Acta Astronautica

Robust Transfer Trajectory Design between Lagrange Points in a Binary Asteroids Based on Reinforcement Learning

--Manuscript Draft--

Manuscript Number:	AA-D-24-02114
Article Type:	Research paper
Section/Category:	Space Technology & Systems Development
Keywords:	Robust trajectory design; binary asteroid system; reinforcement learning
Corresponding Author:	Lie Yang Harbin Institute of Technology CHINA
First Author:	Lie Yang
Order of Authors:	Lie Yang
	Mingying Huo
	Ruhao Jin
	Kang Sun
	Qiufan Yuan
	Ziqi Hao
	Lehan Wang
	Naiming Qi
Abstract:	This study employs Reinforcement Learning (RL) to address the robust optimization challenges in transferring trajectories between Lagrangian points within a binary asteroid system, accounting for uncertainties in dynamic modeling. A second-order gravity field model based on a double ellipsoid model is established, and the ellipsoid shape uncertainties are introduced to represent the dynamic uncertainties. The robust transfer trajectory optimization problem is then formulated as a Markov Decision Process (MDP), with a tailored reward function designed to facilitate RL-based solutions. The study focuses on solving transfer problems between L2-L4 and L2-L3 Lagrangian points, demonstrating through simulations that RL-generated policies, even without considering dynamic uncertainties, yield orbital maneuvering strategies comparable to those derived from indirect methods, thus validating the efficacy of RL in this context. Moreover, Monte Carlo simulations underscore RL's capability to generate robust orbital maneuvering strategies under dynamic uncertainty. Compared to strategies neglecting dynamic uncertainties, these RL-based approaches exhibit a significant reduction in terminal error standard deviation, ensuring more stable transfer trajectories.
Suggested Reviewers:	Gang Zhang zhanggang@hit.edu.cn
	Hongwei Yang hongwei.yang@nuaa.edu.cn
	Jihe Wang wangjihe@mail.sysu.edu.cn
Opposed Reviewers:	

October 31, 2024

Rock Jeng-Shing Chern
Editor-in-Chief
Acta Astronautica

Dear Editor:

I wish to submit a research paper for publication in *Acta Astronautica,* entitled "Robust Transfer Trajectory Design between Lagrange Points in a Binary Asteroids Based on Reinforcement Learning." The paper was coauthored by Mingying Huo, Ruhao Jin, Kang Sun, Qiufan Yuan, Ziqi Hao, Lehan Wang, Naiming Qi.

This paper aims to study the robust optimization design of low-thrust Lagrangian point transfer trajectories using reinforcement learning in the presence of gravitational field modeling uncertainties in binary asteroid systems. We believe that our study makes a significant contribution to the literature.

Further, we believe that this paper will be of interest to your journal readers, as asteroid exploration and near-Earth asteroid defense missions have become popular space exploration projects in recent years, which is one of the motivations behind this study. This research has the potential to optimize trajectory planning near multiple asteroid systems such as binary asteroid exploration and other complex deep space exploration mission designs.

This manuscript has not been published or presented elsewhere in part or in entirety and is not under consideration by another journal. We have read and understood your journal's policies, and we believe that neither the manuscript nor the study violates any of these. There are no conflicts of interest to declare.

Thank you for your consideration. I look forward to hearing from you.

Sincerely,
Lie Yang
School of Astronautics,
Harbin Institute of Technology,
Harbin 150001, China
E-mail: 21B918122@stu.hit.edu.cn

- This study uses reinforcement learning (RL) to solve the problem of robust trajectory optimization between Lagrange points in a binary asteroid system.
- A gravitational field uncertainty model is established for the binary asteroid system and a reward is customized function suitable for this problem.
- The indirect method is employed to compare and verify the accuracy of the RL optimization results, as well as the robustness of the terminal.

Robust Transfer Trajectory Design between Lagrange Points in a Binary Asteroids Based on Reinforcement Learning

Lie Yang[a], Mingying Huo[a,*], Ruhao Jin[a], Kang Sun[b], Qiufan Yuan[c], Ziqi Hao[a], Lehan Wang[a], Naiming Qi[a]

[a]*School of Astronautics, Harbin Institute of Technology, Harbin 150001, China*
[b]*Beijing Institute of Spacecraft System Engineering, Beijing 100094, China*
[c]*Shanghai Institute of Aerospace Systems Engineering, Shanghai 201108, China*

Abstract

This study employs Reinforcement Learning (RL) to address the robust optimization challenges in transferring trajectories between Lagrangian points within a binary asteroid system, accounting for uncertainties in dynamic modeling. A second-order gravity field model based on a double ellipsoid model is established, and the ellipsoid shape uncertainties are introduced to represent the dynamic uncertainties. The robust transfer trajectory optimization problem is then formulated as a Markov Decision Process (MDP), with a tailored reward function designed to facilitate RL-based solutions. The study focuses on solving transfer problems between L2-L4 and L2-L3 Lagrangian points, demonstrating through simulations that RL-generated policies, even without considering dynamic uncertainties, yield orbital maneuvering strategies comparable to those derived from indirect methods, thus validating the efficacy of RL in this context. Moreover, Monte Carlo simulations underscore RL's capability to generate robust orbital maneuvering strategies under dynamic uncertainty. Compared to strategies neglecting dynamic uncertainties, these RL-based approaches exhibit a significant reduction in terminal error standard deviation, ensuring more stable transfer trajectories.

*Corresponding author
Email address: huomingying@hit.edu.cn (Mingying Huo)

Keywords: Robust trajectory design, binary asteroid system, reinforcement learning

1. Introduction

Asteroids are celestial bodies in the solar system that orbit the sun similar to planets, but are much smaller in size and mass. Detecting asteroids is helpful to study the origin and evolution of planetary systems, such as the Hayabusa 1[1], Hayabusa 2[2] and NASA's OSIRIS-REx[3] missions that have already completed asteroid sampling returns. For binary asteroid exploration, NASA's DART mission[4] completed a kinetic impact test of the smaller asteroid in 65803 Didymos in 2022.

Given the unique gravitational field of binary asteroid systems compared to single celestial bodies, extensive research has focused on spacecraft mission orbit designs near such systems. For instance, this includes exploration and maintenance of periodic orbits[5, 6] and bounded orbits[7, 8], studying transfer trajectories utilizing invariant manifolds[9–11], designing landing[12–14] and autonomous impact[15, 16] trajectories, among other areas of investigation. Research into orbital transfers of probes within binary asteroid systems predominantly centers on designing transfer orbits for synchronous binary systems. In such scenarios, the system closely resembles the classical circular restricted three-body problem, enabling solutions through analogous methods employed in the classical problem. Wu[11] design transfer trajectory in the doubly synchronous binary system based on the libration point orbits and their invariant manifolds. After selecting the surfaces of two bodies as Poincaré sections, the L1 and L2/L3 manifolds can be patched together to construct the equatorial and spatial transfer trajectory. Ferrari[17] studied transfer trajectories of spacecraft by assembling different three-body problems. The orbital motion near the primary and away from it are approximated using two circular restricted three-body problem(CRTBP) models. Based on the distribution characteristics of the invariant manifolds of these two three-body problems, a global low-energy

transfer trajectory and trajectories for takeoff or landing on the asteroid surface were designed using manifold stitching techniques. In recent years, due to the development and application of artificial intelligence methods in trajectory design, learning algorithms have solved the problem of trajectory optimization and control near binary asteroid systems. Federici[15] proposed meta-reinforcement learning combined with convolutional neural networks to achieve impact guidance and control for binary asteroids, simulating the final stages of the DART mission. Based on the particle swarm optimization algorithm and neural network, Ambrosio[18] performed neural network fitting on the gravitational acceleration of the binary asteroid system to improve the calculation efficiency, and successfully applied it to the optimization design of celestial body orbital trajectory and transfer trajectory. Parmar[19] generates a decision network for Lagrange point transfer trajectories based on behavioral cloning and long short-term memory neural networks, using the optimization results of traditional direct and indirect methods as training sets. In the design process of the transfer trajectory mentioned above, the uncertainty stemming from the asteroid's approximate shape in the gravitational field is often overlooked. Some studies only perform Monte Carlo simulations and robustness analyses after completing the trajectory design. Therefore, this work will focus on the robust optimization design of transfer trajectories within binary asteroid systems.

