网络首发地址：https://kns.cnki.net/kcms/detail/42.1755.tj.20210824.1515.001.html 期刊网址：www.ship-research.com

引用格式：王兴众, 王敏, 罗威. 基于 SAC 算法的作战仿真推演智能决策技术 [J]. 中国舰船研究, 2021, 16(6): 99–108.
WANG X Z, WANG M, LUO W. Intelligent decision technology in combat deduction based on soft actor-critic algorithm[J]. Chinese Journal of Ship Research, 2021, 16(6): 99–108.

基于 SAC 算法的作战仿真推演智能决策技术

扫码阅读全文

王兴众, 王敏*, 罗威
中国舰船研究设计中心, 湖北 武汉 430064

摘 要: ［**目的**］现有作战推演仿真系统主要基于作战规则和经验知识作决策, 但存在应用场景有限、效率低、灵活性差等问题。为此, 提出一种基于深度强化学习 (DRL) 技术的智能决策模型。［**方法**］首先, 建立仿真推演的最大熵马尔科夫决策过程(MDP); 然后, 以 actor-critic (AC) 体系为基础构建智能体训练网络, 生成随机化策略以提高智能体的探索能力, 利用软策略迭代更新的方法搜索更优策略, 不断提高智能体的决策水平; 最后, 在仿真推演平台上对决策模型进行验证。［**结果**］结果表明, 利用改进 SAC 决策算法训练的智能体能够实现自主决策, 且与深度确定性策略梯度(DDPG)算法相比, 获胜概率约提高了 24.53%。［**结论**］所提出的决策模型设计方案可以为智能决策技术研究提供理论参考, 对作战仿真推演具有借鉴意义。

关键词: 作战推演; 自主决策; 深度强化学习; 软策略迭代; 最大熵
中图分类号: U662.9 **文献标志码:** A **DOI:** 10.19693/j.issn.1673-3185.02099

Intelligent decision technology in combat deduction based on soft actor-critic algorithm

WANG Xingzhong, *WANG Min**, *LUO Wei*
China Ship Development and Design Center, Wuhan 430064, China

Abstract: ［**Objectives**］The existing combat deduction simulation system mainly implements decision-making based on operational rules and experience knowledge, and it has certain problems such as limited application scenarios, low decision-making efficiency and poor flexibility. In view of the shortcomings of conventional decision-making methods, an intelligent decision-making model based on deep reinforcement learning (DRL) technology is proposed. ［**Methods**］First, the maximum entropy Markov decision process(MDP) of simulation deduction is established, and then the agent training network is constructed on the basis of actor-critic architecture to generate randomization policies that improve the agent's exploration ability. At the same time, the soft policy iterative updating method is used to search for better policies and continuously improve the agent's decision-making level. Finally, the simulation is carried out on the Mozi AI platform to validate the model. ［**Results**］The results show that an agent trained with the improved soft actor-critic (SAC) decision-making algorithm can achieve autonomous decision-making. Compared with the deep deterministic policy gradient (DDPG) algorithm, the probability of winning is increased by 24.53%. ［**Conclusions**］The design scheme of this decision-making model can provide theoretical references for research on intelligent decision-making technology, giving it some reference significance for warfare simulation and deduction.

Key words: combat deduction; independent decision making; deep reinforcement learning (DRL); soft policy iteration; maximum entropy

0 引 言

人工智能技术的快速发展正在加速战争形态的演变, 军事智能化已成为新一轮军事变革的核心驱动力[1]。认知速度是智能时代制胜之道的根本。在认知域, 自主决策将逐步取代辅助决策。

收稿日期: 2020–08–31 **修回日期:** 2021–02–04 **网络首发时间:** 2021–08–25 18:48

作者简介: 王兴众, 男, 1979 年生, 博士, 高级工程师
王敏, 女, 1997 年生, 硕士生
罗威, 男, 1980 年生, 博士, 高级工程师
***通信作者:** 王敏

在认知战这种全新智能化战争需求推动下，智能决策问题亟待解决。作战仿真推演系统是体系作战的重要决策判断工具，利用从战争或军事训练中抽象得到的作战规则，对战场环境、军事力量、作战行动等要素进行形式化描述和建模，依据模型，推演分析作战过程及其结果和伤亡等情况，是军队开展模拟训练、科学评估作战预案的有效途径。一旦认知战成为未来体系对抗的主战场，仿真推演系统将成为研究对抗战术的虚拟战场，而智能决策则是影响战争走向的机关枢纽[2]。

传统的作战推演决策主要依靠固化在模型中的决策部件，它可直接实现仿真模型对战场态势的理解，输出对应的决策方案[3]。而外部决策模型则是与之相对的另一种传统决策建模方法，其将指挥员的决策知识和判断经验录入知识库，作为该模型进行仿真推演的依据，例如基于规则和条件的决策建模方法。但是，一旦状态空间规模增加，上述决策方法将难以维护，而行为树的模块化和可复用性使其可成为了强有力的决策工具[4]。2017 年 4 月 26 日，美国国防部率先提出成立“算法战跨功能小组（AWCFT）”，以推动人工智能和大数据等技术加速融入军事领域，正式开启了认知智能军事化应用探索的进程。2019 年 12 月 31 日，美国智库“战略与预算评估中心 (CSBA)”发布了《夺回海上优势：为实施“决策中心战”推进美国水面舰艇部队转型》的报告，指出“决策中心战”作战概念将成为美军智能化转型建设的理论牵引。为此，国内也在积极开展相关研究。国防科技创新特区于 2018 年起举办了战术级别的“先知兵圣”人机对抗挑战赛，对抗双方基于“兵棋陆战”平台，在城市和山地等多地形、多地貌虚拟场景下进行自主推演、决策执行和“夺控点”的争抢，最终以双方对抗的胜负来评估相关人工智能算法在作战决策中的影响程度。

军事对抗面临的突出问题是规则不完备、信息不完全、响应高实时等，而强化学习（RL）中的智能体在与环境交互过程中可以不断试错，以最大化累积回报作为目标，不断探索最优策略，展现出强大的决策力，这为解决上述问题提供了新的有效途径。例如，利用残差网络 (ResNets) 与蒙特卡罗搜索树 (MCTS)[5] 相结合的方式建立决策模型，采用单智能体对回合作战进行决策，使兵棋推演系统具备智能决策能力。然而，机器学习方法一般普遍存在的问题是模型过拟合、泛化能力差；因此，基于深度强化学习 (DRL) 的智能决策框架成为了研究热点。利用深度学习（DL）算法分析处理战场感知数据，有助于指挥员迅速辨别战场态势；利用 RL 算法进行辅助决策，有利于提高指挥员的谋略水平，夺取竞争性优势[6]。未来战争的作战模式是“以快吃慢”，决策模型的策略输出速度同样不容忽视。压缩网络结构（例如采用参数修剪和低秩分解等方法[7]）将深度网络变为轻量级网络，以满足实际作战响应高实时的特点。

