Electronic Research Archive

http://www.aimspress.com/journal/era

ERA, 30(12): 4366–4381.
DOI: 10.3934/era.2022221
Received: 21 June 2022
Revised: 02 September 2022
Accepted: 15 September 2022
Published: 30 September 2022

Research article

Online intelligent maneuvering penetration methods of missile with respect to unknown intercepting strategies based on reinforcement learning

Yaokun Wang[1], Kun Zhao[2], Juan L. G. Guirao[3,4], Kai Pan[1] and Huatao Chen[1,*]

[1] Division of Dynamics and Control, School of Mathematics and Statistics, Shandong University of Technology, Zibo 255000, China
[2] Beijing Electro-Mechanical Engineering Institute, Beijing 100074, China
[3] Department of Applied Mathematics and Statistics, Technical University of Cartagena, Hospital de Marina, Cartagena 30203, Spain
[4] Department of Mathematics, Faculty of Science, King Abdulaziz University, P. O. Box 80203, Jeddah 21589, Saudi Arabia

* **Correspondence:** Email: htchencn@sdut.edu.cn.

Abstract: This paper considers the maneuvering penetration methods of missile which do not know the intercepting strategies of the interceptor beforehand. Based on reinforcement learning, the online intelligent maneuvering penetration methods of missile are derived. When the missile is locked by the interceptor, in terms of the tracking characteristics of the interceptor, the missile carries out tentative maneuvers which lead to the interceptor makes the responses respectively, in the light of the information on interceptor responses which can be gathered by the missile-borne detectors, online game confrontation learning is employed to increase the miss distance of the interceptor in guidance blind area by reinforcement learning algorithm, the results of which are used to generate maneuvering strategies that make the missile to achieve the successful penetration. The simulation results show that, compared with no maneuvering methods or random maneuvering methods, the methods proposed not only present higher probability of successful penetration, but also need less overload and lower command switching frequency. Moreover, the effectiveness of this maneuvering penetration methods can be realized under the condition of limited number of training.

Keywords: penetration of missile; artificial intelligence; Q-learning; reinforcement learning

1. Introduction

The study of maneuvering penetration technologies which can improve the combat effectiveness of missiles is a hot topic in the research field of guidance and control, there are many methods about

midpiece maneuvering penetrations, such as pre-procedural maneuvering penetration strategies [1]; actively evasive maneuvering penetration methods [2]; penetration methods based on flying around the detection interception area [3] and so on.

In the process of pre-procedural maneuvering penetration methods, the trajectories of penetration are preset before launching, the missiles can not make any responses with respect to strategies of the interceptor, alternatively, the missile flies along the preconcerted trajectories and is abided by the preconcerted maneuvering timing. Compared with the pre-procedural maneuvering penetration strategies, the actively evasive maneuvering penetration strategies, which can obtain the the optimal penetration strategies based on the parameters of interceptor detected by missile-borne computer, can increase the rate of successful penetration. There are two mainly maneuvering guidance strategies in the actively evasive maneuver penetrations, one is differential game type [3], the other is matrix game type [4]. From the mathematical standpoint, to derive the differential game typical guidance strategies is equivalent to solve a bilateral extremum problems for the associated functionals, but it is very hard to find the analytical solutions for these kinds of functionals. With respect to achieve the numerical solutions, it costs a lot of computing resource of the missile-borne computer to keep the high precision and real-time performance. In light of the finite two-person zero-sum game theory, matrix game typical guidance strategies consider the missile and interceptor as the players, the target-missing quality and its negative value are regraded as their payments. This method needs a lot of information on the motion of the interceptor, for more details, one can refer to [5].

Recently, with the improvement of the computing ability of computer, the artificial intelligence have been developed rapidly [6], investigations on theory and applications of reinforcement learning (RL) [7–13] is very important. In order to obtain the best operating action of the whole system, the RL can make the intelligent agent to select the behaviors which can gain the maximum reward of the environment state by learning the mappings from the environmental states to the behaviors. There are many literatures on the theory and application of RL. For instance, Bradtke and Duff [14] considered the continuous-time Markov decision problems by using RL. Taking advantage of the RL, some problems in transport were studied by Abdulhai and Kattan [15]. Based on the RL, Lewis et al. [16] designed optimal adaptive controllers by using natural decision methods. In the achievements on dynamics and control problem of robotics, the RL was also employed wildly [17]. As for the applications of RL in economics, one can refer to [18]. With respect to literatures on dynamics and control of missiles or aircrafts by RL, Shalumov [19] proposed cooperative online guide-launch-guide policy for the target-missile-defender engagement. The computational guidance problem of missile was considered in [20]. A homing-phase guidance law of missile was considered in [21]. The scenarios of avoiding Obstacles via missile real-time inference belong to Hong and Park [22]. Gaudet et al. gave an angle-only intercept guidance of maneuvering targets [23]. missile guidance for head-on interception of maneuvering target was established by Li et al. [24]. A planar evasive maneuvering strategy of aircrafts was derived in [24], too name but a few. For more detail, one can refer to [11, 25–28].

Actually, RL is a kind of self-regulated learning method driven by experiences, the maneuvering penetration technologies based on RL are nearly trained with predictable intercepting strategies, which means that the maneuvering penetration technologies are useless if the intercepting strategies can not be acquired beforehand. In order to overcome this obstacle, this paper is devoted to propose online intelligent maneuvering penetration methods with respect to non-predictable intercepting strategies based on RL, the main idea is that: let the tentative maneuvers of missile and line of sight rates (LOS)

rate between missile and interceptor be actions and states respectively. Inducing the interceptor to generate the responses with respect to the actions, which can be captured by the missile-borne detectors. In the light of the increment of LOS rates, the reward function can be designed. Thus, by means of the information gathered by missile-borne detectors, the the maneuvering penetration strategies can be derived by missile-borne computer based on RL.

The rest of this paper is organised as follows. The preliminaries are given in Section 2. Section 3 is devoted to demonstrate the main results, in which the online intelligent maneuvering penetration methods are proposed and some numerical results are accomplished to validate this new method. In Section 4, the conclusions are listed.

