# 基于 DDPG 算法的末制导律设计研究

刘 扬[1)] 何泽众[1)] 王春宇[1)] 郭茂祖[2)]

[1)](哈尔滨工业大学 计算机科学与技术学院, 哈尔滨 150001)

[2)]( 北京建筑大学 电气与信息工程学院 北京 100044)

**摘 要** 末制导律设计是拦截系统中的关键技术，常用的比例制导律及其变型在目标大机动下性能下降，且受到导航比的影响。提出基于 DDPG 算法的末制导律设计方法，通过对拦截问题的环境状态和动作（控制量）设计，实现了从仿真环境交互数据中学习回报最优的制导律；与传统方法相比，该无模型方法更具灵活性；针对强化学习方法动作集假设偏置弱带来训练效率低的问题，进一步提出将导航比作为决策优化参数，加速了训练过程并实现动态调整比例制导律中的导航比。对比实验表明，两种强化学习末制导律设计方法获得了优于比例制导律及其变型的拦截效果，展现出良好的研究前景和潜在的应用价值。

关键词 末制导律；强化学习；确定性策略，归纳偏置

**中图法分类号** TP18

## Terminal guidance law design based on DDPG algorithm

Liu Yang [1)] He Zezhong [1)] Wang Chunyu [1)] Guo Maozu [2)]

[1)](School of Computer Science and Technology, Harbin Institute of Technology, Harbin 150001)

[2)](School of Electrical and Information Engineering, Beijing University of Civil Engineering and Architecture, Beijing 100044)

**Abstract** The design of terminal guidance law is the key technology in interception system. The performance of the commonly used proportional guidance law and its variants will degrade under the condition of large maneuvering target, and will be affected by the navigation ratio. A terminal guidance law design method based on ddpg algorithm is proposed. By designing the environment state and action (control quantity) of interception problem, the guidance law with optimal learning reward from the interactive data of simulation environment is realized. Compared with the traditional method, the model free method is more flexible. Aiming at the problem of low training efficiency caused by weak hypothesis bias of action set in reinforcement learning method, a further proposal is proposed Taking the navigation ratio as the decision optimization parameter, the training process is accelerated and the navigation ratio in proportional guidance law is adjusted dynamically. The comparative experiments show that the two design methods of terminal guidance law based on reinforcement learning obtain better interception effect than proportional guidance law and its variants, showing good research prospects and potential application value.

**Key words** Terminal guidance law, Reinforcement learning, Deterministic policy, Inductive bias

---

本课题部分得到国家自然科学基金可微分深度注意模型及其参数自适应方法研究(No. 61671188)资助. 刘扬(**通信作者**)，博士，副教授，主要研究领域为机器学习、计算机视觉. E-mail: liuyang@hit.edu.cn. 何泽众，硕士研究生，主要研究领域为强化学习及应用. E-mail: hzzssx@foxmail.com. 王春宇，博士，副教授，主要研究领域为机器学习. E-mail: chunyu@hit.edu.cn. 郭茂祖(**通信作者**)，博士，教授，主要研究领域为机器学习、生物信息学. E-mail: guomaozu@bucea.edu.cn.

## 1 研究背景及相关工作

控制高速飞行的导弹击中目标的末段控制规律称为末制导律，是拦截系统中的一项关键技术。制导律给出的控制量是导弹调整自身飞行状态的依据，现在工程上常用制导律多数基于比例制导律或其改进[1]。原理是通过某些控制手段，使导弹与目标之间的视线转率与导弹速度矢量旋转角速度成正比。

在理想情况下，比例制导律可以取得比较好的效果[2]。但是，在考虑弹体空气动力模型的非理想性、自动驾驶仪延迟及目标进行大机动时，这种制导律的效果会明显下降[3]，同时适当的导航比对于该类制导律的应用也至关重要[4]。为克服上述困难，研究人员也从最优控制角度出发获得相应的制导律，但这种方法是在确定性环境下建模的，而实际应用中会有一些随机因素参杂其中；同时，用于评价最优性的耗费函数需要对飞行时间进行准确估计；此外，最优控制方法也需要更多的弹上计算资源，这些都为实际应用带来难题。

随着模拟技术和遥测技术的发展，飞行控制领域可以应用的数据越来越多，这就为引入数据驱动的制导律设计方法提供了契机。相对于上述基于运动学模型的制导律设计方法，机器学习技术为使用这些数据设计末制导律提供了可行途径。

末制导律就是根据当前时刻环境状态给出控制量。这一过程与强化学习解决的问题是相似的；同时，从回报最优角度看，强化学习也是一种最优决策[5]。特别是 DQN 算法[5]在 Atari 的数款游戏中超过了人类水平以来，这也促使相关研究者尝试将深度强化学习应用到不同领域。

与游戏类最优策略求解问题不同是，强化学习也被应用在动力学环境中，其决策变量通常与智能体的运动参数相关。

在车辆的自动驾驶领域，文献[7]设计了一种根据原始视频输入进行强化学习的学习结构，智能体需要从赛车游戏的图像数据中提取出对控制有帮助的信息，同时给出合理的控制方法。文献[8]设计了一种针对自动驾驶的测试方法，由于真实环境测试时间成本较高且事故相对较为罕见，自动驾驶汽车的测试工作效率较低，文章通过强化学习方法来提高事故出现的概率以提升对自动驾驶汽车的测试速度。文献[9]在文章中提出了一个基于视觉的自动驾驶方法，文章中将视觉感知任务与控制任务分为两个部分，并在其提出的赛车模拟环境中进行了实验，结果优于线性二次型调节器和模型预测控制器。强化学习常被应用于机器人控制领域，文献[10]针对现阶段方法针对新目标泛化能力差和数据利用率低等问题，提出了一个适用于目标驱动的视觉导航任务的模型，其模型收敛速度更快且部署到真实机器人时仅需对模型进行微调。文献[11]提出了一种将原始图像数据映射到机器人电机扭矩的控制方法，文章认为相较于单独训练每一个组件，将感知与控制系统相结合得到一个端到端的控制方法效果更好。DDPG 算法[12, 13]是其中一个比较有代表性的，从形式上讲，这种基于确定性策略梯度的方法适于连续值决策变量任务。文献[14]与[15]提出了相应的自动驾驶系统，其使用 DDPG 算法训练智能体做出控制动作，文献[16]则应用 DDPG 算法训练了一个可以在斜坡上稳定行走的一个双足机器人。

在飞行控制领域，文献[17]利用强化学习方法训练滑翔机利用大气中的上升热气流进行自主导航，并通过实验证明了其飞行策略的有效性。与学习飞行策略这样的高层决策相比，一些研究将强化学习应用到动力学环境下的低层控制指令的学习上来。例如 Ng[18]在本世纪初就利用强化学习控制直升机的倒立飞行，该工作利用了飞行员的实际数据。随着四旋翼无人机快速普及，文献[19]使用基于模型的强化学习来控制四旋翼的飞行。Hwangbo[20]则提出一种新的强化学习算法，利用更少的高质量样本，在无模型的情况下，获得更加稳定四旋翼控制算法，并在实际系统中加以演示。Wang 和 Sun[21]则提出采用 DDPG 算法和积分补偿器来控制四旋翼飞行器；此外，模拟环境下离线训练和真实飞行在线训练改进控制策略有效地提升系统性能。

在制导与导航任务领域，文献[22]提出了一种基于强化学习的制导律设计方法，其结果考虑了导弹机身动力学，且对系统噪声有着一定适应性，但并没有考虑目标机动问题。文献[23]给出了一种无人机飞行控制方法，其在利用强化学习方法的同时引入原始状态特征作为先验知识，以提高智能体在训练时的学习效率。

对于拦截任务中，己方弹的末制导律设计是一个控制量决策问题。通过传感器可以获得速度，位置以及部分目标信息，它们可以较好地描述动力学环境的状体。作用在己方弹上的控制量（法向加速度）可以视为动作。这样就可以在强化学习框架下

