北京理工大學
BEIJING INSTITUTE OF TECHNOLOGY

IAA-PDC-21-07-29

HOVERING CONTROL FOR GRAVITY TRACTOR USING ASYNCHRONOUS METHODS FOR REINFORCEMENT LEARNING

Jucheng Lu, Bingwei Wei, Haibin Shang, Pingyuan Cui,
Rui Xu, Shengying Zhu, Ai Gao

Contents

1 Introduction

2 Hovering Problem Formulation

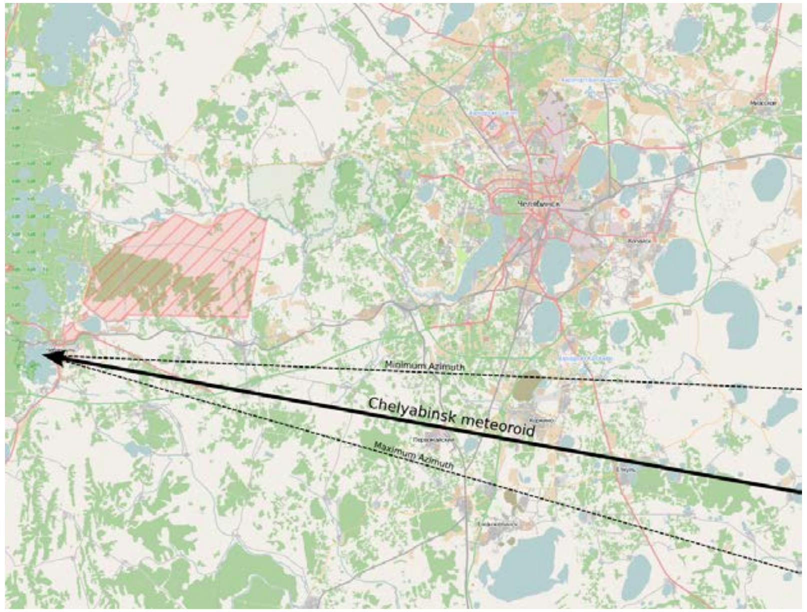
3 Asynchronous Advantage Actor-Critic(A3C)

4 Hovering Control of Gravity Tractor based on A3C

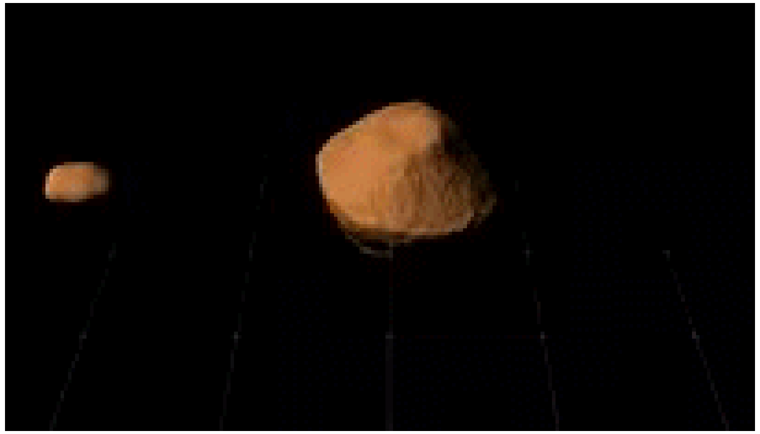
5 Numerical Simulations

6 Conclusion and Discussion

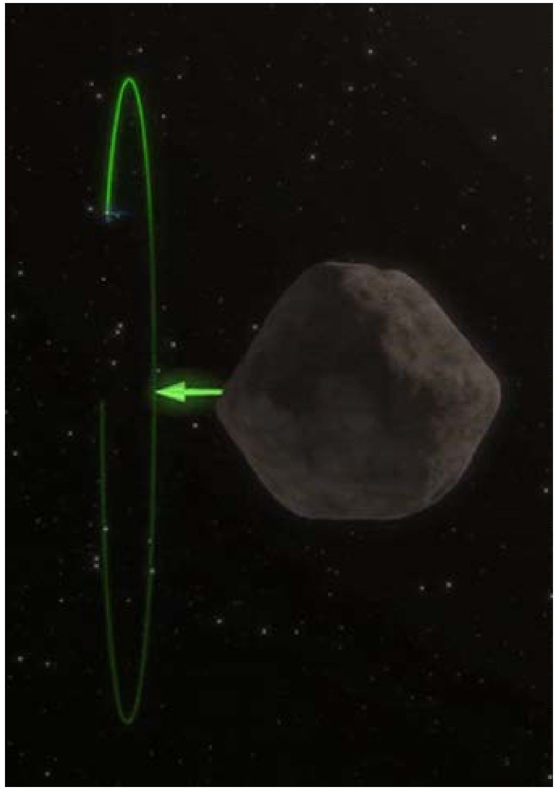
Introduction



Chelyabinsk meteor
[Source:https://en.wikipedia.org/wiki/Chelyabinsk_meteor#/media/File:Trajectory_of_Chelyabinsk_meteoroid_en.png]



Binary Asteroid System
[Source:https://en.wikipedia.org/wiki/66391_Moshup#/media/File:1999_KW4_animated.gif]