Meanwhile, robust trajectory optimization has emerged as hot spot[20–24] due to the inevitable uncertainties in flight missions, such as unmodeled accelerations, navigation errors, errors in onboard mechanism execution, and potential thruster failures. In recent years, there has been a significant rise in interest in deep learning techniques to address optimal and robust control challenges, particularly in the realm of space applications robust space trajectory design, because neural network controllers can establish the mapping between spacecraft states and corresponding thruster actions through basic linear algebra operations without iteration, with high computational efficiency and suitable for onboard autonomous guidance. Feng[25] used a transfer learning deep neural network (TL-DNN) combined with gravity field data collected in a limited

area to perform online modeling of the asteroid's non-uniform gravity field, reducing the uncertainty of the gravity field in the asteroid soft landing mission. Compared with supervised learning in neural network training methods, reinforcement learning (RL) does not require preparing a large amount of data sets in advance. Instead, it models the problem as an Markov Decision Process (MDP), and the agent repeatedly interacts with the environment to maximize the cumulative reward on the trajectory. The established MDP can take into account various uncertainties. At the same time, the agent will randomly explore the entire solution space in the process of interaction with the environment, so the robustness of the neural network controller generated by training can be guaranteed. Recently, there has been various application of RL in robust trajectory optimization design. Zavoli[26] applies RL to design robust trajectories for low-thrust interplanetary missions, addressing significant uncertainties and disturbances. Hu[27] designed dense rewards based on Zavoli's work to solve the multirevolution low-thrust robust trajectory optimization problem. Federici[28] studied the application of meta-reinforcement learning to the robust design of low-thrust interplanetary trajectories. Yuan[29] proposes a deep reinforcement learning-based method for robust navigation and guidance of kinetic impact on near-Earth asteroids, which map angle measurements to guidance maneuvers. Boone[30] employed RL to train a policy network for effectively conducting multi-pulse spacecraft transfer trajectory between periodic orbits in the Earth-Moon circular restricted three-body problem, while considering uncertainties in observational states. Yang[31] proposed robust zero-effort-miss/zero-effort-velocity (R-ZEM/ZEV) guidance to overcome the low terminal accuracy in RL-based transfer trajectory design.

This paper aims to study the robust optimization design of low-thrust Lagrangian point transfer trajectories using reinforcement learning in the presence of gravitational field modeling uncertainties in binary asteroid systems. The binary asteroid system is characterized by a double ellipsoid model, so the uncertainty in the gravitational field is caused by the modeling errors of the three axes of the ellipsoids. As a further contribution to the existing literature, this

paper applies reinforcement learning methods to the design of Lagrangian point transfer trajectories for a binary asteroid system and proposes a reward function suitable for the scenario. The rest of this paper is organized as follows. In Section 2, the dynamics model for a low-thrust spacecraft near a binary asteroid system is proposed and the MDP required for developing a robust trajectory optimization method in an uncertainty scenario is formulated. The reward function design for the Lagrangian point transfer scenario is presented. After that, Section 3 presents the implementation of the training algorithm PPO. Section 4 presents the performance of the proposed method in the L2 to L4 and L2 to L3 transfer tasks in 809 Lundia. Finally, Section 5 concludes this paper.

2. Problem Statement

2.1. Dynamics of Low-thrust Trajectories Near Binary Asteroid System

It is necessary to model the binary asteroid system before describing the dynamic equations of a spacecraft around it. In this paper, a dual ellipsoid model is used to characterize the binary asteroid system. Assuming that the binary asteroid system is tidally locked and stable in its long axis, and considering the spacecraft's negligible mass relative to the binary asteroids, the dynamics of the spacecraft can be accurately described within a rotating coordinate system centered on the center of mass of the binary asteroid system, as shown in Figure 1. The celestial body M_1 is an ellipsoid with a mass m_1, characterized by three semi-major axes α_1, β_1 and γ_1, where $\alpha_1 > \beta_1 > \gamma_1$; the celestial body M_2 is also an ellipsoid with a mass m_2, characterized by three semi-major axesα_2, β_2 and γ_2, where $\alpha_2 > \beta_2 > \gamma_2$. The distance between M_1 and M_2 is L and the mass ratio of the binary asteroid system is $\mu = m_2/(m_1 + m_2)$, so the locations of the primary and secondary can be denoted as $[-\mu L, 0, 0]^T$and $[(1-\mu)L, 0, 0]^T$, respectively. The semi-major axis α_2 of the primary asteroid is taken as the normalized length UL, $UT = \sqrt{\alpha_2^3/[G(m_1 + m_2)]}$ is the normalized time, where G is the gravitational constant. Given the normalized length, the

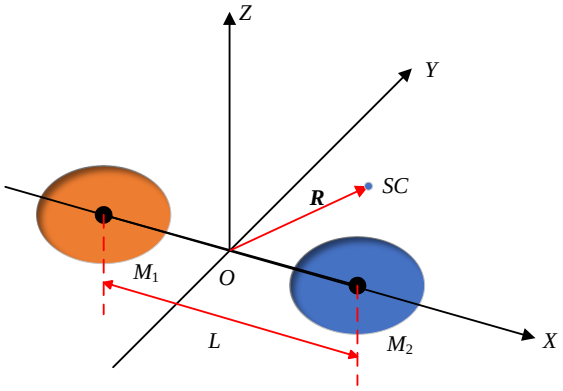

Figure 1: The ellipsoid-ellipsoid model of the binary asteroid system.

orbital trajectory dynamic equation near the binary asteroid system is

$$\begin{cases} \ddot{x} = \dfrac{\partial U_{12}}{\partial x} + 2\omega\dot{y} + f_x \\ \ddot{y} = \dfrac{\partial U_{12}}{\partial y} - 2\omega\dot{x} + f_y \\ \ddot{z} = \dfrac{\partial U_{12}}{\partial z} + f_z \end{cases}, \tag{1}$$

where $\boldsymbol{R} = [x, y, z]^T$ is denoted the position of spacecraft, $\boldsymbol{f} = [f_x, f_y, f_z]^T$ is the 3D vector of the propulsive acceleration. ω is the rotation angular velocity, and U_{12} is the gravitational field of the binary asteroid system. Subsequently, the gravitational field of the binary asteroid system is derived in detail using the double ellipsoid model.

The gravitational field of the binary asteroid system can be expressed as the combination of the gravitational fields of its two constituent single asteroids. In order to balance accuracy and efficiency, a second-order quadratic spherical harmonics model is used to calculate an asteroid gravitational field

$$U = U_0 + U_1 + U_2, \tag{2}$$

where

$$U_0 = \frac{\mu_c}{r}, \tag{3}$$

$$U_1 = \frac{\mu_c}{r^2}\left[C_{10}\sin\delta + \cos\delta\left(C_{11}\cos\lambda + S_{11}\sin\lambda\right)\right], \tag{4}$$

$$U_2 = \frac{\mu_c}{r^3}\left[C_{20}\left(1-\frac{3}{2}\cos^2\delta\right)+\frac{3}{2}\cos 2\delta\left(C_{21}\cos\lambda+S_{21}\sin\lambda\right)+3\cos^2\delta\left(C_{22}\cos 2\lambda+S_{22}\sin 2\lambda\right)\right], \quad (5)$$

μ_c is the gravitational constant of the central gravitational body, δ is the latitude, λ is the longitude. C_{ij}, S_{ij}, $i,j = 0,1,2$ are the second-order quadratic spherical harmonics model parameters. $C_{10} = C_{11} = S_{11} = C_{21} = S_{21} = S_{22} = 0$ when the coordinate origin of the fixed coordinate system of the asteroid is set as the center of mass, and the three-axis direction is along the direction of the main inertia axis of the central body. Eq. (2)˜(5) can be simplified to

$$U = U_0 + U_1 + U_2 = \frac{\mu_c}{r} + \frac{\mu_c}{r^3}\left[C_{20}\left(1-\frac{3}{2}\cos^2\delta\right)+3C_{22}\cos^2\delta\cos 2\lambda\right], \quad (6)$$

