The Journal of Navigation (2022), **75**:3 662–684
doi:10.1017/S0373463321000813

CAMBRIDGE UNIVERSITY PRESS

RESEARCH ARTICLE

A spacecraft attitude manoeuvre planning algorithm based on improved policy gradient reinforcement learning

Bing Hua,* Shenggang Sun, Yunhua Wu, and Zhiming Chen

School of Astronautics, Nanjing University of Aeronautics and Astronautics, Nanjing, China.
*Corresponding author: E-mail: huabing@nuaa.edu.cn

Received: 27 January 2021; **Accepted:** 7 October 2021; **First published online:** 14 December 2021

Keywords: spacecraft attitude manoeuvre; attitude constraints; path planning; reinforcement learning; policy gradient

Abstract

To solve the problem of spacecraft attitude manoeuvre planning under dynamic multiple mandatory pointing constraints and prohibited pointing constraints, a systematic attitude manoeuvre planning approach is proposed that is based on improved policy gradient reinforcement learning. This paper presents a succinct model of dynamic multiple constraints that is similar to a real situation faced by an in-orbit spacecraft. By introducing return baseline and adaptive policy exploration methods, the proposed method overcomes issues such as large variances and slow convergence rates. Concurrently, the required computation time of the proposed method is markedly reduced. Using the proposed method, the near optimal path of the attitude manoeuvre can be determined, making the method suitable for the control of micro spacecraft. Simulation results demonstrate that the planning results fully satisfy all constraints, including six prohibited pointing constraints and two mandatory pointing constraints. The spacecraft also maintains high orientation accuracy to the Earth and Sun during all attitude manoeuvres.

1. Introduction

When performing space missions, spacecraft must often execute many attitude manoeuvres. To gain energy, the solar energy cell array of the spacecraft must always be kept facing the Sun. For Earth-observing satellites, the satellite's camera must maintain its attitude requirement relative to the orientation to the Earth. Although sensors cannot be mounted in the same direction as the solar energy cell array, they must also not be exposed to sunlight. In addition to these attitude constraints, it is also necessary to consider other constraints, such as communication, suppression of stray light and avoidance of space debris. In summary, spacecraft face multiple mandatory pointing constraints and prohibited pointing constraints during orbiting operations. Specifically, when performing Earth observation missions, certain parts of the spacecraft must meet the requirements of orientation to the Earth and the Sun. Concurrently, because the surface of the Earth reflects sunlight, there are multiple reflected light areas in space. The sensors must avoid both direct and reflected sunlight. Considering these constraints, it is necessary to plan attitude change trajectories that satisfy all constraints.

Scholars have investigated the issue of spacecraft attitude manoeuvre planning. Currently, related research can be divided into two categories. One primary concept is online planning methods, such as semidefinite programming (Hablani, 1999), constraint monitoring algorithms (Singh et al., 1997) and randomised attitude planning algorithms (Kim and Mesbahi, 2004; Olivares and Staffetti, 2018). However, these methods have disadvantages such as poor universality and a large computing time, which means that they are difficult to use in practical engineering applications. Celani and Lucarelli (2019) proposed an attitude motion planning approach that uses gradient-based optimisation and special orthogonal group SO(3). However, that method was only used to perform rest-to-rest manoeuvres.

© The Author(s), 2021. Published by Cambridge University Press on behalf of The Royal Institute of Navigation

https://doi.org/10.1017/S0373463321000813 Published online by Cambridge University Press

Xu et al. (2020) proposed a new dynamic-iteration-based multi-objective planning method. However, that method linearises the constraints and relaxes the requirement for the highly accurate satisfaction of the middle target nodes. Cui et al. (2017) proposed an onboard spacecraft slew-planning algorithm. However, that method is not suitable for solving problems facing multiple constraints over a long period of time. Kjellberg and Lightsey (2016) and Tanygin (2017) separately used the A* pathfinding algorithm to find the optimal path away from the forbidden zones but did not consider both mandatory pointing constraints and forbidden pointing constraints. However, the A* algorithm tends to make a spacecraft perform ineffective attitude manoeuvres. The same features of particle swarm optimisation were also used by Spiller et al. (2016), Showalter (2014) and Siliang and Shijie (2010). Miao and Pan (2015) presented a method of encoding based on proportional-differential control. However, that method only considered a few forbidden pointing constraints.

Another primary category is deterministic analytic methods, which primarily includes the potential function method. Lee and Mesbahi (2011) proposed a convex logarithmic potential function based on the parameterised convex description model of attitude constraints established by Kim et al. (2015). Wu et al. (2015) proposed a multi-constraint attitude manoeuvre planning method based on logarithmic potential functions. However, the process of constructing the logarithmic potential function is complex, and the algorithm is only verified in a two-dimensional plane. Ma et al. (2020) built a general description of spacecrafts' conical forbidden pointing attitude set and then transformed it into the convex constraints set. Then, a new potential function was proposed to avoid the local minimum of the traditional potential function. However, that method only considers two forbidden pointing constraints within virtual three-dimensional spheres.

The potential function method can easily fall into a local optimum, and there is the problem of GNRON (goals non-reachable with obstacles nearby) (Guo et al., 2020). It is thus not conducive to the expansion of the attitude constraint model. Most related studies simplify the existence conditions of the repulsive potential function and set the parameters improperly, which easily leads to unnecessary attitude manoeuvres (Chen et al., 2019). Existing research that uses a potential function to solve the attitude manoeuvre planning problem stays in the static pointing constraint in the virtual space, and the total number of constraints is limited to two or three, which is not practical.

The reinforcement learning algorithm has absorbed the advantages of both optimal control theory and adaptive control theory (Cheng, 2019; Huang et al., 2021); it does not require a complex and tedious modelling process and does not require the system to be fully characterised. More importantly, it can be used for online planning, which is suitable for solving nonlinear problems in continuous space (Reed, 2020). The attitude manoeuvre planning problem proposed in this paper is a nonlinear high-dimensional continuous time-varying optimisation problem, which is suitable for solving by reinforcement learning.

A spacecraft in orbit faces many dynamic constraints; thus, there is a need for attitude manoeuvre planning. The problems of existing attitude manoeuvre planning methods include being too complex, having poor generality and yielding low-accuracy solutions. Thus this paper proposes an attitude manoeuvre planning algorithm based on improved policy gradient reinforcement learning. Based on the policy gradient reinforcement learning algorithm, the baseline of return and adaptive policy exploration are introduced. The improved model is simple and has a clear physical meaning, which has strong generality and can provide good expansion. Simulation results demonstrate that the planning results allow the spacecraft to avoid multiple dynamic attitude constraint areas successfully and maintain a high orientation accuracy to the Earth and Sun.

The remainder of this paper is organised as follows. Section 2 describes the attitude manoeuvre planning problem and explains the process of model construction, including reference attitude constraints in detail. Section 3 introduces the scheme of attitude manoeuvre planning and then introduces the process and principles of the attitude manoeuvre planning algorithm based on improved policy gradient reinforcement learning. Section 4 shows the simulation results and analysis. Section 5 concludes the paper.

https://doi.org/10.1017/S0373463321000813 Published online by Cambridge University Press

2. Model construction

The spacecraft in this paper refers to a three-axis stabilisation spacecraft that is used for Earth observation and carries multiple sensors, which cannot work normally when exposed to strong light. The orientation of the solar energy cell array is the same as that of the axis $-\boldsymbol{y}_{\mathrm{B}}$ in the spacecraft body frame $Ox_{\mathrm{B}}y_{\mathrm{B}}z_{\mathrm{B}}$, and the orientation of the camera is the same as that of the axis $\boldsymbol{z}_{\mathrm{B}}$. In Earth observation missions, the observation effect is the most important goal, while the spacecraft's energy acquisition requirements are minor. This paper considers the orientation accuracy of the axis $\boldsymbol{z}_{\mathrm{B}}$ to the Earth. The above attitude constraints can be divided into two types: mandatory pointing constraints and prohibited pointing constraints.

2.1. Reference attitude modelling

The definition of the reference attitude is the attitude considering the spacecraft's orientation accuracy to the Earth and Sun with different priorities.

When the spacecraft is in the reference attitude under ideal conditions, the axis $\boldsymbol{z}_{\mathrm{B}}$ of the spacecraft body frame $Ox_By_Bz_B$ points to the centre of the Earth, and the $-\boldsymbol{y}_{\mathrm{B}}$ axis points to the centre of the Sun. However, the position vector of the Sun $\boldsymbol{R}_{\mathrm{sun}}$ is typically not strictly vertical to the position vector of the Earth $\boldsymbol{R}_{\mathrm{earth}}$. The reference attitude cannot be directly defined as the axis $\boldsymbol{z}_{\mathrm{B}}$ coinciding with the line between the spacecraft and the centre of the Earth and the axis $-\boldsymbol{y}_{\mathrm{B}}$ coinciding with the Sun position vector $\boldsymbol{R}_{\mathrm{sun}}$.

Therefore, when the spacecraft is in the reference attitude, only the direction of the $\boldsymbol{x}_{\mathrm{B}}$ axis is unique:

$$\boldsymbol{x}_{\mathrm{B}} = \frac{-\boldsymbol{R}_{\mathrm{sun}} \times \boldsymbol{z}_O}{|\boldsymbol{R}_{\mathrm{sun}}|} \tag{1}$$

where $\boldsymbol{z}_O$ is the axis $\boldsymbol{z}$ of the orbital coordinate system. The origin of the orbital coordinate system $Ox_Oy_Oz_O$ is located at the mass centre of the spacecraft. The $\boldsymbol{x}_{\mathrm{O}}$ axis points in the speed direction of the satellite. The $\boldsymbol{z}_{\mathrm{O}}$ axis points to Earth's centre, is vertical and points in the direction of spacecraft flight. The $\boldsymbol{y}_{\mathrm{O}}$ axis is defined by the $\boldsymbol{x}_{\mathrm{O}}$ axis and the $\boldsymbol{z}_{\mathrm{O}}$ axis according to the right-hand rule.