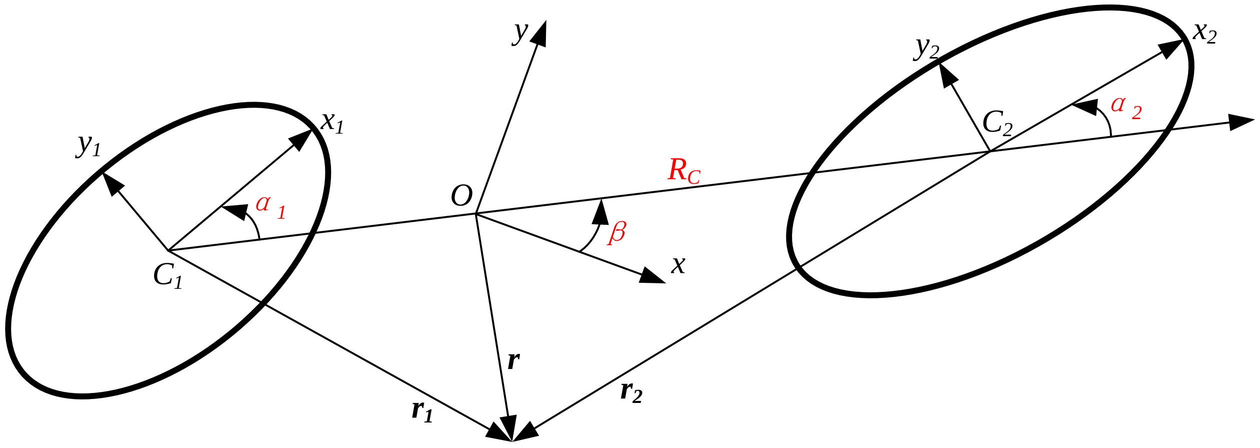
Gravity Tractor
[Source:https://en.wikipedia.org/wiki/Gravity_tractor]

Hovering Problem Formulation

➢ The Equations of Motion

$$
\begin{cases}
\ddot{x} = \dfrac{\partial U_1}{\partial x} + \dfrac{\partial U_2}{\partial x} + p_x + u_x \\
\ddot{y} = \dfrac{\partial U_1}{\partial y} + \dfrac{\partial U_2}{\partial y} + p_y + u_y \\
\ddot{z} = \dfrac{\partial U_1}{\partial z} + \dfrac{\partial U_2}{\partial z} + p_z + u_z
\end{cases}
$$

➢ Planar Motion in the Full Two-Body Problem

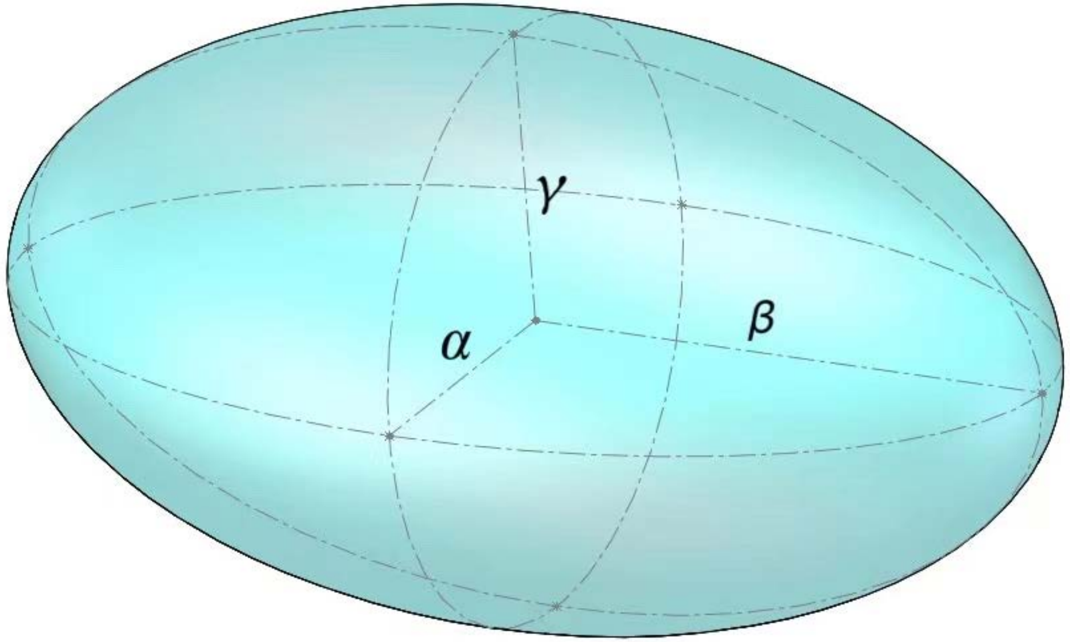


➢ Second Degree and Order Gravity Field

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

$$\begin{cases} C_{20} = -\frac{1}{2}\left(2I_{zz} - I_{xx} - I_{yy}\right) \\ C_{22} = \frac{1}{4}\left(I_{yy} - I_{xx}\right) \end{cases} \qquad \begin{cases} I_{xx} = \frac{\beta^2 + \gamma^2}{5} \\ I_{yy} = \frac{\alpha^2 + \gamma^2}{5} \\ I_{zz} = \frac{\alpha^2 + \beta^2}{5} \end{cases}$$

Triaxial Ellipsoid



Asynchronous Advantage Actor-Critic(A3C)

北京理工大学 BEIJING INSTITUTE OF TECHNOLOGY

➢ Reinforcement Learning Problem

Return:

$$R_t = \sum_{k=0}^{\infty} \gamma^k r_{t+k}$$

Probability Along the Episode:

$$P(\tau | \pi) = \rho_0(s_0) \prod_{t=0}^{T-1} P(s_{t+1} | s_t, a_t) \pi(a_t | s_t)$$

Object Function:

$$J(\pi) = \int_\tau P(\tau | \pi) R(\tau) = E_{\tau \sim \pi}[R(\tau)]$$

Optimal Policy:

$$\pi^* = \arg\max_\pi J(\pi)$$

➢ Value Function

Action Value Function:

$$Q^\pi(s,a) = E[R_t | s_t = s, a]$$

State Value Function:

$$V^\pi(s) = E[R_t | s_t = s]$$

➢ Advantage Function

$$A^\pi(s,a) = Q^\pi(s,a) - V^\pi(s)$$

Asynchronous Advantage Actor-Critic(A3C)

北京理工大学
BEIJING INSTITUTE OF TECHNOLOGY

➢ Policy Gradients

$$\theta_{t+1} = \theta_t + \alpha \nabla_\theta J(\pi_\theta)\big|_{\theta_t}$$

$$\nabla_\theta J(\pi_\theta) \approx \frac{1}{N}\sum_{i=1}^{N}\left[\left(\sum_{t=0}^{T-1}\nabla_\theta \log \pi_\theta(a_t|s_t)\right)\left(\sum_{k=0}^{T-1}\gamma^k r(s_{t+k},a_{t+k})\Big|_{t=0}\right)\right]$$