研究末制导律设计问题，并有望克服人工设计制导律的一些难题，如精准建模需要随机方法或精确的目标运动信息；同时，目标机动以及噪声等的存在，常用的比例制导方法面临性能下降，需要尝试利用强化学习方法设计新的末制导律；应用强化学习方法解决该问题可以跳出比例制导模式的限制，发现潜在的性能更加优异的制导控制方法，也可以为该领域的专家提供参考。

本文利用 DDPG 算法学习可以适应目标复杂机动的末制导律，该方法通过从模拟环境交互数据中学习从状态到控制量的映射规律。首先针对制导控制问题特点进行了马尔科夫决策过程设计，给出了影响环境状态的变量；针对数据利用率较低导致智能体训练效率不高问题，在动作集设计时引入现有制导律信息作为归纳偏置以减小假设空间，进而提升智能体训练效率。

实验结果表明，相对于当前常用的比例制导方法，基于强化学习方法设计的末制导律比取得了更好的实验效果，在目标机动情况下脱靶量更小；也预示着随着人工智能技术的发展，基于数据和学习方法可以设计出更优的末制导律。

## 2 制导控制问题描述

本文选择在二维平面内对仿真拦截场景建立数学模型[24]。在二维平面内仿真使运算过程更加高效，仿真过程中将目标与导弹视为质点，整个仿真过程在惯性参考系下进行。图 1 给出了过程示意图：

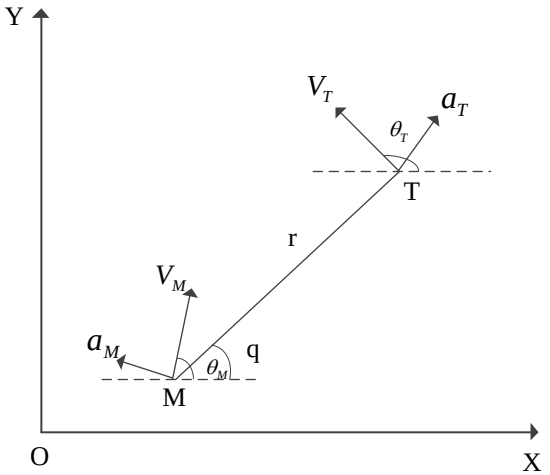


图 1 弹目拦截示意图

上图中 M 和 T 分别为导弹与目标；r 和 q 是导弹与目标相对距离和视线角；$\theta_M$ 和 $\theta_T$ 分别为弹目的弹道倾角；$a_M$ 与 $a_T$ 为对应的法向加速度，可以改变飞行方向。该过程的主要运动学方程如公式 1 所示。

$$\begin{aligned}
\dot{X}_r &= \frac{dX_r}{dt} = V_T\cos\theta_T - V_M\cos\theta_M \\
\dot{Y}_r &= \frac{dY_r}{dt} = V_T\sin\theta_T - V_M\sin\theta_M \\
q &= \arctan\left(\frac{Y_T - Y_M}{X_T - X_M}\right) \\
\dot{q} &= \frac{X_r\dot{Y}_r - Y_r\dot{X}_r}{X_r^2 + Y_r^2} \\
\frac{d\theta_M}{dt} &= \frac{a_M}{V_M}
\end{aligned} \tag{1}$$

公式(1)中的 $X_r$ 和 $Y_r$ 为相对距离在横纵坐标轴方向的分量，当下标为 M 和 T 时表示弹和目的坐标值。$\dot{q}$ 为视线转率。

当前常用的末制导方法为比例制导律，输出为导弹的法向加速度 $a_m$，其描述如下：

$$a_M := N\dot{q}v_c \tag{2}$$

其中 $v_c$ 计算如下：

$$v_c = \frac{X_r\dot{X}_r + Y_r\dot{Y}_r}{\sqrt{X_r^2 + Y_r^2}}$$

其中 $N$ 为导航比，$v_c$ 为闭合速度，$\dot{q}$ 为视线转率。

本文给出了一种制导律学习方法，应用强化学习算法对智能体进行训练。智能体在上述的仿真环境中进行仿真并更新控制方法，期望可以通过该方式得到一种性能满足要求的制导律。

## 3 马氏决策过程设计

由于强化学习方法经常被应用到不同的领域，为了能够使用较为统一的方法分析不同的问题，通常会使用马尔科夫决策过程(Markov Decision Process，以下简称 MDP)将问题进一步抽象以得到一种较为统一的描述方式。在强化学习中 MDP 通常由一个五元组构成，其形式为 $<S,A,R,P,\gamma>$，其中，$S$ 为状态的集合，状态通常用智能体能够在环境中观测到的变量表示，可理解为是对环境的一种描述；智能体的控制变量集合由动作集合 $A$ 给

出；回报函数 $R$： $S\times A\times S\to R$，为状态-动作对或状态的优劣评判，将其映射到一个实数作为其评价；状态转移概率 $P$： $S\times A\times S\to P$，某一状态在执行某一决策后到达另一状态的概率；折扣因子 $\gamma$ 为 0 到 1 之间的实数，控制着对未来回报反馈信号的重视程度；

MDP 设计的合理与否，会直接影响到强化学习算法的最终效果。由于本文选择使用 Model-Free 的强化学习方法，所以 MDP 设计中不涉及状态转移概率的部分，接下来结合待解决问题背景设计相应项。

### 3.1 状态描述

智能体观测到的环境信息来可以在一定程度上刻画环境状态，所以在设计状态集合时应尽可能地利用对解决问题有帮助的信息。而一些可能会对决策任务造成干扰的信息则不应该包含在状态集合中，本文方法中选择视线转率作为状态。

$$S:<\dot{q}> \tag{3}$$

该状态集合的设计参考了比例制导律的原理。视线转率和导弹速度矢量转率是比例制导律中的重要决策因素，本文希望通过这种设计方式为智能体提供合理的环境信息。

### 3.2 动作（控制量）集合

智能体能够给出的控制量由动作集合表示。目前训练过程的低效率限制了强化学习在一些问题中的应用，一些研究者认为目前的强化学习方法的归纳偏置较弱，这导致智能体的学习效率比一般的监督学习方法更低[25]。经典的比例制导律以垂直于导弹速度的加速度作为控制量，导弹飞行轨迹受其控制。设计动作空间时，本文参考了比例制导律的形式，同时希望能够引入一定的归纳偏置来提高智能体训练效率，故给出了如下的两种设计。

针对训练效率较低的问题这里给出了**第一种**动作空间设计。传统的比例制导律中导航比是一个可以调整的超参数，第一种动作设计中希望可以通过深度强化学习方法动态的调整导航比，故选择比例制导律的导航比作为智能体的决策变量，形式如下：

$$A:N=Navigation\quad radio\in[2,6] \tag{4}$$

$N$ 表示导航比，动作集合设计时参考了常用的导航比取值范围。

参考一些传统制导律，例如比例制导律的形式，**第二种**动作集合设计中智能体选择直接输出法向加速度。为了保证各个元器件的正常工作，实际的飞行器需要考虑过载的影响，将其限制在一定范围内。智能体的决策变量为重力加速度的倍数，过载限制在 20 倍的重力加速度之内。这种设计的归纳偏置相较于第一种设计更弱，具体形式如下：

$$A:Ng\quad N\in[-20,20] \tag{5}$$

g 为重力加速度。

### 3.3 回报函数设计与折扣因子

本文选择通过人为设计的方式得到回报函数。本文中涉及到的拦截问题本质上可以看作是一种追踪问题，由人类的先验知识可认为：在同一参考系下，导弹-目标间的视线角与导弹弹道倾角越接近，导弹越倾向于追踪目标。故本文的回报函数如下：

$$\mathrm{reward_t}=-\left(\left|\theta_{\mathrm{t}}^{\mathrm{M}}-\mathrm{q_t}\right|\right)^{\alpha} \tag{6}$$