To calculate the $\boldsymbol{y}_{\mathrm{B}}$ axis and the $\boldsymbol{z}_{\mathrm{B}}$ axis of the spacecraft body frame, two auxiliary coordinate systems $Ox_{\mathrm{earth}}y_{\mathrm{earth}}z_{\mathrm{earth}}$ and $Ox_{\mathrm{sun}}y_{\mathrm{sun}}z_{\mathrm{sun}}$ are defined, shown in Formula (2), where $\boldsymbol{x}_{\mathrm{sun}} = \boldsymbol{x}_{\mathrm{earth}} = \boldsymbol{x}_{\mathrm{B}}$. The origin of auxiliary coordinate systems is the mass centre of the spacecraft. Figure 1 is a graphical representation of the relationship between auxiliary coordinate systems, the Earth and the Sun:

$$\begin{cases} \boldsymbol{z}_{\mathrm{earth}} = \boldsymbol{z}_{\mathrm{O}}, \\ \boldsymbol{y}_{\mathrm{earth}} = \boldsymbol{z}_{\mathrm{earth}} \times \boldsymbol{x}_{\mathrm{B}}, \end{cases} \quad \begin{cases} \boldsymbol{y}_{\mathrm{sun}} = -\dfrac{\boldsymbol{R}_{\mathrm{sun}}}{|\boldsymbol{R}_{\mathrm{sun}}|} \\ \boldsymbol{z}_{\mathrm{sun}} = \boldsymbol{x}_{\mathrm{B}} \times \boldsymbol{y}_{\mathrm{sun}} \end{cases} \tag{2}$$

Therefore, the spacecraft body frame coincides with $Ox_{\mathrm{earth}}y_{\mathrm{earth}}z_{\mathrm{earth}}$ when priority is given to the accuracy of the orientation to the Earth. The spacecraft body frame coincides with $Ox_{\mathrm{sun}}y_{\mathrm{sun}}z_{\mathrm{sun}}$ when the priority is given to the accuracy of the orientation to the Sun.

$$\begin{cases} \boldsymbol{y}_{\mathrm{B}} = \dfrac{\kappa_{\mathrm{sun}}\boldsymbol{y}_{\mathrm{sun}} + \kappa_{\mathrm{earth}}\boldsymbol{y}_{\mathrm{earth}}}{|\kappa_{\mathrm{sun}}\boldsymbol{y}_{\mathrm{sun}} + \kappa_{\mathrm{earth}}\boldsymbol{y}_{\mathrm{earth}}|} \\ \boldsymbol{z}_{\mathrm{B}} = \dfrac{\kappa_{\mathrm{sun}}\boldsymbol{z}_{\mathrm{sun}} + \kappa_{\mathrm{earth}}\boldsymbol{z}_{\mathrm{earth}}}{|\kappa_{\mathrm{sun}}\boldsymbol{z}_{\mathrm{sun}} + \kappa_{\mathrm{earth}}\boldsymbol{z}_{\mathrm{earth}}|} \end{cases} \tag{3}$$

where $\kappa_{\mathrm{earth}} \in [0, 1]$ and $\kappa_{\mathrm{sun}} \in [0, 1]$ are weight coefficients.

https://doi.org/10.1017/S0373463321000813 Published online by Cambridge University Press

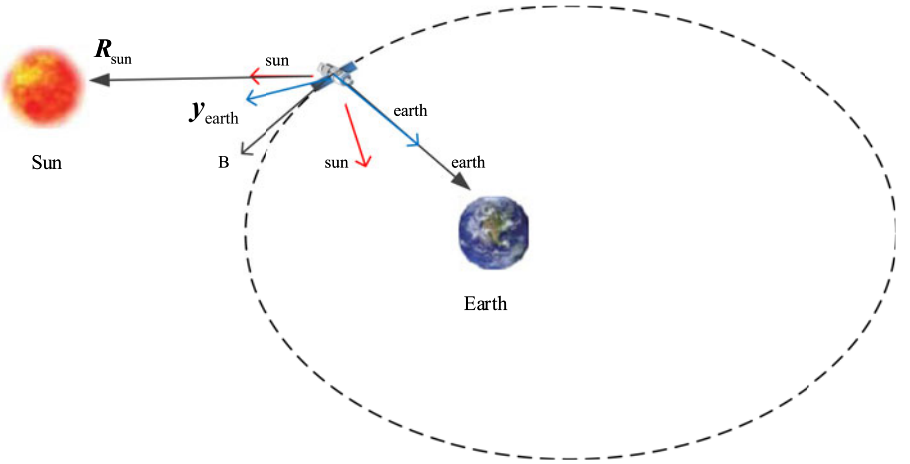


Figure 1. *Graphical representation of auxiliary coordinate systems.*

According to Equations (1), (2) and (3), the reference attitude $(\varphi_d, \theta_d, \psi_d)$ of the spacecraft can be described by:

$$\begin{cases} \varphi_d = \arctan \dfrac{\boldsymbol{y}_B \cdot \boldsymbol{z}_O}{\boldsymbol{z}_B \cdot \boldsymbol{z}_O} \\ \theta_d = \arcsin(-\boldsymbol{x}_B \cdot \boldsymbol{z}_O) \\ \psi_d = \arctan \dfrac{\boldsymbol{x}_B \cdot \boldsymbol{y}_O}{\boldsymbol{x}_B \cdot \boldsymbol{x}_O} \end{cases} \tag{4}$$

To describe the accuracy of the spacecraft's Earth orientation visually, the angle between axis z_B and axis z_O is introduced:

$$\cos\langle z_B, z_O\rangle = z_B \cdot z_O \tag{5}$$

2.2. Attitude constraints modelling

The mandatory pointing constraints faced by the spacecraft include the energy acquisition constraints on the Sun and orientation constraints to the Earth. The prohibited pointing constraints refer to the sensor attitude constraints, including direct and reflected sunlight constraints.

The energy acquisition constraint states that the included angle between the solar light and the solar energy cell array is less than θ_{q1}, and the included angle between the $-\boldsymbol{y}_B$-axis and the solar position vector $\boldsymbol{R}_{sun}$ in the spacecraft body frame is less than θ_{q1}:

$$\cos\langle -\boldsymbol{y}_B, \boldsymbol{R}_{sun}\rangle = \frac{-\boldsymbol{y}_B \cdot \boldsymbol{R}_{sun}}{|\boldsymbol{R}_{sun}|} > \cos(\theta_{q1}) \tag{6}$$

Orientation constraints on the Earth require that the included angle between axis z_B and Earth position vector $\boldsymbol{R}_{earth}$ in the spacecraft body frame be less than θ_{q2}:

$$\cos\langle z_B, \boldsymbol{R}_{earth}\rangle = \frac{z_B \cdot \boldsymbol{R}_{earth}}{|\boldsymbol{R}_{earth}|} > \cos(\theta_{q2}) \tag{7}$$

During attitude manoeuvring, it is necessary to avoid direct sunlight and reflected light areas within the conical fields of view of the sensors. Assuming that the position vector of the ith central axis of the sensor's field of view in the spacecraft body frame is $\boldsymbol{r}^B_{f\text{-}i}$, the position vector of the jth central axis of the reflected light area relative to the spacecraft is expressed as $\boldsymbol{r}^B_{m\text{-}j}$ in the spacecraft body frame.

https://doi.org/10.1017/S0373463321000813 Published online by Cambridge University Press

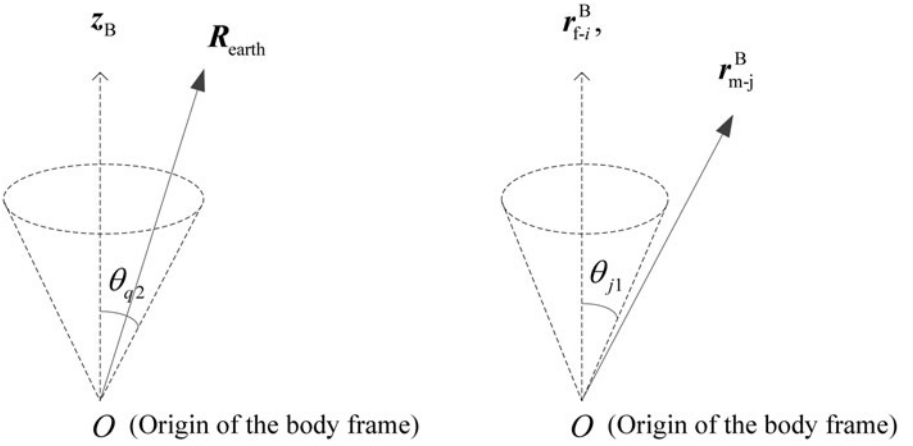


Figure 2. *Graphical representation of two types of constraint.*

Figure 2 is a graphical representation of mandatory and prohibited pointing constraints. The position vector of the Sun is expressed in the spacecraft body frame as $\boldsymbol{r}^{\mathrm{B}}_{\mathrm{sun}}$. The prohibited pointing constraints require the included angle between $\boldsymbol{r}^{\mathrm{B}}_{\mathrm{f}\text{-}i}$ and $\boldsymbol{r}^{\mathrm{B}}_{\mathrm{m}\text{-}j}$ to be greater than θ_{j1} and the included angle between $\boldsymbol{r}^{\mathrm{B}}_{\mathrm{f}\text{-}i}$ and $\boldsymbol{r}^{\mathrm{B}}_{\mathrm{sun}}$ to be greater than θ_{j2}:

$$\cos\left\langle \boldsymbol{r}^{\mathrm{B}}_{\mathrm{f}\text{-}i}, \boldsymbol{r}^{\mathrm{B}}_{\mathrm{m}\text{-}j} \right\rangle = \frac{\boldsymbol{r}^{\mathrm{B}}_{\mathrm{f}\text{-}i} \cdot \boldsymbol{r}^{\mathrm{B}}_{\mathrm{m}\text{-}j}}{|\boldsymbol{r}^{\mathrm{B}}_{\mathrm{f}\text{-}i}||\boldsymbol{r}^{\mathrm{B}}_{\mathrm{m}\text{-}j}|} < \cos(\theta_{j1}) \tag{8}$$

$$\cos\left\langle \boldsymbol{r}^{\mathrm{B}}_{\mathrm{f}\text{-}i}, \boldsymbol{R}_{\mathrm{sun}} \right\rangle = \frac{\boldsymbol{r}^{\mathrm{B}}_{\mathrm{f}\text{-}i} \cdot \boldsymbol{R}_{\mathrm{sun}}}{|\boldsymbol{r}^{\mathrm{B}}_{\mathrm{f}\text{-}i}||\boldsymbol{R}_{\mathrm{sun}}|} < \cos(\theta_{j2}) \tag{9}$$

3. Attitude manoeuvre planning for spacecraft based on reinforcement learning

Reinforcement learning can be divided into three categories: policy-based only, value-based only, policy-based and value-based. For a high-dimensional problem, it is difficult to isolate each state strictly and to calculate the value function of each state. The value-based only method has no explicit representation policy; thus, the change of the value function cannot reflect the change of the policy well. It is difficult to address the problem of random policy (Liu et al., 2019). In contrast, the policy gradient algorithm always improves the policy by improving its direction and can effectively manage problems in the high and continuous action space (Tang et al., 2017). For single-agent reinforcement learning, the optimal baseline and the low variance policy have been proven fruitful (Greensmith et al., 2004; Zhao et al., 2015). Therefore, the policy gradient-based reinforcement learning method is used in this paper to perform attitude manoeuvre planning.

3.1. Attitude manoeuvre planning scheme design

Based on the reference attitude and attitude constraints model, this section describes the attitude manoeuvre planning scheme based on reinforcement learning. First, the reinforcement learning model of attitude planning is established, including the basic concepts of policy, reward function and policy value function. Second, the initial database for reinforcement learning is generated. In this paper, the reference attitude is used to establish the database, which can reduce the calculation of attitude manoeuvre planning and accelerate the convergence.

The total amount of data in the database is set to be N. Then, attitude manoeuvre planning is performed based on the improved policy gradient reinforcement learning algorithm. At the end of planning, $N + k_{\mathrm{tol}}$ attitude manoeuvre trajectories will be obtained, where k_{tol} is the total number of iterations.

https://doi.org/10.1017/S0373463321000813 Published online by Cambridge University Press

The policy gradient reinforcement learning algorithm will update the Euler angle along the direction satisfying all constraints. However, it cannot guarantee that all $N + k_{\text{tol}}$ attitude manoeuvre trajectories meet the attitude constraint requirements. The accuracy is different for different trajectories. Therefore, the acquired attitude manoeuvre trajectories must be screened.

