基于距离信息的追逃策略: 信念状态连续随机博弈

陈灵敏 [1] 冯 宇 [1] 李永强 [1]

摘 要 追逃问题的研究在对抗、追踪以及搜查等领域极具现实意义. 借助连续随机博弈与马尔科夫决策过程 (Markov decision process, MDP), 研究使用测量距离求解多对一追逃问题的最优策略. 在此追逃问题中, 追捕群体仅领导者可测量与逃逸者间的相对距离, 而逃逸者具有全局视野. 追逃策略求解被分为追博弈与马尔科夫决策两个过程. 在求解追捕策略时, 通过分割环境引入信念区域状态以估计逃逸者位置, 同时使用测量距离对信念区域状态进行修正, 构建起基于信念区域状态的连续随机追博弈, 并借助不动点定理证明了博弈平稳纳什均衡策略的存在性. 在求解逃逸策略时, 逃逸者根据全局信息建立混合状态下的马尔科夫决策过程及相应的最优贝尔曼方程. 同时给出了基于强化学习的平稳追逃策略求解算法, 并通过案例验证了该算法的有效性.

关键词 追逃问题, 信念区域状态, 连续随机博弈, 马尔科夫决策过程, 强化学习

引用格式 陈灵敏, 冯宇, 李永强. 基于距离信息的追逃策略: 信念状态连续随机博弈. 自动化学报, 2024, **50**(4): 828−840

DOI 10.16383/j.aas.c230018

Distance Information Based Pursuit-evasion Strategy: Continuous Stochastic Game With Belief State

CHEN Ling-Min[1] FENG Yu[1] LI Yong-Qiang[1]

Abstract The pursuit-evasion problem is of great importance in the fields of confrontation, tracking and searching. In this paper, we are focused on the study of optimal strategies for solving the multi-pursuits and single-evader problem with only measured distances within the framework of continuous stochastic game and Markov decision process (MDP). In such problem, only the leader of pursuits can measure its relative distance with respect to the evader, while the evader has a global view. The strategies of the pursuits and evader are established via two steps: The pursuit game and the MDP. For the pursuits′ strategy, the belief region state is introduced by partitioning the environment to estimate the evader′s position, and the belief region state is further corrected by using the measured distances. A continuous stochastic pursuit game is then formed based on the belief region state, and the existence of stationary Nash equilibrium strategies is shown through the fixed-point theorem. For the evader′s strategy, an MDP with the global states is established and the underlying optimal Bellman equation is devised. Moreover, a reinforcement learning based algorithm is presented for stationary pursuit-evasion strategies computation, and an example is also included to exhibit the effectiveness of the current method.

Key words Pursuit-evasion problem, belief region state, continuous stochastic game, Markov decision process (MDP), reinforcement learning

Citation Chen Ling-Min, Feng Yu, Li Yong-Qiang. Distance information based pursuit-evasion strategy: Continuous stochastic game with belief state. *Acta Automatica Sinica*, 2024, **50**(4): 828−840

近年来, 追逃问题在飞行器、移动机器人等领域一直广受关注, 如无人机围捕搜查[1]、机器人协同对抗[2]、搜索救援[3] 等. 在典型追逃问题中追捕方试图快速捕获或逼近逃逸方, 而逃逸方则试图远离追捕方以避免被捕获. 自二十世纪六十年代提出一对一追逃问题以来[4], 学术界对其进行了充分探索[5−8], 并逐步演变为当下的多对一[9−11]、多对多[12−14] 对抗问题的研究.

追逃问题可视为智能体间的对抗与合作问题, 因此博弈论[15−17] 被广泛用于此类问题的求解[18−20]. 文献 [21] 在追逃双方具有无限视野下建立了线性二次型微分博弈模型, 将多追捕者与多逃逸者问题转化为多组两人零和微分博弈. 文献 [22] 基于非零和博弈框架, 研究了针对三种不同类型追捕者的追

收稿日期 2023-01-12 录用日期 2023-04-04
Manuscript received January 12, 2023; accepted April 4, 2023
国家自然科学基金 (61973276, 62073294), 浙江省自然科学基金 (LZ21F030003) 资助
Supported by National Natural Science Foundation of China (61973276, 62073294) and Natural Science Foundation of Zhejiang Province (LZ21F030003)
本文责任编委 杨涛
Recommended by Associate Editor YANG Tao
1. 浙江工业大学信息工程学院 杭州 313000
1. College of Information Engineering, Zhejiang University of Technology, Hangzhou 313000

逃问题, 并分析了可捕获性、纳什均衡以及捕获时间. 文献 [23] 在确保每个时刻都至少有一个追捕者具有全局视野的情况下, 提出了基于微分模型的追捕策略. 不同于无限视野的结果, 文献 [24] 在有限视野下设计了追捕群体快速逼近逃逸者的分布式算法, 并根据初始分布及速度比推导了捕获条件. 文献 [25] 采用图论方式研究了有限感知的追捕问题, 为每个智能体求解了分布式最优策略.

上述研究均基于模型求解追逃策略, 然而现实中由于不确定因素的存在, 构建准确的模型极为困难, 而强化学习可通过无模型的方式寻求最优策略, 因此其与追逃问题的结合也成为当下研究热点[6−7, 26]. 针对某一方使用固定策略的追捕问题, 文献 [8] 利用视野图像引入逃逸者位置的信念状态, 并基于 Soft actor-critic 算法获取最优追捕策略. 文献 [27] 基于深度 Q 网络, 并借助人工势场法对奖励函数进行改造以获取逃逸策略. 而对于追逃双方通过对抗学习进行智能追捕的问题, 文献 [28] 在无限视野下, 提出了 $Q(\lambda)$-learning 算法以求解追逃策略. 文献 [29] 则在有限视野下基于深度确定性策略梯度, 提出了两种网络拓扑结构来快速求解策略, 降低了多智能体算法的复杂度. 文献 [30] 对深度确定性策略梯度公式进行向量化拓展, 提出了一种多智能体协同目标预测网络, 保证了追捕群体对目标轨迹预测的有效性.

上述绝大多数追逃问题求解均基于定位信息, 但在特定环境下此类信息无法获取. 如水下航行器在固定海域中执行巡航与入侵驱逐任务时, 由于无线电信号在海水中迅速衰减, 此时航行器无法借助无线电导航系统对入侵者实现水下远距离、大范围的定位[31−32], 在此情况下, 借助轻便且低频的测距传感器实现追捕的研究是极为重要的. 文献 [33] 研究了单个追捕者基于距离构造几何图形以估计逃逸者的追逃问题, 并提出了在三维环境下使用两个追捕者估计逃逸者位置的方法. 在固定信标的帮助下, 文献 [34] 基于三角定位进行逃逸者位置估计, 并提出了对测量距离进行去噪处理的方法以获得精准定位. 文献 [35] 借助凸优化方法, 提出一种基于测量距离的梯度算法实现对逃逸者的定位. 文献 [36] 针对固定规则下的单移动机器人目标跟踪问题, 提出了一种利用测量距离与距离变化率求解追捕策略的方法. 此外, 文献 [37] 基于距离变化率提出了自适应切换算法, 证明了该算法稳定性与收敛性, 并在距离变化率不可用时将其扩展为使用观测器补偿的算法, 通过移动机器人围捕实验验证了其有效性.

综上所述, 基于距离的追逃问题已有较多研究成果, 但部分结果仍基于模型求解[32, 35−37], 或只针对固定策略的逃逸者[36], 亦或是需要借助额外设备如信标等[34]. 因此在无模型情况下针对智能逃逸者, 仅利用距离信息来实现追捕的问题仍有待于进一步探索. 本文将基于距离信息的 N 对 1 围捕问题与随机博弈相结合, 研究最优追逃策略. 在此问题中, 追捕群体仅领导者可测量与逃逸者间的相对距离, 其他跟随者通过领导者的共享获取此信息, 而逃逸者则拥有无限视野. 为求解追捕策略, 将环境分割引入信念区域状态以估计逃逸者位置. 同时根据相对距离, 对信念区域状态进行修正. 领导者借助信念引入想象逃逸者, 建立了信念区域状态下的连续随机追博弈, 并使用不动点定理证明此博弈平稳纳什均衡策略的存在性. 为求解逃逸策略, 由于逃逸者具有全局信息优势, 在追捕群体最优策略的基础上, 建立基于混合状态的 MDP 与相应最优的贝尔曼方程. 最后给出了基于强化学习的追逃策略求解算法.

本文结构安排如下: 第 1 节对追逃问题作出具体描述; 第 2 节证明基于信念区域状态的追博弈存在平稳纳什均衡策略, 并构建逃逸者的混合状态 MDP 与最优贝尔曼方程; 第 3 节给出求解追逃问题平稳策略的算法; 第 4 节通过数值仿真与对比, 验证本文方法的有效性; 第 5 节是全文总结.

符号说明. $\mathbf{R}^m$ 表示 m 维欧几里得空间; e_i 表示第 i 个元素为1, 其余为0 的列向量; $\|\cdot\|$ 表示欧几里得范数; $\Delta(A)$ 表示在集合 A 上概率测度的集合.

1　问题描述

本文研究 N 对 1 追逃问题, 将 N 个追捕者表示为 $P_i,\ i=1,\cdots,N$, 其中 P_1 是领导者, 其余为跟随者, 逃逸者表示为 E. 令第 i 个追捕者和逃逸者在 k 阶段的位置分别为 $P_i(k)$ 和 $E(k)$. 具体描述如下:

1) 环境: 如图 1 所示, 二维地图环境由不规则边界和障碍物组成. 追逃双方均可获知环境信息, 且追捕群体和逃逸者都被禁止触碰边界与障碍物.

2) 信息: 追捕群体仅领导者配备测距传感器, 用于测量其与逃逸者间的相对距离 $d(P_1,E)=\|P_1-E\|$, 追捕群体间可共享此信息与各自的位置; 而逃逸者具有全局视野, 可获得追捕群体的定位信息. 为方便起见, 假设即使被障碍物遮挡, 领导者仍可测量相对距离.

3) 捕获条件: k 阶段任意一个追捕者与逃逸者间的相对距离小于设定值 ℓ, 即 $d(P_i(k),E(k))=\|P_i(k)-E(k)\|<\ell$, 则追捕群体捕获逃逸者.

4) 速度与方向约束: 使用 $v_{P_i}(k)$ 表示第 i 个追捕者在阶段 k 的速度, $v_{P_i}(k)\in\mathcal{V}^P:=\{v_j^P,\ j=1,\cdots,M_1\}$. 类似地, 逃逸者速度为 $v_E(k)\in\mathcal{V}^E:=\{v_j^E,\ j=$

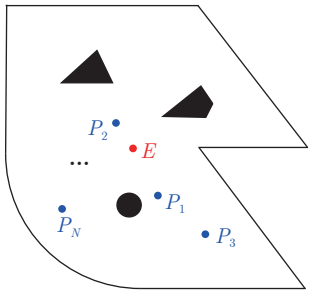


图 1 追逃问题环境

Fig. 1 Environment of pursuit-evasion problem

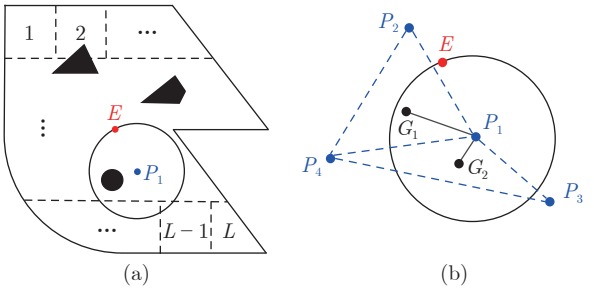


图 2 (a) L 个区域; (b) 追捕群体的划分

Fig. 2 (a) L regions; (b) Division of pursuit group

$1, \cdots, M_2\}$. 追逃双方运动速度具有以下约束: a) 不同追捕者每阶段可选择不同速度; b) 追捕群体和逃逸者在每阶段中的速度是常数; c) $V_{\max}^P < V_{\max}^E$, 其中 $V_{\max}^P = \max \mathcal{V}^P$, $V_{\max}^E = \max \mathcal{V}^E$. 此外, 追捕者与逃逸者移动方向的选择被限定为 D_1 和 D_2 个, 即 $\mathcal{D}^P = \{d_j^P, j = 1, \cdots, D_1\}$, $\mathcal{D}^E = \{d_j^E, j = 1, \cdots, D_2\}$.