其中 $reward_t$ 为 $t$ 时刻的回报反馈信号， $\theta_t^M$ 与 $q_t$ 分别为时刻 $t$ 导弹弹道倾角与导弹-目标视线角。智能体通过逐渐减小二者之间的差异以完成追踪拦截任务。回报函数中的幂次形式可以令智能体更好地关注两个角度差值较小时的回报变化情况。本文中 $\alpha=0.3$，因从实验过程中看，取该值时算法收敛性好。 $\gamma$ 设为 0.95。

## 4 基于 DDPG 的末制导律设计方法

本文应用 DDPG 算法，基于上述 MDP 设计对智能体进行训练，以期通过学习的方式得到制导控制方法。

### 4.1 算法中的优化目标

机器学习方法通常将训练过程看作一个优化问题，强化学习中也希望通过这种方式得到一个最优的决策方法。一般来说最优的决策方法可以最大化回报累加和。当时间无限时，可以认为不同状态的回报累加和均为无穷大，这样无法体现出不同状态间的差异，因此一般会使用折扣因子，以减小距离当前时刻过远回报的重要性。公式(7)描述了回报累加和：

$$G_t = R_t + \gamma R_{t+1} + \gamma^2 R_{t+2} + ... = \sum_{k=0}^{\infty} \gamma^k R_{t+k} \quad (7)$$

其中，$G_t$ 表示时刻 $t$ 的后续回报累加和，$R_t$ 为时刻 $t$ 回报函数的结果，$\gamma$ 则为折扣因子。基于回报累加和的概念，强化学习理论中进一步提出了值函数的定义，值函数可以针对智能体处于的某一状态给出一个评价值，公式如(8)所示：

$$V(s) = E[\sum_{k=0}^{\infty} \gamma^k R_{t+k} \mid s_t = s] = E[G_t \mid s_t = s] \quad (8)$$

在 $t$ 时刻智能体观测到状态 $s$，由于函数 $V(s)$ 是对状态的一种评价，故其可以体现出状态 $s$ 的优劣，并记为值函数。

强化学习方法希望在训练中得到一个针对待解决任务的最优策略。针对相同且合理的回报函数，一个策略能够获得的累计回报越高，说明该策略相比其他策略更具优越性。借助值函数这一形式，最优策略的形式化表述如下：

$$\pi^* \geq \pi' \qquad \forall s \in S, \ V_{\pi^*}(s) \geq V_{\pi'}(s) \quad (9)$$

其中 $\pi^*$ 表示最优策略，$\pi'$ 表示其他任意策略。在 MDP 的背景下引入了决策变量对累积回报的影响，可以给出与状态值函数 V 类似的状态-动作值函数 Q(s,a)，在某一策略的控制下整条 MDP 轨迹的累计回报可以写成如(10)所示的形式：

$$\begin{aligned} J(\theta) &= E_{\tau \sim p_\theta(\tau)}[r(\tau)] = \int p_\theta(\tau) r(\tau) d\tau \\ p(\tau) &= p(s_1) \prod_{t=1}^{T} \pi_\theta(a_t \mid s_t) p(s_{t+1} \mid s_t, a_t) \end{aligned} \quad (10)$$

其中 $\theta$ 表示策略 $\pi$ 的参数，$\tau$ 表示在当前策略下的 MDP 轨迹，其形式为 $(s_0, a_0, r_0, ..., s_T, a_T, r_T)$，$r(\tau)$ 表示整条轨迹的累积回报。

根据上述定义，一般的强化学习方法优化目标为最大化累计回报。描述如下：

$$\theta^* = \arg\max_{\theta} J(\theta) \quad (11)$$

可以通过求梯度的方式求解优化问题，于是便得到了策略梯度的形式：

$$\nabla_\theta J = E_{s \sim p_\pi, a \sim \pi_\theta(a|s)} \left[ \nabla_\theta \log \pi_\theta(a \mid s) Q^\pi(s,a) \right] \quad (12)$$

因为梯度项是对策略函数参数 $\theta$ 求导得到的，与状态转移概率有关的项都被消去，所以 Model-Free 方法不需要状态转移概率。

上述的策略函数 $\pi_\theta(a \mid s)$ 是一个概率分布，智能体通过对这个分布进行采样得到具体的决策值。

DDPG 算法基于确定性策略，即算法中使用的策略输出为确定值而非概率分布。确定性策略被认为是随机策略的一种特例[26]，这种形式被称为确定性策略梯度(Deterministic Policy Gradient)，如公式(13)所示：

$$\nabla_\theta J = E_{s \sim p_\pi} \left[ \nabla_\theta \pi_\theta(s) \nabla_a Q^\pi(s,a) |_{a=\pi_\theta(s)} \right] \quad (13)$$

策略函数参数通过梯度项来更新，从而得到适合任务的决策。DDPG 含有两个网络结构，分别为 Critic 和 Actor，并用 Replay Buffer 和延迟更新降低时序上相关性。

### 4.2 数据驱动的末制导律

本文通过智能体与仿真环境交互收集数据，并根据所获得数据对自身策略进行优化，最终给出的策略函数即为学习方法得到的控制方法。导弹控制方法示意图如下：

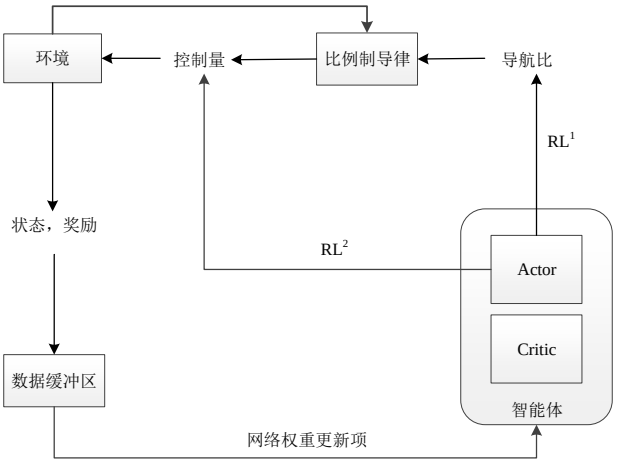


图 2 基于强化学习的末制导律学习示意图

上图中 $RL^1$ 表示决策变量为导航比的强化学习控制方法，$RL^2$ 表示决策变量为法向加速度的强化学习控制方法。图中的两种方法是单独工作的，二者的训练过程相互独立。因为大部分训练步骤较为相似，所以在示意图中将二者绘制在了一起。两种方法的训练方式基本相同，均为与环境交互收集数据，并通过从数据缓冲区随机抽样的方式消除时序数据的相关性，接下来根据数据对神经网络进行训练。二者的区别在于 Actor 网络输出的决策值不同。$RL^1$ 的决策值为用于比例制导律的导航比，强化学习方法数据利用率较低的一个原因是方法的归纳偏置较弱导致假设空间过大。因此在设计 $RL^1$ 的 MDP 动作（控制量）集合时，比例制导律采用的信息被置为归纳偏置，进而使假设空间范围缩小，本质上 $RL^1$ 方法相当于对比例制导律的一种改进。$RL^2$ 方法则采用端到端控制的设计，即决策值直接为系

统控制量，这样的设置下 $RL^2$ 方法可以在训练效率上与 $RL^1$ 方法进行直观的对比。

在网络模型权重更新过程中，为了减弱对应神经网络在更新过程中变化幅度过大对训练稳定性造成的影响，DDPG 中采用了软更新的策略，即对估值网络的 $\pi^e$ 与 $Q^e$ 和目标网络的 $\pi^t$ 与 $Q^t$ 应用反向传播算法，目标网络参照估值网络的权重进行延后更新，以提高训练中的稳定性。