Before screening the results, it is necessary to define the screening index $\chi(k)$ for attitude planning results. The following is an optional screening index $\chi(k)$ for attitude planning results:

$$\begin{aligned}\chi(k) = \upsilon_1 \min\left[\left\langle z_{\text{B},k}(m), z_{\text{O},k}(m)\right\rangle\right] + \upsilon_2 \min\left[\left\langle -y_{B,k}(m), R_{\text{sun},k}(m)\right\rangle\right] \\ + \upsilon_3[\sigma(\varphi_k(m)) + \sigma(\theta_k(m)) + \sigma(\psi_k(m))]\end{aligned} \tag{10}$$

where k represents the current iteration serial number; υ_1,υ_2,υ_3 represent the weight coefficient; and $\sigma(*)$ represents the mean square deviation of $*$. The roll angle $\varphi_k(m)$, pitch angle $\theta_k(m)$ and yaw angle $\psi_k(m)$ are the three Euler angles of the spacecraft at moment m of the kth iteration.

The screening index is not unique and can be designed according to real requirements. For the screening index proposed in this paper, the smaller the value is, the better the attitude planning result will be.

The final selected attitude planning result is the attitude manoeuvre planning result of group k_{opt}:

$$k_{\text{opt}} = \arg\underset{k^*}{\{\chi(k^*) = \min[\chi(k)],\quad k = 1, 2, \ldots, N + k_{\text{tol}}\}} \tag{11}$$

Flutter and other problems may exist in the final selected result, which will improve the difficulty of attitude control. Therefore, it is necessary to smooth the attitude planning results. The smoothing method is not unique. However, the attitude planning results after smoothing should still meet the requirements of attitude constraints. In this paper, the moving average filter is used to smooth the results.

The flow of the attitude manoeuvre planning scheme based on reinforcement learning is shown in Figure 3.

3.2. *Attitude manoeuvre planning algorithm based on improved policy gradient reinforcement learning*

3.2.1. Attitude manoeuvre planning model for reinforcement learning

The purpose of the policy gradient reinforcement learning method based on the finite difference policy gradient is to optimise policy $x_k(m)$ at the current moment m of iteration kth based on policy value function $J[x_k(m)]$:

$$J[x_{q,k}(m)] = \sum_{m=1}^{m_{\max}} \gamma(m) R_{q,k}(m) \tag{12}$$

where $\gamma(m)$ is the discount factor at this moment, and $r_k(m)$ is the reward function at time m.

The policy $x_{k+1}(m)$ is updated as:

$$x_{k+1}(m) = x_k(m) + \eta \nabla J[x_k(m)] = x_k(m) + \eta G \tag{13}$$

where η is the step size of reinforcement learning and G is the policy gradient.

The policy $x_{q,k}(m)(q = 1, 2, 3)$ of the attitude manoeuvre is related to the Euler angles of the spacecraft:

$$\begin{cases} x_{1,k}(m) = \varphi_k(m) \\ x_{2,k}(m) = \theta_k(m) \\ x_{3,k}(m) = \psi_k(m) \end{cases} \tag{14}$$

where $x_{1,k}(m)$, $x_{2,k}(m)$ and $x_{3,k}(m)$ represent the roll angle $\varphi_k(m)$, pitch angle $\theta_k(m)$, and yaw angle $\psi_k(m)$, respectively, of the spacecraft at time m of iteration kth.

https://doi.org/10.1017/S0373463321000813 Published online by Cambridge University Press

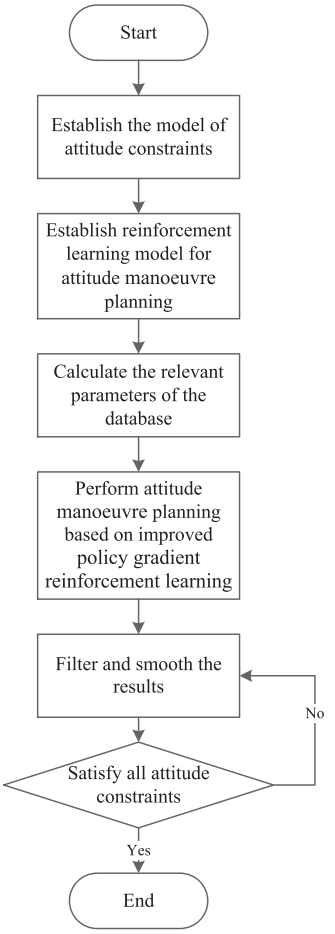


Figure 3. *Flow chart of attitude manoeuvre planning.*

The reward function shown in Equation (15) is related to the accuracy of the orientation to the Earth and the Sun and the prohibited pointing constraints:

$$\begin{cases} R_{\mathrm{d},k}(m) = -\mu_{\mathrm{earth}} \dfrac{1}{\cos^2\left\langle \boldsymbol{z}_{\mathrm{B},k}(m), \boldsymbol{z}_{\mathrm{O},k}(m)\right\rangle} - \mu_{\mathrm{sun}} \dfrac{1}{\cos^2\left\langle -\boldsymbol{y}_{\mathrm{B},k}(m), \boldsymbol{R}_{\mathrm{sun},k}(m)\right\rangle} \\ R_{\mathrm{m}\text{-}i,k}(m) = -\sum\limits_{i=1}^{N_o} \mu_{\mathrm{f}\text{-}j-i} \cos^2\left\langle \boldsymbol{r}^{\mathrm{B}}_{\mathrm{f}\text{-}j,k}(m), \boldsymbol{r}^{\mathrm{B}}_{\mathrm{m}\text{-}j-i,k}(m)\right\rangle \\ R_{\mathrm{s},k}(m) = -\sum\limits_{j=1}^{N_o} \mu_{\mathrm{sun}\text{-}i} \cos^2\left\langle \boldsymbol{r}^{\mathrm{B}}_{\mathrm{sun},k}(m), \boldsymbol{r}^{\mathrm{B}}_{\mathrm{m}\text{-}j\text{-}i,k}(m)\right\rangle \end{cases} \tag{15}$$

where m is the moment, k is the current serial number of iterations, $R_{\mathrm{d},k}(m)$ is the reward function related to the orientation to the Earth and the Sun, and $R_{\mathrm{d},k}(m)$ in Equation (15) is the reward function of mandatory pointing constraints. There are two mandatory pointing constraints in this paper; thus, $R_{\mathrm{d},k}(m)$ consists of two fractions. If we must add new mandatory pointing constraints to the spacecraft, just add new fractions to $R_{\mathrm{d},k}(m)$.

$R_{\mathrm{m}\text{-}i\text{-}j,k}(m)$ and $R_{\mathrm{s},k}(m)$ represent reward functions of two different types of prohibited pointing constraints. N_o represents the total number of sensors carried by spacecraft and N_J represents the total

https://doi.org/10.1017/S0373463321000813 Published online by Cambridge University Press

number of reflected light areas. Therefore, N_o and N_J represent the number of two different types of prohibited pointing constraints. If we must add new prohibited pointing constraints to the spacecraft, just add new items to $R_{\text{m-}i\text{-}j,k}(m)$ or $R_{\text{s},k}(m)$.

$R_{\text{m-}i\text{-}j,k}(m)$ is the reward function related to the angle between the ith central axis of the sensor's field of view and the jth central axis of the reflected light area. $\mu_{\text{f-}j\text{-}i}$ is the weight coefficient of the reward function related to the included angle between the ith central axis of the sensor's field of view and jth the central axis of the reflected light area. $R_{\text{s},k}(m)$ is the reward function related to the included angle between all central axes of the sensor's fields of view and the position vector of the Sun. $\mu_{\text{f-}j\text{-}i}$ is the weight coefficient of the reward function related to the included angle between the ith central axis of the reflected light area and the position vector of the Sun. When the relevant included angle meets the attitude constraint requirements, the weight coefficient can be set to 0.

Because the attitude manoeuvre can be accomplished by circling the two coordinate axes of the spacecraft body frame, the total reward function of the policy is set as follows:

$$\begin{cases} R[x_{1,n}(m)] = R[\varphi_k(m)] = R_{\text{d},k}(m) \\ R[x_{2,k}(m)] = R[\theta_k(m)] = R_{\text{d},k}(m) + \sum_{j=1}^{N_J} R_{\text{m-}i\text{-}j,k}(m) + R_{\text{s},k}(m) \\ R[x_{3,k}(m)] = R[\psi_k(m)] = R_{\text{d},k}(m) + \sum_{j=1}^{N_J} R_{\text{m-}i\text{-}j,k}(m) + R_{\text{s},k}(m) \end{cases} \tag{16}$$

The value function $J[x_{q,k}(m)]$ of $x_{q,k}(m)(q = 1, 2, 3)$ policy is:

$$R_{q,k}(m) = R[x_{q,k}(m)] \tag{17}$$

$$J[x_{q,k}(m)] = \sum_{m=1}^{m_{\max}} \gamma(m) R_{q,k}(m) \tag{18}$$

3.2.2. Attitude manoeuvre planning algorithm based on improved policy gradient reinforcement learning

This section describes the process of the attitude manoeuvre planning algorithm based on improved policy gradient reinforcement learning. The process begins with the data in the database and then the outer iteration starts. First, initialise the serial number of iterations $k = 1$ and time $m = 1$.

To update the policy gradient, it is necessary to set the inner iteration. Let the current one be the k_G. Initialise $k_G = 0$. After the first generation, the initial policy at time m of the kth iteration adopts the policy $x_{q,k}^{(0)}(m)$ at time m of the $k-1$thiteration:

$$x_{q,k}^{(0)}(m) = x_{q,k-1}(m) \tag{19}$$

Calculated reference expected value function $\hat{J}_{q,\text{ref}}^{(k_G)}$ of policy $x_{q,k}^{(k_G)}(m)$:

$$\hat{J}_{q,\text{ref}}^{(k_G)} = E[\hat{J}[x_{q,k}^{(k_G)}(m)]] = E\left[\sum_{m=1}^{m_{\max}} \gamma(m) R_{q,k}(m)\right] \tag{20}$$

where $E[\cdot]$is the expect of $[\cdot]$.$R_{q,n}(m)$ is the reward of the policy of the qth Euler angle at the moment.

ε is a small value greater than zero. The policy $x_{q,k}^{(0)}(m)$ at time m of the kth iteration satisfies $|x_{q,k}^{(0)}(m) - x_{q,k}(m)| \le \varepsilon$.

https://doi.org/10.1017/S0373463321000813 Published online by Cambridge University Press

In the inner iteration, the difference of the policy $\Delta x_{q,k}^{(k_G)}(m)$ must be calculated first, and then the expected value function $\hat{J}_q^{(k_G)}$ $(q = 1, 2, 3)$ of $x_{q,k}^{(k_G)}(m) + \Delta x_{q,k}^{(k_G)}(m)$:

$$\Delta x_{q,k}^{(k_G)}(m) = \begin{cases} x_{q,k-2}(m) - x_{q,k-1}(m), & k_G = 0 \\ \eta G_{q,k}(m) & , \ k_G > 0 \end{cases} \tag{21}$$

$$\hat{J}_q^{(k_G)} = E[\hat{J}[x_{q,k}^{(k_G)}(m) + \Delta x_{q,k}^{(k_G)}(m)]] \tag{22}$$

$$\Delta \hat{J}_{q,k}^{(k_G)}(m) = \hat{J}_q^{(k_G)} - \hat{J}_{q,ref}^{(k_G)} - R_{q,k-1}(m)/k - 1 \tag{23}$$

where $\Delta \hat{J}_{q,k}^{(k_G)}(m)$ is the expected value function difference with the return baseline, and $R_{q,k-1}(m)$ is the return at time m of the $k-1$th iteration, where $R_{q,k-1}(m)/k - 1$ is the return baseline. Generally, reinforcement learning algorithms based on policy gradients lack evaluation of policies. In some problems, the returns of the policy are always positive. Therefore, even a poor policy will be strengthened, resulting in a larger variance (Wang et al., 2015). To solve this problem, the return baseline is introduced to distinguish the quality of the strategy and reduce the variance of the policy gradient. To improve the calculation efficiency, this paper uses $R_{q,k-1}(m)/k - 1$ to approximate the mean value the return baseline at time m of the $1 \sim k - 1$th iteration:

$$\begin{cases} \Delta \boldsymbol{x}_{q,k}(m) = \left[\Delta x_{q,k}^{(0)}(m) \ \Delta x_{q,k}^{(1)}(m) \ \cdots \ \Delta x_{q,k}^{(k_G)}(m) \right] \\ \Delta \hat{\boldsymbol{J}}_{q,k}(m) = \left[\Delta \hat{J}_{q,k}^{(1)}(m) \ \Delta \hat{J}_{q,k}^{(2)}(m) \ \cdots \ \Delta \hat{J}_{q,k}^{(k_G)}(m) \right] \end{cases} \tag{24}$$