Longitude and latitude can be expressed in Cartesian coordinates as

$$\begin{cases} \cos\delta = \dfrac{\sqrt{(x^2+y^2)}}{r} \\ \sin\lambda = \dfrac{y}{\sqrt{(x^2+y^2)}} \\ r = \sqrt{x^2+y^2+z^2} \end{cases} \quad (7)$$

Substitute Eq. (7) into Eq. (6) to derive the gravitational potential of celestial bodies in the Cartesian coordinate system

$$U = \frac{\mu_c}{r} + \left[-\frac{\mu_c C_{20}\left(x^2+y^2-2z^2\right)}{2r^5}+\frac{3\mu_c C_{22}\left(x^2-y^2\right)}{r^5}\right]. \quad (8)$$

The parameters C_{20} and C_{22} determined by the three spindle moments of inertia of the celestial body I_{xx},I_{yy},I_{zz}

$$\begin{cases} C_{20} = \dfrac{1}{2}(2I_{zz} - I_{xx} - I_{yy}) \\ C_{22} = \dfrac{1}{4}(I_{yy} - I_{xx}) \end{cases} \quad (9)$$

When the central celestial body is an ellipsoid and its three main axes are α, β and γ, the moment of inertia of each coordinate axis can be obtained as

$$\begin{cases} I_{xx} = \dfrac{1}{5}\left(\beta^2+\gamma^2\right) \\ I_{yy} = \dfrac{1}{5}\left(\alpha^2+\gamma^2\right) \\ I_{zz} = \dfrac{1}{5}\left(\alpha^2+\beta^2\right) \end{cases}. \quad (10)$$

Therefore, the expression of the gravitational field of the binary asteroid system can be obtained as

$$\left\{\begin{array}{c} U_{12}=U_{M_1}+U_{M_2} \\ U_{M1}=\frac{\mu}{\sqrt{[x-(1-\mu)r]^2+y^2+z^2}}-\frac{\mu C_{20M_1}\left\{[x-(1-\mu)r]^2+y^2-2z^2\right\}}{2\left\{[x-(1-\mu)r]^2+y^2+z^2\right\}^{\frac{5}{2}}}+\frac{3\mu C_{22M_1}\left[[x-(1-\mu)r]^2-y^2\right]}{\left\{[x-(1-\mu)r]^2+y^2+z^2\right\}^{\frac{5}{2}}} \\ U_{M2}=\frac{1-\mu}{\sqrt{(x+\mu r)^2+y^2+z^2}}-\frac{(1-\mu)C_{20M_2}\left[(x+\mu r)^2+y^2-2z^2\right]}{2\left[(x+\mu r)^2+y^2+z^2\right]^{\frac{5}{2}}}+\frac{3(1-\mu)C_{22M_2}\left[(x+\mu r)^2-y^2\right]}{\left[(x+\mu r)^2+y^2+z^2\right]^{\frac{5}{2}}} \end{array}\right. . \tag{11}$$

Similar to CRTBP, the double ellipsoid gravitational field model also features five gravitational equilibrium points known as Lagrange points. At these points, the relative velocity and acceleration of a spacecraft relative to the main celestial bodies are zero. In Eq.(1), let $\dot{x}=0,\dot{y}=0,\dot{z}=0,\ddot{x}=0,\ddot{y}=0,\ddot{z}=0$ and $\|\boldsymbol{f}\|=0$, through the numerical root method, three collinear Lagrangian points $L_1(x_1,0,0), L_2(x_2,0,0), L_3(x_3,0,0)$, and two non-collinear Lagrangian points $L_4(x_4,y_4,0), L_5(x_5,y_5,0)$ can be calculated, where $x_4=x_5, y_4=-y_5$. Through Lyapunov stability analysis, it has been demonstrated that the collinear Lagrange points exhibit instability, whereas the triangular Lagrange points show marginal stability when the mass parameter values are below a critical threshold. These stability properties can be effectively utilized for path-planning applications.

2.2. Markov Decision Process Formulation

To align with the standard formulation of RL, the problem of optimizing Lagrange point transfer trajectories is reformulated as an MDP. An MDP usually consists of state space $\boldsymbol{S}$, action space $\boldsymbol{A}$, state transition matrix $p(\boldsymbol{s}_{k+1}|\boldsymbol{s}_k,\boldsymbol{a}_k)$, reward function r, etc. At each step, an action $\boldsymbol{a}_k$, is taken according to the current observation $\boldsymbol{o}_k$ and policy $\pi(\boldsymbol{o}_k)$, and then the environment returns a subsequent state $\boldsymbol{s}_{k+1}$ and a scalar reward value r_k, The agents interact with the environment iteratively and generate a trajectory made of a sequence of state-action pairs $\tau=\{\ \boldsymbol{s}_1,\boldsymbol{a}_1,r_1,\boldsymbol{s}_2,\boldsymbol{a}_2,r_2,...,\boldsymbol{s}_N,\boldsymbol{a}_N,r_N\}$, where N denotes the total discrete steps. In addition, some uncertainties can be considered in MDP framework, such as those arising from inaccuracies in modeling the gravitational field. where N denotes the total discrete steps. In addition, some uncertainties

can be considered in MDP framework, such as those arising from inaccuracies in modeling the gravitational field.

Then, for the Lagrange point transfer trajectory scenario, the agent state $\boldsymbol{s}_k$ at each step consists of spacecraft position $\boldsymbol{R}_k = [x_k, y_k, z_k]^T$ and velocity $\boldsymbol{V}_k = [v_{xk}, v_{yk}, v_{zk}]^T$, the agent action at each step is $\boldsymbol{a}_k = [f_{xk}, f_{yk}, f_{zk}]^T$, the agent state transition matrix of the spacecraft is obtained by numerical integration of Eq. (1). and the design of reward function r_k refers to the research of Zavoli[26] and Hu[27] et al

$$r_k = -||\Delta v_k|| - \lambda_{e_{f_z}} e_{f_z,k} - \lambda_1 e_{s,k} + c_2 e^{\lambda_2 e_{s,k}}, \tag{12}$$

where

$$\begin{cases} \Delta v_k = \dfrac{||f_k||\Delta t}{f_{\max} T} \\ e_{f_z,k} = \dfrac{||f_{z,k}||\Delta t}{f_{\max} T} \\ e_{s,k} = \begin{cases} 0 & \text{if } k < N \\ \max\{e_r, e_v\} & \text{if } k = N \end{cases} \end{cases} . \tag{13}$$

Δt is duration of each maneuver step for the spacecraft, T is total transfer time for the entire process, $f_{\max}$ is maximum thrust acceleration magnitude for the spacecraft. In Eqs.(12) and (13), the reward r_k of each step consists of a weighted combination of three components: fuel consumption $||\Delta v_k||$, penalties for out-of-plane maneuvers $\lambda_{e_{f_z}} e_{f_z,k}$ in the binary star system, and deviation from the terminal state $\lambda_1 e_{s,k} - c_2 e^{\lambda_2 e_{s,k}}$. $\lambda_{e_{f_z}}$, λ_1, c_2, $\lambda_2$are weight coefficients.

Considering practical engineering applications, the uncertainty in the gravitational field due to modeling errors of binary asteroids is predominantly taken into account. In the double ellipsoid model, the error lies in the three-axis dimensions of the ellipsoid that approximates the asteroid's shape. This study assumes that these errors follow a normal distribution, and the specific expres-

sion is

$$\begin{bmatrix} \delta\alpha_1 \\ \delta\beta_1 \\ \delta\gamma_1 \\ \delta\alpha_2 \\ \delta\beta_2 \\ \delta\gamma_2 \end{bmatrix} \sim N(0_6, \sigma_{s,k}), \sigma_{s,k} = \kappa^2 \begin{bmatrix} \bar{\alpha}_1^2 & 0 & 0 & 0 & 0 & 0 \\ 0 & \bar{\beta}_1^2 & 0 & 0 & 0 & 0 \\ 0 & 0 & \bar{\gamma}_1^2 & 0 & 0 & 0 \\ 0 & 0 & 0 & \bar{\alpha}_2^2 & 0 & 0 \\ 0 & 0 & 0 & 0 & \bar{\beta}_2^2 & 0 \\ 0 & 0 & 0 & 0 & 0 & \bar{\gamma}_2^2 \end{bmatrix}, \tag{14}$$

where $\sigma_{s,k}$ is the covariance matrix, with standard deviations $[\kappa_{11}\bar{\alpha}_1,\ \kappa_{12}\bar{\beta}_1,\ \kappa_{13}\bar{\gamma}_1,\ \kappa_{21}\bar{\alpha}_2,\ \kappa_{22}\bar{\beta}_2,\ \kappa_{23}\bar{\gamma}_2]^T$on the double ellipsoid model, $\kappa_{ij} \in [0,1], i = 1,2, j = 1,2,3$ is design parameters, and $\bar{\alpha}_1, \bar{\beta}_1, \bar{\gamma}_1, \bar{\alpha}_2, \bar{\beta}_2, \bar{\gamma}_2$ are average three-axis length of the double ellipsoid model. In the state transition calculation of the MDP, uncertainties in the gravitational field are consistently accounted for. Through RL algorithm training, the agent interacts iteratively with the environment, ultimately obtaining a neural network controller with robustness. The following is an introduction to the RL algorithm used in this paper.