2. Preliminaries

For brevity, some symbols are introduced firstly, let D and M be the interceptor and the missile respectively, $\mathbf{r} = (r_x, r_y, r_x)$ is the relative distance between the missile and the interceptor. Based on the kinematic theory, the relationship between the missile and the interceptor in attack-defense confrontation can be described by Figure 1. θ_1 is the ballistic inclination angle of the interceptor, the ballistic declination angle of the interceptor is denoted by φ_1. θ_2, φ_2 are the ballistic inclination angle and the ballistic declination angle of missile respectively. v_M is the velocity of missile, and the velocity of interceptor is signified by v_D, thus $v_r = v_D - v_M$ is the relative velocity between the missile and the interceptor. $\mathbf{q} = (q_x, q_y, q_z)$ is LOS angle in the ground coordinate system, the horizontal and vertical LOS angle are denoted by q_φ and q_θ. Let (x_M, y_M, z_M) and (x_D, y_D, z_D) be the positions of missile and interceptor in ground coordinate system respectively.

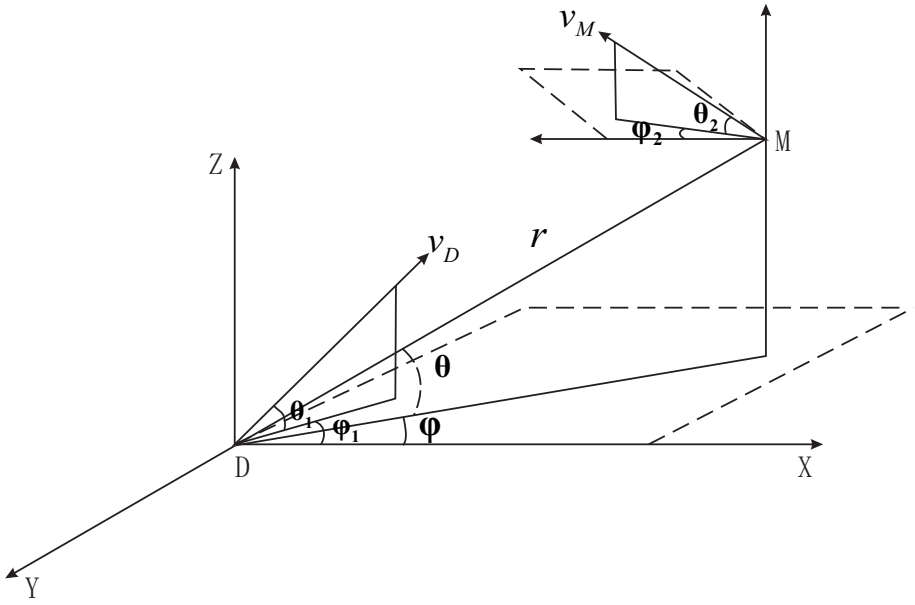


Figure 1. The relative motion of the missile and the interceptor.

By [29], the mathematical model of relative motion between the missile and the interceptor can be described as follows

$$\begin{cases} \frac{dr_x}{dt} = v_D \cos\theta_1 \cos\varphi_1 - v_M \cos\theta_2 \cos\varphi_2 \\ \frac{dr_y}{dt} = v_D \cos\theta_1 \sin\varphi_1 - v_M \cos\theta_2 \sin\varphi_2 \\ \frac{dr_z}{dt} = v_M \sin\theta_2 - v_D \sin\theta_1, \end{cases} \tag{2.1}$$

the equation governing the dynamics of LOS angle is

$$\begin{cases} \frac{dq_\varphi}{dt} = \cos\varphi_2 \frac{q_y}{dt} - \sin\varphi_2 \frac{dq_x}{dt} \\ \frac{dq_\theta}{dt} = \sin\theta_2 \left(\cos\varphi_2 \frac{dq_x}{dt} + \sin\varphi_1 \frac{dq_y}{dt}\right) + \cos\theta_1 \frac{dq_z}{dt}, \end{cases} \tag{2.2}$$

the position of missile can be described by

$$\begin{cases} \frac{dx_M}{dt} = v_M \cos\theta_2 \cos\varphi_2 \\ \frac{dy_M}{dt} = -v_M \cos\theta_2 \sin\varphi_2 \\ \frac{dz_M}{dt} = v_M \sin\theta_2, \end{cases} \tag{2.3}$$

the following equations model the position of the interceptor

$$\begin{cases} \frac{dx_D}{dt} = v_D \cos\theta_1 \cos\varphi_1 \\ \frac{dy_D}{dt} = -v_D \cos\theta_1 \sin\varphi_1 \\ \frac{dz_D}{dt} = v_D \sin\theta_1. \end{cases} \tag{2.4}$$

Next, it is turned to give the description on interceptor guidance blind area.

In general, exoatmospheric kill vehicle (EKV), which consists of guidance systems, transferring orbital control systems, propulsion systems and so on (for more detail, see Figure 2), destroys the missile by collision. The homing guidance system should be interrupted when the distance between EKV and missile is less than or equal to a certain quantity R which is referred as guidance blind area. Obviously, the EKV is uncontrollable in guidance blind area, in addition, the interceptor guidance blind area is about 30 m to 500 m in practice. Figure 3 indicates the graphical representation on instantaneous miss distance d of EKV [30, 31] which can be determined as following,

$$d = \frac{R^2}{|\dot{R}|} \|\dot{\mathbf{q}}\| \tag{2.5}$$

where R is the instantaneous distance between the missile and the interceptor, $\dot{R}$ denotes the instantaneous relative velocity when both the missile and the interceptor just enter the guidance blind area, $\|\dot{\mathbf{q}}\| = \sqrt{\dot{q}_\varphi^2 + \dot{q}_\theta^2}$. Clearly, the instantaneous miss distance is proportional to $\|\dot{\mathbf{q}}\|$, R and is inversely proportional to $|\dot{R}|$. Since the time of both missile and EKV are in guidance blind area is so short that the missile can not maneuver, thus, the instantaneous miss distance can be regarded as the actual miss distance.

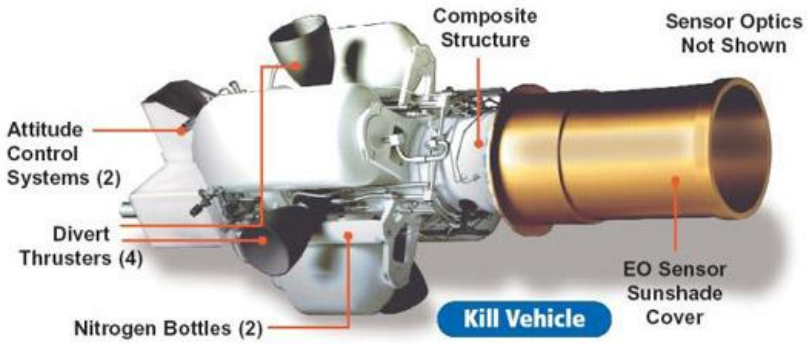


Figure 2. EKV schematic.