目标网络的权重更新遵从如下形式：

$$\begin{aligned} \theta^t &\leftarrow \tau\theta^e + (1-\tau)\theta^t \\ w^t &\leftarrow \tau w^e + (1-\tau)w^t \end{aligned} \tag{14}$$

公式中 $\theta^e$ 与 $w^e$ 表示 Actor 与 Critic 估值网络的的参数，$\theta^t$ 与 $w^t$ 表示 Actor 与 Critic 估值网络目标网络的参数，为了达到目标网络的参数缓慢向估值网络参数靠近的效果 $\tau$ 取值为 0.01。

Critic 网络的训练类似于监督学习中的回归任务。不过这里的标签并不是数据中给出的准确值，而是通过贝尔曼最优方程得到的一个部分准确的标签。即标签 y 的形式为：

$$y_t = r_t + \gamma Q^t(s_{t+1}, \pi^t(s_{t+1})) \tag{15}$$

$r_t$ 是当前时刻的回报值，该值是通过真实采样获取，准确程度较高。由于 Q 值函数在一条轨迹中的状态上存在递推关系，所以可以通过真实采样的回报值和后继状态的 Q 值构造一个相对准确的“标签” $y_t$。使用 $y_t$ 与当前 Q 网络的估计值构造损失函数，Critic 的损失函数形式如下：

$$L^c = \frac{1}{N}\sum_t^N \left(y_t - Q^e(s_t, a_t)\right)^2 \tag{16}$$

损失函数 N 表示样本数量，其形式为由 $y_t$ 估值网络函数的结果 $Q^e$ 构造的均方误差。

Actor 网络利用确定性策略梯度进行权重更新，其形式如公式(17)所示。

$$\nabla_{\theta^e} \approx \frac{1}{N}\sum_t^N \nabla_a Q^e(s_t, a_t)\nabla_{w^e}\pi^e(s_t) \tag{17}$$

这里的确定性策略梯度是根据数据得到的无偏估计值。

强化学习智能体可以在训练过程中根据上述优化目标和优化方法逐步修正自身策略，并最终通过学习得到可用的控制方法。

# 5 实验部分

本节从训练效率、飞行轨迹、控制量、导航比及视线角速率变化对各种方法进行测试，最后从脱靶量的统计分布上分析了各种方法性能。比较的方法包括提出的两种方法，公式（2）描述的比例制导律及下面公式描述的改进后的比例制导律。

$$\frac{N_1\dot{r}\dot{q} + N_2 a_M \cos(eng_T)}{\cos(\theta_M - q)} + 9.8\cos(\theta_M)$$

其中，$eng_T$ 目标弹道倾角与视线角夹角，其余与第 2 节介绍的意义一致。其中，$N_1$ 和 $N_2$ 为 4 和 1。

## 5.1 训练效率对比实验

表 1 给出了实验的具体设置。

**表 1 训练时 DDPG 采用的相关参数**

| | |
|---|---|
| 神经网层个数 | 3 层 |
| 批大小 | 128 个 |
| 执行者学习速率 | $10^{-3}$ |
| 批评家学习速率 | $2x10^{-3}$ |
| $\tau$ | $10^{-2}$ |
| $\gamma$ | 0.95 |
| 采用的梯度优化算法 | Adam |

导弹与目标双方均处于一个二维平面坐标系内，导弹和目标的坐标分别为 (0,0)和(4300,2500)；导弹的入射角度为 60°，目标则以 170°入射；导弹以 $6x10^2$ 米/秒为开始速度，目标则以 $3x10^2$ 米/秒为开始速度，目标做加速度大小为 2 倍重力加速度的常值机动。在训练未开始前，$RL^1$ 与 $RL^2$ 的拦截仿真轨迹图像如下：

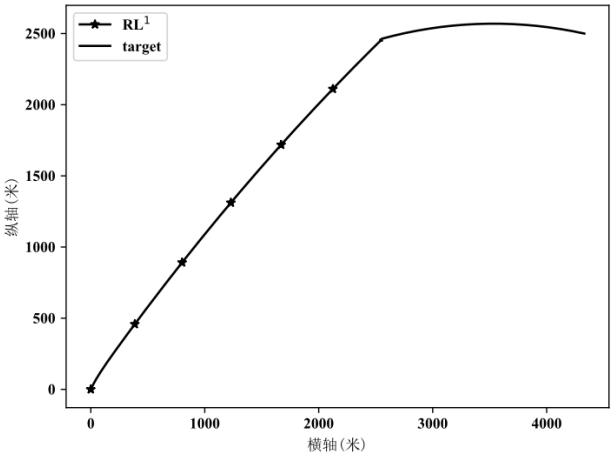


图 3 $RL^1$ 未训练仿真轨迹

图 3 与图 4 中体现出的差异较为明显，由于 $RL^1$ 有着较强的归纳偏置，这导致其在未经训练前也能表现出追踪目标的效果，不过其脱靶量相较于正常的实际要求仍有一定距离，为 14.18 米。$RL^2$ 在未训练前效果较差，可以在轨迹图像上直接观察到其偏离目标较远，脱靶量达到了 958.98 米。

$RL^1$ 引入较强归纳偏置的目标是为了提高数据利用率进而提升训练效率。为了对比两种方法的训练效率，实验中对二者训练过程中的脱靶量下降过程进行了比较。

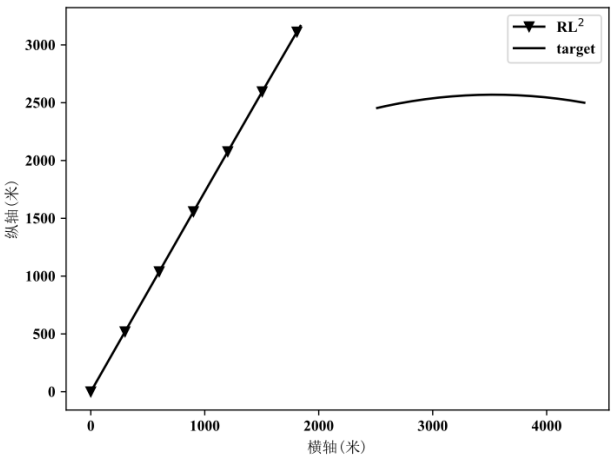


图 4 $RL^2$ 未训练仿真轨迹

$RL^1$ 引入较强归纳偏置的目标是为了提高数据利用率进而提升训练效率。为了对比两种方法的训练效率，实验中对二者训练过程中的脱靶量下降过程进行了比较。

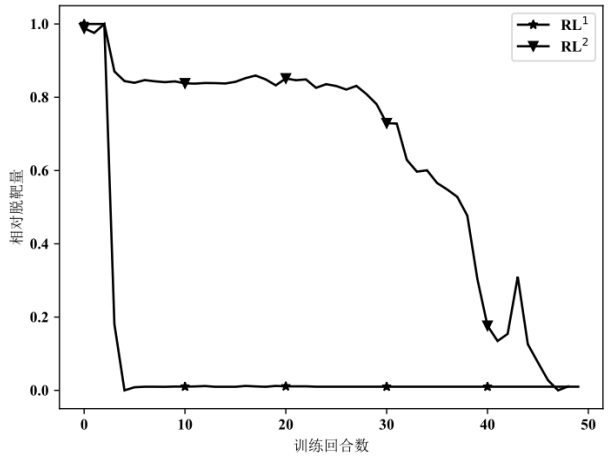


图 5 训练过程中本文所提方法相对脱靶量对比

图中横坐标为训练回合（episode）数，纵坐标为对原始数据缩放后的相对脱靶量。由于两种方法在训练开始前的脱靶量相差较大，所以在图中对其脱靶量进行了缩放，以便更加明显的对比其变化趋势。数据缩放方式为常用的 min-max 标准化，即对原始数据减去数据最小值并除以数据最大值与最小值的差值。$RL^1$ 方法相对于 $RL^2$ 在训练过程中脱靶量下降更快，仅在几个 episode 的训练之后脱靶量便达到较低的水平并收敛。而 $RL^2$ 方法在训练过程中脱靶量下降的速度较为缓慢，最终两者的脱靶量都减小到了一个较小的值，但 $RL^2$ 方法需要训练的 episode 数更高，相对 $RL^1$ 方法训练效率较低。由于 $RL^1$ 方法中引入了一定的先验知识作为归纳偏置，其假设空间和 $RL^2$ 方法相比更小，在训练过程中的收敛速度也更快，需要采样的数据更少，提高了数据利用率与训练效率。

