合作式多智能体系统中的能力类型结构分散式推断

Charles Jin[1] **Zhang-Wei Hong**[1] **Farid Arthaud**[1], * **Idan Orzech**[1], * **Martin Rinard** [1]

[1]CSAIL、麻省理工学院、马萨诸塞州剑桥 02139*贡献相同

arXiv:2304.13957v1 [cs.MA] 27 Apr 2023

摘要

本文研究由计算能力各异的智能体组成的团队中的临时团队协作问题。我们考虑合作式多人游戏，其中每个智能体的策略都受到私有能力参数的约束，并且能力较高的智能体能够模拟能力较低智能体的行为（反之则不然）。为应对这一挑战，我们提出一种算法，用于维护对其他智能体能力的信念，并将该信念纳入规划过程。我们的主要创新是一种基于能力类型结构的新框架，它确保信念更新在不构建无限层级信念的情况下保持一致且具有信息量。我们还将这些技术扩展到智能体的观测受到噪声影响的情形。我们给出了相关游戏示例，说明无知智能体之间的能力差异可能导致任意糟糕的结果，并通过实验验证，在这些玩具环境以及更复杂的合作式跳棋环境中，我们的能力感知算法能够避免朴素方法产生的反合作行为。

1 引言

多智能体系统（MAS）研究日益关注如何开发能够与其他智能体协作以实现共同目标的自主智能体。在许多现实场景中，智能体可能具有不同的计算能力。例如，在一支自主机器人队伍中，有些机器人可能比其他机器人拥有更好的传感器或更强的处理能力，这会影响它们完成共同目标的能力。由于传感器可能发生故障，处理能力也可能波动，因此开发能够根据不同能力进行在线推断和适应的技术至关重要。

临时团队协作 [Mirsky 等人， 2022]已成为研究多智能体系统中协作问题的框架。在这种框架下，智能体在没有事先协调的情况下临时组建团队，以完成给定任务。当智能体之间的通信受限，或智能体需要与陌生或未知的智能体协作时，临时团队协作会很有益处。然而，现有方法没有考虑智能体之间计算能力的差异，而且通常只支持单个自适应智能体。

我们提出一种能力感知型临机团队协作方法，其中智能体必须适应彼此的计算能力。我们的设定假定存在一个能力层级，能力较强的智能体可以模拟能力较弱的智能体所采用的策略。我们的方法基于分散式控制领域的公共智能体方法 [Nayyar 等人， 2013]以及 Harsanyi 关于贝叶斯博弈类型结构的开创性研究 [Harsanyi， 1967] 。此外，我们的方法通过纳入智能体的能力，扩展了以往的临时团队协作方法。我们通过若干玩具设定和一项合作跳棋任务中的实验，展示了该方法的有效性。实验结果表明，我们的能力感知算法优于未考虑智能体计算差异的传统方法。

2 相关工作

临时团队协作的目标是设计能够以有限协调适应新队友的智能体 [Mirsky 等人， 2022] 。然而，大多数研究关注的是单个临时团队协作智能体的设置，其余智能体则不具备适应能力。一个值得注意的例外是 Albrecht 等人的研究 [2016], , 其中队友可以使用各自独立的策略进行适应，从而避免高阶信念的问题。据我们所知，本研究也是首个考虑不同

联系邮箱：ccj@csail.mit.edu。

计算能力。

我们的工作还与协作多智能体规划中的队友/对手建模相关 [Torreno 等人， 2017]。协作多智能体规划旨在协调一个智能体团队，以最大化联合效用函数。经典工作在每个智能体上采用启发式搜索算法，以最大化联合效用。例如， [Desaraju 和 How，2011] 使用快速扩展随机树（RRT）算法，为团队中的智能体规划无碰撞路径。除启发式搜索之外，先前的工作 [Lowe 等人，2017 年，Foerster 等人， 2018] 在多智能体强化学习中 [Sutton 和 Barto， 2018] 联合学习每个智能体的策略，包括多智能体信念系统 [Ouyang 等人，2016 年，Foerster 等人，2017 年、2019 年，Vasal 和 Anastasopoulos， 2022]。相比之下，我们关注在线推断与适应问题，尤其关注当推断过程也受到能力约束时所产生的复杂性。

最后，许多研究探讨了计算能力受限的智能体，尤其是人机交互领域的相关研究。众所周知，人类会表现出次优行为，这给那些将人类建模为完全理性个体的研究带来了挑战。一种常见的方法称为有限理性，它假设智能体在某些约束条件下能够达到（近似）最优 [Russell and Wefald, 1991, Simon，1997]。例如，一些研究探讨了推理其他智能体的递归推理能力有限的智能体 [Fotiadis and Vamvoudakis, 2021, Wen et al.，2019]，，并证明，当群体中的递归推理深度遵循泊松分布时，系统会收敛。[Nikolaidis et al.，2016] 将人类队友建模为具有有限记忆，即他们只会基于最近的 k 次交互进行推理。然而，所有这些研究都将“受限”的程度视为已知量；相比之下，我们研究的是如何适应未知的受限程度。

3 背景与设定

我们考虑一个包含 $N \geq 2$ 名玩家的完全协作随机博弈设定。为对具有异质能力的玩家建模，我们假设其能力取自一个已知的全序集合**能力类型** $\mathcal{C}$。例如，考虑这样一类策略：将穷举前瞻搜索进行到最大深度 k ，并在叶节点处使用价值函数近似。这类策略的能力类型集合可以指定搜索深度的上限，能力越强，对应的搜索深度越大。我们采用以下约定：层级中较低的类型（$<$）在能力方面较弱。

在游戏开始时，每位玩家 i 都会从 $\mathcal{C}$中随机分配到一种能力类型 c_i 。由于玩家的类型会限制其采取甚至考虑某些策略的能力，因此通常而言，在不同能力分配下的最优玩法也可能需要不同的（联合）策略。我们的目标是让玩家在不进行显式通信的情况下，学习彼此的私有能力类型，并收敛到一个信息充分的联合策略。为便于阐述，我们假设基础博弈（具有公开类型）是完全可观测的，并且每个时间步只有一个智能体采取行动。

形式上，该游戏以离散时间、部分可观测的去中心化马尔可夫决策过程（Dec-POMDP） [Littman，1994]，的形式进行，由元组$\langle \mathcal{S}, \mathcal{N}, \mathcal{C}, \mathcal{A}, P, R, \rho_0, \gamma \rangle$给出，其中 $\mathcal{S}$是基础博弈的所有状态集合， $\mathcal{N} = \{1, 2, \ldots, N\}$ 是玩家集合，$\mathcal{C}$ 是能力类型集合， $\mathcal{A}$ 是可用动作集合， P: $\mathcal{S} \times \mathcal{A} \to \Delta(\mathcal{S} \times \mathcal{N})$给出状态转移的分布， R: $\mathcal{S} \times \mathcal{A} \to \Delta([0, 1])$给出（有界）奖励的分布， $\rho_0 = \Delta(\mathcal{S} \times \mathcal{N} \times \mathcal{C}^N)$是初始状态的分布， $\gamma \in$ (0,1)是折扣因子。游戏的每个回合都从初始状态 s(0)、玩家 i(0)以及从 ρ_0 中抽取的能力分配 $C = (c_1, \ldots, c_N)$开始。每位玩家私下观察分配给自己的能力类型。在时间步t，所有玩家观察状态 $s(t)$，随后玩家 $i(t)$选择动作 $a(t) \in \mathcal{A}$；所有玩家获得奖励$r(t) \sim R(s(t)\ a(t))$，并转移至下一个时间步 $t+1$，此时状态为 $s(t+1)\ i\,t+1 \sim P(s(t)\ a(t))$. 目标是最大化期望联合回报$\mathbb{E}\big[\sum_{t=0}^{\infty} \gamma^t r(t)\big]$。

3.1 DEC-POMDP 中的规划

在环境部分可观测的单智能体 POMDP 设置中，智能体维护着对环境真实状态的信念 b，该信念初始化为某个先验分布，并随着游戏的进行依据贝叶斯法则不断更新。这种信念是智能体动作和观测历史的充分统计量[Kaelbling 等人， 1998]。因此，每个动作不仅会对下一状态诱导出一个后验分布，还会对下一信念诱导出一个后验分布。由此，我们可以通过将信念视为状态的一部分来构建相应的 信念状态 *MDP*，从而将 POMDP 化简为标准 MDP，并使用值迭代等标准技术求解 POMDP。