目前，针对决策空间巨大、信息不完全的智能策略对抗技术尚未完全取得突破，基于 DRL 的理论与方法仍处于起步阶段。面对作战需求，决策优势是核心，其将成为现代战争制胜的关键，在 OODA（观察、判断、决策、行动）环路中，决策还是制约循环速度的瓶颈。鉴此，本文将主要从智能决策模型构建角度，研究将深度神经网络（deep neural network ）与 SAC（soft actor-critic）强化学习算法应用于作战仿真推演系统的途径，以及应用仿真推演平台，以舰载直升机反潜想定为例，验证决策模型的有效性和相关算法的适用性。

1 SAC 算法

SAC 算法[8] 包含 3 个关键要素：1）利用最大熵框架增强模型的稳定性，并提高智能体的探索能力；2）采用离线策略更新，重复利用先前收集的数据，提高效率；3）基于 actor- critic (AC) 体系结构，其策略网络与价值网络相互独立，将策略称为 Actor，价值函数称为 Critic。

1.1 最大熵强化学习

马尔科夫决策过程（Markov decision process, MDP）是强化学习问题在数学上的理想化形式[9]。用五元组(S,A,P,r,γ)定义无限视界 MDP，其中：S 为状态空间，A 为动作决策空间，二者的空间连续；$P:S\times S\times A\to[0,\infty)$，为状态转移概率，表示在给定当前时刻的状态$s_t\in S$和探索动作$a_t\in A$时，转移到下一时刻状态$s_{t+1}\in S$的概率密度；$\gamma$为折扣因子，表示每个时刻获得的收益对总回报的影响程度；r为每次状态转换环境给予的有界回报，$r:S\times A\to[r_{\min},r_{\max}]$。

智能体的目标是学习一种策略$\pi:S\to A$，使累积期望回报值最大化，如式（1）所示。

$$\pi_{\text{std}}^{*}=\arg\max_{\pi}\sum_{t}E_{(s_t,a_t)\sim\rho_\pi}[r(s_t,a_t)] \tag{1}$$

式中，ρ_π为策略生成的轨迹$(s_t,a_t,s_{t+1},a_{t+1},s_{t+2},a_{t+2},\cdots)$的边缘分布。

如式（2）所示，最大熵强化学习是在式（1）的基础上增加了可调整熵值项 H，智能体的目标变

为寻找令累积期望回报和熵值同时最大化的最优策略。信息熵越大, 说明分布越均匀。最大化信息熵, 有利于增加模型的探索能力。

$$\pi^{*}_{\mathrm{MaxEnt}} = \arg\max_{\pi} \sum_{t} E_{(s_t,a_t)\sim\rho_\pi}[r(s_t,a_t)+\alpha H(\pi(\cdot|s_t))] \tag{2}$$

式中, α为温度参数, $\alpha \to 0$时, 式(2)等同于式(1)。

1.2　离线策略更新

离线策略更新采用两个策略, 一个策略用于智能体学习, 并最终成为最优策略, 称为目标策略π_{tar}; 另一个策略用于生成智能体轨迹样本, 称为行动策略π_{act}。此时, 由于智能体用于学习的数据与待学习的目标策略相互分离, 因此离线策略更新一般方差大且收敛慢。但是, 这种分离的优点在于, 当行动策略对所有可能的动作继续进行采样时, 可以采用确定性的目标策略。

在处理预测问题时, 目标策略和行动策略固定, 旨在学习状态价值函数$\hat{v} \approx v_\pi$(给定策略π的状态价值函数)或动作价值函数$\hat{q} \approx q_\pi$(给定策略π的动作价值函数)。对于控制问题, 两个策略在智能体学习过程中会不断变化, 目标策略π_{tar}逐渐变成关于$\hat{q}$的贪心策略, 而行动策略π_{act}逐渐变成关于$\hat{q}$的某种探索性策略。

1.3　AC 体系

AC 体系采用时序差分(TD)的方法, 使用独立的模型估计状态−动作序列的长期回报, 而非直接采用真实回报。如图 1 所示, 策略网络被称为 Actor, 用于动作选择; 价值网络称为 Critic, 用于评估动作的优劣。采用式(3)所示的 TD 形式来评估最新选择的探索动作a_t, 其中V为 Critic 的状态价值。若 TD 误差(即δ_t)为正, 则应加强选择探索动作a_t的趋势; 若 TD 误差为负, 则应降低选择探索动作a_t的频率。

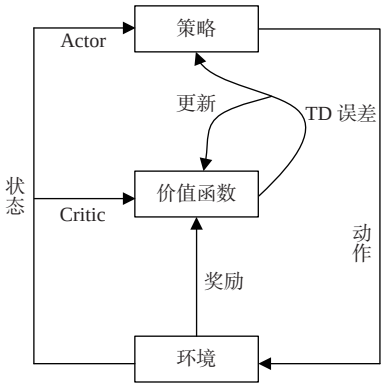

图 1　AC 体系架构图[10]

Fig. 1　The architecture of Actor-Critic[10]

$$\delta_t = r_{t+1} + \gamma V(s_{t+1}) - V(s_t) \tag{3}$$

式中, r_{t+1}为下一时刻环境给予的回报值。

如图 2 所示, 采用离线策略更新的 AC 体系由在线评估网络和目标网络构成, 网络结构和初始化参数均相同。首先, 提取经验缓存数据, 通过目标网络得到目标回报值; 然后, 根据 TD 误差更新评估网络中的 Critic 网络; 最后, 更新评估网络中的 Actor 网络, 其中动作的探索和 Actor 网络的更新分别采用不同策略。

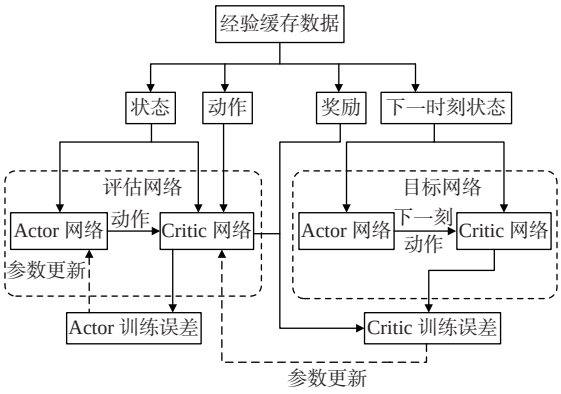

图 2　采用离线策略更新的 AC 体系

Fig. 2　AC architecture updated by off-policy

2　仿真推演任务分析

2.1　反潜想定介绍

导弹驱逐舰搭载直升机反潜想定设定的场景如下: 蓝方海军有 2 艘潜艇停泊在某海域, 红方海军得到情报, 派遣舰载反潜直升机(以下称反潜机)前往搜寻蓝方海军的潜艇。红方海军反潜机的作战目标是通过在目标水域投放声呐浮标, 发现并摧毁蓝方潜艇, 其水面舰船主要为反潜机提供鱼雷和反潜火箭弹等武器支援。而蓝方海军潜艇的作战目标是隐藏航迹, 避免被消灭。红蓝双方海军的兵力编成如表 1 和表 2所示。

