doi:10.11887/j.cn.202402010 http://journal.nudt.edu.cn

导弹突防后弹道机动调整策略强化学习

樊博璇[1,2],陈桂明[1*],韩 磊[2],李 冰[2]
(1. 火箭军工程大学 作战保障学院,陕西 西安 710025;
2. 火箭军装备部驻西安地区第一军事代表室,陕西 西安 710025)

摘 要:针对弹道导弹中段突防后飞行弹道与标准弹道产生较大偏离的弹道机动调整问题,建立了机动调整时机策略最优化模型。设计了机动调整逆序 Q 学习算法,采用 Tile coding 逼近器编码状态特征空间,并对其进行线性逼近。构建了 Q 学习算法与蒙特卡罗方法相结合的逆序更新策略机制,以对导弹机动调整最优时机进行训练。仿真测试分析结果表明,在给定场景参数下,通过 10 000 代强化学习算法训练得到的策略能够可靠地使用最少机动次数控制导弹突防后飞行弹道的调整决策,验证了方法的有效性。

关键词:弹道导弹;中段突防;强化学习;Q 学习;控制决策
中图分类号:TJ765.3 **文献标志码**:A **文章编号**:1001-2486(2024)02-094-10

Reinforcement learning of ballistic maneuver adjustment strategy after missile penetration

FAN Boxuan[1,2], *CHEN Guiming*[1*], *HAN Lei*[2], *LI Bing*[2]
(1. College of Operational Support, Rocket Force University of Engineering, Xi'an 710025, China;
2. The First Military Representative Office of the Rocket Force Equipment Department in Xi'an Region, Xi'an 710025, China)

Abstract: In order to solve the problem of trajectory maneuver adjustment caused by large deviation of flight trajectory after midcourse penetration of ballistic missile, an optimization model of maneuver adjustment timing strategy was established. A reverse sequence Q learning algorithm for maneuver adjustment was designed, and a Tile coding approximator encoding was used to encode the state characteristics space, and the space was linearly approximated. A reverse-order update strategy mechanism combining Q learning algorithm and Monte Carlo method was constructed, the optimal timing of missile maneuvering adjustment was trained. The simulation results show that the strategy obtained by training 10 000 generations of reinforcement learning algorithm can reliably control the adjustment decision of flight trajectory after missile penetration with the minimum maneuver times under given scenario parameters, which verifies the effectiveness of the method.

Keywords: ballistic missile; midcourse penetration; reinforcement learning; Q learning; control decision

预先规划弹道已无法适应未来对抗需要,提升导弹的智能化感知和机动能力,使其实现自主规避突防与弹道调整瞄准是未来发展趋势。近年来,弹道导弹主动规避式机动突防取得了一些研究成果,逐渐具备技术可行性,但突防后会使导弹与标准弹道产生较大偏离[1-3]。为保证导弹在机动突防规避后仍能准确命中目标,导弹应具备适时向适当方向采取机动调整的能力,以保证其在自身不被摧毁的同时准确完成打击任务。因此,提高弹道导弹中段突防后飞行弹道的机动调整能力,对我军实现精确打击、提高威慑能力具有重要意义。

强化学习(reinforcement learning,RL)是人工智能领域新的研究热点,因其具备较强的智能决策能力,在导弹机动规避的控制策略研究中已经取得了实质性的突破,将被应用于导弹武器装备智能机动突防[4]。文献[5]在基于 Markov 决策过程的导弹中段突防控制模型的基础上构建了深度神经网络,在虚拟战场环境中通过强化学习技术不断迭代出最优的侧推发动机点火指令。文献[6]针对载机面对来袭导弹自主规避问题,采取一种改进的基于确定策略梯度优化的深度强化学习方法进行训练学习,得到载机规避来袭导弹的最佳机动策略。文献[7]采用比例-积分-微分(proportion integration differentiation,PID)控制的设计思想,在深度确定性策略梯度(deep

收稿日期:2022-01-15
基金项目:国家自然科学基金资助项目(71601180)
第一作者:樊博璇(1987—),女,吉林扶余人,工程师,博士,E-mail:1092442646@qq.com
***通信作者**:陈桂明(1966—),男,江苏南京人,教授,博士,博士生导师,E-mail:792757066@qq.com

deterministic policy gradient, DDPG)算法基础上,通过训练神经网络输出控制指令,成功地完成追捕任务。文献[8]针对巡飞弹动态突防控制决策问题,提出基于深度强化学习(deep reinforcement learning, DRL)的巡飞弹突防控制决策模型及其求解方法,仿真结果证明了该方法能够使巡飞弹在动态对抗环境中实现自主突防。文献[9]使用 DDPG 算法探究巡航导弹智能突防问题。文献[10]利用了强化学习算法中的决斗双深度 Q 网络(dueling double deep Q network, D3QN)算法解决弹道导弹中段突防问题,该方法训练所得的深度神经网络控制器很好地规避了拦截器的攻击。文献[11]基于连续控制强化学习算法,提出了一个将确定性策略梯度算法与 Actor-Critic 框架相结合的任务无关模型,使用相同参数解决了导弹打击任务不同的连续控制问题。还有很多学者对导弹突防过程中的规避效能评估以及制导律优化等问题进行了研究[12-18]。此外,通过广泛文献调研,发现利用强化学习技术来解决飞机姿态控制及弹道优化等复杂非线性系统控制问题具有显著的效果[19-21]。但目前尚未见到运用强化学习方法对弹道导弹机动突防后的弹道机动调整决策进行训练方面的研究成果。

对于弹道导弹中段突防,机动突防后的弹道调整是一个重难点问题。针对该问题,本文引入了 Tile coding 逼近器,在最优机动调整速度增量计算的基础上,设计了状态空间、动作空间和奖励函数,建立了机动调整时机最优策略模型,构建了一种逆序 Q 学习算法,对导弹的机动调整时机策略进行训练学习,得到最优机动调整时机,给出了算法实现过程,并进行了仿真验证和分析。

1　弹道机动调整时机的马尔可夫决策模型

1.1　问题描述

针对弹道导弹中段自主机动策略,机动调整与机动突防的重难点存在不同。机动突防成功的关键在于及时感知来袭大气层外拦截器(exoatmospheric kill vehicle, EKV)的相对位置和速度,使导弹能够在正确的时间向正确的方向机动,对控制动力精确性要求不高[22],可采用脉冲姿/轨控发动机实现机动突防。但机动调整时的机动控制精度可谓差之毫厘谬以千里,对机动控制能力提出了更高要求。此外,导弹突防成功后可能面临二次拦截的威胁,不应过早耗尽机动能力。因此,为保证弹道导弹具有一定的再入可控机动性,在中段尽量将弹道调整到可控范围的基础上,采用多次点火的脉冲姿/轨控发动机实现机动突防与突防后的机动调整[23]。在此基础上,本文旨在寻求一种弹道导弹中段机动突防后的弹道机动调整策略,以实现导弹对目标的精确打击。基于多脉冲姿/轨控发动机的机动设定和弹载雷达感知来袭 EKV 的基本设定,提出机动突防后的弹道智能机动调整策略。

