Research on Inertial Space Intercept Game based on Deep Reinforcement Learning

Wenshu Wei¹, Bin Wang¹, Linjian Hou², Han Long²*

¹China Academy of Launch Vehicle Technology
²College of Arts and Sciences, National University of Defense Technology

*lonyhan@163.com

Abstract: Aiming at the problem of the intercept game in the inertial space, this paper builds a model of basic actions of the two sides of the intercept game in the inertial space, explored the applicability of Deep Reinforcement Learning in inertial space game problem-solving. Based on Proximal Policy Optimization (PPO), This inertial space game problem is solved and the optimal solution is obtained by the reward designing with cumulative miss distance and minimum distance respectively, and finally, the effectiveness of the algorithm based on PPO is verified through the game simulation and results from comparison.

Keywords: Inertial Space, Deep Reinforcement Learning, Proximal Policy Optimization (PPO), Game

1. Introduction

As a traditional method of studying inertial space game, Differential Game Theory has put forward many important theories and methods for inertial space game through many years of continuous development. Based on Dynamic Game Theory, the guidance law of optimal cooperative strategy is proposed [1]. Based on Cooperative Differential Game Theory, the optimal cooperative tracking strategy and the optimal target avoidance strategy of interceptors are designed [2]. With the application of cooperative guidance, the research on three-system guidance is deepening. Based on the relationship among target, interceptor, and defender, researchers studied and designed the optimal control strategy of the interceptor, the cooperative maneuver strategy of target and defender, and the optimal avoidance strategy of defender [3-5]. These theories greatly promote the research and development of inertial space game. However, the application of Differential Game Theory to game confrontation in inertial space also has some defects. Differential Game Theory requires the completeness of state information, which is reflected in the strict requirement of observation, almost assuming full observation, And some methods need to know the complete game matrix information, which is actually equivalent to knowing all strategies in advance. when solving game theory mainly based on Differential Game, it often falls into local optimal or there are multiple Nash equilibrium solutions.

Deep Reinforcement Learning (DRL) is an algorithm that combines Deep Learning and Reinforcement Learning to realize end-to-end Learning from state to action. In 2013, Google DeepMind put forward Deep Reinforcement Learning for the first time [6], which combined Reinforcement...
Learning and Deep Learning, Agent relies on pure image input to play Atari games completely through self-learning. In 2015, Google Deep Mind published an article in Nature [7], using DRL to achieve human-level control on Atari games. In 2016, AlphaGo, developed by the Google DeepMind team based on the Monte Carlo game tree search and DRL, won the game against the world champion of Go Lee Sedol, which attracted wide attention from the society. DRL has many advantages for solving game confrontation, Deep convolutional network can automatically learn the abstract representation of high-dimensional input data and effectively solve the problem of domain knowledge representation and acquisition in complex tasks. At the same time, its strategy is expressed as a mapping from state to action. The optimal strategy is selected through random sampling and other methods, without knowing the game strategy in advance. Some Reinforcement Learning methods can find better strategies through training and exploration under partial observation conditions. In addition, DRL can be solved by a variety of methods. Under the condition that the excitation function is suitable, there are few cases of multiple solutions and non-convergence. In this paper, basic actions of the two sides of inertial space games are modeled, the applicability of DRL in solving inertial space game confrontation is explored, and the penetration effect of inertial space game is verified by simulation.

2. Basic action modeling of inertial space game

2.1. EKV model

EKV adopts an infrared seeker, and its main performance parameters are field of view Angle 2×2 degrees, capture distance 200KM, frame frequency 25, 50, 100HZ optional, and infrared focal plane array: 256×256[8,9].

EKV guidance equation [10-12]:

\[ a_m = KV\lambda + 0.5Ka_f \]  \hspace{1cm} (1)

Where \( K \) is the proportional guidance coefficient and \( a_f \) is the maneuvering acceleration of the target.