2.2　声呐浮标建模

声呐浮标系统按直线通路探测、水面反射探测和汇聚区探测依次进行判断。以 9.874 73 n mile 为直线探测距离基数, 根据敌我深度 d、目标所处海底地形 m 和目标距海底高度 h, 修正得到直线探测距离R_{D}, 如式(4)所示。

$$R_{\mathrm{D}} = 9.874\,73 f_d f_m f_h \tag{4}$$

式中, 以声呐最大理论探测距离为基数, 根据 d, m, h 和探测方速度 v 及目标信号特征强度T_{s}, 修正得到汇聚区的有效探测范围R_{C}, 如(5)式所示。

表 1 红方海军兵力编成

Table 1 The navy strength of red team

单元类型及名称	航速/（km·h⁻¹）	位置	数量/艘	单元配备的主要武器
MH-60R“海鹰”反潜机	259.28	(34°13'9" E，43°48'37" N)	1	Mk-54轻型鱼雷 ×2 AN/SSQ-62E定向指令主动声呐浮标系统（DICASS系统）×8 AN/SSQ-53F 定向频率分析和记录被动声呐浮标（DIFAR系统）×1
“阿利·伯克”级Flight IIA导弹驱逐舰	0	(33°50'15" E，43°26'30" N)	1	Mk-54轻型鱼雷 ×40 RUM-139C VLA反潜火箭弹 ×8

表 2 蓝方海军兵力编成

Table 2 The navy strength of blue team

单元类型及名称	航速/（km·h⁻¹）	位置	数量/艘	单元配备的主要武器
955A“北风之神”级战略核潜艇	0	(34°65'28" E，43°4'36" N)	1	SS-N-15“海星”反潜导弹×2；USET-80K鱼雷 ×14
21310型“鱼鳃-NN”级常规潜艇	0	(33°84'74" E，43°73'80" N)	1	高性能炸药 ×6

$$R_C = R_{\max} f_d f_m f_h f_v f_{T_s} \tag{5}$$

式中，$R_{\max}$为声呐的最大理论探测距离。通常，直线探测距离远大于汇聚区探测范围，即$R_D \gg R_C$。式(4)与式(5)中f均为与变量相关的修正系数。

对于水面反射，如图 3 所示，设探测点和目标点距海底高度分别为 a 和 b，计算声波通过水面反射的位置，该位置与探测点的水平距离为 w，探测点与目标点的水平距离为 c，根据式(6)计算 w。以探测点所在海面位置处为原点，向目标方向以距离 w 确定反射点P_0。

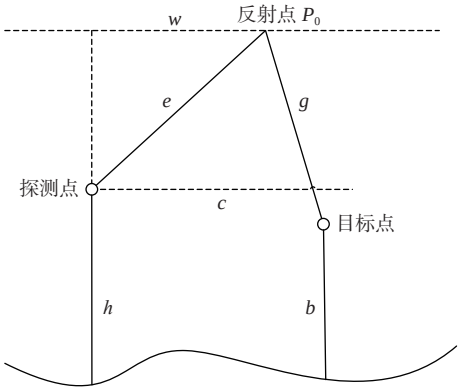

图 3 水面反射探测

Fig. 3 Water surface reflection detection

$$w = \frac{hc}{h+b} \tag{6}$$

计算探测点经反射点到目标点的直线累加距离 D_h:

$$D_h = e + g \tag{7}$$

式中：e 为反射点与探测点的直线距离；g 为反射点与目标点的直线距离。

若修正系数f小于 1，则由式(8)对R_C进行修正：

$$R'_C = R_C\left(1 + \frac{f}{2}\right) \tag{8}$$

比较修正后的探测距离R'_C与反射直线累加距离 D_h。若$D_h < R'_C$，则表明经水面反射有可能探测到目标；若$D_h > R'_C$，则判断为水面反射不可探测。

2.3 红方反潜机 MDP 建模

针对仿真推演想定中的红方作战单元进行深度强化学习训练，使红方反潜机在与环境反复交互中学会自主决策，最终自动摧毁蓝方潜艇。采用五元组(S,A,P_a,J,γ)表示 MDP，其中，S 为实时获取的红方反潜机状态空间，A 为红方反潜机的动作决策空间，$P_a(s',s)$为红方反潜机在状态 s 和动作a下进入下一个状态s'的概率，J为红方反潜机的优化目标，折扣因子$\gamma \in (0,1)$。

1）状态空间设置。

红方反潜机返回的态势数据约有一百多个维度，为了确保模型收敛，选择了影响任务成败的关键因素（例如经、纬度和航向）作为状态信息要素，亦即 S=[经度，纬度，航向]。如图 4 所示，利用经、纬度坐标信息表示反潜机当前的位置，航向β表示红方反潜机当前的飞行方向与蓝方目标潜艇的偏差角度，v_x和v_y分别表示反潜机在 x 和 y 方向的速度。

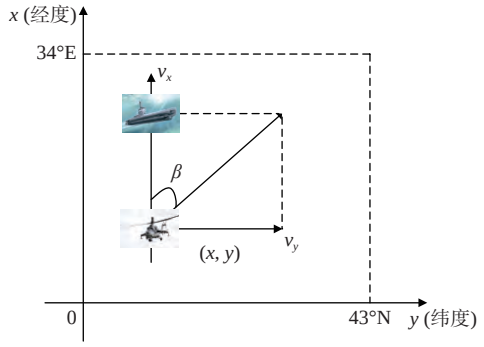

图 4 红方反潜直升机状态分析

Fig. 4 State analysis of red team's ASW helicopter

2) 动作空间设置。

红方反潜机可以执行的动作也有几十种, 动作空间选择最能够影响任务成败的动作。例如, 反潜机的航向以及是否投放声呐探测潜艇, 亦即 A=[航向, 是否投放声呐]。

3) 优化目标设置。

如式(9)所示[8], 红方反潜机的优化目标 J 是同时最大化回报函数 r 和策略的期望熵 H。式中: T 为视界长度; α 为温度参数, 它决定了熵项对回报的相对重要性, 从而控制了最优策略的随机性[11]。与传统的优化累积回报值相比, 增加 H 项将鼓励红方进行更广泛的探索, 放弃无法获得最终胜利的规划路径, 与此同时, 还将显著提高红方反潜机的学习速度。

$$J(\pi)=\sum_{t=0}^{T}E_{(s_t,a_t)\sim\rho_\pi}[r(s_t,a_t)+\alpha H(\pi(\cdot|s_t))] \tag{9}$$