5) 目标: 追捕群体目标为尽快捕捉到逃逸者, 而逃逸者目标为避免被追捕群体抓获.

注 1. 论述 4) 中的约束 b) 限定了双方在每阶段中使用匀速运动进行追逃. 当需考虑变速运动的情况时, 可通过在动作集中引入额外的加速度来实现, 这只会扩大双方的动作集, 对本文结论并无本质影响. 上述追逃问题被定义在二维环境中, 但通过改变相应动作与状态, 可直接将本文结果扩展到三维.

2 非完全信息追逃问题

本节针对此追逃问题, 给出基于信念区域状态的连续随机博弈框架与马尔科夫决策过程. 其中第 2.1 节通过对环境进行区域分割以估计逃逸者的位置, 并令跟随者采取包围行动; 第 2.2 节建立连续随机博弈框架求解追捕策略, 并证明了此博弈平稳纳什均衡策略的存在性; 第 2.3 节建立基于混合状态的马尔科夫决策过程与相应最优贝尔曼方程求解逃逸策略.

2.1 信念区域状态和重心距离

由于追捕群体无法对逃逸者作出定位, 因此对地图进行分割以估计逃逸者位置. 如图 2(a) 所示, 在地图中沿横向和纵向作平行线, 将其分割成 L 个区域. 令逃逸者所处的区域作为状态, 则状态集合 $S := \{s_1, \cdots, s_L, s_i = e_i, i = 1, \cdots, L\}$, 其中 s_i 表示第 i 个区域. 但仅通过测距仍无法获知其具体状态, 因此引入信念区域状态.

定义 1. 信念区域状态集合 $\mathcal{B} = \Delta(S)$ 表示在区域状态集合 $\mathcal{B}$ 上的概率分布集合, 则 k 时刻下信念区域状态 $B_k \in \mathcal{B}$ 表示追捕群体对逃逸者所处区域的概率估计.

基于上述定义, 显然得 $\sum_{i=1}^{L} B_k(s_i) = 1$, 其中 $B_k(s_i)$ 表示 k 阶段逃逸者位于第 i 个区域的概率. 同时, 如图 2(a) 所示, 基于 k 阶段测量距离, 构建以领导者位置 $P_1(k)$ 为圆心, 相对距离 $d(P_1(k), E(k))$ 为半径的圆. 显然只有被此圆穿越的区域, 才可能存在逃逸者, 因此信念区域状态 $B_k \in \mathbf{R}^L$ 是一个稀疏向量. 令 $\hat{L}$ 代表图 2(a) 追逃环境最外层区域个数, 则 B_k 最多有 $\hat{L}$ 个非零元素. 此外, 本文假设逃逸者均匀分布在被圆所穿越的区域中.

追捕群体使用信念区域状态估计逃逸者位置, 然而跟随者未配备测距传感器, 无法测量自身与逃逸者间的距离, 因此它们在追捕过程中使用包围的方式来协助领导者达到捕获的目的. 为实现包围, 将追捕群体划分为多个三角形, 每个跟随者 P_i 与其最近的两个跟随者及领导者构成两个三角形, 即 P_i 和 P_1 是两个三角形的公共点. 如图 2(b), 跟随者 P_4 与 P_2、P_3 以及领导者 P_1 形成三角形 $\triangle P_4P_1P_2$ 和 $\triangle P_4P_1P_3$. 整个追捕群体可构成 f 个三角形, 其中 f 的取值如下

$$f = \begin{cases} 1, & \text{若 } N = 3 \\ N - 1, & \text{其他} \end{cases} \tag{1}$$

根据追捕群体位置可计算出所有三角形的重心 G_i, $i = 1, \cdots, f$. 需要注意的是, 如果某一跟随者 P_i 与任意一个邻居以及领导者 P_1 共线, 则相应的三角形重心变为线段重心.

定义第 i 个重心与领导者间的距离为重心距离, 即 $\|G_i(k) - P_1(k)\|$, $i = 1, \cdots, f$. 显然每个跟随者均存在两个与自身相关的重心距离. 可知当三角形重心越接近上述以领导者位置 $P_1(k)$ 为圆心, 相对距离 $d(P_1(k), E(k))$ 为半径的圆, 跟随者包围效果越好. 因此每阶段任意跟随者 P_i 都试图最小化相应的两个重心距离与 $d(P_1(k), E(k))$ 的差值. 令此差

值为 ϕ_i, 显然其与跟随者的移动方向和速度相关.

此外, 为避免追逃双方触碰边界及障碍物, 引入如图 3 中所示的黄色警戒区域. 地图边界的警戒区域是沿边界向内延伸相应智能体的一步最大距离, 即 $V_{\max}^{P}$ 或 $V_{\max}^{E}$; 而障碍物的警戒区域则是向外延伸相应一步最大距离. 由于追逃双方知道环境信息和自身位置, 因此警戒区域的信息是公开的. 第 2.2 节将通过对警戒区域内智能体奖励的设置, 来规避其碰撞边界及障碍物的风险.

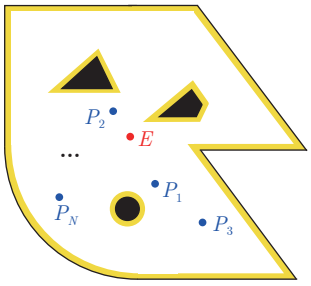


图 3　警戒区域

Fig. 3　Warning area

2.2　基于信念区域状态的追博弈

由于双方获知的位置信息不对称, 因此追捕群体借助信念区域状态引入一个想象逃逸者 $\bar{E}$, 由此建立连续随机追博弈来求解追捕策略.

首先, 定义追捕群体与想象逃逸者的纯动作空间, 由运动方向和速度组成. 令 $A^{P_i}=\{\text{static}\}\cup\{(d_j^{P_i}, v_l^{P_i})\}$, $j\in\{1,\cdots,D_1\}$, $l\in\{1,\cdots,M_1\}$ 为追捕者 P_i 的纯动作集, 其中 "static" 表示静止, $d_j^{P_i}$ 为追捕者第 j 个运动方向, $v_l^{P_i}$ 为第 l 个运动速度, 则追捕群体的纯动作集为 $A^P=\times A^{P_i}$. 类似地, 想象逃逸者的纯动作集为 $A^{\bar{E}}=\{\text{static}\}\cup\{(d_j^{\bar{E}}, v_l^{\bar{E}})\}$, $j\in\{1,\cdots,D_2\}$, $l\in\{1,\cdots,M_2\}$. $\mathcal{A}^P=\Delta(A^P)$ 和 $\mathcal{A}^{\bar{E}}=\Delta(A^{\bar{E}})$ 分别表示追捕群体和逃逸者纯动作集上的概率分布集合.

其次, k 阶段基于信念区域状态 B_k 和想象逃逸者的概率动作 $\sigma^{\bar{E}}\in\mathcal{A}^{\bar{E}}$, 可计算出区域状态转移概率 $\Pr(s_n|s_m,\sigma^{\bar{E}})$. 即在概率动作 $\sigma^{\bar{E}}$ 下, 追捕群体认为想象逃逸者从状态 s_m (第 m 个区域) 到 s_n (第 n 个区域) 的概率为

$$\Pr(s_n|s_m,\sigma^{\bar{E}})=\sum_{a^{\bar{E}}\in A^{\bar{E}}}\sigma^{\bar{E}}(a^{\bar{E}})\Pr(s_n|s_m,a^{\bar{E}})=$$

$$\begin{cases}\displaystyle\sum_{a^{\bar{E}}}\sigma^{\bar{E}}(a^{\bar{E}})\frac{\text{size}(m,a^{\bar{E}})}{\text{area}(m)}, & B_k(s_m)\neq 0\\ 0, & \text{其他}\end{cases}\tag{2}$$

其中 $a^{\bar{E}}\in A^{\bar{E}}$ 表示想象逃逸者纯动作, $\sigma^{\bar{E}}(a^{\bar{E}})$ 表示其采取 $a^{\bar{E}}$ 的概率, $\text{area}(m)$ 表示 m 区域中除去障碍物后的面积, $\text{size}(m, a^{\bar{E}})$ 则表示 m 区域中采取纯动作 $a^{\bar{E}}$ 后可转移到 n 区域的所有点集合.

注 2. 若区域 m 不存在障碍物, 则 $\text{size}(m, a^{\bar{E}})$ 为 m 区域中可转移到区域 n 的极限点所包围而成的面积; 若区域中存在障碍物, 还需去除由障碍物导致的不可行区域. 该不可行区域左右两侧边界线平行于逃逸者运动方向, 并且相切于区域内障碍物左右最外侧点. 如图 4 所示, 设想象逃逸者的纯动作 $a^{\bar{E}}=(d_3^{\bar{E}}, v_1^{\bar{E}})$, 方向角 $\theta_3^{\bar{E}}$ 为运动方向 $d_3^{\bar{E}}$ 与水平线的夹角, 速度 $v_1^{\bar{E}}$ 等于距离 $d(B, J)$. 显然点 A, B, E, F 构成了从 m 区域进入到 n 区域的所有极限点. 因此 $\text{size}(m, a^{\bar{E}})$ 等价于矩形 $ABEF$ 面积减去不可行区域面积 $CDGKHI$. 同时为便于计算, 将各区域东边与北边的边界线认定为本区域.

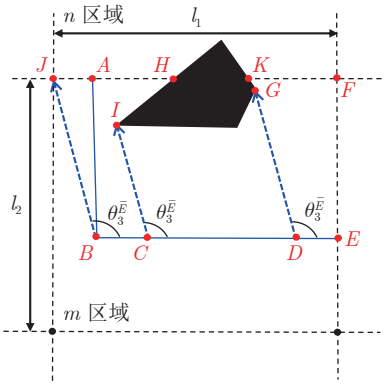


图 4　第 m 个区域

Fig. 4　The m-th area

信念区域状态更新机制.

基于区域状态转移概率 $\Pr(s_n|s_m,\sigma^{\bar{E}})$ 与测量距离 $d(P_1(k), \bar{E}(k))$, 可进行信念区域状态的更新, 其分为修正与估计两个过程. 具体来说, 追捕群体根据测量距离对信念区域状态进行后验修正, 再使用 $\Pr(s_n|s_m,\sigma^{\bar{E}})$ 估计下一阶段的信念区域状态. 为给出信念区域状态修正机理, 做出如下定义.

定义 2. 假设逃逸者均匀分布在区域 m 中, k 阶段的预测距离 $\hat{d}_k(s_m, a^P, a^{\bar{E}})$ 表示追捕群体和逃逸者分别做出纯动作 $a^P=\{a^{P_i}\}\in A^P$ 和 $a^{\bar{E}}\in A^{\bar{E}}$ 后, 逃逸者与领导者间的期望距离.