然而，在多智能体场景中，仅有关于环境的一阶信念并不足够，因为其他智能体也可能具有部分可观测的内部状态，其中包含私有信息。特别是，由于玩家的动作可能取决于其私下持有的信念，为了计算最优策略，其他智能体必须维护关于彼此一阶信念的二阶信念，以此类推。例如，交互式 POMDP [Gmytrasiewicz 和 Doshi2005] 会将此类信念显式建模至给定深度，但一般而言，这种技术会随着信念空间（以及智能体数量）的增大而难以扩展。

另一种方法是构建一个（虚构的）公共代理 [Nayyar 等人 2013], ，其信念以所有智能体均可观测的信息为条件。我们还假设，每个玩家的策略均源自一个公共策略 π，该策略仅取决于私有信息 $s_i^{pri} \in \mathcal{S}_i^{pri}$ 和公共信念 $b^{pub} \in \mathcal{B}^{pub}$。更明确地说，公共策略 π 由一系列按时间步定义的动作选择函数 $\pi = \{\pi(t) : \mathcal{S}^{pri} \times \mathcal{B}^{pub} \to \Delta(\mathcal{A})\}_{t=0}^{\infty}$构成。随后，玩家 i的策略 π_i 定义为 $\pi_i(\cdot) := \pi(s_i^{pri}, \cdot)$。由于不同的私有信息 s_i^{pri} 通常会产生不同的策略，这种表述允许在玩家 i每次采取动作后，依据贝叶斯法则更新关于 s_i^{pri} 的公共信念：

$$P(s_i^{pri}|a, b^{pub}) \propto P(a|s_i^{pri}, b^{pub}) \cdot P(s_i^{pri}|b^{pub}) \quad (1)$$

其中， $P(a|s_i^{pri}, b^{pub})$ 由公共策略$\pi(s_i^{pri}, b^{pub})$给出。由于所有玩家都能计算公共代理的信念更新，即 b^{pub} 是共同知识，因此该策略无需对高阶信念进行建模。

最后，给定每位玩家的一个策略，我们可以构成联合策略 $\Pi = \{\pi_1, \ldots, \pi_N\}$。通过根据每位玩家各自的策略选择其动作，该联合策略会产生一个关于轨迹（以及由此产生的奖励）的分布。我们将联合策略 Π 的期望奖励记为$J(\Pi) := \mathbb{E}\big[\sum_{t=0}^{\infty} \gamma^t r(t)\big]$。我们将玩家 i 的信念状态记为 $s_i := (s_i^{pri}, b^{pub})$。联合信念状态 $s := (s_i)_{i \in \mathcal{N}}$和时间 t 下该策略的价值函数为

$$V^{\Pi}(s, t) := \gamma^{-t}\mathbb{E}\big[\sum_{t'=t}^{\infty} \gamma^{t'} r(t') \mid s\big]. \quad (2)$$

在本文中，我们还将考虑一种源于价值函数近似的双重视角。给定任意价值函数 V，相应的 Q 函数为

$$Q(s, a, t) := \mathbb{E}_{r,s'}\big[r + \gamma V(s', t+1) \mid a\big], \quad (3)$$

其中， s' 是执行动作 a后得到的下一个（随机）信念状态。随后可以将新的贪心策略 Π^V 定义为

$$\Pi^V(s, t) := \arg\max_{a \in \mathcal{A}} Q(s, a, t), \quad (4)$$

并在出现并列时随机打破平局。对于根据公式（2）从某个现有策略 Π 推导出的价值函数 V， [Bellman 的策略改进定理保证，该贪心策略 Π^V 获得的奖励优于原策略 Π。

3.2 能力可行策略

接下来，我们描述如何使用能力类型对计算能力各异的玩家进行形式化建模。在我们的设定中，玩家 i 的私有信息 s_i^{pri} 由其能力类型 $c_i \in \mathcal{C}$ 构成。其在公共智能体方法下的策略定义为 $\pi_i(b^{pub}) := \pi(c_i, b^{pub})$，其中 π 是已知的公共策略。由于策略仅取决于玩家的类型，因此只需考虑公共**类型化策略** $(\pi^c)_{c \in \mathcal{C}}$的集合，其中 $\pi^c(\cdot) := \pi(c, \cdot)$。

如果 $c' \geq c$，则称策略 π^c 对于类型为 c'的玩家是**可行的**。玩家只能评估对于其能力类型可行的策略。这可以防止玩家仅通过评估 π^c，就“假装”以更强的类型 c 进行游戏。例如，再次考虑一个配备了能力类型的规划算法，该能力类型限定了前瞻深度。可行策略集合由前瞻深度等于或小于该上限的策略组成。

引入能力类型给公共智能体方法的直接应用带来了若干挑战。首先，在原始表述中，所有玩家都应能够评估所有公共策略，以计算公式（1）中的公共信念更新。然而，在我们的情形中，玩家只能评估可行策略，这一限制使智能体无法完成完整的信念更新（这需要计算所有 $c \in \mathcal{C}$ 对应的后验概率$P(c|a)$）。更根本的问题是，所有玩家都必须能够计算公共信念 b^{pub} ，尤其是最弱玩家。因此，公共智能体方法最终只能采用最弱玩家的推断。为应对这一挑战，下一节将介绍一种受 *Harsanyi*类型结构启发、对公共智能体方法的推广。这种方法允许我们为不同智能体维护信念，同时避免无限信念层级。

4 能力感知的临时团队协作

本节介绍一种通用的多智能体临时团队协作框架，该框架会在每个回合中根据其他玩家的能力进行调整。首先，每位玩家独立维护对所有玩家能力的信念，并在观察到一个动作后更新信念。随后，玩家会根据自己对其他玩家能力的信念调整行动。我们指出了这一设定中的两个主要挑战。首先，由于玩家只能通过基础博弈进行交互，他们无法协调彼此的信念更新；因此，从任何单个玩家的视角来看，如何克服环境中的非平稳性是一项技术挑战，因为其他玩家也在在线调整其行为。其次，玩家通常可能需要考虑无限层级信念：由于其他玩家的信念会影响他们未来的行动，当前玩家的动作也应纳入二阶信念，即对其他玩家信念的信念；但其他玩家也拥有二阶信念，因此还需要维护三阶信念，以此类推。

为应对这些挑战，我们提出了一种基于

能力类型结构的学习框架，这些结构是保留能力排序的信念：

定义 4.1。设 $(\mathcal{C}, \trianglelefteq)$ 为一组能力类型。给定基集合 X，能力类型结构（记作 $\mathcal{B}_X$）是由实值函数 $B^c_X\colon X \to \mathbb{R}$ 组成的集合，这些函数由 $\mathcal{C}$ 索引，并配有约简算子 $\downarrow_c$，使得若 $c' \trianglelefteq c$，则 $B^c_X \downarrow_{c'} = B^{c'}$。在此语境下，我们将每个 B^c_X 称为类型 c 的信念，或一般意义上的类型化信念。

令 $B_X := (B^i_X)_{i\in\mathcal{N}}$ 表示玩家关于 X 的信念集合。给定一组约简算子，当且仅当对于所有玩家 i，$j \in \mathcal{N}$ 满足以下条件之一时，我们可以将信念 B_X 构成能力类型结构 $\mathcal{B}_X$：(1) $c_i = c_j$ 且 $B^i_X = B^j_X$；或 (2) $c_j \trianglelefteq c_i$ 且 $B^i_X \downarrow_{c_j} = B^j_X$（当 $c_i \trianglelefteq c_j$ 时亦同）。在这种情况下，我们可以将玩家 i 的信念 B^i_X 视为类型化信念 $B^{c_i}_X \in \mathcal{B}_X$。

能力类型结构是公共信念的一种自然推广，因为对于类型为 $c' \trianglerighteq c$ 的所有玩家而言，类型化信念 B^c_X 是共同知识。这一关键点使我们能够避免退化到最弱玩家，同时仍允许玩家以保持一致性的方式更新其信念，从而克服其他智能体信念（以及由此产生的行为）的非平稳性。能力类型结构还与 Harsanyi类型结构有关，因为能力类型结构$\mathcal{B}_X$ 生成了任何玩家针对某一感兴趣属性所持有的所有高阶信念 X。具体而言，如果 B^c_X 是类型为 c 的玩家 i 的信念，那么 i 对“一个类型为 c' 的玩家 j 对 X 的信念”所持有的二阶信念由 $B^c_X \downarrow_{c'}$ 给出；根据构造，这等价于 j 的一阶信念 $B^{c'}_X$；由此可知，所有高阶信念也都等价于 $B^{c'}_X$。相比之下，Harsanyi类型结构通常难以显式构造。