➢ Actor-Critic(AC)

actor:

$$\nabla_\theta J(\pi_\theta) \approx \frac{1}{N}\sum_{n=1}^{N}\sum_{t=0}^{T-1}\left[\nabla_\theta \log \pi_\theta(a_t^n|s_t^n) Q^{\pi_\theta}(s_t^n|a_t^n)\right]$$

critic:

$$loss = \frac{1}{N}\sum_{n=1}^{N}\sum_{t=0}^{T-1}\left[r_t^n + \max_{a_{t+1}^n} Q^{\pi_\theta}(s_{t+1}^n|a_{t+1}^n) - Q^{\pi_\theta}(s_t^n|a_t^n)\right]^2$$

➢ Advantage Actor-Critic(A2C)

actor:

$$\nabla_\theta J(\pi_\theta) = \frac{1}{N}\sum_{n=1}^{N}\sum_{t=0}^{T-1}\left[\left(Q^{\pi_\theta}(s_t^n|a_t^n) - V^{\pi_\theta}(s_t^n)\right)\nabla_\theta \log \pi_\theta(a_t^n|s_t^n)\right]$$

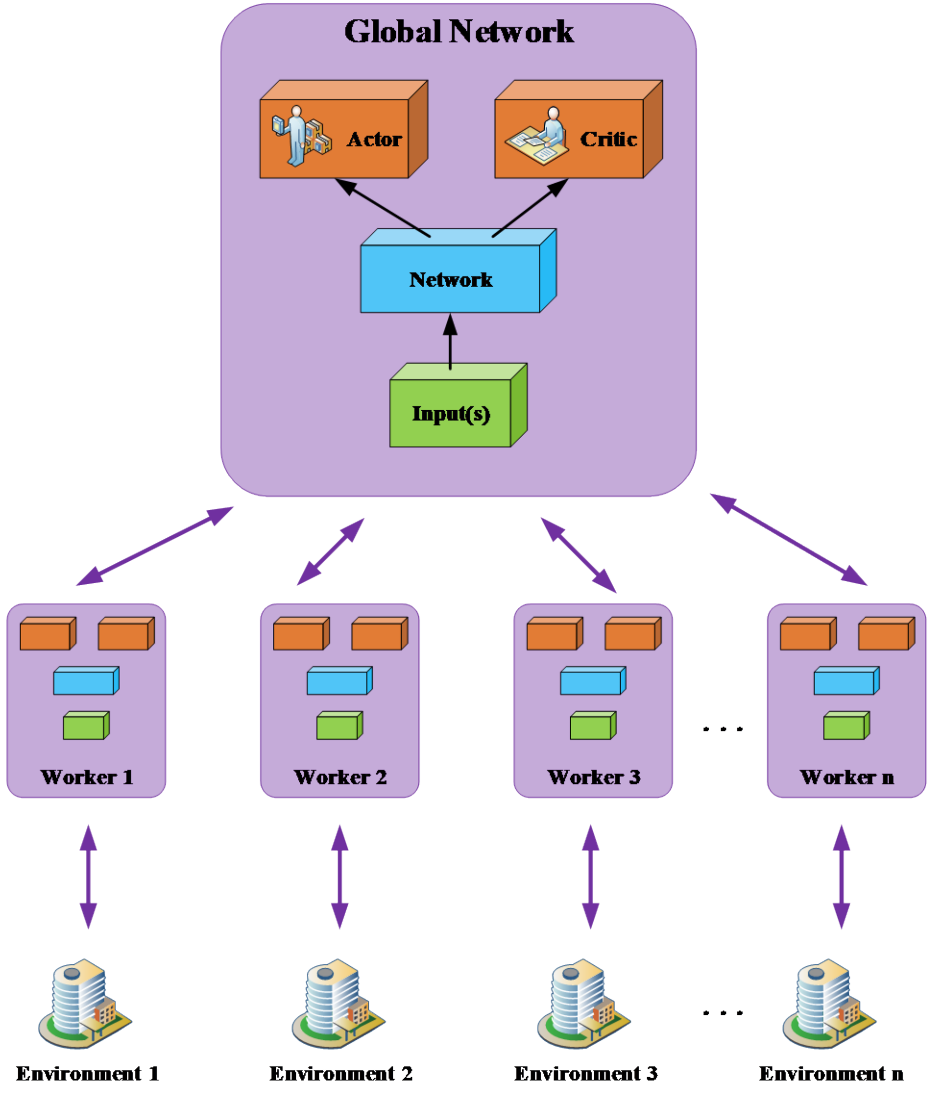
$$Q^{\pi}(s_t,a_t) = r_t + V^{\pi}(s_{t+1})$$

critic:

$$\nabla_\theta J(\pi_\theta) = \frac{1}{N}\sum_{n=1}^{N}\sum_{t=0}^{T-1}\left[\left(r_t^n + V^{\pi_\theta}(s_{t+1}^n) - V^{\pi_\theta}(s_t^n)\right)\nabla_\theta \log \pi_\theta(a_t^n|s_t^n)\right]$$

$$loss = \frac{1}{N}\sum_{n=1}^{N}\sum_{t=0}^{T-1}\left[r_t^n + V^{\pi_\theta}(s_{t+1}^n) - V^{\pi_\theta}(s_t^n)\right]^2$$

Asynchronous Advantage Actor-Critic(A3C)



Markov Decision Process

- **Markov Tuple**

S **State Space**

A **Action Space**

R **Reward**

P **Transition Probability**

Hovering Control Problem

$$S=\begin{bmatrix} x-x_d & y-y_d & z-z_d & \dot{x}-\dot{x}_d & \dot{y}-\dot{y}_d & \dot{z}-\dot{z}_d \end{bmatrix}$$

$$A=\begin{bmatrix} u_x & u_y & u_z \end{bmatrix}$$

$$R(s)=k_r\left\|\mathbf{r}-\mathbf{r}_d\right\|+k_v\left\|\dot{\mathbf{r}}-\dot{\mathbf{r}}_d\right\|$$

$$\ddot{\mathbf{r}}=\frac{\partial U_1}{\partial \mathbf{r}}+\frac{\partial U_2}{\partial \mathbf{r}}+\boldsymbol{u}$$