面对随机的初始攻防双方飞行状态,在设定的约束条件下,任意位置到指定目标位置所需的初速度通过求解 Lambert 方程得到,某时刻导弹的位置速度状态为 $\boldsymbol{d}_0,\boldsymbol{v}_0$,目标点位置为 $\boldsymbol{d}_{\mathrm{T}}$,在弹道轨迹中的某一位置,控制导弹的多脉冲姿/轨控发动机通过对导弹施加所需的速度增量,即可改变弹道轨迹使其飞向目标,弹道机动调整策略问题可表达为如何使用最少的机动次数使导弹再入时的预期落点偏差能够控制在较小范围内的控制策略优化问题。机动越早,相同冲量机动落点可调整范围越大,但机动误差导致的落点误差也越大。基于此,本文主要研究弹道的最优机动调整时机问题。

1.2　弹道机动调整时机策略最优化模型

1.2.1　优化目标

策略 $\boldsymbol{\pi}$ 的优化目标是对于任意导弹成功机动规避 EKV 来袭后的状态 $\boldsymbol{u}$,都能通过最小的机动次数 N(对于多脉冲姿/轨控发动机,脉冲机动次数是有限确定的)将预期落点调整到目标附近,可表示为

$$\min E[N|\boldsymbol{\pi}],\quad \forall\{\boldsymbol{u},\boldsymbol{d}_{\mathrm{T}}\} \tag{1}$$

式中,$E[N|\boldsymbol{\pi}]$ 表示策略 $\boldsymbol{\pi}$ 下机动次数 N 的期望。

1.2.2　约束条件

1)落点偏差小于导弹落点偏差距离的容许量 X,即

$$d<X \tag{2}$$

2)任意相邻两次机动之间多脉冲姿/轨控发动机的调整准备时间充足,即

$$t_{\mathrm{a}}^{i+1}-t_{\mathrm{a}}^{i}\geqslant t_{\min} \tag{3}$$

式中,t_{a}^{i} 表示第 i 次机动的时刻,$t_{\min}$表示两次机动之间的最小时间间隔。

1.2.3　决策变量

决定导弹下一次机动的时刻和速度增量可表示为

$$\boldsymbol{\pi}:t_{\mathrm{a}},\boldsymbol{p}=\boldsymbol{\pi}(\boldsymbol{d},\boldsymbol{v},\boldsymbol{d}_{\mathrm{T}}) \tag{4}$$

式中,t_{a} 表示下一次的机动时刻,$\boldsymbol{p}$ 表示下一次机动的速度增量,决策变量策略 $\boldsymbol{\pi}$ 为由位置速度状

态到下一次机动的时刻和速度增量的函数。

多脉冲姿/轨控发动机的主要功能是机动突防，其控制精度不高，需要在合适时机进行多次机动调整才能有效控制落点误差。假设实际机动加速度大小、机动起始时刻、机动方向与设定值存在一定随机误差，且在式(5)～(7)所示范围内均匀随机分布。若指定的机动速度增量矢量为 $\boldsymbol{p}$，推力起始时刻为 t_i，实际速度增量为 $\breve{\boldsymbol{p}}$，$\breve{\boldsymbol{p}}$ 与 $\boldsymbol{p}$ 间的角度偏差为 θ，实际推力起始时刻为 $\breve{t}_i$，则应满足

$$0.9|\boldsymbol{p}| < |\breve{\boldsymbol{p}}| < 1.1|\boldsymbol{p}| \tag{5}$$

$$\theta < 0.002 \text{ rad} \tag{6}$$

$$t_i - 50 \text{ ms} < \breve{t}_i < t_i + 50 \text{ ms} \tag{7}$$

1.3 状态空间设计

状态空间降维，使用有限离散参数构造价值函数，将9维状态空间用2维状态特征空间近似表示，并采用 Tile coding 编码对状态特征空间做线性逼近。

状态变量包括导弹的位置、速度，目标的位置。状态变量可以使用导弹落地前剩余飞行时间 T_{remain} 和预期落点与目标位置的偏差近似表示。通过分析可知，导弹落地前剩余飞行时间 T_{remain} 是决定机动后落点偏差范围的主要因素。机动误差造成的机动落点不确定性也主要与导弹剩余飞行时间和当前轨道落点偏差有关，因此，选择剩余飞行时间 T_{remain} 和加权落点偏差估计 $\tilde{d}$ 作为用于决策的状态特征，即状态空间

$$s = \{T_{\text{remain}}, \tilde{d}\} \tag{8}$$

1.4 动作空间设计

机动调整时机决策 Agent 在每个决策时刻只需决定是否此刻做机动调整。如果决策为是，则按照通过求解 Lambert 方程计算得到的机动调整速度增量做机动；如果决策为否，导弹在下次决策前不做机动。机动调整时机决策的动作空间为

$$\mathcal{A}^{\text{tac}} = \{0, 1\} \tag{9}$$

式中，0 表示此刻不做机动调整，1 表示立即做机动调整。考虑两次机动调整时间需要的必要导弹姿态控制时长，而机动调整时机不需要像机动突防时机那样精确，因此将决策步长设置为 30 s，保证弹道能够及时得到调整。

2 逆序 Q 学习算法

Q 学习算法是一种基于时序差分(temporal-difference, TD)的强化学习算法，由当前策略 π 对各状态动作进行评价以修正价值函数 $Q^{\pi}(s,a)$，然后反过来借由 $Q^{\pi}(s,a)$ 的变化改进策略 π，依此反复迭代对策略进行优化[24]，以此逼近目标函数。该算法的目标策略为 ε - 贪婪策略，即在训练中的每个时间步，多数情况选择预计能获得最大回报的行为，但以小概率 ε 完全随机选择一个动作以确定动作 a_t，即在状态 s_t 下选择动作 a_t 的概率为

$$p(a_t|s_t) = \begin{cases} 1 - \varepsilon + \dfrac{\varepsilon}{|\mathcal{A}(s_t)|}, & a_t = \underset{a}{\operatorname{argmax}} Q(s_t, a) \\ \dfrac{\varepsilon}{|\mathcal{A}(s_t)|}, & a_t \neq \underset{a}{\operatorname{argmax}} Q(s_t, a) \end{cases} \tag{10}$$

其中，$\mathcal{A}(s_t)$ 表示在状态 s_t 下的可选动作空间，$|\mathcal{A}(s_t)|$ 表示状态 s_t 下的可选动作空间大小。