4.1 使用能力类型结构进行推断

本节将介绍我们的能力自适应多智能体系统框架，该框架利用能力类型结构。我们假设玩家可以访问一组价值函数 $(V^c)_{c\in\mathcal{C}}$，其中 $V^c(s,\ B^c)$ 的输入为[0] 当前（完全可观测的）博弈状态 s，以及一个类型为 c 的信念 B^c，并输出价值估计。类似地，我们根据公式（3）定义一组相应的 Q 函数 $(Q^c)_{c\in\mathcal{C}}$。随后，玩家查询这些 Q 函数，以执行公式（4）中的贪心策略所给出的相应类型化策略π^c。

我们首先描述初始设置。每位玩家都会为每位玩家（包括自己）维护一个信念，总计 N 个信念；这些信念是以基集合 $\mathcal{C}$ 为索引的实数向量。我们将玩家 i 关于玩家 j 的信念记为 B^i_j，并将玩家 i 的汇总信念记为 $B^i := (B^i_j)_{j\in\mathcal{N}}$。每个信念 B^i_j 都包含关于玩家 j 的能力类型的未归一化似然，并以玩家 j 的类型低于玩家 i 的类型为条件。为初始化信念，类型为 c 的玩家在 $c' \trianglelefteq c$ 时将 $B[c'] = 1$ 设为相应值，否则设为 0，对所有 $B \in B^i$ 均如此。此外，我们定义一个信念约简算子 $\downarrow_{c'}$，将输入信念向量中对应于能力 $c > c'$ 的所有条目设为零。请注意，信念集合 $\mathcal{B}_j := (B^i_j)_{i\in\mathcal{N}}$ 在初始化时显然构成一个能力类型结构，其中信念所基于的基集合 $\mathcal{C}$ 表示玩家 j 的类型可能取值。

接下来，我们将描述玩家如何根据其信念调整策略。给定一组信念 B，我们将 $B_{[j=c]}$ 表示为同一组信念，但将 B_j（关于玩家 j 类型的信念）替换为 $B_j[c'] = \infty$（如果 $c' \neq c$，否则为 0）（这一干预将信念限定为玩家 j 的真实类型为 c）。玩家 i 选择使其类型 c 的 Q 函数最大化的动作，从而得到如下贪心策略：

$$\pi^i(s, B^i) := \arg\max_{a\in\mathcal{A}} Q^c(s, B^i_{[i=c]}, a) \tag{5}$$

随机打破平局。

最后，我们描述玩家如何更新其信念。每个回合由两个阶段组成。在第一阶段，行动玩家 i 根据策略 π^i 执行动作 a。在第二阶段，所有玩家 $j \in \mathcal{N}$ 更新其关于玩家 i 的信念 B^j_i。将能力 c 下的最优价值记为

$$v^*(c) := \max_{a'\in\mathcal{A}} Q^c(s, B^j_{[i=c]} \downarrow_c, a'), \tag{6}$$

并将实现 $v^*(c)$ 的动作集合记为

$$A^*(c) := \{a' \in \mathcal{A} | Q^c(s, B^j_{[i=c]} \downarrow_c, a') = v^*(c)\}. \tag{7}$$

然后，类型为 c_j 的玩家 j 执行更新

$$B^j_i[c] \leftarrow B^j_i[c] * \mathbb{1}_{a\in A^*(c)}/|A^*(c)| \tag{8}$$

对于所有 $c \trianglelefteq c_j$。注意，更新因子只是在策略 π^i 下的似然 $P(a|c)$。

将时间 t 时所有玩家关于玩家 j 类型的信念集合记为 $B_j(t)$。我们的关键结果如下：

定理 4.2（信念更新的一致性）。如果 $B_j(t)$ 是一个能力类型结构，那么 $B_j(t+1)$ 也是一个能力类型结构。

我们在附录中给出了完整的证明。作为推论，由于先验信念 $B_j(0)$ 初始化为能力类型结构，因此对于所有 $t \geq 0$，$B_j(t)$ 也可以归入一个类型结构。

我们的下一个结果表明，这些信念是准确的。我们首先需要明确这些信念所跟踪的内容。对于一个信念 B_j

[0]为清晰起见，我们通常会省略对 t 的依赖关系，尤其是价值函数对其的依赖。

类型为 c时，我们可以定义函数

$$P^c(c') := \frac{B_j[c']}{\sum_{c'' \le c} B_j[c'']}, \forall c' \in \mathcal{C} \tag{9}$$

只要分母非零，以下定理就表明， $P^c(c')$ 实际上就是由能够访问所有 Q 函数 $(Q^c)_{c\in\mathcal{C}}$、执行精确 贝叶斯更新的“全能”公共代理计算出的真实（条件）可能性 $P(c_j = c' | c_j \le c)$。

定理 4.3 (信念更新的正确性)**.** 设 j 为一名玩家，其类型记为 c_j。设 B_j 为类型c 在时间 t的信念， $H(t)$ 为截至时间 t的完整动作-观测历史，并按方程(9)定义 P^c 。那么，只要P^c 有定义，就有 $P^c(c') = P(c_j = c' | c_j \le c, H(t))$ 对所有 $c' \le c$均成立。此外，若 P^c 未定义，则 $P(c_j \le c | H(t)) = 0$。

请注意，若基础博弈的马尔可夫决策过程（MDP）缺乏额外的结构性假设，则真实似然一般不保证收敛至真实能力上的狄拉克分布。然而，根据定理 4.3，若真实似然确实收敛，则能力类型结构中的信念也会收敛至正确的条件似然。

4.2 噪声环境下的能力感知临时团队协作

接下来，我们转向这样一种情形：玩家受到独立采样的噪声影响。一般而言，噪声的存在会显著增加多智能体系统中信念更新的复杂性，因为噪声导致的信念初始差异哪怕很小，也可能使观察到的行为偏离预期行为，进而放大信念差异，导致信念变得越来越不一致。核心思路是对方程（8）中的更新进行温和化处理，并证明即使存在噪声，所得“似然”仍保持一致。假设似然的小幅差异只会导致 Q 函数的小幅差异，我们证明信念仍可组装成近似能力类型结构，从而得到定理 4.2 的近似版本。所有证明均见附录。

4.2.1 基于能力类型的临时团队协作

现在，我们介绍临时团队协作的通用模型。我们假设，当类型已知时，游戏便已“解决”，也就是说，存在一个公开的价值函数集合（V^C），其涵盖能力 C的所有可能分配。与之前一样，我们可以使用价值函数 V^C 来定义 Q^C 函数以及单步贪心策略 Π^C。如果对于类型为 c 的玩家 i 而言， Q^C 是可行的，则称$c' \le c$ 对于所有 $c' \in C$ 均成立（也就是说，玩家 i 只有在联合策略中的所有玩家的能力不超过V^C时，才能准确展开该策略），并将玩家限制为具有可行 Q^C 函数的策略。

为了与更强的智能体合作，我们在本文中假设，玩家 i 会采用其可用的最强类型，即自身类型 c。当计算（或表示）价值函数所需的计算资源超出能力较弱玩家的可用资源时，这种设定很自然，因此玩家会对能力更强的玩家采用尽力近似。由于较强玩家能够精确模拟较弱玩家使用的近似方法，我们的设定仍允许团队执行对其类型分配 C而言“最优”的联合策略：最强的玩家执行 Π^C，所有较弱玩家则使用尽力近似。对于类型为 c 的玩家，我们将**前驱集合**定义为 $p(c) := \{c' \mid c' \le c\}$，并将**可行分配**集合记为 $n(c) := p(c)^N$。

为了将信念纳入该框架，我们采用 QMDP近似 [Littman 等人， 1995]。固定一个**广义似然函数** ϕ ，该函数接收一组关于玩家类型的类型 c 信念 $B = (B_j^c)_{j\in\mathcal{N}}$，并输出可用分配的分布： $\phi_B(\cdot) \in \Delta(n(c))$。我们将 ϕ-价值函数和 ϕ-Q 函数定义为

$$V_\phi^c(s, B) := \mathbb{E}_{C\sim\phi_B^c}\left[V^C(s)\right], \tag{10}$$

$$Q_\phi^c(s, B, a) := \mathbb{E}_{r,s'}\left[r + \gamma V_\phi^c(s', B) \mid a\right], \tag{11}$$

分别。在我们的情形中， ϕ 价值函数可以解释为：如果在下一步之前揭示所有类型，某个动作的价值估计。为简单起见，我们只考虑一步贪心策略，但通常可以通过增加前瞻深度 [Bellman， 1957]，获得准确度呈指数级提升的估计。

随后，玩家 i 选择使其类型对应的 ϕ-Q函数最大化的动作，从而得到以下贪心策略：

$$\pi_\phi^i(s, B^i) := \arg\max_{a\in\mathcal{A}} Q_\phi^{c_i}(s, B^i_{[i=c_i]}, a). \tag{12}$$