2.2. Guidance Law modeling for interceptor missiles

Command acceleration of TPN:

\[ a_{TPN} = N\hat{\rho} \dot{\hat{\rho}} \]  \hspace{1cm} (2)

Fig.1 Geometric relationship between the 3D line-of-sight coordinate system and the transmitting inertial system

Geometric relationship between the 3D line-of-sight coordinate system and the transmitting inertial system is shown in the Fig. 1. Relative motion equation of 3d projectile established under the line of sight is as follows:
\begin{align*}
\ddot{r} - r \left( \omega_y^2 + \omega_z^2 \right) &= a_x - a_{mx} \\
\dot{r} \omega_z + 2 \dot{r} \omega_x + r \omega_x \omega_y &= a_y - a_{my} \\
\dot{r} \omega_y + 2 \dot{r} \omega_x - r \omega_x \omega_z &= a_z + a_{mz}
\end{align*}

\text{(3)}

Command acceleration of 3D TPN in the line of sight coordinate system:

\[ a_m = 0, a_{my} = N | \dot{r} | \omega_z, a_{mz} = -N | \dot{r} | \omega_y \]

\text{(4)}

The sum of Command accelerations:

\[ a_m = -N \dot{r} q_x \cdot y_x + -N \dot{r} q_y \cos q_x \cdot z_x = N \dot{r} \cdot (-q_x \cdot y_x + \dot{q}_y \cos q_x \cdot z_x) \]

\text{(5)}

Available:

\[ a_m = -N \dot{r} \omega_x e_\theta \]

\text{(6)}

3. Application of Deep Reinforcement Learning

In this paper, Proximal Policy Optimization (PPO) is adopted. PPO is a new Policy Gradient algorithm. Policy Gradient is very sensitive to the step size, but it is difficult to choose the appropriate step size. If the difference between the old and the new strategies is too large in the training process, it is not conducive to learning. PPO proposed a new objective function that could realize small-batch updates through multiple training steps, thus solving the problem that the step size was difficult to determine in Policy Gradient. The most commonly used gradient estimator has the form:

\[ \hat{g} = \hat{E}_{i} \left[ \nabla_{\theta} \log \pi_{\theta} (a_{i} | s_{i}) \hat{A}_{i} \right] \]

\text{(7)}

The estimator \(\hat{g}\) is obtained by differentiating the objective:

\[ L(\theta) = \hat{E}_{i} \left[ \log \pi_{\theta} (a_{i} | s_{i}) \hat{A}_{i} \right] \]

\text{(8)}

Policy update of PPO-clip:

\[ \theta_{k+1} = \arg \max_{\theta} \left( \frac{1}{E} \left[ L(s, a, \theta_{k}, \theta) \right] \right) \]

\text{(9)}

If the agent maneuvers so that missile-target distance at the final intersection exceeds a certain threshold, we give it a corresponding, appropriate, and positive reward. Thus, \(L\) in formula (10) is increased to the maximum.

\[ L(s, a, \theta_{k}, \theta) = \min \left( \frac{\pi \theta(a | s)}{\pi \theta_k(a | s)} A^{\pi \theta_k(s, a)} \cdot \text{clip} \left( \frac{\pi \theta(a | s)}{\pi \theta_k(a | s)}, 1 - \varepsilon, 1 + \varepsilon \right) A^{\pi \theta_k(s, a)} \right) \]

\text{(10)}

4. Simulation and Analysis

4.1. Simulation Setting

The observable quantities of Intelligent Missile are Position of Missile relative to TPN, TPN Velocity, and Missile Velocity. We assumed that the Intelligent Missile only makes a decision once in each iteration. The contents of Decision include maneuvering position, maneuvering duration, maneuvering Angle, and maneuver force. Initial action is designed as a discrete vector, and the size of the action space is 2. Taking the action as 0 means no maneuver while taking the action as 1 means that Intelligent Missile takes maneuvering action. Training parameters are shown in Tab.1.
4.2. Analysis of results

After training, Intelligent Missile should find a maneuver position around 0.7F. Training results are shown in Fig.2:

![Fig.2 Training results](image)

The green track is the trajectory of the agent after the maneuver, while the blue track is the trajectory without maneuver. The visualization result shows that the maneuver of the agent makes the penetration successful.

The result indicates that Intelligent Missile chooses to maneuver at the beginning, which is bad training. The agent will only choose the endpoint value to maneuver each time, which is a typical sparse reward problem in Reinforcement Learning [13]. With the improved reward factor, the agent can make partially correct decisions, which solves the problem that the agent can only choose to maneuver at the endpoint, but it can only do the one maneuver. However, the fact is that the agent can maneuver multiple times, even though it will be punished for multiple maneuvers. After analysis and research, we have designed a set of mutually exclusive maneuvers.