回报函数 r 设置如下: 红方靠近目标区域(以经纬度坐标(34°17'55" E, 43°48'74" N)为圆心, 半径为 5 km 的范围)会获得正回报, 偏离目标区域则获得负回报, 偏离目标越近或越远获得的回报绝对值越大。具体计算如式(10)所示, 其中 η 表示反潜机目标朝向与当前朝向的绝对偏差。

当红方反潜机与蓝方潜艇间距离小于目标区域半径(5 km)时红方反潜机投放声呐。若在当前态势中红方捕获的蓝方标识符(GUID)与目标潜艇相同, 则判定红方反潜机发现了蓝方潜艇。为缓解强化学习中的稀疏奖励问题, 红方反潜机投放声呐和发现潜艇都将会获得 10 分的回报值。

推演系统的输入是神经网络推算的作战单元动作列表, 依此判断红蓝双方作战单元是否被击落。若未被击落(即状态量非全为 0)且执行的动作存在, 则该作战单元的动作将会被执行; 若红方反潜机被消灭, 将获得 −100 分回报值, 而击毁蓝方潜艇会获得 150 分回报值。

$$r_t=\begin{cases}+10, & \text{if drop sonobuoy}\\ -100, & \text{if aircraft not exist}\\ +10, & \text{if find target}\\ +150, & \text{if target not exist}\\ (10\,000*\cos(\eta))/\lambda, & \text{if cos\_value}\geqslant 0\\ (\lambda*\cos(\eta))/10\,000, & \text{if cos\_value}<0\end{cases} \tag{10}$$

3　基于 SAC 的决策模型构建

3.1　红方决策网络构建

在经典 AC 体系的基础上增加价值网络, 构建红方反潜智能体的决策网络。如图 5 所示, 其主要包括价值网络、策略网络和软 Q 网络这 3 个部分, 分别引入参数化状态价值函数 $V_\psi(s_t)$、策略 $\pi_\phi(a_t|s_t)$ 以及软 Q 函数 $Q_\theta(s_t,a_t)$ 进行描述。3 个网络单元分别具有相同结构和参数的目标网络以及在线网络, 利用经验缓存数据 (s_t,a_t,r_t,s_{t+1}) 对目标网络进行离线策略训练, 通过计算损失函数 L 并求解梯度(价值网络梯度 $\nabla_\psi J_V(\psi)$、策略网络梯度 $\nabla_\phi J_\pi(\phi)$、软 Q 网络梯度 $\nabla_\theta J_Q(\theta)$), 循环更新在线网络参数, 不仅可以降低样本之间的相关性, 而且

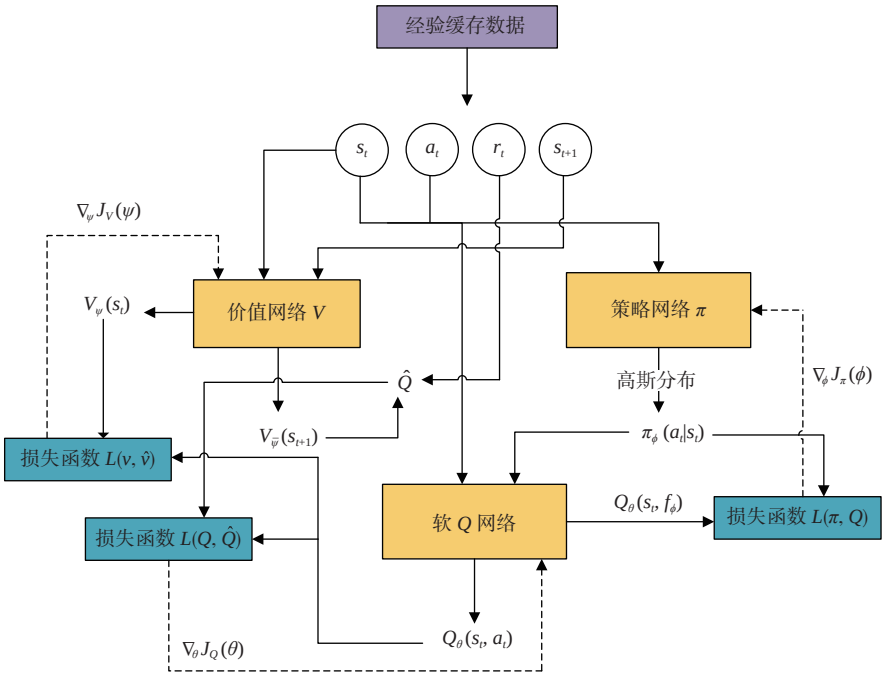

图 5　基于 SAC 的红方直升机决策网络

Fig. 5　Decision-making network of red team's helicopter based on SAC

针对仿真推演此类大规模连续域问题，还可有效提高数据利用率。同时，对目标网络参数进行软更新，可稳定整个训练过程[12]。

引入独立的价值网络可以使红方反潜智能体的训练更加稳定且易于与其他网络同步训练。将式(9)中的熵值项 H 按照期望的形式展开，将其作为价值网络更新的目标之一，得到式(11)所示的状态价值函数。如图 6 所示，价值网络共有 4 层，前 3 层为隐藏层，隐藏单元个数均为 256；神经网络的最后一层是单维度的输出层，输出相应的状态价值 V。隐藏层神经元使用 Relu激活函数，以提升价值网络的非线性建模能力。

$$V(s_t)=E_{a_t\sim\pi}[Q(s_t,a_t)-\log\pi(a_t|s_t)] \tag{11}$$

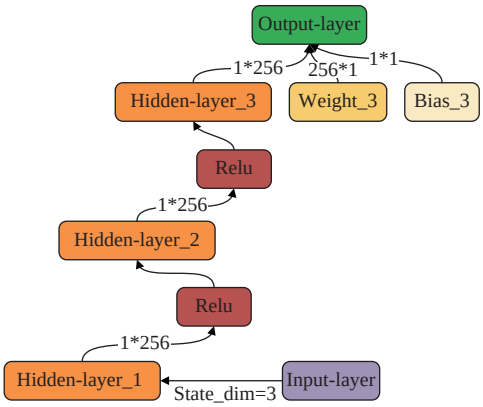

图 6 价值网络结构

Fig. 6 The structure of value network

如图 7 所示，策略网络共有 5 层，前 3 层为隐藏层，隐藏单元个数分别为 256, 128, 64；第 4 层神经网络对前一层的输出信息进行标准化处理，输出包含均值和标准差的高斯分布；策略网络的最后一层是 2 维度的输出层，根据输入状态输出相应的动作。策略网络生成的策略经高斯分布输出后(即 $\pi_\phi(a_t|s_t)$)，将为不同动作分配不同的概率值，这有利于红方智能体探索，提高长期回报；策略网络参数 ϕ 可直接通过最小化 KL 散度进行更新，如式(12)所示。通过重参数化，得到红方智能体要执行的随机动作，并获得新的软 Q 函数值 $Q_\theta(s_t,f_\phi)$，保证目标函数可微及可进行梯度更新。

