arXiv:2305.01869v1 [cs.RO] 3 May 2023

面向协同护送问题的连续动作空间分布式主动感知

Rhett Hull[1,3], Ki Myung Brian Lee[1], Jennifer Wakulicz[1],Chanyeol Yoo[1],James McMahon[2],Bryan Clarke[3],Stuart Anstee[3],Jijoong Kim[3] 和Robert Fitch[1]

摘要——**我们研究协同护送问题，其中一支去中心化的支援机器人团队会隐式协助价值更高的主机器人执行任务。其核心挑战在于如何评估支援机器人的行动对主机器人任务的影响。为刻画这种影响，我们为支援机器人定义了两种新颖的辅助奖励函数，称为满足度提升和满意度熵，分别用于计算任务成功概率的提升幅度及其不确定性。基于这些奖励函数，我们使用去中心化交叉熵方法（Dec-CEM）协调由主机器人和支援机器人组成的整个团队。Dec-CEM 是一种基于乘积分布近似、针对多智能体系统对交叉熵法（CEM）的新扩展。在一个模拟的物体避障场景中，与传统的解耦式信息收集相比，我们的规划框架将任务满足度最多提高了一倍。我们的研究成果意义在于引入了一类新的算法问题，有望推动异构多机器人系统的重要实际应用。**

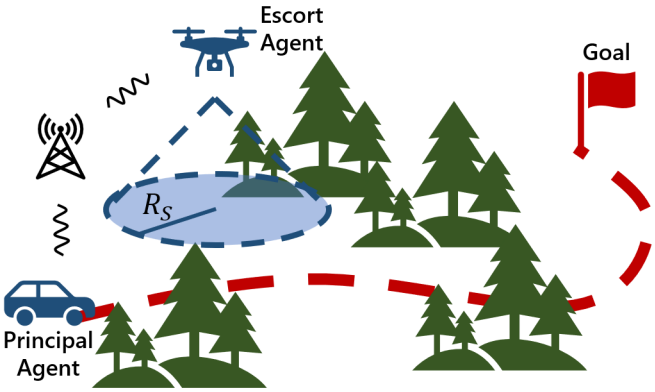

图1. 协同护航示例。没有机载传感器的主体智能体需要避开物体并到达目标。护送智能体配备了传感器，用于获取测量值并更新对物体位置的信念。所有智能体之间都会共享测量值和轨迹意图。我们考虑由最多三个护送智能体组成的护航团队。

I. 引言

协调多机器人系统的应用可能涉及由异构团队组成的系统，其中主体机器人以某种方式得到一个或多个支援机器人的协助，而这些支援机器人可能能力较弱、成本也较低。支援机器人可以发挥多种作用，例如作为远程感知和感知信息的来源，为导航决策提供更充分的信息。设计此类系统所面临的基本挑战之一，是如何协调支援机器人的行为，以促进主体机器人实现其目标。理想情况下，支援机器人能够主动收集与主体机器人目标相关的信息；但另一方面，主体机器人的行动可能取决于这些信息，因此无法预先确定。

为了让机器人团队能够支援主体机器人或主体智能体，协调算法必须能够预测支援团队的测量值或行动对主体智能体任务表现的影响。我们将这类问题称为协同护送问题，以指代这一类协调问题，即支援机器人作为主体机器人／智能体的护航团队。

护送任务在冲突和紧急情况下十分常见；例如，船舶或车辆必须穿越疑似布有水雷的区域，或车队必须沿着可能遭受袭击的路线行进。在这一过程中，一个隐含的事实是，护送机器人会为了自己所护送的对象而承受更高的受伤风险；这推动了关于在此类任务中使用机器人的研究。在这类情况下，人们通常假设护送机器人会在一定程度上独立于其保护对象行动。护送机器人不能假定主智能体会协调甚至理解它们的行为。另一方面，护送机器人必须理解主智能体的目标，才能采取适当行动。护送机器人还必须能够传达有关风险和安全的信息，因为主智能体很可能需要调整自身行为，才能使护送任务切实可行。

本文定义了通用协同护送问题的一个特定实例，并提出 DecCEM——一种基于交叉熵法（CEM）新变体的去中心化解决方案，用于在连续动作空间中进行规划。我们考虑的问题是：护送智能体（EA）必须针对以高斯信念建模的不确定物体位置进行信息收集，以支持主体智能体（PA）。CEM [1, 2] 是一种基于采样的规划算法，可找到一个能以概率最大化给定奖励的控制分布。CEM 允许对控制分布进行任意参数化，因此将基于分布的规划器从离散

本研究得到澳大利亚政府研究培训计划（RTP）奖学金、澳大利亚联邦政府和海军研究办公室（N62909-21-1-2031）的支持。[1]作者就职于悉尼科技大学，地址为澳大利亚新南威尔士州 Ultimo，邮编 2006 {rhett.c.hull, brian.lee, jennifer.wakulicz}@student.uts.edu.au，{chanyeol.yoo, rfitch}@uts.edu.au。[2] James McMahon 就职于美国海军研究实验室，美国，邮箱：james.mcmahon@nrl.navy.mil。[3]作者就职于澳大利亚国防部国防科学技术集团 {rhett.hull1, bryan.clarke, stuart.anstee, jijoong.kim}@defence.gov.au。

动作空间扩展到了连续动作空间。我们开发 DecCEM 的方法是基于乘积分布近似 [3, 4]，将 CEM 扩展到多智能体系统。

DecCEM 的目标是最大化整个团队（包括所有探索智能体和主智能体）的联合奖励函数。为准确反映探索智能体的测量值对主智能体任务表现的影响，我们为探索智能体定义了一个辅助奖励函数。我们提出两种可选的辅助函数：满意度熵（SE）和满足度提升（SI），它们直接衡量主智能体任务满足概率的提升，或任务成功的不确定性。

在图1所示的一个到达-避障场景示例中，我们将所提算法与不考虑主智能体在线动作的传统方法进行了比较。与基础信息收集方法相比，SE 方法的表现最高提升至两倍；与仅基于先验进行规划的主体智能体相比，最高提升至三倍。SI 方法的表现最高是仅基于先验进行规划方法的两倍；有趣的是，其表现与基础信息收集方法相当，甚至更差。最后，我们证明，所提出的目标函数能够自适应地调整对先验的探索与利用，以影响主智能体的轨迹，从而提升整体任务成功率。

本工作的贡献是：1）提出协同护送问题，将其作为一种特定类型的联合优化问题，其中协调目标取决于独立智能体的动作；2）提出 DecCEM 算法，用于解决连续动作空间中的分布式智能体所面临的这一问题。本工作有助于推动多机器人系统的重要实际应用，在这些应用中，机器人发挥着不可或缺的辅助作用。

II. 相关工作

对于护送问题，存在多种不同的解释。早期变体将其定义为一个最优控制问题，其中护送智能体必须围困并引导主体机器人进入目标区域 [5–12]。近年来，出现了更复杂的定义，即协同 护送问题，其中协同处理护送智能体和主体智能体的控制，以实现目标任务 [13, 14]。特别相关的协同护送问题建模主要侧重于护送智能体解决信息收集问题，以辅助完成目标 [15–18]。

典型的信息收集方法，例如 [19–28] ，会利用传感器（移动式或其他类型）进行测量，以降低环境中任意目标系统状态估计的不确定性。其他方法则侧重于降低对环境过程或现象的不确定性 [29–31]。在这两种情况下，通常都通过假设目标或环境具有高斯信念来实现不确定性降低，并使用卡尔曼滤波器或高斯过程的各种变体来维持该信念。在这些研究中，由于所采用的协方差更新过程独立于测量值的具体数值，因此可以直接评估信息增益。我们针对护送智能体的信息收集任务采用了类似的建模方式，以描述一种.....

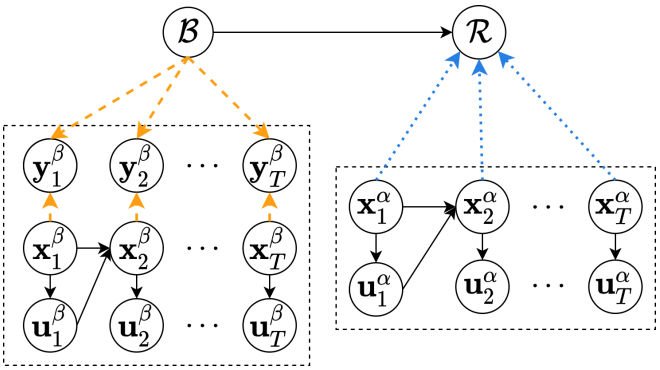

图2. 解耦护送问题 [18] 的概率图模型（PGM）表示。护送智能体（蓝色虚线）勘察环境并测量 $\mathbf{O}$。主体智能体（橙色点线）移动以完成 ϕ，且可能与 $\mathbf{O}$ 碰撞。碰撞会导致任务失败。

不确定环境，并评估护送机器人收集的信息所产生的影响。然而，在我们的研究中，护送智能体的信息收集目标与这些论文不同。

与 [15–18], 类似，我们提出的护送智能体信息收集目标并非着眼于降低整体环境不确定性，而是专门降低不确定性，以帮助主体机器人完成其任务。尽管这些研究依赖探索智能体和保护智能体通过序贯规划以集中式方式解决联合控制问题，但我们的工作提出了一种完全分布式解决方案，其灵感来自 [3] 中的概念。

III. 问题表述

图2所示的协同护送问题由一个机器人团队 $\mathcal{R}$ 组成。每个机器人 $r \in \mathcal{R}$ 在离散时刻 $t+1$ 的状态 $\mathbf{x}_t^r$ 由一个状态转移模型描述：

$$\mathbf{x}_{t+1}^r = \mathbf{f}^r(\mathbf{x}_t^r, \mathbf{u}_t^r), \tag{1}$$

其中，状态$\mathbf{x}_t^r \in \mathbb{R}^N$ 和动作$\mathbf{u}_t^r \in \mathbb{R}^M$ 是连续的。机器人在一个部分已知的环境中运行，该环境包含一组物体，其位置$\mathbf{O} = \{\mathbf{o}_1, \cdots, \mathbf{o}_N\} \subset \mathbb{R}^N$已知得并不精确。

存在两类不同的机器人：主智能体（PAs）和护送智能体（EAs）。PAs 和 EAs 的集合是 $\mathcal{R}$ 的不相交子集，分别记为 A，$B \subset \mathcal{R}$。

每个探索智能体 $\beta \in B$ 都配备了噪声传感器，可测量处于特定径向范围 R_S内的物体的位置。因此，状态 $\mathbf{x}_t^\beta$ 下的测量值 $\mathbf{y}_t^\beta$ 由传感器模型描述：

$$\mathbf{y}_t^\beta(\mathbf{x}_t^\beta) = \{\mathbf{o}_i + \epsilon \mid \mathbf{o}_i \in \mathbf{O}, ||\mathbf{o}_i - \mathbf{x}_t^\beta|| < R_S\}, \tag{2}$$

其中，$\epsilon \sim \mathcal{N}(0, \sigma^2 I)$ 为均值为零的高斯噪声。我们假设，每个探索智能体所获取的所有测量值 $\mathbf{y}_t^\beta$ 均可供其他所有机器人使用。保护智能体未配备用于获取测量值的传感器，因此必须依靠探索智能体提供的测量值来估计环境。

为简洁起见，我们用$\mathbf{X}_t^r = \{\mathbf{x}_t^r, ...\mathbf{x}_{t+T}^r\}$ 表示机器人 $r \in \mathcal{R}$ 从时间范围 T开始的轨迹。

t。控制动作和测量值也遵循相同的规则（即，对于 $r \in \mathcal{R}$, $\mathbf{U}_t^r = \{\mathbf{u}_t^r, \ldots \mathbf{u}_{t+T}^r\}$，$\mathbf{Y}_t^\beta = \{\mathbf{y}_t^\beta, \ldots \mathbf{y}_{t+T}^\beta\}, \beta \in \mathcal{B}$）。类似地，我们将上标替换为机器人集合，以表示该集合中所有机器人的轨迹集合，例如 $\mathbf{X}_t^{\mathcal{R}} = \{\mathbf{X}_t^r \mid r \in \mathcal{R}\}$。

虽然 PAs 无法测量物体位置，但仍需完成一项取决于物体位置$\mathbf{O}$的任务 ϕ 。该任务通过满足可能性 $P(\phi \mid \mathbf{X}_t^A, \mathbf{O})$ 建模，它描述了在完全已知物体位置的情况下成功的概率。例如，对于通用代价函数 C [32]，可以采用玻尔兹曼分布 $\log P(\phi \mid \mathbf{X}_t^A, \mathbf{O}) \propto -C(\mathbf{X}_t^A, \mathbf{O})$；或者采用某个时序逻辑公式 Ψ 的满足概率，使得 $P(\phi \mid \mathbf{X}_t^A, \mathbf{O}) = P(\mathbf{X}_t^A \models \Psi \mid \mathbf{O})$ [33–36]。

由于环境$\mathbf{O}$存在不确定性，无法直接计算 $P(\phi \mid \mathbf{X}_t^A, \mathbf{O})$。因此，我们最大化在探索智能体（ EAs ）的测量值 $\mathbf{Y}_t^B$条件下的满足的后验概率 $P(\phi \mid \mathbf{X}_t^A, \mathbf{Y}_t^B)$。这样做可以根据探索智能体所获得的环境信念，对所有可能的物体位置进行平均。因此，满足的后验概率反映了探索智能体所进行的测量所产生的影响，进而体现了其路径的质量。总体而言，该问题可正式表述如下。

问题 1（协同护送问题）。给定部分已知环境 $\mathcal{E}$ 中的主智能体 A、探索智能体 B 以及隐藏物体 $\mathbf{O} \subset \mathcal{E}$，分别为主智能体和探索智能体求解最优控制序列 $\mathbf{U}^{A*}$、$\mathbf{U}^{B*}$，以最大化主智能体完成任务的概率 ϕ:

$$\mathbf{U}^{A*}, \mathbf{U}^{B*} = \underset{\mathbf{U}^A, \mathbf{U}^B}{\arg\max} P(\phi \mid \mathbf{X}_t^A, \mathbf{Y}_t^B), \tag{3}$$

问题 1 的挑战有两个方面。最直接的是，由于目标函数与探索智能体轨迹之间存在复杂依赖，朴素方法需要采用集中式规划。我们使用交叉熵方法（CEM） [1]，的去中心化变体来解决这一问题；只要能够根据机器人的轨迹评估目标函数，该方法就可以计算出高质量的解。更重要的是，规划时无法获得未来测量值 $\mathbf{Y}_t^B$。我们通过为探索智能体推导替代奖励函数来缓解这一问题，这些奖励函数利用预测的高斯信念来刻画主智能体规划性能的提升。

IV. 协同护航规划

我们提出了一种滚动时域规划框架，用于解决问题 1，同时应对上述去中心化和测量选择方面的挑战。算法 1 概述了单个机器人 r 所采用的框架。在第 1 行，利用最新测量值更新物体位置的信念 $\mathcal{B}_t(\mathbf{O})$。随后使用更新后的信念评估奖励函数 R^r ，并据此优化控制量（第 2—6 行）。为避免测量模拟带来的挑战，探索智能体和主智能体分别使用不同的奖励R^r 。

为了找到使奖励最大化的控制量，我们引入算法 2 中概述的 DecCEM。DecCEM 不再寻找单一的控制量集合，而是寻找一种控制分布 $\mathcal{Q}(\mathbf{U}_t^{\mathcal{R}})$，使

算法 1 去中心化滚动时域规划概述
机器人 r的规划

输入： 物体测量值 $\mathbf{y}_t^B$，奖励函数 R^r
输出： 机器人 r 的**控制**$_t^r$,
1: $\mathcal{B}_t(\mathbf{O}) \leftarrow$ 更新信念$(\mathcal{B}_{t-1}(\mathbf{O}), \mathbf{x}_t^B, \mathbf{y}_t^B)$
2: $\mathcal{Q}(\mathbf{U}_t^r) \leftarrow \mathcal{N}(\mathbf{0}, \sigma_0^2 I)$
3: **对于**固定次数的迭代 **执行**
4: $\mathcal{Q}(\mathbf{U}_t^{\mathcal{R}\setminus r}) \leftarrow$接收分布()
5: $\mathcal{Q}(\mathbf{U}_t^r) \leftarrow$ 分散式交叉熵法$(R^r, \mathcal{Q}(\mathbf{U}_t^{\mathcal{R}\setminus r}))$
6: 广播分布$(\mathcal{Q}(\mathbf{U}_t^r))$
返回 $\mathbf{U}_t^r \sim \mathcal{Q}(\mathbf{U}_t^r)$

算法 2 机器人 r 的分散式交叉熵法过程

输入： 奖励函数 R^r，其他机器人的控制分布-布 $\mathcal{Q}(\mathbf{U}_t^{\mathcal{R}\setminus r})$
输出： 机器人 r 的控制分布， $\mathcal{Q}(\mathbf{U}_t^r)$
1: **对于**固定次数的迭代 **执行**
2: $\hat{\mathbf{U}}_t^r \sim \mathcal{Q}(\mathbf{U}_t^r)$ //采样自身控制
3: $\hat{}\mathbf{U}_t^{\mathcal{R}\setminus r} \sim \mathcal{Q}(\mathbf{U}_t^{\mathcal{R}\setminus r})$ //采样其他机器人控制
//计算自身的精英集
4: $\hat{}\mathbf{U}_t^{r*} = \{\mathbf{U}_t \in \hat{}\mathbf{U}_t^r \mid R^r(\hat{}\mathbf{U}_t^r, \hat{}\mathbf{U}_t^{\mathcal{R}\setminus r}) > \mathrm{R}\}$
//更新自身分布
5: $\mathcal{Q}(\mathbf{U}_t^r) \leftarrow$ 拟合高斯分布$(\hat{\mathbf{U}}_t^{r*})$
返回 $\mathcal{Q}(\mathbf{U}_t^r)$

任务满足概率 $\mathcal{P}(\phi \mid \mathbf{X}_t^A, \mathbf{Y}_t^B)$ 最大化。受 [3, 4]，启发，我们对 $\mathcal{Q}(\mathbf{U}_t^{\mathcal{R}}) = \prod_{r \in \mathcal{R}} \mathcal{Q}(\mathbf{U}_t^r)$ 施加乘积分布分解。这使所有机器人的规划相互解耦，因此每个机器人都会更新并传递自己的控制分布 $\mathcal{Q}(\mathbf{U}_t^r)$（第 6 行和第 4 行）。在第 5 行，每个机器人运行独立的 CEM 循环，依据其他机器人传递的分布更新自身动作 $\mathcal{Q}(\mathbf{U}_t^r)$，以使团队奖励最大化。通信不必同步进行，而第 4 行只返回当前可用的最新分布。

A. 信念更新程序

给定测量值 $\mathbf{Y}_t^B$，我们使用卡尔曼滤波器的信息形式，获得关于物体 $\mathcal{B}_t(\mathbf{O}) = \mathcal{N}(\hat{}\mathbf{O}_t, \Lambda_t^{-1})$ 的高斯信念，其均值为 $\hat{}\mathbf{O}_t$，信息矩阵为 Λ_t，即协方差的逆。采用信息形式的好处是，信念更新具有可加性。

$$\begin{aligned} \Lambda_t &= \Lambda_{t-1} + \sum_{\beta \in B} \mathbf{I}(\mathbf{x}_t^\beta), \\ \hat{\mathbf{O}}_t &= \Lambda_t^{-1}(\Lambda_{t-1}\hat{\mathbf{O}}_{t-1} + \sum_{\beta \in B} \mathbf{I}(\mathbf{x}_t^\beta)\mathbf{y}_t^\beta). \end{aligned} \tag{4}$$

这里，$\mathbf{I}(\mathbf{x}_t^\beta)$ 是新息矩阵。我们通过设置 $\mathbf{I}(\mathbf{x}_t^\beta) = \sigma^{-2} I$ if $\mathbf{y}_t^\beta$ 在感知范围内，否则设为 0（即测量结果是虚假的），以考虑感知范围约束 (2)。这样可以自动忽略感知范围之外的物体，相当于测量误差为无穷大。

B. 去中心化交叉熵方法

根据更新后的信念，DecCEM 过程会调整每个机器人的控制分布 $\mathcal{Q}(\mathbf{U}_t^r)$，以最大化团队奖励。在此过程中，它会使用其他机器人定期通过通信共享的控制分布 $\mathcal{Q}(\mathbf{U}_t^{\mathcal{R}\setminus r})$（算法 1 第 4、6 行）。其他机器人的分布可以异步接收，并且即使通信丢失，该算法也能平稳地降级运行。

机器人 r 的一次DecCEM迭代包括以下步骤。首先，从其自身控制的当前分布中抽取随机样本 $\mathbf{U}_t^r$（第2行）。此外，从已通信的其他机器人控制分布中抽取随机样本 $\hat{}\mathbf{U}_t^{\mathcal{R}\setminus r}$（第3行）。这些样本通过机器人动力学模型进行传播，并用于评估团队奖励函数。利用奖励$\hat{}$值，提取机器人 r自身控制样本的"精英集"$\mathbf{U}_t^{r*}$，其中样本的奖励超过设定阈值 R（第4行）。随后，通过对精英集拟合高斯分布来更新机器人 r的控制分布（第5行）。在广播之前，该过程会重复固定次数的迭代。更新并通信控制分布的外层循环会重复执行，以得到最终的控制分布。

在我们的实现中，我们通过提取最优的 N_e 个样本而非使用阈值来改进交叉熵法（CEM）的迭代过程。此外，当控制动作的平均方差低于预设阈值时，我们终止迭代，以防止过拟合。

虽然我们不提供收敛保证，但算法 2 中 DecCEM 的计算复杂度主要取决于迭代次数和控制分布的选择，$\mathcal{Q}(\mathbf{U}_t^r)$。本文选择高斯分布，因为有许多高效的期望最大化算法可供使用。

C. 奖励函数

DecCEM 使用的奖励函数 R^r 因保护智能体和探索智能体而异。对于保护智能体，奖励简单地设为：在当前物体信念下任务满足的边缘对数概率：

$$R^A(\mathbf{U}_t^A) = \log P(\phi \mid \mathbf{X}_t^A) = \log \mathbb{E}_{\mathbf{O}\sim\mathcal{B}_t(\cdot)}[P(\phi \mid \mathbf{X}_t^A, \mathbf{O})]. \tag{5}$$

值得注意的是，主智能体的奖励函数不依赖探索智能体收集的未来测量值，这促使保护智能体仅依据当前信念进行保守规划。

给定为最大化式 (5) 而计算出的保护智能体控制分布 $\mathcal{Q}(\mathbf{U}_t^A)$，探索智能体应选择自身的控制，以尽可能支持保护智能体完成任务。因此，探索智能体的规划不应仅仅像信息收集任务中的典型做法那样，着眼于提升整个环境 $\mathcal{E}$ 中的 $\mathcal{B}$ [19–21]，而应着重提升 $\mathcal{B}$，以提高保护智能体满足任务的概率。这里，我们介绍两种体现这一思路的奖励函数，以及一种基于互信息上置信界的更传统的信息收集奖励。

1) 满足度提升（*SI*）：最直接的方法是让探索智能体直接最大化保护智能体达到任务满足的概率。为此，SI 方法旨在通过结合 EA测量，对保护智能体达到任务满足的预期概率进行最大幅度的提升。换言之，该方法求解：

$$R_{\mathrm{SI}}^B(\mathbf{U}_t^B) = \mathbb{E}_{\mathbf{X}_t^A\sim\mathcal{Q}(\cdot)}\left[P(\phi \mid \mathbf{X}_t^A, \mathbf{Y}_t^B) - P(\phi \mid \mathbf{X}_t^A)\right], \tag{6}$$

其中，$\mathbf{X}_t^A$ 是通过动态模型从 $\mathcal{Q}(\mathbf{U}_t^A)$中隐式采样得到的。

计算 SI 目标函数 (6) 的一个难点在于$\mathbf{U}_t^B$与条件测量值 $\mathbf{Y}_t^B$之间存在隐式依赖。为缓解这一难题，我们对第一项进行如下变换，该变换可由贝叶斯定理和概率图模型中的条件独立性性质推导得出（图2）：

$$P\left(\phi \mid \mathbf{X}_t^A, \mathbf{Y}_t^B\right) = \mathbb{E}_{\mathbf{O}\sim\mathcal{B}_t(\cdot\mid\mathbf{U}_t^B)}\left[P(\phi \mid \mathbf{X}_t^A, \mathbf{O})\right]. \tag{7}$$

这里，$\mathcal{B}_t(\mathbf{O} \mid \mathbf{U}_t^B)$ 表示探索智能体执行动作$\mathbf{U}_t^B$后物体的预测信念。对于我们考虑的高斯目标，可以通过将信息矩阵 (4) 向前传播，同时保留均值来完成此类预测。换言之，$\Lambda_{t+T} = \Lambda_t + \sum_{\tau\in T,\beta\in B} I(\mathbf{x}_\tau^\beta)$。

2) 满意度熵 *(SE)* 的降低：SI 方法旨在提高由保护智能体控制分布所隐含的任务满足概率，而从信息收集的角度来看，有人可能会认为，降低任务满足概率与提高该概率同样有价值。换言之，测量任务满足的变化可能具有同等价值。为此，我们考虑利用测量集$\mathbf{Y}_t^B$降低满足概率的二元熵。形式上，SE 方法求解：

$$R_{\mathrm{SE}}^B(\mathbf{U}_t^B) = \mathbb{E}_{\mathbf{X}_t^A\sim\mathcal{Q}(\cdot)}\left[h(P(\phi \mid \mathbf{X}_t^A)) - h(P(\phi \mid \mathbf{X}_t^A, \mathbf{Y}_t^B))\right], \tag{8}$$

其中 $h(P) = -P\log P - (1-P)\log P$为二元熵。任务满足的后验概率 $P(\phi \mid \mathbf{X}_t^a)$ 使用公式 (7) 按照计算 SI 的类似方法求得。此处的奖励倾向于在探索智能体测量值之后获得较低的熵。由于二元熵在 $P = 1/2$ 处达到最大值 1，并且在该点两侧对称递减，因此 SE 目标会使从探索智能体抽取的样本的满足概率朝着更确定的方向增加或降低。

3) 互信息上置信界：互信息上置信界根据先验任务满足的期望值和信息增益，为后验任务满足提供上界。使用互信息上置信界可以证明，以 $\geq 1-\delta$ [18] 的概率：

$$P(\phi \mid \mathbf{X}_t^A, \mathbf{Y}_t^B) \leq \log \mathbb{E}_{\mathbf{O}\sim\mathcal{B}_t(\cdot)} \exp P(\phi \mid \mathbf{X}_t^A, \mathbf{O}) + \frac{1}{\delta} I(\mathbf{Y}_t^B; \mathbf{O}). \tag{9}$$

这里，第一项中的期望是相对于当前信念计算的，而 $I(\mathbf{Y}_t^B;\mathbf{O})$ 表示测量值 $\mathbf{Y}_t^B$ 与物体位置 $\mathbf{O}$ 之间的香农信息增益。对于本文所考虑的高斯目标，信息增益可通过信息矩阵表示为 $I(\mathbf{Y}_t^B; \mathbf{O}) = \frac{1}{2}(\log\det\Lambda_{t+T} - \log\det\Lambda_t)$】【。



图 3. 不同 EA 奖励变体下 PA 的失败率。数值越高越差。当 PA 与物体发生碰撞时，记录为失败。**深灰色：**盲目，**浅灰色：**互信息上置信界，**红色：**满意度提升，**绿色：**满意度熵。

MI-UCB（9）的一个显著特点是，PAs 和探索智能体完全解耦，即它们的奖励可以直接相加。由于探索智能体的动作只影响信息增益项，因此信息增益是探索智能体唯一需要最大化的奖励。因此，MI-UCB 方法求解的是：

$$R^{B}_{\text{MI-UCB}}(\mathbf{U}^{B}_{t}) = I\left(\mathbf{Y}^{B}_{t}; \mathbf{O}\right). \tag{10}$$

也就是说，互信息上置信界目标等价于文献中通常考虑的信息最大化 [19, 20]。

V. Results

A. 仿真设置

在以下仿真结果中，给定 PAs 的轨迹和物体 $\mathbf{O}$，PAs 规划以成功概率 $P(\phi \mid \mathbf{X}^{\alpha}_{t}, \mathbf{O})$实现任务 ϕ。我们考虑了一部分任务类别，其目标是在避开物体$\mathbf{O}$的同时到达目的地$\mathbf{d}$。受 [33], 启发，我们将此到达-避障任务建模为到达任务和避障任务的合取：

$$P(\phi \mid \mathbf{X}^{\alpha}_{t}, \mathbf{O}) = \prod_{\tau,i} P\left(\phi^{O} \mid \mathbf{x}^{\alpha}_{\tau}, \mathbf{o}_{i}\right) P\left(\phi^{D} \mid \mathbf{x}^{\alpha}_{\tau}\right), \tag{11}$$

其中，到达（ϕ^{D}）任务和避障（ϕ^{O}）任务建模为：

$$\begin{aligned} P(\phi^{O} \mid \mathbf{x}^{\alpha}_{t}, \mathbf{o}_{i}) &= 1 - P_{O} \exp\left(-\frac{||\mathbf{x}^{\alpha}_{t} - \mathbf{o}_{i}||^{2}}{2r^{2}_{O}}\right), \\ P(\phi^{D} \mid \mathbf{x}^{\alpha}_{t}) &\propto \exp\left(-\frac{||\mathbf{x}^{\alpha}_{t} - \mathbf{d}||^{2}}{2r^{2}_{D}}\right), \end{aligned} \tag{12}$$

其中，r_{D}、$_{O}$ 是控制接受半径/碰撞半径的参数，而 P_{O} 控制峰值碰撞概率。

所有智能体均遵循自行车运动学模型：

$$\begin{bmatrix} \dot{x} \\ \dot{y} \\ \dot{\theta} \end{bmatrix} = \begin{bmatrix} v\cos\theta \\ v\sin\theta \\ u \end{bmatrix}, \tag{13}$$

其中，x, y 和 θ 分别表示位置和航向。v 和 u 表示线速度和角速度。 v 对每个智能体分别固定，而 u可控。

仿真在大小为 $100 \times 100m$ 的环境中进行，每个环境中均有 20 个物体在区域 $[20, 80] \times [20, 80]$ 内均匀随机生成。所有机器人均从 $[10, 50]$ 开始。主智能体的目标位置为 $[90, 50]$。关于.....的信念每个目标都以较高的 $25m^{2}$方差进行初始化，初始均值也相应地受到扰动。主智能体以 $v_{PA} = 2ms^{-1}$的速度移动，探索智能体以 $v_{EA} = 4ms^{-1}$的速度移动。

B. 演示与比较

我们在第 IV-C 节中，通过考察采用每个探索智能体奖励的重复仿真中主智能体的任务成功率，验证了我们的方法。我们引入了仅由主智能体组成的“盲目”变体作为基准。盲目 PA 仅根据先验信念规划路径，不使用探索智能体提供的任何测量值。为确保公平比较，我们生成了 10 个随机环境，并考察主智能体轨迹失败的比例；轨迹失败定义为轨迹与真实值中的物体发生碰撞。结果如图 3 所示。

图 3 显示，SE 方法（第 IV-C.1 节）通常表现最佳，其次是互信息上置信界和 SI（第 IV-C.3 节和第 IV-C.2 节），然后是盲目方法。不出所料，由于无法消除环境不确定性，盲目方法在所有配置下的失败率最高，表现最差。特别是，SE 方法相较于盲目基准始终能带来超过两倍的提升。与互信息上置信界方法相比，在护送机器人数量较少时，SE 表现更好；在机器人数量较多时，两者表现相当。这是因为 SE 方法能够更好地利用有限信息，引导探索智能体前往与任务更相关的物体。对 PA 的任务表现进行显式考量所带来的更紧密协调，相较于互信息上置信界实现的探索-利用平衡，能够取得更好的表现。有趣的是，SI 方法的表现与互信息上置信界方法相当或更差。这是因为 SI 方法仅尝试验证 当前保护智能体的控制分布，因此无法区分事实上会失败的规划方案。

为了更好地理解所观察到的趋势，我们考察了系统行为随时间的变化，如图 4 所示。红色热力图表示对障碍物（黑色十字）的信念，并已归一化至信念的最高概率。因此，观测到障碍物时，对环境其余部分的信念看起来会降低，而障碍物本身的信念保持不变。

图 3 中盲目 PA 的糟糕表现可由第一行中 $t = 30s$ 处的失败作为例证。该失败是由于盲目 PA 受到物体较大的初始不确定性和受污染的均值的阻碍，而这些问题始终未得到消解，因此只能选择一条物体避让程度不足的路径。换言之，不确定性消解对于提升规划性能是必要的。

采用互信息上置信界方法（第二行）时，探索智能体会朝着降低目标不确定性的方向行动。在此过程中，在$t = 30s$，探索智能体会覆盖靠近 PA 的轨迹的障碍物簇，因为这些区域的信息丰富；然而，它们未能考虑与当前 PA 轨迹相关的特定信息，因此导致了与盲目情况类似的碰撞。

采用 SE 和 SI 方法时，探索智能体通常会降低更接近 PA 的轨迹的目标的不确定性。具体而言，可以看出，SI 方法覆盖了 $t = 30s$ 处附近的障碍物，使主智能体能够在刚刚超出碰撞范围的位置避开它们。

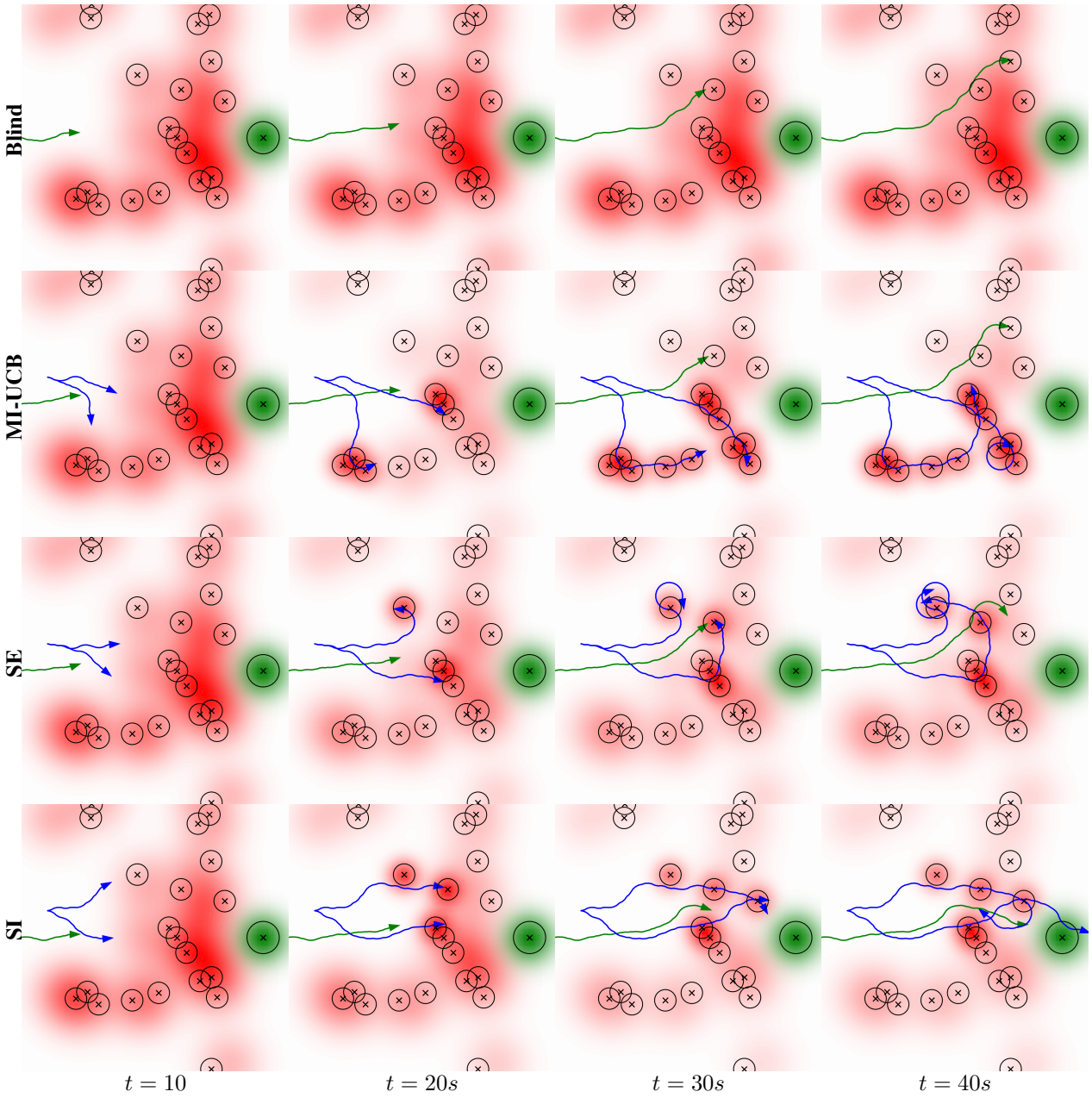

图 4. 两个护送智能体（蓝色）和一个主体智能体（绿色）使用四种奖励函数进行二维物体避障的比较。实线表示路径历史，三角形表示给定时刻 t 机器人所处的位置和航向。物体的真实值位置以 “x” 表示，圆圈表示这些物体的边缘。目的地 d 以绿色热力图表示。红色热力图表示对这些障碍物的信念。**注**：颜色强度按信念的最大概率归一化（即使观察到障碍物，也不会影响其他区域的信念概率，尽管其显示出来似乎有所降低）。

VI. 结论与未来工作

我们提出了协同护送问题，这是一种新颖的联合优化问题表述，其协调目标取决于独立智能体的行动。我们提出了 DecCEM，这是一种针对协同护航联合连续控制问题的新颖分布式解决方案。基于乘积分布近似，DecCEM 支持对总体控制分布进行任意参数化，从而将现有基于乘积分布的规划器从离散控制空间扩展至连续控制空间。

我们提出并评估了两种新的信息收集目标函数，供护送智能体使用，其任务是提高保护智能体的任务满足概率。与通用信息收集目标相比，这些目标函数实现了更高或相当的任务成功率。未来的工作将研究动态对抗性物体情况下的协同护送问题，以及 DecCEM 收敛性和最优性的理论分析。

参考文献

[1] D. P. Kroese、S. Porotsky 和 R. Y. Rubinstein，“连续多极值优化的交叉熵方法”，Methodol. Com- put. Appl. Probab.，第 8 卷，第 3 期，第 383–407 页，2006 年。[2] M. Kobilarov，“交叉熵运动规划”，Int. J. of Rob. Res.，第 31 卷，第 7 期，第 855–871 页，2012 年。[3] G. Best、O. Cliff、T. Patten、R. R. Mettu 和 R. Fitch，“Dec- MCTS：面向多机器人主动感知的去中心化规划”，Int. J. Robot. Res.，第 38 卷，第 2–3 期，第 316–337 页，2019 年。[4] D. H. Wolpert 和 C. E. M. Strauss，“利用概率集体推进分布式优化”，Adv. Complex Syst.，第 9 卷，2011 年。[5] Y. Lan、Z. Lin、M. Cao 和 G. Yan，“使用单轮车团队执行护送和巡逻任务的分布式可重构控制律”，载于 IEEE CDC 会议论文集，2010 年，第 5456–5461 页。[6] J.-M. Lien、O. B. Bayazit、R. T. Sowell、S. Rodriguez 和 N. M. Amato，“牧群行为”，载于 IEEE ICRA 会议论文集，2004 年。[7] E. Montijano 和 A. R. Mosteo，“利用分布式优化实现高效的多机器人编队”，载于 IEEE CDC 会议论文集，2014 年，第 6167–6172 页。[8] E. Montijano、D. Zhou、M. Schwager 和 C. Sagues，“无需全局参考系的分布式编队控制”，载于 ACC 会议论文集，2014 年，第 3862–3867 页。[9] D. Str¨ombom、R. P. Mann、A. M. Wilson、S. Hailes、A. J. Morton、D. J. T. Sumpter 和 A. J. King，“解决牧群问题：自主交互智能体的驱赶启发式方法”，J. R. Soc. Interface，第 11 卷，第 100 期，2014 年。[10] G. Antonelli、F. Arrichiello 和 S. Chiaverini，“多机器人系统的围捕/护送任务：理论与实验”，载于 IEEE/ASME AIM 会议论文集，2007 年。[11] ——，“围捕/护送任务”，IEEE Robot. Auto m. Mag.，第 15 卷，第 1 期，第 22–29 页，2008 年。[12] I. Mas、S. Li、J. Acain 和 C. Kitts，“多机器人集群空间控制中的围捕/护送与巡逻任务”，载于 IEEE/RSJ IROS 会议论文集，2009 年。[13] Y. Wu、S. Wu 和 X. Hu，“面向城市环境中持续监视任务的无人机与无人地面车辆协同路径规划”，IEEE Internet Things J.，第 8 卷，第 6 期，第 4906–4919 页，2021 年。[14] S. G. Manyam、K. Sundar 和 D. W. Casbeer，“空地车辆团队的协同路径规划——精确算法、变换方法与启发式方法”，IEEE Trans. Autom. Sci. Eng.，第 17 卷，第 1 期，第 537–547 页，2020 年。[15] P. Nilsson、S. Haesaert、R. Thakker、K. Otsu、C.-I. Vasile、A.-A. Agha-Mohammadi、R. M. Murray 和 A. D. Ames，“面向协作机器人团队的规范引导式火星主动探测”，载于 RSS 会议论文集，2018 年。[16] T. Sasaki、K. Otsu、R. Thakker、S. Haesaert 和 A.-a. Agha-Mohammadi，“在哪里进行测绘？面向火星探测的迭代式漫游车—直升机路径规划”，IEEE Robot Autom. Lett.，第 5 卷，第 2 期，第 2123–2130 页，2020 年。[17] L. Folsom、M. Ono、K. Otsu 和 H. Park，“不确定环境中漫游车—直升机团队的可扩展信息论路径规划”，Int. J. Adv. Robot Syst.，第 18 卷，第 2 期，2021 年。[18] K. M. B. Lee、F. H. Kong、R. Cannizzaro、J. L. Palmer、D. Johnson、C. Yoo 和 R. Fitch，“异构多机器人系统中同时探索与利用的上置信界”，载于 IEEE ICRA 会议论文集，2021 年。[19] N. Atanasov、J. Le Ny、K. Daniilidis 和 G. J. Pappas，“感知机器人进行信息获取：算法与误差界”，载于 IEEE ICRA 会议论文集，2014 年。[20] ——，“去中心化主动信息获取：理论及其在多机器人同步定位与地图构建中的应用”，载于 IEEE ICRA 会议论文集，2015 年。[21] B. Schlotfeldt、D. Thakur、N. Atanasov、V. Kumar 和 G. J. Pappas，“面向去中心化多机器人主动信息收集的随时规划”，IEEE Robot Autom. Lett.，第 3 卷，第 2 期，第 1025–1032 页，2018 年。[22] J. Wakulicz、H. Kong 和 S. Sukkarieh，“任意未知扰动下的主动信息获取”，载于 IEEE ICRA 会议论文集，2021 年。[23] A. Krause 和 C. Guestrin，“利用子模函数进行近最优观测选择”，载于 AAAI 会议论文集，第 7 卷，2007 年，第 1650–1654 页。[24] A. Hashemi、M. Ghasemi、H. Vikalo 和 U. Topcu，“面向大规模传感器网络中近最优传感器调度的随机贪心算法”，载于 ACC 会议论文集，2018 年，第 1027–1032 页。

[25] M. Ghasemi and U. Topcu，《有限预算下部分可观测马尔可夫决策过程的在线主动感知》，载于*IEEE CDC* 论文集，2019年，第6169–6174页。[26] M. Shamaiah、S. Banerjee 和 H. Vikalo，《贪心传感器选择：利用子模性》，载于*IEEE CDC* 论文集，2010年，第2572–2577页。[27] Y. Chen、L. Zhao、K. M. B. Lee、C. Yoo、S. Huang 和 R. Fitch，《广播你的弱点：多机器人的协作式主动位姿图SLAM》，*IEEE*机器人与自动化快报，第5卷，第2期，第2200–2207页，2020年。[28] K. C. To、F. H. Kong、K. M. B. Lee、C. Yoo、S. Anstee 和 R. Fitch，《根据集合预报和在线测量值估计空间相关海流》，载于*IEEE ICRA* 论文集。IEEE，2021年，第2301–2307页。[29] K. M. B. Lee、J. J. H. Lee、C. Yoo、B. Hollings 和 R. Fitch，《使用水下滑翔机进行羽流源定位的主动感知》，载于*ARAA ACRA* 论文集，2018年。[30] K. M. B. Lee、C. Yoo、B. Hollings、S. D. Anstee、S. Huang 和 R. Fitch，《利用稀疏 GPS 数据在线估计水下航行器的海流》，载于*IEEE ICRA* 论文集，2019年，第3443–3449页。[31] J. Wakulicz、K. M. B. Lee、C. Yoo、T. Vidal-Calleja 和 R. Fitch，《稀疏高斯过程回归中最坏情况误差最小化的信息规划》，载于*IEEE ICRA* 论文集，2022年。[32] S. Levine，《作为概率推断的强化学习与控制：教程与综述》，*arXiv*预印本 *arXiv:1805.00909*，2018年。[33] K. M. B. Lee、C. Yoo 和 R. Fitch，《作为概率推断的信号时序逻辑综合》，载于*IEEE ICRA* 论文集，2021年。[34] C. Yoo 和 C. Belta，《基于概率信号时序逻辑的控制》，*arXiv*预印本 *arXiv:1510.08474*，2015年。[35] C. Yoo、R. Fitch 和 S. Sukkarieh，《具有安全约束且可证明正确的随机运动规划》，载于*IEEE ICRA* 论文集。IEEE，2013年，第981–986页。[36] ——，《风场中燃料受限空中机器人的在线任务规划与控制》，国际机器人研究杂志，第35卷，第5期，第438–453页，2016年。