The policy gradient is updated to:

$$G_{\text{temp},q,k}(m) = \begin{cases} 0, & \Delta \boldsymbol{x}_{q,k}(m) = 0 \\ \text{sgn}\{-[\Delta \boldsymbol{x}_{q,k}^{\text{T}}(m) \Delta \boldsymbol{x}_{q,k}(m)]^{-1} \Delta \boldsymbol{x}_{q,k}^{\text{T}}(m) \Delta \hat{\boldsymbol{J}}_{q,k}(m)\}, & \text{else} \end{cases} \tag{25}$$

We now introduce the error between policy $x_{q,n}(m)$ and reference attitude $x_{q,n}^{\text{e}}(m)$, $(q = 1, 2, 3)$:

$$\begin{cases} x_{1,k}^{\text{e}}(m) = \varphi_k^{\text{e}}(m) = \varphi_k(m) - \varphi_d(m) \\ x_{2,k}^{\text{e}}(m) = \theta_k^{\text{e}}(m) = \theta_k(m) - \theta_d(m) \\ x_{3,k}^{\text{e}}(m) = \psi_k^{\text{e}}(m) = \psi_k(m) - \psi_d(m) \end{cases} \tag{26}$$

If $G_{q,k}(m)$ converges, reset$k_G = 0$, and $\mu_{\text{e}} \geq 0$ is the weight coefficient. Then, calculate the final policy gradient $G_{q,k}(m)$:

$$G_{q,k}(m) = G_{\text{temp},q,k}(m) + \mu_{\text{e}} x_{q,k}^{\text{e}}(m) \tag{27}$$

If $G_{q,k}(m)$ does not converge, add $k_G + 1$ then repeat Equations (21)–(26) until the policy gradient converges.

In Equation (27), μ_{e} is a positive error weight coefficient less than 1, and $x_{q,n}^{\text{e}}(m)$ can guide the attitude of the spacecraft towards the reference attitude, which can mitigate any non-directionality in attitude updates. After obtaining the converged policy gradient at the current time, the time is updated until all times of the current outer iteration are completed.

Policy exploration is an important step in policy updating. If only the current obtained optimal policy gradient is used to guide updating, it can easily fall into a local optimum when the policy gradient is insufficient. A good policy exploration can consider other actions while learning the optimal actions and may find a better policy than the current one. Based on Equation (27), this paper proposes an adaptive exploration method for policy updating, as shown in Equation (28). c_1 is the fixed policy weight coefficient, $c_{q1} * \sin(k/k_{\text{tol}}) * (\pi/2)$ is the dynamic policy weight coefficient, c_2 is the fixed exploration coefficient and $c_{q2} * \cos(k/k_{\text{tol}}) * (\pi/2)$ is the dynamic exploration coefficient.

https://doi.org/10.1017/S0373463321000813 Published online by Cambridge University Press

Table 1. *Parameters of Sun-synchronous orbit.*

Semimajor axis	Eccentricity	Inclination	Ascending node	Argument of perigee	True anomaly
6,978 km	0	97°	125°	0°	0°

Note: Time 1 January 2021 04:00:00·000 Universal Time Coordinated.

In the early stage of the iteration, to keep the change in attitude stable, c_1 cannot be too large. However, c_2 can be marginally larger to enable the policy to gain stronger exploration ability and avoid falling into the local optimum prematurely. $c_{q2} * \cos(k/k_{\text{tol}}) * (\pi/2)$ should increase gradually, and c_2 decreases gradually as the iteration progresses. The intention of the change is to balance the ability to follow the policy gradient and free exploration, which can speed up the process of convergence:

$$\begin{aligned} x_{q,k}(m) = x_{q,k-1}(m) - (c_1 + c_{q1} * \sin\left(\frac{k}{k_{\text{tol}}} * \frac{\pi}{2}\right) * G_{q,k}(m) \\ + (c_2 + c_{q2} * \cos\left(\frac{k}{k_{\text{tol}}} * \frac{\pi}{2}\right) * (x_{q,k}(m) - x_{q,k-1}(m)) \end{aligned} \tag{28}$$

After updating the policy according to Equation (28), the outer iteration continues until the maximum number of iterations is reached. Finally, multiple groups of attitude manoeuvre planning trajectories were obtained to be screened.The flow chart of the algorithm proposed in this paper is shown in Figure 4.

4. Simulation and analysis

This paper analyses spacecraft operating in a Sun-synchronous orbit. The parameters of the orbit are shown in Table 1.

To gain a more intuitive understanding of the problem of attitude manoeuvre planning, this paper analyses a typical scene of spacecraft in orbit under complex and multi-constraint conditions first.

In this paper, sunlight is reflected by five reflection points on the surface of the Earth; thus, five reflected light areas will be formed. In the spacecraft body frame, the position vectors of the central axes of the sensor's field of view are:

$$\begin{cases} \boldsymbol{r}_{\text{f-1}}^{B} = C_z(-50\ \text{deg})C_y(50\ \text{deg})\left[\,1\ 0\ 0\,\right]^T \\ \boldsymbol{r}_{\text{f-2}}^{B} = C_z(-50\ \text{deg})C_y(-50\,\text{deg})\left[\,1\ 0\ 0\,\right]^T \\ \boldsymbol{r}_{\text{f-3}}^{B} = C_z(50\ \text{deg})C_y(50\ \text{deg})\left[\,1\ 0\ 0\,\right]^T \\ \boldsymbol{r}_{\text{f-4}}^{B} = C_z(50\ \text{deg})C_y(-50\ \text{deg})\left[\,1\ 0\ 0\,\right]^T \\ \boldsymbol{r}_{\text{f-5}}^{B} = \left[\,0\ \sin(60\ \text{deg})\ \cos(60\ \text{deg})\,\right]^T \end{cases} \tag{29}$$

where $\boldsymbol{C}_z(\alpha)$ and $\boldsymbol{C}_y(\alpha)$ are the direction cosine matrices around the axis $\boldsymbol{z}_B$ and the axis $\boldsymbol{y}_B$, respectively.

After preliminary analysis, the relative position relationship between the spacecraft and the reflected light areas changes periodically. We can thus draw the following conclusion: sometimes all fields of view for the sensors do not face any reflected light areas; sometimes sensors simultaneously face several reflected light areas in a short period of time, which means prohibited pointing constraints are violated. The selected most 'dangerous' situation is called the intensive pointing situation shown in Figure 5: in a short period of time, the fields of view of different sensors are successively directed to the reflected light areas. The next part of this paper focuses on the situation shown in Figure 5 (26 May 2021 23:00:26~23:03:51) to develop the attitude manoeuvre planning strategy.

The simulation parameters are shown in Table 2. The position vectors of the central axes of reflected light areas and the Sun are both unit vectors in the Earth-centred inertial system.

https://doi.org/10.1017/S0373463321000813 Published online by Cambridge University Press

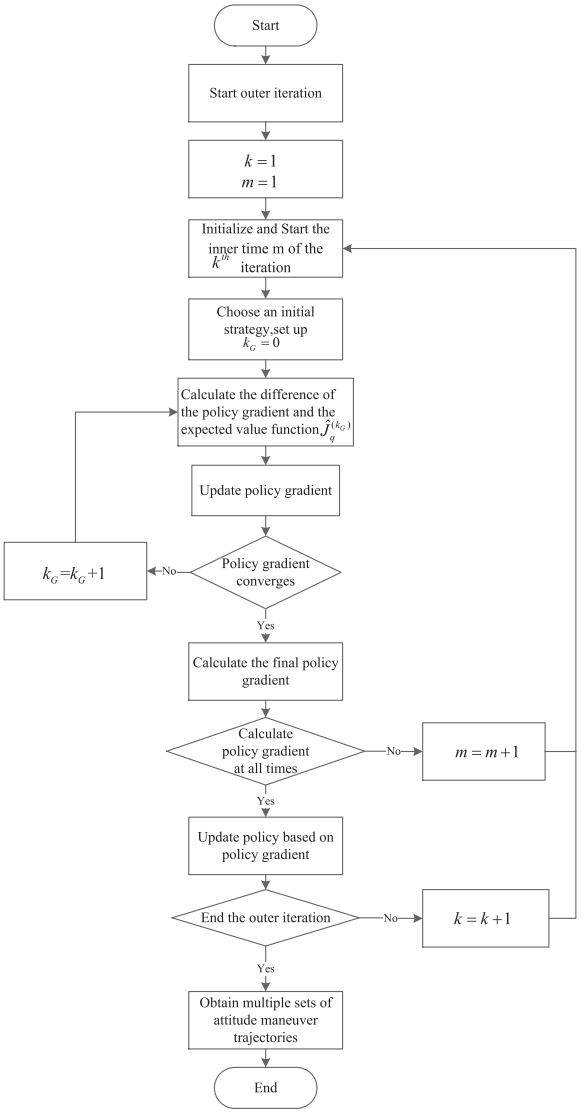


Figure 4. *Flow chart of attitude manoeuvre planning based on improved policy gradient reinforcement learning.*

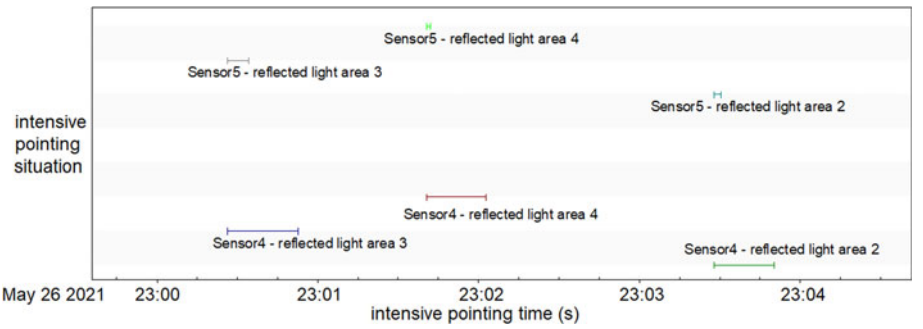


Figure 5. *Intensive pointing situation.*

https://doi.org/10.1017/S0373463321000813 Published online by Cambridge University Press

Table 2. *Simulation parameters.*

Simulation duration (s)	220
Semimajor axis a (km)	6,978
Eccentricitye	0
Inclination i(°)	97
Initial ascending node ω_0(°)	125
Initial argument of perigee Ω_0(°)	0
Initial true anomaly M_0(°)	0
Initial value of position vector of reflection point 1	(0·7313, −0·6712,0·1217)
Initial value of position vector of reflection point 2	(0·8132, −0·5755,0·0864)
Initial value of position vector of reflection point 3	(0·7954, −0·5801,0·1756)
Initial value of position vector of reflection point 4	(0·7900, −0·5851,0·1393)
Initial value of position vector of reflection point 5	(0·7463, −0·6455,0·1623)
Initial value of the position vector of the Sun	(0·4129, 0·8357,0·3623)