## 5.2 强化学习制导律与经典制导律对比实验

为综合评估并对比本文的 $RL^1$ 方法、$RL^2$ 方法与比例制导律（导航比为 5）、改进的比例制导律等四种方法的表现，本文在相同的初始实验条件下分别使用上述四种方法进行了实验，并对其中的一些指标进行了对比。为检验本文提出方法的泛化能力，实验的初始条件与训练时不同，如表 2 所示：

表 2 三种方法对比实验初始条件

| | |
|---|---|
| 初始导弹位置（米） | (0.0, 0.0) |
| 初始目标位置（米） | ($4x10^3$, $2.3x10^3$) |
| 被拦截目标加速度（g） | 5 |
| 被拦截目标所采用机动 | 常定值或者正弦机动 |
| 入射角（拦截方）(度) | 4x10 |
| 入射角（被拦截目标）(度) | $1.8x10^2$ |

表 2 给出初始实验条件下的对比实验，其对应的脱靶量结果在下表中给出。

表 3 四种方法对比实验脱靶量结果

| 被拦截目标机动 | 制导方法 | 脱靶量(米) |
|---|---|---|
| 常定值机动 | PNG | 5.77 |
| | 改进的 PNG | 1.01 |
| | $RL^1$ | 1.60 |
| | $RL^2$ | **0.17** |
| 正弦机动 | PNG | 2.15 |
| | 改进的 PNG | 2.02 |
| | $RL^1$ | 2.32 |
| | $RL^2$ | **1.15** |

根据表 3 中给出的脱靶量数据，在当前实验初始条件下：**$RL^2$ 在两种目标机动方式下的脱靶量均优于其他三种方法。**

**在常值机动的情况下**，比例制导律的表现较差，脱靶量达到了 5.774 米，远大于另外几种方法；**在正弦机动的情况下**，$RL^1$ 方法与两种比例制导方法效果相近但都不如 $RL^2$ 方法。

$RL^1$ 在面对两种目标机动方式时表现均稍逊于 $RL^2$ 方法，虽然 $RL^1$ 方法在训练过程中效率更高，由于其假设空间被限制在了比例制导的范围内，这可能导致了最终得到的控制方法也会拥有比例制导律自身固有的缺点。该组实验结果体现出在当前实验初始条件下，本文提出的 $RL^1$ 方法与 $RL^2$ 方法相较于比例制导律，效果均有一定提升。**尤其 $RL^2$ 方法由于其归纳偏置较弱，训练效率较低，但相对于 $RL^1$ 方法灵活度也更高，在智能体训练过程中可以突破比例制导律的限制，得到更好的控制策略。**