执行动作后的下一时刻，环境会反馈状态 s_{t+1} 和奖励 r_{t+1}，$\langle s_t, a_t, s_{t+1}, r_{t+1}\rangle$ 即为一组训练样本。定义最优策略下状态 s_t 的价值为

$$V^*(s_t) = \max_{a_t}[r_{t+1} + \gamma V^*(s_{t+1})] \tag{11}$$

式中，$\gamma \in [0,1]$ 为折扣因子，值越大越注重奖励的长期累积。改进策略的 Q 值函数更新公式为

$$Q(s_t, a_t) \leftarrow Q(s_t, a_t) + \alpha[r_{t+1} + \gamma \max_{a} Q(s_{t+1}, a) - Q(s_t, a_t)] \tag{12}$$

式中，α 为学习率，使用增量执行方法进行计算。

在机动调整问题中，假设从导弹机动突防结束时刻到再入大气层之前的每个飞行过程为一个片段(episode)。

如果直接使用 TD 算法，算法的 TD 误差从终止态向前传递的速度比较慢，因此本文引入 Tile coding 编码方法对状态特征空间做线性逼近，并将 Q 学习算法与蒙特卡罗算法相结合，采用片段结束后使用 Q 学习算法的逆序更新策略机制，实现 TD 误差的快速传递。

导弹进行智能机动调整时机决策是 Agent 与环境交互的过程中不断进行策略迭代改进，Agent 在训练过程中学习每个动作对的价值 $Q^{\pi}(s,a)$，在决策时基于对价值的估计选择动作。

2.1 奖励函数设计

奖励通常根据问题特点人为设定。奖励是评价策略的唯一标准，必须保证更好的策略能获得更高的奖励，最优策略获得最高的奖励，才能使策略可靠收敛。

本文将折扣因子设置为 $\gamma = 1$，即无折扣。这是因为首要的控制目的是将落点偏差控制在一定范围内，落点偏差要在一次飞行结束时才能获得。而折扣奖励会使距离片段结束时间较长时，Agent 由于不重视落点偏差奖励而不能及时进行机动调整。

奖励函数设置为

$$r=\begin{cases}0, & a=0 \text{ 且 } t\neq t_{\text{end}}\\ -1, & a=1 \text{ 且 } t\neq t_{\text{end}}\\ \min(-\ln d+\ln 100-1,0), & a=0 \text{ 且 } t=t_{\text{end}}\\ \min(-\ln d+\ln 100-1,0)-1, & a=1 \text{ 且 } t=t_{\text{end}}\end{cases} \tag{13}$$

式中,t_{end}表示片段结束的时刻,由此时导弹状态计算落点偏差 d。

奖励函数由两部分构成。一部分是片段结束时根据落点偏差的大小给出奖励 $\min(-\ln d+\ln 100-1,0)$。为确保将导弹的落点偏差调整到允许范围内(本文取 100 m),当落点偏差 $d<100$ m 时,落点偏差奖励大于 -1。这样设计使 Agent 在预期落点偏差 $d<100$ m 时就不再机动,同时又鼓励调整到落点偏差更小的状态。另一部分是机动时给出的即时奖励 -1,这是为了调整机动次数尽可能少。

这两部分的奖励的具体形式须综合考虑,以使两方面目的同时实现。根据前期研究分析发现,在合适的时机进行机动调整,一般能将调整后的偏差降低至原偏差的 1/10 以内,因此设定落点偏差的对数形式作为落点偏差奖励函数。

2.2　策略函数

机动调整时机策略在每个决策时刻只有机动和不机动两个可选动作,为此刻不做动作与做动作建立独立的 Q 值近似函数,分别记为 Q^0,Q^1。函数的输入为状态特征值,输出为 Q 值,即此时刻选择该动作,随后使用 ε - 贪心策略的回报估计。两个 Q 函数都使用相同的 Tile coding 编码对状态特征空间做线性逼近,在训练过程中使用交互中获得的经验,通过调整 Tile coding 编码对应的权值不断改进策略。

对于编码空间内的任意位置,每个 Tiling 中有且仅有 1 个 Tile 被激活,图 1 为 3×3 的 Tiling 组成的 Tile coding 逼近器,所示的点编码为 Index1 = 3,Index2 = 5,Index3 = 6。

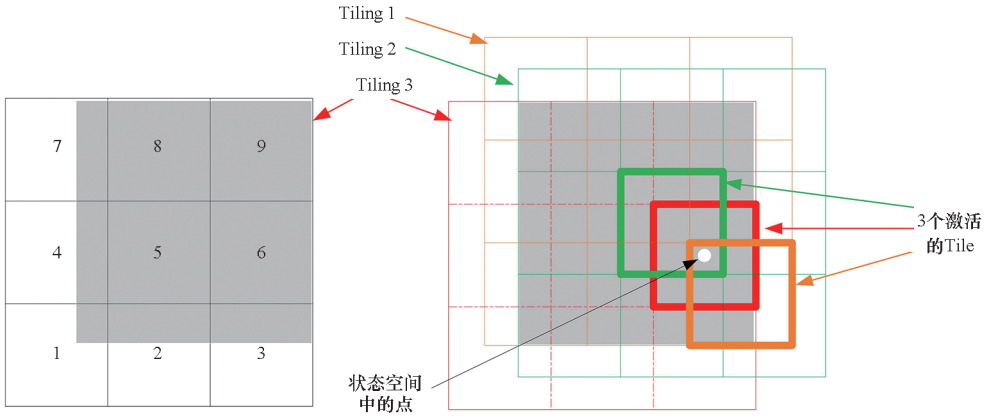


图 1　Tile coding 编码示意图

Fig. 1　Tile coding schematic diagram

其中,每个 Tile 对应一个参数值 w,价值函数 $Q(s)$ 简单地定义为对应激活 Tile 的参数值求和形式。对于使用 N 个分辨率为 $m_1\times m_2$ 的 Tiling 组成的 Tile coding 逼近器构造的价值函数,一共包含 $N\times m_1\times m_2$ 个参数,价值函数为

$$Q(s)=w_{\text{Index1}(s)}+w_{\text{Index2}(s)}+\cdots+w_{\text{Index}n(s)} \tag{14}$$

参数的更新公式为

$$w_{\text{Index}i(s)}\leftarrow w_{\text{Index}i(s)}+\frac{\alpha}{n}[G(s)-Q(s)] \tag{15}$$

二维 Tile coding 编码的两个维度分别设置为加权落点偏差的对数 lg$d$ 和剩余飞行时间 T_{remain}。Tiling 数量设置为 $N=8$,每个 Tiling 的 lg$d$ 维度的分辨率取 $m_1=5$,T_{remain} 维度的分辨率取 $m_2=10$。因为弹道导弹中段调整可能的时间范围不会超出落地前 100 s 到 1 700 s 的范围,这样设置使实际空间中 T_{remain} 维度的最小分辨率为 (1 700 s − 100 s)/(10 × 8) = 20 s,小于决策时间步长。

