0y(t_\mathrm{f})}$, $|\theta(t_\mathrm{f})+90^\circ|$, $|x(t_\mathrm{f})|$ and $m(t_\mathrm{f})$, respectively. In details, the first stage of curriculum RL stopped after 46,426 episodes of training. We found that the performance of curriculum RL improves slowly after 100,000 episodes of training, so we terminate the training process.

As shown in Fig. 2, the standard RL could not find an acceptable policy within 100,000 episodes of training, while the curriculum RL finds an acceptable policy, which satisfies all the terminal constraints. The goodness of the curriculum RL will be demonstrated in the next subsection. For the terminal position error, it is very large in the first stage, because it is not included in the objective of the first stage. And it decreases quickly in the final stage since it is included in the objective of the final stage.

B. Policy Evaluation

The policy learned by curriculum RL is evaluated in two different settings in this subsection. The two settings differ in initial states and disturbance, which are introduced as follows:

In the first setting, we consider the maximum and minimum initial states in TABLE III. In case 1, the values of initial states are the medium value. In case 2, the values of initial states are the upper bound value. In case 3, the values of initial states are the lower bound value. The wind velocity and thrust disturbance are zero in this test.

The rocket could land successfully from all of those initial states, and the terminal states are listed in TABLE V. The optimal fuel usage is optimized by open-loop trajectory optimization. The profiles of the three trajectories are plotted in Fig. 3, which clearly show that the trained policy could guide the rocket landing on the target set. And the thrust profiles are close to the Bang-Bang profile, which is the fuel-optimal thrust profile. Furthermore, the rocket automatically keeps some thrust margin to overcome disturbance in the future. As shown in the TABLE V. , the fuel usage is 3.8% higher than the open-loop optimal value, while existing RL based guidance and control method is 18% higher than the open-loop optimal value [20].

TABLE V. FUEL CONSUMPTION AND TOUCHDOWN PRECISION

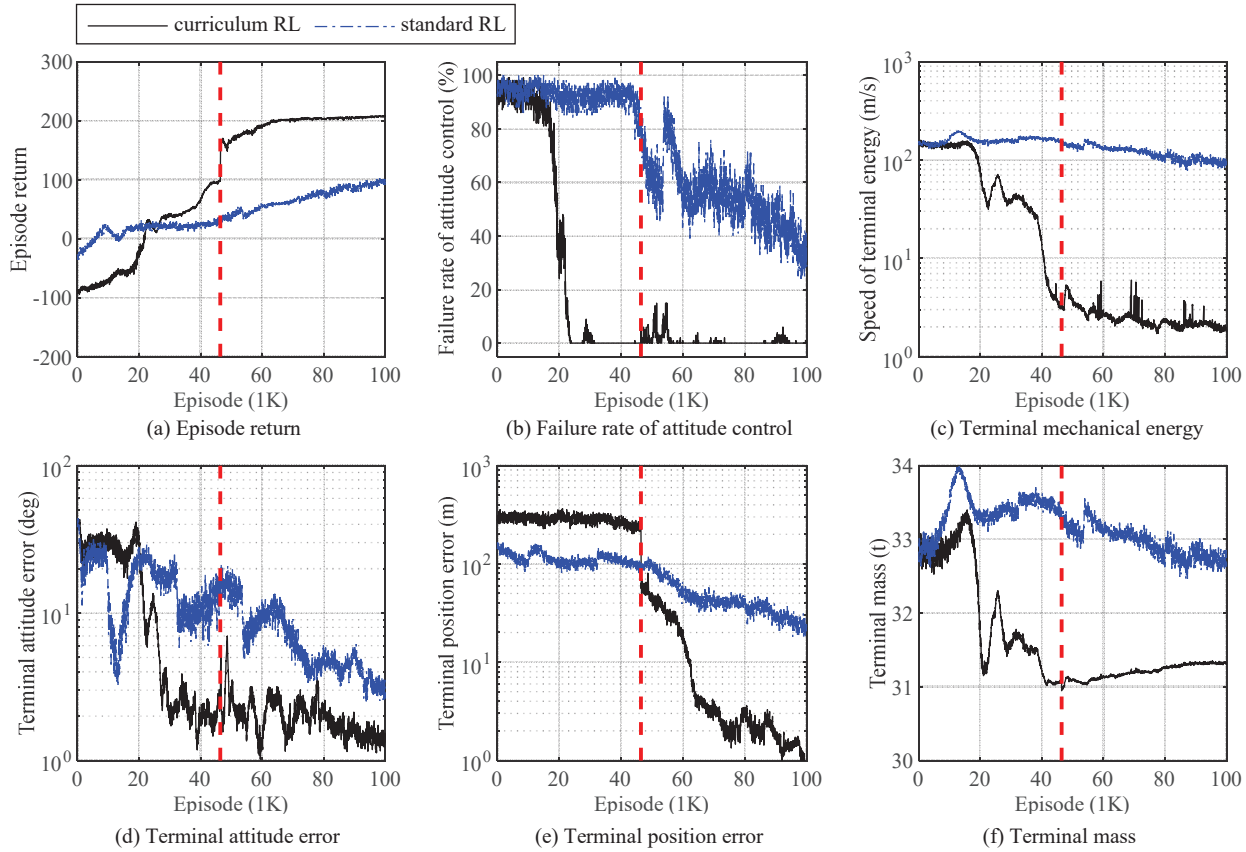
Initial states	Position (m)	Velocity (m/s)	Attitude angle (deg)	Fuel usage (kg)	
				Learning	*Optimal*
Case1	-1.54	1.43	-89.7	3669.2	3590.3
Case2	2.06	0.81	-88.4	3935.6	3791.6
Case3	-0.41	2.12	-85.5	3540.9	3485.1