基于预测距离与相对距离两者间的差异来修正信念区域状态. 如图 5 所示, 令领导者与想象逃逸者在 k 阶段做出动作前的位置分别为 (x_{P_1}, y_{P_1}) 和 (x, y), 因此预测距离为

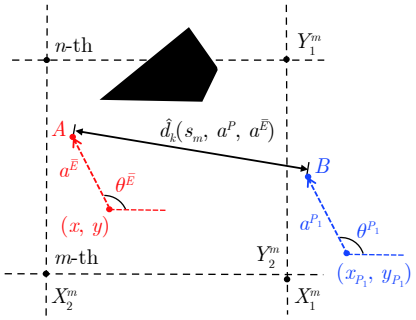


图 5 预测距离

Fig. 5 Prediction distance

$$\begin{cases}\hat{d}_k(s_m, a^P, a^{\bar{E}}) = d(A, B) = \dfrac{1}{\text{area}(m)} \times \\ \qquad \displaystyle\int_{Y_2^m}^{Y_1^m}\int_{X_2^m}^{X_1^m} \sqrt{x^2+y^2-2\xi_1 x-2\xi_2 y+\xi_3}\mathrm{d}x\mathrm{d}y \\ \xi_1 = x_{P_1} + v^{P_1}\cos\theta^{P_1} - v^{\bar{E}}\cos\theta^{\bar{E}} \\ \xi_2 = y_{P_1} + v^{P_1}\sin\theta^{P_1} - v^{\bar{E}}\sin\theta^{\bar{E}} \\ \xi_3 = x_{P_1}^2 + y_{P_1}^2 + (v^{P_1})^2 + (v^{\bar{E}})^2 + \\ \qquad 2x_{P_1}(\xi_1 - x_{P_1}) + 2y_{P_1}(\xi_2 - y_{P_1})\end{cases} \tag{3}$$

其中, θ^{P_1} 与 $\theta^{\bar{E}}$ 分别表示领导者与想象逃逸者的方向角; v^{P_1} 和 $v^{\bar{E}}$ 分别为领导者与想象逃逸者的运动速度; X_1^m, X_2^m 为区域 m 横坐标上下限; Y_1^m, Y_2^m 为区域 m 纵坐标上下限. 如图 5 所示, 区域 m 中存在障碍物, 则积分区域需去除障碍物部分.

令 k 阶段的信念区域状态为 $B_k = b$, 则 k 阶段逃逸者位于 m 区域的先验概率为

$$B_k(s_m) = \Pr(s_m|b) \tag{4}$$

令 k 阶段追捕群体的纯动作为 a^P, 想象逃逸者的概率动作为 $\sigma^{\bar{E}}$, 做出此动作后领导者与想象逃逸者之间的距离为 $d_k = d(P_1(k), \bar{E}(k))$. 基于此, 可获取信念区域状态的后验概率, 即

$$\tilde{B}_k(s_m) = \Pr(s_m|b, a^P, \sigma^{\bar{E}}, d_k) \tag{5}$$

引理 1. k 阶段信念区域状态的后验概率为

$$\tilde{B}_k(s_m) = \sum_{a^{\bar{E}}} \frac{\sigma^{\bar{E}}(a^{\bar{E}})b(s_m)\Pr(d_k|s_m, b, a^P, a^{\bar{E}})}{\sum\limits_{s'_m} b(s'_m)\Pr(d_k|s'_m, b, a^P, a^{\bar{E}})} \tag{6}$$

证明. 根据信念区域状态后验概率的定义, 可得

$$\tilde{B}_k(s_m) = \Pr(s_m|b, a^P, \sigma^{\bar{E}}, d_k) =$$

$$\sum_{a^{\bar{E}}} \sigma^{\bar{E}}(a^{\bar{E}})\Pr(s_m|b, a^P, a^{\bar{E}}, d_k) =$$

$$\sum_{a^{\bar{E}}} \frac{\sigma^{\bar{E}}(a^{\bar{E}})\Pr(s_m, d_k|b, a^P, a^{\bar{E}})}{\sum\limits_{s'_m\in S}\Pr(s'_m, d_k|b, a^P, a^{\bar{E}})} =$$

$$\sum_{a^{\bar{E}}} \frac{\sigma^{\bar{E}}(a^{\bar{E}})\Pr(s_m|b, a^P, a^{\bar{E}})\Pr(d_k|s_m, b, a^P, a^{\bar{E}})}{\sum\limits_{s'_m\in S}\Pr(s'_m|b, a^P, a^{\bar{E}})\Pr(d_k|s'_m, b, a^P, a^{\bar{E}})} =$$

$$\sum_{a^{\bar{E}}\in A} \frac{\sigma^{\bar{E}}(a^{\bar{E}})b(s_m)\Pr(d_k|s_m, b, a^P, a^{\bar{E}})}{\sum\limits_{s'_m\in S} b(s'_m)\Pr(d_k|s'_m, b, a^P, a^{\bar{E}})} \tag{7}$$

第三个等式由贝叶斯公式得出, 其中 $\Pr(s_m, d_k|b, a^P, a^{\bar{E}})$ 表示在信念区域状态 b 下, 追捕群体与逃逸者分别做出动作 a^P, $a^{\bar{E}}$ 后, 逃逸者位于 m 区域以及其与领导者间的距离为 d_k 的概率; 使用条件概率公式可得第四个等式; 根据式 (4), 可知逃逸者位于 m 区域的先验概率值仅与 b 有关, 因此 $\Pr(s_m|b, a^P, a^{\bar{E}}) = \Pr(s_m|b) = b(s_m)$, 所以第五个等式成立. □

对于 $\Pr(d_k|s_m, b, a^P, a^{\bar{E}})$, 其值与预测距离和测量距离间的差异有关, 即

$$\Pr(d_k|s_m, b, a^P, a^{\bar{E}}) = \frac{\ln\left(\frac{d_k}{b(s_m)\hat{d}_k(s_m, a^P, a^{\bar{E}})}\right)}{b(s_m)\hat{d}_k(s_m, a^P, a^{\bar{E}})} \tag{8}$$

值得一提的是, 若地图中只存在一个区域, 则逃逸者位于 m 区域的先验概率为 1, 因此无需进行修正, 在此情况下式 (7) 变为 $\tilde{B}_k(s_m) = \sum_{a^{\bar{E}}} \times \frac{\sigma^{\bar{E}}(a^{\bar{E}})b(s_m)\Pr(d_k|s_m, b, a^P, a^{\bar{E}})}{b(s_m)\Pr(d_k|s_m, b, a^P, a^{\bar{E}})} = \sum_{a^{\bar{E}}}\sigma^{\bar{E}}(a^{\bar{E}}) = 1$, 与上述分析一致.

基于区域状态转移概率可获取下一信念区域状态, 因此 $k+1$ 阶段的信念区域状态 B_{k+1} 为

$$B_{k+1} = \tilde{B}_k\Pr(s_n|s_m, \sigma^{\bar{E}}) \tag{9}$$

其中 $\Pr(s_n|s_m, \sigma^{\bar{E}})$ 为式 (2).

连续随机追博弈框架.

信念区域状态的转移满足马尔科夫特性[38]. 将追捕群体视为一个整体, 并且由于信念区域状态与追捕群体位置的连续性, 因此可建立由六元组组成的基于信念区域状态的连续随机博弈 $\mathcal{G} = (\mathcal{I}, \mathcal{A}, \mathcal{U}, \tilde{T}, T, R)$. 不失一般性, 假设所有追捕者和逃逸者是完全理性的, 并且互相知道其他人是完全理性的[15]. 以下是追博弈六元组具体信息.

1) 参与人: 令 $\mathcal{I} = \{P, \bar{E}\}$ 表示理性玩家集合, 其中 $P = \{P_i, i = 1, \cdots, N\}$ 和 $\bar{E}$ 分别表示追捕者群体和想象逃逸者.

2) 动作: $\mathcal{A} = \mathcal{A}^P \times \mathcal{A}^{\bar{E}}$ 表示在追捕群体和逃逸者纯动作集上的联合概率分布集合, 其中元素 $\sigma = (\sigma^P, \sigma^{\bar{E}}) \in \mathcal{A}$ 被称为追捕群体和逃逸者的联合概率动作.

3) 联合状态: 联合状态集合 $\mathcal{U}=\{\mathcal{P}os, \mathcal{B}\}$ 由追捕群体位置集合 $\mathcal{P}os$ 和信念区域状态集合 $\mathcal{B}$ 组成, 其中 $\mathcal{P}os=(\mathcal{X}^P, \mathcal{Y}^P)$ 为追捕群体横坐标与纵坐标集合.

4) 联合状态修正概率: 根据修正机制可得联合状态修正概率, 即 $\tilde{T}(\tilde{u}|u, \sigma, d_k)=\Pr(\tilde{U}_k=\tilde{u}|U_k=u, A_k=\sigma, d(P(k), \bar{E}(k))=d_k)$, 这表示 k 阶段基于测量距离 d_k, 联合概率动作 $\sigma\in\mathcal{A}$, 状态 $u\in\mathcal{U}$ 转移到修正状态 $\tilde{u}\in\mathcal{U}$ 的概率.

5) 联合状态转移概率: 根据 k 阶段联合概率动作 $\sigma\in\mathcal{A}$, 已修正的联合状态 $\tilde{u}\in\mathcal{U}$ 转移到下一联合状态 $u'\in\mathcal{U}$ 的概率为 $T(u'|\tilde{u}, \sigma)=\Pr(U_{k+1}=u'|\tilde{U}_k=\tilde{u}, A_k=\sigma)$.

6) 收益: 令 $R=\{r^P(u, \sigma), r^{\bar{E}}(u, \sigma)\}$ 表示在联合状态 u, 联合概率动作 σ 下, 各参与人的期望收益集合. 其中 $r^P(u, \sigma)=\sum_{i=1}^N r^{P_i}(u, \sigma^P, \sigma^{\bar{E}})$. 领导者单阶段收益为

$$r^{P_1}(u, \sigma^P, \sigma^{\bar{E}})=\sum_{a^P\in A^P}\sum_{a^{\bar{E}}\in A^{\bar{E}}}\sigma^P(a^P)\sigma^{\bar{E}}(a^{\bar{E}})\times (\gamma_1\varphi+\sum_{s_m\in S}b(s_m)\hat{d}(s_m, a^P, a^{\bar{E}})) \tag{10}$$

其中 $\gamma_1\in\{0, 1\}$ 表示领导者是否触碰边界及障碍物, 若领导者发生碰撞则受到惩罚 φ. 跟随者 P_i 单阶段收益为

$$r^{P_i}(u, \sigma^P, \sigma^{\bar{E}})=\sum_{a^P\in A^P}\sum_{a^{\bar{E}}\in A^{\bar{E}}}\sigma^P(a^P)\sigma^{\bar{E}}(a^{\bar{E}})\times (\gamma_i\varphi+\phi_i(a^P)) \tag{11}$$

其中 $\gamma_i\in\{0, 1\}$ 表示跟随者是否触碰边界及障碍物, $\phi_i(a^P)$ 表示跟随者的包围目标. 追捕群体的损失为想象逃逸者的收益, 因此逃逸者单阶段收益为

$$r^{\bar{E}}(u, \sigma^P, \sigma^{\bar{E}})=-r^P(u, \sigma^P, \sigma^{\bar{E}}) \tag{12}$$

根据追博弈框架中联合状态修正概率与转移概率定义, 可得出如下引理.