4.2.2 噪声模型

我们的噪声模型允许：(1) 每个玩家对状态 s 的观测受到独立噪声影响；(2) 玩家使用不同的 V^C 函数私有版本私有，这涵盖了许多自然的噪声来源：例如，对状态 s的观测可能受到传感器噪声的影响；允许玩家使用私有的 V^C，则使我们的结果适用于更一般的分散式训练和分散式执行情形，因为独立求解近似价值函数的玩家之间的差异至多为其近似误差之和。

我们唯一的假设是偏差有界。设s_1， s_2 和 V_1^C， V_2^C分别为任意两个玩家的带噪状态和私人价值函数。我们假设

$$|V_1^C(s_1) - V_2^C(s_2)| \le \epsilon \tag{13}$$

对于一个已知常数 $\epsilon \geq 0$。请注意，无噪声设定是 $\epsilon = 0$ 时的特殊情况，在这种情况下，前一节的结果适用。

4.2.3 温度化信念更新

本节为玩家的观测和价值函数受到（有界）噪声影响的设定提供了一个通用框架。从无噪声设定转变而来时，主要问题在于，公式（8）中的硬更新对噪声并不具有鲁棒性，若直接应用，实际上会导致错误的推断。一种方法是在贝叶斯更新中显式纳入噪声，但这并不切实际，因为其精确形式同样会产生无限层级信念；此外，为了执行闭式更新，我们还需要假设价值函数层面的噪声分布是已知的，而这一点在实践中很难保证，尤其是在分散式训练中。

相反，我们提出了一种基于 **温度化信念** 的技术，通过温度参数对信念更新进行平滑处理。与无噪声情形下维护显式似然不同，使用信念来跟踪充分统计量会更加方便。所有信念均初始化为全零向量。在观察到玩家 i 执行动作 a后，类型为 c_j 的玩家 j 执行更新

$$B_i^j[c] \leftarrow B_i^j[c] + \ell_a(c) \tag{14}$$

对于所有 $c \leq c_j$，其中 ℓ 是损失：

$$\ell_a(c) := \max_{a' \in \mathcal{A}} Q_\phi^c(s, B^j_{[j=c]}, a') - Q_\phi^c(s, B^j_{[j=c]}, a). \tag{15}$$

玩家 j 在时间 t 计算广义似然，结果为

$$P_i^j(c) := \frac{\exp(-B_i^j[c]/T(t))}{\sum_{c' < c_j} \exp(-B_i^j[c']/T(t))}, \tag{16}$$

$$\phi_B^c(C) \propto \prod_{i \in \mathcal{N}} P_i^j(C[i]), \tag{17}$$

其中 T： $t \to \mathbb{R}^+$ 是温度，并根据公式（12）执行策略。注意，当 $T \to 0$ 时，如果能力 c 在观察到玩家 i 的动作后曾经历过正损失，则 $P_i^j(c)$ 会收敛到 0；这等价于公式（8）中的硬更新（除归一化因子外）。相反，当 $T \to \infty$ 时，更新会变得无信息，似然会收敛到初始均匀分布。因此，我们选择尽可能小的温度（以提供信息），同时保留平滑噪声的好处。

4.2.4 对抗性噪声

如果噪声具有对抗性，那么对噪声的唯一约束由公式（13）中的 ϵ 界给出。在这种情况下，我们将温度 $T(t) = 6tN$，从而得到以下广义似然：

$$\phi_B^c(C) \propto \prod_{i \in \mathcal{N}} \exp(-B_i^j[c_i]/6tN) \tag{18}$$

定理 4.4（信念更新的一致性，对抗性噪声）。设 B_i^j 和 B_i^k 分别表示玩家 j 和 k 对玩家 i 的类型所持有的类型 c' 信念。那么，对于所有 $c \in \mathcal{C}$ 和时间步 $t \geq 0$，$|B_i^j[c] - B_i^k[c]|/t \leq 3\epsilon$。

4.2.5 随机噪声

对于对抗性噪声，广义似然中会产生一个线性因子 t，实质上相当于取平均损失。我们的下一个结果针对随机噪声给出了 PAC 保证，其中温度中的因子为次线性。具体而言，我们还将假设，对于每次实现，价值函数中的观测误差都是独立同分布采样的。

给定 $\delta > 0$。我们选择温度 $T(t) = \sqrt{dt^{2/3}}$，得到如下广义似然：

$$\phi_B^c(C) \propto \prod_{i \in \mathcal{N}} \exp(-B_i^j[c_i]/\sqrt{d}t^{2/3}) \tag{19}$$

其中，d 是一个依赖于 $1/\delta$ 的对数函数的常数。

定理 4.5 (信念更新的一致性，随机噪声**)**。令 B_i^j 和 B_i^k 表示玩家 j 和 k 对玩家 i 类型的类型 c' 信念。那么，以至少 $1 - \delta$ 的概率，对于所有 i、j、$k \in \mathcal{N}$、$c \in \mathcal{C}$ 以及时间步 $t \geq 0$，$|B_i^j[c] - B_i^k[c]|/t^{2/3} \leq \sqrt{d\epsilon}/2N$。

5 实验结果

我们展示了若干实验结果，这些实验使用通过深度受限在线树搜索进行协作的玩家。具体而言，能力为 d 的玩家旨在最大化接下来 d 步的联合奖励。每项任务由两名交替采取动作的玩家组成。一名玩家是专家，搜索深度较大；另一名玩家是新手，搜索深度较小。除此之外，两名玩家完全相同（即拥有相同的动作空间、共享相同的奖励函数、使用相同的搜索超参数，并且完全了解环境以及另一名玩家采取的动作）。

为了进行规划，智能体在深度受限树上使用蒙特卡洛树搜索（MCTS）算法。由于叶节点处的价值估计取决于蒙特卡洛估计，因此在所有实验中，我们都使用第 4.2.5 节所述的温度化信念的随机版本。我们的附录更详细地说明了 MCTS策略和信念更新的定义方式。

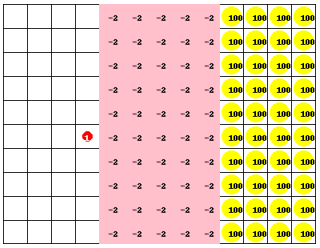

图 1：火墙任务的初始状态。

团队组成	奖励
新手 + 新手	0
专家 + 专家	1490
专家 +新手	-20
CA-专家 + 新手	-2

表 1：不同团队在火墙任务上的表现。结果为 5 次运行的中位数。

5.1 玩具环境

我们首先展示两个玩具环境中的结果，以说明合作环境中能力感知的影响。在**火墙**任务中，两名玩家轮流控制一个化身。化身可以沿四个基本方向之一移动。每次移动后，玩家根据自己的位置获得奖励：红色的“火焰”方块会带来 -2 的惩罚，而黄色的“硬币”方块会带来 +100的奖励。每枚硬币只能收集一次。我们将每个回合设为 20 步（每名玩家 10 步）。我们使用深度为 2 的新手和深度为 20 的专家。图 1 展示了玩家和棋盘的初始状态。表 1 报告了各种团队组合的表现。

专家 + 专家团队采用最优策略，先用 5 步穿越火墙，再用剩余步数收集金币。相反，新手团队无法“看见”火墙另一侧，因此一直待在火墙左侧的中立区域。然而，当无感知专家与新手搭档时，两名玩家在整个回合中轮流进出火墙，导致奖励大幅为负（专家 + 新手，表 1）。相比之下，具备能力感知的专家只需一次交互，就能推断出新手无法看见火墙另一侧；它只承受一次惩罚，之后便与新手合作“探索”棋盘左侧的中立区域（CA-专家 +新手，表 1）。

在**狭窄隧道**任务中，两名玩家分别控制二维棋盘上的一个化身。动作空间包括不移动或向四个基本方向之一移动。红色硬币由红色化身收集时价值为 +30，蓝色硬币由蓝色化身收集时价值为 +1；否则硬币没有价值，并且无论由哪个化身收集，都会消失-