3. Reinforcement Learning Algorithm

Reinforcement learning enables the agent to learn what actions to choose in different states to maximize the long-term cumulative reward by learning and making decisions under the MDP framework. Specifically, the RL algorithm trains an "expert" DNN through the interaction between the agent and the environment. The DNN can generate the optimal action sequence based on the state of each step of the agent. This section uses the PPO algorithm in the policy gradient algorithm to train the decision DNN. The policy gradient algorithm and the PPO algorithm are introduced below.

3.1. Policy Gradient Algorithm

In an episode of MDP sequence $\tau = \{\ \boldsymbol{s}_1, \boldsymbol{a}_1, r_1, \boldsymbol{s}_2, \boldsymbol{a}_2, r_2, ..., \boldsymbol{s}_N, \boldsymbol{a}_N, r_N\}$, the state-action value function $Q(\boldsymbol{s}, \boldsymbol{a})$ is used to evaluate the policy network π_θ, where θ is the network parameters. The $Q(\boldsymbol{s}, \boldsymbol{a})$ refers to the expectation of the

sum of reward values that can be obtained after executing action $\boldsymbol{a}$ from the current state $\boldsymbol{s}$

$$Q^{\pi}(\boldsymbol{s}, \boldsymbol{a}) = E_{\pi}\left\{\sum_{k=0}^{\infty} \gamma^{k}\, r(\boldsymbol{s}_{t+k}, \boldsymbol{a}_{t+k}) | \left(\boldsymbol{s}_{t} = \boldsymbol{s}, \boldsymbol{a}_{t} = \boldsymbol{a}\right)\right\}. \tag{15}$$

The impact of actions taken in the present moment on the future will gradually decay. Therefore, the reward value attenuation factor $\gamma = [0, 1]$ is introduced into the state-action value function. $\gamma = 1$ means that the impact of the current moment on the future does not decay with time, $\gamma = 0$ means that the current moment has no impact on the future.

The probability $\pi_{\theta}(\tau)$ that the policy network π_{θ} generates a certain action sequence τ is

$$\pi_{\theta}(\tau) = p(\boldsymbol{s}_1) \prod_{t=1}^{T} \pi_{\theta}(\boldsymbol{a}_t \mid \boldsymbol{s}_t) p(r_t, \boldsymbol{s}_{t+1} \mid \boldsymbol{s}_t, \boldsymbol{a}_t). \tag{16}$$

In Eq.(16), $p(\boldsymbol{s}_1)$ represents the probability that the environment state is $\boldsymbol{s}_1$, $p(r_t, \boldsymbol{s}_{t+1} \mid \boldsymbol{s}_t, \boldsymbol{a}_t)$ represents the probability that the environment state is $\boldsymbol{s}_t$, and after executing the action $\boldsymbol{a}_t$, the environment state becomes $\boldsymbol{s}_{t+1}$ and the reward value r_t is obtained.

Thus, the reward expectation obtained by the policy network in multiple action sequences is

$$J = \int \pi_{\theta}(\tau)\, Q^{\pi}(\tau)\, d\tau. \tag{17}$$

The optimization goal of the policy network is to maximize the reward expectation, that is

$$\theta^{*} = \arg\max J. \tag{18}$$

The above formula can be obtained by calculating the gradient of the network parameters θ

$$\nabla_{\theta} J(\theta) = \int \nabla_{\theta} \pi_{\theta}(\tau)\, Q(\tau)\, d\tau. \tag{19}$$

Noticed

$$\nabla_{\theta} \pi_{\theta}(\tau) = \pi_{\theta}(\tau) \frac{\nabla_{\theta} \pi_{\theta}(\tau)}{\pi_{\theta}(\tau)} = \pi_{\theta}(\tau) \nabla_{\theta} \log \pi_{\theta}(\tau), \tag{20}$$

substituting into the gradient formula Eq.(20), we have

$$\nabla_\theta J(\theta) = \int \pi_\theta(\tau) \nabla_\theta \log \pi_\theta(\tau) Q^\pi(\tau) d\tau = E_{\tau \in \pi_\theta} [\nabla_\theta \log \pi_\theta(\tau) Q(\tau)] \quad (21)$$

From Eq.(16), we can get

$$\nabla_\theta \log \pi_\theta(\tau) = \sum_{t=1}^{T} \nabla_\theta \log \pi_\theta(\boldsymbol{a}_t|\boldsymbol{s}_t). \quad (22)$$

In the actual training process, using N rounds of data to estimate the expected value of the gradient, we can get

$$\nabla_\theta J(\theta) = \frac{1}{N} \sum_{n=1}^{N} \sum_{i=1}^{T} \nabla_\theta \log \pi_\theta(\boldsymbol{a}_t^n \mid \boldsymbol{s}_t^n) Q(\boldsymbol{s}_t^n, \boldsymbol{a}_t^n).. \quad (23)$$

After calculating the gradient, back propagation updates the parameters of the policy network π_θ

$$\theta^{new} \leftarrow \theta^{old} + \eta \nabla_\theta J\left(\theta^{old}\right). \quad (24)$$

As training progresses, the probability that an action with a larger state-action value function will be selected will increase, while the probability that an action with a smaller state-action value function will be selected will decrease.

3.2. Actor-Critic Method and PPO Algorithm

If the variance of the cumulative reward output by the environment is large, the policy network will converge slowly. Add a baseline value to the reward can solve this problem. The Actor-Critic method is to subtract the state value function $V^\pi(\boldsymbol{s}_t)$ as a baseline from the action value function $Q^\pi(\boldsymbol{s}_t, \boldsymbol{a}_t)$ to obtain the advantage function

$$A^\pi(\boldsymbol{s}_t, \boldsymbol{a}_t) = Q^\pi(\boldsymbol{s}_t, \boldsymbol{a}_t) - V^\pi(\boldsymbol{s}_t), \quad (25)$$

where

$$V^\pi(\boldsymbol{s}_t) = E_\pi \left\{ \sum_{k=0}^{\infty} \gamma^k \, r(\boldsymbol{s}_{t+k}, \boldsymbol{a}_{t+k}) | \, (\boldsymbol{s} = \boldsymbol{s}_t) \right\}. \quad (26)$$

Substituting Eq.(26) into Eq. (25), we can get

$$\nabla_\theta J(\theta) = \frac{1}{N} \sum_{n=1}^{N} \sum_{t=1}^{T} \nabla_\theta \log \pi_\theta(\boldsymbol{a}_t^n \mid \boldsymbol{s}_t^n) A(\boldsymbol{s}_t^n, \boldsymbol{a}_t^n). \quad (27)$$

The idea of Actor-Critic is to use the value network to fit the state value function $V^{\pi}(\boldsymbol{s}_t)$. The loss value of the value network is

$$L(\phi) = \sum_{t} \|r_t + \gamma V\left(\boldsymbol{s}_{t+1}\right) - V\left(\boldsymbol{s}_t\right)\|. \tag{28}$$

Actor-Critic method is obtained by combining the value network and the policy network, as shown in Figure 2. The value network is equivalent to the scorer of the strategy network, giving the average score of the policy network. The policy network updates the probability of selecting an action based on the difference between the reward value of the real environment and the average score.