Numerical Simulations

Table.1 Physical parameters of the binary asteroid system and GT

Physical parameter	Magnitude	Unit
Thrust	[-1,1]	N
Mass of GT	10000	kg
Hovering Position	[6000,0,0]	m
Perturb	$[2,-3,4]\times10^{-5}$	m/s^2
Semi-axis of asteroid 1	[1.417,1.361,1.183]	km
Semi-axis of asteroid 2	[0.595,0.450,0.343]	km
Density of asteroid 1	1.97×10^{15}	kg/km^3
Density of asteroid 2	2.81×10^{15}	kg/km^3
Period of asteroid 1	2.7645	h
Period of asteroid 2	17.4223	h
Period of system	17.4223	h

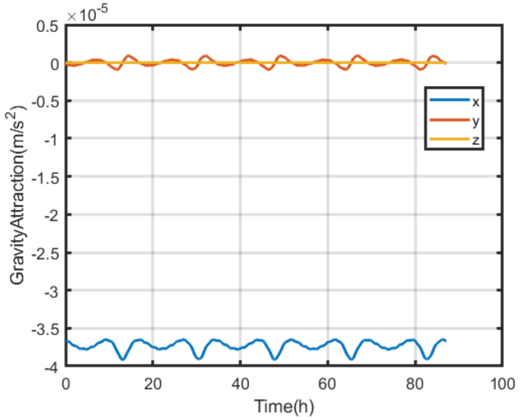


Fig.1 The gravity acceleration on the desire hovering position

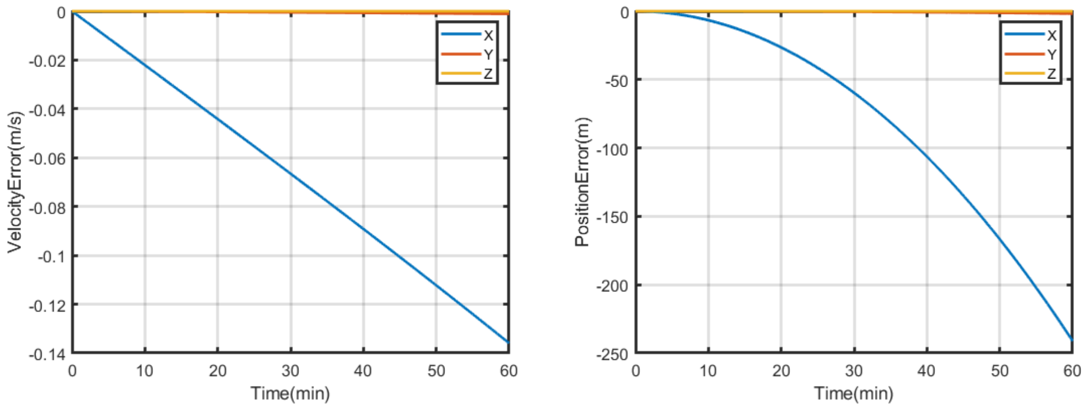


Fig.2 The deviation of the state without control

Numerical Simulations

Table 2 The NN architecture of the actor-critic frame

	Actor		Critic	
	units	activation	units	activation
Input Layer	6	/	6	/
Layer1	200	tanh	100	tanh
Layer2	200	tanh	100	tanh
Output Layer	3 3	tanh softplus	1	None