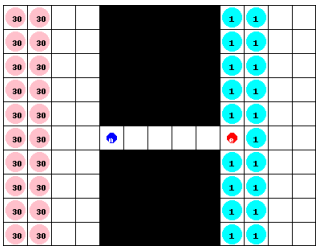

图2：狭窄隧道任务的初始状态。在此情况下，新手控制蓝色化身，专家控制红色化身。

蓝色化身	红色化身	奖励
新手	新手	4
专家	专家	90
新手	专家	0
新手	CA-专家	4

表2：不同团队在狭窄隧道任务中的表现。结果为5次运行的中位数。

无论由哪个化身收集。每个回合运行20步。我们使用深度为10的新手和深度为30的专家。图2展示了玩家和棋盘的初始状态。表2报告了该任务中各种团队组合的表现。

由于奖励分配不均，最优策略是让蓝色化身把狭窄隧道让给红色化身，使其能够收集价值更高的红色硬币（专家 +专家，表2）。相反，新手玩家无法获得红色奖励，因此新手团队改为派蓝色化身穿过狭窄隧道（新手 +新手，表2）。然而，当无感知专家控制红色化身、新手控制蓝色化身时，两名玩家在隧道中央陷入僵局，谁也不愿让步（新手 +专家，表2）。使用能力感知算法的专家只需经历1回合僵局，便能推断出新手的深度不足以采用最优策略，因此将隧道让给蓝色玩家（新手 +CA-专家，表2）。

5.2 合作跳棋

接下来，我们报告合作版跳棋的实验结果。在这种玩法中，同一团队的玩家轮流移动棋子。由于底层游戏本身复杂，合作跳棋为研究临时团队协作带来了颇具挑战性的情境。例如，Schaeffer [2007] 报告称，一盘跳棋平均持续约 50 步，平均分支因子约为 6；因此，其博弈树复杂度约为 10^{40}（相比之下，国际象棋为 10^{83}）。

每个多智能体团队由一名专家与

Δ	#运行次数	得分（%）
2	480	1.0
4	320	6.4
6	160	4.4
总计	960	3.3

表3：CA团队对阵OBL团队的表现，得分以CA团队的视角计算（正得分表示CA强于OBL）。 Δ 表示专家与新手之间的能力差异。

一名新手组成，其能力涵盖集合 $\{2, 4, 6, 8\}$ 中所有不同的配对。轮到团队移动时，玩家交替选择动作。如果某个团队无棋可走或没有剩余棋子，游戏便结束，对手获胜。如果总移动次数达到 120 次，或连续 40 次移动没有任何奖励，我们也会终止游戏，并宣布累积奖励最高的团队获胜。玩家在规划时并不知道终止条件，以防止拖延。附录中包含有关设置的更多细节以及其他实验结果。

5.2.1 能力感知合作与无感知合作

在第一组实验中，我们让两支团队相互对战：能力感知（CA）团队和无感知（OBL）团队。两支团队均由一名专家和一名新手组成，且两支团队中的专家和新手具有相同的深度。CA团队中的专家采用能力感知策略，而OBL团队中的专家采用无感知策略；两名新手均采用无感知策略。对于每种能力组合，我们针对团队间的颜色分配以及各团队内部的玩家顺序，使用不同的随机种子进行20场游戏，总计进行160场游戏。我们报告汇总后的**得分**，其定义为（#胜场 - #损失）/#游戏场数；得分越高，表示表现越强。

表3展示了结果。能力感知策略团队始终优于 OBL 团队，而且能力差异 Δ 越大，优势往往也越大。这与预期一致：随着 Δ 增加，无感知专家更频繁地误判新手的策略。

5.2.2 多智能体能力感知协作

我们还比较了采用多智能体能力推断算法（MA）的团队与由两名采用单智能体推断算法（SA）的自适应玩家组成的团队的表现。在前一种情况下，每位玩家都会维护一个能力类型结构，用于建模队友的信念并确保信念更新一致。在后一种情况下，每位玩家都会尝试推断其队友，但将队友建模为无感知地进行游戏。虽然这种做法忽略了队友行为的非平稳性，从而简化了更新规则，但我们的主要假设是，在信念收敛到真实能力之前，这种近似会导致更差的表现。

对手	Team	#运行次数	得分（%）	d_{exp}	d_{nov}
新手	MA	600	27.0	1.1	0.8
	SA	600	21.0	1.2	0.9
专家	MA	600	-19.7	1.2	0.8
	SA	600	-21.8	1.3	0.9
总计	MA	1200	3.7	1.1	0.8
	SA	1200	-0.4	1.2	0.9

表 4：多智能体（MA）和单智能体（SA）自适应团队对阵深度为新手或专家的对手时的表现。得分以团队的视角计算（得分为正表示该团队强于单个对手）。d_{nov} 和 d_{exp} 分别表示新手和专家最终后验分布的偏差。

团队与一名玩家对手进行游戏，该对手的能力等于团队中的新手或专家。对于每种团队颜色分配和团队内玩家顺序的组合，我们使用不同的随机种子进行了 50 局游戏，因此每种能力与对手的组合共进行了 200 局游戏。除得分外，我们还报告了最终后验分布 p 相对于队友真实深度d^* 的**偏差** d，其定义为 $\sqrt{\sum_d p(d)(d - d^*)^2}$。对于新手，我们将 d^* 设为尽力近似（即新手自身的深度）。将 100% 密度置于 d^* 上的“真实”后验分布，其最优偏差为 0。

表 4 展示了结果。采用正确多智能体更新的多智能体团队始终优于采用单智能体更新的单智能体团队，尤其是在与新手对战时更为明显。我们还注意到，即使采用单智能体更新，最终后验分布也相当准确。因此，我们将表现差异归因于游戏早期后验分布的分歧：即使游戏已经输掉，后验分布也可能逐渐变得准确。这些结果表明，基于能力类型结构的多智能体框架能够使后验分布快速收敛到真实值，这对于实现良好表现至关重要，相比之下，朴素的单智能体更新则无法做到这一点。

6 结论

本研究探讨了计算能力不同的智能体之间的临时团队协作。我们的主要贡献是提出了一种基于能力类型结构的框架，使多个计算能力不同的智能体能够在无需明确协调的情况下调整行为，

以适应其能力。 此外，通过使用温度化信念，当智能体的观测和价值估计受到噪声影响时，我们的方法也能适用。实验表明，在若干简单场景以及更复杂的跳棋环境中，我们的能力感知算法能够提升能力各异的团队的表现。

参考文献

Stefano V Albrecht、Jacob W Crandall 和 Subramanian Ramamoorthy。假设行为中的信念与真相。*ArtificialIntelligence*，235：63–94，2016。

Peter Auer、Nicolo Cesa-Bianchi 和 Paul Fischer。多臂老虎机问题的有限时间分析。*Machinelearning*，47(2)：235–256，2002。

Richard Bellman。*DynamicProgramming*。美国新泽西州普林斯顿，普林斯顿大学出版社，第 1 版，1957 年。URL http://books.google. com/books?id=fyVtp3EMxasC&pg=PR5& dq=dynamic+programming+richard+e+bell man&client=firefox-a#v=onepage&q=dynamic% 20programming%20richard%20e% 20bellman&f=false。

Cameron B Browne、Edward Powley、Daniel Whitehouse、Simon M Lucas、Peter I Cowling、Philipp Rohlfshagen、Stephen Tavener、Diego Perez、Spyridon Samothrakis 和 Simon Colton。蒙特卡罗树搜索方法综述。*IEEETransactionson ComputationalIntelli-genceand AI ingames*，4(1)：1–43，2012。

Vishnu R Desaraju和Jonathan P How。使用快速扩展随机树在复杂环境中进行多智能体团队的分散式路径规划。载于*2011 IEEE*机器人与自动化国际会议，第4956–4961页。IEEE，2011年。

Jakob Foerster、Gregory Farquhar、Triantafyllos Afouras、Nantas Nardelli和Shimon Whiteson。反事实多智能体策略梯度。载于*AAAI*人工智能会议论文集，第32卷，2018年。

Jakob Foerster、Francis Song、Edward Hughes、Neil Burch、Iain Dunning、Shimon Whiteson、Matthew Botvinick和Michael Bowling。用于深度多智能体强化学习的贝叶斯动作解码器。载于机器学习国际会议，第1942–1951页。PMLR，2019年。

Jakob N Foerster、Richard Y Chen、Maruan Al-Shedivat、Shi- mon Whiteson、Pieter Abbeel 和 Igor Mordatch。具有对手学习感知的学习。*arXiv* 预印本 *arXiv:1709.04326*，2017 年。

Filippos Fotiadis 和 Kyriakos G. Vamvoudakis。多智能体非均衡博弈学习系统中有限理性的递归推理。载于 *2021 IEEE* 控制技术与应用会议（*CCTA*），第 741–746 页，2021 年。doi：10.1109/CCTA48906.2021. 9658965。Piotr J Gmytrasiewicz 和 Prashant Doshi。多智能体场景中序贯规划的框架。人工智能研究杂志，24：49–79，2005 年。John C Harsanyi。不完全信息下由“贝叶斯”玩家进行的游戏，I–III 部分 I：基本模型。管理科学，14(3)：159–182，1967 年。