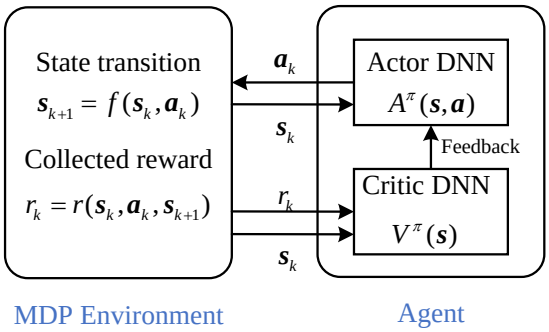

Figure 2: The Actor-Critc method.

If the single-step update of the policy network is too large, the network will not converge. Assume that the policy network before the update is π, and the policy network after the update is π', the objective function before and after the update changes to

$$L_{\pi}(\pi') = J(\pi') - J(\pi) = E_{\tau \in \pi'}\left[\sum_{t=1}^{T} A^{\pi}\left(\boldsymbol{s}_t, \boldsymbol{a}_t\right)\right]. \tag{29}$$

Approximately

$$L_{\pi}\left(\pi'\right) = E\left[\sum_{t=1}^{T} \frac{\pi'(\boldsymbol{a}_t \mid \boldsymbol{s}_t)}{\pi(\boldsymbol{a}_t \mid \boldsymbol{s}_t)} A^{\pi}\left(\boldsymbol{s}_t, \boldsymbol{a}_t\right)\right]. \tag{30}$$

Therefore, the training objective function of the policy network π_{θ} can also be expressed as maximizing the expected growth of the reward value after the

update

$$\theta_{k+1} = \arg\max L_{\theta_k}(\pi_\theta). \tag{31}$$

Add a constraint on the update step size to limit the update of the policy network to the following range

$$1 - \varepsilon \leqslant r_t(\theta) = \frac{\pi_{k+1}(\boldsymbol{a}_t \mid \boldsymbol{s}_t)}{\pi_k(\boldsymbol{a}_t \mid \boldsymbol{s}_t)} \leqslant 1 + \varepsilon. \tag{32}$$

Substitute constraints into the objective function and improve it to truncate the optimization objective

$$L_{\theta_k}^{CLIP}(\theta) = E\left[\sum_{t=1}^{T} \min\left[r_t(\theta) A^{\pi_k}(\boldsymbol{s}_t, \boldsymbol{a}_t), \operatorname{clip}\left[r_t(\theta), 1-\varepsilon, 1+\varepsilon\right] A^{\pi_k}(\boldsymbol{s}_t, \boldsymbol{a}_t)\right]\right]. \tag{33}$$

4. Numerical simulations

In order to verify the effectiveness of the reinforcement learning algorithm in the orbit transfer task in the double asteroid system, this section uses the double asteroid system 809 Lundia as the basis to optimize the simulation calculation of the Lagrangian point transfer trajectory. physical parameters of 809 Lundia is shown in Table 1.

Table 1: Physical parameters of 809 Lundia

parameters	the primary asteroid M_1	the secondary asteroid M_2
The distance L between the centers of mass of the asteroids(km)	15.87	
Density of the asteroids(g/cm^3)	1.67	
Semi-major axis(km)	3.9	3.5
Semi-middle axis(km)	3.3	2.9
Semi-minor axis(km)	3.2	2.8

It can be determined from Table 1 that the mass ratio μ of the binary asteroid system is 0.4083, and the normalized length UL is 3.9 km, the normalized time UT is 1351.0816s. the normalized position of M_1 and M_2 are

[- 1.6615,0,0]^T and [2.4077,0,0]^T .The rotational angular velocity of the binary asteroid system in equilibrium is 9.1105 $\times$ 10^5 rad/s. The dimensionless angular velocity of the binary asteroid system is 0.1231. The position of the equilibrium point is solved by the method in Section 2.1, and the result is $L_1(0.5253, 0, 0)$,$L_2(4.9980, 0, 0)$,L_3(- 4.7373, 0, 0),L_4(0.3668,3.4956,0), L_5(0.3668, - 3.4956,0), See the Figure 3.

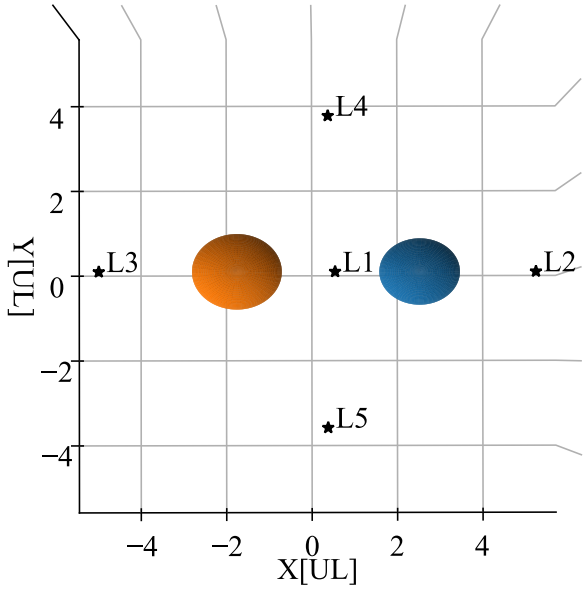

Figure 3: The Lagrangian points of 809 Lundia.

The transfer trajectory optimization simulation scenarios are selected as L2 to L4 and L2 to L3. The acceleration of the detector is a constant value of 5 mm/s^2, and the transfer time of the two messions is set to 5000s and 8000 s respectively.

In implementing PPO using Stable-Baselines, an open-source RL library, default network architectures for the Actor and Critic are employed based on the PPO2 algorithm. Following Zavoli[26] 's method for adjusting RL hyperparameters, this study fine-tunes PPO hyperparameters using Optuna[32], an open-source framework for automated hyperparameter search. Table 2 presents the hyperparameter configurations for the PPO reinforcement learning algo-

rithm.

Table 2: Hyperparameter settings of PPO

Hyperparameter	γ	λ	α	ε	β_1	β_2	n_{opt}
Value	0.9999	0.99	2.5×10^{-4}	0.3	0.5	4.75×10^{-8}	30

To assess the effectiveness of RL, the solution under nominal gravitational conditions is initially compared with the optimal solution derived from an indirect method. Subsequently, the policy network's robustness and optimality under various uncertain scenarios are evaluated using Monte Carlo simulations.

4.1. L2-L4 Transfer

4.1.1. Performance in the deterministic scenario

In the L2-L4 transfer problem, the transfer time T is set to 5000s. The optimization result of is 0.01254 km/s, and the transfer trajectory is shown in Figure 4. The result of the indirect method optimization is 0.011883 km/s. The thrust curves of the two methods are shown in Figures 5 and 6. It can be seen from the figures that the fuel consumption and thrust acceleration curves of RL are closer to those of the indirect method. The difference in fuel consumption is about 6e-4km/s, because RL is a piecewise constant thrust. It can be proved that the optimization results of RL are effective. However, the indirect method cannot take into account the uncertainty in dynamic modeling. Next, the uncertainty in dynamic modeling is considered to perform robust optimization of the spacecraft's transfer trajectory.

4.1.2. Performance in uncertain scenario

For convenience, the design parameters $\kappa_{ij} \in [0, 1], i = 1, 2, j = 1, 2, 3$ of shape uncertainty of the double ellipsoid model are equal during simulation. The standard deviation of the modeling uncertainty of the double ellipsoid models set to $\kappa = 0.1\%\tilde{}0.5\%$ average three-axis length. RL is used to train the decision network for robust trajectory optimization, and 500 Monte Carlo simulations

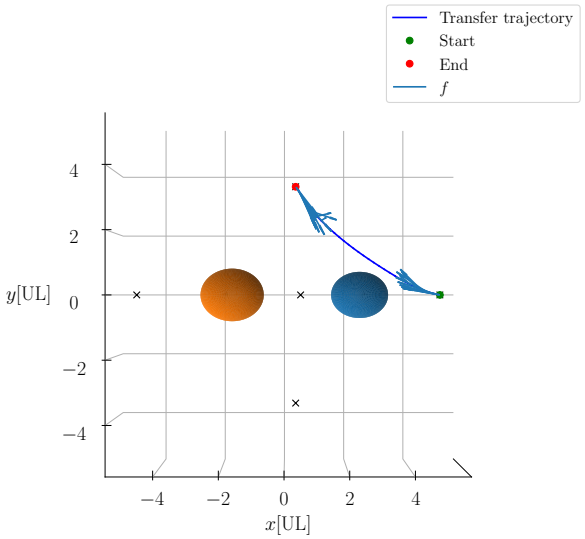

Figure 4: The L2-L4 transfer trajectory of RL in the deterministic scenario.