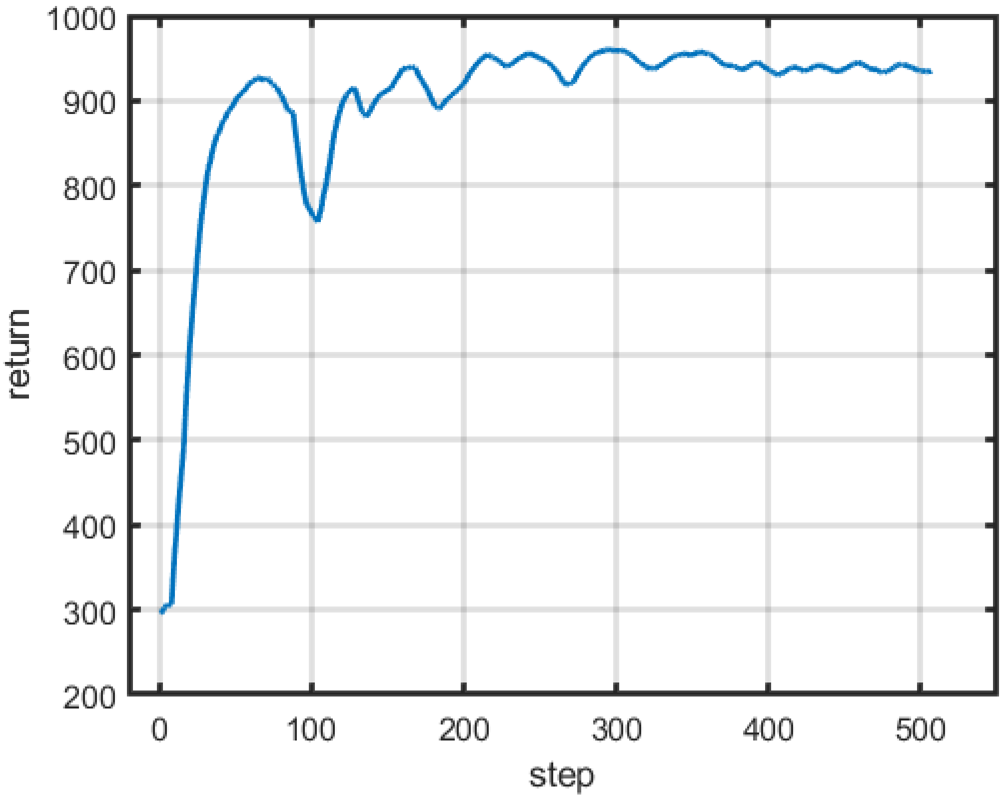


Fig.3 Policy optimization evolution

Numerical Simulations

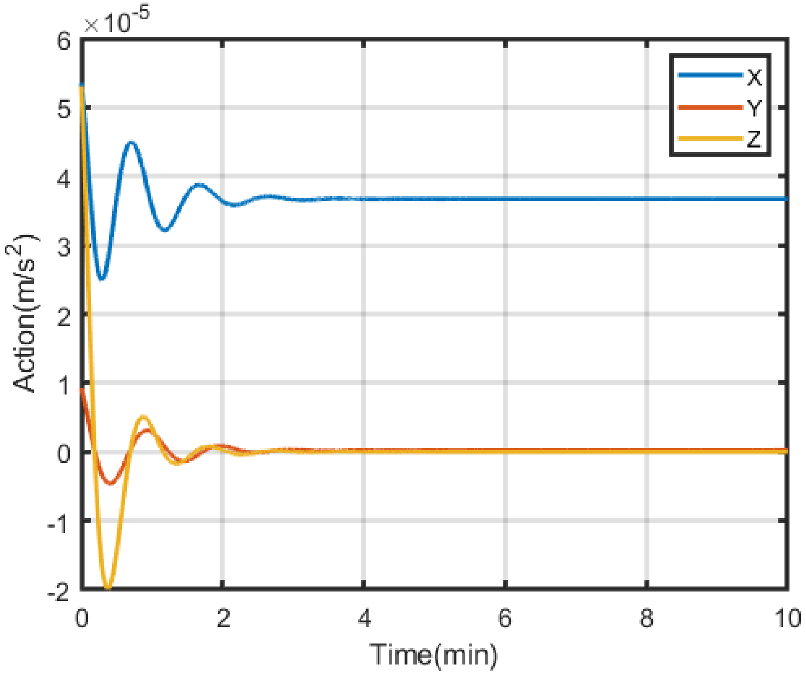
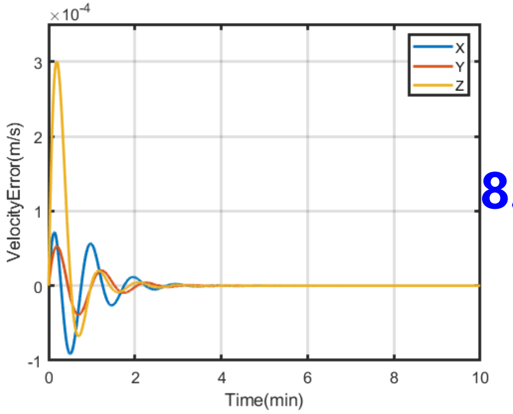
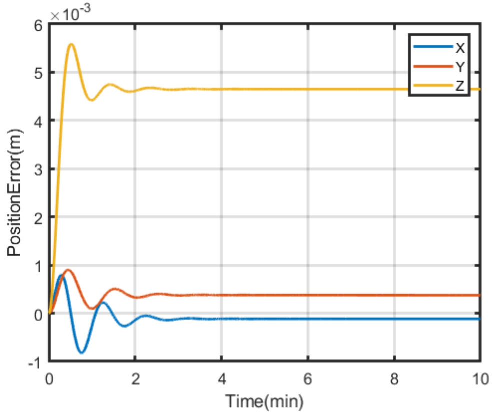