Leslie Pack Kaelbling、Michael L Littman和Anthony R Cassandra。在部分可观测随机领域中的规划与行动。人工智能，101(1-2)：99–134，1998年。

Levente Kocsis 和 Csaba Szepesvári。基于赌博机的蒙特卡洛规划。载于欧洲机器学习会议，第 282–293 页。Springer，2006 年。

Michael L Littman。将马尔可夫博弈作为多智能体强化学习的框架。载于机器学习论文集 *1994*，第 157–163 页。Elsevier，1994 年。

Michael L Littman、Anthony R Cassandra 和 Leslie Pack Kaelbling。部分可观测环境中的策略学习：扩展规模。载于机器学习论文集 *1995*，第 362–370 页。Elsevier，1995 年。

Ryan Lowe、Yi I Wu、Aviv Tamar、Jean Harb、OpenAI 的 Pieter Abbeel 和 Igor Mordatch。混合合作竞争环境中的多智能体行动者-评论家。神经信息处理系统进展，30，2017 年。

Reuth Mirsky、Ignacio Carlucho、Arrasy Rahman、Elliot Fosong、William Macke、Mohan Sridharan、Peter Stone 和 Stefano V Albrecht。临时团队协作研究综述。载于多智能体系统：第 *19* 届欧洲会议，*EUMAS 2022*，杜塞尔多夫，德国，*2022* 年 *9* 月 *14–16* 日，论文集，第 275–293 页。Springer，2022 年。

Ashutosh Nayyar、Aditya Mahajan 和 Demosthenis Teneketzis。具有部分历史共享的去中心化随机控制：一种公共信息方法。*IEEE* 自动控制学报，58(7)：1644–1658，2013 年。

Stefanos Nikolaidis、Anton Kuznetsov、David Hsu 和 Siddhartha Srinivasa。《人机相互适应形式化：有界记忆模型》。载于 *2016*年第*11*届*ACM/IEEE*人机交互国际会议（*HRI*），第75–82页。IEEE，2016年。

Yi Ouyang、Hamidreza Tavafoghi 和 Demosthenis Teneketzis。《不对称信息下的动态游戏：基于共同信息的完美贝叶斯均衡与序贯分解》。*IEEE*自动控制汇刊，62(1)：222–237，2016年。

Stuart Jonathan Russell 和 Eric Wefald。做正确的事：有限理性研究。麻省理工学院出版社，1991 年。

Jonathan Schaeffer。《游戏结束：跳棋中黑方走棋并和棋》。*ICGAJournal*，30(4):187–197，2007 年。

David Silver、Aja Huang、Chris J Maddison、Arthur Guez、Laurent Sifre、George Van Den Driessche、Julian Schrittwieser、Ioannis Antonoglou、VedaPanneershelvam、Marc Lanctot 等。通过深度神经网络和树搜索掌握围棋。《自然》，529(7587):484–489，2016 年。

Herbert Alexander Simon。有限理性模型：以实证为基础的经济理性，第 3 卷。麻省理工学院出版社，1997 年。

Richard S Sutton 和 Andrew G Barto。强化学习：导论。麻省理工学院出版社，2018 年。

Alejandro Torreno、Eva Onaindia、Antonín Komenda 和 Michal Štolba。协作多智能体规划：一项综述。*ACM Computing Surveys*（*CSUR*），50(6):1–32，2017 年。

Deepanshu Vasal 和 Achilleas Anastasopoulos。研究具有策略性智能体的去中心化贝叶斯学习的框架。*StochasticSystems*，12(3):274–292，2022 年。

Ying Wen、Yaodong Yang、Rui Luo 和 Jun Wang。通过广义递归推理对多智能体交互中的有限理性进行建模。*arXiv* 预印本 *arXiv:1901.09216*，2019 年。

延后证明

定理 *4.2* 的证明。设 i 设其为当前行动者，并令 j 和 k 分别为具有能力类型 c_j、c_k 的任意两个玩家 . 固定一个类型 $c \le c_j$ 和 $c \le c_k$。首先，我们证明玩家 j 和 k 为 $v^*(c)$ 计算出相同的价值，该价值是在能力 c 处的最优价值。记 $v^j(c)$ 和 $v^k(c)$ 为玩家 j 和 k分别计算出的能力 c 处的最优价值。根据假设，对于所有 $j \in \mathcal{N}$， $B_j(t)$ 都是一个能力类型结构，因此（稍微滥用记号）有$B^j_{[i=c]} \downarrow_c = B^k_{[i=c]} \downarrow_c$。由于 Q^c 和 s 是公共知识，

$$v^j(c) = \max_{a' \in \mathcal{A}} Q^c(s, B^j_{[j=c]} \downarrow_c, a') \tag{20}$$
$$= \max_{a' \in \mathcal{A}} Q^c(s, B^k_{[j=c]} \downarrow_c, a') \tag{21}$$
$$= v^k(c). \tag{22}$$

类似地，设 $A^j(c)$ 和 $A^k(c)$ 分别为玩家 j 和 k计算出的最优行动 $A^*(c)$。于是

$$A^j(c) = \{a \in \mathcal{A} | Q^c(s, B^j_{[j=c]} \downarrow_c, a) = v^j(c)\} \tag{23}$$
$$= \{a \in \mathcal{A} | Q^c(s, B^k_{[j=c]} \downarrow_c, a) = v^k(c)\} \tag{24}$$
$$= A^k(c)。 \tag{25}$$

再次利用 B_j 构成能力类型结构这一事实，$B^j_i[c] = B^k_i[c]$，并且

$$B^j_i[c] * \mathbb{1}_{a \in A^j(c)} / |A^j(c)| = B^k_i[c] * \mathbb{1}_{a \in A^k(c)} / |A^k(c)|. \tag{26}$$

由于这对所有满足 $c \le c_i$ 和 $c \le c_j$的 i， $j \in \mathcal{N}$和 $c \in \mathcal{N}$都成立，因此我们得出结论： $B_j(t+1)$ 是一个能力类型结构。

□

定理 *4.3* 的证明。 该结果直接源于这样一个事实：信念更新项 $\mathbb{1}_{a \in A^*(c')} / |A^*(c')|$ 恰好就是似然 $P(a|c')$。更具体地说，设 $a(t_1)$ $a(t_2) \ldots, a(t_T)$ 是玩家 j截至目前所采取的行动序列。根据贝叶斯定理，

$$B^j_i[c'] = \frac{\mathbb{1}_{a(t_1) \in A^*_1(c')}}{|A^*_1(c')|} \frac{\mathbb{1}_{a(t_2) \in A^*_2(c')}}{|A^*_2(c')|} \cdots \frac{\mathbb{1}_{a(t_T) \in A^*_T(c')}}{|A^*_T(c')|} \tag{27}$$
$$= P(a(t_1)|c_j = c') * P(a(t_2)|c_j = c') * \cdots * P(a(t_T)|c_j = c') \tag{28}$$
$$\propto P(c_j = c' | a(t_1) \ a(t_2) \ \ldots, a(t_T)) . \tag{29}$$

其中， A^*_n 是玩家 i 在能力 c 下采取动作 $a(t_n)$ 时的最优行动集合；此处省略了其对状态观测历史的依赖。

接下来，我们考虑 $P^c(c')$ 的分母。具体而言，若 $\sum_{c'' \le c} B_j[c''] = 0$，则 P^c 未定义，但这也意味着对于所有 $c'' \le c$， $P(c_j = c'' | a(t_1) \ a(t_2) \ldots, a(t_T)) = 0$，即 $c_j > c$，正如所述。另一方面，若 $\sum_{c'' \le c} B_j[c''] > 0$，则这会计算出 $P(\cdot | c_j \le c)$ 的归一化因子；因此 $P^c(c') = P(c_j = c' | c_j \le c, \ a(t_1) \ a(t_2) \ldots, a(t_T))$，由此完成证明。

□

在证明近似一致性结果之前，我们先陈述并证明一个有用的引理：

引理 A.1. 令 i 为正在行动的玩家，将其动作记为 a，并令 B^j 和 B^k 分别表示类型至少为 c的任意两个玩家 j和 k 的类型 c 信念。将由 j和 k计算出的损失分别记为 $\ell^j_a(c)$ 和 $\ell^k_a(c)$。若对所有玩家 m 和所有能力 $c' \le c$，均有 $|B^j_m[c'] - B^k_m[c']| / T(t) \le \epsilon/2N$，则 $|\ell^j_a(c) - \ell^k_a(c)| \le 3\epsilon$。