In the second setting, we use 1000 episodes of Monte-Carlo simulations to test the policy. The dispersions of initial

states and disturbances in thrust and wind are considered, as shown in TABLE III. The statistics results are presented in TABLE VI. All those simulations are terminated by touchdown criterion, which means $y(t_f) \le 0$, so $y(t_f)$ is not listed. The statistical results show that all terminal constraints in (4) are satisfied. The results demonstrate the policy learned by curriculum RL is robust under such disturbances.

TABLE VI. STATISTICS OF TOUCHDOWN STATES

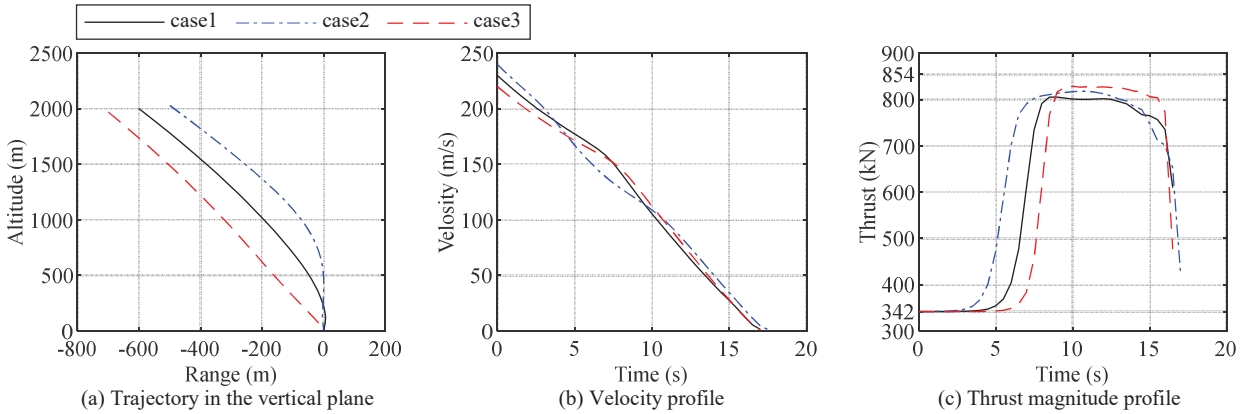
	Velocity (m/s)	**horizontal Velocity error (m/s)**	**Attitude angle error (deg)**	**Position error (m)**	**Fuel usage (kg)**
variables	$v(t_f)$	$\lvert v_x(t_f)\rvert$	$\lvert\theta(t_f)+90^\circ\rvert$	$\lvert x(t_f)\rvert$	$m_0 - m(t_f)$
main	1.61	0.15	0.75	1.10	3733.1
SD	0.336	0.072	0.514	0.535	143.4
Max	2.43	0.41	2.45	2.53	3971.7



(a) Episode return (b) Failure rate of attitude control (c) Terminal mechanical energy

(d) Terminal attitude error (e) Terminal position error (f) Terminal mass

Fig. 2. Learning curves. The red dash line denotes the switch from task 1 to task 2 in curriculum RL.



(a) Trajectory in the vertical plane (b) Velocity profile (c) Thrust magnitude profile

Fig. 3. Trajectory of typical cases

V. CONCLUSIONS

In this work, we proposed a novel curriculum learning method for reinforcement learning based powered landing guidance and control, to improve sample efficiency and performance. The plane rigid dynamic model with unstable aerodynamic is adopted, and the wind disturbance and thrust disturbance are considered. The simulation results demonstrate the advantage of curriculum RL in such task with multiple goals. Compared to standard RL, our method has higher sample efficiency and easily learns a good policy.

