arXiv:2301.01219v1 [cs.LG] 30 Dec 2022

可扩展的任务引导型部分可观测马尔可夫决策过程逆向强化学习

Franck Djeumou*、Christian Ellis**、Murat Cubuktepe*、Craig Lennon、Ufuk Topcu*

摘要

在逆向强化学习（逆向强化学习）中，学习智能体利用专家提供的示范，推断出编码了底层任务的奖励函数。然而，许多现有的逆向强化学习技术都作出一个往往不切实际的假设，即智能体可以获得有关环境的完整信息。我们通过为部分可观测马尔可夫决策过程开发一种逆向强化学习算法，摒弃了这一假设。我们解决了现有逆向强化学习技术的两个局限。首先，由于专家与学习者之间存在信息不对称，这些技术需要过量的数据。其次，这些逆向强化学习技术大多需要在部分可观测马尔可夫决策过程中求解计算上难以处理的前向问题——即给定奖励函数计算最优策略。我们开发的算法将以时序逻辑表达的任务规范融入逆向强化学习，从而在减少信息不对称的同时提高数据效率。除示范外，这些规范还可被视为学习者事先可获得的辅助信息。此外，该算法以因果熵而非熵作为衡量示范似然的指标，从而避免了一种常见的算法复杂度来源。尽管如此，由于所谓的前向问题，由此产生的问题仍然是非凸的。我们通过序贯线性规划方案，以可扩展的方式解决前向问题固有的非凸性，并保证收敛到局部最优策略。在一系列示例中，包括在高保真 Unity模拟器中进行的实验，我们展示了即使数据量有限且部分可观测马尔可夫决策过程包含数万个状态，我们的算法也能利用所提供的辅助信息，学习出满足任务要求且行为与专家相似的奖励函数和策略。

1. 引言

机器人可以通过奖励函数描述期望行为，从而完成某些人类指定的任务。然而，设计这样的奖励函数并非易事。逆向强化学习（IRL）是一种成熟的技术，它利用专家示范推断出编码底层任务的奖励函数。逆向强化学习

*德克萨斯大学奥斯汀分校

**马萨诸塞大学达特茅斯分校 美国陆军研究实验室电子邮箱：
fdjeumou@utexas.edu (Franck Djeumou), cellis3@umassd.edu (Christian Ellis), mcubuktepe@utexas.edu (Murat Cubuktepe), craig.t.lennon.civ@army.mil (Craig Lennon),utopcu@utexas.edu (Ufuk Topcu)

提交至爱思唯尔的预印本 *2023*年*1*月*4*日

技术已在各种定义域中得到广泛应用，例如特技直升机飞行、 [1], 推断人们未来的行动 [2], 人类-自主系统交互、 [3, 4], 机器人手术 [5, 6], 以及机器人操作任务 [7]。大多数现有研究 [1, 8, 9, 10, 3, 7]都聚焦于马尔可夫决策过程（ MDP），并假设学习者能够完全观测环境状态和专家示范。然而，在许多应用中，学习者无法获得完整的状态观测。例如，机器人永远无法了解其环境中的所有信息 [11, 12, 13] ，也可能无法观测与其协作的人类的内部状态 [14, 15]。此类信息限制违背了大多数现有逆向强化学习技术所依赖的内在假设。

我们研究部分可观测马尔可夫决策过程（POMDP）中的逆向强化学习，这是一种广泛用于不完全信息下决策的模型。部分可观测性给逆向强化学习带来了两个关键挑战。第一个挑战源于专家与学习者之间所谓的信息不对称。专家通常掌握关于环境的全部或部分信息，而学习者只能部分观察状态和专家示范。即使假设学习者已知底层奖励函数，由于这种信息不对称，在信息受限的情况下，其最优策略也未必能产生与信息完备的专家相同的行为。

第二个挑战源于部分可观测马尔可夫决策过程中策略综合的计算复杂度。事实上，许多标准的逆向强化学习技术都依赖一个子程序来求解所谓的前向问题，即针对给定奖励计算最优策略。与马尔可夫决策过程相比，求解部分可观测马尔可夫决策过程的前向问题，无论在理论上还是实践中都困难得多 [16]。部分可观测马尔可夫决策过程的最优策略可能需要无限记忆来记录观测 [17], ，而马尔可夫决策过程只需无记忆策略即可。

现有逆向强化学习技术的另一项局限在于，基于状态的奖励函数表达能力有限，而且往往不切实际，难以表示复杂任务 [18]。例如，要定义一个纯粹基于状态的奖励函数来描述“前往目标位置并返回家中时，不要驶离道路”或“按特定顺序监测多个位置”等要求，将极其困难。然而，这类要求可以通过时序逻辑简洁而精确地表示 [19, 20]。因此，近期研究已展示了将时序逻辑规范融入马尔可夫决策过程中的逆向强化学习的实用性 [21, 22]。

在这项工作中，我们通过研究以下问题，着手应对最先进的逆向强化学习技术中的这些挑战与局限。

任务引导的部分可观测马尔可夫决策过程逆向强化学习：给定一个部分可观测马尔可夫决策过程、一组专家示范，以及（如果有的话）以时序逻辑表示的任务规范，学习一个策略及其底层奖励函数，使诱导的随机过程的因果熵最大化、产生与专家相似的行为，并确保满足该规范。

我们重点介绍问题陈述中的两个部分。将因果熵 作为优化准则，而非传统熵，能够得到一种最少承诺的策略：该策略所产生的行为获得的累积奖励与专家示范相同，同时不对示范作出任何额外假设。任务规

范 以任务要求的形式给出，通过描述可行行为来指导学习过程，并使学习者能够学习出符合任务要求且表现良好的策略。这类规范可被视为学习者事先可获得的辅助信息；除了示范之外，它们还旨在部分缓解专家与学习者之间的信息不对称。

具体而言，我们将问题重新表述为最大因果熵（MCE）问题，以解决部分可观测马尔可夫决策过程上的逆向强化学习问题。随后，我们为该问题开发了一种新求解器，相较于现有方法提高了计算上的可处理性。该求解器能够施加以时序逻辑规范表示的先验任务知识，从而引导学习、提高数据效率，并在一定程度上缓解信息不对称问题。

现有大多数逆向强化学习工作都依赖熵 来衡量示范的似然；然而，将其应用于随机马尔可夫决策过程时，必须处理非凸优化问题 [8, 10]。另一方面，采用因果熵作为似然度量的逆向强化学习技术，则可基于凸优化进行建模 [9, 10, 23]。我们证明，最大因果熵逆向强化学习在算法上的类似优势也适用于部分可观测马尔可夫决策过程。

不过，最大因果熵逆向强化学习中，马尔可夫决策过程与部分可观测马尔可夫决策过程之间的一个主要区别，源于部分可观测马尔可夫决策过程中策略综合固有的非凸性，这使得任务引导的逆向强化学习问题被表述为一个非凸优化问题。众所周知，这种非凸性严重限制了部分可观测马尔可夫决策过程中的综合可扩展性 [16]。我们通过改进序列凸规划（SCP），开发了一种迭代算法，能够以可扩展的方式求解由此产生的非凸问题 [24, 25]。每次迭代中，该算法都会围绕上一次迭代得到的解对底层非凸问题进行线性化。该算法还引入了若干扩展，以减轻线性化所导致的误差。其中一项扩展是现有SCP方案中没有的验证步骤。我们证明，所提出的算法能够计算出任务引导问题的可靠局部最优解。

此外，我们通过实验证明，该算法可扩展至包含数万个状态的部分可观测马尔可夫决策过程，而现有大多数工作只能处理包含几十个状态的问题。在部分可观测马尔可夫决策过程中，基于观测历史的有限记忆策略优于无记忆策略 [26]。此外，为部分可观测马尔可夫决策过程计算有限记忆策略，等价于在更大的乘积部分可观测马尔可夫决策过程中计算无记忆策略 [27]。因此，我们利用算法的可扩展性，通过有限状态控制器计算出性能更优且能利用记忆的策略 [28, 29]。另一方面，前文提到的现有部分可观测马尔可夫决策过程逆向强化学习技术无法有效利用记忆，因为它们无法扩展到大型部分可观测马尔可夫决策过程。

我们通过若干示例展示了该方法的适用性，其中包括一个在高保真、连续的 3D Unity 仿真环境中运行的轮式地面机器人。我们表明，在没有任务规范的情况下，所开发的算法能够计算出性能优于将原始生成对抗模仿学习 GAIL [30] 直接适配到部分可观测马尔可夫决策过程中的策略。随后，我们展示了将任务规范纳入逆向强化学习流程后，学习到的奖励函数和策略能够准确描述专家的行为，同时优于未使用任务规范时得到的策略。我们观察到，在数据更有限时，使用和不使用任务规范所学习到的策略之间的性能差距会更加显著。

最重要的是，我们通过与多种最先进的部分可观测马尔可夫决策过程求解器进行广泛比较，实证展示了我们的方法在求解前向问题 时的可扩展性，并表明在规模更大的部分可观测马尔可夫决策过程中，该算法能够以显著更短的时间计算出性能更优的策略。

2. 预备知识

以下章节将回顾关于部分可观测马尔可夫决策过程、其相应策略，以及如何使用贝叶斯技术更新部分可观测马尔可夫决策过程对状态的信念等必备知识。

记号。 我们用 $\mathbb{R}_+$表示非负实数集，用 Distr($\mathcal{X}$)表示有限或可数无限集合 $\mathcal{X}$ 上的所有概率分布构成的集合，用 ($\mathcal{X}$)* 表示某个集合 $\mathcal{X}$上的所有（无限或空）序列 x_0， x_1， $\ldots$， x_∞（其中 $x_i \in \mathcal{X}$），并用$\mathbb{E}_{X,\ Y}[g(X,\ Y)]$表示关于联合分布的随机变量 X 和 Y的函数 g 的期望。

2.1. 部分可观测马尔可夫决策过程

部分可观测马尔可夫决策过程（POMDP）是一种用于建模智能体与部分可观测环境之间序贯交互的框架。在这种环境中，智能体无法感知其底层状态，必须根据给定的含噪观测推断该状态。

部分可观测马尔可夫决策过程。我们将部分可观测马尔可夫决策过程定义为一个元组 $\mathcal{M} = (\mathcal{S}, \mathcal{A}, \mathcal{P}, \mathcal{Z}, \mathcal{O}, \mathcal{R}, \mu_0, \gamma)$，其中 $\mathcal{S}$、$\mathcal{A}$ 和 $\mathcal{Z}$ 分别是有限的状态、动作和观测集合。函数μ_0： $\mathcal{S} \mapsto \mathbb{R}_+$ 给出智能体状态的初始分布，而 $\gamma \in [0, 1)$ 是适用于可能无限的规划时域的折扣因子。在每个决策时刻，智能体选择一个动作 $\alpha \in \mathcal{A}$，转移函数 $\mathcal{P}$： $\mathcal{S} \times \mathcal{A} \mapsto$ Distr($\mathcal{S}$) 定义了到达状态 $s' \in \mathcal{S}$ 的概率 $\mathcal{P}(s'|s, \alpha)$，其中给定当前状态 $s \in \mathcal{S}$和动作 α。状态转移后，智能体根据函数 $\mathcal{O}$： $\mathcal{S} \mapsto$ Distr($\mathcal{Z}$) 接收一个观测 $z' \in \mathcal{Z}$，该函数定义了在状态 s'下感知到 z' 的概率 $\mathcal{O}(z'|s')$。智能体还从函数 $\mathcal{R}$： $\mathcal{S} \times \mathcal{A} \mapsto \mathbb{R}$ 接收奖励函数 $\mathcal{R}(s, \alpha)$，该函数对任务规范进行编码。下文中，我们不失一般性地考虑无限时域问题。

策略。对于 POMDP，基于观测的策略 $\sigma : (\mathcal{Z} \times \mathcal{A})^* \times \mathcal{Z} \mapsto$ Distr($\mathcal{A}$) 将观测和动作序列映射为动作分布。记忆有限状态控制器（M-FSC）是一个元组 $\mathcal{C} = (\mathcal{Q}, q_I, \eta, \delta)$，其中 $Q = \{q_1, q_2, \ldots, q_M\}$ 是有限的记忆状态集合， q_I 是初始记忆状态， η: $\mathcal{Q} \times \mathcal{Z} \mapsto$Distr($\mathcal{A}$) 是决策函数， δ: $\mathcal{Q} \times \mathcal{Z} \times \mathcal{A} \mapsto$ Distr($\mathcal{Q}$) 是记忆转移函数。动作映射 $\eta(n, z)$ 接收一个 FSC 记忆状态 n 和一个观测 $z \in \mathcal{Z}$，并返回 POMDP 动作的分布。记忆更新 $\delta(n, z, \alpha)$ 返回记忆状态的分布，其取值取决于由 η 选出的动作 α。FSC 通过在 POMDP 轨迹上联合执行这两个函数，诱导出一个基于观测的策略。如果 FSC 仅有一个

记忆状态。 无记忆有限状态控制器，记作 σ: $\mathcal{Z} \to \mathrm{Distr}(\mathcal{A})$，是基于观测的策略，其中 $\sigma(\alpha|z) = \sigma_{z,\alpha}$ 表示仅根据观测 z 时采取动作 α 的概率。

备注 1 (R化简为 M无记忆 策略)。 在本文余下部分中，为了简化记法，我们通过计算较大且有理论依据的部分可观测马尔可夫决策过程上的无记忆策略 σ，来合成部分可观测马尔可夫决策过程的最优M-*FSC*（即所谓的前向问题）。这些较大的部分可观测马尔可夫决策过程由记忆更新 δ 与原始部分可观测马尔可夫决策过程的所谓乘积构造而成。事实上， *[27]* 的作者给出了乘积部分可观测马尔可夫决策过程，其规模仅随 δ 的定义域大小呈多项式增长。