证明。 首先，我们证明对于任意 $m \in \mathcal{N}$， P_m^j 和 P_m^k 之间的全变差距离很小。固定 $c' \leq c$，并且不失一般性地假设 $P_m^k(c') > P_m^j(c')$。于是

$$P_m^k(c') - P_m^j(c') = \frac{\exp(-B_m^j[c']/T(t)}{\sum_{c'} \exp(-B_m^j[c']/T(t))} - \frac{\exp(-B_m^k[c']/T(t))}{\sum_{c'} \exp(-B_m^k[c']/T(t))} \tag{30}$$

$$\leq \frac{\exp(-B_m^j[c']/T(t))}{\sum_{c'} \exp(-B_m^j[c']/T(t))} - \frac{\exp((-B_m^j[c'] - \epsilon/2T(t)N)/T(t))}{\sum_{c''} \exp((-B_m^j[c''] + \epsilon/2T(t)N)/T(t))} \tag{31}$$

$$= (1 - e^{-\epsilon/N}) \frac{\exp(-B_m^j[c']/T(t))}{\sum_{c'} \exp(-B_m^j[c']/T(t))} \tag{32}$$

$$= (1 - e^{-\epsilon/N}) P_m^k(c') \tag{33}$$

$$\approx \epsilon P_m^k(c')/N \tag{34}$$

由于这对所有 $c' \leq c$都成立，因此 P_m^j 与 P_m^k 之间的全变差距离至多为

$$d_{TV}(P_m^j, P_m^k) = \sum_{c' \leq c} |P_m^j(c') - P_m^k(c')|/2 \tag{35}$$

$$\leq \sum_{c' \leq c} \epsilon P_m^k(c')/2N \tag{36}$$

$$= \epsilon/2N \tag{37}$$

分别用 $\phi_{B^j}^c(C)$ 和 $\phi_{B^k}^c(C)$ 表示玩家 j和 k计算得到的广义似然。由于每个似然分别是 N个分布 $\prod_{m \in \mathcal{N}} P_m^j$ 和 $\prod_{m \in \mathcal{N}} P_m^k$的乘积，且它们的全变差距离均不超过$\epsilon/2N$，因此我们有

$$d_{TV}(\phi_{B^j}^c(C), \phi_{B^k}^c(C)) \leq \epsilon/2 \tag{38}$$

令 $V_{\phi,\ j}^c(s,\ B)$ 和 $V_{\phi,\ k}^c(s,\ B)$ 分别为玩家 j和 k对应的结果价值函数。于是

$$|V_{\phi,j}^c(s,B) - V_{\phi,k}^c(s,B)| = |\mathbb{E}_{C \sim \phi_{B_j}^c}\left[V_j^C(s_j)\right] - \mathbb{E}_{C \sim \phi_{B_k}^c}\left[V_k^C(s_k)\right]| \tag{39}$$

$$\leq \epsilon/2 + |\mathbb{E}_{C \sim \phi_{B_k}^c}\left[V_j^C(s_j)\right] - \mathbb{E}_{C \sim \phi_{B_k}^c}\left[V_k^C(s_k)\right]| \tag{40}$$

$$= \epsilon/2 + \mathbb{E}_{C \sim \phi_{B_k}^c}\left[|V_j^C(s_j) - V_k^C(s_k)|\right] \tag{41}$$

$$\leq 3\epsilon/2 \tag{42}$$

It f由此可知， ϕ-Q函数的差异也至多为 $3\epsilon/2$，因此 $|\ell_a^j(c) - \ell_a^k(c)| \leq 3\epsilon$.

□

定理 *4.4* 的证明。 令 B_i^j 和 B_i^k 为任意两个玩家 j 和 k 的类型 c' 信念。我们将通过归纳法证明，对于所有 $c \leq c'$，$|B_i^j[c] - B_i^k[c]|/t \leq 3\epsilon$，其中 t 是更新次数， ϵ 是私人价值函数之间的误差界。

由于信念初始化为相同的值，基例 $t = 0$ 显然成立。对于归纳步骤，根据假设，$|B_i^j[c] - B_i^k[c]|/t \leq 3\epsilon$，即 $|B_i^j[c] - B_i^k[c]|/T(t) \leq \epsilon/2N$。直接应用该引理可知，它们对损失的估计最多相差 3ϵ。由于信念累积了损失，归纳证明完成。

□

定理 *4.5* 的证明。 设 B_i^j 和 B_i^k分别是任意两个玩家 j 和 k的类型 c' 信念。注意， $B_i^j[c]$ 和 $B_i^k[c]$ 是随机变量，并且由于噪声独立性的假设，对所有 j 和 k，都有 $\mathbb{E}[B_i^j[c]] = \mathbb{E}[B_i^k[c]]$。将该期望记为 $B_i[c]$。我们将证明，以至少 $1 - \delta$ 的概率，

$$|B_i^j[c] - B_i[c]|/\sqrt{d}t^{2/3} \leq \epsilon/4N \tag{43}$$

对所有 i、 $j \in \mathcal{N}$、 $c \in \mathcal{C}$ 以及时间步 t，均有此结论。因此，以至少 $1-\delta$的概率，对所有 i、 j、 $k \in \mathcal{N}$、 $c \in \mathcal{C}$ 以及时间步 t，都有 $|B_i^j[c] - B_i^k[c]|/\sqrt{dt^{2/3}} \leq \epsilon/2N$，由此完成证明。

我们将对 t 进行归纳。由于所有信念均初始化为零，基例 $t = 0$ 成立。因此假设$|B_i^j[c] - B_i^k[c]|/\sqrt{dt^{2/3}} \leq \epsilon/2N$，或等价地， $|B_i^j[c] - B_i^k[c]|/T(t) \leq \epsilon/2N$。根据该引理，损失与真实损失之差至多为 3ϵ。由 Hoeffding 不等式

$$P(|B_i^j[c] - B_i[c]| \geq \sqrt{d}\epsilon t^{2/3}/4N) \leq 2\exp\big(-\frac{2d\epsilon^2 t^{4/3}}{(4N)^2 t(3\epsilon)^2}\big) \tag{44}$$

$$= 2\exp\big(-d' t^{1/3}\big) \tag{45}$$

其中， $d' = d/72N^2$和 t是总和 $B_i^j[c]$中的损失估计数量。共有 N^2个信念 B_i^j（每个 $i, j \in \mathcal{N}$对应一个），且每个信念的支撑集至多为 $c_{max} := \max_{i \in \mathcal{N}} |p(c_i)|$。因此，为应用并集界，我们需要

$$\sum_{n=1}^{\infty} 2\exp(-d' t^{1/3}) \leq \delta/N^2 c_{max}. \tag{46}$$

由于相应的积分收敛，我们发现取 $d' = \ln(20N^2 c_{max}/9\delta)$即可。根据并集界，我们得出：以至少 $1-\delta$ 的概率，公式 (43) 对所有智能体 $i, j \in \mathcal{N}$、所有 $c \in \mathcal{C}$ 以及所有时间步 t均成立。

□

B 能力感知型蒙特卡洛树搜索

近年来，蒙特卡洛树搜索（MCTS）算法因成为 AlphaGo [Silver 等人， 2016] 的重要组成部分，在游戏和规划领域迅速流行。本节将简要介绍蒙特卡洛树搜索算法。接下来，我们将介绍一种经过修改的蒙特卡洛树搜索，使其能够处理在可探索的游戏树深度方面受到限制的智能体。随后，我们提出一种能力感知蒙特卡洛树搜索算法，用于根据队友的能力进行调整。最后，我们将介绍如何在能力感知蒙特卡洛树搜索算法中进行推断。

B.1 蒙特卡洛树搜索

本节将简要介绍蒙特卡洛树搜索算法。有关更全面的综述，请参阅读者可参考的 [Browne 等人， 2012]。

算法 1 概述了蒙特卡洛树搜索算法的主要步骤。搜索从一棵仅包含单个根节点 $\mathcal{T} = s(0)$ 的树开始，该根节点包含当前状态。在每一步中，我们**选择** 一条通往叶节点 s 的路径，向叶节点添加所有合法转移作为子节点以**扩展** 该叶节点，**模拟** s 的初始值估计，然后**反向传播** s 的估计值至根节点，并使用递归公式更新其前驱节点的值。具体而言，**模拟** 会将随机动作向下执行m 次，直到最大深度 d （或游戏终止），并将累计奖励的平均值作为初始值估计。**选择** 需要在探索与利用之间进行权衡（探索倾向于选择访问次数较少的节点；利用倾向于选择估计值较高的节点）；流行的上置信树（UCT）变体 [Kocsis 和 Szepesvári， 2006] 将每次选择视为一个多臂老虎机问题，并使用广为人知的 UCB1 方法 [Auer 等人， 2002]从而实现良好的理论保证（以及在实践中出色的表现）。