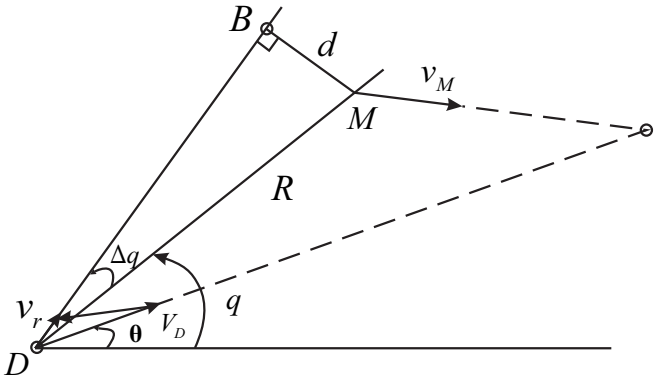


Figure 3. Graphical representation on instantaneous miss distance.

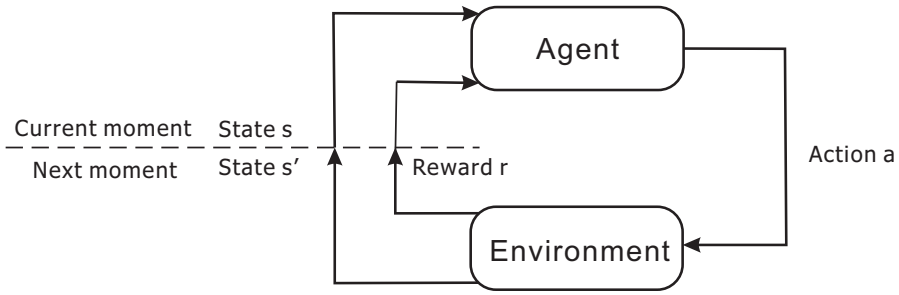


Figure 4. Interaction between agent and environment.

The RL can be divided into two kinds, one is using the same strategy to update the value function and select a new action, such as Sarsa method. The other is utilizing the different strategies to do these things, for instance, Q-learning (QL) method. In QL, the interactions between agent and environment are modelled by the markov decision process $(\mathcal{S}, \mathscr{A}, \mathbb{P}, \overline{R}, \gamma)$, where $\mathcal{S}$ is the states, $\mathscr{A}$ is a $\sigma-$ algebra generated by the sets which is composed of all possible actions, $\mathbb{P}$ is the transition probability of one state to another state, $\overline{R}$ is the reward function, γ is discount factor. Let $S_t \in \mathcal{S}$ be the state at time t,

$A_t \in \mathscr{A}(S_t)$ is the action, here $\mathscr{A}(S_t)$ is the set of actions available in state S_t, for simplicity, we use $\mathscr{A}(t)$ instead of $\mathscr{A}(S_t)$. $\overline{R}_t \in \overline{R}$ is the reward, and S_{t+1} is the next state of the agent (see Figure 4). Let

$$G_t = \overline{R}_{t+1} + \gamma \overline{R}_{t+2} + \gamma^2 \overline{R}_{t+3} + \cdots = \sum_{k=0}^{\infty} \gamma^k \overline{R}_{t+k+1}, \tag{2.6}$$

where $0 \leq \gamma \leq 1$ is the discount rate. For all $s', s \in \mathcal{S}$, $r \in \overline{R}$ and $a \in \mathscr{A}(s)$, let

$$p(s', \overline{r}|s, a) = \mathbb{P}\{S_t = s', \overline{R}_t = r|S_{t-1} = s, A_{t-1} = a\}, \tag{2.7}$$

suppose π is the policy, π_* is the optimal policy, $\pi(a|s)$ is the probability of getting the action a at time s, then the associated state-value function is given as following

$$v_\pi(s) = \mathbb{E}_\pi[G_t|S_t = s], \tag{2.8}$$

the action-value function is

$$\overline{q}_\pi(s, a) = \mathbb{E}_\pi[G_t|S_t = s, A_t = a]. \tag{2.9}$$

Together with (2.6)–(2.9) and the markov property, we have

$$v_\pi(s) = \sum_{a \in \mathscr{A}(s)} \pi(a|s) \sum_{s' \in \mathcal{S}, \overline{r} \in \overline{R}} p(s', \overline{r}|s, a)[\overline{r} + \gamma v_\pi(s')], \tag{2.10}$$

and

$$\overline{q}_\pi(s, a) = \sum_{s' \in \mathcal{S}, \overline{r} \in \overline{R}} p(s', \overline{r}|s, a)[\overline{r} + \gamma \sum_{a \in \mathscr{A}(s)} \pi(a'|s') \overline{q}_\pi(s', a')], \tag{2.11}$$

thus, the optimal state-value function is

$$v_{\pi_*}(s) = \max_{a \in \mathscr{A}(s)} \sum_{s' \in \mathcal{S}, \overline{r} \in \overline{R}} p(s', \overline{r}|s, a)[\overline{r} + \gamma v_{\pi_*}(s')], \tag{2.12}$$

and the optimal action-value function is

$$\overline{q}_{\pi_*}(s, a) = \sum_{s' \in \mathcal{S}, \overline{r} \in \overline{R}} p(s', \overline{r}|s, a)[\overline{r} + \gamma \max_{a' \in \mathscr{A}(s)} \overline{q}_{\pi_*}(s', a')]. \tag{2.13}$$

Generally, the following $\epsilon - greedy$ strategy is employed in QL

$$A = \begin{cases} \arg\max_{a \in \mathscr{A}} Q(s, a) & \text{with probability } 1 - \varepsilon \\ \text{a random action} & \text{with probability } \varepsilon \end{cases},$$

where $Q(s, a)$ is the estimated value of the action value function $\overline{q}_{\pi_*}(s, a)$).