Our empirical analysis and existing studies suggest that it is hard to learn a good policy directly in such multiple goals task. And it is effective to learn the complex task (target task) via a series of subtasks (source tasks) from simple to

complex by adding more goals, and learn them in a gradual way. Therefore, each task provides a good initial policy for the next task, making it easier to train the policy. But our approach still has limitations in terms of generalizability. The design of sub-tasks depends on the transition structures of MDP and the relationship between different goals. Therefore, the design relies on expert experience.

This work adopted the 3DoF plane rigid model, in our future work, we will explore a curriculum RL approach for 6DoF model of the rocket landing problem.

REFERENCES

[1] A. R. Klumpp. "Apollo lunar descent guidance," Automatica. vol. 10, No. 3, pp. 133-146, 1975.

[2] H. ZHANG, Y. GUAN, M. CHENG, et al. "Guidance navigation and control for Chang'E-4 lander," SCIENTIA SINICA Technologica. Vol. 49, No 12, pp.1418-1428, 2019.

[3] M. S. Martin , G. Mendeck, P. Brugarolas, et al. "In-flight experience of the Mars Science Laboratory Guidance, Navigation, and Control system for Entry, Descent, and Landing," Ceas Space Journal. Vol. 7, No.2, pp. 119-142, 2015.

[4] A. Nelessen, C. Sackier, I. G. Clark, et al. "Mars 2020 Entry, Descent, and Landing System Overview, " in IEEE Aerospace Conference, Big Sky, MT, USA: IEEE, March 2019.

[5] P. Lu. "Introducing Computational Guidance and Control, " Journal of Guidance Control & Dynamics. Vol. 40, No. 2, pp. 193, 2017.

[6] P. Lu. "Propellant-optimal powered descent guidance, " Journal of Guidance Control & Dynamics. Vol. 41, No. 4, pp. 813-826, 2018.

[7] L. Ma, K. Wang, Z. Shao, et al. "Direct trajectory optimization framework for vertical takeoff and vertical landing reusable rockets: case study of two-stage rockets, " Engineering Optimization. Vol. 51, No. 4, pp. 627-645, 2019.

[8] X. Liu. "Fuel-optimal rocket landing with aerodynamic controls," Journal of Guidance Control and Dynamics. Vol. 42, No. 1, pp. 65-77, 2017.

[9] J. Wang, N. Cui, C. Wei, et al. "Optimal rocket landing guidance using convex optimization and model predictive control," Journal of Guidance Control and Dynamics. Vol. 42, No. 5, pp. 1078-1092, 2019.

[10] T. Reynolds, D. Malyuta, M. Mesbahi, et al. "A real-time algorithm for non-convex powered descent guidance," in AIAA SciTech Forum, Orlando, Florida, pp. 0844, January 6-10, 2020.

[11] Y. Li, W. Chen, H. Zhou, et al. "Conjugate gradient method with pseudospectral collocation scheme for optimal rocket landing guidance," Aerospace Science and Technology. Vol.104, pp. 105999, 2020.

[12] I. Ross. "How to find minimum-fuel controllers, " in AIAA Guidance, Navigation, and Control Conference and Exhibit, Rhode Island, USA, pp. 5346, August 16-19, 2004.

[13] I. Exarchos, E. A. Theodorou, P. Tsiotras, et al. "Optimal thrust profile for planetary soft landing under stochastic disturbances, " Journal of Guidance Control and Dynamics. Vol. 42, No.1, pp. 209-216, 2019.

[14] C. Wang, Z. Song, G. Shi, et al. "Trajectory planning for landing phase of reusable rocket with high thrust-to-weight ratio," in International Conference on Guidance, Navigation and Control, Tianjin, China, October 23-25, 2020.

[15] X. Guo, Z. Ren, Z. Wu, et al. "A Deep Reinforcement Learning Based Approach for AGVs Path Planning, " in 2020 Chinese Automation Congress, Shanghai, China, pp. 6833-6838, November 6-8, 2020.

[16] X. Liao, Y. Wang, Y. Xuan, et al. "AGV Path Planning Model based on Reinforcement Learning," in 2020 Chinese Automation Congress, Shanghai, China, pp. 6722-6726, November 6-8, 2020.

[17] Y. Lv, D. Hao, Y. Gao, et al. "Q-Learning Dynamic Path Planning for an HCV Avoiding Unknown Threatened Area, " in 2020 Chinese Automation Congress, Shanghai, China, pp. 271-274, November 6-8, 2020.

