学习基于密度的马尔可夫博弈相关均衡

Libo Zhang[1], 2, *, Yang Chen[2], *, †, Toru Takisaka[1], Bakh Khoussainov[1], Michael Witbrock[2],和 Jiamou Liu[2], †1 中国电子科技大学，中国 {takisaka,bmk}@uestc.edu.cn[2] 奥克兰大学，新西兰
lzha797@aucklanduni.ac.nz,{yang.chen,jiamou.liu,m.witbrock}@auckland.ac.nz

arXiv:2302.08001v1 [cs.AI] 16 Feb 2023

摘要
相关均衡（CE）是一种成熟的解概念，能够刻画智能体之间的协调，并具有良好的算法特性。在实际的多智能体系统中，人们通常期望智能体的策略不仅处于均衡状态，还能满足有关安全性和公平性的要求。这些附加要求通常可以通过状态密度来表达，该密度衡量博弈过程中各状态的访问频率。然而，现有的相关均衡概念或相关均衡求解方法无法明确指定满足特定状态密度属性的相关均衡；它们只能通过修改奖励函数或使用价值函数作为选择标准来隐式实现。因此，得到的相关均衡可能无法完全满足状态密度要求。本文提出基于密度的相关均衡（DBCE），这是一种将状态密度明确作为选择准则的新型相关均衡概念。具体而言，我们根据实际应用中的需求，指定不同的状态密度要求，以实例化 DBCE。为求解 DBCE，我们针对底层控制问题提出基于密度的相关策略迭代算法。我们在多种博弈中进行了实验，结果表明，在关注状态密度的场景中，与现有方法相比，我们的相关均衡求解方法具有优势。

关键词
相关均衡；状态密度；马尔可夫博弈

ACM 引用格式：
Libo Zhang[1]*, YangChen *, †、Toru Takisaka[1],BakhKhoussainov[1], Michael Witbrock[2],, 和 Jiamou Liu[2], †。2023年。学习基于密度的马尔可夫博弈相关均衡。载于第*22*届国际自主智能体与多智能体系统会议论文集（*AAMAS 2023*），伦敦，英国，*2023*年*5*月*29*日至*6*月*2*日，IFAAMAS，共11页。

1 引言
多智能体系统研究中的一个核心问题，是为理性博弈玩家寻找能够达到特定形式均衡的策略。近年来，针对这一问题的研究趋势是将策略的副作用纳入考量 [23]。事实上，在许多现实世界场景中，很难定义一个能够涵盖智能体期望输出所有方面的奖励函数。例如，机器人系统可能通过执行某种不太理想的高风险操作，或采取被视为不道德的行动，来获得高奖励 [31]。因此，为智能体寻找策略时，仅仅让智能体获得尽可能高的奖励并非最佳选择。更重要的是，这些过程还必须满足其他理想属性，例如奖励未能体现的安全性和公平性。

为简单起见，我们将这类问题表述为：以一个N人马尔可夫博弈为输入，同时要求智能体的策略满足两类要求：

(1) 奖励要求：首先，我们期望理性的智能体不会因效用考量而单方面偏离其策略；(2) 非奖励要求： 其次，策略必须满足某些约束多智能体系统运行过程的非效用要求。

本文研究上述一般问题的一个实例。(1) 对于奖励要求，我们指定了一种会考虑智能体之间可能存在的协调的解概念。更具体地说，我们采用相关均衡 (CE) [17] 作为解概念。与该领域广泛采用的纳什均衡（NE）相比， [19], CE 不要求智能体之间相互独立，因此适用于更广泛的实际场景。此外，CE 的集合构成一个凸多胞体，因此可以通过线性规划轻松求得。许多自适应过程已被证明会收敛到 CE，而非限制性更强的 NE [13, 16]。(2)对于非奖励要求，我们考虑一些重要属性，它们可以粗略地表述为，例如“某种情形不应发生”、“某种情形应以预先规定的频率发生” 以及“两种情形应以相同频率发生”。这些属性可以通过考察运行来刻画，即智能体在博弈过程中所处状态构成的序列。更准确地说，它们是状态分布要求，其定义基于博弈中对各状态的访问：

• 安全要求。这些条件要求某些坏状态不应被访问。许多工业应用都涉及绝不应发生的危险状态。例如，机器人系统处于低电量状态 [29]。类似的问题也可能从伦理视角引发关注 [31]。