The flow chart of QL algorithm is as following, where $\mathcal{S}^+$ is the state space containing the termination state, and

Algorithm 1. (Q-learning (off-policy TD control) for estimating $\pi \approx \pi_*$)

Q-learning (off-policy TD control) for estimating $\pi \approx \pi_*$
Algorithm parameters: step size $\alpha \in (0, 1]$, small $\epsilon > 0$
Initialize $Q(s, a)$, for all $s \in \mathcal{S}^+$,$a \in \mathcal{A}(s)$,arbitrarily except that $Q(terminal, \cdot) = 0$
Loop for each episode:
    Initialize S
    Loop for each step of episode:
        Choose A from S using policy derived from $Q(e.g., \epsilon - greedy)$
        Take action A, observe R, S'
        $Q(S, A) \leftarrow Q(S, A) + \alpha[R + \gamma max_a Q(S', a) - Q(S, A)]$
        $S \leftarrow S'$
    until S is terminal

3. Main results

3.1. Design on the intelligent maneuvering penetration strategy

Firstly, we propose the flow chart of the maneuvering penetration methods of the missile with respect to unknown intercepting strategies by QL methods (see Figure 5).

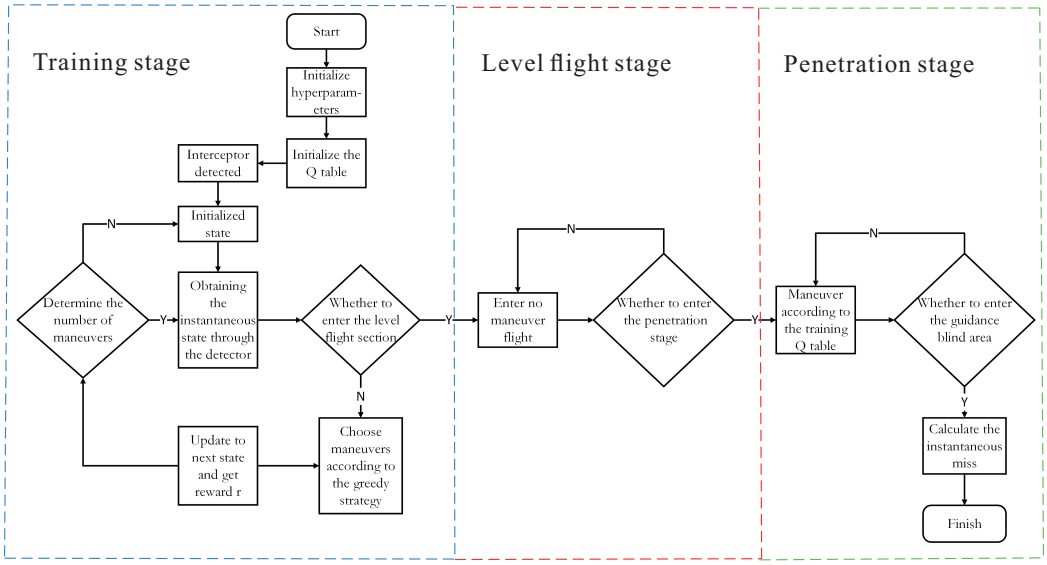


Figure 5. The flow chart of the intelligent maneuvering penetration strategies proposed.

Dividing the combat process between the missile and the interceptor into three stages. After the missile is locked by the interceptor, let the missile make tentative maneuvers to induce the interceptor to correspondingly generate the responses, based on the information of responses of the interceptor, maneuvering penetration strategies of the missile are trained by QL method. This is the training stage.

After that, the missile do not make any tentative maneuver, thus, the second stage is named as level flight stage. The third stage is referred to as penetration stage, the missile produces the maneuvering penetration strategies in terms of the training results and calculate instantaneous miss distance of EKV in guidance blind area to implement successful penetration.

It is emphasized the maneuvering penetration strategies indicated above do not need to know the law of the interceptor beforehand, alternatively, the necessary information for QL training is acquired by missile-borne detectors, and the maneuvering penetration strategies based on QL training is done by the missile-borne computer. Thus, the maneuvering penetration strategies are intelligent and real-time.

In the following, we realize the flow chart described by Figure 5. To begin with, the following markov decision process model is established.

1) The states space: the instantaneous antagonistic states between the missile and the interceptor are depended on the LOS rates, thus, let the LOS rates be the state with the following form

$$\begin{aligned}\mathcal{S} = [&-0.8, -0.4, -0.06, -0.03, -0.02, -0.01, -0.009, -0.006, -0.00188,\\ &-0.00162, -0.00138, 0.00138, 0.00162, 0.00188, 0.006, 0.009, 0.01,\\ &0.02, 0.03, 0.06, 0.1, 0.4, 0.8],\end{aligned} \tag{3.1}$$

   the unit is rad/s.
2) The actions space: the offense-defense confrontation between the missile and the interceptor can be formulated by pursuit-evasion game, the actions of which consist of up, down, left and right. In order to consider the penetration ability of missile with lower maneuverability, we set the optional overloads are 2 and 3.5 along with each direction of the up, down, left and right, the ultimate overload of interceptor is supposed to be 8, thus, we establish the actions of missile as follows

$$\mathscr{A} = [2, 3.5, -2, -3.5, 2, 3.5, -2, -3.5]. \tag{3.2}$$

3) The reward function: the purpose of the test maneuvers of missile is to find the sequences which can result in the increasing of LOS rates. In order to generate a large value of LOS rate during the penetration, which can lead to the failure of proportional guidance method of EKV, the value of reward should be big when LOS rate is big and vice versa. Together with (2.5) and the amount of instantaneous off-target required, the critical value of LOS rate can be derived. If the LOS rate is larger than this critical value, the bigger value of reward function should be given. In practice, the LOS rates is very small (its magnitude is less than 10^{-1}), thus, the rewards are set to be the absolute value of the LOS rates when the LOS rates are less than the critical value. In other cases, let the value of reward be 10. As indicated above, the reward function is designed as the following form

$$r = \begin{cases} |\dot{q}|, \text{if } \dot{q} < 0.6\text{r}ad/s \\ 10, \text{i}f \ \dot{q} \geq 0.6\text{r}ad/s. \end{cases} \tag{3.3}$$

4) The behavioral strategies: following the $\varepsilon-$ greedy strategy, the behavioral strategies of missile can be set up.