信念更新。 给定部分可观测马尔可夫决策过程上的一段历史，即感知到的观测和已执行的动作序列 $\tau = \{(z_0, \alpha_0), (z_1, \alpha_1) \ldots\ldots, (z_T, \alpha_T)\}$，其中 $z_i \in \mathcal{Z}$、$\alpha_i \in \mathcal{A}$、$i \in \{0, \ldots, T\}$ 和 T 是轨迹的长度，信念状态表示在给定初始信念 $b_0 = \mu_0$ 的情况下，处于部分可观测马尔可夫决策过程各个状态的概率。每个时间步都可以使用以下贝叶斯规则更新该信念状态

$$b_{t+1}(s') = \frac{\mathcal{O}(z_t|s') \sum_{s \in \mathcal{S}} \mathcal{P}(s'|s, \alpha_t) b_t(s)}{\sum_{s'' \in \mathcal{S}} \mathcal{O}(z_t|s'') \sum_{s \in \mathcal{S}} \mathcal{P}(s''|s, \alpha_t) b_t(s)}. \tag{1}$$

2.2. 部分可观测马尔可夫决策过程中的因果熵。

对于部分可观测马尔可夫决策过程 $\mathcal{M}$，策略 σ 诱导出随机过程 $S^\sigma_{0:\infty} := (S^\sigma_0, \ldots, S^\sigma_\infty)$、$A^\sigma_{0:\infty} := (A^\sigma_0, \ldots, A^\sigma_\infty)$ 和 $Z^\sigma_{0:\infty} := (Z^\sigma_0, \ldots, Z^\sigma_\infty)$。在每个时间索引 t，随机变量 S^σ_t、A^σ_t 和 Z^σ_t 分别取值于 $s_t \in \mathcal{S}$、$\alpha_t \in \mathcal{A}$ 和 $z_t \in \mathcal{Z}$。给定 $S_{0:T}$ 时，$A_{0:T}$ 关于 $S_{0:T}$ 的因果条件概率 $P(A_{0:T}||S_{0:T})$ 由 [10, 31, 32] $P(A_{0:T}||S_{0:T}) := \prod_{t=0}^{T} P(A_t|S_{0:t}, A_{0:t-1})$ 给出，它定义了这些随机过程之间的相关性，其中每个变量 A_t 仅受较早预测的变量 $S_{0:t}$、$A_{0:t-1}$条件影响，而不受未来变量 $S_{t+1:T}$影响。令 $H(A|S) \triangleq \mathbb{E}_{A,S}[-\log P(A|S)]$ 表示给定随机变量 S时随机变量A 的条件熵。在有限时域情形下，给定策略 σ所诱导的因果熵 H_σ 定义为 $H_\sigma := \mathbb{E}_{A^\sigma_{0:T}, S^\sigma_{0:T}}[-\log \mathbb{P}(A^\sigma_{0:T}||S^\sigma_{0:T})] = \sum_{t=0}^{T} H(A^\sigma_t|S^\sigma_{0:t}, A^\sigma_{0:t-1})$。因此，无限时域情形下的因果熵，即折扣因果熵 [9, 33], ，定义为

$$H^\gamma_\sigma := \sum\nolimits_{t=0}^{\infty} \gamma^t H(A^\sigma_t|S^\sigma_{0:t}, A^\sigma_{0:t-1}) = \sum\nolimits_{t=0}^{\infty} \gamma^t \mathbb{E}_{A^\sigma_t, S^\sigma_t}[-\log \mathbb{P}(A^\sigma_t|S^\sigma_t)], \tag{2}$$

其中，第二个等式是由马尔可夫性质得出的。

备注 2. 部分可观测马尔可夫决策过程（或马尔可夫决策过程）的熵 涉及未来的策略决策 *[8],*，即 $S^\sigma_{t+1:T}$，在时间索引 t 处的决策，这与部分可观测马尔可夫决策过程（或马尔可夫决策过程）中的因果熵不同。因此， *[8]* 的作者证明，即使在马尔可夫决策过程中，计算使熵最大化的策略也是一个非凸问题。最大化策略熵的逆向强化学习技术依赖于近似方法，或假设马尔可夫决策过程的转移函数是确定性的。另一方面，在马尔可夫决策过程中，计算使因果熵最大化的策略可以表述为一个凸优化问题 *[10, 9]*。

2.3. 线性时序逻辑规范。

我们使用一般线性时序逻辑（LTL）来表达部分可观测马尔可夫决策过程 $\mathcal{M}$ 上的复杂任务规范。给定原子命题集合 AP，即对于给定状态 s 或观测 z 具有真值的布尔变量，LTL 公式按如下方式归纳构造：

$$\varphi := \text{true} \mid a \mid \neg\varphi \mid \varphi_1 \wedge \varphi_2 \mid \mathbf{X}\varphi \mid \varphi_1 \mathbf{U} \varphi_2,$$

其中，$a \in$ AP、φ、φ_1和 φ_2 是 LTL 公式，$\neg$和 $\wedge$分别表示逻辑否定和合取，而 **X** 和**U** 是下一步算子 和直到算子 这两种时序算子。此外，诸如总是算子 **(G)**和最终算子 **(F)**这样的时序算子可分别定义为$\mathbf{F}\varphi := \text{true}\mathbf{U}\varphi$ 和$\mathbf{G}\varphi := \neg\mathbf{F}\neg\varphi$。我们用 $\Pr^{\sigma}_{\mathcal{M}}(\varphi)$ 表示遵循策略时满足线性时序逻辑公式的概率 φ σ 在部分可观测马尔可夫决策过程 $\mathcal{M}$上。关于 LTL 语法和语义的详细说明超出了本文的讨论范围，详见 [20, 19]。

3. 问题表述

本节将表述部分可观测马尔可夫决策过程中的任务引导逆向强化学习（IRL）问题。给定一个奖励函数未知 的部分可观测马尔可夫决策过程 $\mathcal{M}$ $\mathcal{R}$，我们旨在学习奖励函数 $\mathcal{R}$ 以及相应的策略 σ，使其诱导出的行为与专家示范相似。

我们将部分可观测马尔可夫决策过程 $\mathcal{M}$上的专家轨迹定义为感知到的观测和执行的动作序列 $\tau = \{(z_0, \alpha_0), (z_1, \alpha_1) \ldots\ldots, (z_T, \alpha_T)\}$，其中 $z_i \in \mathcal{Z}$ 且$\alpha_i \in \mathcal{A}$ 对所有 $i \in \{0, \ldots, T\}$均成立，T表示轨迹长度。与[34], 类似，我们假设已知信念轨迹，或者可以根据 τ（通过贝叶斯信念更新（1））构造信念轨迹 $b^{\tau} = \{b_0 := \mu_0, \ldots\ldots, b_T\}$，其中 $b_i(s)$是在时间索引 i处处于状态 s 的估计概率。下文假设我们已获得一组由轨迹 $\tau_1, \ldots\ldots, \tau_N$生成的信念轨迹 $\mathcal{D} = \{b^{\tau_1}, \ldots, b^{\tau_N}\}$，其中 N表示底层轨迹的总数。

我们用关于参数可微的函数 $\mathcal{R}^{\theta}$ 对未知奖励函数 $\mathcal{R}$ 进行参数化：$\mathcal{S} \times \mathcal{A} \mapsto \mathbb{R}^d$，其中 $\theta \in \mathbb{R}^F$ 是唯一确定奖励函数的参数。这种编码涵盖了奖励函数的传统表示形式，例如 $\mathcal{R}^{\theta}(s, \alpha) = g_{\theta}(\phi(s, \alpha))$，其中 $\phi \colon \mathcal{S} \times \mathcal{A} \mapsto \mathbb{R}^d$是由基函数构成的已知向量，其分量称为特征函数，$d$ 是特征数量，而 g_{θ} 可以是任意函数逼近器，例如神经网络。例如，在传统的线性编码中，我们有 $g_{\theta}(z) = \theta^{\mathrm{T}} z$。

具体而言，我们要寻找一个定义 $\mathcal{R}^{\theta}$ 的参数 θ 和一个策略 σ，使其折扣回报期望 R^{θ}_{σ} 与专家示范 $\mathcal{D}$ 的经验折扣回报期望 $\bar{R}^{\theta}$ 相匹配。也就是说，我们有 $R^{\theta}_{\sigma} = \bar{R}^{\theta}$，其中

$$R^{\theta}_{\sigma} := \sum_{t=0}^{\infty} \gamma^t \mathbb{E}_{S^{\sigma}_t, A^{\sigma}_t}[\mathcal{R}^{\theta}(S^{\sigma}_t, A^{\sigma}_t) | \sigma] \text{ and } \bar{R}^{\theta} = \frac{1}{N} \sum_{b^{\tau} \in \mathcal{D}} \sum_{b_i \in b^{\tau}} \gamma^i \sum_{s \in \mathcal{S}} b_i(s) \mathcal{R}^{\theta}(s, \alpha_i).$$

在奖励的线性编码情况下，上述条件称为特征匹配期望，可以通过将 $\mathcal{R}^{\theta}$ 替换为特征函数 ϕ 来简化。

然而，这个问题是不适定的，可能存在无限多个奖励函数和策略满足上述匹配条件。为消除歧义，我们寻求一个策略 σ，使其同时最大化折扣因果熵 H_σ^γ。现在我们来定义所关注的问题。

问题1. 给定一个无奖励部分可观测马尔可夫决策过程 $\mathcal{M}$、一组示范 $\mathcal{D}$ 和一个特征 ϕ，计算策略 σ 和权重 θ，使得：*(a)* 满足匹配条件；*(b)* σ使(2) 给出的因果熵 H_σ^γ 最大化。

此外，如果有可用信息，我们还希望在奖励和策略的设计中纳入关于专家所示范任务的先验高层辅助信息。

问题2. 给定一个线性时序逻辑公式 φ，计算策略 σ 和权重θ，使得满足问题*1* 中的约束*(a)*和*(b)*，并且对于给定参数 $\lambda \geq 0$，$\Pr_{\mathcal{M}}^{\sigma}(\varphi) \geq \lambda$。

虽然假定给定用于指定 φ 满足阈值的参数 λ，但该方法可以轻松调整，以计算最优的 λ。

4. 部分可观测马尔可夫决策过程中逆向强化学习的非凸表述

在本节中，我们将问题1和问题2表述为寻找非凸函数的鞍点。随后，我们提出一种基于求解非凸优化问题的算法，以计算这些鞍点。我们强调（见备注 1），通过在更大的乘积部分可观测马尔可夫决策过程中计算无记忆策略 σ，来计算部分可观测马尔可夫决策过程的记忆有限状态控制器。事实上，在本文余下部分，我们将直接针对乘积部分可观测马尔可夫决策过程进行推理；它是一个部分可观测马尔可夫决策过程与一个有限状态控制器的乘积，并产生具有状态—记忆对 [27]的部分可观测马尔可夫决策过程。

替换访问计数。我们通过替换为由策略诱导的折扣状态访问计数 μ_σ^γ：$\mathcal{S} \mapsto \mathbb{R}_+$ 以及状态-动作访问计数 ν_σ^γ:$\mathcal{S} \times \mathcal{A} \mapsto \mathbb{R}_+$，消除了 H_σ^γ 中对（无限）时间的依赖以及匹配条件。对于策略 σ、状态 s 和动作 α，折扣状态访问计数和状态-动作访问计数定义为

$$\mu_\sigma^\gamma(s) := \mathbb{E}_{S_t}[\sum_{t=1}^{\infty} \gamma^t \mathbb{1}_{\{S_t=s\}} | \sigma] \text{ and } \nu_\sigma^\gamma(s,\alpha) := \mathbb{E}_{A_t,S_t}[\sum_{t=1}^{\infty} \gamma^t \mathbb{1}_{\{S_t=s,A_t=\alpha\}} | \sigma],$$

其中 $\mathbb{1}_{\{.\}}$ 是指示函数。根据这些定义，可以直接推导出 $\nu_\sigma^\gamma(s,\ \alpha) = \pi_{s,\ \alpha}\mu_\sigma^\gamma(s)$，其中 $\pi_{s,\ \alpha} = \mathbb{P}[A_t = a | S_t = s]$。同样可以直接验证，对于所有 $s \in \mathcal{S}$ 和 $\alpha \in \mathcal{A}$，$\mu_\sigma^\gamma(s) \geq 0$，$\nu_\sigma^\gamma(s,\ \alpha) \geq 0$，且$\mu_\sigma^\gamma(s) = \sum_{\alpha \in \mathcal{A}} \nu_\sigma^\gamma(s,\ \alpha)$。

我们首先给出折扣因果熵 H_σ^γ 的一个凹表达式，形式如下：

访问计数 μ_σ^γ 和 ν_σ^γ的函数:

$$\begin{aligned} H_\sigma^\gamma &:= \sum_{t=0}^{\infty} \gamma^t \mathbb{E}_{S_t^\sigma, A_t^\sigma}[-\log(\pi_{s_t,\alpha_t})] \\ &= \sum_{t=0}^{\infty} \sum_{(s,\alpha)\in\mathcal{S}\times\mathcal{A}} -(\log \pi_{s,\alpha})\pi_{s,\alpha}\gamma^t \mathbb{P}[S_t^\sigma = s] \\ &= \sum_{(s,\alpha)\in\mathcal{S}\times\mathcal{A}} -(\log \pi_{s,\alpha})\pi_{s,\alpha}\mu_\sigma^\gamma(s) \\ &= \sum_{(s,\alpha)\in\mathcal{S}\times\mathcal{A}} -\log \frac{\nu_\sigma^\gamma(s,\alpha)}{\mu_\sigma^\gamma(s)}\nu_\sigma^\gamma(s,\alpha), \end{aligned} \tag{3}$$

其中，第一个等号来自折扣因果熵 H_σ^γ的定义，第二个等号是通过展开期望得到的。第三个和第四个等号分别依据状态访问计数 μ_σ^γ和状态-动作访问计数 ν_σ^γ 的定义得出。我们在附录中证明，上述表达式确实关于访问计数是凹函数。接下来，我们得到关于 ν_σ^γ 的折扣回报期望 R_σ^θ 的一个线性表达式:

$$\begin{aligned} R_\sigma^\theta &= \sum_{t=0}^{\infty} \sum_{(s,\alpha)\in\mathcal{S}\times\mathcal{A}} \mathcal{R}^\theta(s,\alpha)\gamma^t \mathbb{P}[S_t^\sigma = s, A_t^\sigma = \alpha] \\ &= \sum_{(s,\alpha)\in\mathcal{S}\times\mathcal{A}} \mathcal{R}^\theta(s,\alpha)\nu_\sigma^\gamma(s,\alpha), \end{aligned} \tag{4}$$