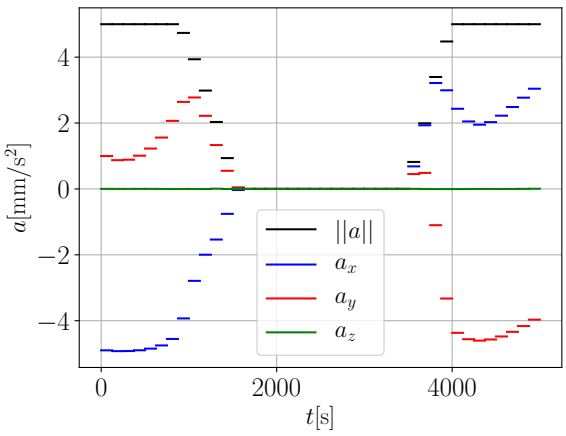

Figure 5: The L2-L4 thrust curve of RL in the deterministic scenario.

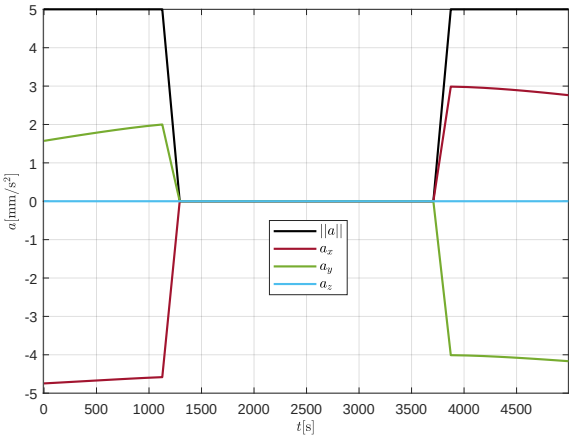

Figure 6: The L2-L4 thrust curve of indirect method.

are performed. The simulation results are shown in Table 3. As can be seen from Table 3, the fuel consumption and terminal error are positively correlated with the modeling error, but the standard deviation of the fuel consumption and terminal error always remains at the same order of magnitude.

Table 3: Monte Carlo simulation results

Uncertainty of Dynamic model (%)	Fuel (km/s)		Position error (km)		Velocity error (km/s)	
	Mean	Std	Mean	Std	Mean	Std
0.1	0.0134	5.10E-06	0.0228	3.30E-03	9.60E-06	2.60E-06
0.2	0.0153	5.80E-06	0.0399	3.90E-03	3.40E-05	2.50E-06
0.3	0.0168	2.70E-06	0.1731	3.30E-03	1.00E-04	3.30E-06
0.4	0.0191	4.40E-06	0.2009	4.80E-03	2.50E-04	3.70E-06
0.5	0.0184	7.00E-06	0.3054	2.70E-03	1.80E-04	5.00E-07

Taking the uncertainty of dynamic model of 0.5% as an example, Figure 7 displays the trajectories from 500 Monte Carlo simulations. The differences between each Monte Carlo sample trajectory and the corresponding nominal trajectory are magnified by a factor of 50 for clarity. Additionally, Figure 8 depicts the thrust curve of the nominal trajectory. It can be seen that compared with the trajectory Figure 4 under the deterministic scenario, the trajectory of spacecraft in Figure 7 is farther away from the secondary celestial body. It can be understood that the farther away from the binary asteroid system, the smaller

the influence of the modeling error on the spacecraft, so that the terminal error stability can be maintained.

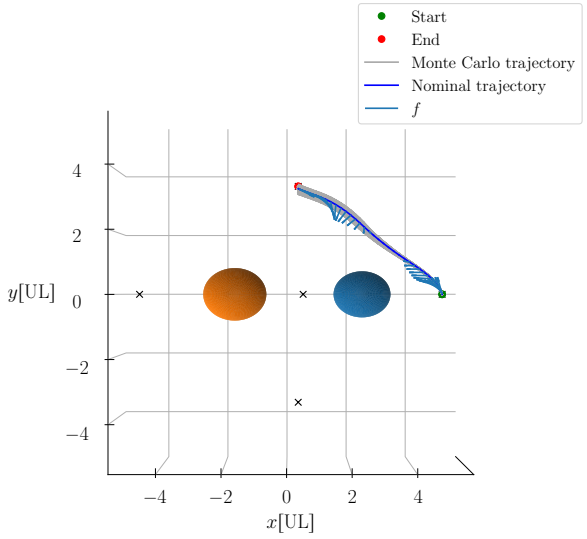

Figure 7: The L2-L4 trajectory of 500 Monte Carlo simulations with uncertainty of dynamic model of 0.5%.

The policy network of the deterministic scenario is placed in a dynamic environment with uncertainty 0.5% for comparison, and the Monte Carlo simulation is also performed. The simulation results are shown in Table 4. It can be seen that compared with the policy network considering uncertainty, although the fuel consumption and terminal error of the decision network of the deterministic scenario are less, the standard deviation of the fuel consumption and terminal error is 1 order of magnitude higher, indicating that the decision network considering the uncertainty of dynamic modeling generated based on RL training can generate a more robust trajectory control strategy.

4.2. L2-L3 Transfer

4.2.1. Performance in the deterministic scenario

In the L2-L3 transfer problem, the indirect method cannot handle the collision between the spacecraft and the binary asteroid system, and the performance of RL in solving the optimal transfer trajectory of fuel has been verified

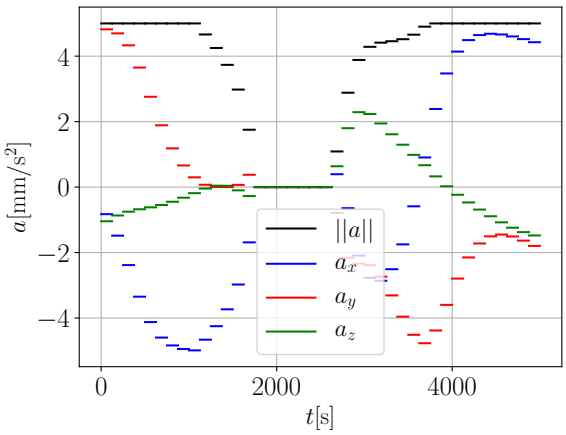

Figure 8: The L2-L4 thrust curve of the nominal trajectory with uncertainty of dynamic model of 0.5%.

Table 4: The L2-L4 Monte Carlo simulation results of the policy network trained in deterministic scenario.

Fuel(km/s)		Position error (km)		Velocity error(km/s)	
Mean	Std	Mean	Std	Mean	Std
0.0125	1.41E-04	0.033	2.60E-05	1.20E-05	1.00E-05

by the indirect method in L2-L4, so only RL is used here to solve the transfer trajectory. In this problem, the transfer time T is set to 8000s. The optimization result of RL is 0.01563 km/s, and the transfer trajectory and the thrust curve are shown in Figure 9 and Figures 10. Next, the uncertainty in dynamic modeling is considered and the transfer trajectory of the spacecraft is robustly optimized.

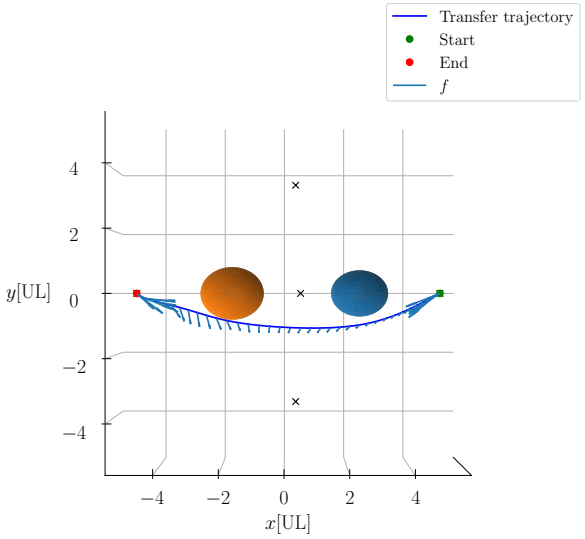

Figure 9: The L2-L3 transfer trajectory of RL in the deterministic scenario.