Fig.4 Acceleration command as a function of time in short-term



$8.6\times10^{-9}m/s^2$

Fig.5 The deviation of the velocity with policy π_1 in env_1



$4.7\times10^{-3}m$

Fig.6 The deviation of the state with policy π_1 in env_1

Numerical Simulations

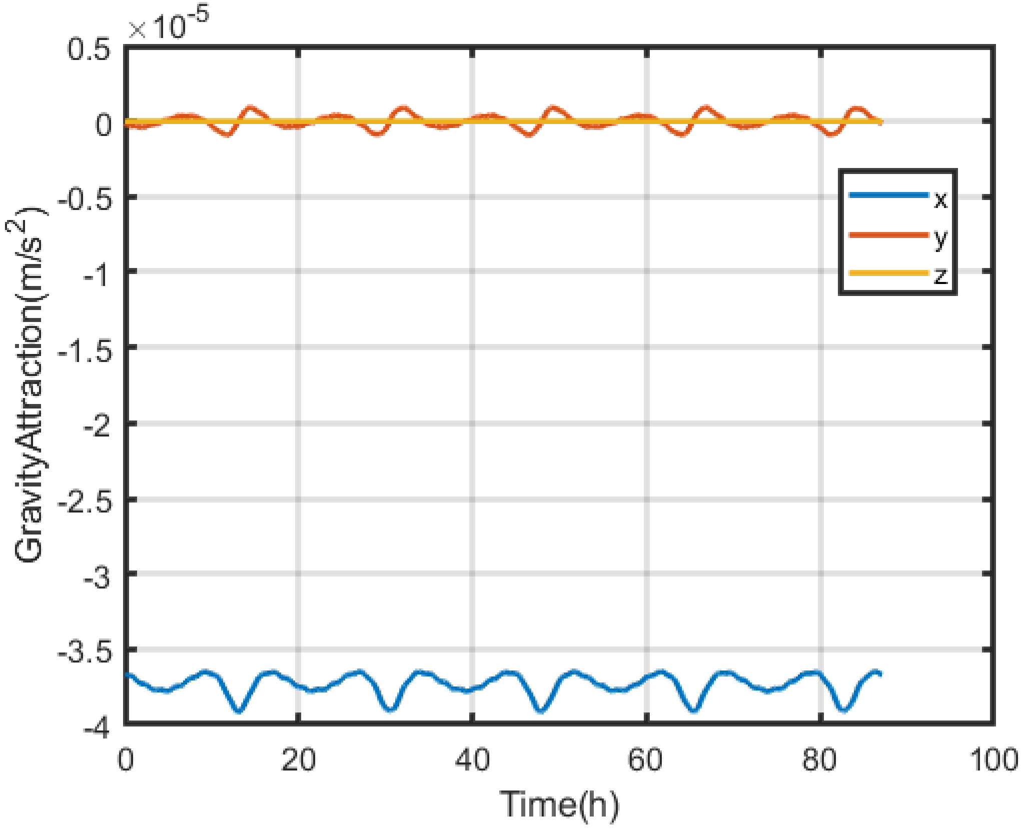


Fig.7 The gravity acceleration on the desire hovering position

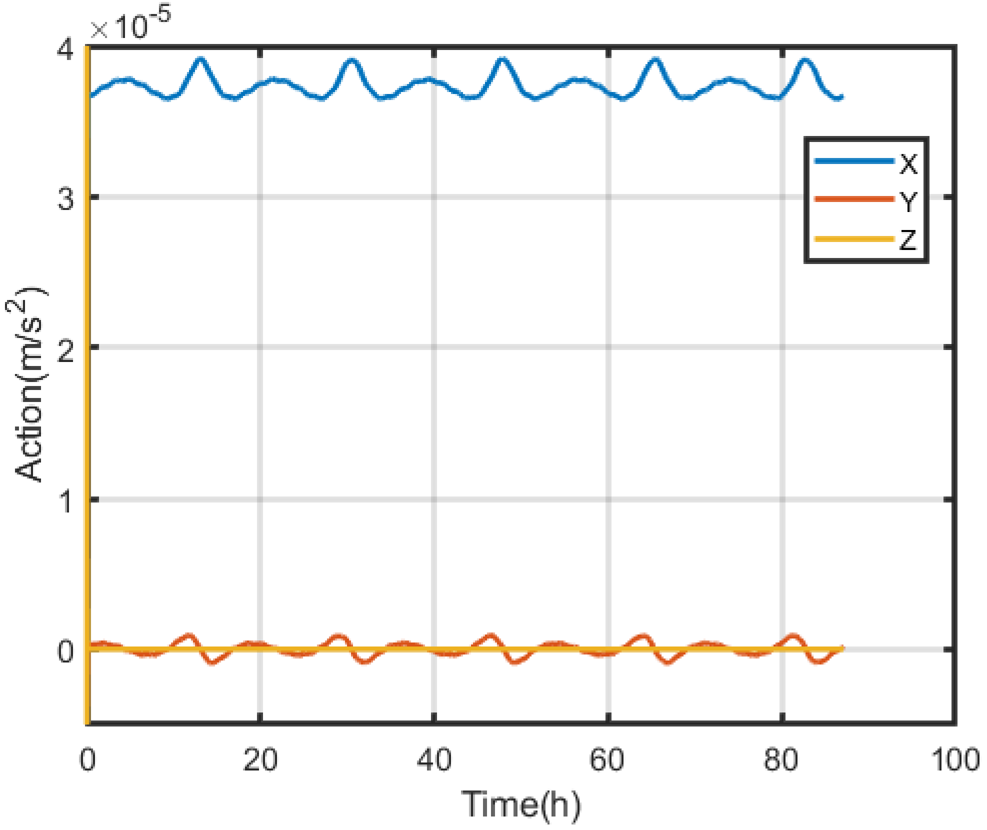


Fig.8 Acceleration command as a function of time in long-term

Numerical Simulations

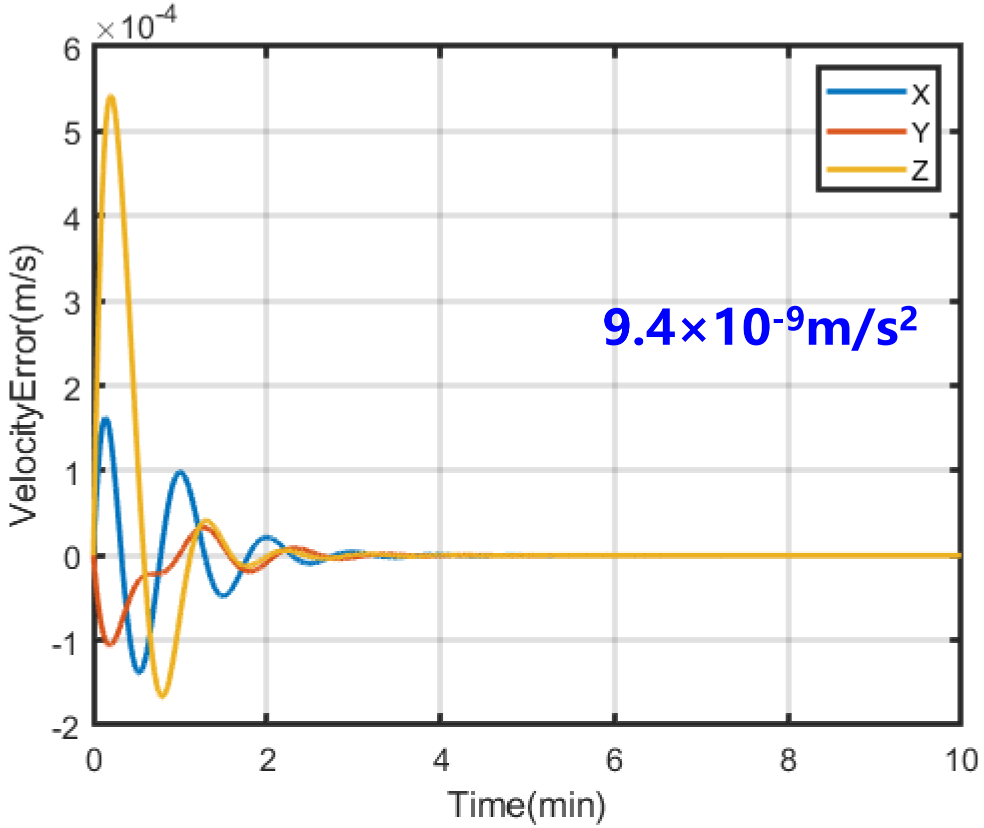


Fig.9 The deviation of the velocity with policy π_1 in env_2

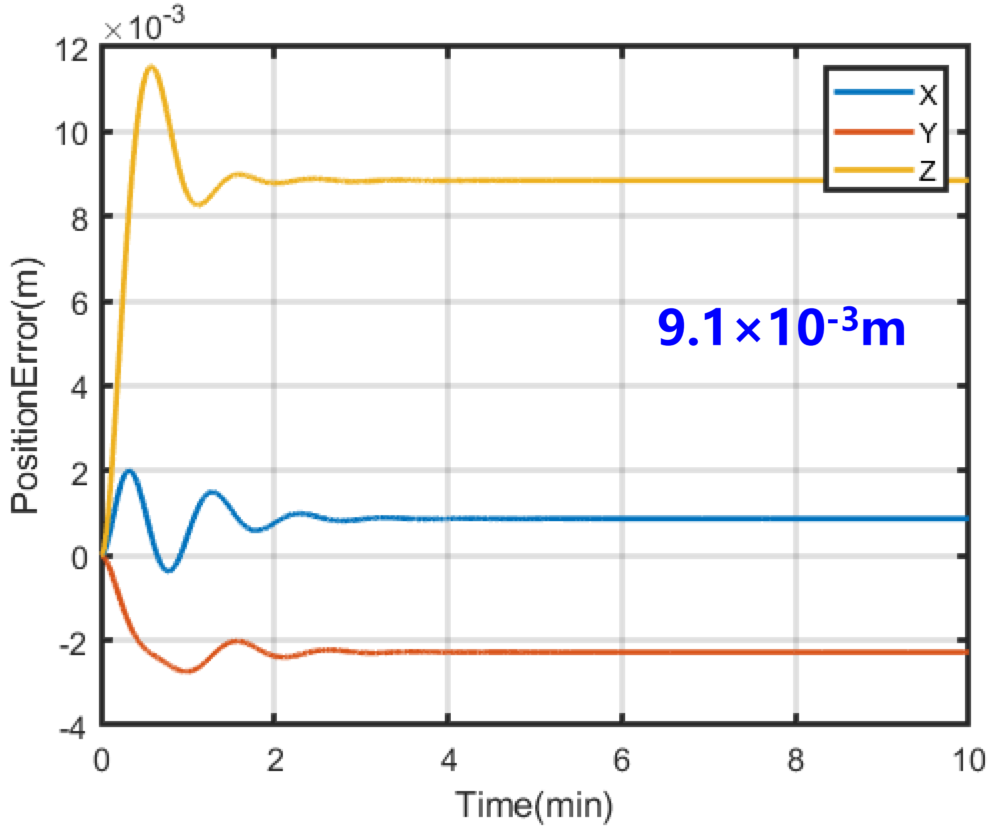