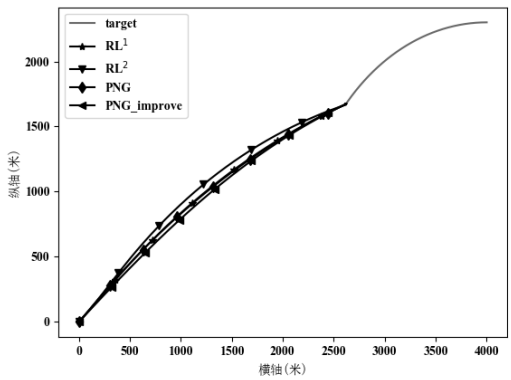


图 6 -1 目标常值机动仿真轨迹

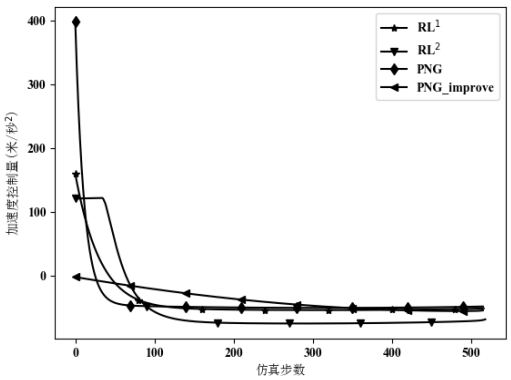


图 6-2 目标常值机动控制量

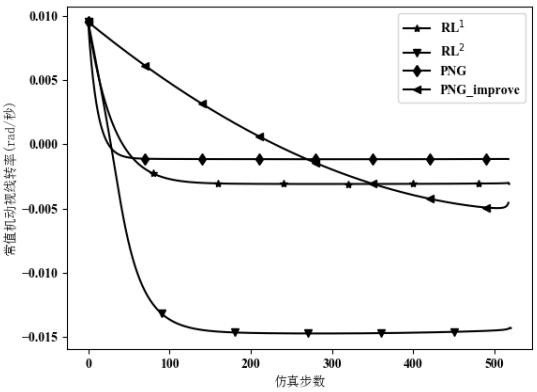


图 6-3 常值机动弹目视线角速率

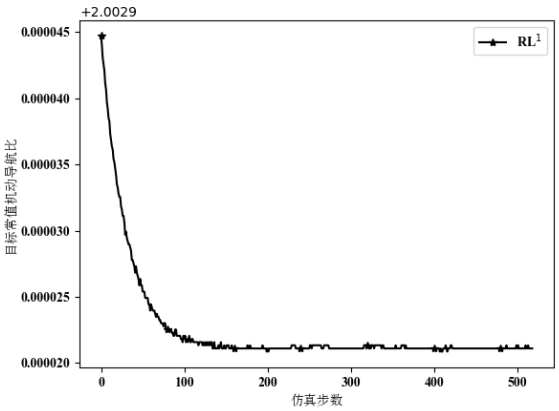


图 6-4 目标常值机动 $RL^1$ 导航比

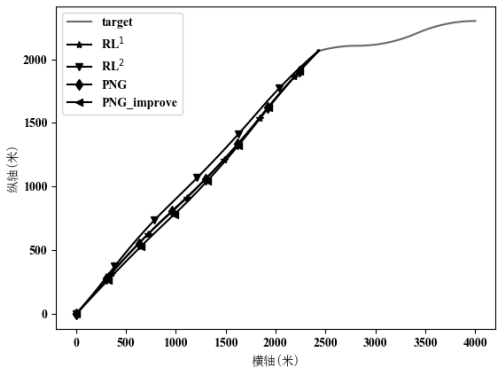


图 6-5 目标正弦机动仿真轨迹

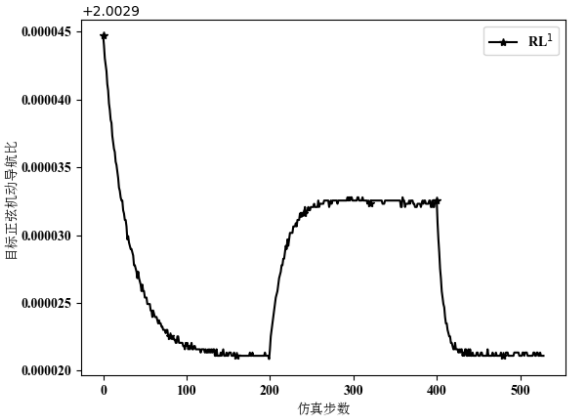


图 6-8 目标正弦机动 $RL^1$ 导航比

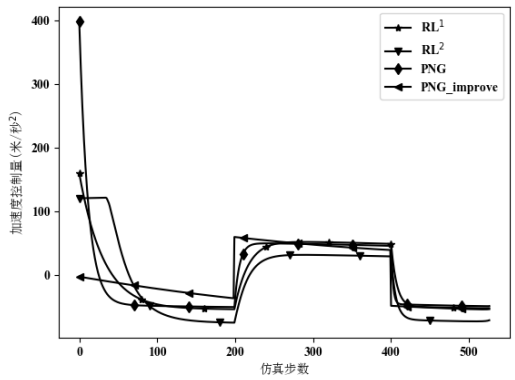


图 6-6 目标正弦机动控制量

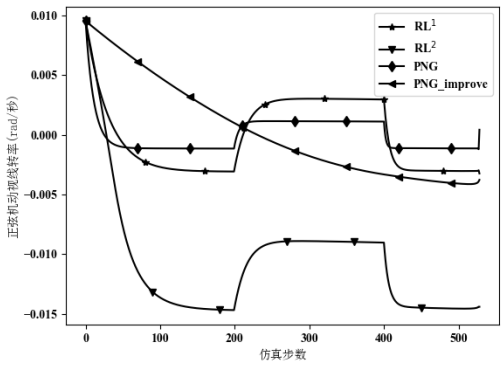


图 6-7 目标正弦机动弹目视线角速率

此外，在两种机动情况下，实验还对四种方法的仿真轨迹、输出的控制量、弹目视线角速度和 $RL^1$ 方法给出的导航比进行了对比。

图 6-1 和图 6-5 分别给出了各方法在目标常值机动和正弦机动的拦截轨迹。可以看到，$RL^1$ 方法作为一种针对比例制导律的改进方法，其轨迹与另外两种比例制导方法相近；而 $RL^2$ 方法的仿真轨迹则与其他两种方法有一定区别。这也为后续观察人工智能方法得到解的特点提供了基础，可以在人类专家设计新型制导律时起到借鉴作用。

图 6-2 和图 6-6 给出两种机动下的各方法的控制量。图中控制量的正负分别表示与导弹飞行速度矢量正交的两个方向，而非传统数值中的正负含义。在控制量曲线中有着类似的情况，不过 $RL^1$ 方法输出的控制量相对于比例制导律变化更加平滑。实验中比例制导律给出的结果超出了合理的过载范围，而本文提出的两种方法对控制量的范围做出了限制，更接近实际的飞行器工作状态。

图 6-3 和图 6-7 给出了目标两种机动下各方法弹目视线角速率随时间的收敛情况。从中可以看到比例制导方法视线角速率收敛到更接近 0 的位置。造成这一结果的原因如下：比例制导方法本身以抑制视线角速率为目标，而本文提出的基于强化学习的方法，优化目标并不是视线角速率为 0；同时目标机动可能也是造成基于强化学习方法的视线角速率收敛到非 0 位置的原因；此外，从改进的比例制导律的弹目视线角速度变化看，其也没有收敛到 0，但其效果更好。

图 6-4 和图 6-8 为 $RL^1$ 方法得到的动态调节导航比。可以看出其数值变化不大，但从脱靶量的统

计分析上看，动态调节导航比有助于提升比例制导的性能。可见这个导航比对性能还是有较大影响，如能合适地调节该值，对改进传统方法也是有帮助的。

### 5.3 脱靶量分布对比实验

单一情况实验难以较全面地评价末制导律的性能，为了验证提出的两种控制方法的鲁棒性与综合表现，本节实验对比 $RL^1$、$RL^2$、比例制导律及改进的比例制导律在多种复杂情况下的脱靶量分布。实验初始条件如下表所示：

**表 4 四种方法对比实验初始条件**

| | |
|---|---|
| 初始导弹位置（米） | (0.0, 0.0) |
| 初始目标位置（米） | ($8.6x10^3$, $4x10^3$) |
| 被拦截目标加速度（g） | 7/ 5/ 3 |
| 被拦截目标所采用机动 | 常定值机动 |
| 入射角（拦截方）(度) | 6x10 |
| 入射角（被拦截目标）(度) | $1.40x10^2$ - $2.2x10^2$ |

由于本文给出的两种方法限制了控制量的范围，所以当目标入射角度发生变化时可能出现无法捕获目标的情况，实验中增加了初始状态下导弹与目标之间的相对距离。目标的入射角度在 140°到 220°之间随机选取，该角度为整数。在该实验设置下对四种制导方法进行了对比，每种方法进行了 1000 次实验并对结果进行了统计。结果下表给出：

**表 5-1 四种方法目标常值机动脱靶量统计**

| 待对比方法 | 加速度（g） | 脱靶量(米) 均值 | 最大值 | 最小值 | 标准差 |
|---|---|---|---|---|---|
| $RL^1$（提出方法 1） | 3 | 1.90 | 4.30 | 0.02 | 1.08 |
| $RL^2$（提出方法 2） | | 1.83 | 3.89 | 0.16 | 1.03 |
| PNG（比例制导） | | 2.00 | 4.35 | 0.00 | 1.20 |
| 改进的 PNG | | **1.80** | 4.25 | 0.23 | 1.12 |
| $RL^1$（提出方法 1） | 5 | 1.27 | 3.63 | 0.04 | 0.93 |
| $RL^2$（提出方法 2） | | **1.16** | 3.30 | 0.04 | 0.81 |
| PNG（比例制导） | | 1.31 | 3.61 | 0.03 | 0.79 |
| 改进的 PNG | | 1.23 | 3.56 | 0.06 | 0.86 |
| $RL^1$（提出方法 1） | 7 | **0.90** | 2.24 | 0.06 | 0.54 |
| $RL^2$（提出方法 2） | | 0.94 | 2.65 | 0.06 | 0.62 |
| PNG（比例制导） | | 0.99 | 2.16 | 0.00 | 0.51 |
| 改进的 PNG | | 0.94 | 2.67 | 0.02 | 0.60 |

**表 5-2 四种方法目标正弦机动脱靶量统计**

| 待对比方法 | 加速度（g） | 脱靶量(m) 均值 | 最大值 | 最小值 | 标准差 |
|---|---|---|---|---|---|
| $RL^1$（提出方法 1） | 3 | 2.16 | 9.80 | 0.03 | 1.66 |
| $RL^2$（提出方法 2） | | **1.96** | 4.27 | 0.03 | 1.13 |
| PNG（比例制导） | | 2.20 | 9.81 | 0.11 | 1.66 |
| 改进的 PNG | | 2.15 | 4.19 | 0.12 | 1.24 |
| $RL^1$（提出方法 1） | 5 | **2.07** | 15.74 | 0.02 | 1.57 |
| $RL^2$（提出方法 2） | | 2.22 | 4.42 | 0.04 | 1.21 |
| PNG（比例制导） | | 2.15 | 15.71 | 0.02 | 1.57 |
| 改进的 PNG | | 2.14 | 4.39 | 0.02 | 1.17 |
| $RL^1$（提出方法 1） | 7 | **1.98** | 6.49 | 0.07 | 1.42 |
| $RL^2$（提出方法 2） | | 2.02 | 4.26 | 0.04 | 1.24 |
| PNG（比例制导） | | 2.11 | 6.13 | 0.04 | 1.44 |
| 改进的 PNG | | 2.02 | 4.34 | 0.10 | 1.17 |

根据不同情况下的脱靶量均值数据，本文提出的两种方法脱靶量均值均小于两种比例制导律，在目标加速度为 7g 时 $RL^1$ 方法的脱靶量均值小于 $RL^2$ 方法的脱靶量均值，其余两种情况均略大于 $RL^2$ 方法的脱靶量均值。从结果的稳定性来看，比例制导律在多数情况下相对好一些，其标准差小于本文给出的两种方法，这很可能是由于其系统的线性假设造成的，也是后续强化学习需要改进的地方。