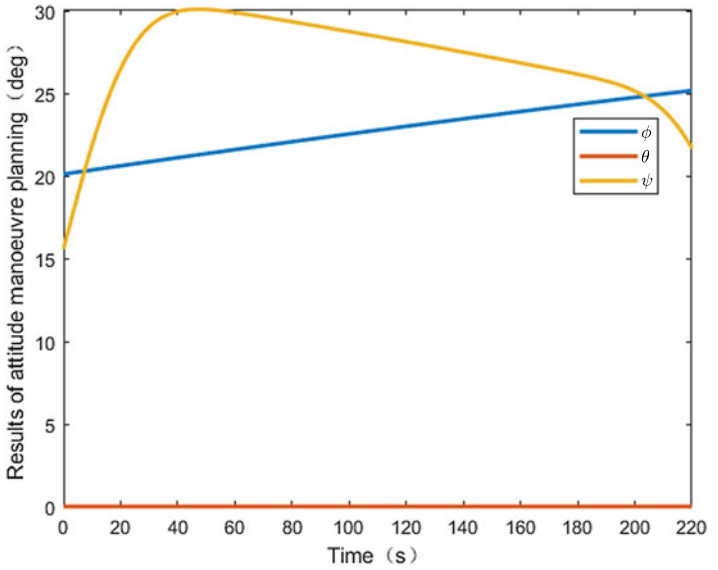


Figure 6. *Results of attitude manoeuvre planning.*

This paper requires that the included angle between the $-\boldsymbol{y}_B$ axis and the position vector of the Sun be less than 20°. In the spacecraft body frame, the included angles between the central axes of the sensor's fields of view and the central axes of the reflected light areas must be greater than 30°. The included angles between all the central axes of the sensor's fields of view and the position vector of the Sun must be greater than 40°. After using the proposed algorithm to plan the spacecraft's attitude manoeuvre, the results of the attitude manoeuvre planning are finally obtained using Equation (10) as the screening index, as shown in Figure 6. The results of attitude manoeuvre planning in this paper leave a time margin of 10 seconds, which means that the attitude manoeuvre planning in all simulation results actually begins from the 10th second. Because the attitude adjustment of spacecraft can be accomplished by circling the two coordinate axes of the spacecraft body frames, Figure 6 shows that the change in pitch angle θ is nearly zero.

During the entire attitude manoeuvre planning time, the roll angle ϕ varies from 20·37° (in 10th second) to 25·19° (in 220th second). The yaw angle ψ changes greatly. The minimum yaw angle is 21·86°

https://doi.org/10.1017/S0373463321000813 Published online by Cambridge University Press

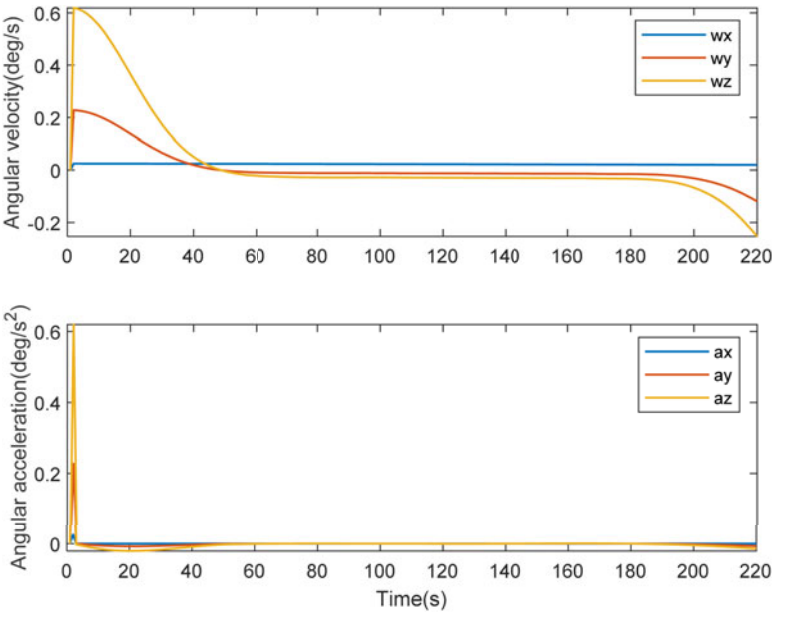


Figure 7. *Angular velocity and angular acceleration.*

(in 10th second), and the maximum yaw angle is 30·12° (in 46th second). Throughout the planning process, the change range is only 8·26°.

Figure 7 shows the components of angular velocity $\boldsymbol{w}$ and angular acceleration $\boldsymbol{a}$ of the attitude manoeuvre along the coordinate axes of the spacecraft body frame. The maximum value of the components of angular velocity is $\boldsymbol{w}_z$ in the 10th second, which is $0{\cdot}55°/s$. The maximum absolute value of angular acceleration is $\boldsymbol{a}_z$ in the 20th second, which is $-0{\cdot}021°/s^2$; this result indicates that in the planning time, the changing ranges of angular velocity and angular acceleration are very low. The changes in angular velocity and angular acceleration are small and slow, which is convenient for micro spacecraft.

Although the changes in Euler angles are small, spacecraft still maintain a high orientation accuracy to the Earth and to the Sun, as shown in Figure 8. In most cases, the included angle between the $-\boldsymbol{y}_{\mathrm{B}}$ axis and Sun is less than 10°, which ensures the ability to gather solar energy. The included angles between the central axes of all reflected light areas and all central axes of the sensor's fields of view are shown in Figures 9–13. All fields of view for the sensors are far away from reflected light areas. The smallest included angle is 41·69° (the included angle between the centre axis of the sensor's field of view 2 and the reflected light area 2), which is still greater than 30°. In addition, as shown in Figure 14, the angle between the central axes of all sensor fields of view and the Sun position vector is larger than 40°, which effectively avoids sunlight.

Using the improved policy gradient reinforcement learning proposed in this paper, the spacecraft can effectively avoid dynamic direct and reflected sunlight while maintaining a high accuracy of pointing to the Earth and to the Sun. All sensors also avoid exposure to strong light with a large safety margin. The above results are not at the cost of a large attitude manoeuvre, and the process of changing the Euler angles in the planning results is smooth and slow, which conforms to the manoeuvrability of the micro spacecraft and can be executed smoothly by a control mechanism.

To demonstrate the originality and advantage of improved policy gradient reinforcement learning, the proposed method is compared with other relevant methods, which are shown in Table 3. Paper 5 in Table 3 shows the improved policy gradient reinforcement learning method proposed in this paper. The four methods in Table 3 represent different ideas to solve the problem of attitude manoeuvre planning.

Table 3 shows that the first four methods only consider static constraints. The focus of the first three methods is the time-optimal reorientation manoeuvre problem of the spacecraft. In the above

https://doi.org/10.1017/S0373463321000813 Published online by Cambridge University Press

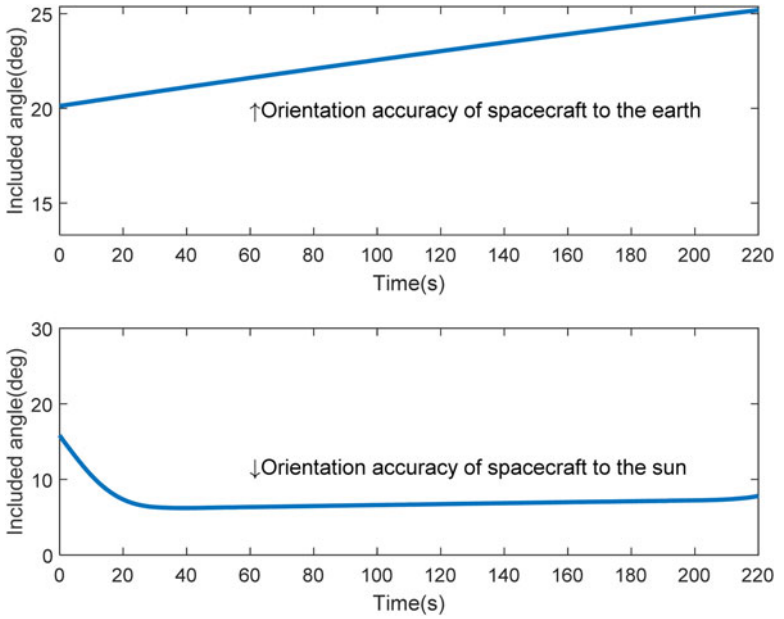


Figure 8. *Orientation accuracy of spacecraft to the Earth and to the Sun.*

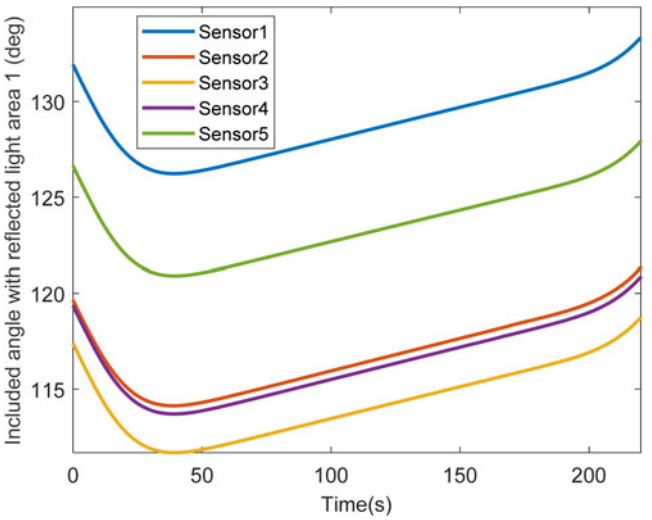


Figure 9. *Included angle between the central axes of the sensor field of view and the central axis of reflected light area 1.*

four methods, few static constraints exist on the virtual sphere, which makes the problem model too idealistic and far from the real-world situation. The start and end points of attitude are often known quantities when planning. These methods thus equate the known end point with mandatory pointing constraints, which cannot solve the problem of multiple mandatory pointing constraints concurrently. Compared with that method, the proposed method can successfully manage two dynamic mandatory pointing constraints concurrently in Sun-synchronous orbit, which is a difficult problem to solve.