$$J_\pi(\phi)=E_{s_t\sim D}\left[D_{\mathrm{KL}}\left(\pi_\phi(\cdot|s_t)\left\|\frac{\exp(Q_\theta(s_t,\cdot))}{Z_\theta(s_t)}\right.\right)\right] \tag{12}$$

$$a_t=f_\phi(\varepsilon_t;s_t) \tag{13}$$

式中：$Z_\theta(s_t)$为配分函数，使分布标准化；D_{KL} 为 KL 距离；$f_\phi(\varepsilon_t;s_t)$为经神经网络变换后的重参数化策略。

将式(13)的动作函数代入式(12)，策略网络

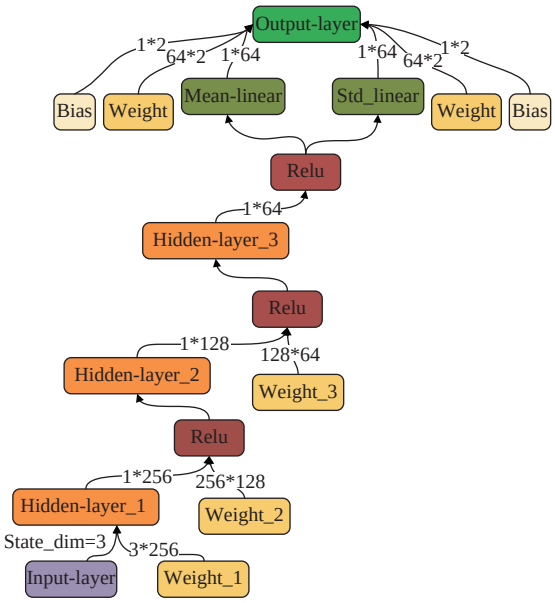

图 7 策略网络结构

Fig. 7 The structure of policy network

的目标表达式可写为

$$J_\pi(\phi)=E_{s_t\sim D,\varepsilon_t\sim N}[\log\pi_\phi(f_\phi(\varepsilon_t;s_t)|s_t)-Q_\theta(s_t,f_\phi(\varepsilon_t;s_t))] \tag{14}$$

式中：E 表示期望函数；ε_t为输入的噪声；N 表示噪声呈高斯分布。

软 Q 网络交替进行策略评估和策略改进环节，使红方智能体获得更优策略，增强决策水平。在策略评估步骤，根据式(9)中的最大熵目标计算智能体采用当前策略π的价值。对于固定策略，可以从任何函数 $Q:S\times A\to R$开始(R 表示回报集)，利用修正贝尔曼辅助算子T^π[13]及根据软贝尔曼方程[14]，迭代计算动作价值函数 Q，如式(15)所示。在动作集 A 有界的情况下，迭代次数 $k\to\infty$时，序列Q^k将收敛到Q_π。

$$T^\pi Q(s_t,a_t)\triangleq r(s_t,a_t)+\gamma E_{s_{t+1}\sim p}[V(s_{t+1})] \tag{15}$$

式中，p 为状态转移概率。

策略改进步骤中，策略集Π中策略π'朝与新 Q 函数的指数分布呈正比的方向离线更新。首先，每个状态根据式(16)更新红方智能体策略，以保证新策略优于旧策略，即$Q^{\pi_{\mathrm{new}}}(s_t,a_t)\geqslant Q^{\pi_{\mathrm{old}}}(s_t,a_t)$。然后，以最小化 KL 散度形式减小 2 个分布间的差异，其中$Z^{\pi_{\mathrm{old}}}(s_t)$为对 Q 值进行归一化分布。最后，经策略迭代，红方智能体找到最优策略π^*。

$$\pi_{\mathrm{new}}=\arg\min_{\pi'\in\Pi}D_{\mathrm{KL}}\left(\pi'(\cdot|s_t)\left\|\frac{\exp(Q^{\pi_{\mathrm{old}}}(s_t,\cdot))}{Z^{\pi_{\mathrm{old}}}(s_t)}\right.\right) \tag{16}$$

如图 8 所示，软 Q 网络有状态和动作 2 个输入维度，输入状态经过 4 层隐藏层，隐藏单元数分别为(256, 128, 256, 128)，输入动作经过 3 层隐藏

层, 隐藏单元数分别为(128, 256, 128)。进入第 3 层隐藏层前, 将输入状态和动作 2 个分支的输出结果合并, 经过最后单维度的输出层, 输出动作状态价值 Q。

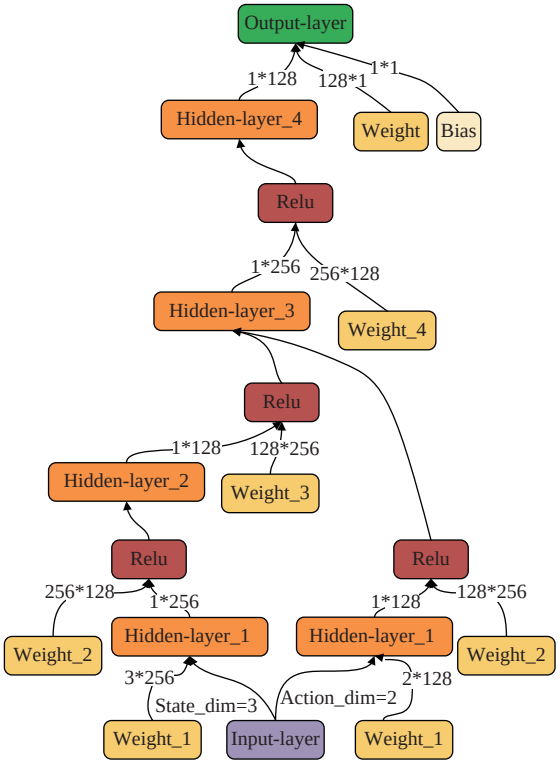

图 8　Soft Q 网络结构

Fig. 8　The structure of soft Q network

3.2　红方智能体探索与利用的平衡

若红方智能体单纯执行当前策略选择回报值最大的动作, 则会因探索不充分而陷入局部最优困境。为了尽快发现蓝方目标以及提高长期回报, 在红方智能体训练过程中, 要扩大对动作状态空间的探索。鉴于 OU 随机过程(Ornstein-Uhlenbeck process)时序相关性较好, 故在策略网络输出确定性动作后引入 OU 噪声[15], 将策略随机化, 再从当前策略中对动作值进行采样, 得到探索动作a_t, 如图 9 所示。