其中，第二个等式根据访问计数 ν_σ^γ的定义得出。以下关于 $\mu_\sigma^\gamma(s)$ 和 $\sigma_{z,\ \alpha}$ 的非凸约束确保策略基于观测:

$$\nu_\sigma^\gamma(s,\alpha) = \mu_\sigma^\gamma(s) \sum_{z\in\mathcal{Z}} \mathcal{O}(z|s)\sigma_{z,\alpha}. \tag{5}$$

最后，折扣访问计数的变量必须满足所谓的贝尔曼流约束 [9]，以确保策略定义良好。对于每个状态$s \in \mathcal{S}$,

$$\mu_\sigma^\gamma(s) = \mu_0(s) + \gamma \sum_{s'\in\mathcal{S}} \sum_{\alpha\in\mathcal{A}} \mathcal{P}(s|s',\alpha)\nu_\sigma^\gamma(s',\alpha). \tag{6}$$

鞍点公式。由于 $\bar{R}^\theta$ 是根据有限数量的示范 $\mathcal{D}$得出的经验估计，计算满足回报匹配约束 $R_\sigma^\theta = \bar{R}^\theta$ 的策略 σ 可能不可行。此外，由于专家与学习者之间存在信息不对称，特征匹配约束也可能不可行，例如，专家拥有完整观测。

我们基于鞍点计算问题，将回报匹配约束纳入前向问题的目标函数，这与文献中的其他逆向强化学习算法类似。具体而言，问题1和问题2中所需的权重向量 θ 和策略 σ 是以下问题的解：$\min_\theta f(\theta) := \max_\sigma H_\sigma^\gamma + (R_\sigma^\theta - \bar{R}^\theta)$。函数$f$对应于奖励参数固定时的内部优化问题。也就是说， $f(\theta)$会计算一个策略 σ ，使因果

算法 1 计算拉格朗日

松弛逆向强化学习问题的权重向量 {v3}和策略 {v4}解。

输入： 来自 $\mathcal{D}$的特征期望 $\bar{R}^{\phi}$ 、初始权重 θ^0、步长 η： $\mathbb{N} \mapsto \mathbb{R}^+$，以及（若可用）先验侧信息 φ 和 $\lambda \in [0,1]$，满足 $\Pr^{\sigma}_{\mathcal{M}}(\varphi) \geq \lambda$。

$\sigma^0 \leftarrow$均匀策略 ▷ 初始化均匀策略

对于 $k = 0, 1, \ldots,$ **执行** ▷ 通过梯度下降计算 θ

  $\sigma^{k+1} \leftarrow$ SCPForward(θ^k, σ^k, φ, λ) ▷ 求解前向问题(7)–(9)

  可选地使用 φ 和 λ

  $\theta^{k+1} \leftarrow \theta^k - \eta(k)\nabla_\theta f(\theta^k; \sigma^{k+1})$ ▷ 梯度步

结束循环

返回 σ^k , θ^k

熵与当前奖励函数估计值之和最大。换句话说， $f(\theta)$返回前向问题的解，即在去除熵项后，求解部分可观测马尔可夫决策过程上的最优策略。

算法 1 使用梯度下降更新奖励权重。初始时，策略 σ^0 是一个随机均匀变量，权重 θ^0 是一个非零向量。在第 $k \geq 0$ 次迭代中，策略 $\sigma^{k+1} = \arg\max_\sigma H^\gamma_\sigma + (R^{\theta^k}_\sigma - \bar{R}^{\theta^k})$ 是当前奖励估计 $\mathcal{R}^{\theta^k}$ （由 θ^k 给出）下部分可观测马尔可夫决策过程上的最优策略。也就是说， σ^{k+1} 是前向问题的解。然后，为了更新权重 θ，算法 1 计算相对于 θ 的梯度 $\nabla_\theta f$，如下所示：

$$\nabla_\theta f(\theta;\sigma) = \sum_{s,\alpha \in \mathcal{S}\times\mathcal{A}} \nu^\gamma_\sigma(s,\alpha)\nabla_\theta \mathcal{R}^\theta(s,\alpha) - \frac{1}{N}\sum_{b^\tau \in \mathcal{D}}\sum_{b_i \in b^\tau}\gamma^i \sum_{s\in\mathcal{S}} b_i(s)\nabla_\theta \mathcal{R}^\theta(s,\alpha_i).$$

我们将在下一节介绍算法 SCPForward，以高效且可扩展的方式求解前向问题，即在给定 θ^k 的情况下计算 σ^{k+1} ，同时纳入高层任务规范来指导学习。

前向问题的非凸形式。 给定权重向量 θ^k，我们利用通过期望访问计数获得的替代形式，将与问题1相关的前向问题表述为以下非凸优化问题：

$$\underset{\mu^\gamma_\sigma, \nu^\gamma_\sigma, \sigma}{\text{maximize}} \sum_{(s,\alpha)\in\mathcal{S}\times\mathcal{A}} -\log\frac{\nu^\gamma_\sigma(s,\alpha)}{\mu^\gamma_\sigma(s)}\nu^\gamma_\sigma(s,\alpha) + \sum_{(s,\alpha)\in\mathcal{S}\times\mathcal{A}} \mathcal{R}^{\theta^k}(s,\alpha)\nu^\gamma_\sigma(s,\alpha) \quad (7)$$

满足以下约束 (5) –(6),

$$\forall (s,\alpha) \in \mathcal{S}\times\mathcal{A},\ \ \mu^\gamma_\sigma(s) \geq 0,\ \ \nu^\gamma_\sigma(s,\alpha) \geq 0, \quad (8)$$

$$\forall (s,\alpha) \in \mathcal{S}\times\mathcal{A},\ \ \mu^\gamma_\sigma(s) = \sum\nolimits_{\alpha\in\mathcal{A}} \nu^\gamma_\sigma(s,\alpha), \quad (9)$$

非凸性来源于式 (5)，我们从上述优化问题的代价函数中去掉常数项 $-\bar{R}^{\theta^k}$ 。

5. 序贯线性规划公式

我们开发了一种算法 SCPForward，通过调整序列凸规划（SCP）方案，高效求解非凸前向问题 (7)–(9)。事实上，SCPForward 包含一个验证步骤，用于计算可靠的策略和访问次数，而现有的 SCP方案中并没有这一步。此外，我们将在下一节说明，如何通过对 SCPForward 求解得到的优化问题进行少量修改，利用高层任务规范（问题 2）。

5.1. 非凸优化问题线性化 *blem*

SCPForward 会围绕先前的解对 (5) 中的非凸约束进行迭代线性化。然而，线性化可能导致线性子问题 [25] 不可行或无界。我们首先在经线性化的约束中添加 松弛变量，以确保可行性。如果该问题的解与先前的解相差很大，线性化问题可能无法准确近似非凸问题。因此，我们使用信赖域约束 [25]，确保线性化能够准确近似非凸问题。在每次迭代中，我们都会引入一个验证步骤，以确保计算得到的策略和访问计数不仅仅是近似值，而是确实满足非凸策略约束 (5)，改善历次迭代中的实际代价函数，并满足时序逻辑规范（如有）。

非凸约束线性化与松弛变量添加。 我们围绕先前计算得到的解（记为 $\hat{\sigma}$, $\mu^{\gamma}{}_{\hat{\sigma}}$ 和 $\nu^{\gamma}{}_{\hat{\sigma}}$），对非凸约束(5)进行线性化；该约束关于 $\mu^{\gamma}_{\sigma}(s)$ 和 $\sigma_{z,\,\alpha}$ 为二次形式。然而，线性化后的约束可能不可行。我们通过添加松弛变量 $k_{s,\,\alpha} \in \mathbb{R}$（针对 $(s,\ \alpha) \in \mathcal{S} \times \mathcal{A}$）来缓解这一问题，由此得到仿射约束：

$$\nu^{\gamma}_{\sigma}(s,\alpha) + k_{s,\alpha} = \mu^{\gamma}_{\hat{\sigma}}(s) \sum\nolimits_{z\in\mathcal{Z}} \mathcal{O}(z|s)\sigma_{z,\alpha} + \big(\mu^{\gamma}_{\sigma}(s) - \mu^{\gamma}_{\hat{\sigma}}(s)\big) \sum\nolimits_{z\in\mathcal{Z}} \mathcal{O}(z|s)\hat{\sigma}_{z,\alpha}. \tag{10}$$

信赖域约束。 如果解与先前的解偏差显著，线性化可能不准确。我们添加以下信赖域约束以缓解这一缺陷：

$$\forall (z,\alpha) \in \mathcal{Z} \times \mathcal{A},\quad \hat{\sigma}_{z,\alpha}/\rho \le \sigma_{z,\alpha} \le \hat{\sigma}_{z,\alpha}\rho, \tag{11}$$

其中，ρ 是信赖域的大小，用于限制线性化问题中允许的策略集合。我们在式（7）的代价函数中加入项 $-\beta \sum_{(s,\alpha)\in\mathcal{S}\times\mathcal{A}} k_{s,\alpha}$，以确保将违反线性化约束的程度降至最低，其中 β 是一个较大的正数。

线性化问题。最后，通过对 $x \mapsto x\log x$ 和 $y \mapsto x\log(x/y)$求导，我们得到将式（7）中的凸因果熵成本函数线性化所需的系数。因此，我们得到以下线性规划（LP）：

$$\begin{aligned}\underset{\mu_\sigma^\gamma, \nu_\sigma^\gamma, \sigma}{\text{maximize}} \sum\nolimits_{(s,\alpha)\in\mathcal{S}\times\mathcal{A}} -\Bigg(\beta k_{s,\alpha} - \Big(\frac{\nu_{\hat{\sigma}}^\gamma(s,\alpha)}{\mu_{\hat{\sigma}}^\gamma(s)}\Big)\mu_\sigma^\gamma(s) \\ + \Big(\log\frac{\nu_{\hat{\sigma}}^\gamma(s,\alpha)}{\mu_{\hat{\sigma}}^\gamma(s)} + 1\Big)\nu_\sigma^\gamma(s,\alpha)\Bigg) + \sum_{(s,\alpha)\in\mathcal{S}\times\mathcal{A}} \mathcal{R}^{\theta^k}(s,\alpha)\nu_\sigma^\gamma(s,\alpha) \quad (12)\end{aligned}$$

满足 (6)，(8) – (11)。

验证步骤。每次迭代后，线性化可能不准确，即所得策略 $\tilde{\sigma}$ 以及可能不准确的访问计数 $\tilde{\nu}_{\tilde{\sigma}}^\gamma$，$\tilde{\mu}_{\tilde{\sigma}}^\gamma$ 可能不满足非凸策略约束 (5)。由于可能不可行，当前得到的（线性化）最优成本可能与 $\tilde{\sigma}$的可行访问计数所产生的实际成本有显著差异。此外，现有的 SCP方案会围绕 $\tilde{\nu}_{\tilde{\sigma}}^\gamma$ 和 $\tilde{\mu}_{\tilde{\sigma}}^\gamma$ 先前得到的不准确解对非凸问题进行线性化，从而进一步放大误差。本文提出的验证步骤解决了这些问题。给定计算得到的策略 $\tilde{\sigma}$，SCPForward 通过求解相应的贝尔曼流约束，计算访问计数 $\mu_{\tilde{\sigma}}^\gamma$ 的唯一且可靠的解：

$$\mu_{\tilde{\sigma}}^\gamma(s) = \mu_0(s) + \gamma \sum_{s'\in\mathcal{S}} \sum_{\alpha\in\mathcal{A}} \mathcal{P}(s|s',\alpha)\mu_{\tilde{\sigma}}^\gamma(s') \sum_{z\in\mathcal{Z}} \mathcal{O}(z|s)\tilde{\sigma}_{z,\alpha}, \quad (13)$$

对所有 $s \in \mathcal{S}$，且其中$\mu_{\tilde{\sigma}}^\gamma \geq 0$ 是该线性规划中的唯一变量。然后，SCPForward 计算 $\nu_{\tilde{\sigma}}^\gamma(s,\ \alpha) = \mu_{\tilde{\sigma}}^\gamma(s') \sum_{z\in\mathcal{Z}} \mathcal{O}(z|s)\tilde{\sigma}_{z,\ \alpha}$，并将当前迭代的实际成本定义为

$$\mathrm{C}(\tilde{\sigma}, \theta^k) = \sum_{(s,\alpha)\in\mathcal{S}\times\mathcal{A}} -\log\frac{\nu_{\tilde{\sigma}}^\gamma(s,\alpha)}{\mu_{\tilde{\sigma}}^\gamma}\nu_{\tilde{\sigma}}^\gamma(s,\alpha) + \sum_{(s,\alpha)\in\mathcal{S}\times\mathcal{A}} \mathcal{R}^{\theta^k}(s,\alpha)\nu_{\tilde{\sigma}}^\gamma(s,\alpha), \quad (14)$$

其中我们假设 $0\log 0 = 0$。最后，如果实际成本 $\mathrm{C}(\tilde{\sigma},\ \theta^k)$ 未能优于先前的成本 $\mathrm{C}(\hat{\sigma},\ \theta^k)$，验证步骤将拒绝所得策略 $\tilde{\sigma}$，收缩信赖域，并让 SCPForward 使用先前的解 $\hat{\sigma}$, $\mu^\gamma{}_{\hat{\sigma}}$ 和 $\nu^\gamma{}_{\hat{\sigma}}$ 进行迭代。否则，线性化足够准确，信赖域将扩大，SCPForward 将使用 $\tilde{\sigma}$、$\mu_{\tilde{\sigma}}^\gamma$ 和 $\nu_{\tilde{\sigma}}^\gamma$ 进行迭代。通过纳入这一验证步骤，我们确保 *SCPForward* 总是算子在满足非凸约束的解附近对非凸优化问题进行线性化 (5)。

5.2. 纳入高层任务规范

给定关于智能体任务的高层侧信息，即线性时序逻辑公式 φ，我们首先计算部分可观测马尔可夫决策过程与表示 φ 的 ω-自动机的乘积，以找出状态集合 $\mathcal{T} \subseteq \mathcal{S}$；这些状态称为目标状态或可达状态，它们以概率 1 满足 φ。这一过程通过