引理 2. k 阶段联合概率动作为 $\sigma=\{\sigma^P, \sigma^{\bar{E}}\}$, 联合状态 u 转移到下一状态 u' 概率为

$$\Pr(u'|u, \sigma)=\prod_{i=1}^N\sigma^{P_i}(a^{P_i})\sum_{a^{\bar{E}}}\sigma^{\bar{E}}(a^{\bar{E}})\sum_{s_m}b(s_m)\times \frac{\sigma^{P_1}(a^{P_1})\sum\limits_{a^{\bar{E}}}\sigma^{\bar{E}}(a^{\bar{E}})\Pr(d_k|s_m, a^{P_1}, a^{\bar{E}})}{\sum\limits_{a'^{P_1}}\sigma^{P_1}(a'^{P_1})\sum\limits_{a^{\bar{E}}}\sigma^{\bar{E}}(a^{\bar{E}})\Pr(d_k|s_m, a^{P_1}, a^{\bar{E}})} \tag{13}$$

证明. 由更新机制可知状态 u 转移到 u' 分为两个部分: 修正与更新, 并且联合状态 $u'=\{pos', b'\}$, $\tilde{u}=\{pos, \tilde{b}\}$, $pos\in\mathcal{P}os, pos'\in\mathcal{P}os$, $b\in\mathcal{B}, b'\in\mathcal{B}$. 因此 u 转移到 u' 的概率为

$$\Pr(u'|u, \sigma)=\Pr(u'|\tilde{u}, \sigma)\Pr(\tilde{u}|u, \sigma, d_k) \tag{14}$$

其中联合状态的修正过程只涉及到信念区域状态 b, 因此如下等式成立

$$\Pr(\tilde{u}|u, \sigma, d_k)=\Pr(\tilde{b}|b, \sigma^P, \sigma^{\bar{E}}, d_k) \tag{15}$$

同时, 修正过程仅与领导者纯动作有关, 所以基于想象逃逸者概率动作与测量距离 d_k, 将式 (15) 转化为

$$\Pr(\tilde{b}|b, \sigma^P, \sigma^{\bar{E}}, d_k)=\Pr(a^{P_1}|b, \sigma^{P_1}, \sigma^{\bar{E}}, d_k)=\sum_{s_m\in S}b(s_m)\Pr(a^{P_1}|s_m, \sigma^{P_1}, \sigma^{\bar{E}}, d_k)=\sum_{s_m\in S}b(s_m)\frac{\Pr(a^{P_1}, d_k|s_m, \sigma^{P_1}, \sigma^{\bar{E}})}{\sum\limits_{a'^{P_1}\in A^{P_1}}\Pr(a^{P_1}, d_k|s_m, \sigma^{P_1}, \sigma^{\bar{E}})}=\sum_{s_m\in S}\frac{b(s_m)\sigma^{P_1}(a^{P_1})\Pr(d_k|s_m, \sigma^{P_1}, \sigma^{\bar{E}}, a^{P_1})}{\sum\limits_{a'^{P_1}\in A^{P_1}}\sigma^{P_1}(a'^{P_1})\Pr(d_k|s_m, \sigma^{P_1}, \sigma^{\bar{E}}, a'^{P_1})}=\sum_{s_m}\frac{b(s_m)\sigma^{P_1}(a^{P_1})\sum\limits_{a^{\bar{E}}}\sigma^{\bar{E}}(a^{\bar{E}})\Pr(d_k|s_m, a^{P_1}, a^{\bar{E}})}{\sum\limits_{a'^{P_1}}\sigma^{P_1}(a'^{P_1})\sum\limits_{a^{\bar{E}}}\sigma^{\bar{E}}(a^{\bar{E}})\Pr(d_k|s_m, a^{P_1}, a^{\bar{E}})}$$

然后在状态更新过程中, 追捕群体位置状态 pos 与信念区域状态 b 是互相独立的, 因此

$$\Pr(u'|\tilde{u}, \sigma)=\Pr(pos', b'|pos, \tilde{b}, \sigma^P, \sigma^{\bar{E}})=\Pr(pos'|pos, \tilde{b}, \sigma^P, \sigma^{\bar{E}})\Pr(b'|pos, \tilde{b}, \sigma^P, \sigma^{\bar{E}})=\sigma^P(a^P)\sum_{a^{\bar{E}}\in A^{\bar{E}}}\sigma^{\bar{E}}(a^{\bar{E}})=\prod_{i=1}^N\sigma^{P_i}(a^{P_i})\sum_{a^{\bar{E}}\in A^{\bar{E}}}\sigma^{\bar{E}}(a^{\bar{E}})$$

其中第二个等式中 $\Pr(pos'|pos, \tilde{b}, \sigma^P, \sigma^{\bar{E}})$ 为追捕群体的位置状态转移概率, 由于位置状态只与追捕群体自身动作有关, 因此 $\Pr(pos'|pos, \tilde{b}, \sigma^P, \sigma^{\bar{E}})=\sigma^P(a^P)$; 而 $\Pr(b'|pos, \tilde{b}, \sigma^P, \sigma^{\bar{E}})$ 则表示修正的信念区域状态到下一阶段状态的转移概率, 由于追捕群体无法获知想象逃逸者的纯动作, 因此使用全概率公式, 由此第三个等式成立. □

令连续随机博弈 $\mathcal{G}$ 中平稳策略为 $\pi=\{\pi^P, \pi^{\bar{E}}\}$, 其表示联合状态到联合动作的映射, 即 $\pi: \mathcal{U}\to\mathcal{A}$. 则在联合状态 u 下, $\pi(u)$ 实际上是在追捕群体和想象逃逸者联合纯动作空间 $A^P\times A^{\bar{E}}$ 上的概率分布. 设初始联合状态为 u_0, 在平稳策略 π 下追捕群体目标函数表达式如下

$$J^P(u_0, \pi)=\sum_{k=0}^{\infty}\rho^k r^P(u, \{\pi^P(u), \pi^{\bar{E}}(u)\}) \tag{16}$$

其中 $0<\rho<1$ 为折扣因子. 根据想象逃逸者奖励函数定义, 它的收益是追捕群体的损失, 因此

$$J^{\bar{E}}(u_0,\pi)=-J^{P}(u_0,\pi) \tag{17}$$

定义 3. 若存在一个平稳策略 $\pi^*=\{\pi^{P*},\pi^{\bar{E}*}\}$ 使得追捕群体与想象逃逸者累积期望收益分别满足

$$J^{P}(u_0,\pi^{P},\pi^{\bar{E}*})\le J^{P}(u_0,\pi^{P*},\pi^{\bar{E}*}) \tag{18}$$

$$J^{\bar{E}}(u_0,\pi^{P*},\pi^{\bar{E}})\le J^{\bar{E}}(u_0,\pi^{P*},\pi^{\bar{E}*}) \tag{19}$$

则称 π^* 是博弈 $\mathcal{G}$ 的平稳纳什均衡策略[15].

追捕群体与想象逃逸者都是最大化自身累积收益的理性参与人, 因此解决此最大化问题就转变为寻求连续随机博弈 $\mathcal{G}$ 的平稳纳什均衡策略问题, 定理 1 证明了此博弈存在平稳纳什均衡策略.

定理 1. 追博弈 $\mathcal{G}$ 存在平稳纳什均衡策略.

证明. 追捕群体使用平稳策略 π, k 阶段联合状态为 u, 获得联合概率动作 σ, 可写出追捕群体的折扣收益

$$\begin{aligned}&F_{\pi}^{P}(u,\sigma)=r^{P}(u,\sigma)+\rho\sum_{u'}\Pr(u'|u,\sigma)J^{P}(u')=\\&r^{P}(u,\sigma^{P},\sigma^{\bar{E}})+\rho\sum_{u'}\sum_{a^{\bar{E}}}\sum_{s_m}\sigma^{P}(a^{P})\sigma^{\bar{E}}(a^{\bar{E}})\times\\&\frac{b(s_m)\sigma^{P_1}(a^{P_1})\sum\limits_{a^{\bar{E}}}\sigma^{\bar{E}}(a^{\bar{E}})\Pr(d_k|s_m,a^{P_1},a^{\bar{E}})}{\sum\limits_{a'^{P_1}}\sigma^{P_1}(a'^{P_1})\sum\limits_{a^{\bar{E}}}\sigma^{\bar{E}}(a^{\bar{E}})\Pr(d_k|s_m,a^{P_1},a^{\bar{E}})}=\\&r^{P}(u,\sigma^{P},\sigma^{\bar{E}})+\rho\sum_{a^{P}}\sum_{a^{\bar{E}}}\sigma^{P}(a^{P})\sigma^{\bar{E}}(a^{\bar{E}})J^{P}(u')\end{aligned} \tag{20}$$

其中 $J^P(u')$ 表示联合状态为 u' 时追捕群体的累积期望收益, $\sigma=\{\sigma^P,\sigma^{\bar{E}}\}$. 将转移概率 (13) 代入, 因此等式二成立; 在当前联合状态 u 确定时, 下一联合状态 u' 由追捕群体与想象逃逸者的纯动作决定, 且追逃双方纯动作集有限, 则下一联合状态 u' 是有限的, 因此等式三成立.

由于区域状态集合 S 是有限集, 它的子集是有限的. 令 $\boldsymbol{S}$ 为在区域状态集合上的 Borel-σ 代数, 则 $\boldsymbol{S}$ 中具有有限个元素, 同时在 $\boldsymbol{S}$ 上定义一个概率测度 p, 因此 $(S,\boldsymbol{S},p)$ 表示概率测度空间. $(S,\boldsymbol{S},p)$ 的任意开覆盖都存在有限子覆盖, 因此概率测度空间 $(S,\boldsymbol{S},p)$ 是一个紧度量空间, 上文定义信念区域状态集合 $\mathcal{B}$ 是 S 上的概率分布集合, 由此可知 $\mathcal{B}$ 是一个紧度量空间. 类似地, 可证得连续随机博弈中动作空间 $\mathcal{A}^P$ 和 $\mathcal{A}^{\bar{E}}$ 也是紧度量空间. 又因为坐标空间 $\mathcal{P}os$ 是二维有限连续区域, 也为紧度量空间, 而联合状态空间 $\mathcal{U}=\{\mathcal{P}os,\mathcal{B}\}$ 是 $\mathcal{P}os$ 空间与 $\mathcal{B}$ 空间的乘积, 因此 $\mathcal{U}$ 是紧度量空间.

接下来证明 $F_{\pi}^{P}(u,\cdot,\cdot)$ 在紧度量空间 $\mathcal{A}^P\times\mathcal{A}^{\bar{E}}$ 上的连续性. 定义 $(u^n,(\sigma^P)^n,(\sigma^{\bar{E}})^n)\in\mathcal{U}\times\mathcal{A}^P\times\mathcal{A}^{\bar{E}}$ 和 $(u,\sigma^P,\sigma^{\bar{E}})\in\mathcal{U}\times\mathcal{A}^P\times\mathcal{A}^{\bar{E}}$. 当 $n\to\infty$ 时, $(u^n,(\sigma^P)^n,(\sigma^{\bar{E}})^n)\to(u,\sigma^P,\sigma^{\bar{E}})$ 意味着 $u^n\to u$, $(\sigma^P)^n\to\sigma^P$, $(\sigma^{\bar{E}})^n\to\sigma^{\bar{E}}$. 由于追捕群体的奖励函数是有界的, 因此根据式 (10) 与式 (11), 可知当 $(u^n,(\sigma^P)^n,(\sigma^{\bar{E}})^n)\to(u,\sigma^P,\sigma^{\bar{E}})$ 时, $r^P(u^n,(\sigma^P)^n,(\sigma^{\bar{E}})^n)\to r^P(u,\sigma^P,\sigma^{\bar{E}})$. 所以对于每个联合状态 u 来说, $r^P(u,\cdot,\cdot)$ 在 $\mathcal{A}^P\times\mathcal{A}^{\bar{E}}$ 上连续. 并且根据表达式 (16) 知其有界, 因此累积期望收益 $J^P(u)$ 在 $\mathcal{A}^P\times\mathcal{A}^{\bar{E}}$ 上连续, 继而由式 (20) 得 $F_{\pi}^{P}(u,\cdot,\cdot)$ 在 $\mathcal{A}^P\times\mathcal{A}^{\bar{E}}$ 上连续.