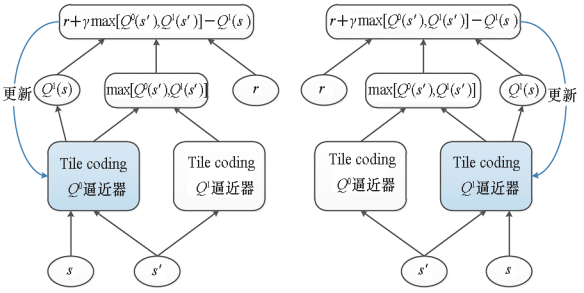
2.3　策略更新

片段中的每个决策时刻,均以当前状态特征值作为 Q 值函数的输入,获得机动和不机动两个动作的 Q 值。根据 Q 值,依照 ε - 贪心策略选择动作。记录每一步决策的状态、动作、奖励和结果状态,作为一个学习样本,记为 s,a,r,s'。按顺序记录一次仿真生成的所有样本,一次仿真结束时可获得样本序列$\{s,a,r,s'\}$。逆序使用这些样本更新策略参数,$a=0$ 的样本用于更新 Q^0 函数的参数,$a=1$ 的样本用于更新 Q^1 函数的参数。每个片段结束时,对整个片段中的记录$\{s,a,r,s'\}$,采用逆序 Q 学习算法更新两个 Q 值函数的参数,

更新方程为

$$\begin{cases} Q^0(s)\leftarrow Q^0(s)+\alpha\{\max[Q^0(s'),Q^1(s')]+ \\ r-Q^0(s)\},\ a=0 \\ Q^1(s)\leftarrow Q^1(s)+\alpha\{\max[Q^0(s'),Q^1(s')]+ \\ r-Q^1(s)\},\ a=1 \end{cases} \tag{16}$$

其中，$\gamma=1$，策略更新算法流程图如图 2 所示。



（a）$a=0$　　（b）$a=1$

图 2　机动调整策略强化学习流程图

Fig. 2　Maneuver adjustment strategy RL flow chart

2.4　Tile coding 强化学习的增量执行方法

使用函数逼近形式的价值函数，通常采用固定学习率或随机梯度下降（stochastic gradient descent，SGD）、Adam 优化算法等随机梯度算法更新参数。但状态空间中的部分区域可能在学习中几乎不会经历到，导致这些方法在基于 Tile coding 的强化学习中的效果并不理想，Tile coding 对应这些区域的参数初值对于学习过程影响很大。增量执行技术根据所记录的每一个状态的经历次数调整学习率，能够快速消除设置参数的初值影响，但对于函数逼近形式的价值函数，状态的经历次数难以确定。Tile coding 构造的价值函数是各被激活的 Tile 的简单求和形式，根据这一特点，本文设计了一种对各个 Tile 分别记录经历次数的方法，使 Q 值函数中的各个 Tile 每次更新为历次回报对应分量的均值，构造了针对 Tile coding 的增量执行算法。

每当更新状态价值时，对激活的每个 Tile 分别更新，同时分别维护各个 Tile 的经历次数，当第 i 个 Tile 被激活，对其参数的更新公式为

$$w_i^0\leftarrow w_i^0+\frac{1}{n_i^0 N}\{\gamma\max[Q^0(s'),Q^1(s')]+r-Q^0(s)\} \tag{17}$$

$$w_i^1\leftarrow w_i^1+\frac{1}{n_i^1 N}\{\gamma\max[Q^0(s'),Q^1(s')]+r-Q^1(s)\} \tag{18}$$

其中，n_i^0 和 n_i^1 分别为 Q^0 和 Q^1 逼近函数第 i 个 Tile 对应参数值 w_i^0 和 w_i^1 更新的次数。

在 Q 学习算法中，当学习初期更新某些状态价值时，更新运算所需的 $Q^0(s')$、$Q^1(s')$ 对应的参数值没有全部被更新过。这种情况下 n_i^0 或 n_i^1 在更新后重新置 0，因为这次更新是不可信的。

这对于 Tile coding 形式的逼近器是必要的，因为不同于表格式方法能保证每个参数值只对应一个状态，每个 Tile 参数值同时对应多个（不同编码）状态，而每个状态被经历到的概率不同，若使用严格的增量执行方法会导致被经历概率更大的状态对 Tile 参数值的影响更大。因此，当参数值更新次数 n_i^0 或 n_i^1 累积到 50 次时则不再继续累加，以保证至少 0.02 的学习率，使参数值能够保持有效更新。保证更新时至少 0.02 的学习率在一定程度上能够克服这一问题。

2.5　经验复用技术

导弹按照 2.1 节的奖励设计进行机动调整动作，除了直接到达终止状态的情况，都将固定获得 $r=-1$ 的奖励。因此，对于结果状态为非终止态的情况，只需记录每个机动步的当前状态 s 和结果状态 s' 即可。本文所使用的 Tile coding 编码形式，实际上一共对应着 $[(m_1-1)N+1]\times[(m_2-1)N+1]=33\times73=2\ 409$ 个不同编码的状态。使用 $2\ 409\times2\ 409$ 的二维矩阵记录机动状态转移发生次数显然过于笨拙。借鉴 2.4 节的思路，分别记录不同编码的各状态到各 Tile 之间的状态转移关系，这样只需要 $[(m_1-1)\times N+1]\times(m_1\times m_2\times N)=2\ 409\times400$ 个的存储矩阵，记为 $\boldsymbol{T}$。$T_{i,j}$ 记录从第 i 个状态机动后转移到第 j 个 Tile 被激活的状态的次数。虽然依然很难收集到足以准确刻画状态转移概率分布的状态转移记录，但是对于状态之间本身的空间分布关系，机动状态转移记录矩阵 $\boldsymbol{T}$ 是能够在一定程度上反映状态转移概率分布规律的。

对于结果状态为终止态的情况，则采用 2.4 节增量执行的方法，维护激活第 i 个状态机动后到达终止态的发生次数 n_i^{F} 和对应的平均奖励 $\bar{r}_i$ 数据。

通过上述记录可以完整刻画估计各状态机动价值所需的全部经历信息。机动动作价值函数的参数估计可表示为

$$w_i^1\leftarrow w_i^1+\frac{\sum_j[T_{i,j}\max(w_j^0,w_j^1)]+\dfrac{n_i^{\mathrm{F}}\bar{r}_i}{N}}{\dfrac{\sum_j T_{i,j}}{N}+n_i^{\mathrm{F}}} \tag{19}$$