使用标准的基于图的算法，作为预处理步骤的一部分。关于如何将线性时序逻辑规范归约为由 $\mathcal{T}$ 给出的可达性规范的详细介绍，请参阅 [19] 。

因此，满足 φ 的概率是到达目标状态 $s \in \mathcal{T}$的概率之和，这些概率由未折扣状态访问计数 μ_σ^{sp}给出。也就是说，$\Pr_{\mathcal{M}}^{\sigma}(\varphi) = \sum_{s\in\mathcal{T}} \mu_\sigma^{\mathrm{sp}}(s)$。除非 $\gamma = 1$，$\mu_\sigma^{\mathrm{sp}} \neq \mu_\sigma^{\gamma}$。因此，我们引入新的变量 μ_σ^{sp}，ν_σ^{sp}，以及线性化问题（12）中的适当约束。

将未折扣访问变量纳入线性化问题。我们将类似于（8）、（9）和（10）的新约束附加到线性化问题（12）中，其中变量 μ_σ^{γ}，ν_σ^{γ}，$k_{s,\,\alpha}$，$\mu^{\gamma}{}_{\hat{\sigma}}$、$\nu^{\gamma}{}_{\hat{\sigma}}$ 分别替换为 μ_σ^{sp}，ν_σ^{sp}、$k_{s,\,\alpha}^{\mathrm{sp}}$，$\mu^{\mathrm{sp}}{}_{\hat{\sigma}}$、$\nu^{\mathrm{sp}}{}_{\hat{\sigma}}$。此外，我们加入约束

$$\mu_\sigma^{\mathrm{sp}}(s) = \mu_0(s) + \sum_{s'\in\mathcal{S}\setminus\mathcal{T}} \sum_{\alpha\in\mathcal{A}} \mathcal{P}(s|s',\alpha)\nu_\sigma^{\mathrm{sp}}(s',\alpha), \tag{15}$$

这是一种对 贝尔曼流约束 的修改，使得对于所有 $s \in \mathcal{T}$，$\mu_\sigma^{\mathrm{sp}}(s)$ 仅统计来自非目标状态的状态转移。最后，我们将项 $-\beta \sum_{(s,\,\alpha)\in\mathcal{S}\times\mathcal{A}} k_{s,\,\alpha}^{\mathrm{sp}}$ 加入代价函数，以惩罚为保证线性化问题可行而引入的松弛变量。

放宽规范约束。为纳入满足规范的概率，我们添加以下约束到线性化问题中:

$$(\mathrm{spec}) := \sum_{s\in\mathcal{T}} \mu_\sigma^{\mathrm{sp}}(s) + \Gamma^{\mathrm{sp}} \geq \lambda, \tag{16}$$

其中，我们引入 $\Gamma^{\mathrm{sp}} \geq 0$ 作为松弛变量，以确保线性化问题总是可行的。此外，我们在代价函数中加入 $-\beta^{\mathrm{sp}}\Gamma^{\mathrm{sp}}$，以惩罚违反 φ 的情况，其中 β^{sp} 是一个正的超参数。

更新验证步骤。我们修改先前引入的实际成本 $\mathrm{C}(\tilde{\sigma},\ \theta^k)$，以惩罚所得策略不满足规范 φ 的情况。该代价还考虑了新策略约束因 σ、μ_σ^{sp} 和 ν_σ^{sp} 而产生的线性化误差。在每次迭代中，SCPForward 通过求解一个带有由修改后的贝尔曼流约束 (15)给出的约束的可行性线性规划，来计算当前策略 $\tilde{\sigma}$ 的准确 $\mu_{\tilde{\sigma}}^{\mathrm{sp}}$。随后，它将 $\mathrm{C}_{\tilde{\sigma}}^{\mathrm{sp}} = \min\{0, (\sum_{s\in\mathcal{T}} \mu_{\tilde{\sigma}}^{\mathrm{sp}}(s) - \lambda)\beta^{\mathrm{sp}}\}$ 加入实际成本，以纳入规范约束。

收敛到局部最优解。所提出的带信赖域的序列凸规划方案的收敛保证，可直接由序列凸规划（SCP）方案的一般收敛性得出，相关结论已在 [25] 的定理 3。14 和定理 4。7 中得到证明。具体而言，随着 SCP 算法生成一组收敛子序列，且这些子序列均满足一阶条件 [25]，弱收敛便得以保证。由于算法可能在若干极限点之间振荡，这并不符合严格意义上的收敛。不过，令人惊讶的是，大多数关于收敛的

算法 2 SCPForward：用于求解前向问题 (7)–(9) 的基于线性规划的算法，即计算使因果熵最大化、考虑匹配约束并满足规范（若有）的策略 σ^{k+1}。输入：~~当前权重估计 θ^k、当前最优策略 $\hat{}\sigma = \sigma^k$、辅助信息 φ 和 λ、信赖域 $\rho > 1$、~~惩罚系数 β、 $\beta^{\text{sp}} \geq 0$、用于扩大或缩小信赖域的常数 ρ_0，以及用于缩小信赖域的阈值 ρ_{lim}。1：通过线性约束 (13) 求得 $\mu^{\gamma}\hat{\sigma}$，以及 $\nu^{\gamma}\hat{\sigma} = \mu_{\sigma}^{\gamma}\hat{}(s')$ $\sum_{z\in\mathcal{Z}}\mathcal{O}(z|s)\hat{}\sigma_{z,\alpha}$，给定 $\hat{}\sigma$ ▷实际访问计数。2：通过线性约束 (15) 求得 $\mu^{\text{sp}}\hat{}$ $_{\sigma}$，以及 $\nu^{\text{sp}}\hat{\sigma} = \mu_{\sigma}^{\text{sp}}\hat{}(s')\sum_{z\in\mathcal{Z}}\mathcal{O}(z|s)\hat{}\sigma_{z,\alpha}$，给定 $\hat{}\sigma$▷ 若 φ 可用。3：计算实际成本 $\text{C}(\hat{}\sigma, \theta^k) \leftarrow$ (14) $+ \text{C}^{\text{sp}}\hat{\sigma}$，给定 $\hat{}\sigma$ ▷；加入规范违规项。4：当 $\rho > \rho_{\text{lim}}$ 时执行▷ 信赖域阈值。5：通过 $\hat{}\sigma, \mu^{\gamma}\hat{\sigma}, \nu^{\gamma}\hat{\sigma}, \mu^{\text{sp}}\hat{\sigma}, \nu^{\text{sp}}\hat{\sigma}$ ▷，求增广线性规划 (12) 的最优 $\tilde{\sigma}$。我们使用由 $\mu_{\sigma}^{\text{sp}}, \nu_{\sigma}^{\text{sp}}$ 导出的约束 (8)、(9)、(10)、(15) 和 (16) 扩充线性规划，并将 $-\beta\sum_{(s,\alpha)\in\mathcal{S}\times\mathcal{A}} k_{s}^{\text{sp}},_{\alpha} - \beta^{\text{sp}}\Gamma^{\text{sp}}$ 加入成本 (12)。6：按照第 1–3 行，通过 $\tilde{\sigma}$ 计算实际 $\mu_{\tilde{\sigma}}^{\gamma}, \nu_{\tilde{\sigma}}^{\gamma}$,$\mu_{\tilde{\sigma}}^{\text{sp}}, \nu_{\tilde{\sigma}}^{\text{sp}}$ 以及 $\text{C}(\tilde{\sigma}, \theta^k)$。7：$\{\hat{}\sigma \leftarrow \tilde{\sigma}$；若 $\text{C}(\tilde{\sigma}, \theta^k) \geq \text{C}(\hat{}\sigma, \theta^k)$，则 $\{\rho \leftarrow \rho/\rho_0\}$ ▷验证步骤。8：结束循环。9：返回 $\sigma^{k+1} := \hat{}\sigma$

非线性优化方案的论断都属于这一类。此外，在代价函数满足适当正则性假设的条件下， [25] 的作者证明，带信赖域的 SCP方案可以超线性收敛速率收敛到局部最优解。

6. 数值实验

我们在 [35],中的多个 POMDP 实例，以及一个在高保真、连续的三维 Unity 仿真环境中运行的轮式地面机器人上，对所提出的逆向强化学习算法进行了评估。我们首先将该算法与一种针对部分可观测马尔可夫决策过程改编的直接版本的生成对抗模仿学习进行比较 [30]。随后，我们展示了利用辅助信息时，所提出方法的数据效率。最后，我们通过与 SolvePOMDP、SARSOP 和 PRISM-POMDP 等先进求解器进行比较，展示了 SCPForward 例程在求解前向 问题方面的可扩展性。我们在 https://github.com/wuwushrek/MCE IRL POMDPS 提供了本文的代码，以便复现结果。

6.1. 手工构造的 *POMDP* 实例

例

我们首先在从文献 [35] 中提取的多个 POMDP 实例上评估所提出的逆向强化学习算法。

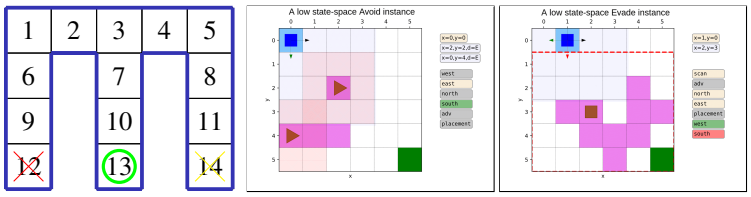


图1： [35] 中基准集合提供的一些示例。从左到右，依次为 迷宫、*Avoid* 和躲避环境。

基准集合。这些 POMDP 实例如下。躲避是一种回合制游戏，智能体必须在被速度更快的玩家拦截之前到达目的地。在 *Avoid* 中，智能体必须避免被另外两名按照某些预设但未知路线移动的玩家发现。在 拦截中，智能体必须拦截另一名试图离开网格世界的玩家。在岩石中，智能体必须在多块岩石中至少采样到一块优质岩石，且不能失败。在障碍物中，智能体必须在不与任何静态障碍物碰撞的情况下，在网格世界中找到出口。在这些实例中，智能体只能观察其当前位置周围固定半径范围内的区域，见图1。最后，在迷宫中，智能体必须尽可能快地走出迷宫，同时只能观察周围的墙壁，并且不能陷入任何陷阱状态。

学习到的策略与专家的变体。我们讨论四种类型的策略。策略类型取决于是否使用来自时序规范的辅助信息 φ，以及使用的记忆大小是 $\mathrm{M} = 1$ 还是 $\mathrm{M} = 10$。我们还考虑两种类型的专家。第一种专家拥有关于环境的完整信息，并在底层马尔可夫决策过程中计算最优策略。第二种专家只能部分观测，并在记忆大小为 $\mathrm{M} = 15$ 的部分可观测马尔可夫决策过程中计算局部最优策略。请注意，智能体始终只能获得部分信息。因此，第一种专家对应于学习智能体与专家之间存在信息不对称的情况。此外，我们将 *GAIL* 的一种变体作为基线：在不使用辅助信息的马尔可夫决策过程中学习策略，并通过离线计算状态的信念，将其扩展到部分可观测马尔可夫决策过程。具体而言，我们通过求解对应于马尔可夫决策过程前向问题的凸优化问题，在马尔可夫决策过程中找到最优策略。所得策略是基于状态的策略，需要经过转换才能用于部分可观测马尔可夫决策过程。转换时利用专家示范构建信念状态。也就是说，专家的轨迹 τ 用于贝叶斯信念更新 (1)，以估计处于部分可观测马尔可夫决策过程各个状态的概率。因此，将计算得到的信念与基于状态的策略相结合，即可得到部分可观测马尔可夫决策过程的基于观测的策略。这样做可能会让 *GAIL* 变体具有显著优势，因为基于状态的策略是马尔可夫决策过程中的最优策略。然而，尽管该策略在实践中性能很高，但部分可观测马尔可夫决策过程中的策略通常并非最优，即使马尔可夫决策过程中的策略是最优的也是如此。

我们讨论辅助信息和记忆对相应策略的影响。虽然我们只详细介绍迷宫 示例（其中智能体必须尽可能快地走出迷宫），但我们发现其他示例也呈现出类似的规律。其他示例的详细结果见附录。

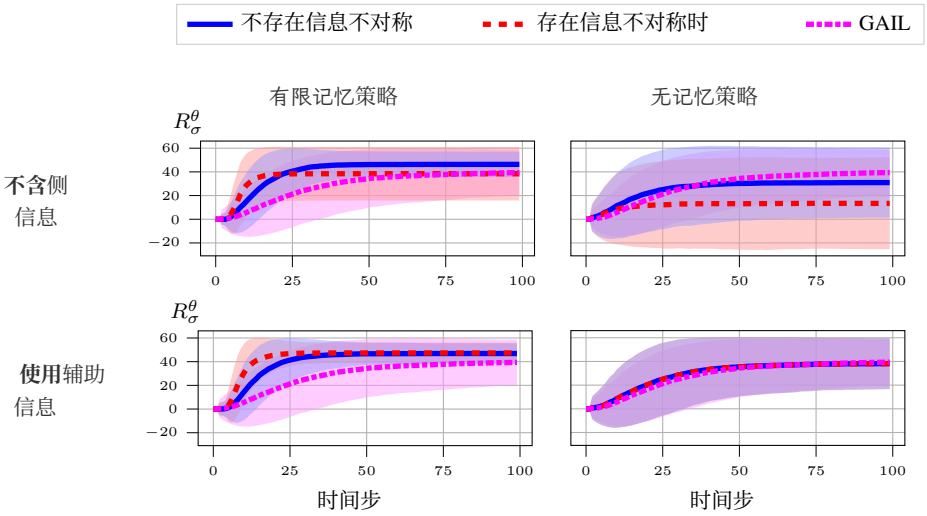


图2：迷宫示例的代表性结果；每个子图表示在真实奖励函数 (R_σ^θ) 下，1000 次运行的平均累积奖励随时间的变化。比较两行：顶行中未利用辅助信息的策略在信息不对称的情况下性能下降。另一方面，在底行中，将辅助信息纳入学习的策略在信息不对称的情况下性能不会下降。比较两列：左列中有限记忆策略的性能显著优于无记忆策略。除不使用辅助信息的无记忆策略外，我们的算法优于生成对抗模仿学习。在马尔可夫决策过程上，专家的平均奖励为 48.22，而专家在部分可观测马尔可夫决策过程上行动时，我们得到的数值为 47.83。