同时定义 $M(u)$ 为联合状态空间 $\mathcal{U}$ 上所有有界 Borel 函数集合, 则 $J^P(u)\in M(u)$. 定义算子 Q 为

$$\begin{aligned}Q(J^P(u))=&\,r^P(u,\sigma^P,\sigma^{\bar{E}})+\rho\sum_{a^P}\sigma^P(a^P)\times\\&\sum_{a^{\bar{E}}}\sigma^{\bar{E}}(a^{\bar{E}})J^P(u')=F_{\pi}^{P}(u,\sigma^P,\sigma^{\bar{E}})\end{aligned} \tag{21}$$

因此 $Q(J^P(u))$ 是连续有界的, 即算子 Q 是自映射的. 又折扣因子 $0<\rho<1$, 容易看出算子 Q 是 $M(u)$ 上的一个压缩映射. 又空间 $M(u)$ 是所有有界 Borel 函数的集合, 因此其是完备度量空间. 根据 Banach 不动点定理[39], 算子 Q 存在一个唯一的不动点 $J^{P*}(u)$, 并且满足 $Q(J^{P*}(u))=J^{P*}(u)=J^P(u,\pi^{P*},\pi^{\bar{E}*})=F_{\pi}^{P}(u,\pi^{P*},\pi^{\bar{E}*})$.

$\mathcal{A}^P$ 和 $\mathcal{A}^{\bar{E}}$ 分别是集合 A^P 和 $A^{\bar{E}}$ 上面的概率分布, 因此 $\mathcal{A}^P$ 和 $\mathcal{A}^{\bar{E}}$ 分别是 M_1D_1 和 M_2D_2 维单纯型, 显然它们是紧凸集. 并且零和博弈下, $F_{\pi}^{P}(u,\cdot,\cdot)$ 是双线性函数, 即 $F^P(u,\cdot,\cdot)$ 在 $\mathcal{A}^P$ 上是凸的, 在 $\mathcal{A}^{\bar{E}}$ 上是凹的, 根据最大最小定理得

$$\max_{\sigma^P}\min_{\sigma^{\bar{E}}}F_{\pi}^{P}(u,\sigma^P,\sigma^{\bar{E}})=\min_{\sigma^{\bar{E}}}\max_{\sigma^P}F_{\pi}^{P}(u,\sigma^P,\sigma^{\bar{E}})$$

而平稳策略 π 是联合状态空间 $\mathcal{U}$ 到动作空间 $\mathcal{A}$ 的可测映射, 根据选择定理[39], 可知

$$\max_{\sigma^P}\min_{\sigma^{\bar{E}}}F_{\pi}^{P}(u,\sigma^P,\sigma^{\bar{E}})=\max_{\sigma^P}F_{\pi}^{P}(u,\sigma^P,\pi^{\bar{E}}(u))$$

因此

$$\begin{aligned}&\max_{\sigma^P}\min_{\sigma^{\bar{E}}}F_{\pi}^{P}(u,\sigma^P,\sigma^{\bar{E}})=\max_{\sigma^P}F_{\pi}^{P}(u,\sigma^P,\pi^{\bar{E}*})=\\&\quad F_{\pi}^{P}(u,\pi^{P*},\pi^{\bar{E}*})\end{aligned}$$

所以 $F_{\pi}^{P}(u,\sigma^P,\pi^{\bar{E}*})\le F_{\pi}^{P}(u,\pi^{P*},\pi^{\bar{E}*})$, 也就是说 $J_{\pi}^{P}(u,\sigma^P,\pi^{\bar{E}*})\le J_{\pi}^{P}(u,\pi^{P*},\pi^{\bar{E}*})$.

由于是零和博弈, 因此 $J^{\bar{E}}(u)=-J^P(u)$, $F_{\pi}^{\bar{E}}(u,\sigma^P,\sigma^{\bar{E}})=-F_{\pi}^{P}(u,\sigma^P,\sigma^{\bar{E}})$. 即逃逸者满足

$$\begin{aligned}J^{\bar{E}*}(u) = & -J^{P*}(u) = -F_{\pi}^{P}(u, \pi^{P*}, \pi^{\bar{E}*}) = \\ & -\max_{\sigma^{P}} \min_{\sigma^{\bar{E}}} (F_{\pi}^{P}(u, \sigma^{P}, \sigma^{\bar{E}})) = \\ & -\min_{\sigma^{\bar{E}}} \max_{\sigma^{P}} F_{\pi}^{P}(u, \sigma^{P}, \sigma^{\bar{E}}) = \\ & \max_{\sigma^{\bar{E}}} \min_{\sigma^{P}} F_{\pi}^{\bar{E}}(u, \sigma^{P}, \sigma^{\bar{E}}) = \\ & \max_{\sigma^{\bar{E}}} F_{\pi}^{\bar{E}}(u, \pi^{P*}, \sigma^{\bar{E}}) \end{aligned} \tag{22}$$

因此, $J_{\pi}^{\bar{E}}(u, \pi^{P*}, \sigma^{\bar{E}}) \leq J_{\pi}^{\bar{E}}(u, \pi^{P*}, \pi^{\bar{E}*})$. 综上所述, 博弈 $\mathcal{G}$ 存在平稳纳什均衡策略π^{P*} 和$\pi^{\bar{E}*}$. □

注 3. 本文采取连续随机博弈框架与分级制决策过程的主要原因是, 追捕群体无法获得逃逸者位置信息, 而逃逸者则拥有全局信息, 追逃双方信息不对称. 追捕群体通过引入想象逃逸者来构建基于信念区域状态的连续随机博弈框架, 以此实现自身利益最大化. 逃逸者由于信息占优, 可在获取追捕群体均衡策略的基础上, 进一步通过构建马尔科夫框架求解最优策略.

经典贝叶斯博弈亦可处理本文追逃问题, 然而每阶段最多会产生 $\hat{L}$ 种区域状态, 并且追逃过程是多阶段持续进行的, 因此 k 阶段可能出现与 $\hat{L}^{k}$ 成比例的状态数量, 这显然会导致维度灾难. 故本文使用基于信念区域状态的随机博弈以避免此类情况发生.

2.3　逃逸者的决策过程

与追捕群体使用连续随机博弈不同, 由于真实逃逸者具有全局信息, 因此它的最优策略求解可转变为一个马尔科夫决策过程. 真实逃逸者纯动作集与想象逃逸者相同, $A^{E} = A^{\bar{E}}$, 用 $\mathcal{A}^{E} = \Delta(A^{E})$ 表示 A^{E} 上的概率分布. 用四元组 $< \mathcal{H}, \check{\mathcal{A}}, \check{T}, \check{R} >$ 表示 MDP, 具体如下.

1) 混合状态: 混合状态 $\mathcal{H} = \{\mathcal{U}, \mathcal{P}os^{E}\}$, 由追博弈中联合状态$\mathcal{U}$ 与逃逸者自身坐标$\mathcal{P}os^{E}$ 组成.

2) 动作: $\check{\mathcal{A}} = \mathcal{A}^{P} \times \mathcal{A}^{E}$ 表示决策过程中的动作空间, 概率动作 $\check{\sigma} = \{\sigma^{P}, \sigma^{E}\} \in \check{\mathcal{A}}$ 为动作空间中的元素, 其中 $\sigma^{E} \in \mathcal{A}^{E}$.

3) 混合状态转移概率: k 阶段概率动作 $\check{\sigma} \in \check{\mathcal{A}}$, 则混合状态 $h \in \mathcal{H}$ 转移到下一阶段状态 $h' \in \mathcal{H}$ 的概率为 $\check{T}(h'|h, \check{\sigma}) = \Pr(H_{k+1} = h'|H_{k} = h, \check{A}_{k} = \check{\sigma})$.

4) 收益: 令 $r^{E}(h, \check{\sigma}) = r^{E}(h, \sigma^{P}, \sigma^{E})$ 表示逃逸者的期望收益, 由与所有追捕者的相对距离和环境的触碰惩罚组成.

$$\begin{aligned} r^{E}(h, \sigma^{P}, \sigma^{E}) = & \sum_{a^{P} \in A^{P}} \sum_{a^{E} \in A^{E}} \sigma^{P}(a^{P}) \sigma^{E}(a^{E}) \times \\ & \sum_{i=1}^{N} d(P_{i}, E) + \gamma_{E} \varphi \end{aligned} \tag{23}$$

其中, $d(P_{i}, E)$ 表示任意追捕者 P_{i} 与逃逸者间的相对距离, $\gamma_{E} \in \{0, 1\}$ 表示逃逸者是否触碰地图边界及障碍物.

令逃逸者平稳策略为 π^{E}, 它是混合状态 $\mathcal{H}$ 到动作空间 $\mathcal{A}^{E}$ 的映射. 在给定混合状态 h 时, $\pi^{E}(h)$ 实际上等价于当前状态下的概率动作 σ^{E}. 基于连续随机博弈的追捕群体最优策略 π^{P*}, 给出逃逸者的累积收益函数

$$J^{E}(h_{0}, \check{\pi}) = \sum_{k=0}^{\infty} \rho^{k} r^{E}(h, \{\pi^{P*}(u), \pi^{E}(h)\}) \tag{24}$$

其中, $h_{0} = \{u_{0}, pos_{0}^{E}\}$ 为初始混合状态, $u_{0} \in \mathcal{U}$, $pos_{0}^{E} \in \mathcal{P}os^{E}$.

逃逸者寻求自身累积收益最大化, 令 $\check{\pi}^{*} = \{\pi^{P*}, \pi^{E*}\}$ 表示最优平稳策略, 即满足

$$J^{E*}(h_{0}, \check{\pi}^{*}) \geq J^{E}(h_{0}, \pi^{P*}, \pi^{E}) \tag{25}$$

为获得马尔科夫决策过程中的相应的贝尔曼最优方程[38], 给出如下定义

$$J^{E*}(h, \check{\pi}^{*}) = \max_{\sigma^{E}} Q(h, \sigma^{P*}, \sigma^{E}) \tag{26}$$

$$\begin{aligned} Q(h, \sigma^{P*}, \sigma^{E}) = & r^{E}(h, \sigma^{P*}, \sigma^{E}) + \\ & \rho \sum_{h'} \check{T}(h'|h, \sigma^{P*}, \sigma^{E}) J^{E*}(h') \end{aligned} \tag{27}$$

其中 $Q(h, \sigma^{P*}, \sigma^{E})$ 表示状态动作对价值函数, h 为 k 时刻下的混合状态, h' 为下一时刻的混合状态, $J^{E*}(h')$ 为逃逸者在状态 h' 下的期望累积收益.