[18] D. Gao, H. Zhang, C. Li, et al. "Satellite Attitude Control with Deep Reinforcement Learning, " in 2020 Chinese Automation Congress, Shanghai, China, pp. 4095-4101, November 6-8, 2020.

[19] X. Qiu, Z. Yao, F. Tan, et al. "One-to-one Air-combat Maneuver Strategy Based on Improved TD3 Algorithm, " in 2020 Chinese Automation Congress, Shanghai, China, pp. 5719-5725, November 6-8, 2020.

[20] B. Gaudet and R. Furfaro. "Adaptive pinpoint and fuel efficient mars landing using reinforcement learning," in IEEE/CAA Journal of Automatica Sinica. Vol. 1, No. 4, pp. 397-411, 2014.

[21] Y. Chen and L. Ma. "Rocket powered landing guidance using proximal policy optimization," in International Conference on Automation, Control and Robotics Engineering, Shenzhen, China, July 19-21, 2019.

[22] B. Gaudet, R. Linares, R. Furfaro. "Deep reinforcement learning for six degree-of-freedom planetary landing," Advances in Space Research. Vol. 65, No. 7, pp. 1723-1741, 2020.

[23] B. Gaudet, R. Linares, R. Furfaro. "Adaptive guidance and integrated navigation with reinforcement meta-learning," Acta Astronautica. Vol. 169, pp. 180-190, 2020.

[24] D. Silver, G. Lever, N. Heess, et al. "Deterministic policy gradient algorithms," in Proceedings of the 31st International Conference on Machine Learning, Beijing, China, pp. I-387-395, June 21-26, 2014.

[25] T. P. Lillicrap, J. J. Hunt, A. Pritzel, et al. "Continuous control with deep reinforcement learning, " in Proceedings of the 4th International Conference on Learning Representations, San Juan, Puerto Rico, May 2-4, 2016.

[26] S. Fujimoto, H. Hoof, D. Meger. "Addressing Function Approximation Error in Actor-Critic Methods, " in Proceedings of the 35th International Conference on Machine Learning, Stockholm Sweden, pp. 1587-1596, July 10-15, 2018.

[27] V. Mnih, A. P. Badia, M. Mirza, et al. "Asynchronous methods for deep reinforcement learning," in Proceedings of the 33rd International conference on machine learning, New York city, NY, USA, pp. 1928-1937, June 19-24, 2016.

[28] J. Schulman, S. Levine, P. Abbeel, et al. "Trust region policy optimization," in Proceedings of the 32nd International Conference on Machine Learning, Lille, France, pp. 1889-1897, July 06-11, 2015.

[29] J. Schulman, F. Wolski, P. Dhariwal, et al. "Proximal policy optimization algorithms," arXiv preprint, arXiv:1707.06347, 2017.

[30] Y. Bengio, J. Louradour, R. Collobert, et al. "Curriculum learning," in Proceedings of the 26th International Conference on Machine Learning, Montreal, Canada, pp. 41–48, June 14-18, 2009.

[31] S. Narvekar, J. Sinapov, M. Leonetti, et al. "Autonomous Task Sequencing for Customized Curriculum Design in Reinforcement Learning," in International Joint Conferences on Artificial Intelligence, Melbourne, Australia, pp. 2536-2542, August 19-25, 2017.

[32] S. Narvekar, B. Peng, M. Leonetti, et al. "Curriculum Learning for Reinforcement Learning Domains: A Framework and Survey," Journal of Machine Learning Research. Vol. 21, pp.1-50, 2020.

[33] F. L. Da Silva and A. H. R. Costa. "Object-Oriented Curriculum Generation for Reinforcement Learning," in Proceedings of the 17th International Conference on Autonomous Agents and Multi Agent Systems, Stockholm, Sweden, pp. 1026-1034, July 10-15, 2018.

[34] J. Lee, J. Hwangbo, L. Wellhausen, et al. "Learning quadrupedal locomotion over challenging terrain," Science Robotics, Vol. 5, No. 47, pp. eabc5986, 2020.

[35] Y. Song, H. Lin, E. Kaufmann, et al. "Autonomous Overtaking in Gran Turismo Sport Using Curriculum Reinforcement Learning," in 2021 IEEE International Conference on Robotics and Automation, Xi'an, China, May 30 - June 5, 2021.

[36] T. Pollack and E. J. Van Kampen. "Safe Curriculum Learning for Optimal Flight Control of Unmanned Aerial Vehicles with uncertain system dynamics", in AIAA Scitech Form, Orlando, FL, USA, January 6-10, 2020.