The policy gradient reinforcement learning proposed in this paper is compared with the convex potential function (CPF) from Ma et al. (2020) to demonstrate its effectiveness. The relevant parameters are still set as shown in Table 2. To meet the need for orientation to the Sun, the desired attitude is

https://doi.org/10.1017/S0373463321000813 Published online by Cambridge University Press

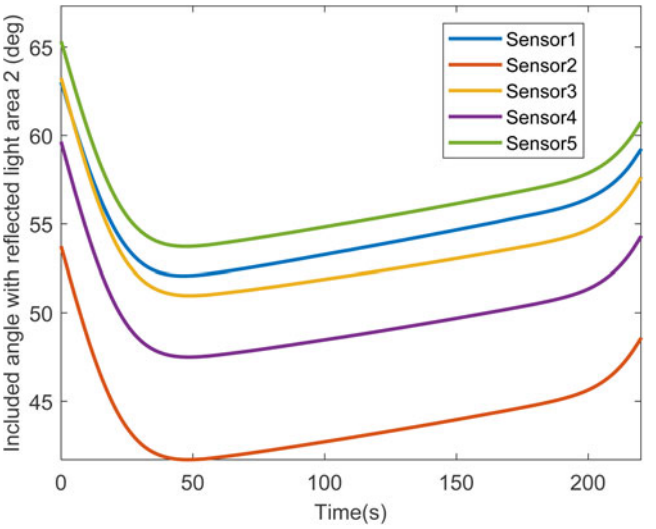


Figure 10. *Included angle between the central axes of the sensor field of view and the central axis of the reflected light area 2.*

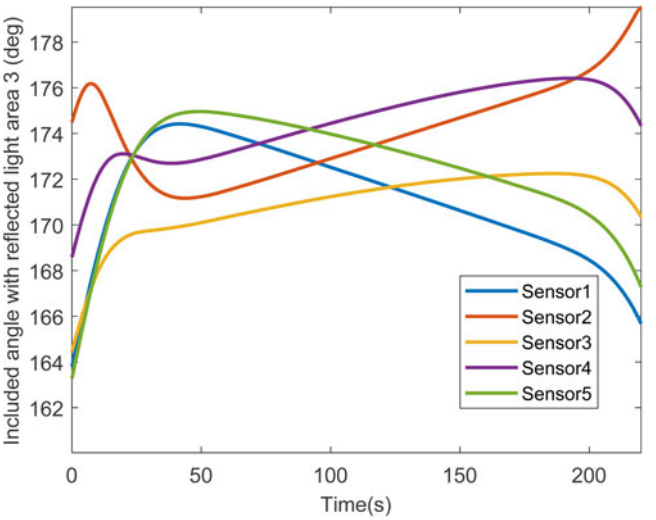


Figure 11. *Included angle between the central axes of the sensor field of view and the central axis of the reflected light area 3.*

defined: the included angle between the $-\boldsymbol{y}_{\mathrm{B}}$ axis and the Sun is 0° if the spacecraft is defined in the desired attitude. The desired attitude expressed by Euler angles is shown in Figure 15. The results of attitude manoeuvre planning using the method of Ma et al. (2020) are shown in Figure 16. The orientation accuracies to the Sun and to the Earth are shown in Figure 17 and Figure 18, respectively.

The planning results clearly do not track with the desired attitude. The included angle between the $-\boldsymbol{y}_{\mathrm{B}}$ axis and Sun is much larger than 20°, which cannot meet the needs of power generation. The included angles between all central axes of reflected light areas and all central axes of the sensor's fields of view are shown in Figures 19–23. Note that the smallest included angle is smaller than 30°.

Figures 15–23 show that the CPF from Ma et al. (2020) does not seem to be suitable for solving the problems raised in this paper. The underlying reason for this result lies in Formula (16) of Ma et al.

https://doi.org/10.1017/S0373463321000813 Published online by Cambridge University Press

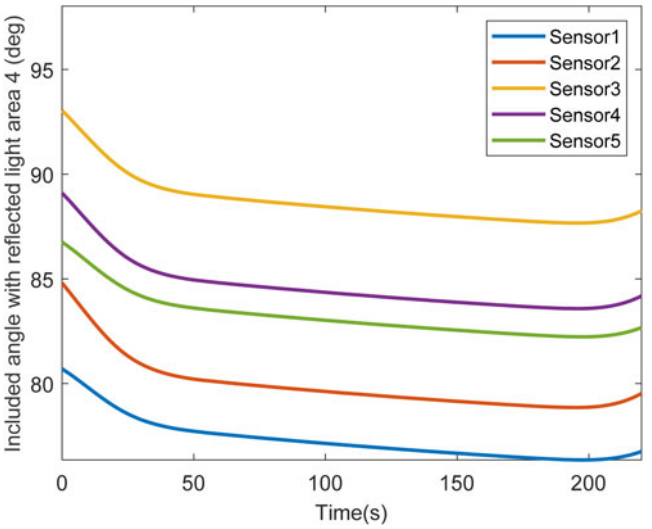


Figure 12. *Included angle between the central axes of the sensor field of view and the central axis of the reflected light area 4.*

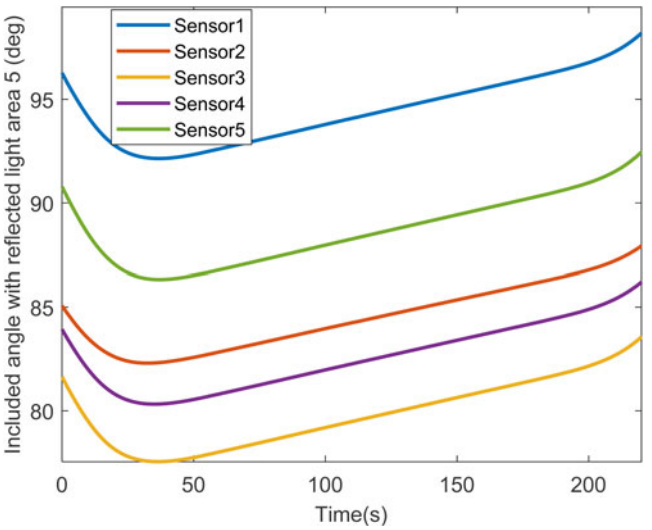


Figure 13. *Included angle between the central axes of the sensor field of view and the central axis of the reflected light area 5.*

(2020), which is shown here as Equation (30):

$$V(\widetilde{\boldsymbol{e}}) = \sum_{i=1}^{n} \frac{k_i ||\boldsymbol{e}||^2}{\mu_i - \widetilde{\boldsymbol{e}}^T \overline{\boldsymbol{A_i}} \widetilde{\boldsymbol{e}}} \tag{30}$$

where $\widetilde{\boldsymbol{e}}$ is the error quaternion when the desired attitude is given, and $k_i > 0$ are weight coefficients of different prohibited pointing areas in the potential function. Every prohibited pointing area has a threat zone, which is a set of attitudes, and weight coefficients are greater than zero ($k_i > 0$) only if the attitude of the spacecraft is within the threat zone. However, once the attitude is adjusted outside the threat zone, the coefficient is equal to zero ($k_i = 0$). When the attitude is outside all threat zones, all weight coefficients k_i are equal to zero. When all k_i are equal to zero, Equation (30) will equal zero because k_i

https://doi.org/10.1017/S0373463321000813 Published online by Cambridge University Press

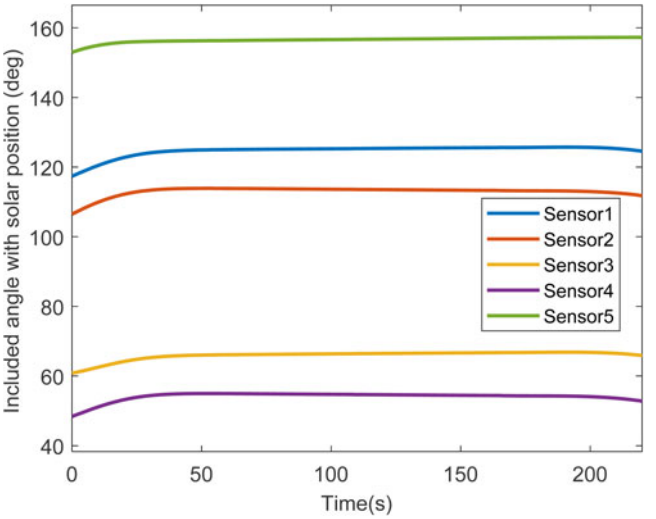


Figure 14. *Included angles between central axes of sensor attitude forbidden zones and the position vector of the Sun.*

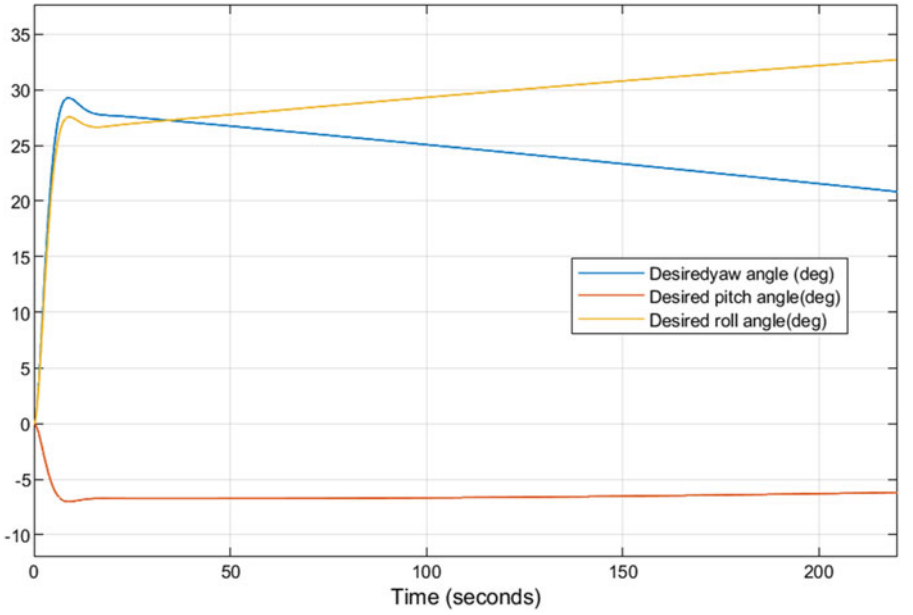


Figure 15. *Desired attitude (CPF).*

is part of the numerator. If the above situation occurs during the attitude manoeuvre planning process, the potential function will be invalid and the spacecraft will perform redundant invalid manoeuvres. In severe cases, the attitude control will be lost.

The first three methods in Table 3 require long computation times compared with the proposed improved policy gradient reinforcement learning method. In contrast, in the fourth method, the potential function combined with the backstepping method is efficient, and its focus is on using the control law to plan the attitude manoeuvre. The first step is to design an appropriate potential function and control law based on constraints. At this step, there is an inevitable problem in which the control parameters are difficult to design. On the premise of obtaining a suitable control law, the second step is to calculate

https://doi.org/10.1017/S0373463321000813 Published online by Cambridge University Press

Table 3. *Comparison with other methods.*

Publication	Method	Attitude constraints	Constraint type/ quantity	Manoeuvre time	Computing time
Olivares and Staffetti (2018)	Mixed-integer nonlinear programming	Mandatory pointing	Static/1	744·03 s	23·27 s
Celani and Lucarelli (2019)	Gradient ascent and inverse attitude dynamics	Mandatory pointing	Static/2	473 s	47 min
Spiller et al. (2016)	Particle swarm optimisation	Mandatory pointing	Static/2	107·60 s	232·23 s
Ma et al. (2020)	Convex potential function (CPF) and backstepping method	Mandatory pointing	Static/2	60 s	Not mentioned
This paper	Improved policy gradient reinforcement learning	Mandatory pointing and prohibited pointing	Dynamic/8	220 s	8·38 s