引理 3. 逃逸者的最优贝尔曼方程为

$$\begin{aligned} J^{E*}(h, \check{\pi}^{*}) = & \max_{\sigma^{E}} \{r^{E}(h, \sigma^{P*}, \sigma^{E}) + \\ & \rho \sum_{a^{E}} \sum_{a^{P}} \sum_{a^{\bar{E}}} \sigma^{E}(a^{E}) \sigma^{P*}(a^{P}) \sigma^{\bar{E}*}(a^{\bar{E}}) J^{E*}(h')\} \end{aligned} \tag{28}$$

证明. 由于概率动作 $\check{\sigma} = \{\sigma^{P*}, \sigma^{E}\}$, 状态 $h = \{u, pos^{E}\}$, $h' = \{u', pos'^{E}\}$, 其中 $pos^{E} \in \mathcal{P}os^{E}$ 为 k 时刻下逃逸者的位置, $pos'^{E} \in \mathcal{P}os^{E}$ 为下一时刻的位置, 因此混合状态 h 转移到下一状态 h' 的概率为

$$\begin{aligned} & \check{T}(h'|h, \check{\sigma}) = \Pr(u', pos'^{E}|u, pos^{E}, \sigma^{P*}, \sigma^{E}) = \\ & \Pr(u'|u, pos^{E}, \sigma^{P*}, \sigma^{E}) \Pr(pos'^{E}|u, pos^{E}, \sigma^{P*}, \sigma^{E}) = \\ & \sigma^{E}(a^{E}) \Pr(u'|u, pos^{E}, \sigma^{P*}, \sigma^{E}) = \\ & \sigma^{E}(a^{E}) \Pr(u'|u, pos^{E}, \sigma^{P*}, \sigma^{\bar{E}*}) \end{aligned} \tag{29}$$

由于联合状态 u 的转移只与追捕群体与想象逃逸者的策略有关, 因此第四个等式成立. 将式 (29) 代入式 (27) 可得

$$Q(h,\sigma^{P*},\sigma^{E})=r^{E}(h,\sigma^{P*},\sigma^{E})+\rho\sum_{h'}\sigma^{E}(a^{E})\times$$
$$\Pr(u'|u,pos^{E},\sigma^{P*},\sigma^{\bar{E}*})J^{E*}(h')=$$
$$r^{E}(h,\sigma^{P*},\sigma^{E})+\rho\sum_{pos'^{E}}\sum_{u'}\sigma^{E}(a^{E})\times$$
$$\Pr(u'|u,pos^{E},\sigma^{P*},\sigma^{\bar{E}*})J^{E*}(h')=$$
$$r^{E}(h,\sigma^{P*},\sigma^{E})+\rho\sum_{a^{E}}\sum_{a^{P}}\sum_{a^{\bar{E}}}\sigma^{E}(a^{E})\sigma^{P*}(a^{P})\times$$
$$\sigma^{\bar{E}*}(a^{\bar{E}})J^{E*}(h')$$

由于当前状态 pos^{E} 与 u 确定, 所以下一状态是由追捕群体与逃逸者的纯动作决定, 等式三成立. □

3 策略求解

本节基于强化学习算法 MAPPO (Multi-agent proximal policy optimization)[40], 给出了追捕群体平稳纳什均衡策略与逃逸者最优策略的求解算法. 与传统算法相比, MAPPO 主要基于中心化训练, 去中心化执行, 每个智能体都具有单独的 Actor-critic 结构. 并且目标函数在训练中进行小批量更新, 既避免了过多策略更新, 又提高了训练稳定性.

在连续随机博弈中, 追捕群体与想象逃逸者采取左右互搏的训练方式. 以追捕者 P_i 为例, 定义两个 Actor 网络: $A^{P_i}_{\text{new}}$, $A^{P_i}_{\text{old}}$, 和一个 Critic 网络: C^{P_i}. 让 $A^{P_i}_{\text{old}}$ 网络中的参数 θ^{i}_{old} 与追逃环境相交互, 以此来获得联合状态 u 和动作 a^{P_i}. θ^{i}_{old} 是前一次迭代训练所获得的策略参数, 在训练中不参与更新. 再根据采样数据对 $A^{P_i}_{\text{new}}$ 网络中的 θ^{i} 参数进行更新, 实现追捕者 P_i 的策略更新. $\pi^{P_i}_{\theta^{i}}(a^{P_i}|u)$ 为 $A^{P_i}_{\text{new}}$ 网络中的当前行为策略分布, $\pi^{P_i}_{\theta^{i}_{\text{old}}}(a^{P_i}|u)$ 为 $A^{P_i}_{\text{old}}$ 网络中的旧行为策略分布, 由于进行策略更新的 $A^{P_i}_{\text{new}}$ 所需的样本数据源于 $A^{P_i}_{\text{old}}$ 网络, 因此两者间的行为策略分布相差不可过大, 需进行取值限制处理, 实现重要性采样. 追捕者 P_i 的网络目标函数如式 (30) 所示.

$$\hat{J}^{P_i}(\theta)=\mathrm{E}\left[\min\left(\frac{\pi^{P_i}_{\theta^{i}}(a^{P_i}|u)}{\pi^{P_i}_{\theta^{i}_{\text{old}}}(a^{P_i}|u)}\hat{A}^{P_i},\ \text{clip}\left(\frac{\pi^{P_i}_{\theta^{i}}(a^{P_i}|u)}{\pi^{P_i}_{\theta^{i}_{\text{old}}}(a^{P_i}|u)},\ 1-\epsilon,\ 1+\epsilon\right)\hat{A}^{P_i}\right)\right] \tag{30}$$

其中优势函数为智能体累积收益与值函数 V^{P_i} 间的差, $\hat{A}^{P_i}=\sum_{k}\rho^{k}r^{P_i}-V^{P_i}$, ϕ^{i} 为 Critic 网络的参数; clip 表示行为策略分布取值切割操作, 使其分布值限制在 $1-\epsilon$ 与 $1+\epsilon$ 之间, ϵ 为切割系数. 通过目标函数不断更新参数 θ^{i} 和 ϕ^{i}, 直至目标函数达到最优. 类似地, 连续随机博弈中想象逃逸者也可定义出 $A^{\bar{E}}_{\text{new}}$, $C^{\bar{E}}$, $A^{\bar{E}}_{\text{old}}$, $\hat{J}^{\bar{E}}(\theta)$ 等.

真实逃逸者具有全局信息, 根据追捕群体的平稳纳什均衡策略 π^{P*}, 进行 MDP 最优策略求解. 同理, 逃逸者可定义出 A^{E}_{new}, A^{E}_{old}, C^{E}, $\hat{J}^{E}(\theta)$ 等参数, 并不断更新网络, 直至找到最优策略.

算法 1 是追逃问题中追捕群体与逃逸者最优平稳策略的求解过程, 第 1) 行到第 14) 行求解连续随机博弈中追捕群体的平稳纳什均衡策略 π^{P_i*}. 为通过数据抽样实现智能体的策略更新, 将 T 条包含状态、动作、奖励、优势函数以及值函数的序列分别存入追捕者与逃逸者记忆库中, 即第 8) 行所示. 同时, 为提高数据的可用性与训练效率, 进行 K 次更新, 如第 9) 行所示. 基于策略 π^{P_i*}, 可进行逃逸策略的求解, 即第 15) 行到第 24) 行. 类似地, 第 19) 行表示将序列存入记忆库中.

算法 1. 最优平稳追逃策略求解算法.

1) 初始化追捕群体与想象逃逸者的 $A^{P_i}_{\text{new}}$, $A^{P_i}_{\text{old}}$, $A^{\bar{E}}_{\text{new}}$, $A^{\bar{E}}_{\text{old}}$, C^{P_i}, $C^{\bar{E}}$ 网络, θ^{P_i}, $\theta^{\bar{E}}$, ϕ^{P_i}, $\phi^{\bar{E}}$, 记忆库 D^{P_i}, $D^{\bar{E}}$

2) Repeat

3) for $t\in\{1,\cdots,T\}$ do

4) 环境状态输入 $A^{P_i}_{\text{old}}$, $A^{\bar{E}}_{\text{old}}$, 根据 $\pi^{P_i}_{\theta^{i}_{\text{old}}}$, $\pi^{\bar{E}}_{\theta_{\text{old}}}$ 获得联合状态 u_t, 动作 $a^{P_i}_t$, $a^{\bar{E}}_t$, 以及奖励 $r^{P_i}_t$, $r^{\bar{E}}_t$

5) end for

6) 将 $\{u_t,a^{P_i}_t,r^{P_i}_t\}^{T}_{t=1}$ 输入 C^{P_i}, $C^{\bar{E}}$ 网络, 分别获得值函数 $\{V^{P_i}_t\}^{T}_{t=1}$, $\{V^{\bar{E}}_t\}^{T}_{t=1}$

7) 计算优势函数 $\{\hat{A}^{P_i}_t\}^{T}_{t=1}$, $\{\hat{A}^{\bar{E}}_t\}^{T}_{t=1}$

8) $\{u_t,a^{P_i}_t,r^{P_i}_t,\hat{A}^{P_i}_t,V^{P_i}_t\}^{T}_{t=1}$, $\{u_t,a^{\bar{E}}_t,r^{\bar{E}}_t,\hat{A}^{\bar{E}}_t,V^{\bar{E}}_t\}^{T}_{t=1}$ 分别存入记忆库 D^{P_i}

9) for $k\in\{1,\cdots,K\}$ do

10) 抽取样本 $\{u,a^{P_i},\hat{A}^{P_i},V^{P_i}\}$, $\{u,a^{\bar{E}},\hat{A}^{\bar{E}},V^{\bar{E}}\}$ 分别输入 $A^{P_i}_{\text{new}}$, $A^{\bar{E}}_{\text{new}}$, 计算目标函数 $\hat{J}^{P_i}$, $\hat{J}^{\bar{E}}$

11) 更新梯度参数 θ^{i}, $\theta^{\bar{E}}$, ϕ^{i}, $\phi^{\bar{E}}$, 获得 $\pi^{P_i}_{\theta^{i}}$, $\pi^{\bar{E}}_{\theta}$

12) end for

13) $\pi^{P_i}_{\theta^{i}_{\text{old}}}\leftarrow\pi^{P_i}_{\theta^{i}}$, $\pi^{\bar{E}}_{\theta_{\text{old}}}\leftarrow\pi^{\bar{E}}_{\theta}$

14) 获得追捕群体平稳纳什均衡策略 π^{P*}

15) 初始化逃逸者 A^{E}_{new}, A^{E}_{old}, C^{E}, θ^{E}, ϕ^{E}, 记忆库 D^{E}

16) Repeat

17) 基于追捕群体策略 π^{P*}, 使用 $\pi^{E}_{\theta^{i}_{\text{old}}}$ 运行 T 次, 获得 $\{u_t,a^{E}_t,r^{E}_t\}^{T}_{t=1}$, 及值函数 $\{V^{E}_t\}^{T}_{t=1}$

18) 计算优势函数 $\{\hat{A}^{E}_t\}^{T}_{t=1}$

19) 将 $\{u_t,a^{E}_t,r^{E}_t,V^{E}_t,\hat{A}^{E}_t\}^{T}_{t=1}$ 存入记忆库 D^{E}

20) for $k\in\{1,\cdots,K\}$ do

21) 计算目标函数 J^{E}, 更新梯度参数 θ^{E}, ϕ^{E}

22) end for

23) $\pi^{E}_{\theta^{E}_{\text{old}}}\leftarrow\pi^{E}_{\theta^{E}}$

24) 获得逃逸者最优策略 π^{E*}

4　数值仿真

本节通过三对一的案例来说明本文方法有效性. 仿真环境为 Windows10, 搭载的 CPU 为 AMD Ryzen 4800H, 显卡为 AMD Radeon Graphics 512 MB. 基于 Python3.6 搭建封闭二维空间, 同时使用 Pytorch1.8 深度学习框架进行训练. 追逃环境尺寸及障碍物如图 6 所示, 环境中存在三个黑色障碍物, 逃逸者为红色质点 E, 追捕群体为蓝色质点 P_1, P_2, P_3. 环境地图被切割为 16 个区域, 区域状态集为 $S=\{s_1,\cdots,s_{16}\}$, 相应的信念区域状态是一个 16 维的向量. 追捕者运动方向被均匀划分为 8 个: 东、东南、南、西南、西、西北、北、东北; 运动速度有两种: 0.4 m/s, 0.5 m/s, 因此结合静止动作, 所有追捕者均可采取 17 个动作. 逃逸者运动方向与追捕者一致, 而运动速度有三种: 0.4 m/s, 0.5 m/s, 0.6 m/s, 因此结合静止动作, 逃逸者可采取 25 个动作. 不失一般性, 令追捕群体与逃逸者的初始位置分别在地图的四个方位内随机产生. 抓捕成功的最短距离 ℓ 设置为追捕群体的最短步长, 即 0.4 m.