其中,w_i^1 指代第 i 个状态激活的各 Tile 对应的 Q^1 逼近函数权值。

将经验复用方法与增量执行方法相结合,按照式(17)更新 w_i^0,按照式(19)更新 w_i^1。在更新 w_i^1 时,如果更新运算所需的参数没有全部被更新过,n_i^1 在更新后置 0,否则置 1。

3　仿真实验

3.1　场景参数设置

选择一条初始弹道作为示例,将导弹初始位置速度设定为 $\boldsymbol{d}_0$,$\boldsymbol{v}_0$,地球惯性坐标系下,$\boldsymbol{d}_0$ = (−1 121 544 m　4 188 439 m　5 794 074 m),$\boldsymbol{v}_0$ = (1 211.10 m/s　−4 197.37 m/s　−4 197.37 m/s),目标位置设定为 $\boldsymbol{d}_T$ = (421 035 m　−4 706 216 m　4 271 385 m),初始状态的剩余飞行时间估计为 T_{remain} = 1 492 s。

弹道导弹中段反导的可能拦截点不包括其整个弹道区间,EKV 不会过早或过晚来袭,因此,导弹初始状态为已扣除不可能遭遇 EKV 的前半段飞行弹道的剩余弹道范围,EKV 的可能来袭时间均匀分布于(0 s,1 492 s) 区间。此外,一个片段的终止条件为预期落点偏差小于 100 m,或者剩余飞行时间小于 100 s。在这样的设定下,每个片段的决策步数最多为 50 个。

3.2　仿真环境设计

下一时刻的状态 s' 和获得的奖励 r 只与当前状态 s 和当前动作 a 有关。即对于指定的 s' 和 r,在给定状态 s 和动作 a 下发生的概率是确定的,均由仿真环境实现,可表示为 $p(s',r|s,a)$。

仿真环境参数设置为额定机动加速度 a_c = 100 m/s^2,推力持续时间为 1 s,推力加速度大小的不确定性为 ε_a = 0.1 m/s^2,推力加速度方向的不确定性为 ε_p = 0.002 rad,推力开始时刻的不确定性为 ε_t = 0.1 s。

3.3　仿真结果分析

强化学习的行为策略采用 ε − 贪心策略,参数取 ε = 0.9。为确保导弹不会过多机动,设置机动调整次数上限为 10 次。使用增量执行 Tile coding 逆序 Q 学习算法(未使用经验复用技术)运行 10 000 个片段。

学习过程中回报(即奖励和)的变化如图 3 所示。可以看到,在强化学习 1 000 个片段时就已经获得性能较好的策略,之后对策略的微调过程中,策略性能相对保持稳定。

同样地,训练过程中更新 Q^0、Q^1 参数值时的 TD 误差也逐渐稳定(因为过程的随机性不可能完全稳定),TD 误差随学习片段数的变化如图 4 所示。

学习开始前以及学习至第 50、200、1 000、10 000个片段时导弹的机动调整策略和 Q 函数分别如图 5 ~ 9 所示,其中子图(a)为策略的可视化表达,二维平面表示状态空间,白色部分对应不做机动的策略,灰色部分对应做机动。使用学习得到的策略分别进行 5 次随机测试,测试结果用不同颜色的点实线表现在策略图中。策略是通过比较 Q^0、Q^1 两个价值函数的取值获得的,这两个函数分别展示在图 5 ~ 9 的子图(b) ~ (c)中,下方的等高线分别对应 Q 值为 −1, −2, −3,…,即预计还需进行 1 次、2 次、3 次机动才能将偏差调整到小于 100 m 的范围内。

算法中的训练代数有限且机动效果的随机性较强,而且状态空间中的部分区域在仿真训练中几乎没有遇到,导致策略的分布并不规整,但是已能确保机动调整策略具备较优性能。使用不同颜色的点实线展现在策略图中的 5 组随机测试算例,都在 3 次以内成功调整到落点偏差 d < 100 m 的范围。

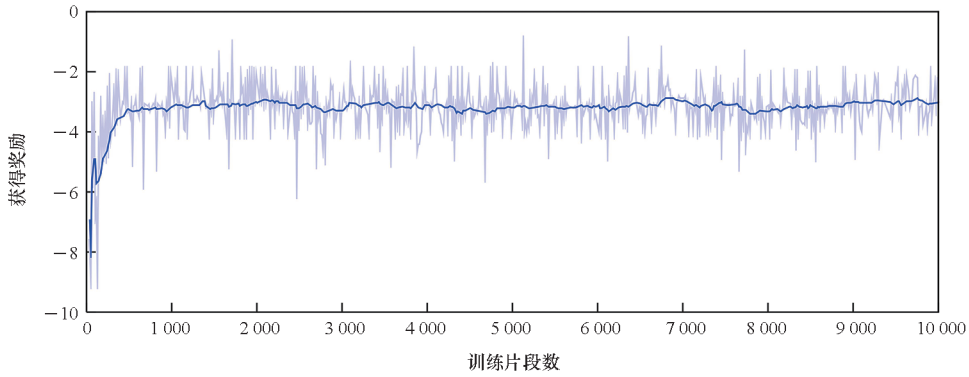
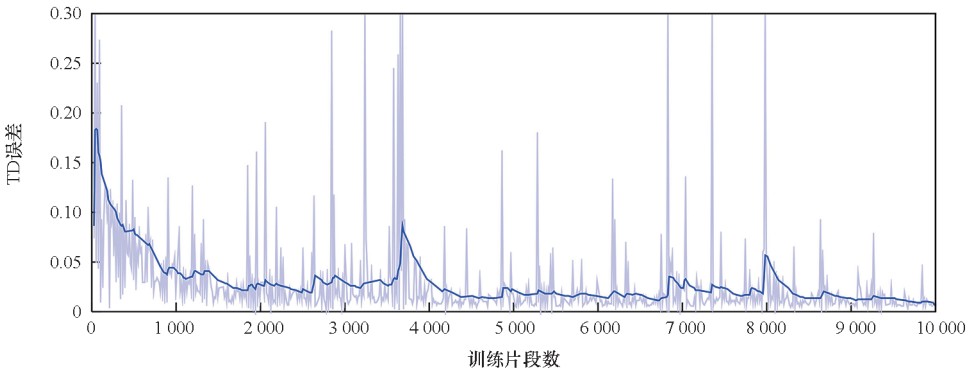