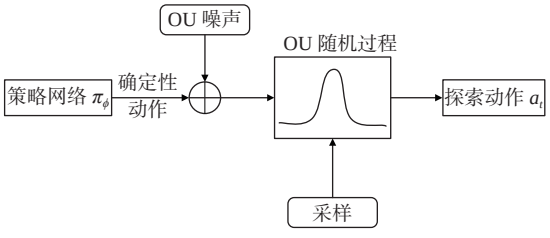

图 9　红方智能体的探索动作选择

Fig. 9　The action choice of red team agent for exploration

3.3　算法设计

算法伪代码包括如下步骤:

步骤 1: 初始化价值网络参数ψ、策略网络参数ϕ以及软 Q 网络参数θ。

步骤 2: 初始化目标网络参数$\bar{\psi} \leftarrow \psi$, $\bar{\phi} \leftarrow \phi$, $\bar{\theta} \leftarrow \theta$。

步骤 3: 初始化经验缓存空间 D 和标志位 done。

步骤 4: 对每个 episode 重复执行如下操作。

1) 获取环境状态s_0。

2) 若红方反潜机存在(未被蓝方潜艇摧毁或未摧毁蓝方潜艇结束推演)并且执行步数小于最大规定步数, 循环执行如下操作:

(1) 将当前状态s_t输入策略网络, 选择动作并加入 OU 噪声, 即$a_t \sim \pi_\phi(s_t)+N$;

(2) 执行动作a_t, 获得回报r_t并进入下一状态 s_{t+1};

(3) 将(s_t, a_t, r_t, s_{t+1})存入经验缓存空间 D;

(4) 若数据量大于经验回放缓存空间容量 N_s, 则从 D 中采集 N_s 个样本$(s_t^i, a_t^i, r_t^i, s_{t+1}^i)$, $i = 1, 2, \cdots, N_s$;

(5) 根据目标价值函数$V_{\bar{\psi}}$, 计算软 Q 目标网络的期望回报值$\hat{Q}$:

$$\hat{Q}(s_t, a_t) = r(s_t, a_t) + \gamma E_{s_{t+1}\sim p}[V_{\bar{\psi}}(s_{t+1})]$$

(6) 计算价值网络的损失函数$J_V(\psi)$, 计算梯度$\hat{\nabla}_\psi J_V(\psi)$, 更新价值网络的所有参数$\psi$, 其中

$$J_V(\psi) = E_{s_t\sim D}\left[\frac{1}{2}(V_\psi(s_t) - E_{a_t\sim\pi_\phi}[Q_\theta(s_t, a_t) - \log\pi_\phi(a_t|s_t)])^2\right]$$

(7) 计算软 Q 网络的损失函数$J_Q(\theta)$, 计算梯度$\hat{\nabla}_\theta J_Q(\theta)$, 更新软 Q 网络的所有参数θ, 其中

$$J_Q(\theta) = E_{(s_t,a_t)\sim D}\left[\frac{1}{2}(Q_\theta(s_t, a_t) - \hat{Q}(s_t, a_t))^2\right]$$

(8) 计算策略网络的损失函数$J_\pi(\phi)$, 计算梯度$\hat{\nabla}_\phi J_\pi(\phi)$, 更新策略网络的所有参数$\phi$;

(9) 软更新目标价值网络参数$\bar{\psi}$、目标软 Q 网络参数$\bar{\theta}$、目标策略网络参数$\bar{\phi}$:

$$\bar{\psi} \leftarrow \tau\psi + (1-\tau)\bar{\psi}, \bar{\theta} \leftarrow \tau\theta + (1-\tau)\bar{\theta}, \bar{\phi} \leftarrow \tau\phi + (1-\tau)\bar{\phi}$$

3)若蓝方潜艇被消灭(done=1)或者达到最大的规定步数, 结束 2) 循环。

步骤 5: 若达到最大规定的 episode 数, 结束整个循环。

4　作战仿真推演实验及分析

4.1　作战仿真推演环境

利用作战仿真推演平台验证决策模型的适用性, 其整体框架如图 10 所示。作战仿真推演系统具备数据管理、指挥控制、效能评估等功能。人工智能研究平台包括 Python 开发包和 AI 处理模

块两部分，与作战仿真推演系统协同配合，实现具体研究案例的智能学习。其中，人工智能研究平台的具体程序开发环境和相关接口如图 11 所示，主要由 AI 业务界面、仿真交互接口、命令池、态势池、态势适配器、算法库、通信界面组成。利用 PyTorch 框架编写 SAC 决策算法并存储到算法库，在 AI 业务界面中创建智能体对象和环境对象，调用算法库中的决策算法，将仿真操作命令传递给服务端执行，实时接收通信界面返回的服务端态势信息，使反潜智能体在与环境信息的不断交互中学会自主决策。