4.2.2. Performance in uncertain scenarios

The standard deviation of the modeling uncertainty of the double ellipsoid models set to $\kappa = 0.1\%\tilde{}0.5\%$ average three-axis length. RL is used to train the policy network for robust trajectory optimization and 500 Monte Carlo simulations are performed. The simulation results are shown in Table 5. As can be seen from Table 5, the fuel consumption and terminal error are positively correlated with the modeling error, but the standard deviation of the fuel consumption and terminal error always remains at the same order of magnitude. At the same time, it can be seen that under the same dynamic modeling error, when the fuel consumption is small, the terminal error will be larger.

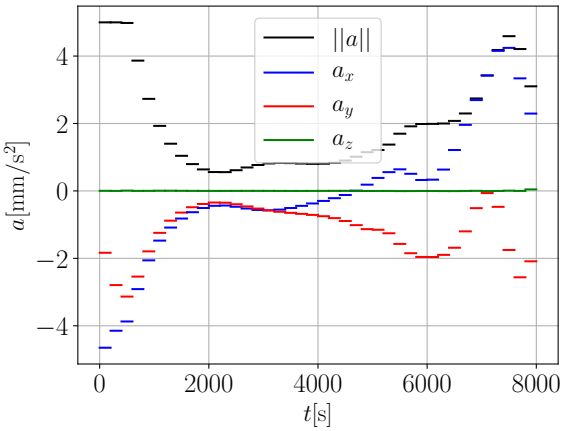

Figure 10: The L2-L3 thrust curve of RL in the deterministic scenario.

Table 5: Monte Carlo simulation results

Dynamic modeling error (%)	Fuel(km/s)		Position error(km)		Velocity error(km/s)	
	Mean	Std	Mean	Std	Mean	Std
0.1	0.017	1.70E-05	0.0585	7.70E-03	5.70E-05	3.60E-06
0.2	0.0189	2.60E-05	0.0707	3.10E-03	2.10E-05	3.30E-06
0.3	0.0245	2.80E-05	0.1058	2.80E-03	1.50E-04	2.80E-06
0.4	0.0222	1.40E-05	0.2355	2.30E-03	1.60E-04	3.30E-06
0.5	0.0238	7.80E-06	0.6259	2.80E-03	7.40E-04	6.20E-06

Taking the dynamic modeling error of 0.5% as an example, the trajectory of 500 Monte Carlo simulations is shown in Figure 11, and the thrust curve of the nominal trajectory is shown in Figure 12. It can be seen that compared with the trajectory Figure 9 under the deterministic scenario, the trajectory of spacecraft in Figure 11 is farther away from the secondary celestial body. It can be understood that the farther away from the binary asteroid system, the smaller the impact of the modeling error on the spacecraft, so that the terminal error stability can be maintained.

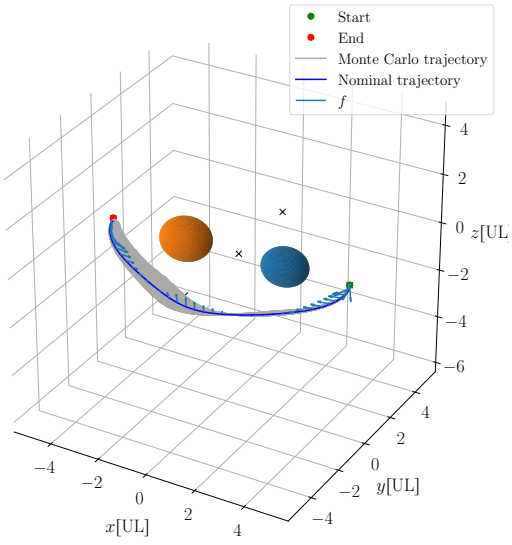

Figure 11: The L2-L3 trajectory of 500 Monte Carlo simulations with uncertainty of dynamic model of 0.5%.

The policy network of the deterministic scenario is placed in a dynamic environment with a modeling error of 0.5% for comparison, and a Monte Carlo simulation is performed. The simulation results are shown in Table 6. It can be seen that compared with the decision network considering uncertainty, although the fuel consumption and terminal error of the decision network of the deterministic scenario are smaller, the standard deviation of the fuel consumption and terminal error is 1 to 2 orders of magnitude higher, indicating that the

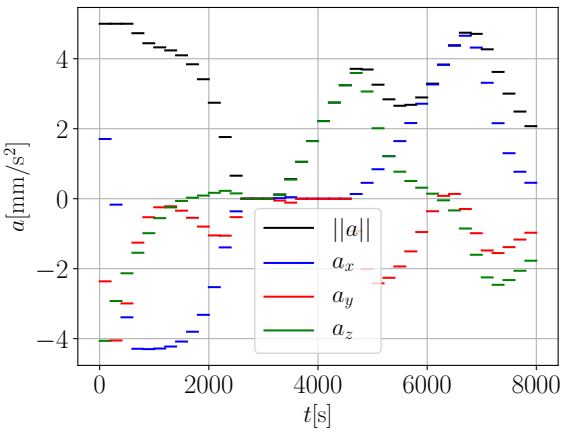

Figure 12: The L2-L3 thrust curve of the nominal trajectory with uncertainty of dynamic model of 0.5%.

policy network considering dynamic modeling uncertainty generated based on RL training can generate a more robust orbit control strategy.

Table 6: The L2-L3 Monte Carlo simulation results of the policy network trained in deterministic scenario.

Fuel(km/s)		Position error (km)		Velocity error(km/s)	
Mean	Std	Mean	Std	Mean	Std
0.0155	1.70E-04	0.097	8.90E-02	5.20E-05	3.10E-05

5. Conclusion

This study applies RL to solve the robust optimization problem of transfer trajectories between Lagrangian points in a binary asteroid system, considering uncertainties in dynamic modeling. Initially, a second-order gravitational field model for the binary asteroid system, based on a double ellipsoid model, was established. Dynamic uncertainties were modeled as Gaussian uncertainties across the ellipsoid's three axes.

Subsequently, the orbit optimization problem was formulated as a MDP, and a suitable reward function was devised to enable RL-based solution methods. The study focused on solving transfer problems between L2-L4 and L2-L3 Lagrangian points, demonstrating through simulations that the RL-based policy network, when dynamic uncertainty is disregarded, can produce orbital maneuvering strategies akin to those derived from indirect methods, thus validating the efficacy of RL.

Furthermore, the study explores robust orbit optimization under dynamic uncertainty. Monte Carlo simulations reveal that RL can generate more resilient orbital maneuvering strategies. Compared to strategies ignoring dynamic uncertainty, these RL-based strategies exhibit an order of magnitude reduction in terminal error standard deviation, ensuring more stable transfer trajectories.

Acknowledgment

This work was supported in part by the National Natural Science Foundation of China (grant number U23B6001) and in part by the National Natural Science Foundation of China (grant number U22B2013).

References

[1] M. Yoshikawa, J. Kawaguchi, A. Fujiwara, A. Tsuchiyama, Hayabusa sample return mission, *Asteroids*, IV 1 (1) (2015).

[2] S.-i. Watanabe, Y. Tsuda, M. Yoshikawa, S. Tanaka, T. Saiki, S. Nakazawa, Hayabusa2 mission overview, *Space Science Reviews*, 208 (2017) 3–16, https://doi.org/10.1007/s11214-017-0377-1.

[3] D. S. Lauretta, S. S. Balram-Knutson, E. Beshore, W. V. Boynton, C. Drouet d'Aubigny, D. N. DellaGiustina, H. L. Enos, D. R. Golish, C. W. Hergenrother, E. S. Howell, OSIRIS-REx: sample return from asteroid (101955) Bennu, *Space Science Reviews*, 212 (2017) 925–984, https://doi.org/10.1007/s11214-017-0405-1.