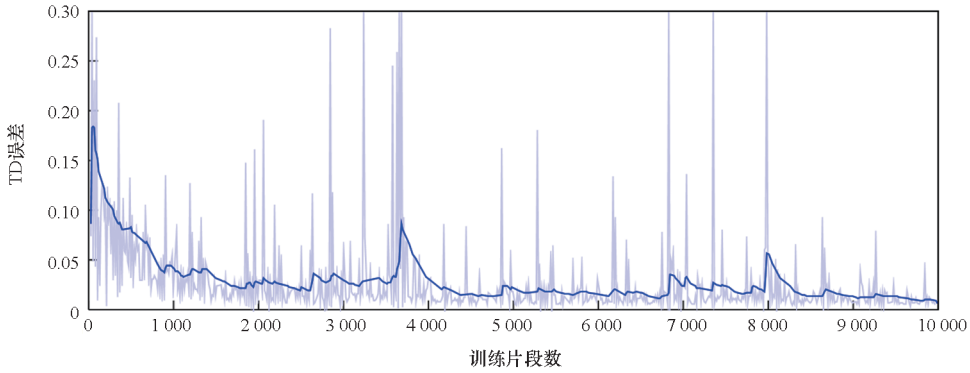
图 3　强化学习训练过程中获得奖励值随训练片段数的变化

Fig. 3　Change of reward value with the training episodes during the RL training process



（a）Q^0 TD 误差

（a）Q^0 TD error

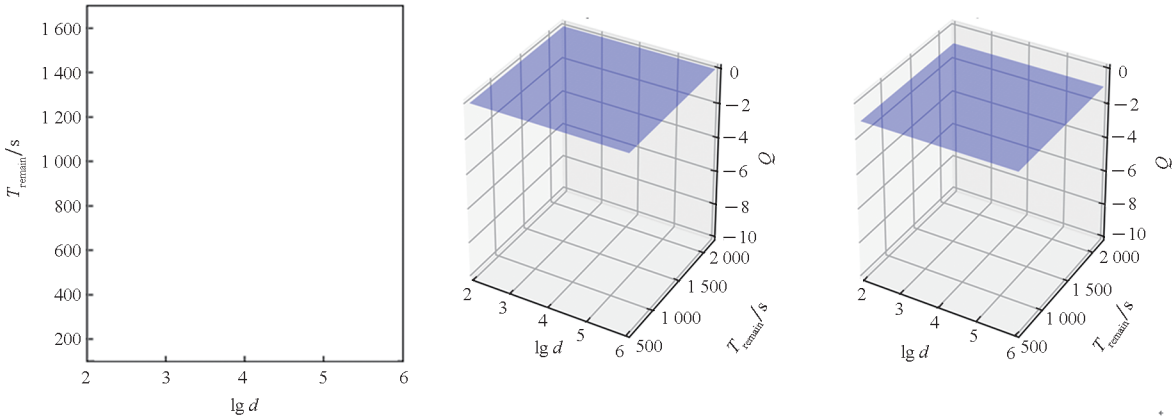


（b）Q^1 TD 误差

（b）Q^1 TD error

图 4　强化学习训练过程中 TD 误差随训练片段数的变化

Fig. 4　Variation of TD error with the training episodes during the RL training process



（a）策略的可视化

（a）Visualization of policies

（b）Q^0 函数

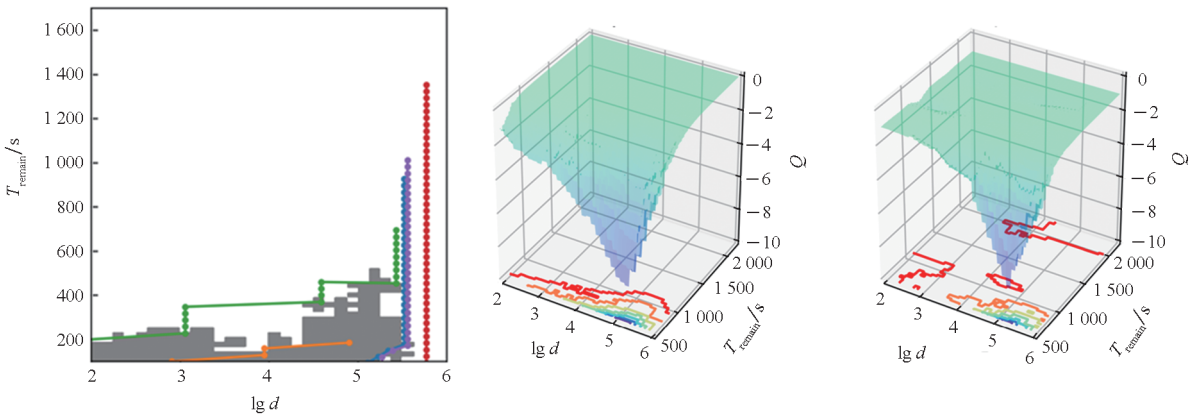
（b）Q^0 function

（c）Q^1 函数

（c）Q^1 function

图 5　学习前导弹的机动调整策略和 Q 函数图像

Fig. 5　Maneuver adjustment strategy of missile and Q function image before learning

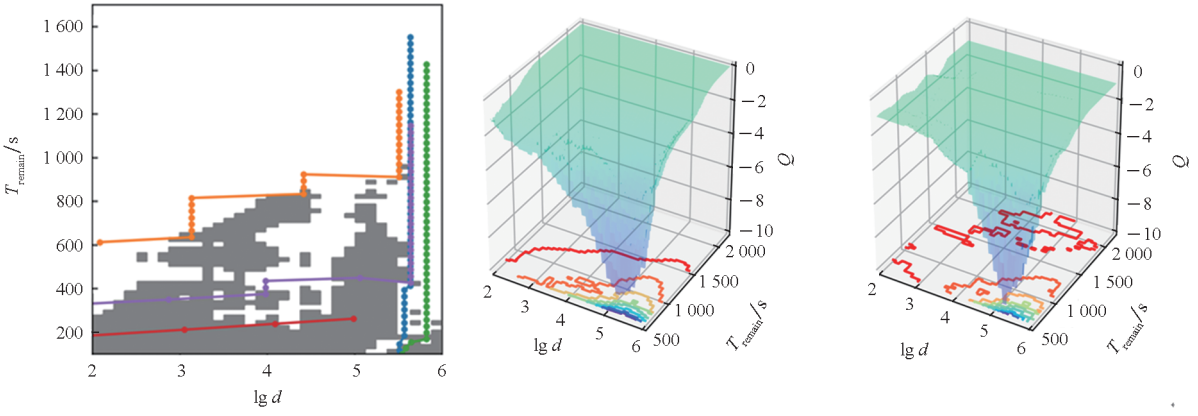


（a）策略的可视化　（b）Q^0 函数　（c）Q^1 函数

（a）Visualization of policies　（b）Q^0 function　（c）Q^1 function

图 6　学习至第 50 个片段时导弹的机动调整策略和 Q 函数图像

Fig. 6　Maneuver adjustment strategy of missile and Q function image after 50 episodes learning