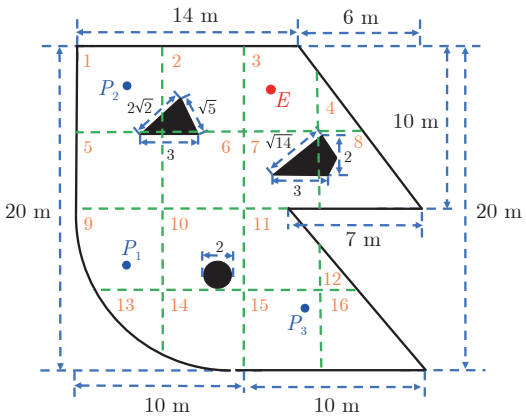


图 6　地图尺寸

Fig. 6　Size of map

为获取追逃问题的最优策略, 使用算法 1 进行求解. 在此算法中 Actor 网络与 Critic 网络使用两个全连接层作为隐藏层, 每层神经元个数分别为 64, 32, 神经网络使用 Adam 的梯度更新方式, 学习率为 0.000 1. 算法中追捕群体与逃逸者的记忆库均为 500, 策略更新次数 $K=20$, 折扣因子 $\rho=0.99$, 切割系数 $\epsilon=0.2$. 此外, 对追逃问题重复训练 20 000 局, 每一局最多运行 400 个阶段, 这样的训练独立进行 20 次.

追捕策略的训练过程如图 7 所示, 红色曲线为追捕群体的平均累积收益曲线, 阴影为其训练方差. 可从图中看出, 追捕群体累积收益呈上升趋势, 在训练 10 000 局时逐渐趋于收敛. 同时未采取信念修正的追捕训练效果如图 7 蓝色曲线所示. 由图可知, 追捕群体收益收敛后约为 −360, 而未修正的收益约为 −420, 修正状态的收益提升了近 15%, 且红色阴影小于蓝色阴影, 即修正状态的追捕训练更为稳定.

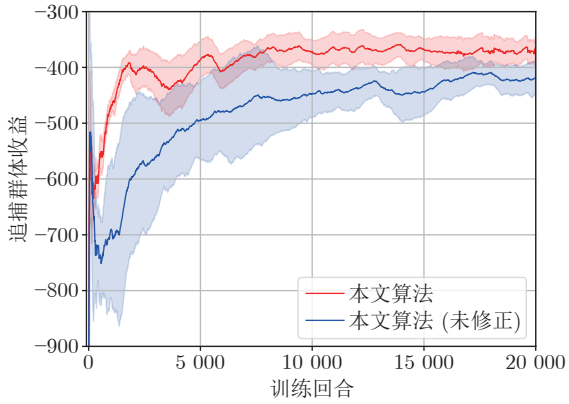


图 7　追博弈中追捕群体的收益

Fig. 7　Pursuits′ reward in the pursuit game

逃逸者策略的训练过程如图 8 所示, 其中红色曲线为逃逸者的平均累积收益, 阴影为其训练方差. 从图中可以看出逃逸者的收益在 13 000 局时趋于收敛, 最终稳定, 收益约为 380. 而蓝色曲线为未修正状态的收益, 稳定收益约为 500. 修正状态的收益较未修正的低了近 30%, 则使用修正状态的逃逸者弱于未修正的, 即修正状态的追捕群体强于未修正的. 同时两条曲线方差阴影的对比, 说明了使用修正状态的训练过程更为稳定.

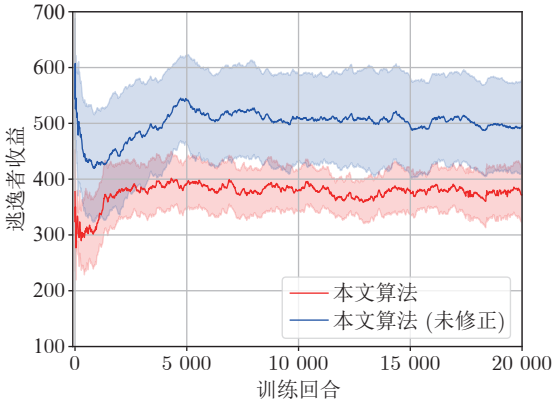


图 8　MDP 中逃逸者的收益

Fig. 8　Evader′s reward in MDP

经统计, 训练完成后追捕群体捕获成功的平均步数为 41 步左右, 成功率为 95%; 未进行状态修正的追捕群体其捕获成功的平均步数为 43 步, 成功率为 87%, 较修正机制下低了 8%, 可见使用测量距离进行信念修正是行之有效的. 为进一步展示追捕训练的效果, 在上述计算最优策略过程中, 每隔

100 局保存一次模型, 即保存了 200 个不同模型. 同时, 对每个模型进行 1 000 局的追捕测试, 具体测试结果如图 9 所示. 从图中可以看出, 随着逃逸者训练的进行, 其逃脱能力逐步上升, 因此被成功捕捉步数相应增加, 当对训练至 15 000 局时保存的模型进行测试时, 成功捕获步数已基本趋于稳定.

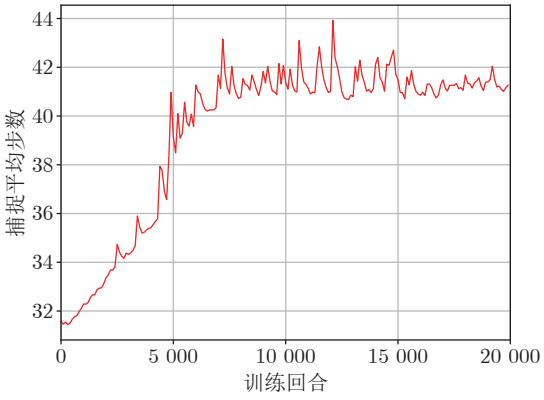


图 9 算法测试过程

Fig. 9 Algorithm testing process

为了验证本文方法的优越性, 使用如下几种算法进行对比: MAPPO[40], MASAC[41], MADDPG[42], 几何估计追捕[33], 基于三角定位追捕[34], 至少一人全局视野追捕[23], 自动追踪追捕[36], 自适应切换追捕[37] 以及随机策略. 在对比中, 固定所有算法中的逃逸者策略 (本文算法 1 所求得的逃逸策略). 同时为了适应此三对一的例子, 对上述部分算法做相应改进, 具体如下.

1) 几何估计追捕算法: 文献 [33] 聚焦一对一追逃问题, 为适应本算例, 将其改写为三对一追捕问题, 即追捕群体共享领导者所估计到的位置.

2) 基于三角定位的追捕算法: 文献 [34] 中追捕者利用信标, 也就是说使用了额外的传感器对逃逸者进行定位. 然而本文中不存在信标, 为进行定位, 准许每个阶段中领导者移动三次, 以三次不同的测距信息进行定位, 并将此信息共享给跟随者.

3) 自动追踪追捕算法与自适应切换追捕算法: 文献 [36] 和 [37] 研究了一对一追逃问题, 为适应本文算例, 将其改写为三对一追捕问题, 即所有追捕者均使用距离与距离变化率构建模型以求解追捕策略.

表 1 记录了使用不同算法的追逃测试结果. 从此表中发现, 当使用 MAPPO、MASAC、MADDPG 算法时, 智能体进行不断地试错与学习, 虽具有一定的训练效果, 但由于以上三种算法均未对距离信息进行有效的利用与处理, 导致捕获平均步数较高, 并且抓捕成功率低. 其中 MADDPG 算法使用同策略, 并且因为以确定性策略的方式, 无法获得随机均衡策略, 而 MAPPO 与 MASAC 算法均使用异策略, 并且采取了随机策略的方式, 所以捕获步数多于其余两种算法. 本文决策机制虽基于 PPO 算法, 但其结果优越性主要源于建立了基于信念区域状态的博弈框架与马尔科夫决策过程, 从表 1 可知与仅使用 MAPPO 相比, 本文算法捕获平均步数减少了 43 步, 成功率提高了 36%.

同时, 几何估计追捕算法[33] 的成功率较本文算法低 23%, 且所花费的步数是本文算法的近两倍, 可见在本文环境下, 该算法对逃逸者的位置估计效果较差. 而使用三角定位的追捕算法[34] 可精准定位逃逸者的位置, 因此捕捉成功率与本文算法接近, 但由于追捕群体为获得定位所需信息, 进行了额外的运动, 因此抓捕步数多于本文算法. 至少一人全局视野的追捕算法[23] 在视野范围内使用了最优追捕策略, 使得捕获成功率较高, 但对于视野范围外的情况, 追捕群体没有作出更为有效处理, 从而捕获步数高出本文算法 21 步. 并且, 通过距离与距离变化率求解追捕策略的自动追踪算法[36] 与自适应切换追捕算法[37], 均未直接对逃逸者的位置做估计定位, 导致追捕效果较差. 最后使用随机策略进行测试, 与预期一致, 由于追捕群体未采取任何智能策略而导致其效果最差. 此外, 通过对比可知, 即使是未进行信念状态修正的本文算法, 其测试效果仍优于绝大部分对比算法, 体现了使用博弈框架求解平稳纳什均衡策略的有效性.

最后, 在本文算法的多次测试中随机抽取 4 局, 画出追逃双方的运动轨迹图, 如图 10 所示. 图中蓝色三角形与红色三角形分别表示为追逃双方的初始位置, 蓝色圆点与红色圆点则分别代表追捕群体与逃逸者的运动轨迹, 颜色越深, 代表轨迹越新. 从图中看出追捕群体都在朝向逃逸者对其形成合围之势, 而逃逸者为逃脱追捕, 整体运动过程均朝着追捕群体相反的方向运动.

5 结论

本文针对仅有距离信息的多智能体追逃问题, 提出了一种基于连续随机博弈与马尔科夫决策过程的最优策略求解方法. 在求解追捕策略中, 为了弥补位置信息的缺失, 通过引入信念区域状态对逃逸者位置进行估计, 并且使用测量距离对信念区域状态进行修正. 由此搭建了基于信念区域状态的连续随机博弈, 并证明了此博弈平稳纳什均衡策略的存在性. 在求解逃逸者策略时, 根据追捕群体的最优策略与混合状态, 建立了最优贝尔曼方程, 并给出了基于强化学习的追逃策略求解算法. 通过与已有算法的对比, 验证了本文方法的有效性. 此外, 通过

表 1　结果对比

Table 1　Result comparison

算法	捕捉平均步数	捕捉成功率
本文算法	41	95%
本文算法 (未修正)	43	87%
MAPPO[40]	88	59%
MASAC[41]	85	61%
MADDPG[42]	99	56%
几何估计追捕[33]	78	72%
基于三角定位追捕[34]	61	94%
至少一人全局视野追捕[23]	62	85%
自动追踪追捕[36]	82	71%
自适应切换追捕[37]	65	66%
随机策略	152	10%

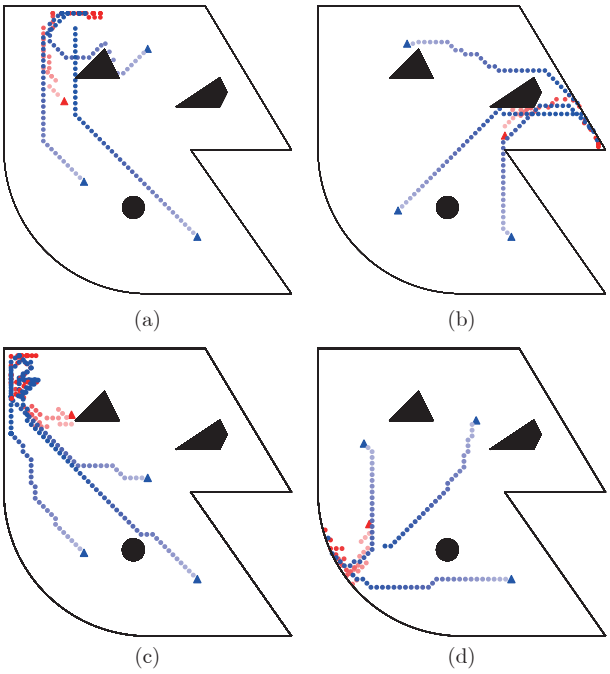


图 10　追捕群体与逃逸者的运动轨迹图

Fig. 10　Trajectories of pursuits and evader

追逃群体间简单的任务分配, 可将本文算法直接应用于多对多的追捕问题. 但如何在围捕过程中构建有效的智能体交互与任务切换机制, 以实现多对多环境下的高效追捕还有待于进一步研究.