3.2. Numerical results

This subsection is devoted to list the numerical results to validate the intelligent maneuvering penetration strategies proposed. Let the initial velocity of both the missile and the interceptor are 900 m/s,

the initial positions are $(120,000\,m, -120,000\,m, 0\,m)$ and $(0\,m, 0\,m, 120,000\,m)$ respectively. Suppose strategy of the interceptor is proportional navigation method [32], the coefficient of which is $Ne = 3$, it is mentioned that the missile is not trained with respect to this strategy before launching. We can get tangential acceleration of interceptor by proportional navigation method, and the tangential acceleration of missile can be derived from the actions in QL. Set the learning rate $\alpha = 0.5$, discount rate $\gamma = 0.9$ in QL, the initial of ε in $\varepsilon-$greedy strategy is 0.6, which is decreased with the increasing of the frequency of training.

From the missile is locked by interceptor to the relative distance is $100,000\,m$ is the training stage, the relative distance between $25,000\,m$ to $100,000\,m$ is the level flight stage, the penetration stage is ended at the missile enter to the guidance blind area and calculate out the instantaneous miss distance or the relative distance is less than or equal to $300\,m$. In order to make comparison, the no maneuvering strategies and random maneuvering strategies with the same initial condition for missile are also considered. With these two strategies, dividing the combat process between missile and interceptor into two stages: from the missile is locked by interceptor to the relative distance is $25,000\,m$ is the level flight stage, after that, the missile enter the penetration stage.

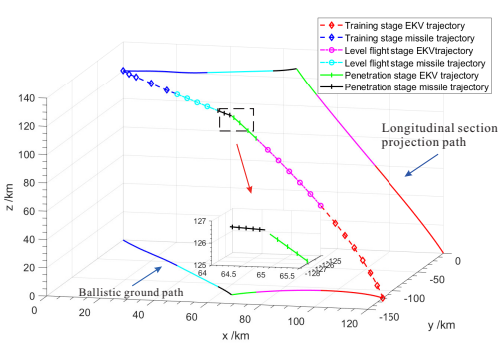
Table 1. Comparison of miss distance of three penetration strategies.

Strategy	Miss Distance (m)	horizontal/vertical overload of EKV (g)	Horizontal/vertical Los-rate (rad/s)
No maneuver	0.09	4.8/6.6	0.02/0.19
Random maneuver	0.14	5.7/4.6	0.02/0.18
QL maneuver[a]	64.04	8/8	0.29/0.49

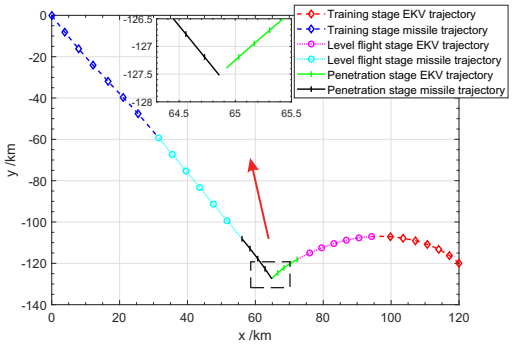
[a] The strategies derived in this paper, for convenience, we call it QL maneuver in the rest of this paper.

From Table 1, we can assert that, under the QL maneuver, the missile produces large LOS rates to induce the overload of the interceptor to exceed the ultimate value, which gives rise to that the miss distance is $64.04\,m$. The Figure 6 illustrates the flight trajectory of the missile and the interceptor EKV with QL maneuver. Figures 7 and 8 show the variations of the overloads for the missile and the interceptor. Figure 7(a) interprets that, in the case of no maneuver, the missile can be effectively intercepted by switching the overload of EKV smoothly. In the circumstance of random maneuver, although the maneuvers produced by missile induce the associated responses of the interceptor, the overload of the interceptor do not over the ultimate value, which indicates that the interceptor have enough residual capacity to intercept the missile and the instantaneous miss distance is small, see Figure 7(b). By the lateral and longitudinal overload in the case of QL maneuver painted in Figure 8, the overload of EKV can exceed the ultimate value in the case of QL maneuver and the instantaneous miss distance is large, which means that the interceptor can not intercept missile.

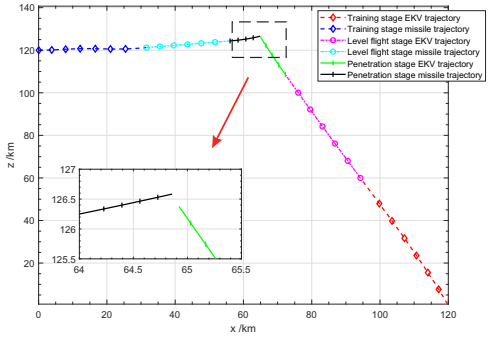
The changes of LOS rates are described by Figures 9 and 10. With the QL maneuvering strategies, we see that the LOS rates increase dramatically in the penetration stage, increasing the overload of interceptor to the ultimate value can not restrain the divergence of the LOS rates, which results in the off-target of the interceptor, see Figure 10. From Figure 9, we find that the LOS rates are almost unchanged, thus, the miss distance is small.



(a) Attack-defense confrontation process between the missile and the EKV.

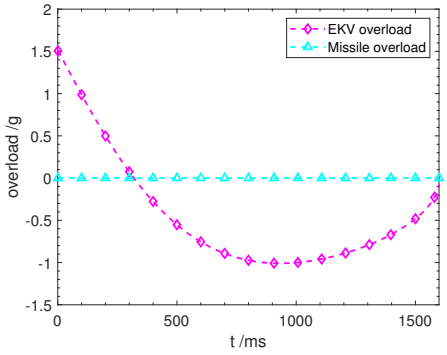


(b) Projection of trajectory of confrontation between the missile and the EKV on xoy.

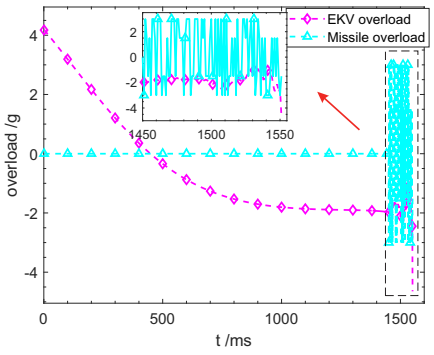


(c) Projection of trajectory of confrontation between the missile and the EKV on xoz.