Retrain after improvement, After 12,000 iterations of training, the mean reward of Agent tends to be stable and the std of the reward tends to 0.

Tab.2 shows some local optimal solutions found by agents based on PPO with cumulative miss distance as the variable reward value and discrete action space.

| Maneuver Position | Duration | X-force | Y-force | Z-force | Minimum Distance |
|-------------------|----------|---------|---------|---------|------------------|
| Experiment 1      | 1320     | 1       | 2       | 3       | 0                | 21.92174 |
| Experiment 2      | 740      | 1       | 3       | 3       | 1                | 34.75   |
| Experiment 3      | 1210     | 1       | 3       | 3       | -1               | 34.3302 |

**Tab.2** With cumulative miss distance as reward value
As shown in Tab.3. With cumulative miss distance as reward value, with the minimum distance as the reward value, discrete action space, and some local optimal solutions found by Agents based on PPO.

| Maneuver Position | Duration | X-force | Y-force | Z-force | Minimum Distance |
|-------------------|----------|---------|---------|---------|-----------------|
| Experiment 1      | 410      | 3       | 0       | -2      | 31.32491        |
| Experiment 2      | 1160     | 2       | 1       | -2      | 20.22684        |
| Experiment 3      | 1470     | 2       | -2      | -1      | 18.61787        |

**Tab.3** With cumulative miss distance as reward value With the minimum distance as the reward value

5. Conclusion

Through simulation, PPO model based on deep reinforcement learning shows good applicability in the application of the inertial space intercept game. Under the conditions of cumulative miss distance and minimum distance as reward values, the optimal solutions close to expert-level can be obtained, and the effectiveness of Deep Reinforcement Learning is verified.

References

[1] Rusnak I 2010 Guidance Laws in Defense against Missile Attack with Acceleration Constrained Players *J. AIAA Guidance, Navigation and Control Conference* **8057**

[2] Perelman A and Shima T 2011 Cooperative Differential Games Strategies for Active Aircraft Protection from a Homing Missile *J. AIAA Guidance, Navigation and Control Conference* **34(3)** pp 761-773

[3] Shima T and Perelman A 2011 Optimal Cooperative Pursuit and Evasion Strategies Against a Homing Missile *J. AIAA Journal of Guidance Control and Dynamics* **34(2)** pp 414-425

[4] Shima T and WeissH 2002 Differential Game Guidance Considering Interceptor’s Direct Lift and Range Dependent Estimation Dealy *J. Proceedings of the 42th Conference on Decision and Control* **2** pp 1281-1286

[5] Ganebny S A, Kumkov S S and Menec S L 2012 Model Problem in a Line with Two Pursuers and One Evader *J. Dynamic Games and Applications* **2(2)** pp 228-257

[6] Mnih V, Kavukcuoglu K and Silver D 2013 Playing Atari with Deep Reinforcement Learning *J. Computer Science* **154(4)** pp 67-69

[7] Mnih V, Kavukcuoglu K and Silver D 2015 Human-level control through deep reinforcement learning *J. Nature* **518(7540)** pp 529-533

[8] Ahmet Tarik Aydin 2005 Orbit Selection and EKV Guidance for Spacebased ICBM Intercept *D. California, USA: Naval Postgraduate School*

[9] J Herring, L Bollengier and D Madajian 1998 Staring 256x256 LWIR Focal Plane Array Performance of the Raytheon Exoatmospheric Kill Vehicle

[10] Pin-Jar Y, Ming-Ghows C and Men-Jyue L 2001 Optimal Guidance of Extended Proportional Navigation *C.AIAA Guidance,Navigation, and Control Conference and Exhibit*

[11] Steve Rogers Missle guidance comparison *C.AIAA Guidance,Navigation, and Control Conference and Exhibit*

[12] QIANG S, Jing Y and Dongyun Y 2007 Mathematical Modeling and Simulations of Tactical Ballistic Missile Interception *J. Journal of System Simulation* **(S1004-731X)** **19(7)** pp 1424-1427+1484.

[13] Weiyi Y and Chenjia B 2020 Survey on Sparse Reward in Deep reinforcement Learning *J. Computer Science* **47**