6.1.1. 迷宫示例

部分可观测马尔可夫决策过程 $\mathcal{M}$ 由与图1中单元格标签对应的 $\mathcal{S} = \{s_1, \ldots, s_{14}\}$ 指定。迷宫中的智能体只能观测相邻单元格中是否有墙（蓝色）。也就是说，观测集合为 $\mathcal{O} = \{o_1, \ldots, o_6, o_7\}$。例如，$o_1$ 对应于观测到西侧和北侧的墙（s_1），o_2 对应于北侧和南侧的墙（s_2, s_4），而 o_5 对应于东侧和西侧的墙（s_6，s_7，s_8，s_9，s_{10}，s_{11}）。观测o_6 和 o_7 分别表示目标状态（s_{13}）和不良状态（s_{12}, s_{14}）。转移模型是随机的，滑移概率为 $p = 0.1$。此外，状态 s_{13} 和 s_{14} 会导致仿真结束（陷阱状态）。

在逆向强化学习实验中，我们考虑三个特征函数。对于所有 s、α，我们通过 $\phi^{\text{time}}(s, \alpha) = -1$ 来惩罚采取更多步骤。当到达 s_{13} 时，如果 $s = s_{13}$，则通过 $\phi^{\text{target}}(s, \alpha) = 1$ 提供正奖励，否则通过 $\phi^{\text{target}}(s, \alpha) = 0$ 提供正奖励。对于不良状态，如果 $s = s_{12}$ 或 $s = s_{14}$，则通过 $\phi^{\text{bad}}(s, \alpha) = -1$ 惩罚不良状态 s_{12} 和 s_{14}，否则通过 $\phi^{\text{bad}}(s, \alpha) = 0$ 惩罚。最后，我们将 LTL 公式 $\varphi = \text{G}\ \neg\,\text{bad}$ 作为任务规范，其中 bad 是一个原子命题；当当前状态 $s = s_{12}$ 或 $s = s_{14}$ 时，该命题为真。我们约束学习到的策略，使其满足 $\Pr_{\mathcal{M}}^{\sigma}(\text{G}\ \neg\text{bad}) \geq 0.9$。

侧信息缓解信息不对称。 图2显示，如果学习智能体与专家之间存在信息不对称，不利用辅助信息的策略会遭受显著的性能下降。这些策略

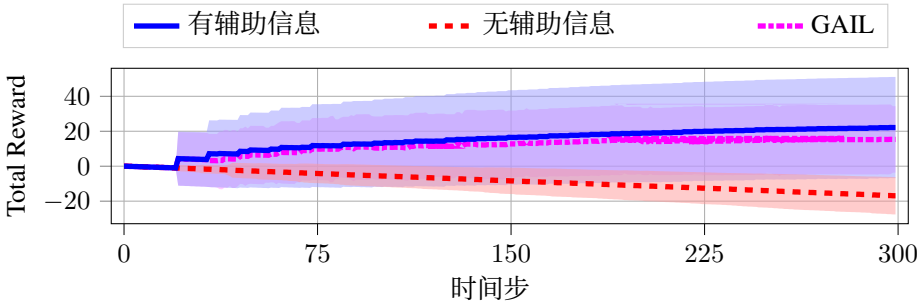


图3：Avoid 示例的代表性结果，展示策略在真实奖励函数 (R_o^g) 下的奖励随时间步的变化。

在学习过程中未纳入辅助信息，在存在信息不对称时性能会降低57%，如图2顶行所示。另一方面，如图2底行所示，使用辅助信息的策略几乎不受信息不对称的影响。

记忆带来性能更佳的策略. 图2中的结果表明，在所有示例中，无论是求解前向问题，还是根据专家示范学习策略，将记忆纳入策略都能提升性能，即获得的奖励。在信息不对称的情况下，有限记忆策略的性能下降了18%，而无记忆策略则下降了57%，这表明纳入记忆在一定程度上缓解了信息不对称的影响。

我们可以看到，在表1中，对于迷宫和岩石基准测试，将记忆纳入策略后，SCPForward 能够计算出近乎最优的策略；其证据是获得的奖励几乎与求解器SARSOP相同。

辅助信息提升数据效率。 图4显示，即使在数据量较少的情况下，使用任务规范进行学习的性能也显著优于不使用任务规范。

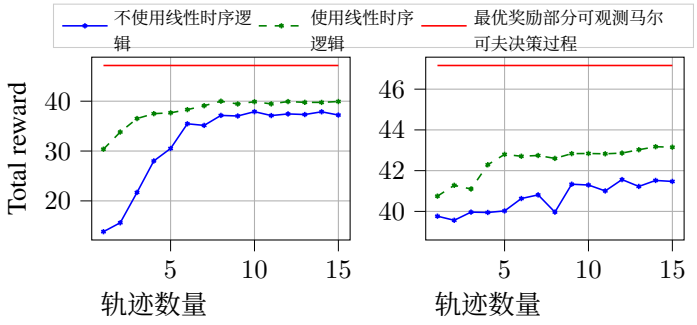


图4：我们通过学习到的策略所获得的总奖励随专家示范数量的变化，展示了所提方法的数据效率（无信息不对称）。左图展示了学习无记忆策略的性能，右图展示了 5-FSC 的性能。

问题	$\|\mathcal{S}\|$	$\|\mathcal{S}\times\mathcal{O}\|$	$\|\mathcal{O}\|$	SCPForward R^{θ}_{σ}	SCPForward 时间（秒）	SARSOP R^{θ}_{σ}	SARSOP 时间（秒）	SolvePOMDP R^{θ}_{σ}	SolvePOMDP 时间（秒）
Maze	17	162	11	39.24	**0.1**	**47.83**	0.24	47.83	0.33
迷宫（3-FSC）	49	777	31	44.98	**0.6**	NA	NA	NA	NA
迷宫（10-FSC）	161	2891	101	46.32	2.04	NA	NA	NA	NA
障碍物[10]	102	1126	5	19.71	8.79	**19.8**	**0.02**	5.05	3600
障碍物[10](5-FSC)	679	7545	31	19.77	38	NA	NA	NA	NA
障碍物[25]	627	7306	5	19.59	14.22	**19.8**	**0.1**	5.05	3600
Rock	550	4643	67	19.68	12.2	**19.83**	**0.05**	–	–
岩石（3-FSC）	1648	23203	199	19.8	15.25	NA	NA	–	–
岩石（5-FSC）	2746	41759	331	19.82	97.84	NA	NA	–	–
拦截$[5,2,0]$	1321	5021	1025	**19.83**	**10.28**	**19.83**	13.71	–	–
拦截$[5,2,0.1]$	1321	7041	1025	**19.81**	**13.18**	**19.81**	81.19	–	–
躲避$[5,2,0]$	2081	13561	1089	**97.3**	**26.25**	**97.3**	3600	–	–
躲避$[5,2,0.1]$	2081	16761	1089	**96.79**	**26.25**	95.28	3600	–	–
躲避$[10,2,0]$	36361	341121	18383	**94.97**	**3600**	–	–	–	–
Avoid$[4,2,0]$	2241	5697	1956	**9.86**	34.74	**9.86**	**9.19**	–	–
Avoid$[4,2,0.1]$	2241	8833	1956	**9.86**	**14.63**	**9.86**	210.47	–	–
Avoid$[7,2,0]$	19797	62133	3164	**9.72**	**3503**	–	–	–	–

表1：求解前向问题的基准集结果。对于较大的基准测试（例如 Evade 和 Avoid），SCPForward 能够计算局部最优策略，而其他求解器无法在给定的时间限制内提供解。在环境 Obstacle$[n]$、Intercept$[n,r$,slip$]$, Evade$[n,r$,slip$]$,和 Avoid$[n,r$,slip$]$, 中，参数 n、 r 和 slip 分别表示网格世界的大小、智能体的视野半径以及湿滑概率。我们将超时设为 3600 秒。空单元格（记作 –）表示求解器在超时前未能计算出任何策略，而 NA 表示不适用，因为该方法基于信念更新。

辅助信息提升性能。 此外，在 Avoid 这样更复杂的环境中，图3显示任务规范至关重要，甚至是学习任务的必要条件。具体来说，Avoid$[n,r,slip]$ 是一款回合制游戏，智能体必须在避免被另外两名按照特定预设但未知路线移动的玩家发现的同时，到达出口点。只有在执行扫描 动作时，如果玩家位于距离智能体当前位置固定半径的范围内，智能体才能观察到他们。此外，由于玩家的速度不确定，智能体无法推断他们在路线上的位置。参数 n、 r和 $slip$ 分别表示网格尺寸、视野半径和滑动概率。

我们使用四个特征函数对未知奖励进行参数化。第一个特征在智能体到达出口点时给予正奖励。第二个特征在智能体与玩家发生碰撞时施加惩罚。第三个特征在智能体被玩家发现时施加惩罚。第四个特征对每个执行的动作施加惩罚成本。我们将辅助信息编码为时序逻辑任务规格避免被发现，直到以大于0.98 的概率到达出口点。

图3显示，该算法在没有辅助信息时无法学习，而辅助信息则会引导其学习到最优策略。具体而言，没有辅助信息时，学习到的策略似乎只关注避免被发现和发生碰撞，因为相应的学习到的特征值接近于零。

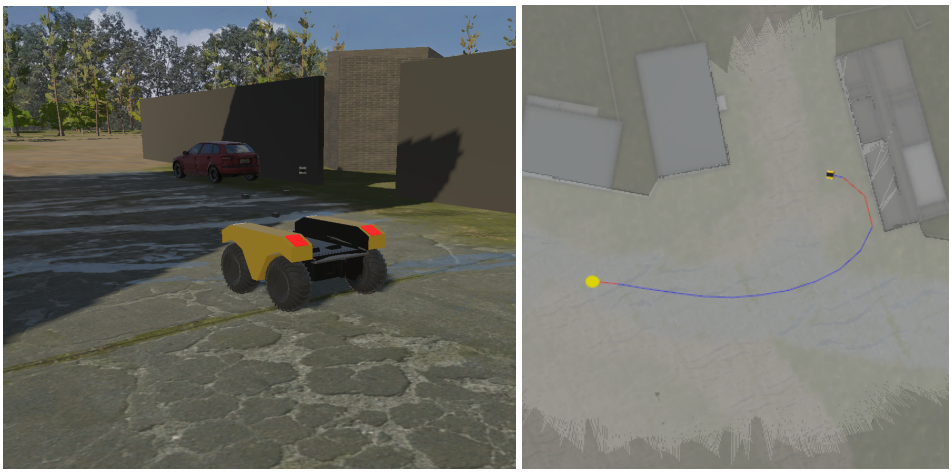


图 5：左：在 Unity 仿真中运行的 Clearpath Warthog 仿真模型。右：专家提供的演示。

*6.1.2. SCPForward*带来更好的可扩展性

我们强调有关 SCPForward 可扩展性的三点观察。首先，表1的结果表明，在较大的部分可观测马尔可夫决策过程中，只有 SARSOP 的表现能与 SCPForward 相媲美。SolvePOMDP 除了在最小的基准测试中外，都会超时；PrismPOMDP 则在所有基准测试中都会耗尽内存。这些方法大多基于对状态的信念进行更新，而对于较大的状态空间，这一过程的计算成本可能极高。

其次，在状态空间较小的基准测试中，例如 迷宫 和 岩石，SARSOP 能够在更短时间内计算出性能更好的策略。这是因为基于信念的方法在小规模问题上效率更高。另一方面，SARSOP 无法扩展到状态和观测数量更多的大型部分可观测马尔可夫决策过程。例如，将 拦截 基准测试中的状态转移数量从 5021 增加到 7041，SARSOP 的计算时间便增加了 516%。相比之下，SCPForward 的计算时间仅增加了 28%。

第三，在包含数万个状态和观测的最大规模基准测试中，SARSOP 在超时前未能计算出任何策略，而 SCPForward 找到了解决方案。最后还要指出，与 SARSOP 不同，SCPForward 也能计算出最大化因果熵且满足线性时序逻辑规范的策略。

6.2. 地面机器人仿真

我们展示了所提出算法在一个连续三维 Unity 环境中的应用。该环境包含一辆 ClearPath Warthog，它在半结构化村庄中运行。图 5 展示了机器人在该环境中运行的屏幕截图及其对应轨迹。该环境中有多种障碍物，包括建筑物、树木和车辆，还有三种用于描述特征的地形类型，ϕ、草地、砾石和道路。仿真环境的状态空间包含 3350 个状态、33254 个状态转移和 944 种观测。此仿真用于

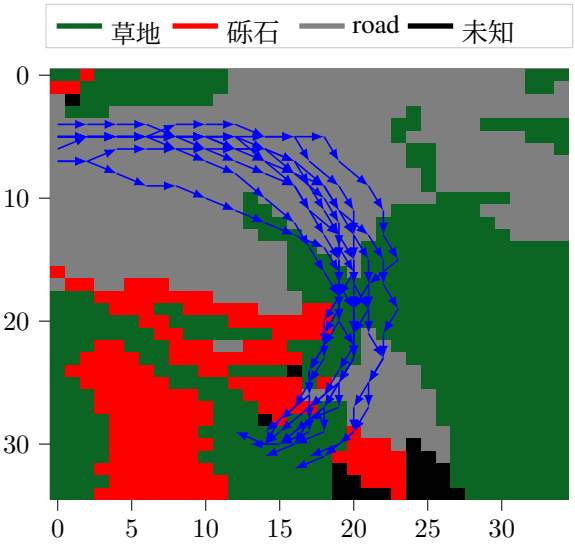
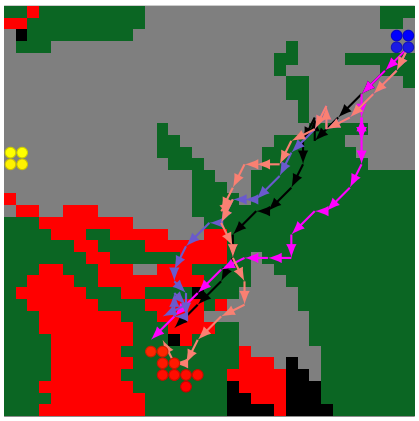