* 贡献相同
† 通讯作者

第*22*届国际自主智能体与多智能体系统会议论文集（*AAMAS 2023*），*A. Ricci*、*W. Yeoh*、*N. Agmon*、*B. An*（编），*2023*年*5*月*29*日至*6*月*2*日，英国伦敦。© 2023年国际自主智能体与多智能体系统基金会（www.ifaamas.org）。保留所有权利。

• 频率要求。这些条件要求以固定频率访问某些状态。为了推广安全性，我们可能期望系统从长期来看按一定比例访问某些状态。例如，我们可能希望机器人系统有 30% 的时间以高效模式运行，其余 70% 的时间以正常模式运行。

• 公平性要求。这些条件要求两个状态应以相同的频率被访问。例如，为了系统稳定性等目的，也可能希望平衡系统中两个不同状态的访问频率。比如，一支无人驾驶航空器团队有两个充电站时，人们可能希望平衡它们的使用率，以避免不必要的排队。又比如，在十字路口，两个方向的交通信号灯应以相同的频率亮起绿灯。

使用策略 [29] 在环境中导航时，可以利用状态密度函数来衡量状态访问频率。该函数能够表达上述状态分布要求。然而，迄今为止，尚无关于相关均衡或相关均衡求解算法的工作明确地纳入由状态密度函数定义的要求。另一方面，已有方法通过对奖励施加额外约束，来隐式地表达这些非奖励要求。然而，这些方法可能不足以满足这些期望的要求。具体而言，现有技术分为两类：

(1) 风险敏感奖励修改。该方法通过添加额外项来调整奖励结构，例如对非期望状态施加负奖励，或将方差作为风险项 [5, 12, 25, 32]。当目标简单且能够高效设计调整策略时，这种方法较为适用。然而，由于最优策略对奖励设置较为敏感，因此需要进行参数微调。在计算相关均衡时，该方法会改变原始博弈的 CE 集合形状。因此，在修改后的博弈中找到的最优策略可能并非原始博弈的相关均衡。此外，为诸如价值目标要求之类的复杂目标设计奖励修改，需要专家级的领域知识，这在实际应用中颇具挑战。

(2) 约束方法。该方法通过添加显式约束（例如受约束马尔可夫博弈 [3]），直接选取策略子集。不过，该方法需要进行参数微调

> 因为约束的阈值会直接影响博弈的性能和可行性。在求解博弈之前，设计者无法得知最优解，因此设定正确的阈值颇具挑战。此外，在计算相关均衡时，引入的约束可能会缩小原始博弈的 CE 集合。

上述两种方法都可能改变原始博弈的可行 CE 集合的形状或大小。这些变化如图 1 所示

在本文中，我们在马尔可夫博弈的背景下引入基于密度的相关均衡（DBCE）。通过使用密度函数作为选择准则，DBCE 将状态分布要求（非奖励要求）和奖励要求明确整合到一种新颖的 CE 概念中，同时避免现有相关均衡概念或相关均衡求解方法所存在的上述问题。

然而，存在均衡概念并不一定意味着能够有效地找到均衡。由于策略的状态密度与累积奖励采用的度量不一致，直接计算 DBCE

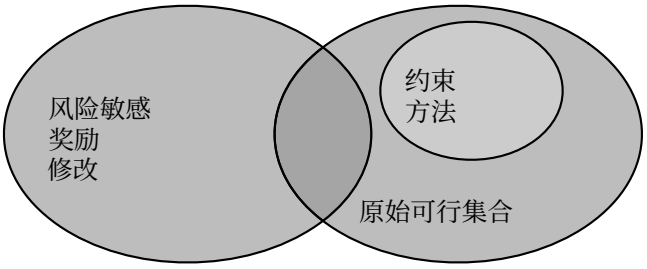


图1：该图展示了现有两种方法对可行相关均衡集合所造成的变化。约束方法会导致可行集合缩小；如果附加约束不可行，可行集合甚至可能为空。风险敏感奖励修改会生成一个新的博弈，从而使可行集合发生偏移。此外，不恰当的奖励修改可能导致修改后的博弈与原始博弈的可行集合之间的交集为空。