为了更加直观的对比四种种方法的脱靶量分布情况，实验根据结果统计了不同方法脱靶量的分布密度，并根据数据绘制了累积分布曲线，如下图所示。

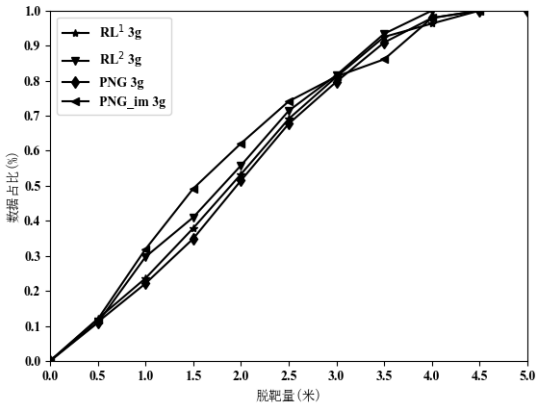


图 7-1 目标常值机动 3g 脱靶量累积分布曲线

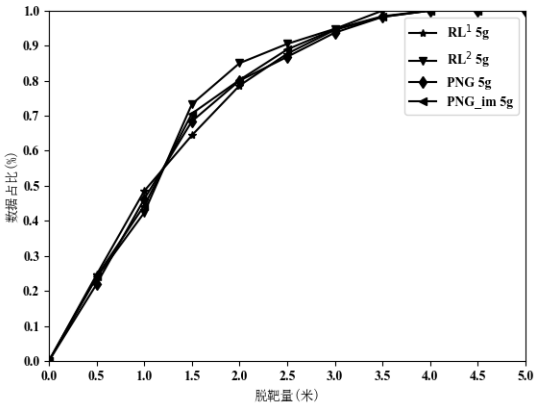


图 7-2 目标常值机动 5g 脱靶量累积分布曲线

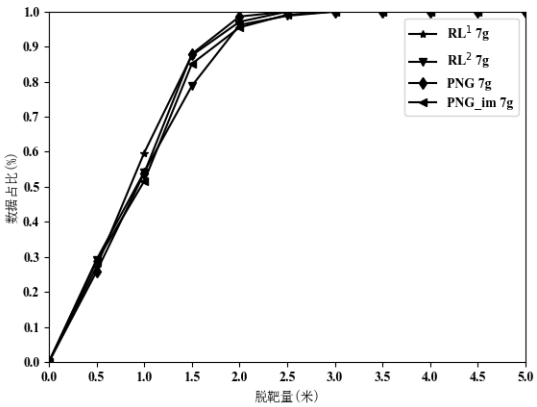


图 7-3 目标常值机动 7g 脱靶量累积分布曲线

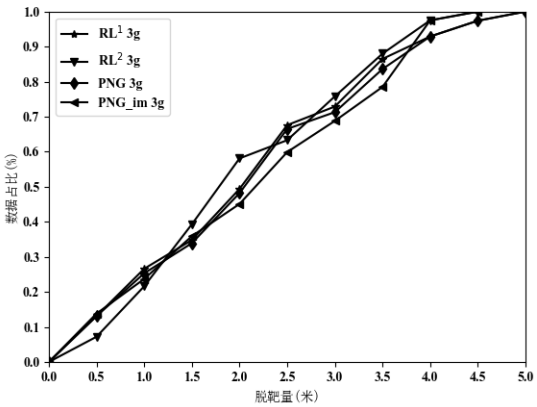


图 7-4 目标正弦机动 3g 脱靶量累积分布曲线

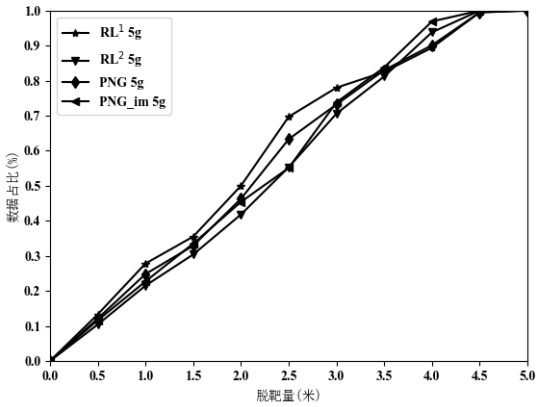


图 7-5 目标正弦机动 5g 脱靶量累积分布曲线

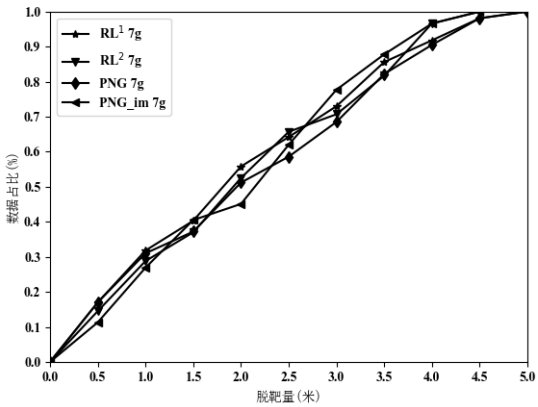


图 7-5 目标正弦机动 7g 脱靶量累积分布曲线

在图 7 所示的曲线中，可以认为其含义类似于 ROC 曲线，曲线与 x 坐标轴包围的面积越大，可以认为其脱靶量更多的分布在较小的范围内，对应制导方法效果更好。在上图曲线中，随着目标加速度在 3 倍、5 倍和 7 倍重力加速度之间变化时，四种种方法的整体效果逐步提升，曲线与坐标轴包围的面积越来越大。

对于目标常值机动时，对应 5g 和 7g 时，本文给出的 $RL^2$ 方法效果优于另外三种方法，但与 $RL^1$，接近。当目标加速度为 3g 时， $RL^2$ 方法的 1 米内脱靶量最好，但在 1-2.5 时改进的比例制导较好，但 5 米内情况又弱于 $RL^2$。

对于正弦机动情况，3g 时，改进后的比例制导效果最好，与本文提出的方法与其相近，差别不大，而且在某些区间分布上更好。当目标加速度为 5g 和 7g 时。$RL^1$ 效果最佳，$RL^2$ 与其相当，好于比例制导律及改进方法。

综上结果，可以看到，基于强化学习的末制导律方法表现出更好的性能，也为后续改进奠定基础。

# 6 结论

本文给了一种基于 DDPG 算法的导弹控制方法。首先将制导控制问题抽象为适合强化学习方法解决的形式，接下来对导弹控制方法进行设计，并在仿真环境中对得到的控制方法进行了性能测试，得到的结论如下：

(1) 本文针对制导控制问题特点，将整个过程抽象为马尔科夫决策过程。同时针对强化学习方法在应用中数据利用率较低的问题，在动作集合设计中引入了较强的归纳偏置，与一般的端到端控制设计相比，智能体的训练效率得到了提升。

(2) 本文基于 DDPG 算法给出了一种导弹控制方法，可以通过智能体与环境在不断交互的过程中修正自身策略进而得到控制方法。在智能体训练过程中基于马尔科夫决策过程中不同的动作集合设计得到了两种控制方法，对比了训练效率，并进行了仿真实验以检验其实际性能。

(3) 对本文得到的两种控制方法进行了性能测试，在仿真环境下与比例制导律进行了对比。针对目标的不同机动幅度、机动方式，实验部分使用上述三种控制方法进行了实验，并对比相应结果，采用端到端控制的强化学习导弹控制方法效果更好。为了更加全面的测试本文方法的性能，实验部分对不同初始情况的仿真结果进行了统计了，并给出了脱靶量分布结果，本文给出的导弹控制方法在多数情况下性能更加优秀。