图 6：环境的网格世界表示。图中展示了我们应用所开发算法的 Unity 环境区域。

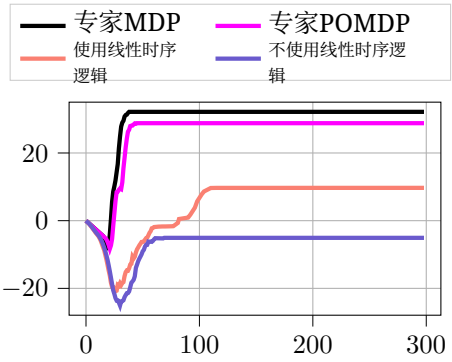
收集训练数据，并在两个实验场景中测试智能体根据学习到的奖励函数遵循策略的能力。在此实验中，我们展示了智能体从相对于已知真实奖励函数而言并非最优的示范中学习奖励函数的能力。我们还展示了学习到的策略与通过使用真实奖励函数求解马尔可夫决策过程或部分可观测马尔可夫决策过程问题而获得的、基于完整观测和部分观测的最优策略相比，表现如何。

地面车辆包含一个自主系统栈，由三个主要子系统组成——建图、感知和规划。基于 Omni-Mapper[?] 的建图子系统使用激光雷达和惯性测量单元传感器进行同步定位与建图（SLAM），并提供规划过程中使用的地图。感知子系统将来自 RGB 摄像机视频流的每张图像进行像素级语义分割，标注为地形和对象类别本体中的类别。每张语义图像都会传入地形投影算法，以构建已知环境的 N 二值占用特征图，用于奖励学习，其中 N 为特征数量。规划子系统使用前述子系统生成的地图，以及学习到的策略生成的轨迹，自主导航至航点。

专家示范与奖励特征编码。 我们收集了 10 次专家遥操作机器人前往预定航点的示范（见图 6）。专家隐含地偏好先走道路，其次是草地，最后是砾石。因此，我们将未知奖励函数编码为已知特征的线性组合：
$\mathcal{R}^\theta = \theta_1\phi^{\text{road}} + \theta_2\phi^{\text{gravel}} + \theta_3\phi^{\text{grass}} + \theta_4\phi^{\text{time}} + \theta_5\phi^{\text{goal}}$，其中$\phi^i$ 在对应状态的特征不是特征 i时返回 0，否则返回 1。为鼓励采用最短路径，时间特征会根据到达航点前在环境中采取的动作数量施加惩罚。此外，



(a) 执行各策略后得到的轨迹，分别展示了使用和不使用任务规范的情况。利用任务规范的学习者（橙色）能够到达某个目标状态，同时避开沿途的砾石。相比之下，没有辅助信息的学习者（紫色）未能避开砾石。



(b) 学习者获得的累积奖励随环境交互次数的变化。对于真实奖励函数，专家MDP和专家POMDP分别是在MDP和POMDP上的最优策略。

图 7：将任务规范纳入奖励学习的影响 。

目标特征会在到达航点时提供正奖励。为了比较学习到的策略，我们使用值 $\theta = [0.2, -30, -2, -0.5, 50]$ 作为真实奖励权重向量。需要强调的是，相对于上述真实奖励，这些示范并非最优，因为车辆经常驶过砾石路段，对应着较高的惩罚奖励。

将机器人动力学建模为部分可观测马尔可夫决策过程。 根据仿真环境的真实地图，我们获得了学习者在整个状态空间中的行为所对应的高层MDP抽象。随后，我们按如下方式引入机器人的部分可观测性：机器人无法看到整个世界地图，只能看到其当前位置周围固定半径 $r = 4$（以网格单元格数量计）范围内的区域。此外，我们还对地形特征的传感器分类引入不确定性，使得预测正确的概率为 $p = 0.9$ 。

任务规范。 除了专家示范之外，我们还约束学习到的策略，使其满足 $\Pr^{\sigma}_{\mathcal{M}}(\neg$砾石 **U** 目标$) \geq 0.9$，其中，砾石是一个原子命题，对于以砾石为特征的状态为真；目标是一个原子命题，在每个目标状态下为真。请注意，这些辅助信息并不一定要求学习者到达目标状态集合。相反，如果学习者到达目标状态，则其以至少.....的概率不应行驶在砾石上

结果。 图 7a 展示了与不使用辅助信息的学习者相比，使用辅助信息的学习者如何避开砾石。图 7b 进一步说明了这一结果，并通过实证表明，所提出的方法能够有效利用辅助信息，计算出符合专家期望行为的策略。具体而言，图 7b 显示，不使用辅助信息的学习者的总奖励增幅比使用辅助信息的学习者高 294%。此外，值得注意的是，图 6 中示范者轨迹的初始状态分布与评估学习到的策略时的初始状态分布（图 7a）并不相同。

尽管存在这些差异，学习到的策略仍能有效地导航至专家示范中出现的点，并最大程度地模仿这些轨迹。

7. 相关工作。

与我们最接近的工作是 [34], ，他们将经典的、基于最大间隔的马尔可夫决策过程逆向强化学习技术扩展到了部分可观测马尔可夫决策过程。然而，即使在马尔可夫决策过程中，基于最大间隔的方法也无法消除次优示范造成的歧义；当存在一个明显优于其他选择的单一奖励函数时，这些方法才表现良好 [39]。相比之下，我们采用了因果熵，已有研究表明它能够缓解这些方法在马尔可夫决策过程中的局限性 [39, 10]。此外，[34] 依赖高效的现成求解器来处理前向问题。与此不同，本文还提出了一种算法，其性能优于现成求解器，并且能够扩展到规模比 [34] 中的示例大几个数量级的部分可观测马尔可夫决策过程。此外， [34] 的公式中未纳入任务规范。

逆向强化学习面临的一项基本挑战是：寻找一个能够诱导出与专家相似行为的奖励函数和策略，这是一个定义不明确的问题。以往工作采用最大间隔公式 [40, 41, 42], ，以及概率模型来计算专家示范的似然 [43, 8, 10]，从而应对这一挑战。我们基于后一种方法，并进一步采用最大因果熵逆向强化学习 [9,10, 23], ；如引言所述，这种方法为部分可观测马尔可夫决策过程中的逆向强化学习带来了算法上的优势。需要指出的是，这些最大因果熵逆向强化学习技术假设专家和智能体都能够完全观测环境，因此这些方法仅适用于马尔可夫决策过程，而不适用于部分可观测马尔可夫决策过程。

先前的研究探讨了部分信息条件下的逆向强化学习 [2, 44, 45, 46, 47]。文献 [44] 考虑的是奖励函数的特征仅被部分指定的情形，而非环境状态信息不完整的情形。 [2]的工作考虑了部分可观测马尔可夫决策过程的一种特殊情况。它仅根据示范推断专家未来轨迹的分布，而不计算能够诱导出与专家相似行为的策略。 [45, 46, 47] 中的研究假设环境状态要么完全可观测，要么对学习智能体完全隐藏。因此，与 [2]类似，这些方法也考虑了部分可观测马尔可夫决策过程的一种特殊情况。我们还注意到，这些方法均未将辅助信息纳入逆向强化学习，也未能保证策略在任务规范下的性能。

利用时序逻辑表达的辅助信息来引导和增强逆向强化学习的想法，已在一些先前的工作中得到探索。在 [48, 22], 中，作者将辅助信息纳入时序逻辑规范，以学习能够产生与专家示范相似行为且满足该规范的策略。文献 [21] 通过迭代推断与专家示范一致的底层任务规范，并学习满足该任务

规范的策略和奖励函数。然而，这些方法也假设专家和智能体都拥有完整信息。

8. 结论

我们提出了一种用于部分观测下逆向强化学习的算法。通过实证，我们表明，将任务规范纳入学习过程，能够缓解专家与学习者之间的信息不对称，同时提高学习方案的数据效率。此外，我们还通过实证表明，在逆向强化学习算法中使用的核心例程 SCPForward 能够以可扩展的方式求解前向问题，并且在状态、观测和状态转移数量较多的实例上，其表现优于最先进的部分可观测马尔可夫决策过程求解器。

工作局限性。 本研究假设算法已知部分可观测马尔可夫决策过程的状态转移函数和观测函数。未来的工作将探索如何取消这一假设，并开发基于无模型的方法。我们还将把该框架与表达能力更强的基于神经网络的奖励函数相结合。

致谢。。 本研究由美国陆军研究实验室和美国海军研究办公室赞助，并依据合作协议编号 ARL W911NF-20-2-0132、ARL W911NF-19-2-0285 和 ONR N00014-22-1-2254 开展。本文所述观点和结论仅代表作者，不应被解读为美国陆军研究实验室、美国海军研究办公室或美国政府明示或暗示的官方政策。尽管本文有任何版权声明，美国政府仍有权为政府用途复制和分发重印本。

参考文献

[1] P. Abbeel, A. Coates, A. Y. Ng, 通过学徒学习实现自主直升机特技飞行，《国际机器人研究杂志》29 (13) (2010) 1608–1639。[2] K. M. Kitani, B. D. Ziebart, J. A. Bagnell, M. Hebert, 活动预测，载于：欧洲计算机视觉会议，施普林格，2012年，页码 201–214。[3] D. Hadfield-Menell, S. J. Russell, P. Abbeel, A. D. Dragan, 协作逆向强化学习，载于：神经信息处理系统大会，2016年。[4] A. D. Dragan, S. S. Srinivasa, 共享控制的策略融合形式体系，《国际机器人研究杂志》32 (7) (2013) 790–805。[5] T. Osa, N. Sugita, M. Mitsuishi, 动态环境中的外科任务自动化在线轨迹规划，载于：机器人学：科学与系统，Citeseer，2014年，页码 1–9。

[6] T. Osa, N. Sugita, M. Mitsuishi, 外科任务自动化的在线轨迹规划与力控制，IEEE自动化科学与工程汇刊 15 (2) (2017年) 675–691。[7] C. Finn, S. Levine, P. Abbeel, 引导式代价学习：通过策略优化实现深度逆最优控制，载于：国际机器学习会议，PMLR，2016年，页码 49–58。[8] B. D. Ziebart, A. L. Maas, J. A. Bagnell, A. K. Dey, 最大熵逆强化学习，载于：AAAI，第 8 卷，2008年，页码 1433–1438。[9] Z. Zhou, M. Bloem, N. Bambos, 无限时域最大因果熵逆强化学习，IEEE自动控制汇刊 63 (9) (2017年) 2787–2802。[10] B. D. Ziebart, J. A. Bagnell, A. K. Dey, 基于最大因果熵原理的交互建模。[11] S. C. Ong, S. W. Png, D. Hsu, W. S. Lee, 面向混合可观测性机器人任务的部分可观测马尔可夫决策过程，载于：机器人学：科学与系统，第 5 卷，2009年，页码 4。[12] H. Bai, D. Hsu, W. S. Lee, 连续空间中的感知与规划集成：部分可观测马尔可夫决策过程方法，《国际机器人研究杂志》33 (9) (2014年) 1288–1302。[13] S. Zhang, J. Sinapov, S. Wei, P. Stone, 使用部分可观测马尔可夫决策过程进行机器人行为探索与多模态感知，载于：AAAI具身智能体交互式多感官物体感知春季研讨会，2017年。[14] K. Akash, K. Polson, T. Reid, N. Jain, 通过基于透明度的反馈改善人机协作——第一部分：人类信任与工作负荷模型，IFAC-PapersOnLine 51 (34) (2019年) 315–321。[15] X. Liu, A. Datta, 使用隐马尔可夫模型对上下文感知动态信任进行建模，载于：AAAI人工智能会议论文集，第 26 卷，2012年。[16] N. Vlassis, M. L. Littman, D. Barber, 部分可观测马尔可夫决策过程中随机控制器优化的计算复杂度，ACM计算理论汇刊 4 (4)。[17] O. Madani, S. Hanks, A. Condon, 概率规划与无限时域部分可观测马尔可夫决策问题的不可判定性，载于：AAAI，AAAI出版社，1999年，页码 541–548。[18] M. L. Littman, U. Topcu, J. Fu, C. Isbell, M. Wen, J. MacGlashan, 通过 GLTL 实现与环境无关的任务规范，arXiv 预印本 arXiv:1704.04341。[19] C. Baier, J.-P. Katoen, 模型检验原理，麻省理工学院出版社，2008年。

[20] A. Pnueli, 《程序的时序逻辑》, 载于: 第18届计算机科学基础年度研讨会论文集, 美国华盛顿哥伦比亚特区, 1977年, 第46–57页。doi:10.1109/SFCS.1977.32。[21] F. Memarian、Z. Xu、B. Wu、M. Wen、U. Topcu, 《主动任务推断引导的深度逆强化学习》, 载于: 2020年IEEE第59届决策与控制会议（CDC）, IEEE, 2020年, 第1932–1938页。[22] M. Wen、I. Papusha、U. Topcu, 《利用高层辅助信息从示范中学习》, 载于: 第二十六届国际人工智能联合会议论文集, 2017年。[23] B. D. Ziebart、J. A. Bagnell、A. K. Dey, 《用于估计交互过程的最大因果熵原理》, IEEE 信息论汇刊 59 (4) (2013) 1966–1980。[24] Y.-x. Yuan, 《信赖域算法的最新进展》, 《数学规划》151 (1) (2015) 249–281。[25] Y. Mao、M. Szmuk、X. Xu、B. Acıkmese, 《序列凸化: 一种针对非凸最优控制问题的超线性收敛算法》, arXiv预印本 arXiv:1804.06539。[26] H. Yu、D. P. Bertsekas, 《平均成本部分可观测马尔可夫决策过程中有限状态控制器集合的近似最优性》, 《运筹学数学》33 (1) (2008) 1–11。[27] S. Junges、N. Jansen、R. Wimmer、T. Quatmann、L. Winterer、J.-P. Katoen、B. Becker, 《利用参数综合构造部分可观测马尔可夫决策过程的有限状态控制器》, 载于: UAI论文集, 2018年, 第519–529页。[28] N. Meuleau、K.-E. Kim、L. P. Kaelbling、A. R. Cassandra, 《通过搜索有限策略空间求解部分可观测马尔可夫决策过程》, 载于: UAI, 1999年, 第417–426页。[29] C. Amato、D. S. Bernstein、S. Zilberstein, 《优化部分可观测马尔可夫决策过程和分散式部分可观测马尔可夫决策过程的定长随机控制器》, AAMAS 21 (3) (2010) 293–320。[30] J. Ho、S. Ermon, 《生成对抗模仿学习》, 《神经信息处理系统进展》29 (2016) 4565–4573。[31] J. Massey, 《因果性、反馈与定向信息》, 载于: 国际信息理论应用研讨会论文集（ISITA-90）, Citeseer, 1990年, 第303–305页。[32] G. Kramer, 《带反馈信道的定向信息》, Citeseer, 1998年。[33] L. P. Hansen、T. J. Sargent、G. Turmuhambetova、N. Williams, 《鲁棒控制与模型设定错误》, 《经济理论杂志》128 (1) (2006) 45–90。[34] J. Choi、K.-E. Kim, 《部分可观测环境中的逆向强化学习》, 《机器学习研究杂志》12 (2011) 691–730。