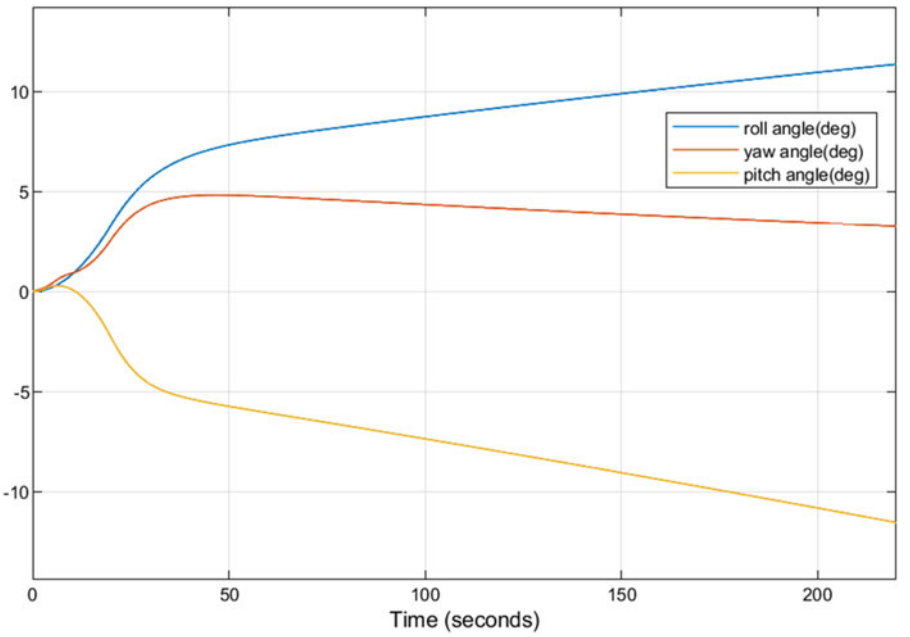


Figure 16. *Results of attitude manoeuvre planning (CPF).*

https://doi.org/10.1017/S0373463321000813 Published online by Cambridge University Press

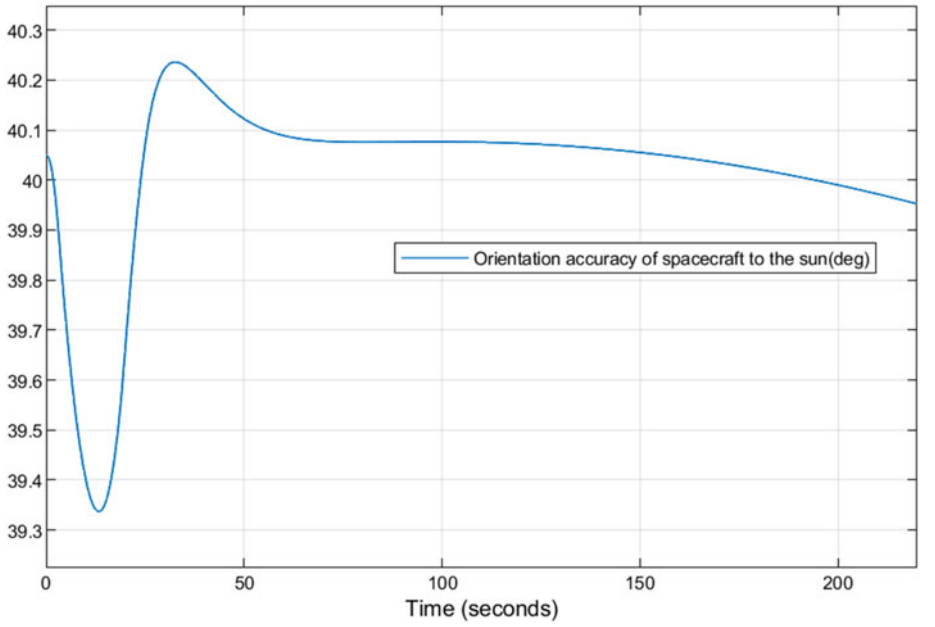


Figure 17. *Orientation accuracy to the Sun (CPF).*

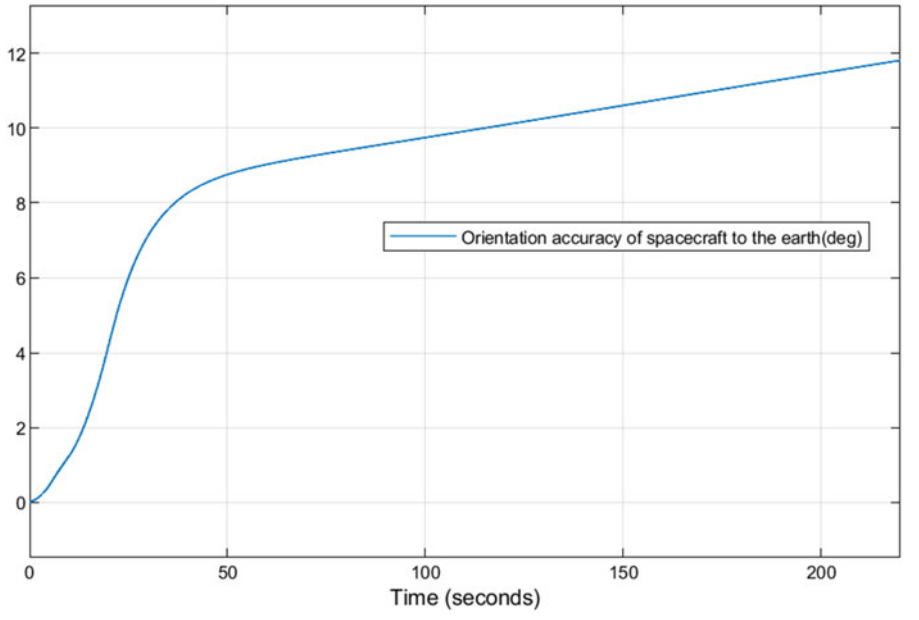


Figure 18. *Orientation accuracy to the Earth (CPF).*

the attitude manoeuvre trajectories based on dynamics and kinematics. However, the method proposed in this paper directly calculates the attitude manoeuvre trajectories, which indicates one step less than the control law method. From this perspective, this paper can be considered to provide a new idea for solving the problem of attitude manoeuvre planning.

https://doi.org/10.1017/S0373463321000813 Published online by Cambridge University Press

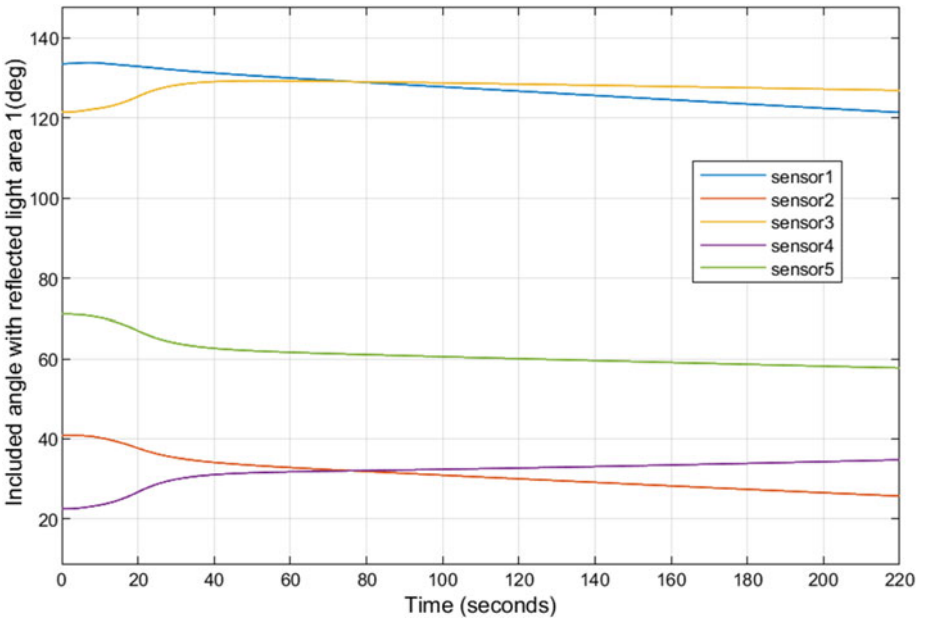


Figure 19. *Included angle with reflected light area 1 (CPF).*

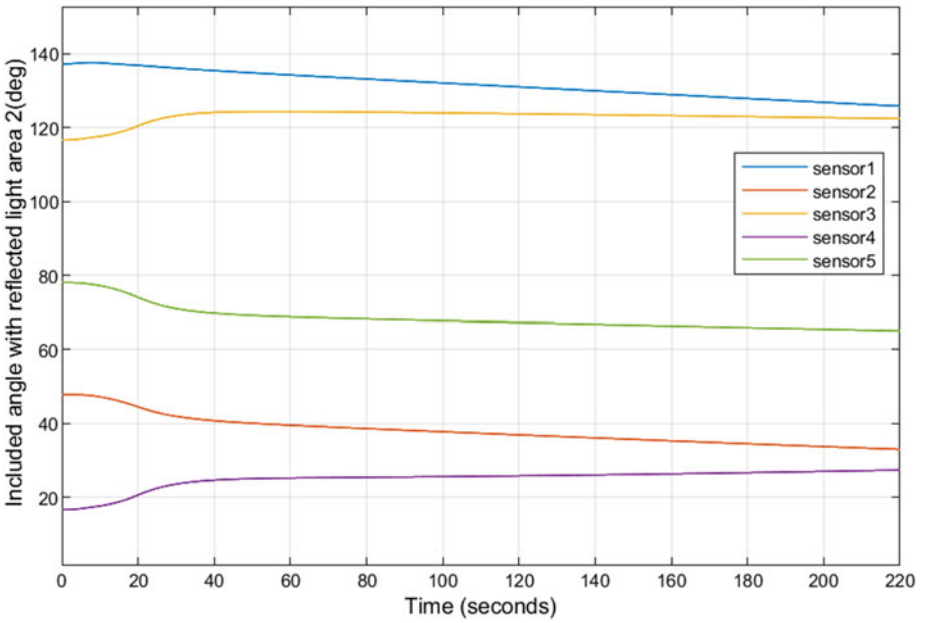


Figure 20. *Included angle with reflected light area 2(CPF).*

5. Conclusions

This paper studies the problem of attitude manoeuvre planning when spacecraft face multiple dynamic constraints. By introducing return baseline and adaptive policy exploration methods, it proposes an attitude manoeuvre planning algorithm based on improved policy gradient reinforcement learning. The attitude manoeuvre planning model proposed in this paper is concise and clearly defined, and can accurately describe multiple dynamic constraints of the spacecraft in real three-dimensional space.

https://doi.org/10.1017/S0373463321000813 Published online by Cambridge University Press