Figure 6. Flight trajectory of the missile and the interceptor EKV with QL maneuver.

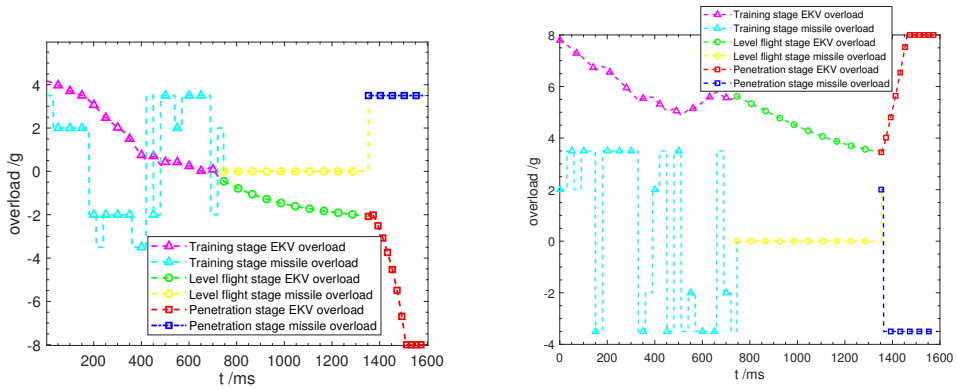


(a) The case of no maneuver.



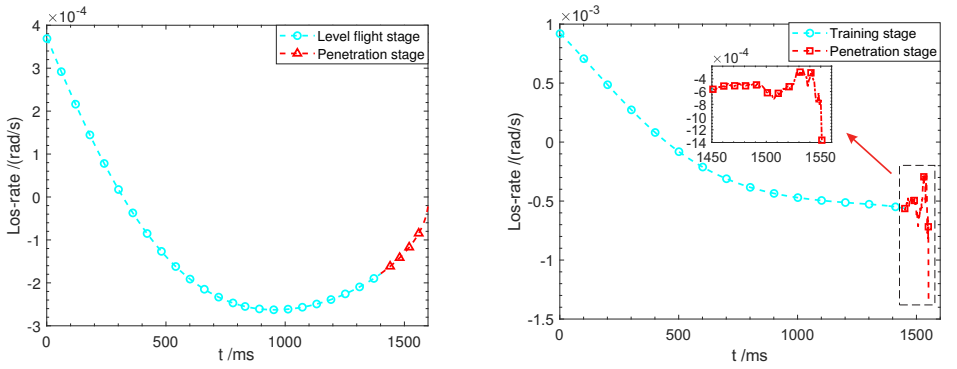
(b) The case of random maneuver.

Figure 7. Transverse overload in the cases of no maneuver and random maneuver.



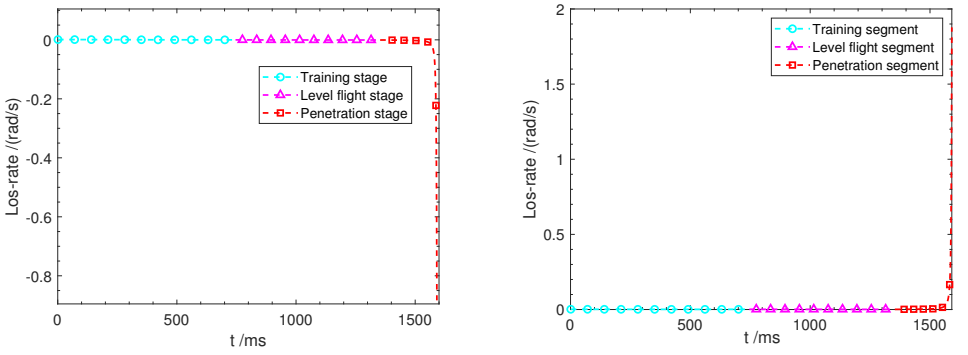
(a) Lateral overload. **(b)** longitudinal overload.

Figure 8. Transverse and longitudinal overload in the case of QL maneuver.



(a) The case of no maneuver. **(b)** The case of random maneuver.

Figure 9. Transverse LOS rates in the cases of no maneuver and random maneuver.



(a) The transverse LOS rate. **(b)** The longitudinal LOS rate.

Figure 10. Transverse and longitudinal LOS rates in the case of QL maneuver.

In order to formulate the probability of successful penetration with the three maneuvering strategies, let $d_1 = 6.5$, here d_1 is the critical value of miss distance, if the miss distance $d_0 > d_1$, it accounts for the penetration is successful, otherwise, the penetration is failed. With the same initial condition, take 1000 numerical results as a group, 10 groups of numerical simulation are accomplished for each strategy. The numerical results state that the probability of successful penetration with no maneuvering strategies or random maneuvering strategies is 0, the probability of successful penetration with QL maneuvering strategies is about 80% (see Table 2).

Table 2. The probability of successful penetration with QL maneuver strategies.

Group of numerical simulation	The probability				
1–5	81.95%	82.3%	81.5%	84.6%	82.3%
6–10	81.6%	81.7%	81.9%	80.6%	81.7%

From Figure 11, we obtain the mean of miss distances with random maneuver is about 0.2 m, the variance is small, it is the reason that the probability of successful penetration is nearly 0 when we set the critical value of miss distance is 6.5 m. When the missile with the QL maneuvering strategy, the mean of miss distance is about 60 m, although the variance of miss distance is larger than in random maneuver, combined with Table 2, we can assert that the QL maneuvering strategies can produce effective penetration with high probability. Furthermore, the Figure 12 describes that the miss distances are distributed in $[0, 200\ m]$ and concentrated around 50 m, thus, we can guarantee that the QL maneuvering strategy can leads the missile to realize successful penetration steady.

The Figure 13 illustrates the target chart of miss distances in the cases of random maneuvering strategies and QL maneuvering strategies. In the case of random maneuver, the distribution of target points are random, no obvious regularity can be found intuitively, see Figure 13(a). Because of the limit on the times of training, the most of target points are around one of the coordinate axis under the QL maneuvering strategies.