[35] S. Junges、N. Jansen、S. A. Seshia，《在部分可观测马尔可夫决策过程中确保几乎必然可达性》，载于：国际计算机辅助验证会议，2020年。[36] E. Walraven、M. Spaan，《用于最优部分可观测马尔可夫决策过程求解器的加速向量剪枝》，载于：AAAI人工智能会议论文集，第31卷，2017年。[37] H. Kurniawati、D. Hsu、W. S. Lee，《SARSOP：通过近似最优可达信念空间实现高效的基于点的部分可观测马尔可夫决策过程规划》，载于：机器人学：科学与系统，第2008卷，Citeseer，2008年。[38] G. Norman、D. Parker、X. Zou，《部分可观测概率系统的验证与控制》，《实时系统》53 (3) (2017) 354–402。[39] T. Osa、J. Pajarinen、G. Neumann、J. Bagnell、P. Abbeel、J. Peters，《模仿学习的算法视角》，《机器人学基础与趋势》7 (1-2) (2018) 1–179。[40] N. D. Ratliff、J. A. Bagnell、M. A. Zinkevich，《最大间隔规划》，载于：第23届国际机器学习会议论文集，2006年，第729–736页。[41] P. Abbeel、A. Y. Ng，《通过逆向强化学习进行学徒学习》，载于：第21届国际机器学习会议论文集，2004年，第1页。[42] A. Y. Ng、S. J. Russell等，《逆向强化学习算法》，载于：第17届国际机器学习会议论文集，第1卷，2000年，第2页。[43] D. Ramachandran、E. Amir，《贝叶斯逆强化学习》，载于：IJCAI，第7卷，2007年，第2586–2591页。[44] A. Boularias、O. Kr¨omer、J. Peters，《结构化学徒学习》，载于：欧洲机器学习与数据库知识发现联合会议，施普林格，2012年，第227–242页。[45] K. Bogert、P. Doshi，《存在遮挡和交互时的多机器人逆强化学习》，载于：2014年国际自主智能体与多智能体系统会议论文集，Citeseer，2014年，第173–180页。[46] K. Bogert、P. Doshi，《面向遮挡条件下多机器人逆强化学习中他者转移模型的估计》，载于：第二十四届国际人工智能联合会议，2015年。[47] K. Bogert、J. F.-S. Lin、P. Doshi、D. Kulic，《利用期望最大化从隐藏数据中进行逆强化学习》，载于：2016年国际自主智能体与多智能体系统会议论文集，2016年，第1034–1042页。

[48] I. Papusha、M. Wen、U. Topcu，《带正则语言规范的逆最优控制》，载于：2018年美国控制会议（ACC），IEEE，2018年，第 770–777 页。

附录

本附录为正文中的结果提供补充推导，并详细介绍数值实验。

A. 因果熵的凹性与贝尔曼约束的推导

本节首先回顾因果熵 H_σ^γ 关于访问计数 μ_σ^γ 和 ν_σ^γ的表达式。然后证明因果熵的凹性，这使得能够采用基于凸优化的表述来处理任务引导的逆向强化学习（逆向强化学习）问题。接着，进一步说明由贝尔曼流约束所蕴含的仿射约束的推导。

凹因果熵。 我们首先回顾状态访问计数和状态-动作访问计数的定义。对于策略 σ、状态 s和动作 α，折扣状态访问计数定义为 $\mu_\sigma^\gamma(s)$ $\triangleq \mathbb{E}_{S_t}[\sum_{t=1}^{\infty}\gamma^t \mathbb{1}_{\{S_t=s\}}]$，折扣状态-动作访问计数定义为 $\nu_\sigma^\gamma(s,\ \alpha) \triangleq \mathbb{E}_{A_t,\ S_t}[\sum_{t=1}^{\infty}\gamma^t \mathbb{1}_{\{S_t=s,A_t=\alpha\}}]$，其中 $\mathbb{1}_{\{\cdot\}}$为指示函数，t为时间步。根据状态访问计数和状态-动作访问计数 μ_σ^γ和 ν_σ^γ的定义，可以直接推得 $\nu_\sigma^\gamma(s,\ \alpha) = \sigma_{s,\ \alpha}\mu_\sigma^\gamma(s)$，其中 $\sigma_{s,\ \alpha} = \mathbb{P}[A_t = a|S_t = s]$。我们利用访问计数在第 4 节证明

$$H_\sigma^\gamma = \sum_{(s,\alpha)\in\mathcal{S}\times\mathcal{A}} -(\log \pi_{s,\alpha})\pi_{s,\alpha}\mu_\sigma^\gamma(s) = \sum_{(s,\alpha)\in\mathcal{S}\times\mathcal{A}} -\log\frac{\nu_\sigma^\gamma(s,\alpha)}{\mu_\sigma^\gamma(s)}\nu_\sigma^\gamma(s,\alpha),$$

其中，最后一个不等式是利用 $\pi_{s,\alpha} = \nu_\sigma^\gamma(s,\alpha)/\mu_\sigma^\gamma(s)$ 得到的。我们声称，H_σ^γ 是关于访问计数的凹函数。因此，我们要证明函数 $f(\nu_\sigma^\gamma,\mu_\sigma^\gamma) = \sum_{(s,\alpha)\in\mathcal{S}\times\mathcal{A}} -\log\ {}^{\nu_\sigma^\gamma(s,\alpha)}{}_{\mu_\sigma^\gamma(s)}\ \nu_\sigma^\gamma(s,\alpha)$ 是凹函数。为此，考虑任意的 $\lambda\in(0,1)$ 以及两组变量 $\nu_\sigma^\gamma,\mu_\sigma^\gamma$和 $\bar{\nu}_\sigma^\gamma,\bar{\mu}_\sigma^\gamma$。于是，我们有

以下结果:

$$
\begin{aligned}
&f(\lambda\nu_\sigma^\gamma + (1-\lambda)\bar{\nu}_\sigma^\gamma, \lambda\bar{\mu}_\sigma^\gamma + (1-\lambda)\bar{\mu}_\sigma^\gamma) \\
&\quad= \sum_{(s,\alpha)\in\mathcal{S}\times\mathcal{A}} -\log\frac{\lambda\nu_\sigma^\gamma(s,\alpha) + (1-\lambda)\bar{\nu}_\sigma^\gamma(s,\alpha)}{\lambda\mu_\sigma^\gamma(s) + (1-\lambda)\bar{\mu}_\sigma^\gamma(s,\alpha)}(\lambda\nu_\sigma^\gamma(s,\alpha) + (1-\lambda)\bar{\nu}_\sigma^\gamma(s,\alpha)) \\
&\quad\geq \sum_{(s,\alpha)\in\mathcal{S}\times\mathcal{A}} -\lambda\nu_\sigma^\gamma(s,\alpha)\log\frac{\lambda\nu_\sigma^\gamma(s,\alpha)}{\lambda\mu_\sigma^\gamma(s,\alpha)} - (1-\lambda)\bar{\nu}_\sigma^\gamma(s,\alpha)\log\frac{(1-\lambda)\bar{\nu}_\sigma^\gamma(s,\alpha)}{(1-\lambda)\bar{\mu}_\sigma^\gamma(s,\alpha)} \\
&\quad= \sum_{(s,\alpha)\in\mathcal{S}\times\mathcal{A}} -\lambda\nu_\sigma^\gamma(s,\alpha)\log\frac{\nu_\sigma^\gamma(s,\alpha)}{\mu_\sigma^\gamma(s,\alpha)} - (1-\lambda)\bar{\nu}_\sigma^\gamma(s,\alpha)\log\frac{\bar{\nu}_\sigma^\gamma(s,\alpha)}{\bar{\mu}_\sigma^\gamma(s,\alpha)} \\
&\quad= \lambda f(\nu_\sigma^\gamma, \mu_\sigma^\gamma) + (1-\lambda) f(\bar{\nu}_\sigma^\gamma, \bar{\mu}_\sigma^\gamma),
\end{aligned}
$$

其中，第一个不等式是通过应用众所周知的对数和不等式得到的，即

$$
x_1\log\frac{x_1}{y_1} + x_2\log\frac{x_2}{y_2} \geq (x_1+x_2)\log\frac{x_1+x_2}{y_1+y_2},
$$

对于非负数 x_1，x_2，y_1，y_2。具体而言，我们进行如下代换：$x_1 = \lambda\nu_\sigma^\gamma$, $y_1 = \lambda\mu_\sigma^\gamma$, $x_2 = (1-\lambda)\bar{\nu}_\sigma^\gamma$，以及 $y_2 = (1-\lambda)\bar{\mu}_\sigma^\gamma$。注意，该不等式

$$
f(\lambda\nu_\sigma^\gamma + (1-\lambda)\bar{\nu}_\sigma^\gamma, \lambda\bar{\mu}_\sigma^\gamma + (1-\lambda)\bar{\mu}_\sigma^\gamma) \geq \lambda f(\nu_\sigma^\gamma, \mu_\sigma^\gamma) + (1-\lambda) f(\bar{\nu}_\sigma^\gamma, \bar{\mu}_\sigma^\gamma)
$$

意味着 $f(\nu_\sigma^\gamma,\ \mu_\sigma^\gamma)$ 关于 ν_σ^γ、以及 μ_σ^γ 是凹函数。

贝尔曼流约束。为了使访问计数变量对应于在部分可观测马尔可夫决策过程 $\mathcal{M}$ 中生成动作的有效策略，ν_σ^γ 和 μ_σ^γ 必须满足以下贝尔曼流约束:

$$
\begin{aligned}
\mu_\sigma^\gamma(s) &= \mathbb{E}_{S_t^\sigma}\Big[\sum_{t=0}^{\infty}\gamma^t \mathbb{1}_{\{S_t^\sigma = s\}}\Big] \\
&= \mu_0(s) + \gamma\mathbb{E}_{S_t^\sigma}\Big[\sum_{t=0}^{\infty}\gamma^t \mathbb{1}_{\{S_{t+1}^\sigma = s\}}\Big] \\
&= \mu_0(s) + \gamma\sum_{t=0}^{\infty}\sum_{s'\in\mathcal{S}}\sum_{\alpha\in\mathcal{A}}\gamma^t\mathcal{P}(s|s',\alpha)\mathbb{P}[S_t^\sigma = s', A_t^\sigma = \alpha] \\
&= \mu_0(s) + \gamma\sum_{s'\in\mathcal{S}}\sum_{\alpha\in\mathcal{A}}\mathcal{P}(s|s',\alpha)\nu_\sigma^\gamma(s',\alpha).
\end{aligned}
$$

B. 实验任务

在本节中，我们首先详细介绍基准测试中使用的部分可观测马尔可夫决策过程模型。基准测试示例的仿真结果通过实证表明，辅助信息能够缓解信息不对称，而更多的记忆能够带来性能更好的策略。随后，我们提供了更多数值仿真，以支持以下主张：对于前向问题，即针对给定奖励函数计算部分可观测马尔可夫决策过程上的最优策略，SCPForward 是可靠的，且比现成求解器具有更好的可扩展性。

B.1. 计算资源和外部资源

本文的所有实验均在一台配备 Intel Core$i9$-9900 CPU、3.1 GHz $\times 16$ 处理器和 31.2 GB 内存的计算机上完成。所有实现均使用 Python 3.8 编写并经过测试，代码随补充材料一并提供。

所需工具。我们的实现需要使用*Storm*的*Stormpy* [?] 以及*Gurobi*的*Gurobipy* 9.1 [?]。一方面，我们使用*Storm*这一模型检测工具来解析部分可观测马尔可夫决策过程文件规范，计算与有限状态控制器的乘积部分可观测马尔可夫决策过程，从而将合成问题简化为无记忆策略的合成，并通过图预处理计算满足规范 φ 的目标状态集合 $\mathcal{T}$。另一方面，我们使用*Gurobi* 求解式（7）中的线性化问题，以及验证步骤所需的贝尔曼流约束的可行解。

用于前向问题的现成求解器。为了展示所开发算法 SCPForward 的可扩展性，我们将其与最先进的部分可观测马尔可夫决策过程求解器 SolvePOMDP [36], SARSOP [37], 和 PRISM-POMDP [38]进行比较。求解器 SolvePOMDP 通过增量剪枝 [?] 结合最先进的向量剪枝方法 [36]，实现精确值迭代和近似值迭代。最后，PrismPOMDP 对信念状态进行离散化，并采用有限记忆策略来求解前向问题的近似解。对于上述所有求解器，我们均采用默认设置，但将超时限制设为 3600 秒。这些求解器不包含在我们的实现中。不过，我们提供了各求解器可直接使用的部分可观测马尔可夫决策过程模型。更多详情请参阅我们实现中的 README 文件。

B.2. 基准集合

我们在多个由 [35] 改编而来的部分可观测马尔可夫决策过程实例上评估所提出的学习算法。我们在代码中附上了修改后的实例，以及为每个现成求解器自动生成的模型，读者可以直接使用这些模型复现表1。读者可参阅表1，了解基准集合中各环境的状态数、观测数和状态转移数。在所有示例中，我们从一位能够完全观测环境当前状态的专家，以及一位只能部分观测环境的专家那里收集了10条轨迹。我们的算法根据这些轨迹，在不同记忆策略和高级侧信息的条件下学习奖励函数。