**致 谢** 本文研究工作还受到了国家自然科学基金项目 61976071 的资助；作者感谢哈尔滨工业航天学院班晓军教授对该研究工作提出的建议。

## 参 考 文 献

[1] Madany Y. M., El-Badawy A. and Soliman A. M.. Optimal Proportional Navigation Guidance Using Pseudo Sensor Enhancement Method (PSEM) for Flexible Interceptor Applications. International Conference on Computer Modelling and Simulation, Cambridge, UK. 2016: 372-377

[2] Yanushevsky R.. Modern Missile Guidance. Boca Raton, USA: CRC Press, 2008

[3] Nesline N., Zarchan P.. Why Modern Controllers Can Go Unstable in Practice. Journal of Guidance, 1984，7(4): 495-500

[4] Ulybyshev Y.. Terminal Guidance Law Based on Proportional Navigation[J]. Journal of Guidance Control & Dynamics, 2005, 28(4):821-824

[5] Sutton R. S., Barto, A. G.. Reinforcement Learning: An Introduction (2nd). Cambridge, USA: MIT Press, 2018

[6] Mnih V., Kavukcuoglu K., Silver D.，et al. Human-level control through deep reinforcement learning. Nature, 2015, 518(7540): 529-533

[7] Lange S., Riedmiller M. and Voigtländer A.. Autonomous reinforcement learning on raw visual input data in a real world application. International Joint Conference on Neural Networks. Brisbane, Australia, 2012: 1-8

[8] O'Kelly M., Sinha A., Namkoong H. et al. Scalable End-to-End Autonomous Vehicle Testing via Rare-event Simulation. Neural Information Processing Systems. Montreal, Canada, 2018: 9849-9860

[9] Li D., Zhao D, Zhang Q., et al. Reinforcement Learning and Deep Learning based Lateral Control for Autonomous Driving. IEEE Computational Intelligence Magazine, 2018, 14: 83-98

[10] Zhu Y., Mottaghi R., Kolve E., et al. Target-driven visual navigation in indoor scenes using deep reinforcement learning. IEEE International Conference on Robotics and Automation. Singapore, 2017: 3357-3364

[11] Sergey L., Chelsea F., Trevor D., Pieter A.. End-to-end training of deep visuomotor policies. The Journal of Machine Learning Research, 2016, 17: 1334-1373

[12] Timothy P. L., Jonathan J.H., et al. Continuous control with deep reinforcement learning. arXiv preprint arXiv:1509.02971. 2015

[13] Liu Quan, Zhai Jian-Wei, Zhang Zong-Zhang, Zhong Shan, Zhou Qian, Zhang Peng, Xu Jin, A Survey on Deep Reinforcement Learning, 2017,Vol.40: 1-27
(刘全,翟建伟,章宗长,钟珊,周倩,章鹏,徐进,深度强化学习综述,2017，Vol.40: 1-27)

[14] Zuo Si-Xiang. Intelligent control of autonomous driving based on Deep Reinforcement Learning [Master's thesis]. Harbin Institute of Technology, 2018
(左思翔. 基于深度强化学习的无人驾驶智能决策控制研究[硕士学位论文]. 哈尔滨工业大学, 2018)

[15] Li Guo-Hao. End-to-End autonomous driving using deep deterministic policy gradient based on 3D convolutional neural network. Electronic Design Engineering, 2018, 26(22):162-165+174
(李国豪. 基于 3D CNN-DDPG端到端无人驾驶控制. 电子设计工程, 2018, 26(22):162-165+174)

[16] Wu X., Liu S, Zhang T., et al. Motion Control for Biped Robot via DDPG-based Deep Reinforcement Learning. WRC Symposium on Advanced Robotics and Automation. Beijing, China, 2018: 40-45

[17] Gautam, R., Jerome, et al. Glider soaring via reinforcement learning in the field. Nature, 2018, 562: 236-239

[18] Ng A. Y., Coates A., Diel M., et al. Autonomous inverted helicopter flight via reinforcement learning. International Symposium on Experiment Robotics. Singapore, 2004: 363-372

[19] Waslander S. L., Hoffmann G. M., Jang J. S., and Tomlin C. J.. Multiagent quadrotor testbed control design: Integral sliding mode vs. reinforcement learning. IEEE/RSJ International Conference on Intelligent Robots and Systems, Edmonton, Canada, 2005: 468-473

[20] Hwangbo J., Sa I., Siegwart R., and Hutter M.. Control of a quadrotor with reinforcement learning. IEEE Robotics and Automation Letters, 2017, 2(4): 2096–2103

[21] Wang Y., Sun J., He H., and Sun C.. Deterministic Policy Gradient With Integral Compensator for Robust Quadrotor Control. IEEE Transactions on Systems, Man, Cybernetics: Systems, DOI: 10.1109/TSMC.2018.2884725

[22] Gaudet B., Furfaro R.. Missile Homing-Phase Guidance Law Design Using Reinforcement Learning. AIAA Guidance Navigation and Control Conference. Minneapolis, USA, 2012: 1-20

[23] Wei H., Research of UCAV air combat based on Reinforcement Learning[Master's thesis]. Harbin Institute of Technology., 2015
(魏航. 基于强化学习的无人机空中格斗算法研究[硕士学位论文]. 哈尔滨工业大学, 2015)

[24] KIm B.S., Lee J.G., Han HS. Biased PNG law for impact with angular constraint. IEEE Transactions on Aerospace and Electronic Systems, 1998, 34(1):277-288.

[25] Matthew B., Sam R., Wang J. X., Zeb Kurth-Nelson, Charles Blundell, Demis Hassabis. Reinforcement Learning, Fast and Slow. Trends in Cognitive Sciences, 2019, 23: 408-422

[26] Silver D., Lever G., et al. Deterministic Policy Gradient Algorithms. International Conference on International Conference on Machine Learning. Beijing, China, 2014, 32: 387-395

**LIU Yang**, Ph.D., associate professor. His research interest covers machine learning, image processing, and computer vision.

**HE Zezhong**, M.S. candidate. His research interest covers r einforcement learning and machine learning.

**WANG Chun-Yu** Ph.D., associate professor. His research interest covers machine learning and bioinformatics.

**GUO Mao-Zu** Ph.D., Professor. His research interests include machine learning, Bioinformatics and urban computing.

**Background**

The guidance law is the core of unmanned aerial vehicle control, and its main effect is to help the unmanned aerial vehicle to plan a reasonable flight path. With the development of technology, the maneuverability of the aircraft has been greatly improved. When the target is faced with an interception strike, it can evade more flexibly. When facing high maneuverability targets, the current guidance law exposes many deficiencies. The current guidance law usually requires manual compensation for different maneuvers of the target. When the target makes a large-scale maneuver at close range, the current method is often difficult to quickly adapt to the maneuver method, and the miss distance increases. In response to the above problems, many researchers have improved the classical guidance law to obtain better performance. With the development of artificial intelligence, related technologies have also provided new ideas for the design and improvement of guidance laws. Under the current experimental conditions, the results obtained by the reinforcement learning method in this paper are superior to the proportional navigation method. Reinforcement learning is a branch of artificial intelligence and a general framework for solving sequence decision problems. The agent updates its own policy by interacting with the environment, and finally achieves the purpose of completing the task. Controlling the aircraft to complete the tracking task can also be regarded as a sequence decision process, and the reinforcement control method can be used to give the control amount of the aircraft. At the same time, aircraft control is also an excellent research problem for reinforcement learning. Reinforcement learning algorithms can check their performance in aircraft control tasks, discover potential deficiencies and make

improvements. It also helps reinforcement learning methods improve sample utilization to reduce deployment costs in real environments and expands the application field of reinforcement learning. This paper provides a new testing environment for the theoretical study of reinforcement learning. According to the characteristics of the reinforcement learning agent training process, the simulation environment in this paper has made corresponding improvements to the general simulation process., so that it can be flexibly adapted to most reinforcement learning algorithms. Based on the above, the research results of this paper can provide a reference for the design of new guidance laws, and also help the theoretical research of reinforcement learning.