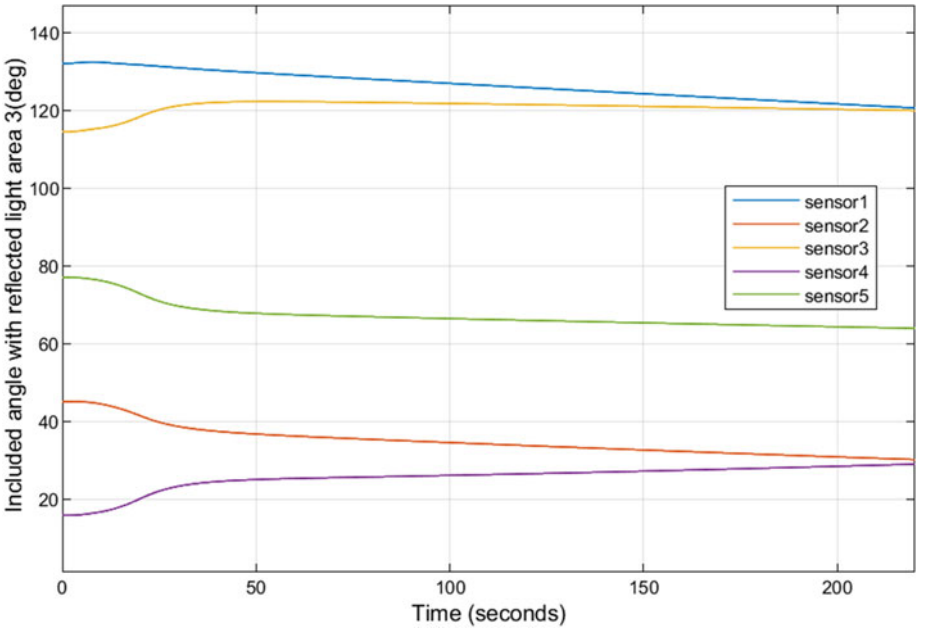


Figure 21. *Included angle with reflected light area 3 (CPF).*

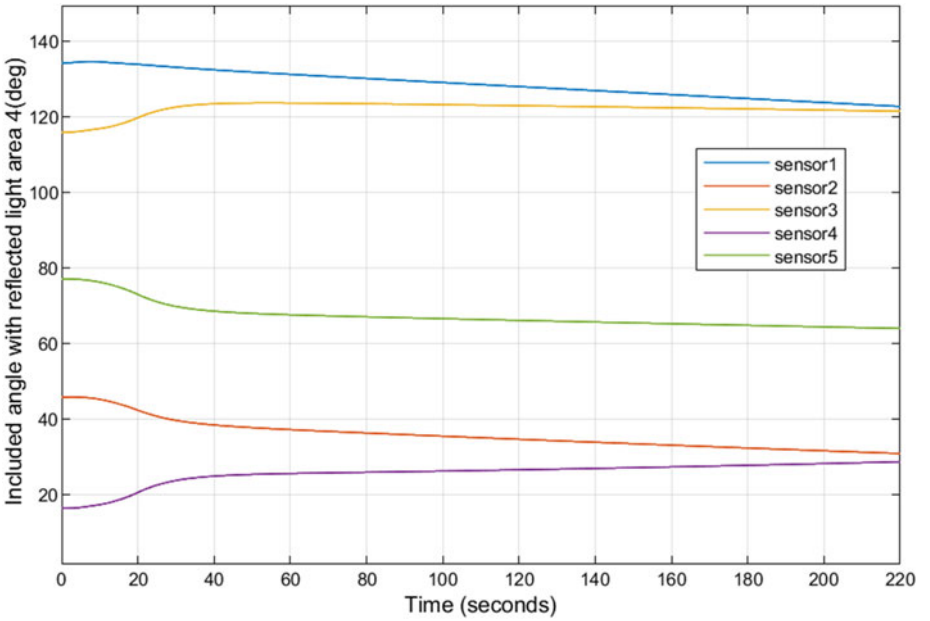


Figure 22. *Included angle with reflected light area 4 (CPF).*

The model exhibits strong practicability and is convenient to expand to include more constraints in the future. Simulation results show that the proposed method can simultaneously meet the requirements of multiple attitude constraints while ensuring high orientation accuracy to the Earth and the Sun at a small cost of attitude manoeuvres.

https://doi.org/10.1017/S0373463321000813 Published online by Cambridge University Press

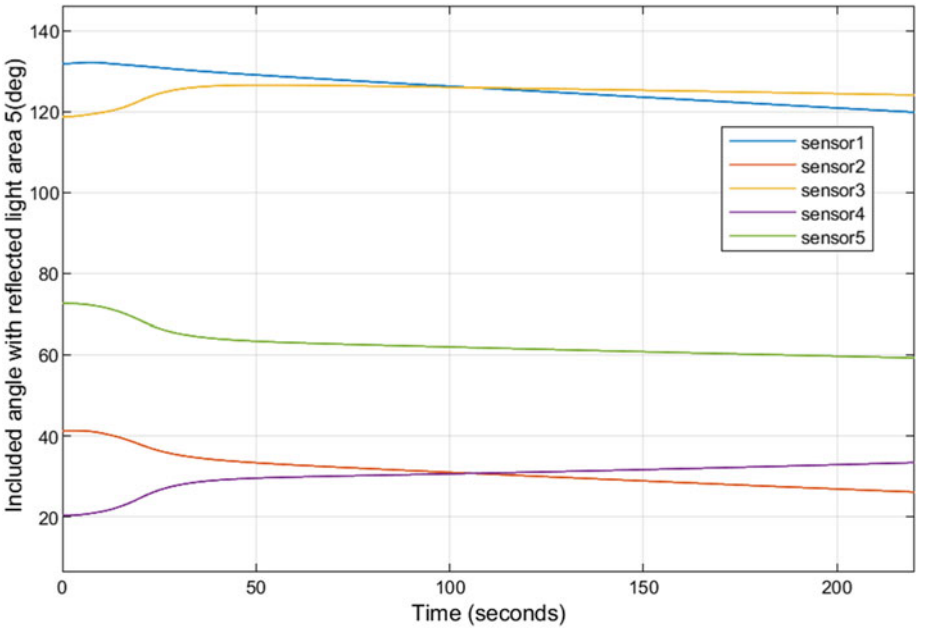


Figure 23. *Included angle with reflected light area 5(CPF).*

There are some limitations to this study:

1. A few iterations are required before obtaining the results of attitude manoeuvre planning. The planning method thus must be performed in advance.
2. The initial attitude of the attitude manoeuvre planning does not allow for Euler angles that equal zero, indicating that there are restrictions on the initial attitude. This feature may require additional attitude adjustments before attitude manoeuvres.

The future directions of this research are:

1. Design a more accurate value function to evaluate the strategy to avoid unnecessary posture manoeuvres.
2. Add the execution ability of the attitude control mechanism as a constraint to the value function.
3. Reduce computation time while considering the computing power of a satellite-based computer.

Funding statement

This work was supported by the National Natural Science Foundation of China (61973513).

References

Celani, F and Lucarelli, D. (2019). Spacecraft Attitude Motion Planning Using Gradient-Based Optimization. *Journal of Guidance, Control, and Dynamics*, **43**(6), 1–6.

Chen, T. D., Huang, Y. Y. and Zhang, Y. L. (2019). Non-trap dynamic path planning based on collision risk. *Systems Engineering and Electronics*, **41**(11), 2496–2506.

Cheng, Y. (2019). Research on Spacecraft Attitude Control Methods with State Constraints (Doctoral dissertation, Harbin Institute of Technology).

Cui, P., Zhong, W. and Cui, H. (2017). Onboard Spacecraft Slew-Planning by Heuristic State-Space Search and Optimization. International Conference on Mechatronics and Automation, IEEE Publ, Piscataway, NJ, USA, 2115–2119.

Greensmith, E., Bartlett, P. L. and Baxter, J. (2004). Variance reduction techniques for gradient estimates in reinforcement learning. *Journal of Machine Learning Research*, **2004**(5), 1471–1530.

https://doi.org/10.1017/S0373463321000813 Published online by Cambridge University Press

Guo, Y. J., Liu, Q., Bao, J. K., Xu, F. and Lv, W. H. (2020). Overview of AUV obstacle avoidance algorithm based on artificial potential field method. *Computer Engineering and Applications*, **56**(04), 16–23.

Hablani, H. B. (1999). Attitude commands avoiding bright objects and maintaining communication with ground station. *Journal of Guidance, Control, and Dynamics*, **22**(6), 759–767.

Huang, X. X., Li, S., Sun, P., Liu, X. W., Yang, B. and Liu, X. Y. (2021). Review of spacecraft guidance and control based on artificial intelligence. *Acta Aeronautica et Astronautica Sinica*, **42**(X), 524201 (in Chinese).

Kim, Y. and Mesbahi, M. (2004). Quadratically constrained attitude control via semidefinite programming. *IEEE Transactions on Automatic Control*, **49**(5), 731–735.

Kim, J. Y., Kim, Y. S., Kim, Y. K., Park, H. J., Kim, S. J., Kang, J. H., Wang, Y. P., Jang, H. S., Lee, S. N. and Yoon, S. C. (2015). On the convex parametrization of spacecraft orientation in presence of constraints and its applications. *IEEE Transactions on Aerospace and Electronic Systems*, 2010, **46**(3), 1097–1109.

Kjellberg, H. C. and Lightsey, E. G. (2016). Discretized quaternion constrained attitude path finding. *Journal of Guidance, Control, and Dynamics*, **39**(3), 713–718.

Lee, U. and Mesbahi, M. (2011) Spacecraft Reorientation in Presence of Attitude Constraints via Logarithmic Barrier Potentials. The 37th American Control Conference, San Francisco, USA, 29 June–1 July.

Liu, J. W., Gao, F. and Luo X. L. (2019). Survey of deep reinforcement learning based on value function and policy gradient. *Chinese Journal of Computers*, **42**(06), 1406–1438.

Ma, G. F., Liu, M. M., Wang, L. Y. and Guo, Y. N. (2020). Spacecraft backstepping attitude control considering multiple forbidden pointing regions. *Journal of Astronautics*, **41**(08), 1042–1048.

Miao, Y. M. and Pan, T. (2015). Path planning method for spacecraft attitude slew to avoid forbidden celestial. *Spacecraft Engineering*, **24**(04), 33–37.

Olivares, A. and Staffetti, E. (2018). Time-optimal attitude scheduling of a spacecraft equipped with reaction wheels. *International Journal of Aerospace Engineering*. 2018, 1–14.

Reed, R. S. (2020). A Machine Learning Approach to Enable Mission Planning of Time-Optimal Attitude Maneuver. Naval Postgraduate School, Monterey, California; Naval Postgraduate School.

Showalter, D. J. and Black, J. T. (2014). Responsive theater maneuvers via particle swarm optimization. *Journal of Spacecraft and Rockets*, **51**(6), 1976–1985.

Siliang, Y. and Shijie, X. (2010). Spacecraft attitude maneuver planning based on particle swarm optimization. *Journal of Beijing University of Aeronautics and Astronautics*, **2010**(1), 48–51.

Singh, G., Macala, G., Wong, E. and Rasmussen, R. D. (1997). A Constraint Monitor Algorithm for the Cassini Spacecraft. AIAA Journal, 1997.

Spiller, D., Ansalone, L. and Curti, F. (2016). Particle swarm optimization for time-optimal spacecraft reorientation with keep-out cones. *Journal of Guidance, Control, and Dynamics*, **39**(2), 312–325.

Tang, Z. T., Shao, K., Zhao, D. B. and Zhu, Y. H. (2017). Recent progress of deep reinforcement learning: from AlphaGo to AlphaGo zero. *Control Theory & Applications*, **34**(12), 1529–1546.

Tanygin, S. (2017). Fast autonomous three-axis constrained attitude pathfinding and visualization for boresight alignment. *Journal of Guidance, Control, and Dynamics*, **40**(2), 358–370.

Wang, G. F., Fang, Z. and Li, P. (2015). Natural actor-critic based on batch recursive least-squares. *Journal of Zhejiang University (Engineering Science)*, **49**(07), 1335–1342.

Wu, C. Q., Xu, R. and Zhu, S. Y. (2015). Deep space explorer attitude planning and control method based on logarithmic potential function. *Journal of Deep Space Exploration*, **2**(04), 365–370.

Xu, R., Wang, H., Zhu, S., Jiang, H. and Li, Z. (2020). Multiobjective planning for spacecraft reorientation under complex pointing constraints. *Aerospace Science and Technology*, **104**, 106002.

Zhao, T., Gang, N., Ning, X., Yang, J. C. and Masashi, S. (2015). Regularized policy gradients: direct variance reduction in policy gradient estimation. *JMLR: Workshop and Conference Proceedings*, **45**, 333–348.

Cite this article: Hua B, Sun S, Wu Y, Chen Z (2022). A spacecraft attitude manoeuvre planning algorithm based on improved policy gradient reinforcement learning. *The Journal of Navigation* **75**: 3, 662–684. https://doi.org/10.1017/S0373463321000813

https://doi.org/10.1017/S0373463321000813 Published online by Cambridge University Press