岩石实例。 在岩石环境中，智能体在网格世界中移动，目标是在两块可能具有危险性的岩石中，至少采集一块有价值的岩石（如果网格中有有价值的岩石），且不能发生任何失败。采集到至少一块有价值的岩石后，或智能体确认所有岩石都具有危险性后，就需要前往出口点以结束任务。部分可观测性源于智能体只能观测到当前位置是否为出口点或危险岩石。此外，智能体配备了噪声传感器，可以对相邻单元格进行采样。

我们考虑三个特征函数。第一个特征函数会在到达出口点时给予正奖励，条件是至少持有一块有价值的岩石；如果所有岩石都很危险，则不持有任何岩石也可以。第二个特征函数会在智能体位于危险岩石所在位置时给予负奖励。最后，第三个特征函数会对每个采取的动作给予负奖励，以促使智能体尽快到达出口点。

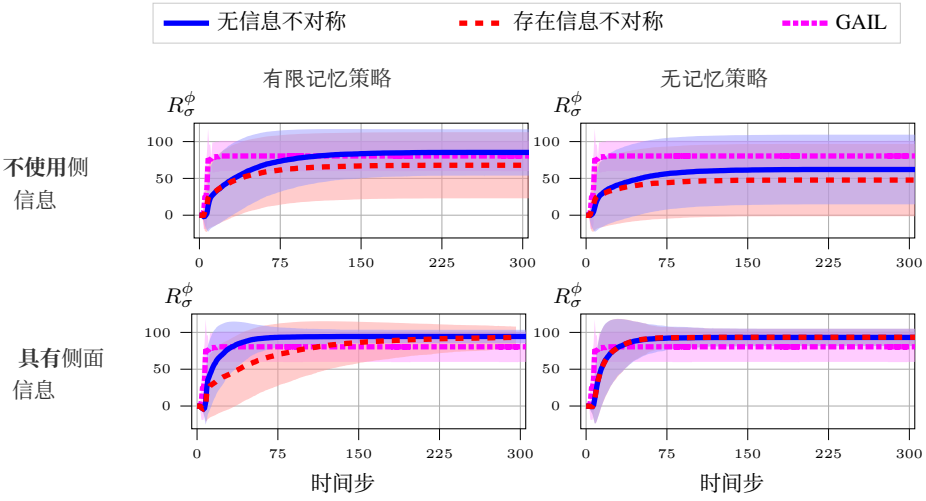


图 8：Rock 示例中的代表性结果，展示策略在真实奖励函数下的奖励（R_σ^ϕ）随时间步的变化。

我们比较了没有辅助信息的情形，以及已知任务先验知识的情形，例如智能体最终到达出口点的概率大于0。995。 图 8 支持了我们的观点：辅助信息确实缓解了专家与智能体之间的信息不对称。此外，我们还观察到，加入记忆后，策略在平均累积奖励方面表现更佳。

障碍物实例。. 在环境 Obstacle$[n]$ 中，智能体必须在网格世界中找到出口，同时避免与网格中的五个静态障碍物发生碰撞。智能体只能观察到当前位置是障碍物还是出口状态。参数n 指定网格的尺寸。

与 theRocks 示例类似，如果智能体成功离开网格世界，则会获得正奖励；每执行一次动作或与障碍物碰撞，都会受到负奖励。

对于辅助信息，我们用时序逻辑规定，在学习奖励的过程中，智能体在以大于0.9的概率到达出口点之前，不应与任何障碍物碰撞。

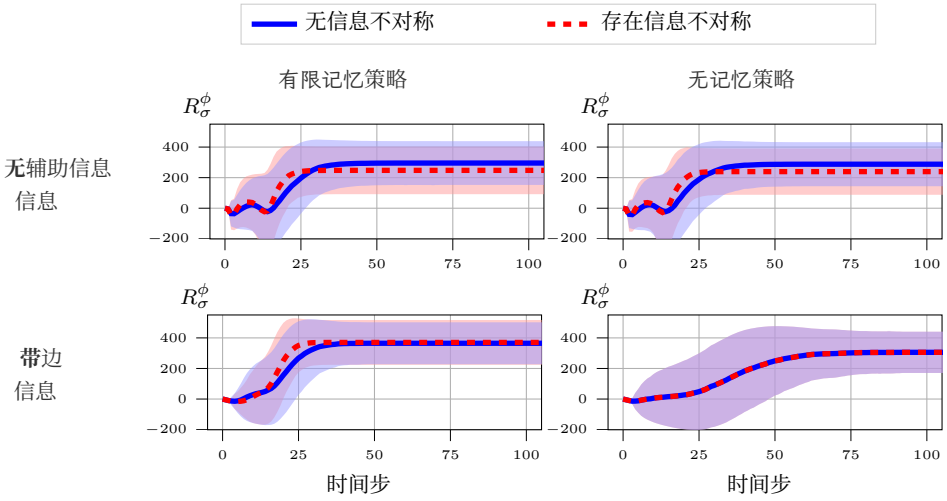


图 9：障碍物示例中的代表性结果，展示了策略在真实奖励函数 (R_σ^ϕ) 下的奖励随时间步的变化。

与迷宫和岩石示例类似，图 9 支持了我们的观点：侧面信息可以缓解信息不对称，而记忆有助于提升策略的表现。

躲避实例。 躲避$[n, r, slip]$是一种回合制游戏，智能体必须在被速度更快的玩家拦截之前抵达目的地。玩家无法进入网格的顶行。此外，只有当玩家位于智能体当前位置的固定半径范围内，且智能体调用扫描动作时，智能体才能观察到玩家。参数 n、 r 和 $slip$分别指定网格尺寸、视野半径和滑动概率。

特征函数定义为：智能体到达目的地时获得正奖励，被玩家拦截时获得较高的负奖励，并且每执行一个动作（包括扫描动作）都会获得少量负奖励。

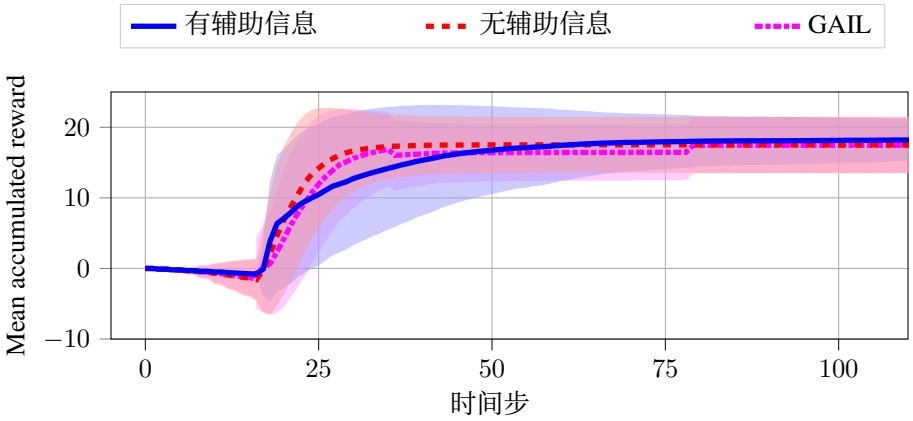


图10：躲避示例中的代表性结果，展示了策略在真实奖励函数 (R_S^{ϕ}) 下的奖励与时间步的关系。

对于辅助信息，我们用时序逻辑规定，在学习奖励的过程中，智能体必须以大于0.98的概率到达一个出口点。

图10显示，利用辅助信息进行学习所获得的奖励高于不使用辅助信息的情况。此外，与不使用辅助信息的策略相比，使用辅助信息的策略随机性更低。具体而言，使用辅助信息的策略的标准差显著小于不使用辅助信息的策略。

本例中我们没有讨论不同记忆大小策略的影响，因为无记忆策略的性能已接近最优；该策略获得的奖励与 SARSOP 相同（参见表1）。具体而言，我们观察到，底层马尔可夫决策过程上的最优策略与部分可观测马尔可夫决策过程上的最优无记忆策略性能相当。因此，我们也观察到，专家与智能体之间的信息不对称在此处同样不存在，且学习到的策略取得了相似的性能。

拦截实例。 拦截$[n, r, slip]$是躲避的一种变体，其中智能体必须拦截另一名试图离开网格世界的玩家。智能体可以向8个方向移动，并且只有在执行扫描 动作时，且玩家处于距离智能体当前位置固定半径范围内，才能观察到玩家。此外，智能体配备了摄像机，可以从网格世界中心由西向东观察所有单元格。相比之下，玩家只能向4个方向移动。参数 n、 r和 $slip$ 分别指定网格尺寸、视野半径和滑动概率。

我们使用三个特征函数对未知奖励进行参数化。第一个特征在智能体拦截玩家时给予其正奖励。第二个特征在玩家离开网格世界时惩罚智能体。第三个特征会对每个执行的动作施加惩罚成本。

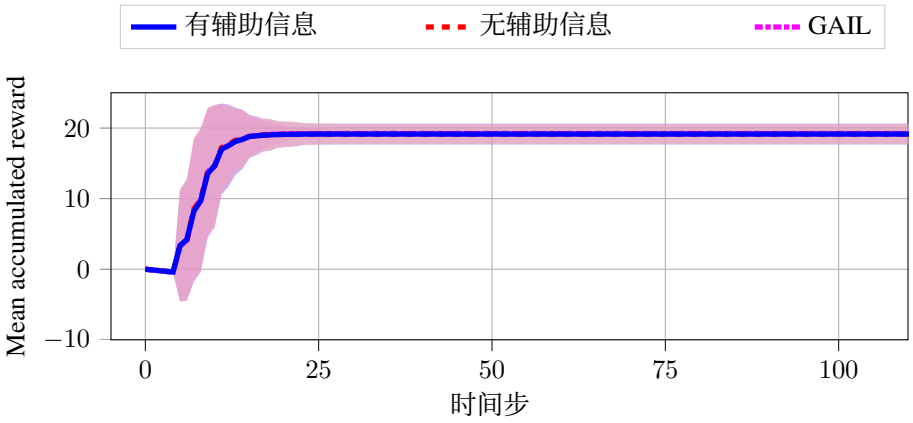


图11：拦截示例的代表性结果，展示了策略在真实奖励函数下的奖励（R_σ^ϕ）与时间步之间的关系。

我们将高级侧信息编码为时序逻辑任务规格以大于 0.98 的概率最终算子拦截玩家，即智能体最终应到达一个自身位置与玩家位置重合的观测状态。

图11表明，辅助信息并不能提升策略的性能。这是因为在此示例中，无记忆策略是最优的，而且给定奖励特征的组合可以完美地编码时序逻辑规范，与“躲避”示例类似。

B.3. 侧信息的影响

本节将通过更多实验，展示辅助信息如何加快学习过程，包括缩短计算时间以及加快收敛到最优策略和奖励参数。随后，我们仅使用任务规范而不使用示范来求解部分可观测马尔可夫决策过程，从而量化辅助信息的影响。所有这些实验均在迷宫示例上进行，这是一个规模相对较小的部分可观测马尔可夫决策过程示例。

有侧信息与无侧信息的学习过程的收敛。 在迷宫实验中，我们通过实验证明，与不使用辅助信息的学习相比，辅助信息使学习算法所需的迭代次数减少到八分之一。此处的迭代次数既指序列凸规划方案中的线性化次数，也指奖励更新期间的梯度步数。然而，计算时间的提升不如迭代次数的减少那么显著。事实上，由于辅助信息几乎使凸优化问题中的变量数量翻倍，使用辅助信息的学习速度仅约为不使用辅助信息时的三倍。

无任何示范时侧信息的影响。

- 实验 *1*。 我们考虑迷宫示例中的相同设定：没有示范，线性时序逻辑规范为 $\Pr^{\sigma}_{\mathcal{M}}(\mathbf{G}\ \neg\text{bad}) \geq 0.9$。本质上，我们要寻找能够以较高概率避开陷阱状态的策略。若没有任何额外奖励，最优策略恰如我们所料：选择不移动动作的概率很高，而选择其他动作的概率很低。就真实奖励而言，这显然不是最优的，因为由于没有尽量缩短在环境中花费的时间，奖励会持续下降。现在，我们针对满足线性时序逻辑规范的策略优化问题，同时依据真实奖励对在环境中花费的时间进行惩罚，因为在环境中采取更多步会受到惩罚。得到的策略具有 47。83 的最优奖励，这相当于在部分可观测马尔可夫决策过程中优化真实奖励。事实上，最快通关迷宫的方法是在不陷入陷阱的情况下经由目标状态离开。

- 实验 *2.* 我们考虑迷宫示例中的相同设置，不提供示范，并采用线性时序逻辑规范 $\Pr^{\sigma}_{\mathcal{M}}(\mathbf{E}\ \text{目标}) \geq 0.95$。本质上，我们要寻找能够以较高概率最终到达目标状态（迷宫出口）的策略。若不添加任何额外奖励，相对于给定真实奖励时部分可观测马尔可夫决策过程上的最优策略而言，该策略是次优的。事实上，由于到达目标所花费的时间，该策略可获得的最优奖励为 28。63。将环境中耗费时间的奖励加入线性时序逻辑规范后，我们便能再次获得部分可观测马尔可夫决策过程上的最优奖励。

B.4. 结果总结

辅助信息缓解信息不对称 。 正如已提交的手稿中所述，辅助信息确实能够缓解信息不对称。具体而言，我们发现，如果在前向问题中存在信息不对称，即底层部分可观测马尔可夫决策过程上的最优策略所获得的奖励低于底层全可观测马尔可夫决策过程上的最优策略所获得的奖励，那么将辅助信息纳入时序逻辑规范能够缓解专家与智能体之间的信息不对称。例如，我们可以在迷宫、岩石、障碍物和 Avoid 示例中看到这种信息不对称的影响。在

这些示例中，部分可观测性会降低前向问题中获得的奖励。在信息不对称的情况下，未将辅助信息纳入学习过程的策略所获得的奖励也较低。

记忆带来性能更优的策略。与辅助信息类似，我们也观察到，在迷宫、岩石和障碍物实例中，引入记忆不仅能提升学习到的策略的性能，还能提高前向问题中获得的奖励。从表1中还可以看出，在这些示例中，引入记忆有助于计算出更好的最优策略，而计算无记忆策略则不然。