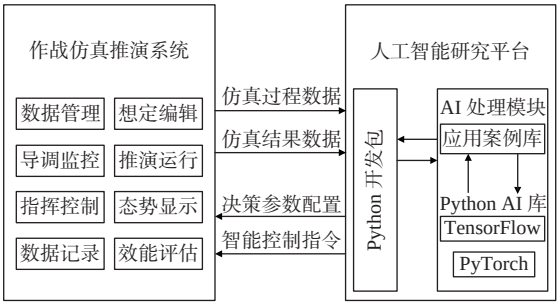

图 10 仿真推演平台总体框架图

Fig. 10 Overall framework of simulation platform

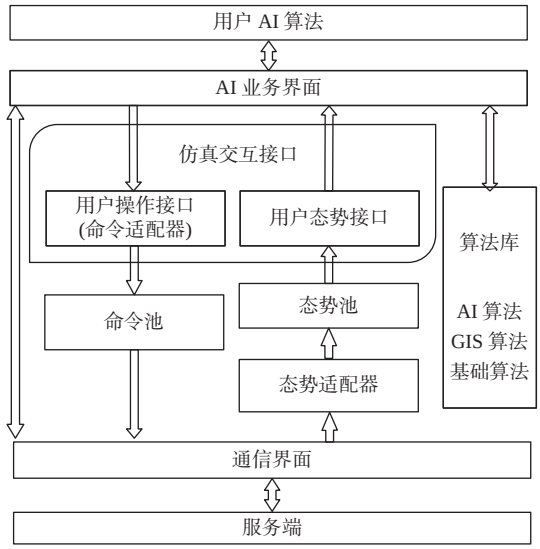

图 11 AI 开发平台总体框架图

Fig. 11 Overall framework of AI development platform

4.2 推演结果及分析

在 4.1 节介绍的仿真实验环境下开展仿真推演，验证训练的红方智能体能否作出智能决策。如图 12 所示，在红方反潜智能体训练开始阶段，获得的回报值相对较低，每轮回报值的波动性较大。随着迭代 episode 数的增加，回报值相应上升，迭代进行到 600 轮左右时，回报值逐渐趋于平稳。相应地，在训练初期，红方反潜机不断投放声呐，四处搜寻蓝方潜艇，随机性较强，经过不断的训练学习，其发现潜艇所用的时间逐渐减少，最终实现自主决策，找到目标后自动击毁。具体训练超参数设置如表 3 所示。

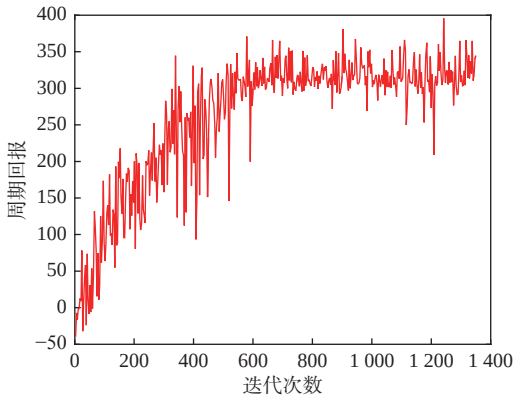

图 12 红方智能体训练过程

Fig. 12 The training process of red team agent

表 3 超参数设置

Table 3 The hyper-parameter settings

参数	数值
学习率	0.001
折扣因子γ	0.99
软更新率τ	0.001
温度参数α	1.00
经验回放缓存空间容量N_s	1 000 000
每批训练样本数/个	128
最大训练episode数/个	5 000
每个episode的最大训练步数	30

图 13 所示为“海鹰”反潜机摧毁“鱼鳃-NN”级常规潜艇的过程。图中，左上角是“鱼鳃-NN”级常规潜艇，左下角是“阿利·伯克”级导弹驱逐舰，中间是红方反潜机为搜索蓝方潜艇而布放的声呐浮标。图 14 所示为对应的航迹图。图中，“海鹰”反潜机从 (34°13'E, 43°48'N) 位置处出发，不断投放声呐来搜寻蓝方潜艇，最终在目标位置处 (33°85'E, 43°74'N) 摧毁“鱼鳃-NN”级常规潜艇。

图 15 所示为“海鹰”反潜机摧毁“北风之神”级战略核潜艇的过程。图中，右下角是“北风之

图 13 红方击毁蓝方常规潜艇

Fig. 13 Red team destroys blue team's conventional submarines

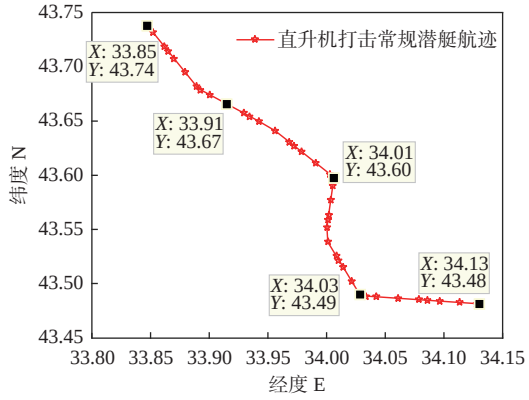

图 14　红方击毁蓝方常规潜艇航迹图

Fig. 14　Track map of red team destroying blue team's conventional submarine

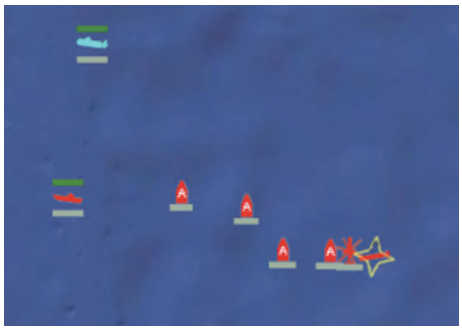

图 15　红方击毁蓝方核潜艇

Fig. 15　Red team destroys blue team's nuclear submarine

神”级战略核潜艇级，左下角是“阿利·伯克”级导弹驱逐舰。图 16 为对应的航迹图，图中，“海鹰”反潜作战机从 (34°13'E, 43°48'N) 位置处出发，不断投放声呐搜寻蓝方潜艇，最终在目标位置处 (34°65'E, 43°4'N) 摧毁“北风之神”级战略核潜艇。

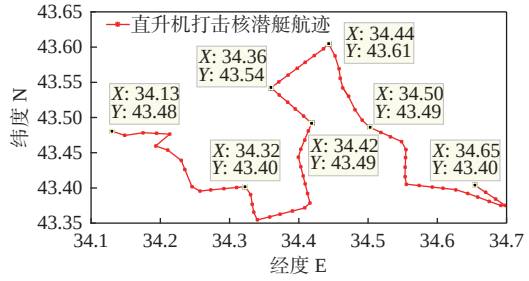

图 16　红方击毁蓝方核潜艇航迹图

Fig. 16　Track map of red team destroying blue team's nuclear submarine

将改进 SAC 决策算法与 DDPG 决策算法对比分析，结果如图 17 所示。由图可见：在训练开始阶段，两种算法的平均回报均较快增长。其中，DDPG 算法训练的智能体采用确定性策略作决策，经迭代 300 次左右开始收敛，回报值趋于稳定；而改进 SAC 算法采用的是探索性策略，收敛相对较慢，随着迭代次数的增加，其平均回报明显高于 DDPG 算法。此外，在两种算法训练的智能体性能趋于稳定后，各自随机进行了 60 次仿真推演实验，其中每 6 次实验划分为一组，共计 10 组实验。基于此，对 10 组获胜概率进行比较分析，结果如图 18 所示。由图可见，采用 DDPG 算法，红方反潜智能体的平均获胜概率为 49.32%；采用改进 SAC 决策算法，红方平均获胜概率为 73.85%，比前者提高近 24.53%。