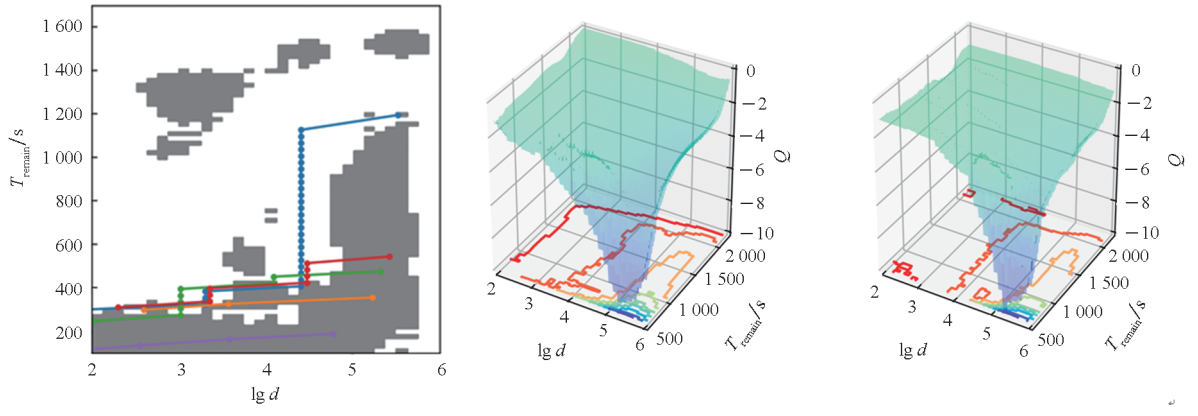


（a）策略的可视化　（b）Q^0 函数　（c）Q^1 函数

（a）Visualization of policies　（b）Q^0 function　（c）Q^1 function

图 7　学习至第 200 个片段时导弹的机动调整策略和 Q 函数图像

Fig. 7　Maneuver adjustment strategy of missile and Q function image after 200 episodes learning

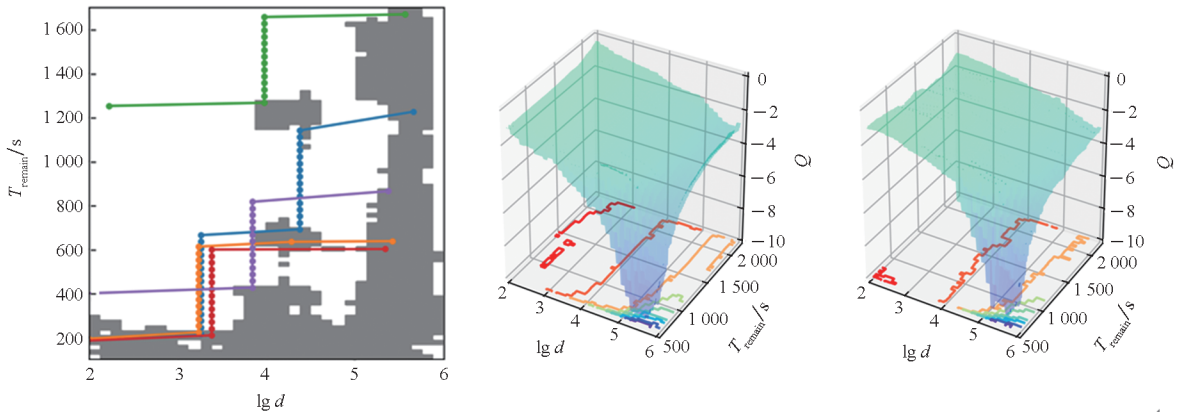


（a）策略的可视化　（b）Q^0 函数　（c）Q^1 函数

（a）Visualization of policies　（b）Q^0 function　（c）Q^1 function

图 8　学习至第 1 000 个片段时导弹的机动调整策略和 Q 函数图像

Fig. 8　Maneuver adjustment strategy of missile and Q function image after 1 000 episodes learning



(a) 策略的可视化
(a) Visualization of policies

(b) Q^0 函数
(b) Q^0 function

(c) Q^1 函数
(c) Q^1 function

图 9 学习至第 10 000 个片段时导弹的机动调整策略和 Q 函数图像

Fig. 9 Maneuver adjustment strategy of missile and Q function image after 10 000 episodes learning

对机动调整逆序 Q 学习算法获得的策略进行 1 000 次测试,结果如图 10 所示,绝大部分的测试至多进行 4 次机动调整。其中有 10 组没有将落点偏差调整至 100 m 以内,最大偏差为 154 m,这是由于图 9(a)中左下角的策略并未收敛到最优,落点偏差接近符合要求时,没有选择进行机动调整,体现了算法策略的正确性和有效性。

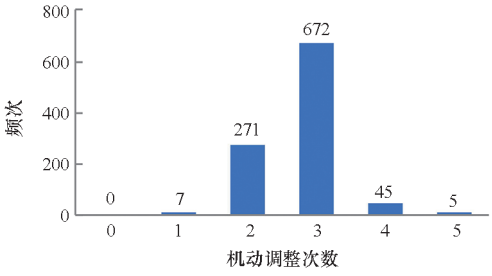


图 10 强化学习机动调整策略测试的机动调整次数分布

Fig. 10 Distribution of the number of maneuver adjustments in the RL maneuver adjustment strategy test

4 结论

本文针对利用多脉冲短时大幅轨控发动机规避来袭 EKV 后进行弹道机动调整的问题,构建导弹机动突防后的弹道机动调整最优策略模型。提出了一种高效可靠的逆序 Q 学习算法,设计了 Tile coding 强化学习的增量执行方法,在奖励函数的构造中综合考虑了控制落点偏差与降低机动次数,结合经验复用技术,对导弹的机动调整策略进行训练。

通过仿真结果可知,强化学习训练算法能够成功得到弹道机动调整的最优时机,设计的奖励函数和输入参数也可以起到相应正确的作用,证明了模型及算法的有效性。该算法对不同战术技术参数普遍适用,具备良好的适应性和延展性,可广泛用于作战场景中,解决了弹道导弹机动突防后的弹道偏离问题,保证了导弹的打击精度。

后续将进一步挖掘强化学习算法在解决导弹飞行中途临时改变打击目标点以及遭遇敌方多次拦截威胁问题中的策略潜力。

参考文献(References)

[1] YAVIN Y. A pursuit-evasion differential game with noisy measurements of the evader's bearing from the pursuer[J]. Journal of Optimization Theory and Applications, 1986, 51: 161 – 177.