Fig.10 The deviation of the position with policy π_1 in env_2

Numerical Simulations

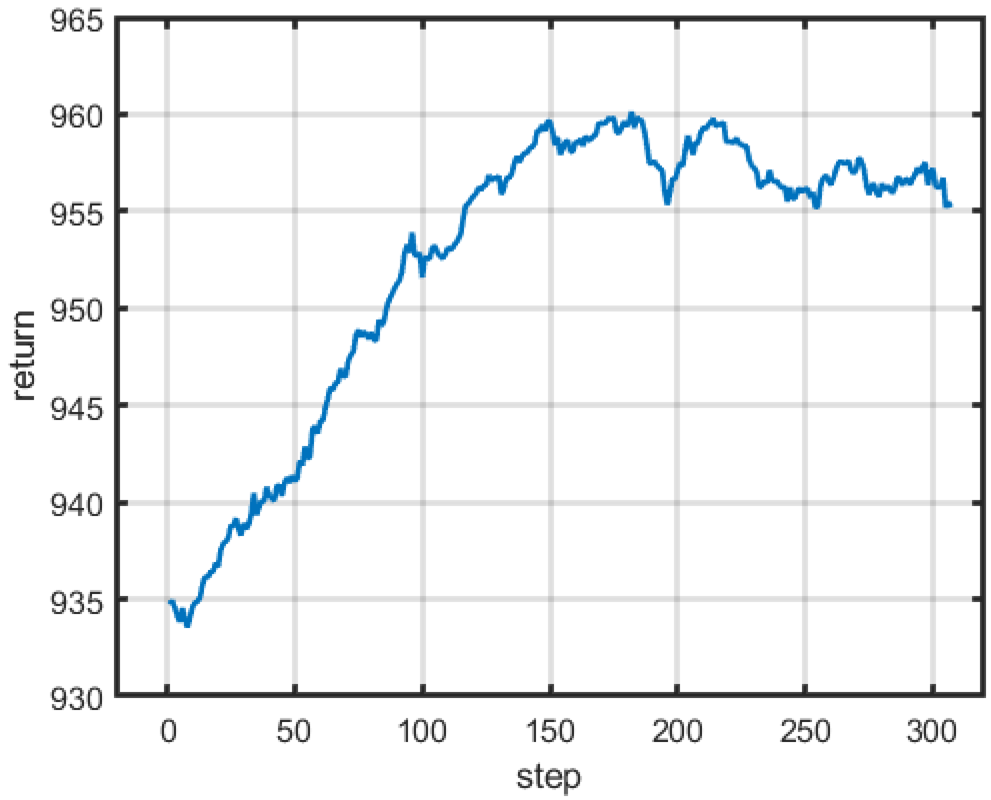


Fig.11 Policy optimization evolution

Numerical Simulations

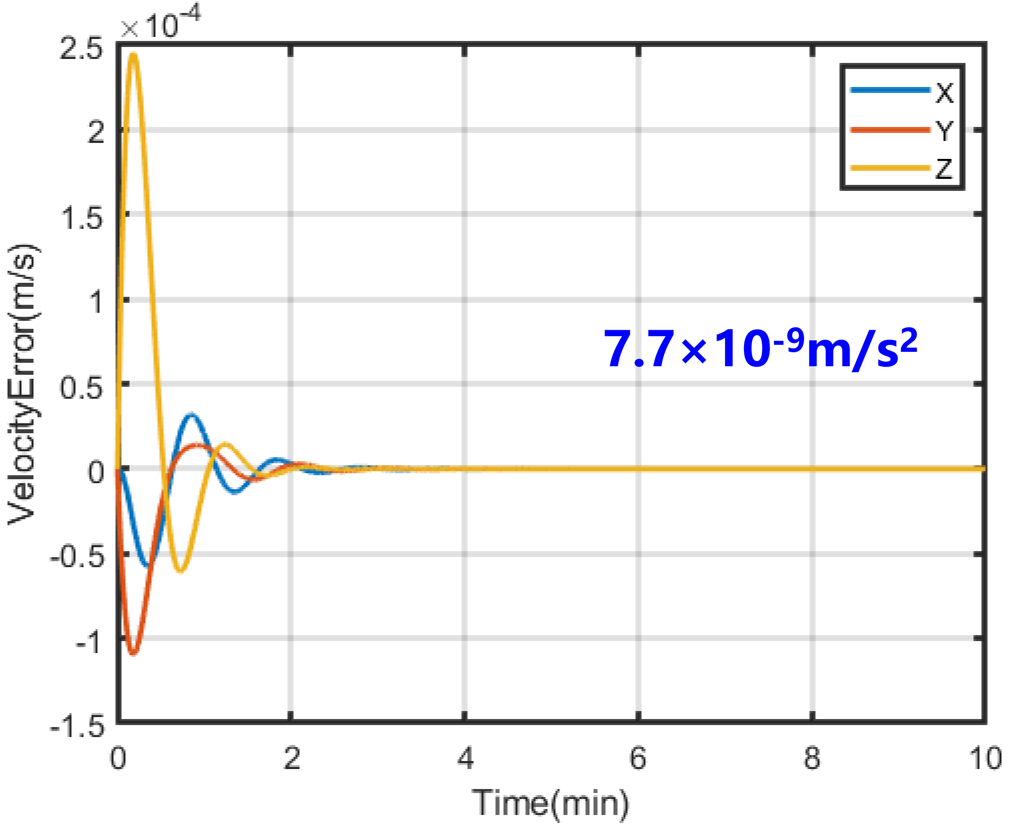


Fig.12 The deviation of the velocity with policy π_2 in env_2

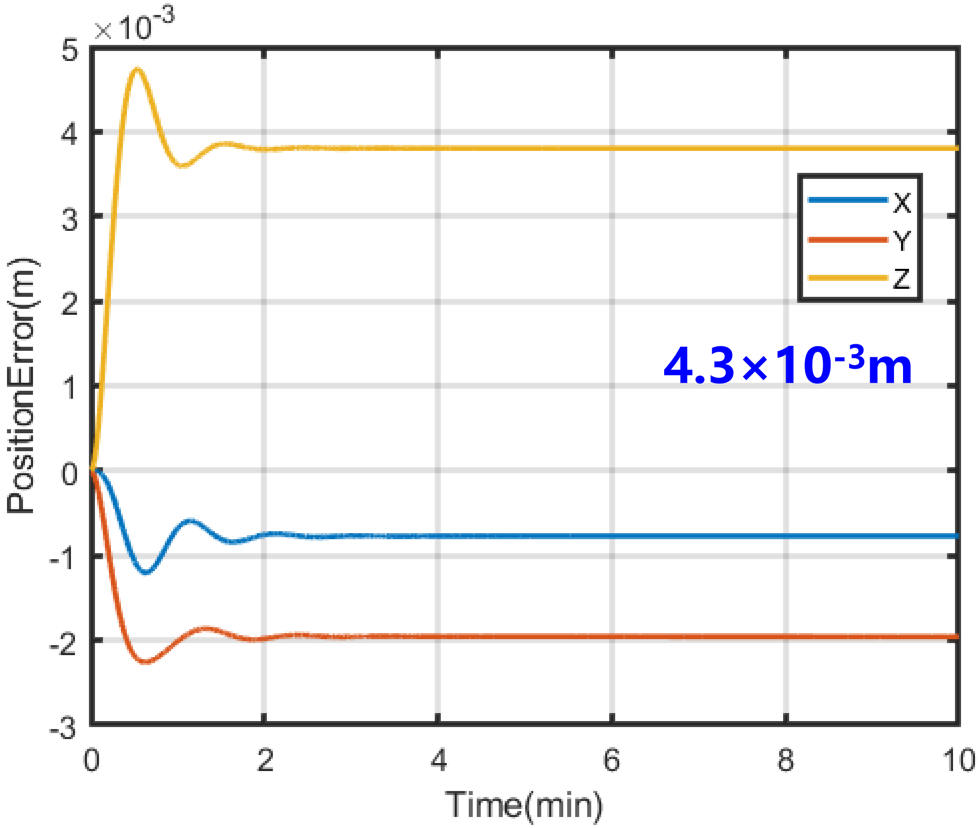


Fig.13 The deviation of the position with policy π_2 in env_2

Conclusion and Discussion

- This paper proposes that Reinforcement Learning(RL) could help the Gravity Tractor(GT) to maintain the hovering state and adapt to the change of the environment. The relationship mapping the Markov Decision Process(MDP) and the hovering control problem is established.
- The simulation results have demonstrated that the RL model could adapt to the change of the attraction on the hovering position. The RL algorithm employed here is Asynchronous Advantage Actor-Critic.
- A3C belongs to on-policy algorithm, which supports learn the data and update the policy during the mission. As a long-term mission, this operation can produce lots of samples to train the model. On the other hand, learning online helps the agent to maintain the control accuracy. The RL model could adapt the evolution of the environment.

北京理工大学
BEIJING INSTITUTE OF TECHNOLOGY

Thanks For Watching

德以明理 学以精工