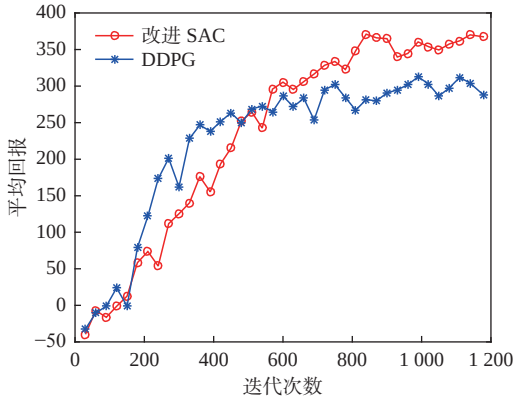

图 17　两种决策算法的平均回报对比

Fig. 17　Comparison of average return of two decision algorithms

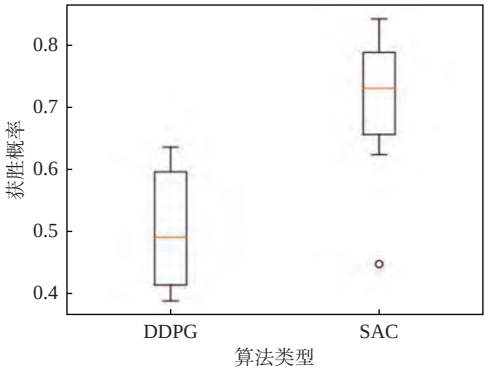

图 18　红方获胜概率对比

Fig. 18　Comparison of winning probability of red team

5　结　语

本文针对作战仿真推演中的决策问题，提出了一种基于 SAC 算法的智能决策模型，使红方智能体在不具备先验知识的情况下，通过与环境不断交互，学会发现蓝方目标，自主决策，最终获得胜利。基于舰载反潜机反潜想定进行仿真推演，结果表明，本文所构建的决策模型在利用经验缓存数据开展离线策略训练后，展现出了较强的学习和判断能力，红方反潜智能体成功探测到了蓝方潜艇并进行了摧毁，从而证明了智能决策模型的有效性。在未来研究中，可以增加智能体数量，通过分析多智能体博弈特点来改进算法，以实现基于多智能体 DRL 的作战推演智能决策。

参考文献：

[1]　胡荟, 吴振齐. 人工智能技术在美国军事情报工作中的

当前应用及发展趋势探析 [J]. 国防科技, 2020, 41(2): 15–20.
HU H, WU Z Q. Research on the current application and development trend of artificial intelligence technology in US military intelligence work[J]. National Defense Science & Technology, 2020, 41(2): 15–20 (in Chinese).

[2] 付长军, 郑伟明, 葛蕾, 等. 人工智能在作战仿真中的应用研究 [J]. 无线电工程, 2020, 50(4): 257–261.
FU C J, ZHENG W M, GE L, et al. Application of artificial intelligence in combat simulation[J]. Radio Engineering, 2020, 50(4): 257–261 (in Chinese).

[3] 孙鹏, 谭玉玺, 李路遥. 基于态势描述的陆军作战仿真外部决策模型研究 [J]. 指挥控制与仿真, 2016, 38(2): 15–19.
SUN P, TAN Y X, LI L Y. Research on external decision model of army operational simulation based on situation description[J]. Command Control & Simulation, 2016, 38(2): 15–19 (in Chinese).

[4] 董倩, 纪梦琪, 朱一凡, 等. 空中作战决策行为树建模与仿真 [J]. 指挥控制与仿真, 2019, 41(1): 12–19.
DONG Q, JI M Q, ZHU Y F, et al. Behavioral tree modeling and simulation for air operations decision[J]. Command Control & Simulation, 2019, 41(1): 12–19 (in Chinese).

[5] 彭希璐, 王记坤, 张昶, 等. 面向智能决策的兵棋推演技术 [C]//2019 第七届中国指挥控制大会论文集. 北京: 中国指挥与控制学会, 2019: 193–198.
PENG X L, WANG J K, ZHANG C, et al. The technology of wargame based on intelligent decision[C]//Proceedings of the 7th China Command and Control Conference in 2019. Beijing: Chinese Institute of Command and Control, 2019: 193–198 (in Chinese).

[6] 廖馨, 孙峥皓. 作战推演仿真中的智能决策技术应用探索 [C]//第二十届中国系统仿真技术及其应用学术年会论文集. 乌鲁木齐: 中国自动化学会系统仿真专业委员会, 2019: 368–374.
LIAO X, SUN Z H. Exploration on application of intelligent decision-making in battle deduction simulation[C]//Proceedings of the 20th China Annual Conference on System Simulation Technology and its Application. Urumqi: System Simulation Committee of China Automation Society, 2019: 368–374 (in Chinese).

[7] 崔文华, 李东, 唐宇波, 等. 基于深度强化学习的兵棋推演决策方法框架 [J]. 国防科技, 2020, 41(2): 113–121.
CUI W H, LI D, TANG Y B, et al. Framework of wargaming decision-making methods based on deep reinforcement learning[J]. National Defense Science & Technology, 2020, 41(2): 113–121 (in Chinese).

[8] HAARNOJA T, ZHOU A, ABBEEL P, et al. Soft actor-critic: off-policy maximum entropy deep reinforcement learning with a stochastic actor[C]//Proceedings of the 35th International Conference on Machine Learning. Stockholm, Sweden: ACM Press, 2018.

[9] SUTTON R S, BARTO A G. Reinforcement Learning: An Introduction[M]. Cambridge: MIT Press, 1998.

[10] SPIELBERG S, GOPALUNI R, LOEWEN P. Deep reinforcement learning approaches for process control[C]// 2017 6th International Symposium on Advanced Control of Industrial Processes, [S. 1.]: IEEE, 2017: 201–203.

[11] HAARNOJA T, ZHOU A, HARTIKAINEN K, et al. Soft actor-critic algorithms and applications [EB/OL]. ArXiv:1812.05905,2018(2018-12-13)[2020-08-30].https://arxiv.org/abs/1812.05905.

[12] MNIH V, KAVUKCUOGLU K, SILVER D, et al. Human-level control through deep reinforcement learning[J]. Nature, 2015, 518(7540): 529–533.

[13] SCHULMAN J, CHEN X, ABBEEL P. Equivalence between policy gradients and soft Q-learning[EB/OL]. ArXiv: 1704.06440, 2017. (2017-4-21)[2020-08-30]. https://arxiv.org/pdf/1704.06440.pdf.

[14] HAARNOJA T, TANG H, ABBEEL P, et al, Reinforcement learning with deep energy-based policies[C]//Proceedings of the 34th International Conference on Machine Learning. Sydney, Australia: ACM Press: MLR. org, 2017: 1352–1361.

[15] LILLICRAP T P, HUNT J J, PRITZEL A, et al. Continuous control with deep reinforcement learning[C]// Proceedings of the 4th International Conference on Learning Representations. San Juan, Puerto Rico: Elsevier, 2016.