难以实现，因为这种不一致使我们无法在同一空间中优化两者。为解决这一挑战，我们采用占用度量这一概念，即状态—动作的累积访问频率。借助这一度量，状态密度和累积奖励都可以表示出来，从而能够以统一的方式对两者进行优化。基于这一方法，我们提出了用于计算 DBCE 的算法，称为基于密度的相关策略迭代 (DBCPI)。具体而言，DBCPI 以迭代方式运行，交替更新智能体的策略和占用度量：通过寻找由当前占用度量诱导且满足非奖励要求的相关均衡来更新策略；随后根据更新后的策略更新当前占用度量。此外，我们还为 DBCPI 提供了理论依据，并给出了其收敛条件。

我们的主要贡献总结如下：
(1) 我们为马尔可夫博弈提出了一种新的相关均衡概念——基于密度的相关均衡（DBCE），该概念利用状态密度函数显式刻画非奖励要求，同时不改变所有可行相关均衡的集合。(2) 为计算 DBCE，我们提出了基于密度的相关策略迭代（DBCPI）。我们证明，在某些假设下，该机制会收敛到有效的 DBCE。(3) 我们在受实际应用启发的不同模拟场景中，将 DBCPI 与现有方法进行了测试。实验结果表明，我们的方法在寻找满足上述额外非奖励要求（即安全性、频率和公平性）的相关均衡方面具有优势。

2 相关研究

均衡概念。在多人博弈中，尤其是非合作博弈中，求解博弈就意味着寻找一个均衡。纳什Q学习 [20] 将经典Q学习扩展到一般和马尔可夫博弈中，以寻找纳什均衡。至于寻找相关均衡（CE），一些研究 [7, 26] 尝试通过确定或近似所得

期望奖励空间的边界来计算相关均衡的全集；研究已证明，该空间是一个凸多胞体。另一种方法是，Greenwald 等人 [14] 提出了一种类Q学习算法，用于在马尔可夫博弈中寻找一个相关均衡实例，而不是求出全集。还有一类研究着重于挖掘相关均衡在现实世界场景中的应用价值，例如，Yu 等人 [37] 和 Han 等人 [15] 利用相关均衡协调工业场景中的设备；Jin 等人 [22] 则将相关均衡作为外包任务定价问题的解决方案。一系列研究通过从全集中选取相关均衡的特定子类，刻画其某些特定性质。例如，Ortiz 等人 [28] 和 Ziebart 等人 [38] 选择策略熵最大的相关均衡，以确保博弈解的唯一性。

非奖励要求。在实际应用中，非奖励要求不可避免。一些研究通过修改奖励函数，以隐式方式满足这些要求 [9]。一种常见方法是在奖励函数中加入风险敏感项，例如方差 [24] 和指数效用函数 [6]。除了隐式调整奖励之外，能够明确描述目标的逻辑指令 [18] 也被视为修改奖励函数的一种方法。近年来，研究者从多个方面研究了采用不同风险敏感因素的强化学习算法 [5, 12, 25, 32]。沿着这一思路，一些研究在学习得到的策略中加入约束，以体现对安全性的考量：Altman [2] 研究了约束马尔可夫决策过程，随后 [10] 将 Q 学习扩展到了约束马尔可夫决策过程。之后，约束方法又进一步扩展到了多智能体环境中的马尔可夫博弈 [1, 3, 11, 21]。在这些约束中，状态密度尤为特殊。典型研究包括 [12]，其直接指定需要避开的不期望状态；Qin 等人 [29] 则提出将密度函数用作约束，以指导强化学习中最优策略的寻找。

3 预备知识

所有自然数、实数和非负实数的集合分别记为 $\mathbb{N}$、$\mathbb{R}$ 和 $\mathbb{R}_{\geq 0}$。对于自然数 N，集合 $\{1, \ldots, N\}$ 记为 $[N]$。

马尔可夫博弈。马尔可夫博弈也称为随机博弈，是马尔可夫决策过程在多智能体环境中的扩展，其中一组智能体在随机环境中进行行动，每个智能体都旨在最大化其累积奖励。

定义 1。 N智能体马尔可夫博弈 是一个元组

$$(\mathcal{S}, \{\mathcal{A}_i\}_{i=1}^N, P, \{r_i\}_{i=1}^N, \eta, \gamma), \ where$$