References

1　Du Yong-Hao, Xing Li-Ning, Cai Zhao-Quan. Survey on intelligent scheduling technologies for unmanned flying craft clusters. *Acta Automatica Sinica*, 2020, **46**(2): 222−241
(杜永浩, 邢立宁, 蔡昭权. 无人飞行器集群智能调度技术综述. 自动化学报, 2020, **46**(2): 222−241)

2　Kou Li-Wei, Xiang Ji. Target fencing control of multiple mobile robots using output feedback linearization. *Acta Automatica Sinica*, 2022, **48**(5): 1285−1291
(寇立伟, 项基. 基于输出反馈线性化的多移动机器人目标包围控制. 自动化学报, 2022, **48**(5): 1285−1291)

3　Ferrari S, Fierro R, Perteet B, Cai C H, Baumgartner K. A geometric optimization approach to detecting and intercepting dynamic targets using a mobile sensor network. *SIAM Journal on Control and Optimization*, 2009, **48**(1): 292−320

4　Isaacs R. *Differential Games*. New York: Wiley, 1965.

5　Osborne M J, Rubinstein A. *A Course in Game Theory*. Cambridge: MIT Press, 1994.

6　Shi Wei, Feng Yang-He, Cheng Guang-Quan, Huang Hong-Lan, Huang Jin-Cai, Liu Zhong, et al. Research on multi-aircraft cooperative air combat method based on deep reinforcement learning. *Acta Automatica Sinica*, 2021, **47**(7): 1610−1623
(施伟, 冯旸赫, 程光权, 黄红蓝, 黄金才, 刘忠, 等. 基于深度强化学习的多机协同空战方法研究. 自动化学报, 2021, **47**(7): 1610−1623)

7　Geng Yuan-Zhuo, Yuan Li, Huang Huang, Tang Liang. Terminal-guidance based reinforcement-learning for orbital pursuit-evasion game of the spacecraft. *Acta Automatica Sinica*, 2023, **49**(5): 974−984
(耿远卓, 袁利, 黄煌, 汤亮. 基于终端诱导强化学习的航天器轨道追逃博弈. 自动化学报, 2023, **49**(5): 974−984)

8　Engin S, Jiang Q Y, Isler V. Learning to play pursuit-evasion with visibility constraints. In: Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS). Prague, Czech Republic: IEEE, 2021. 3858−3863

9　Al-Talabi A A. Multi-player pursuit-evasion differential game with equal speed. In: Proceedings of the IEEE International Automatic Control Conference (CACS). Pingtung, Taiwan, China: IEEE, 2017. 1−6

10　Selvakumar J, Bakolas E. Feedback strategies for a reach-avoid game with a single evader and multiple pursuers. *IEEE Transactions on Cybernetics*, 2021, **51**(2): 696−707

11　de Souza C, Newbury R, Cosgun A, Castillo P, Vidolov B, Kulić D. Decentralized multi-agent pursuit using deep reinforcement learning. *IEEE Robotics and Automation Letters*, 2021, **6**(3): 4552−4559

12　Zhou Z J, Xu H. Decentralized optimal large scale multi-player pursuit-evasion strategies: A mean field game approach with reinforcement learning. *Neurocomputing*, 2022, **484**: 46−58

13　Garcia E, Casbeer D W, Von Moll A, Pachter M. Multiple pursuer multiple evader differential games. *IEEE Transactions on Automatic Control*, 2021, **66**(5): 2345−2350

14　Pierson A, Wang Z J, Schwager M. Intercepting rogue robots: An algorithm for capturing multiple evaders with multiple pursuers. *IEEE Robotics and Automation Letters*, 2017, **2**(2): 530−537

15　Gibbons R. *A Primer in Game Theory*. Harlow: Prentice Education Limited, 1992.

16　Parthasarathy T. Discounted, positive, and noncooperative stochastic games. *International Journal of Game Theory*, 1973, **2**(1): 25−37

17　Maitra A, Parthasarathy T. On stochastic games. *Journal of Optimization Theory and Applications*, 1970, **5**(4): 289−300

18　Liu S Y, Zhou Z Y, Tomlin C, Hedrick K. Evasion as a team against a faster pursuer. In: Proceedings of the American Control Conference. Washington, USA: IEEE, 2013. 5368−5373

19　Huang L N, Zhu Q Y. A dynamic game framework for rational and persistent robot deception with an application to deceptive pursuit-evasion. *IEEE Transactions on Automation Science and Engineering*, 2022, **19**(4): 2918−2932

20　Qi D D, Li L Y, Xu H L, Tian Y, Zhao H Z. Modeling and solving of the missile pursuit-evasion game problem. In: Proceedings of the 40th Chinese Control Conference (CCC). Shanghai, China: IEEE, 2021. 1526−1531

21　Liu Kun, Zheng Xiao-Shuai, Lin Ye-Ming, Han Le, Xia Yuan-Qing. Design of optimal strategies for the pursuit-evasion prob-

lem based on differential game. *Acta Automatica Sinica*, 2021, **47**(8): 1840−1854
(刘坤, 郑晓帅, 林业茗, 韩乐, 夏元清. 基于微分博弈的追逃问题最优策略设计. 自动化学报, 2021, **47**(8): 1840−1854)

22 Xu Y H, Yang H, Jiang B, Polycarpou M M. Multiplayer pursuit-evasion differential games with malicious pursuers. *IEEE Transactions on Automatic Control*, 2022, **67**(9): 4939−4946

23 Lin W, Qu Z H, Simaan M A. Nash strategies for pursuit-evasion differential games involving limited observations. *IEEE Transactions on Aerospace and Electronic Systems*, 2015, **51**(2): 1347−1356

24 Fang X, Wang C, Xie L H, Chen J. Cooperative pursuit with multi-pursuer and one faster free-moving evader. *IEEE Transactions on Cybernetics*, 2022, **52**(3): 1405−1414

25 Lopez V G, Lewis F L, Wan Y, Sanchez E N, Fan L L. Solutions for multiagent pursuit-evasion games on communication graphs: Finite-time capture and asymptotic behaviors. *IEEE Transactions on Automatic Control*, 2020, **65**(5): 1911−1923

26 Zheng Yan-Bin, Fan Wen-Xin, Han Meng-Yun, Tao Xue-Li. Multi-agent collaborative pursuit algorithm based on game theory and Q-learning. *Journal of Computer Applications*, 2020, **40**(6): 1613−1620
(郑延斌, 樊文鑫, 韩梦云, 陶雪丽. 基于博弈论及 Q 学习的多 Agent 协作追捕算法. 计算机应用, 2020, **40**(6): 1613−1620)

27 Zhu J G, Zou W, Zhu Z. Learning evasion strategy in pursuit-evasion by deep Q-network. In: Proceedings of the 24th International Conference on Pattern Recognition (ICPR). Beijing, China: IEEE, 2018. 67−72

28 Bilgin A T, Kadioglu-Urtis E. An approach to multi-agent pursuit evasion games using reinforcement learning. In: Proceedings of the International Conference on Advanced Robotics (ICAR). Istanbul, Turkey: IEEE, 2015. 164−169

29 Wang Y D, Dong L, Sun C Y. Cooperative control for multi-player pursuit-evasion games with reinforcement learning. *Neurocomputing*, 2020, **412**: 101−114

30 Zhang R L, Zong Q, Zhang X Y, Dou L Q, Tian B L. Game of drones: Multi-UAV pursuit-evasion game with online motion planning by deep reinforcement learning. *IEEE Transactions on Neural Networks and Learning Systems*, DOI: 10.1109/TNNLS.2022.3146976

31 Coleman D, Bopardikar S D, Tan X B. Observability-aware target tracking with range only measurement. In: Proceedings of the American Control Conference (ACC). New Orleans, USA: IEEE, 2021. 4217−4224

32 Chen W, Sun R S. Range-only SLAM for underwater navigation system with uncertain beacons. In: Proceedings of the 10th International Conference on Modelling, Identification and Control (ICMIC). Guiyang, China: IEEE, 2018. 1−5

33 Bopardikar S D, Bullo F, Hespanha J P. A pursuit game with range-only measurements. In: Proceedings of the 47th IEEE Conference on Decision and Control. Cancun, Mexico: IEEE, 2008. 4233−4238

34 Lima R, Ghose D. Target localization and pursuit by sensor-equipped UAVs using distance information. In: Proceedings of the International Conference on Unmanned Aircraft Systems (ICUAS). Miami, USA: IEEE, 2017. 383−392

35 Fidan B, Kiraz F. On convexification of range measurement based sensor and source localization problems. *Ad Hoc Networks*, 2014, **20**: 113−118

36 Chaudhary G, Sinha A. Capturing a target with range only measurement. In: Proceedings of the European Control Conference (ECC). Zurich, Switzerland: IEEE, 2013. 4400−4405

37 Güler S, Fidan B. Target capture and station keeping of fixed speed vehicles without self-location information. *European Journal of Control*, 2018, **43**: 1−11

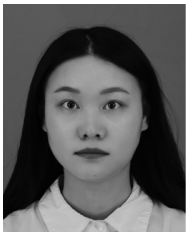
38 Sutton R S, Barto A G. *Reinforcement Learning: An Introduction* (Second edition). Cambridge: MIT Press, 2018.

39 Kreyszig E. *Introductory Functional Analysis With Applications*. New York: John Wiley & Sons, 1991.

40 Yu C, Velu A, Vinitsky E, Gao J X, Wang Y, Bayen A, et al. The surprising effectiveness of PPO in cooperative multi-agent games. In: Proceedings of the 36th International Conference on Neural Information Processing Systems. New Orleans, USA: NIPS, 2022.

41 Haarnoja T, Zhou A, Abbeel P, Levine S. Soft actor-critic: Off-policy maximum entropy deep reinforcement learning with a stochastic actor. In: Proceedings of the 35th International Conference on Machine Learning. Stockholm, Sweden: PMLR, 2018. 1861−1870

42 Lillicrap T P, Hunt J J, Pritzel A, Heess N, Erez T, Tassa Y, et al. Continuous control with deep reinforcement learning. In: Proceedings of the 4th International Conference on Learning Representations. San Juan, Puerto Rico: ICLR, 2015.



陈灵敏 浙江工业大学信息工程学院硕士研究生. 2020 年获得绍兴文理学院学士学位. 主要研究方向为博弈论与机器学习在决策问题中的应用.
E-mail: 2112003096@zjut.edu.cn
(**CHEN Ling-Min** Master student at College of Information Engineering, Zhejiang University of Technology. She received her bachelor degree from Shaoxing University in 2020. Her research interest covers game theory and machine learning in decision-making.)

冯 宇 浙江工业大学信息工程学院教授. 2011 年获得法国南特矿业大学博士学位. 主要研究方向为网络化控制系统, 分布式滤波, 不确定系统的鲁棒分析与控制, 以及博弈论与机器学习在决策问题中的应用. 本文通信作者. E-mail: yfeng@zjut.edu.cn
(**FENG Yu** Professor at College of Information Engineering, Zhejiang University of Technology. He received his Ph.D. degree from Ecole des Mines de Nantes in 2011. His research interest covers networked control systems, distributed filtering, and robust analysis and control for uncertainty systems, and applications of game theory and machine learning in decision-making. Corresponding author of this paper.)

李永强 浙江工业大学信息工程学院副教授. 2014 年获得北京交通大学博士学位. 主要研究方向为强化学习, 非线性控制以及深度学习.
E-mail: yqli@zjut.edu.cn
(**LI Yong-Qiang** Associate professor at College of Information Engineering, Zhejiang University of Technology. He received his Ph.D. degree from Beijing Jiaotong University in 2014. His research interest covers reinforcement learning, nonlinear control and deep learning.)