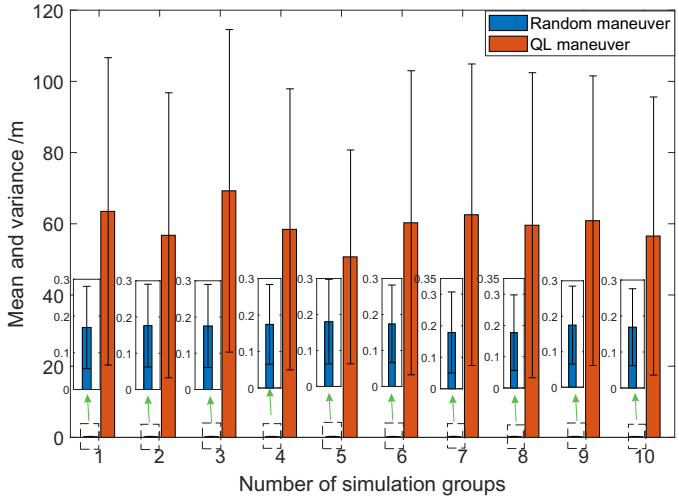


Figure 11. Mean and variance of the miss distance under random maneuver and QL maneuver.

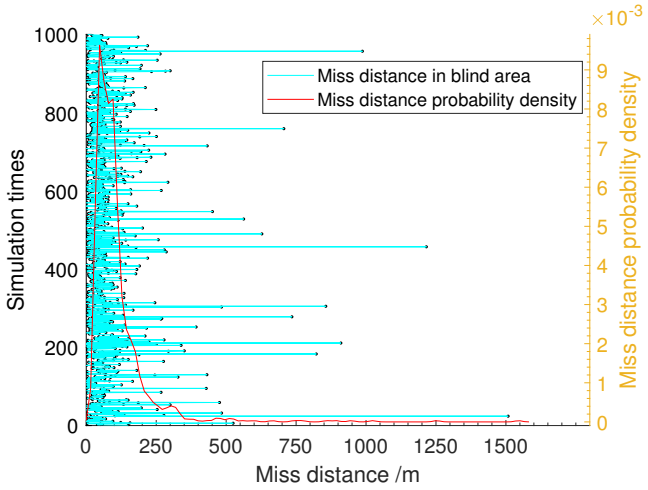
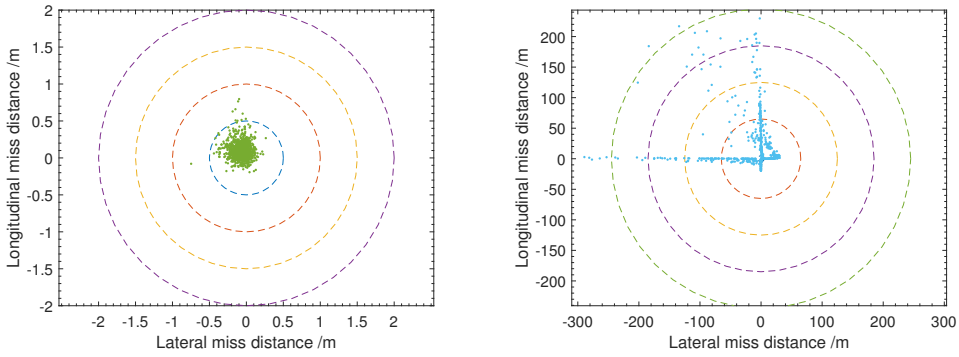


Figure 12. Miss distance and its probability density under QL maneuver.



(a) In the case of random maneuver.

(b) In the case of QL maneuver.

Figure 13. The target figure of miss distances under random maneuver and QL maneuver.

In order to illustrate the intensively effectiveness of maneuvering strategies proposed, we set proportional guidance coefficients Ne of EKV to be 2.3, 2.5, 2.7 and 3.3 respectively, under which the similar numerical investigations are considered. The related results are given in Table 3.

Table 3. The probability of successful penetration under QL maneuver strategies with respect to the different Ne.

	$Ne = 2.3$	$Ne = 2.5$	$Ne = 2.7$	$Ne = 3.2$
Probability	94.24%	91.27%	88.35%	75.25%

As indicated above, we can conclude that, in the confrontation between the missile and the interceptor, the missile trained by QL maneuver based on the information of tentative maneuver of missile and the responses of interceptor induced can autonomously select a suitable maneuvering strategy to implement penetration successfully.

4. Conclusions

When the intercepting manners are unknown in the confrontation between the missile and the interceptor, this paper proposes the intelligent maneuvering penetration strategies of the missile based on reinforcement learning, which is referred as QL maneuver. Compared with the no maneuvering and random maneuvering methods, the QL maneuver possesses the advantage of high probability of successful penetration. To be important, it is only needed the information of positions and LOS rates in the process of training the QL maneuvering strategies, furthermore, the time of training is not very long and all the training is done by the missile-borne computer when the missile are flying.

Acknowledgments

This work have been supported by National Natural Science Foundation of China (No. 12072178; No. 12002194; N0. 12202250), Project (No. ZR2020MA054; No. ZR2020QA037) supported by Shandong Provincial Natural Science Foundation.

Conflict of interest

The authors declare there is no conflicts of interest.

References

1. P. Zarchan, Proportional navigation and weaving targets, *J. Guid. Control Dyn.*, **18** (1995), 969–974. https://doi.org/10.2514/3.21492
2. J. I. Lee, C. K. Ryoo, Impact angle control law with sinusoidal evasive maneuver for survivability enhancement, *Int. J. Aeronaut. Space Sci.*, **19** (2018), 433–442. https://doi.org/10.1007/s42405-018-0042-2
3. E. Garcia, D. W. Casbeer, M. Pachter, Design and analysis of state-feedback optimal strategies for the differential game of active defense, *IEEE Trans. Autom. Control*, **64** (2019), 553–568. https://doi.org/10.1109/TAC.2018.2828088
4. J. Ding, C. L. Li, G. S. Zhu, Two-person zero-sum matrix games on credibility space, in *2011 Eighth International Conference on Fuzzy Systems and Knowledge Discovery (FSKD)*, (2011), 912–916. https://doi.org/10.1109/FSKD.2011.6019721
5. T. Yang, L. Geng, M. Duan, K. Zhang, X. Liu, Research on the evasive strategy of missile based on the theory of differential game, in *2015 34th Chinese Control Conference (CCC)*, (2015), 5182–5187. https://doi.org/10.1109/ChiCC.2015.7260447
6. M. Flasiński, Symbolic artificial intelligence, in *Introduction to Artificial Intelligence*, (2016), 15–22. https://doi.org/10.1007/978-3-319-40022-8_2
7. L. P. Kaelbling, M. L. Littman, A. W. Moore, Reinforcement learning: a survey, *J. Artif. Intell. Res.*, **4** (1996), 237–285. https://doi.org/10.1613/jair.301
8. R. S. Sutton, A. G. Barto, *Reinforcement Learning: An Introduction*, MIT press, 2018.