[4] R. T. Daly, C. M. Ernst, O. S. Barnouin, N. L. Chabot, A. S. Rivkin, A. F. Cheng, E. Y. Adams, H. F. Agrusa, E. D. Abel, A. L. Alford, Successful kinetic impact into an asteroid for planetary defence, *Nature*, 616 (7957), https://doi.org/10.1038/s41586-023-05810-5.

[5] I. Jean, A. K. Misra, A. Ng, Controlled spacecraft trajectories in the context of a mission to a binary asteroid system, *The Journal of the Astronautical Sciences*, 68 (2021) 38–70, https://doi.org/10.1007/s40295-021-00248-1.

[6] R. Zhang, Y. Wang, Y. Shi, S. Xu, Libration points and periodic orbit families near a binary asteroid system with different shapes of the secondary, *Acta Astronautica*, 177 (2020) 15–29, https://doi.org/10.1016/j.actaastro.2020.07.006.

[7] X. Li, D. Qiao, P. Li, Bounded trajectory design and self-adaptive maintenance control near non-synchronized binary systems comprised of small irregular bodies, *Acta Astronautica*. 152 (2018) 768–781, https://doi.org/10.1016/j.actaastro.2018.09.028.

[8] I. Fodde, J. Feng, A. Riccardi, M. Vasile, Robust stability and mission performance of a CubeSat orbiting the Didymos binary asteroid system, *Acta Astronautica*, 203 (2023) 577–591, https://doi.org/10.1016/j.actaastro.2022.12.021.

[9] G. Gómez, W. S. Koon, M. W. Lo, J. E. Marsden, J. Masdemont, S. D. Ross, Connecting orbits and invariant manifolds in the spatial restricted three-body problem, *Nonlinearity*, 17 (5) (2004) 1571, https://doi.org/10.1088/0951-7715/17/5/002.

[10] D. J. Scheeres, The dynamical evolution of uniformly rotating asteroids subject to YORP, *ICARUS*, 188 (2) (2007) 430–450, https://doi.org/10.1016/j.icarus.2006.12.015.

[11] X. Wu, H. Shang, X. Qin, Transfer trajectory design about doubly synchronous binary asteroid system, in: *54th AIAA Aerospace Sciences Meeting*, 2016, p. 0478.

[12] O. Çelik, J. P. Sánchez, Opportunities for ballistic soft landing in binary asteroids, *Journal of Guidance, Control, and Dynamics*, 40 (6) (2017) 1390–1402, https://doi.org/10.2514/1.G002181.

[13] F. Ferrari, M. Lavagna, Ballistic landing design on binary asteroids: The AIM case study, *Advances in Space Research*, 62 (8) (2018) 2245–2260, https://doi.org/10.1016/j.asr.2017.11.033.

[14] T. Wen, X. Zeng, Landing simulation in the full two-body problem of binary asteroids, *Journal of Guidance, Control, and Dynamics*, 46 (5) (2023) 885–899, https://doi.org/10.2514/1.G006526.

[15] L. Federici, A. Scorsoglio, L. Ghilardi, A. D'Ambrosio, B. Benedikter, A. Zavoli, R. Furfaro, Image-Based Meta-Reinforcement Learning for Autonomous Guidance of an Asteroid Impactor, *Journal of Guidance, Control, and Dynamics*, 45 (11) (2022) 2013-2028, https://arc.aiaa.org/doi/10.2514/1.G006832.

[16] C. C. Merrill, C. J. Geiger, A. T. M. Tahsin, D. Savransky, M. Peck, Creating a contact binary via spacecraft impact to near-Earth binary asteroid (350751) 2002 AW,*Acta Astronautica*, 214 (2024) 629–640, https://doi.org/10.1016/j.actaastro.2023.11.030.

[17] F. Ferrari, M. Lavagna, K. C. Howell, Dynamical model of binary asteroid systems through patched three-body problems, *Celestial Mechanics and Dynamical Astronomy*, 125 (2016) 413–433, https://doi.org/10.1007/s10569-016-9688-x.

[18] A. D'Ambrosio, A. Carbone, F. Curti, Optimal Maneuvers Around Binary Asteroids Using Particle Swarm Optimization and Machine Learn-

ing, *Journal of Spacecraft and Rockets*, 60 (5) (2023): 1458-1472, https://arc.aiaa.org/doi/10.2514/1.A35317.

[19] K. Parmar, E. Taheri, D. Guzzetti, Comparison of LearningSpacecraft Path-Planning Solutions from Imitation in Three-Body Dynamics, *Journal of Spacecraft and Rockets*, 60 (3) (2023) 699–715, https://arc.aiaa.org/doi/10.2514/1.A35458.

[20] S. da Graça Marto, M. Vasile, robust trajectory optimisation under epistemic uncertainty and imprecision,*Acta Astronautica*, 191 (2022) 99–124, https://doi.org/10.1016/j.actaastro.2021.10.022.

[21] W. Li, Y. Song, L. Cheng, S. Gong, Closed-loop deep neural network optimal control algorithm and error analysis for powered landing under uncertainties, *Astrodynamics*, 7 (2) (2023) 211–228, https://doi.org/10.1007/s42064-022-0153-1.

[22] S. Kelly, D. Geller,Robust Cislunar Trajectory Optimization in the Presence of Stochastic Errors, *The Journal of the Astronautical Sciences*, 71 (4) (2024) 30, https://doi.org/10.1007/s40295-024-00450-x.

[23] C. Greco, S. Campagnola, M. Vasile, Robust space trajectory design using belief optimal control, *Journal of Guidance, Control, and Dynamics*, 45 (6) (2022) 1060–1077, https://doi.org/10.2514/1.G005704.

[24] H. Yuan, D. Li, G. He, J. Wang, Uncertainty-resilient constrained rendezvous trajectory optimization via stochastic feedback control and unscented transformation, *Acta Astronautica*, 219 (2024) 264–277, https://doi.org/10.1016/j.actaastro.2024.03.017.

[25] W. Feng, M. Huo, Y. Xu, L. Mo, W. Ke, Y. Ma, H. Su, N. Qi, A framework of gravity field online modeling and trajectory optimization in asteroid soft-landing mission scenarios, *Aerospace Science and Technology*, 143 (2023): 108656. https://doi.org/10.1016/j.ast.2023.108656.

[26] A. Zavoli, L. Federici, Reinforcement learning for robust trajectory design of interplanetary missions, *Journal of Guidance, Control, and Dynamics*, 44 (8) (2021) 1440–1453, https://doi.org/10.2514/1.G005794.

[27] J. Hu, H. Yang, S. Li, Y. Zhao, Densely rewarded reinforcement learning for robust low-thrust trajectory optimization, *Advances in Space Research*, 72 (4) (2023) 964–981, https://doi.org/10.1016/j.asr.2023.03.050.

[28] L. Federici, A. Zavoli, Robust interplanetary trajectory design under multiple uncertainties via meta-reinforcement learning, *Acta Astronautica*, 214 (2024) 147-158, https://doi.org/10.1016/j.actaastro.2023.10.018.

[29] H. Yuan, D. Li, J. Wang, Integrated robust navigation and guidance for the kinetic impact of near-earth asteroids based on deep reinforcement learning,*Aerospace Science and Technology*, 142 (2023) 108666, https://doi.org/https://doi.org/10.1016/j.ast.2023.108666.

[30] S. Boone, S. Bonasera, J. W. McMahon, N. Bosanac, N. R. Ahmed, Incorporating observation uncertainty into reinforcement learning-based spacecraft guidance schemes, in: *AIAA SciTech 2022 Forum*, 2022, p. 1765.

[31] H. Yang, J. Hu, S. Li, X. Bai, Reinforcement-Learning-Based Robust Guidance for Asteroid Approaching, *Journal of Guidance, Control, and Dynamics*, (2024) 1–15, https://doi.org/10.2514/1.G008085.

[32] T. Akiba, S. Sano, T. Yanase, T. Ohta, M. Koyama, Optuna: A next-generation hyperparameter optimization framework, in: *Proceedings of the 25th ACM SIGKDD international conference on knowledge discovery & data mining*, 2019, pp. 2623–2631.

Declaration of interests

☒ The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

☐The authors declare the following financial interests/personal relationships which may be considered as potential competing interests: