arXiv:2304.01094v2 [cs.RO] 1 Aug 2023

通过基于子目标的抽象，在部分地图中实现数据高效的导航策略选择

Abhishek Paudel 和 Gregory J. Stein

摘要——我们提出了一种新方法，用于在部分地图中快速、可靠地选择导航策略。我们利用近期提出的学习增强型、基于模型的子目标学习规划（LSP）抽象进行规划，使机器人能够重用导航过程中收集的数据，通过一种称为离线替代策略回放的过程，评估其他替代策略可能取得的表现。在部署期间，离线替代策略回放的成本可用于约束基于 LSP 的策略之间的选择，从而提升收敛速度、降低累积遗憾和平均导航成本。即使对未探索环境的性质仅有有限的先验知识，在模拟迷宫和类似办公室的环境中进行的实验表明，与基线老虎机方法相比，我们的累积遗憾至少降低了 67%，最高降低了 96%。

I. 引言

我们使机器人能够在部分建图环境中部署并导航至未探索目标时，快速从一组策略中找出表现最佳的策略。面向目标的部分建图环境导航是自主机器人的一项重要能力。若机器人需要在未知办公楼中前往远处的目标，就必须在行进过程中构建地图、规划穿越地图中的未知区域，并在这些区域显现后决定如何行动。在此类情境中实现高效导航，要求机器人能够推理其行动对遥远未来的影响，这通常称为长时域规划问题，可将其表述为部分可观测马尔可夫决策过程（POMDP） [1]。求解此类 POMDP 在计算上难以处理，因此，许多现有的此类情境规划方法都利用学习来引导机器人的行为 [2]–[7]。

在局部地图中导航时，学习通常依赖机器人搭载的传感器所采集的观测（例如图像和激光扫描数据），据此预测机器人动作的优劣，从而引导其采取在训练环境中表现出色的动作。然而，通常机器人可能拥有一系列由学习提供信息的策略，每种策略都在不同的环境中训练而成，并且必须从中选择一种，在部署期间引导其行动。使用其中一种策略进行规划后，机器人必须根据该策略的累积表现，判断是再次选择该策略，还是切换到另一种策略，以期降低导航成本：这就是将模型选择 应用于策略选择的一个实例。

Abhishek Paudel 和 GregoryJ. Stein 任职于美国弗吉尼亚州费尔法克斯的乔治梅森大学计算机科学系。（{apaudel4, gjstein}@gmu.edu）

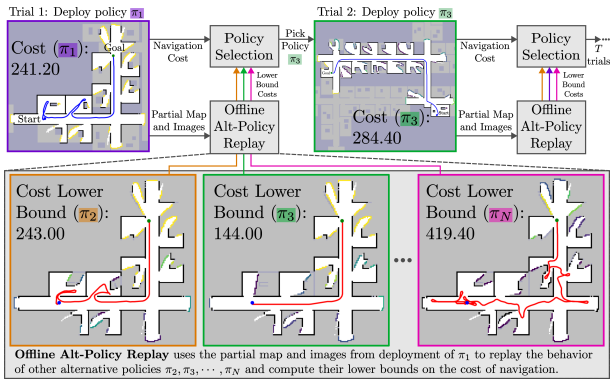


图 1。**我们用于在部分地图中进行数据高效策略选择的方法概览**。我们的方法依赖于离线替代策略回放，这是一种在部署后计算替代策略导航成本下界的过程，并利用这些下界约束选择。

模型选择方法 [8]–[15] 旨在通过多次部署每种策略并评估其表现，从一组预定义行为中找出表现最佳的模型。例如，多臂老虎机算法 [11]–[13], 会在利用 （选择迄今表现最佳的学习型策略）与探索 之间权衡，以挑选另一种可能 提升表现的策略。上置信界（UCB）老虎机算法 [11] 根据每种可用策略的观测表现及其部署次数，计算其潜在表现的统计界限。然而，多臂老虎机算法通常是黑箱，因此，收紧这些界限往往要求机器人进行多次规划尝试，并在排除这些策略之前承受糟糕的行为表现。由于耗时且表现不佳的导航试验代价高昂，数据效率对于此类选择方法在实践中发挥作用至关重要。

相反，我们重用在一次试验中收集的数据，评估其他备选策略本可以取得的表现，并利用这项评估收紧策略选择的界限：这一过程称为离线替代策略回放，如图 1 所示。机器人根据每次试验中收集的信息，无需从环境中收集额外信息，就能回放其每个其他备选策略本可以取得的表现。我们的离线替代策略回放方法对导航试验中未见过的空间作出乐观假设（即机器人只会在未探索空间中花费最少的时间），从而能够计算每项策略可能表现的严格下界。这些信息对类似多臂老虎机的选择形成约束——探索绝不应选择一项迄今为止在部署过程中不可能改善表现的策略——并提高策略选择的数据效率。

我们的离线替代策略回放方法依赖于这样一种规划策略：仅凭迄今为止机器人各次试验中收集的有限信息，就能准确界定该策略本可取得的表现。并非所有规划策略都适合这种回放，包括通过深度强化学习训练的无模型方法 [16], [17], ，这类方法可能会因输入变化而变得脆弱 [18]–[20]。本研究的一项关键洞见是，近期提出的基于模型、由学习增强的基于子目标的学习规划 （ LSP）方法（Stein 等人 [2] ）非常适合这一用途。在 LSP 抽象下，学习仅用于对未探索空间的统计量作出与机器人姿态无关的预测：例如，某个特定区域通向未探索目标的概率。由于其学习方法能够适应机器人位置的变化，对 LSP 策略进行离线回放，可以得到准确的成本下界，并据此约束类似多臂老虎机的选择。

本文提出了一种名为离线替代策略回放的数据高效策略选择方法。该方法利用导航过程中收集的信息，评估其他备选策略可能取得的表现，从而约束对一组基于子目标的学习规划器 [2]进行的类似多臂老虎机的选择。我们的机器人部署在模拟迷宫和类似办公室的环境中，使用多种学习信息驱动的策略：每种策略都依赖于基于子目标的学习规划抽象，并在不同环境中训练。实验表明，与通常收敛较慢且累积遗憾高于我们方法的基线 UCB老虎机方法相比，我们的方法能够以少得多的试验次数快速降低平均导航成本。结果验证了我们的方法优于 UCB 多臂老虎机，并且通常能够更快地完成选择。

II. 相关工作

不确定性下的规划 部分可观测马尔可夫决策过程 [1], [21] 为表示不确定性下的导航和探索提供了通用框架 [22], [23]。然而，这类问题在计算上难以求解，因此会利用学习进行规划 [2]–[7]。深度强化学习方法也广泛用于部分已知地图环境中的规划 [24]–[26],，但通常局限于短时域规划。

模型选择 强化学习中的在线模型选择方法将多臂老虎机推广开来，旨在从一组可用模型中找出最佳模型。多臂老虎机算法 [11]–[13], 通常是黑箱方法，在利用与探索之间进行权衡，因此机器人需要经历多次表现不佳的试验，才能找到更好的策略。

运行时监控 所谓运行时监控方法旨在评估学习引导的机器人行为在部署时的可靠性 [16], [29]–[31]，以决定是否应继续采用学习方法，或在性能下降时改用备用策略。然而，这类方法大多着眼于解决与局部感知不足相关的问题，或旨在监控底层学习模型的原始预测性能，而非整体任务表现 [30], [31], ，因此限制了它们评估长时域行为优劣的能力。

III. 问题表述

A. 部分地图中的目标导向导航

我们的机器人置于部分建图环境中，任务是在最小化期望距离的前提下，到达未探索空间中的一个点目标。我们将此问题表述为部分可观测马尔可夫决策过程（ POMDP） [1]。信念状态 b_t 包含部分观测到的地图以及机器人在时间 t 的位姿。机器人配备了平面激光扫描仪，因此能够可靠地绘制已到访区域周围的地图，并以高精度确定自身位置；这属于近期提出的一类 POMDP，称为局部 可观测马尔可夫决策过程 [32]。机器人的行为由策略 π 引导，该策略根据信念 b_t指定机器人应采取的动作；信念由部分占据地图 m_t 和机器人收集的视觉观测集合构成。在部署期间，性能以多次试验中到达目标所行进距离的平均值衡量；一次试验是指在此前未见过的地图中，从起点到目标的一次完整行程。

B. 部署期间的策略选择

我们假设机器人可以使用一组策略$\mathcal{P} = \{\pi_1, \pi_2, \cdots, \pi_N\}$，其目标是在部署期间选择使期望成本最小的策略，

$$\pi^* = \underset{\pi \in \mathcal{P}}{\operatorname{argmin}} \, \mathbb{E}[C(\pi)] \tag{1}$$

其中， $\mathbb{E}[C(\pi)]$ 是策略 $C(\pi)$在部署期间的期望成本。然而，通常我们无法直接获取期望成本，因此必须在部署期间通过执行多次试验来估算。

C. 黑盒策略选择在局部地图规划中的局限

黑盒策略选择方法（例如许多多臂老虎机算法）旨在通过平衡最小化迄今观察到的平均成本与探索来选择策略，从而偶尔选择有望提升性能的策略。对于试验 $k+1$，上置信界（UCB）[1] [11], [33] 多臂老虎机算法根据以下规则选择下一个策略$\pi^{(k+1)}$

$$\pi^{(k+1)} = \underset{\pi \in \mathcal{P}}{\operatorname{argmin}} \left[\bar{C}_k(\pi) - c\sqrt{\frac{\ln k}{n_k(\pi)}} \right] \tag{2}$$

其中， $\bar{C}_k(\pi)$是策略 π 在第 1 至k 次试验中被选中时的平均成本， $n_k(\pi)$是截至第 k次试验，策略π 被选中的次数， $c > 0$ 则是控制探索与利用之间平衡的参数。

[1]式（2）使用下置信界（LCB）而非上置信界（UCB），因为我们的性能估计以代价 而非奖励 表示，因此需要最小化。因此，我们使用更常见的术语 UCB 泛指这种方法，而非仅指上置信界本身。

然而，对于不确定性下的导航，每次试验的执行成本都很高，而且试验结果取决于环境地图，因此方差通常很大。因此，通过式（2）进行策略选择往往难以快速收敛，这限制了该方法在实践中的实用性。要缩小各策略预期性能的置信界，就必须部署这些策略，且往往需要多次部署；因此，在排除表现不佳的策略之前，机器人可能不得不反复尝试表现不佳的行为。为此，我们旨在通过白盒选择策略提高选择的数据效率：该策略利用离线替代策略回放，计算各策略在迄今为止的试验中本可以取得的表现的界，并以此约束选择过程。

IV. 概述：通过离线替代策略回放实现数据高效策略选择

若要加快策略选择，我们必须能够快速收紧每个策略预期性能的界限，优先以更少的试验选择最有前景的策略。因此，我们希望通过确定某个替代策略原本可能取得怎样的表现（即由它来控制机器人时的表现），来约束策略选择。虽然我们无法重新部署机器人，以另一种策略重复同一试验，但可以利用试验期间收集的信息，仔细分析替代行为，并在不进行部署的情况下确定其性能下界。

我们希望利用部署时离线替代策略回放，计算试验期间未控制机器人行为的策略的性能下界。利用策略 π 在试验期间收集的信息，我们可以回放所有其他替代策略 $\pi' \in \mathcal{P} \setminus \pi$原本可能取得的表现。离线替代策略回放对导航试验期间未探索的空间作出乐观假设（即机器人在未探索空间中只会花费极少时间），从而计算迄今为止每种策略平均性能的严格下界（见第 VII-C 节、VII-D 节）。我们利用这一¯均值 C^{lb} 下界来约束 UCB 多臂老虎机选择，从而避免选择那些无法提升性能的策略。我们的受约束UCB 多臂老虎机算法根据以下方式进行选择：

$$\pi^{(k+1)} = \underset{\pi \in \mathcal{P}}{\operatorname{argmin}} \left[\max \left(\bar{C}_k^{\text{lb}}(\pi), \bar{C}_k(\pi) - c\sqrt{\frac{\ln k}{n_k(\pi)}} \right) \right] \quad (3)$$

直观而言，式 (3) 总是选取¯以下两种下界中较紧（即较高）的一种：(i) C_k^{lb}通过离线替代策略回放计算出的下界，以及 (ii)UCB 算法计算出的下界，如式 (2) 所示；随后在该下界上进行最小化，以选出下一次试验所用的策略。

V. 离线策略回放需要一种能够适应视角变化的规划方法

为了通过我们的约束UCB多臂老虎机算法（式（3））进行选择，需要审视替代行为，以确定代价下界 $C_k^{\text{lb}}(\pi)$；这要求我们能够在不实际部署机器人的情况下，离线回放替代策略下机器人的行为。一般而言，离线回放策略要求能够生成机器人可能从未到访过的位姿所对应的观测；对于该领域的许多学习驱动的规划策略而言，这些观测无法准确反映该策略若曾控制机器人时的行为。许多不确定性下的视觉驱动导航方法，尤其是依赖深度强化学习 [7]‚[17],的方法，需要获取机器人在原始试验中未曾到访的位姿和视点处的观测（图像），而且即使发生细微变化也可能变得脆弱 [18], ，因此回放此类策略不太可能得到准确的成本下界。因此，我们转而需要一种能够适应视角变化的稳健规划方法，以及一种能够适应机器人位姿及相应观测的细微变化、并能可靠抵达目标的策略，即使是在学习导致行为不佳的环境中也如此。

VI. 预备知识：基于子目标学习的基于模型规划抽象

一个关键见解是，Stein 等人的子目标学习规划（LSP）非常适合离线替代策略回放，同时也适用于在部分建图环境中进行长时域规划。LSP 是一种高层规划框架，其中部分建图环境中的导航被表述为部分可观测马尔可夫决策过程（POMDP）。在这一抽象中，子目标代表机器人前往某个前沿的高层动作；前沿即地图中自由与未观测区域之间的边界。每个与子目标对应的动作 a_t 都具有三个相关属性：子目标通向目标的概率 P_S (a_t)、经由该子目标到达目标的期望成本 R_S(a_t)，以及若子目标未能通向目标时，失败或探索所产生的期望成本R_E(a_t)。一个以 θ为参数的神经网络通过监督学习进行训练，以根据以子目标为中心的图像观测来估计子目标属性 $\mathcal{R}_\theta = \{P_{S,\theta}, R_{S,\theta}, R_{E,\theta}\}$ 。式 (4) 所示的贝尔曼方程分解形式用于计算每个动作的期望成本，

$$\begin{aligned} Q_\theta(b_t, a_t) &= D(b_t, a_t) + P_{S,\theta}(a_t) R_{S,\theta}(a_t) \\ &+ (1 - P_{S,\theta}(a_t)) \left[R_{E,\theta}(a_t) + \min_{a \in \mathcal{A}(b_t) \setminus a_t} Q_\theta(b_t, a) \right] \end{aligned} \quad (4)$$

其中， D(b_t， a_t) 是前往某个前沿的已知代价。机器人的策略是最小化期望成本：

$$\pi_\theta(b_t) = \underset{a \in \mathcal{A}(b_t)}{\operatorname{argmin}} Q_\theta(b_t, a) \quad (5)$$

由于学习仅用于利用朝向子目标的全景图像，对与机器人姿态无关的子目标属性进行预测，因此 LSP 能够适应机器人位置及相应图像观测的变化，因而适用于离线替代策略回放。

A. 基于子目标的学习规划

在每个时间步，机器人通过式（4）选择一个高层动作 a_t ，以指定机器人将前往的子目标。选择高层动作后，

机器人通过 A* [34] 在已观测地图上计算代价网格，并选择一个能够最大程度推进至子目标的低层（短时域）运动基元。机器人随后（i）执行该基元动作，（ii）根据激光扫描仪的观测更新局部地图，（iii）重新计算子目标（并在必要时根据机器人搭载设备采集的全景图像重新计算子目标属性 P_S、 R_S 和 R_E），最后（iv）通过式（4）重新规划。由于 LSP 规划会持续探索未探索区域，直至到达目标，因此只要存在可行路径，就能保证到达目标，即使学习过程导致了不佳行为也是如此。

B. 网络架构与训练

神经网络的输入包括以子目标为中心的 128×512 RGB全景图像、子目标的相对位置以及目标的相对位置。我们的神经网络架构和训练流程与Bradley等人 [35]的方法基本一致。输入图像首先经过4个卷积层进行编码，然后与表示子目标和目标相对位置的特征拼接。拼接后的特征经过9个卷积层和5个全连接层，输出3个子目标属性： P_S、 R_S和 R_E。

为生成训练数据，我们在此前未见过的地图中进行了数百次试验。在这些试验中，机器人通过非学习型启发式规划器从起点导航至目标，并在每一步收集环境图像。对于观测到的子目标， P_S、 R_S和 R_E 的目标标签通过底层已知地图计算得出。 P_S 的标签表示是否存在一条经由某个子目标通往目标的路径。代价标签表示：若可经由某个子目标到达目标，则为通过未知空间到达目标的行驶距离（对应 R_S）；若无法经由某个子目标到达目标，则为启发式代价，用于近似机器人遇到死胡同前需要行驶的距离（对应 R_E）。有关数据收集和训练的更多细节，请参见 [2], [35]。

VII. 方法：通过对子目标的学习进行离线替代策略回放

我们的方法利用试验期间收集的信息，对策略进行离线回放。在离线替代策略回放期间，我们对未探索空间作出乐观假设或单连通假设，以计算这些策略的近似成本下界；随后利用该下界约束 UCB 多臂老虎机，从而加快策略选择。

A. 试验期间的信息收集

在一次试验（即从起点到目标的一次导航）期间，我们记录离线回放其他策略所需的信息。对于试验 k，策略 $\pi^{(k)}$ 决定行为，并生成一份记录 $\mathcal{Z}_k$，其中包含一个元组列表（X_t， I_t）： $X_t = (x_t, y_t, \phi_t)$ 是时间 t 时的机器人位姿，I_t 是机器人在时间 t采集的全景图像。记录 $\mathcal{Z}_k$ 还包括到达目标后获得的部分已知占据栅格地图 m_{final}。随后，记录$\mathcal{Z}_k = \{(X_t, I_t)_{t\in 1\cdots M}, m_{\text{final}}\}$ 用于对由其他每个策略引导的机器人行为执行离线回放。

B. 离线替代策略回放概述

一次试验完成后（例如，试验 k），机器人使用其在试验期间收集的记录 $\mathcal{Z}_k$，尝试回放所有其他 LSP策略 $\pi' \in \mathcal{P} \setminus \pi^{(k)}$ 的行为。离线替代策略回放与通过 LSP 进行在线规划的过程类似，如第六节-A所述：在每个回放时间步 τ，机器人使用模拟激光扫描更新其地图；该激光扫描通过在二维局部地图 m_{final}中进行射线投射计算得到。随后，机器人重新计算可用的子目标，每个子目标都对应于自由空间与未知空间之间的边界。

为计算与探索未探索空间特定区域相关的子目标属性 P_S, R_S, R_E，机器人需要将全景图像输入其神经网络（第VI-B节）。尽管在离线替代策略回放期间，环境中的每个位置都无法获取图像，但子目标属性可以根据存储在 $\mathcal{Z}_k$ 中的附近图像进行估计。具体而言，我们从 $\mathcal{Z}_k$ 中找到能够看到目标子目标、且位姿 X_{nearest} 距当前（回放）位姿 X_τ 最近的位置，然后使用该位置关联的图像 I_{nearest} 估计子目标属性。利用估计出的子目标属性，机器人通过公式（4）选择一个高层动作，执行一个低层运动基元以朝选定的子目标动作前进，并重复这一过程。关键在于，如果使用已部署策略 $\pi^{(k)}$ 自身的记录 $\mathcal{Z}_k$ 对其进行回放，我们就能精确复现其行为。

C. 计算近似成本下界

在离线替代策略回放期间，机器人导航至目标时，可能会尝试进入机器人使用策略 $\pi^{(k)}$ 进行规划时未探索的空间，因此该空间在局部地图 m_{final} 中是未知的。由于我们不了解这片空间，无法准确知道其中会发生什么，因此我们对可能发生的情况做出保守假设：假设机器人要么（i）通过未探索空间，以最短可能距离到达目标；要么（ii）立即掉头，并必须选择另一条路线前往目标。根据我们对未探索空间的了解，例如环境是否已知为单连通，或是否存在通往目标的未发现的捷径，我们可以做出不同的假设，并据此计算下界代价。此处（并在图2中进行了可视化）我们将讨论这些不同的假设、它们的影响，以及各自适用的情形。

乐观成本下界 如果我们对未探索空间的结构没有任何先验信息，就会对未探索空间做出乐观假设：所有未探索空间都是畅通的，因此可能提供一条通往目标的路线。在这一假设下，每当离线回放的策略中的动作导致机器人离开已知空间（经由子目标 s'）时，我们就假设机器人可能 已经通过 m_{final} 中经过 s'的最短可能路径到达目标。图2(b)展示了两条此类替代路径的示例。机器人尝试离开已知空间时，我们会计算乐观路径成本，将该前沿屏蔽（以便机器人必须继续在已知空间中探索），然后继续导航。我们会生成此类乐观

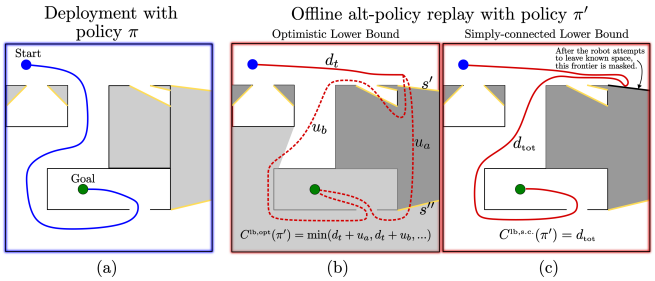


图2。**下界成本近似**：(a) 策略 π 在一次试验中引导机器人。(b) 在离线替代策略回放期间，策略 π' 通过子目标 s' 尝试离开已知空间，并沿提议路径 u_a尝试到达目标。回放期间获得的此类路径中的最小值给出了策略 π' 的乐观下界。(c) 在单连通假设下，策略 π' 在离线回放期间行进的净距离就是下界代价。

路径成本：每当离线回放的机器人尝试离开已知空间时，都会生成一次。这些路径中最短的一条就是试验 k中回放策略π' 的乐观下界代价 $C^{\text{lb, opt}}_k(\pi')$。

单连通成本下界 如果我们事先知道环境是单连通的，就可以得出更紧的代价下界，因为到达目标的最短路径应存在于 m_{final} 的已知空间内，而离开观测空间的路线将无法到达目标。在离线回放导航期间，每当机器人尝试离开已知空间时，它试图通过的边界都会在允许其通过之前被遮蔽为障碍物，迫使机器人回头并寻找通往目标的替代路线，如图2(c)所示。到达目标所行进的总距离就是回放策略 π'的单连通下界代价 $C^{\text{lb, s.c.}}_k(\pi')$。

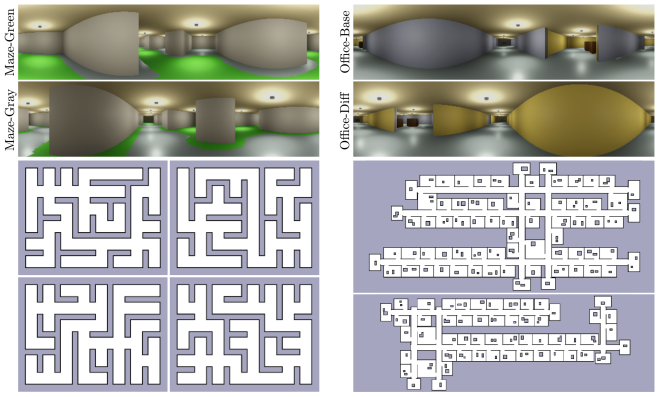
加权近似成本下界在许多环境中，乐观下界并不特别紧，因此无助于加快选择；然而，如果无法确定环境是否为单连通环境，单连通下界可能过高，因而不再是成本下界。通常，我们可以利用对环境的先验认知，判断机器人在试验过程中未能看到的区域内存在通往目标的替代捷径的概率。我们引入一个参数 p_{short} ，表示存在一条通往目标的更短路径的概率。此参数用于计算第三种近似下界，其定义为乐观下界与单连通下界的加权和：

$$C^{\text{lb,wgt}}_{k,p}(\pi') = p_{\text{short}} C^{\text{lb,opt}}_k(\pi') + (1 - p_{\text{short}}) C^{\text{lb,s.c.}}_k(\pi') \quad (6)$$

虽然加权近似成本下界可能并非严格意义上的下界，因此可能违反 UCB 多臂老虎机选择所提供的渐近遗憾界保证，但我们将表明，这种近似下界代价有助于取得良好的经验性能（第八节 B）。

D. 合并观测与回放下界成本

随着试验的进行，我们计算均值 $\bar{C}^{\text{lb}}$的近似下界，该下界结合了策略 π 截至试验 k 的平均性能 $\bar{C}_k(\pi)$ 以及回放得到的均值下界 $\bar{C}^{\text{lb, rep}}$，即通过离线



(a) Maze Environments (b) Office Environments

图 3. **模拟环境**。来自仿真环境的机器人视角全景图像（上方两行），以及来自 (a) 迷宫和 (b) 类似办公室环境的地图样例。我们的所有实验均在使用 Unity游戏引擎渲染的模拟环境中进行。

替代策略回放，依据第 VII-C 节定义的某一下界计算出的平均成本：

$$\bar{C}^{\text{lb}}_k(\pi) = \frac{n_k(\pi)\bar{C}_k(\pi) + n^{\text{rep}}_k(\pi)\bar{C}^{\text{lb,rep}}}{n_k(\pi) + n^{\text{rep}}_k(\pi)} \quad (7)$$

其中， $n_k(\pi)$ 和 $n^{\text{rep}}_k(\pi)$ 分别表示每种策略被部署和离线重放的次数。

基于第四节所讨论的经修改的 UCB 多臂老虎机公式的受约束策略选择，利用这个下界 ¯bound $\bar{C}^{\text{lb}}_k(\pi)$ 快速识别性能最佳的策略。

VIII. 实验结果

我们在模拟迷宫和类似办公室的环境中进行实验（图 3），评估我们的受约束的策略选择方法 Const-UCB（公式 (3) ），并将其与基线 UCB 多臂老虎机选择（公式 (2) ）进行比较。我们进行了 200 次部署，每次部署包含 100 次导航试验，每次试验均在机器人尚未见过的不同程序生成地图中进行。每次部署开始时，机器人随机选择一种策略。后续试验所采用的策略通过我们的 Const-UCB 方法或 UCB 多臂老虎机来选择。我们的方法 Const-UCB 和基线 UCB老虎机方法在所有实验中均使用探索参数 $c = 100$ ；该参数是在留出的测试集上通过经验选择的，以使基线 UCB老虎机方法取得良好表现。我们注意到，结果对 c 的变化并不特别敏感。此外，我们还展示了在不进行策略选择的情况下单独部署每种策略的结果，以说明部署时进行策略选择对于在不同环境中取得良好表现的必要性。

A. 以迷宫为中心的结果

我们的模拟迷宫环境由随机生成的单连通迷宫组成，机器人需要从起点导航至未探索目标。每张生成地图都各不相同，起点和目标位姿也经过随机设定。为研究我们方法的通用性和有效性，我们设计了三种迷宫环境变体，并分别在其中训练基于 LSP 的策略（第六节）。如图 3 所示，这三种迷宫环境变体如下：

	部署环境	Maze-Green			Maze-Gray			Maze-Random		
	试验次数 (k)	$k=10$	$k=40$	$k=100$	$k=10$	$k=40$	$k=100$	$k=10$	$k=40$	$k=100$
Average Nav. Cost (mean)	UCB-Bandit（基线）	250.04▲	190.08◆	170.88■	230.66▲	187.71◆	172.10■	290.89▲	216.87◆	194.20■
	Const-UCB:$C^{\text{lb, opt}}$（我们的）	175.38▲	159.14◆	155.08■	173.96▲	163.50◆	158.81■	209.44▲	186.21◆	181.20■
	Const-UCB:$C^{\text{lb}_{\text{s.c.}}}$（我们的方法）	**163.56**▲	**153.12**◆	**152.13**■	**165.62**▲	**158.06**◆	**155.94**■	**197.30**▲	**182.14**◆	**179.66**■
	最佳单一策略[2]	150.39	150.39	150.39	154.03	154.03	154.03	177.31	177.31	177.31
Cumul. Regret (mean)	UCB-Bandit（基线）	1327.0△	3105.9◇	4761.4□	994.1△	2444.8◇	3879.4□	1472.0△	3391.9◇	4892.1□
	Const-UCB：$C^{\text{lb, opt}}$（我们的方法）	369.1△	813.4◇	1180.5□	278.2△	698.9◇	1076.9□	417.8△	918.6◇	1249.1□
	Const-UCB:$C^{\text{lb}_{\text{s.c.}}}$（我们的方法）	**259.2**△	**446.6**◇	**573.3**□	**208.4**△	**419.1**◇	**565.0**□	**309.5**△	**600.6**◇	**786.9**□

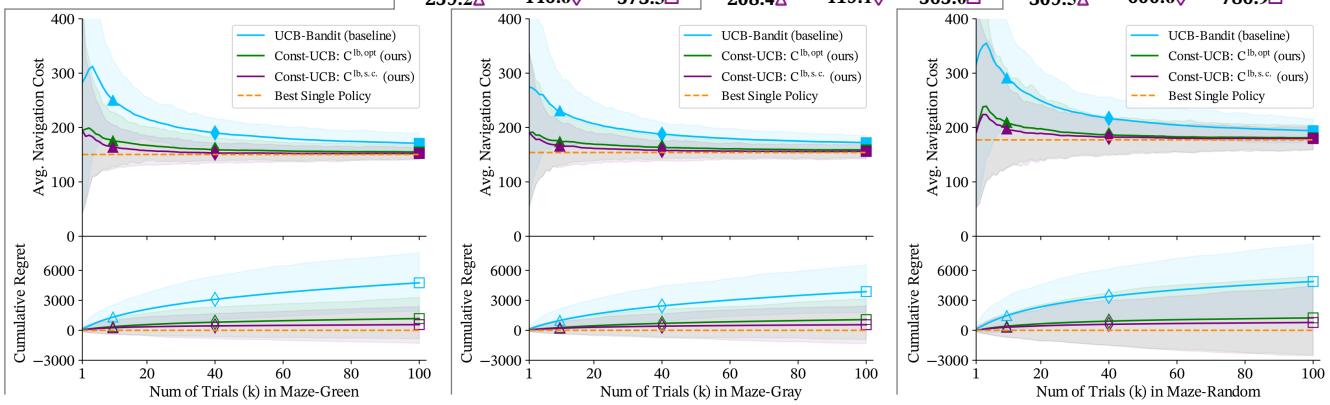


图4。**图3(a)中迷宫环境部署的平均导航成本（均值）和累积遗憾（均值）**。每次部署由100次随机导航试验组成，每次试验均在此前未见过的迷宫中进行。均值成本和遗憾是在200次随机部署上计算得到的。对于我们的方法 Const-UCB，我们展示了如第 VII-C 节所述的乐观 $C^{\text{lb, opt}}$ 和单连通 $C^{\text{lb, s.c.}}$ 下界的结果。实线表示均值，阴影区域表示 10^{th} 至 90^{th} 百分位数。符号：三角形、菱形和正方形分别表示在每个环境的表格和图中第 10^{th}、40^{th} 和 100^{th} 次试验时的平均成本（实心）和累积遗憾（空心）。

表I 各策略在无策略选择的以迷宫为中心的环境中的平均导航成本

	LSP-Maze-Green	Maze-Gray	Maze-Random
非学习型	206.05	194.37	177.31
LSP-Maze-Green	150.39	483.20	557.99
LSP-迷宫-灰色	618.87	154.03	418.98
LSP-迷宫-随机	231.71	238.22	180.23

Maze-Green 地面通常为灰色，但地面上有一条绿色路径连接起点和目标。因此，绿色是一个信号，经验丰富的智能体应能识别出它指向未探索目标。

Maze-Gray 与迷宫-绿色环境类似，但地板和路径的颜色互换：地板为绿色，通往目标的路径为灰色。因此，这种环境应会误导在 Maze-Green 中训练的策略。

Maze-Random灰色地板上的绿色路径随机放置，并非通往目标的可靠路线。

我们按照第VI-B节所述流程，在每种环境中训练一个基于 LSP 的策略，得到 LSP-Maze-Green、LSP-迷宫-灰色和 LSP-迷宫-随机。每个基于 LSP 的策略都在 500 个随机生成的迷宫上进行训练。导航试验中的每张迷宫地图也各不相同，且不属于训练集。我们还部署了一种非学习型乐观基线策略，供我们的方法选择；该非学习型策略不使用神经网络，而是利用关于环境的乐观启发式来计算子目标属性，即每个子目标都可能通向未探索目标：即 $P_S = 1$。

我们在每种迷宫变体中进行了100次试验，以评估策略选择方法。每次试验都包含一个随机生成的迷宫，该迷宫未出现在相应策略的训练过程中。每次试验的策略选择评估[3]汇总了200次随机部署的结果以计算统计量；我们计算均值和上下 10^{th}百分位数，并在图4中展示结果。我们展示各方法截至 k^{th} 试验的平均导航成本和累积遗憾。UCB-Bandit 是采用式（2）的基线策略选择方法，Const-UCB 则对应采用式（3）的受约束策略选择方法。

图4中的结果表明，在所有环境中，我们的受约束 UCB（Const-UCB）老虎机方法都能加快策略选择速度，与仅使用UCB-Bandit相比，只需少得多的试验即可降低平均导航成本。因此，与UCB 多臂老虎机相比，我们的 Const-UCB方法累积的平均累积遗憾显著更低：在 Maze-Green中低88%，在Maze-Gray中低85%，在 Maze-Random中低84%，这是因为它能迅速排除表现不佳的策略。迷宫环境被构建为单连通环境，采用单连通下界 $C^{\text{lb, s.c.}}$ 进行选择可取得最佳表现。即使我们事先不知道这些环境是单连通的，使用乐观下界 $C^{\text{lb, opt}}$ 的选择过程仍优于基线UCB多臂老虎机，展现了我们方法的实用性。

[2]对于每种环境，“最佳单一策略”指在该环境中部署时产生最低成本的策略。最佳单一策略产生的成本在表I和表II中以带下划线标出。

[3]为提高计算统计量的计算效率，每次部署都会从 150 个评估场景中随机抽取 100 个不同的场景，并在为每种策略选择模型之前预先运行这些场景。

	部署环境	Maze-Green			Office-Base			Office-Diff		
	试验次数（k）	$k=10$	$k=40$	$k=100$	$k=10$	$k=40$	$k=100$	$k=10$	$k=40$	$k=100$
Average Navigation Cost (mean)	UCB-Bandit（基线）	213.90▲	184.62◆	168.87■	648.02▲	503.31◆	470.43■	638.02▲	508.72◆	480.22■
	Const-UCB:$C^{\text{lb},\text{opt}}$（我们的方法）	174.03▲	158.88◆	155.03■	642.21▲	505.00◆	469.48■	622.17▲	508.17◆	475.46■
	Const-UCB:$C^{\text{lb},\text{wgt}}_{p=0.5}$（我们的方法）	165.65▲	154.23◆	152.63■	467.63▲	451.30◆	446.58■	**476.77**▲	**456.97**◆	**452.43**■
	Const-UCB:$C^{\text{lb},\text{s.c.}}$（我们的方法）	**163.01**▲	**152.91**◆	**152.03**■	**454.22**▲	**446.93**◆	**444.77**■	483.54▲	458.35◆	454.27■
	最佳单一策略[2]	150.39	150.39	150.39	442.95	442.95	442.95	424.84	424.84	424.84
Cumul. Regret (mean)	UCB-Bandit（基线）	736.7△	2138.7◇	3603.1□	3180.7△	6186.5◇	8495.4□	3389.6△	6978.6◇	10948.8□
	Const-UCB:$C^{\text{lb},\text{opt}}$（我们的方法）	332.2△	760.7◇	1121.3□	2003.2△	5053.6◇	7309.7□	2374.3△	5848.0◇	9615.9□
	Const-UCB:$C^{\text{lb},\text{wgt}}_{p=0.5}$（我们的方法）	274.9△	504.0◇	674.8□	222.1△	568.5◇	847.5□	**719.6**△	**1783.2**◇	**3599.2**□
	Const-UCB:$C^{\text{lb},\text{s.c.}}$（我们的方法）	**249.8**△	**428.9**◇	**546.4**□	**74.5**△	**230.9**◇	**327.4**□	749.3△	1887.6◇	3737.3□

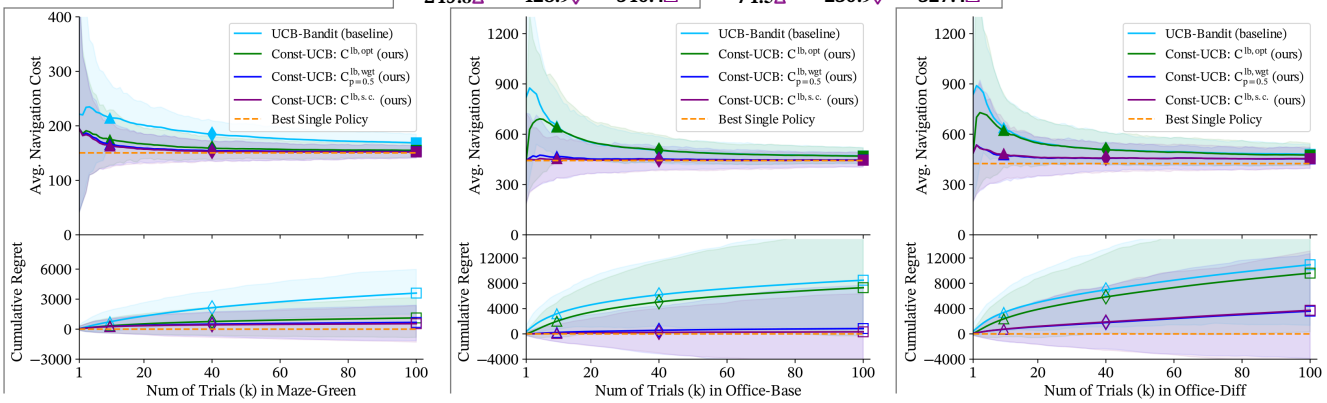


图 5. **以办公室为中心的环境中部署的平均导航成本（均值）和累积遗憾（均值），图 3(b)**。每次部署包含 100 次随机导航试验，每次试验均在此前未见过的地图中进行。均值成本和遗憾是在 200 次随机部署中计算得出的。对于我们的方法 Const-UCB，我们展示了第 VII-C 节所述的乐观 $C^{\text{lb},\text{opt}}$ 加权 $C^{\text{lb},\text{wgt}}$ 和单连通 $C^{\text{lb},\text{s.c.}}$ 下界的结果。实线表示均值，阴影区域表示从 10^{th} 到 90^{th} 百分位数。符号：三角形、菱形和正方形分别表示表格和图中各环境在第 10^{th}、40^{th} 和 100^{th} 次试验时的平均成本（实心）和累积遗憾（空心）。

表 II 未进行策略选择时以办公室为中心的环境中各策略的平均导航成本

	Maze-Green	Office-Base	Office-Diff
非学习型	206.05	448.60	476.88
LSP-Maze-Green	150.39	1667.70	1759.76
LSP-Office-Base	258.76	442.95	641.88
LSP-Office-Diff	306.45	942.62	424.84

B. 以办公室为中心的结果

我们在程序生成的类办公室环境中模拟导航。这些环境由随机生成、彼此连通的走廊组成，走廊中设有房间，旨在模拟摆放着家具类杂物的办公室。每张生成地图都是独一无二的，起点和目标位姿均为随机设定。图3展示了视觉模拟器中的办公室示例。我们针对两种办公室环境变体进行设计和实验：

Office-Base 走廊墙壁涂成灰色，房间墙壁则涂成黄色。

Office-Diff走廊与房间的墙壁颜色互换，使其在视觉上与 Office-Base 不同。

根据第VI-B节中的流程，在各自对应的环境中训练了两种基于 LSP 的策略：LSP-Office-Base 和 LSP-Office-Diff。与迷宫实验类似，办公室环境中的每种基于 LSP 的策略均在 500 个随机生成的办公室中进行训练，同时部署于另一组不同的地图上。除这两种策略外，我们还部署了乐观的非学习策略和 LSP- Maze-Green 策略（两者均见第VIII-A节），后者在 Maze-Green 中训练，用于展示跨环境策略选择，从而体现我们方法的灵活性。部署和统计数据生成遵循迷宫实验中所述的相同流程（第VIII-A节）。

图 5 的结果表明，在我们所有的实验中，与基线 UCB 多臂老虎机方法相比，我们的受约束UCB（Const-UCB）多臂老虎机方法只需更少的试验即可降低平均导航成本，更快地收敛到接近表现最佳策略的成本水平，而且表现从不逊于多臂老虎机。因此，与 UCB 多臂老虎机相比，我们的 Const-UCB 方法累积的平均累积遗憾显著更低：在 Maze-Green 中低 85%（界限为 $C^{\text{lb},\text{s.c.}}$），在 Office-Base 中低 96%（界限为 $C^{\text{lb},\text{s.c.}}$），在 Office-Diff 中低 67%（界限为 $C^{\text{lb},\text{wgt}}_{p=0.5}$），这得益于我们能迅速排除表现不佳的策略。

我们的办公室环境并非单连通，因此只有乐观下界 $C^{\text{lb},\text{opt}}$ 能够保证长期收敛；在这些下界中，它最不紧，仅当在迷宫环境中部署于办公室训练的策略时，才会带来非平凡的性能提升（即限制策略选择）。然而，即使是对未探索空间结构的一些假设，也有助于更有力地约束选择。尽管这些假设可能会违反一般情况下关于渐近性能的理论保证，我们展示了使用加权下界代价 $C^{\text{lb},\text{wgt}}_{p=0.5}$时，有 $p=50\%$ 的概率存在一条更短的替代路径；而非单连通假设使得在两种办公室环境中，性能均优于无约束 UCB老虎机。

IX. 结论与未来工作

我们提出了一种数据高效策略选择方法，该方法利用支持基于子目标的学习规划的离线替代策略回放，根据导航过程中从环境收集的局部地图和图像，计算策略性能的下界，并使用类似多臂老虎机的方法快速识别表现最佳的策略。我们的方法使学习引导的机器人能够通过仅选择已知表现更好或有潜力表现更好的策略，降低其在各种局部地图环境中的平均导航代价，从而显著减少相较于基线 UCB 多臂老虎机的累积遗憾。未来，我们希望扩展这项工作，在新环境中通过在线再训练或策略适应来实现策略选择。

X. 致谢

我们谨向 Jana Koˇseck´a、George Konidaris 和 Kevin Doherty 致谢，感谢他们对本工作的宝贵反馈。本材料基于美国国家科学基金会第 2232733 号拨款资助的工作。

参考文献

[1] L. P. Kaelbling、M. L. Littman 和 A. R. Cassandra，《部分可观测随机领域中的规划与行动》，人工智能，1998年。[2] G. J. Stein、C. Bradley 和 N. Roy，《通过子目标学习，高效导航结构化未知环境》，载于机器人学习会议。PMLR，2018年。[3] C. Richter、J. Ware 和 N. Roy，《利用学习得到的碰撞概率实现未知环境中的高速自主导航》，载于*2014*年*IEEE*机器人与自动化国际会议（*ICRA*），2014年。[4] G. Wayne、C.-C. Hung、D. Amos、M. Mirza、A. Ahuja、A. Grabska-Barwinska、J. Rae、P. Mirowski、J. Z. Leibo、A. Santoro等，《目标导向智能体中的无监督预测记忆》，*arXiv*预印本 *arXiv:1803.10760*，2018年。[5] Y. Zhu、R. Mottaghi、E. Kolve、J. J. Lim、A. Gupta、L. Fei-Fei 和 A. Farhadi，《利用深度强化学习在室内场景中实现目标驱动的视觉导航》，载于*2017*年*IEEE*机器人与自动化国际会议（*ICRA*），2017年。[6] G. Kahn、A. Villaflor、B. Ding、P. Abbeel 和 S. Levine，《利用广义计算图实现机器人导航的自监督深度强化学习》，载于*2018*年*IEEE*机器人与自动化国际会议（*ICRA*），2018年。[7] P. Mirowski、R. Pascanu、F. Viola、H. Soyer、A. J. Ballard、A. Banino、M. Denil、R. Goroshin、L. Sifre等，《学习在复杂环境中导航》，*arXiv*预印本 *arXiv:1611.03673*，2016年。[8] J. Lee、A. Pacchiano、V. Muthukumar、W. Kong 和 E. Brunskill，《基于函数逼近的强化学习在线模型选择》，载于国际人工智能与统计会议。PMLR，2021年。[9] J. Reisinger、P. Stone 和 R. Miikkulainen，《贝叶斯强化学习的在线核选择》，载于第*25*届国际机器学习会议论文集，2008年。[10] T. Lattimore 和 C. Szepesv´ari，《多臂老虎机算法》。剑桥大学出版社，2020年。[11] T. L. Lai、H. Robbins等，《渐近高效的自适应分配规则》，应用数学进展，1985年。[12] J. C. Gittins，《多臂老虎机过程与动态分配指数》，英国皇家统计学会会刊：*B*辑（方法论），1979年。[13] W. R. Thompson，《根据两组样本的证据，一个未知概率超过另一个的可能性》，*Biometrika*，1933年。[14] V. Kuleshov 和 D. Precup，《多臂老虎机问题的算法》，*arXiv*预印本 *arXiv:1402.6028*，2014年。[15] S. Mannor 和 J. N. Tsitsiklis，《多臂老虎机问题中的探索样本复杂度》，机器学习研究杂志，2004年。[16] P. Mallozzi、E. Castellano、P. Pelliccione、G. Schneider 和 K. Tei，《用于对探索复杂环境的强化学习智能体实施不变式的运行时监控框架》，载于2019年IEEE/ACM第二届机器人软件工程国际研讨会（RoSE），2019年。[17] J. Kulh´anek、E. Derner、T. De Bruin 和 R. Babuˇska，《基于视觉的深度强化学习导航》，载于2019年欧洲移动机器人会议（ECMR），2019年。[18] P. Henderson、R. Islam、P. Bachman、J. Pineau、D. Precup 和 D. Meger，《真正重要的深度强化学习》，载于AAAI人工智能会议，2018年。[19] S. Huang、N. Papernot、I. Goodfellow、Y. Duan 和 P. Abbeel，《针对神经网络策略的对抗攻击》，arXiv预印本 arXiv:1702.02284，2017年。[20] Y. Xiang、W. Niu、J. Liu、T. Chen 和 Z. Han，《一种基于主成分分析的模型，用于预测路径搜索中Q学习的对抗样本》，载于2018年IEEE第三届网络空间数据科学国际会议（DSC），2018年。[21] M. L. Littman、A. R. Cassandra 和 L. P. Kaelbling，《部分可观测环境中的策略学习：扩展规模》。摩根·考夫曼出版社，1997年。[22] S. Candido 和 S. Hutchinson，《利用信息引导的POMDP规划实现最小不确定性机器人导航》，载于2011年IEEE国际机器人与自动化会议，2011年。[23] H. Kurniawati、Y. Du、D. Hsu 和 W. S. Lee，《面向长时间跨度机器人任务的不确定性下运动规划》，国际机器人研究杂志，2011年。[24] S. Gupta、J. Davidson、S. Levine、R. Sukthankar 和 J. Malik，《面向视觉导航的认知制图与规划》，载于IEEE计算机视觉与模式识别会议（CVPR）论文集，2017年7月。[25] J. Zhang、J. T. Springenberg、J. Boedecker 和 W. Burgard，《利用后继特征实现跨相似环境导航的深度强化学习》，载于2017年IEEE/RSJ智能机器人与系统国际会议（IROS），2017年。[26] L. Tai、G. Paolo 和 M. Liu，《虚实迁移深度强化学习：移动机器人无地图导航的连续控制》，载于2017年IEEE/RSJ智能机器人与系统国际会议（IROS），2017年。[27] A. Ghosh 和 S. R. Chowdhury，《基于一般函数逼近的强化学习模型选择》，arXiv预印本 arXiv:2207.02992，2022年。[28] A. Pacchiano、C. Dann、C. Gentile 和 P. Bartlett，《用于多臂老虎机和强化学习模型选择的遗憾界平衡与消除》，arXiv预印本 arXiv:2012.13045，2020年。[29] Q. M. Rahman、P. Corke 和 F. Dayoub，《机器人感知中机器学习的运行时监控：新兴趋势综述》，IEEE Access，2021年。[30] W. Zhou、J. S. Berrio、S. Worrall 和 E. Nebot，《自动驾驶中语义分割鲁棒性的自动评估》，IEEE智能交通系统汇刊，2019年。[31] S. Daftry、S. Zeng、J. A. Bagnell 和 M. Hebert，《内省式感知：学习预测视觉系统中的故障》，载于2016年IEEE/RSJ智能机器人与系统国际会议（IROS），2016年。[32] M. Merlin、N. Parikh、E. Rosen 和 G. Konidaris，《局部可观测马尔可夫决策过程》，载于ICRA 2020感知、行动与学习研讨会，2020年。[33] R. S. Sutton 和 A. G. Barto，《强化学习：导论》。麻省理工学院出版社，2018年。[34] P. E. Hart、N. J. Nilsson 和 B. Raphael，《启发式确定最小代价路径的理论基础》，IEEE系统科学与控制论汇刊，1968年。[35] C. Bradley、A. Pacheck、G. J. Stein、S. Castro、H. Kress-Gazit 和 N. Roy，《在未知环境中针对时间扩展任务进行学习与规划》，载于2021年IEEE机器人与自动化国际会议（ICRA），2021年。