[2] 吴启星, 张为华. 弹道导弹中段机动突防研究[J]. 宇航学报, 2006, 27(6): 1243 – 1247.
WU Q X, ZHANG W H. Research on midcourse maneuver penetration of ballistic missile[J]. Journal of Astronautics, 2006, 27(6): 1243 – 1247. (in Chinese)

[3] 范玉珠, 张为华, 王中伟, 等. 面向突防效能的超声速巡航导弹总体设计技术初步研究[J]. 国防科技大学学报, 2012, 34(2): 125 – 129.
FAN Y Z, ZHANG W H, WANG Z W, et al. The preliminary research of supersonic cruise missile master design based on penetration effectiveness[J]. Journal of National University of Defense Technology, 2012, 34(2): 125 – 129. (in Chinese)

[4] 闫双卡, 谭守林, 滕和平, 等. 提高巡航导弹突防能力的技术途径[J]. 飞航导弹, 2009(4): 26 – 29.
YAN S K, TAN S L, TENG H P, et al. Technical ways to improve the penetration ability of cruise missiles[J]. Winged Missiles Journal, 2009(4): 26 – 29. (in Chinese)

[5] 南英, 蒋亮. 基于深度强化学习的弹道导弹中段突防控制[J]. 指挥信息系统与技术, 2020, 11(4): 1 – 9, 27.
NAN Y, JIANG L. Midcourse penetration and control of ballistic missile based on deep reinforcement learning[J]. Command Information System and Technology, 2020, 11(4):

1 – 9, 27. (in Chinese)

[6] 范鑫磊, 李栋, 张尉, 等. 基于深度强化学习的导弹规避决策训练研究[J]. 电光与控制, 2021, 28(1): 81 – 85.
FAN X L, LI D, ZHANG W, et al. Missile evasion decision training based on deep reinforcement learning[J]. Electronics Optics & Control, 2021, 28(1): 81 – 85. (in Chinese)

[7] 谭浪, 巩庆海, 王会霞. 基于深度强化学习的追逃博弈算法[J]. 航天控制, 2018, 36(6): 3 – 8, 19.
TAN L, GONG Q H, WANG H X. Pursuit-evasion game algorithm based on deep reinforcement learning [J]. Aerospace Control, 2018, 36(6): 3 – 8, 19. (in Chinese)

[8] 高昂, 董志明, 叶红兵, 等. 基于深度强化学习的巡飞弹突防控制决策[J]. 兵工学报, 2021, 42(5): 1101 – 1110.
GAO A, DONG Z M, YE H B, et al. Loitering munition penetration control decision based on deep reinforcement learning[J]. Acta Armamentarii, 2021, 42(5): 1101 – 1110. (in Chinese)

[9] 马子杰, 高杰, 武沛羽, 等. 用于巡航导弹突防航迹规划的改进深度强化学习算法[J]. 电子技术应用, 2021, 47(8): 11 – 14, 19.
MA Z J, GAO J, WU P Y, et al. An improved deep reinforcement learning algorithm for cruise missile penetration path planning [J]. Application of Electronic Technique, 2021, 47(8): 11 – 14, 19. (in Chinese)

[10] JIANG L, NAN Y, LI Z H. Realizing midcourse penetration with deep reinforcement learning[J]. IEEE Access, 2021, 9: 89812 – 89822.

[11] LILLICRAP T P, HUNT J J, PRITZEL A, et al. Continuous control with deep reinforcement learning[EB/OL]. (2019 – 07 – 05) [2022 – 01 – 02]. https://arxiv.org/abs/1509.02971.

[12] VAN HOORN M. Optimizing air-to-air missile guidance using reinforcement learning [D]. Delft, Netherlands: Delft University of Technology, 2019.

[13] SHALUMOV V. Cooperative online Guide-Launch-Guide policy in a target-missile-defender engagement using deep reinforcement learning [J]. Aerospace Science and Technology, 2020, 104: 105996.

[14] GAUDET B, FURFARO R. Missile homing-phase guidance law design using reinforcement learning[C]//Proceedings of the AIAA Guidance, Navigation, and Control Conference, 2012.

[15] 李文振. 基于 Agent 战场仿真建模与想定的研究与实现[D]. 南京: 南京航空航天大学, 2013.
LI W Z. Agent-based research and implementation on simulation modeling and scenario in battle field environment[D]. Nanjing: Nanjing University of Aeronautics and Astronautics, 2013. (in Chinese)

[16] 王光辉, 吕超, 谢宇鹏, 等. 歼击机规避空空导弹的评价算法[J]. 系统工程与电子技术, 2016, 38(11): 2561 – 2566.
WANG G H, LYU CHAO, XIE Y P, et al. Evasive maneuver model of a fighter against air-to-air missiles[J]. Systems Engineering and Electronics, 2016, 38(11): 2561 – 2566. (in Chinese)

[17] 白金鹏, 李天. 面向指标论证的战斗机突防效能评估[J]. 航空学报, 2016, 37(1): 122 – 132.
BAI J P, LI T. Evaluation of penetration mission effectiveness oriented to fighter performance parameter analysis[J]. Acta Aeronautica et Astronautica Sinica, 2016, 37(1): 122 – 132. (in Chinese)

[18] 邵彦昊, 朱荣刚, 贺建良, 等. 基于深度学习的不可逃逸区内的规避决策研究[J]. 电光与控制, 2019, 26(11): 60 – 64.
SHAO Y H, ZHU R G, HE J L, et al. Evasive decision-making in inescapable areas based on deep learning[J]. Electronics Optics & Control, 2019, 26(11): 60 – 64. (in Chinese)

[19] ANDRADE P M C. Reinforcement learning for predictive aircraft maintenance[D]. Coimbra, Portugal: University of Coimbra, 2020.

[20] SHAYAN K, VAN KAMPEN E J. Online actor-critic-based adaptive control for a tailless aircraft with innovative control effectors[C]//Proceedings of the AIAA Scitech 2021 Forum, 2021.

[21] KONATALA R, VAN KAMPEN E J, LOOYE G. Reinforcement learning based online adaptive flight control for the cessna citation II(PH-LAB) aircraft[C]//Proceedings of the AIAA Scitech 2021 Forum, 2021.

[22] 樊博璇, 陈桂明, 林洪涛. 弹道导弹中段反应式机动突防规避策略[J]. 兵工学报, 2022, 43(1): 69 – 78.
FAN B X, CHEN G M, LIN H T. Mid-course reactive maneuver penetration and evading strategy of ballistic missile[J]. Acta Armamentarii, 2022, 43(1): 69 – 78. (in Chinese)

[23] 杨涛. 基于微分对策理论的大气层外弹道导弹弹头机动突防策略研究[D]. 长沙: 国防科学技术大学, 2015.
YANG T. Research on the evasive strategy of exo-atmospheric ballistic missile based on the theory of differential games[D]. Changsha: National University of Defense Technology, 2015. (in Chinese)

[24] 张秦浩, 敖百强, 张秦雪. Q-learning 强化学习制导律[J]. 系统工程与电子技术, 2020, 42(2): 414 – 419.
ZHANG Q H, AO B Q, ZHANG Q X. Reinforcement learning guidance law of Q-learning[J]. Systems Engineering and Electronics, 2020, 42(2): 414 – 419. (in Chinese)

(编辑: 王颖娟, 杨琴)