算法 1蒙特卡洛树搜索（MCTS）

过程 MCTS($\mathcal{T}$， n， d， m)
    对于 $i = 1, ..., n$ **do**
        $s \leftarrow$ **选择**($\mathcal{T}$)
        扩展(s)
        模拟(s， d， m)
        反向传播($\mathcal{T}$， s)
    结束循环
结束过程

B.2 深度受限蒙特卡洛树搜索

为了针对搜索深度受限 d 的智能体调整核心蒙特卡洛树搜索算法，我们增加了一项限制：**选择** 在到达叶节点或路径达到最大深度 d 时终止搜索。由于同一参数 d 也适用于模拟深度，深度为 d 的智能体能够显式存储一棵深度为 d 的树，并通过蒙特卡洛模拟访问接下来的 d 层。更新后的算法见算法 2；请注意，与算法 1 相比，唯一的区别是**选择**中的参数 d 。

算法 2深度受限蒙特卡洛树搜索

过程 BOUNDEDMCTS($\mathcal{T}$， n， d， m)
  对于 $i = 1, ..., n$ **do**
    $s \leftarrow$ **选择**($\mathcal{T}$， d)
    扩展(s)
    模拟(s， d， m)
    反向传播($\mathcal{T}$， s)
  结束循环
结束过程

最后，我们假设智能体逐步搜索树的各个深度。这一附加结构简化了推断，并且根据经验，在我们的设置中不会显著影响智能体的能力。算法 3 展示了具有能力类型 d 的无知智能体用于执行搜索的算法。请注意，智能体在每一层搜索时所花费的时间也会逐渐增加。

算法 3深度受限蒙特卡洛树搜索智能体（无知）

过程 OBLIVIOUSSEARCH($\mathcal{T}, n, d, m$)
  对每个 $i = 1, ..., d$ **do**
    BOUNDEDMCTS($\mathcal{T}, n * i, i, m$)
  结束循环
结束过程

B.3 能力感知型蒙特卡洛树搜索

通常，我们可以将游戏树中的节点分为两类：一类代表玩家所属团队采取的动作，另一类代表对手采取的动作。处理对手节点的标准方法是执行极大极小搜索，假定对手会选择对玩家而言价值最差的动作。相反，对于由玩家所属团队采取行动的节点，朴素策略（无认知玩家所采用的策略）是假设队友会选择最佳动作。然而，如果该节点实际上由能力不同的队友行动，他们可能不同意你对应该采取哪个动作的判断，从而导致反向传播的数值不正确。

为了解决这个问题，我们提出修改蒙特卡洛树搜索算法，使其为每种能力类型分别维护一个数值（和访问次数），而不是像标准算法那样只使用一组统计数据。具体而言，每当玩家将搜索深度推进到下一层时，也会转而存储与该搜索深度对应类型的数值。对于深度 > 1，这些数值初始化为上一深度的数值。

为了利用类型化的值，在 BOUNDEDMCTS 每次迭代开始时，我们根据信念为队友采样一组深度。在选择过程中，我们继续使用自身节点和对手节点的最佳（即最深）值估计；但对于队友节点，我们使用与采样类型相对应的值。此外，我们还会递归运行另一个 BOUNDEDMCTS 实例，采用对队友深度的假设，以模拟队友进行探索。为了限制递归模拟带来的开销，我们只允许初始调用生成额外的模拟。因此，修改后的蒙特卡洛树搜索算法引入的开销与原智能体的深度呈线性关系。

算法 4 展示了核心蒙特卡洛树搜索算法的能力感知版本，该版本额外接收输入 B，表示智能体的信念。新的子程序 **SampleDepths** 为智能体的队友采样一组深度

基于这些信念，**Select** 还将采样深度 D 作为输入，而 **Backpropogate** 使用当前最大深度 d 来存储适当类型的值。

算法 4能力感知型蒙特卡洛树搜索

过程 能力感知蒙特卡洛树搜索($\mathcal{T}$，n，d，m，B)
    对 $i = 1, ..., n$ **do**
        $D \leftarrow$**SampleDepths**$(B \downarrow_d)$
        $s \leftarrow$ **选择**（$\mathcal{T}$，d，D）
        扩展（s）
        模拟(s, d, m)
        反向传播$(\mathcal{T}, s, d)$
    结束循环
结束过程

最后，我们介绍如何实现推断。每次队友采取动作之前，玩家首先从当前状态运行一个全新的能力感知蒙特卡洛树搜索实例。由于搜索具有渐进性，在深度为 d 时暂停搜索，可以得到深度为 d 的智能体行为估计。因此，玩家只需读取每种能力类型对应的 Q 值，因为这些值已经存储在每个节点中。随后，在观察到队友执行动作 a 后，玩家可以计算损失 $\ell_a(d)$，以更新其信念。请注意，标准的 UCT 动作选择会选择访问次数最多的动作，因此在实践中，我们使用访问次数替代 Q 函数中的平均值。

C 附加实验结果与详细信息

在本节中，我们提供了跳棋环境的补充实验结果和详细信息。

C.1 超参数

对于蒙特卡洛树搜索，所有智能体均使用 $n = 200$ 作为搜索参数（因此，深度为 d 的智能体总共搜索 $100d$（$d + 1$）次迭代），并使用 $m = 5$ 作为模拟次数。规划时，我们采用 0.9 的折扣率，但确定获胜者时使用未折扣的累积奖励。对于信念更新，我们将温度设为 T（t）$= .1$，并将损失裁剪至最高 0.5。能力感知智能体的所有信念均初始化为均匀分布（除非另有说明）。

C.2 消融研究

我们使用以下基线方法对能力感知蒙特卡洛树搜索算法进行了额外评估：

1. 能力感知（**能力感知策略**）：运行能力感知蒙特卡洛树搜索，信念初始化为均匀分布，并执行信念更新。2. 预言机（**ORA**）：运行能力感知蒙特卡洛树搜索，信念初始化为真实信念，且不执行任何信念更新。3. 无感知（**OBL**）：运行标准蒙特卡洛树搜索。4. 无更新（**NU**）：运行能力感知蒙特卡洛树搜索，信念初始化为均匀分布，且不执行任何信念更新。5. 最小化者（**MIN**）：运行标准蒙特卡洛树搜索，但将队友视为对手（即价值最小化者）。

每个团队由两名玩家组成：一名专家使用上述 5 种策略之一，另一名是无感知新手。其余设置与第 5.2.2 节相同：从 $\{2, 4, 6, 8\}$ 中（不放回地）抽取新手和专家的深度，他们与一名单独的玩家对战，该玩家的深度等于新手或专家的深度。

表5报告了结果。总体而言，预言机表现最佳，总得分为 5.4，这符合预期。该结果证实，我们的能力感知蒙特卡洛树搜索算法能够根据队友的深度正确调整其行为。

我们的能力感知推断智能体以总得分 2.4 位居第二，表现与预言机对阵新手智能体时一样出色。然而，该团队对阵专家时表现不佳，我们将其归因于此前的误判

对手	Team	#运行次数	得分（%）
新手	CA	600	**26.0**
	ORA	600	25.3
	OBL	600	12.8
	NU	600	19.0
	MIN	600	15.2
专家	CA	600	-21.2
	ORA	600	**-14.5**
	OBL	600	-17.7
	NU	600	-18.3
	MIN	600	-17.8
总计	CA	1200	2.4
	ORA	1200	**5.4**
	OBL	1200	-2.4
	NU	1200	0.3
	MIN	1200	-1.3

表5：预言机（**ORA**）、能力感知（**CA**）、无感知（**OBL**）、无更新（**NU**）和最小团队（**MIN**）策略专家与无感知新手配对时的表现，对手的深度为新手或专家。得分从团队的角度计算（正得分表示团队强于单个对手）。

信念收敛。请注意，每场游戏都是具有挑战性的“一次性”设置，信念不会在不同游戏之间延续。在重复博弈环境中，我们预计信念会收敛，预言机团队与能力感知团队之间的差异也会消失。

表现最差的是无感知团队，这证实了我们的动机：在具有异质能力的合作环境中，具备能力感知非常重要。让专家将新手建模为对手的团队表现同样不佳，这是由于其行动过于保守。最后，我们注意到，仅使用均匀先验并采用能力感知蒙特卡洛树搜索算法，其结果略优于无感知策略或完全最小化策略。这种策略对应于这样一种智能体：它会调整自身行为，以应对队友可能具备不同能力的可能性，但不会尝试推断真实深度。