9. S. Lei, Y. Lei, Z. Zhu, Research on missile intelligent penetration based on deep reinforcement learning, *J. Phys. Conf. Ser.*, **1616** (2020), 012107. https://doi.org/10.1088/1742-6596/1616/1/012107
10. X. Wang, Y. Cai, Y. Fang, Y. Deng, Intercept strategy for maneuvering target based on deep reinforcement learning, in *2021 40th Chinese Control Conference (CCC)*, (2021), 3547–3552. https://doi.org/10.23919/CCC52363.2021.9549458
11. L. Jiang, Y. Nan, Z. H. Li, Realizing midcourse penetration with deep reinforcement learning, *IEEE Access*, **9** (2021), 89812–89822. https://doi.org/10.1109/ACCESS.2021.3091605
12. K. Arulkumaran, M. P. Deisenroth, M. Brundage, A. A. Bharath, Deep reinforcement learning: a brief survey, *IEEE Signal Process Mag.*, **34** (2017), 26–38. https://doi.org/10.1109/MSP.2017.2743240
13. S. Arora, P. Doshi, A survey of inverse reinforcement learning: challenges, methods and progress, *Artif. Intell.*, **297** (2021), 103500. https://doi.org/10.1016/j.artint.2021.103500
14. S. Bradtke, M. Duff, Reinforcement learning methods for continuous-time markov decision problems, in *Advances in Neural Information Processing Systems*, **7** (1994). Available from: https://proceedings.neurips.cc/paper/1994/file/07871915a8107172b3b5dc15a6574ad3-Paper.pdf.
15. B. Abdulhai, L. Kattan, Reinforcement learning: introduction to theory and potential for transport applications, *Can. J. Civ. Eng.*, **30** (2003), 981–991. https://doi.org/10.1139/l03-014
16. F. L. Lewis, D. Vrabie, K. G. Vamvoudakis, Reinforcement learning and feedback control: using natural decision methods to design optimal adaptive controllers, *IEEE Control Syst. Mag.*, **32** (2012), 76–105. https://doi.org/10.1109/MCS.2012.2214134
17. J. Kober, J. A. Bagnell, J. Peters, Reinforcement learning in robotics: a survey, *Int. J. Rob. Res.*, **32** (2013), 1238–1274. https://doi.org/10.1177/0278364913495721
18. A. Mosavi, Y. Faghan, P. Ghamisi, P. Duan, S. F. Ardabili, E. Salwana, et al., Comprehensive review of deep reinforcement learning methods and applications in economics, *Mathematics*, **8** (2020), 1640. https://doi.org/10.3390/math8101640
19. V. Shalumov, Cooperative online guide-launch-guide policy in a target-missile-defender engagement using deep reinforcement learning, *Aerosp. Sci. Technol.*, **104** (2020), 105996. https://doi.org/10.1016/j.ast.2020.105996
20. S. He, H. S. Shin, A. Tsourdos, Computational missile guidance: a deep reinforcement learning approach, *J. Aerosp. Inf. Syst.*, **18** (2021), 571–582. https://doi.org/10.2514/1.I010970
21. B. Gaudet, R. Furfaro, Missile homing-phase guidance law design using reinforcement learning, in *AIAA Guidance, Navigation, and Control Conference*, (2012), 4470. https://doi.org/10.2514/6.2012-4470
22. D. Hong, S. Park, Avoiding obstacles via missile real-time inference by reinforcement learning, *Appl. Sci.*, **12** (2022), 4142. https://doi.org/10.3390/app12094142
23. B. Gaudet, R. Furfaro, R. Linares, Reinforcement learning for angle-only intercept guidance of maneuvering targets, *Aerosp. Sci. Technol.*, **99** (2020), 105746. https://doi.org/10.1016/j.ast.2020.105746

24. W. Li, Y. Zhu, D. Zhao, Missile guidance with assisted deep reinforcement learning for head-on interception of maneuvering target, *Complex Intell. Syst.*, **8** (2022), 1205–1216. https://doi.org/10.1007/s40747-021-00577-6

25. X. Qiu, C. Gao, W. Jing, Maneuvering penetration strategies of ballistic missiles based on deep reinforcement learning, *Proc. Inst. Mech. Eng., Part G: J. Aerosp. Eng.*, 2022 (2022), 09544100221088361.

26. B. Gaudet, R. Furfaro, Integrated and adaptive guidance and control for endoatmospheric missiles via reinforcement learning, preprint, arXiv:2109.03880.

27. A. Candeli, G. de Tommasi, D. G. Lui, A. Mele, S. Santini, G. Tartaglione, A deep deterministic policy gradient learning approach to missile autopilot design, *IEEE Access*, **10** (2022), 19685–19696. https://doi.org/10.1109/ACCESS.2022.3150926

28. C. Yang, J. Wu, G. Liu, Y. Zhang, Ballistic missile maneuver penetration based on reinforcement learning, in *2018 IEEE CSAA Guidance, Navigation and Control Conference (CGNCC)*, (2018), 1–5. https://doi.org/10.1109/GNCC42960.2018.9018872

29. R. T. Yanushevsky, *Modern Missile Guidance*, CRC Press, 2018. https://doi.org/10.1201/9781351202954

30. A. Wong, F. Nitzsche, M. Khalid, Formulation of reduced-order models for blade-vortex interactions using modified volterra kernels, 2004. Available from: https://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.632.1295&rep=rep1&type=pdf.

31. S. H. Hollingdale, The mathematics of collision avoidance in two dimensions, *J. Navig.*, **14** (1961), 243–261. https://doi.org/10.1017/S037346330002960X

32. Y. Chen, J. Wang, C. Wang, J. Shan, M. Xin, A modified cooperative proportional navigation guidance law, *J. Franklin Inst.*, **356** (2019), 5692–5705. https://doi.org/10.1016/j.jfranklin.2019.04.013

AIMS Press