- $\mathcal{S}$ 是状态集合，
- $\mathcal{A}_i$ 是第 i 个智能体的动作 集合，
- $P: \mathcal{S} \times \boldsymbol{\mathcal{A}} \to \Delta(\mathcal{S})$ 是转移函数，它用于指定给定联合动作$\boldsymbol{a} = (a_1, \ldots, a_n)$时两个状态之间的转移概率，其中$\boldsymbol{\mathcal{A}} = \times_{i=1}^N \mathcal{A}_i$是联合动作空间，而 $\Delta(\mathcal{S})$ 表示$\mathcal{S}$上的概率分布集合，
- $r_i: \mathcal{S} \times \boldsymbol{\mathcal{A}} \to \mathbb{R}$ 是奖励函数，用于确定智能体 i在某一状态下采取联合动作所获得的即时奖励，
- $\eta \in \Delta(\mathrm{S})$ 是状态的初始分布，
- $\gamma \in (0, 1)$ 是折扣因子。

全文中，我们使用不带下标的粗体变量表示所有智能体对应变量的串联并使用下标 $-i$ 表示除 i 之外的所有智能体，例如，$\boldsymbol{a} = (a_1, \ldots\ldots, a_n) = (a_i, \boldsymbol{a}_{-i})$ 表示所有智能体的联合动作。

定义 2。 智能体的（平稳）联合策略是一个函数

$$\boldsymbol{\pi}: \mathcal{S} \to \Delta(\boldsymbol{\mathcal{A}})$$

它规定了智能体根据当前状态概率性地选择动作。所有联合策略的集合记为 $\boldsymbol{\Pi}$。

每个智能体都旨在找到一种策略，以最大化其在整个博弈过程中的累积奖励：$\sum_{t=0}^{\infty} \gamma^t r_i(s^t, \boldsymbol{a}^t)$。对于每个智能体 i，联合策略 $\boldsymbol{\pi}$ 下状态-联合动作对的期望回报定义如下：

$$Q_i^{\boldsymbol{\pi}}(s, \boldsymbol{a}) \triangleq \mathbb{E}\left[\sum_{t=0}^{\infty} \gamma^t r_i(s^t, \boldsymbol{a}^t) \middle| s^0 = s, \boldsymbol{a}^0 = \boldsymbol{a}, P, \boldsymbol{\pi}\right].$$

相关均衡。 马尔可夫博弈的解称为均衡，其对应一个联合策略，在该策略下，没有任何智能体会为了获得奖励而有动机单方面偏离。关于智能体策略之间不同程度独立性的假设，有两种经典的均衡概念。广为人知的 纳什均衡 (NE) [8] 要求智能体之间相互独立，即$\boldsymbol{\pi} = \times_{i=1}^N \pi_i$其中 $\pi_i: \mathrm{S} \to \Delta(\mathrm{A}_i$代表单个智能体的策略。相比之下，相关均衡 (CE) [4] 通过体现智能体之间的协调，对 NE 进行了推广，因此更适用于智能体需要协调行动的多智能体系统。从概念上讲，智能体由一个相关装置进行协调，该装置向每个智能体 i推荐一个动作 $a_i \in \mathrm{A}_i$；每个智能体都知晓其他所有智能体的条件分布 $\boldsymbol{\pi}_{-i}(\boldsymbol{a}_{-i}|s, a_i)$。若要达到 CE，每个智能体都没有违背建议的动机，即选择另一个动作 $a_i' \in \mathrm{A}_i$，该动作称为偏离动作。

定义 3. 马尔可夫博弈的相关均衡*(*相关均衡*)*是满足以下条件的联合策略 $\boldsymbol{\pi}$：

$$\forall i \in [N], s \in \mathcal{S}, a_i, a_i' \in \mathcal{A}_i, \quad \mathrm{reg}_{\boldsymbol{\pi}}(s, i, a_i, a_i') \leq 0. \tag{1}$$

这里，遗憾 $\mathrm{reg}_{\boldsymbol{\pi}}(s, i, a_i, a_i')$体现了转向偏离动作所带来的期望奖励增益：

$$\mathrm{reg}_{\boldsymbol{\pi}}(s, i, a_i, a_i') \triangleq \mathbb{E}_{\boldsymbol{a}_{-i} \sim \boldsymbol{\pi}_{-i}(\cdot|s, a_i)}\left[Q_i^{\boldsymbol{\pi}}(s, a_i', \boldsymbol{a}_{-i}) - Q_i^{\boldsymbol{\pi}}(s, a_i, \boldsymbol{a}_{-i})\right].$$

纳什均衡（NE）的一般存在性 [8] 意味着相关均衡（CE）的存在。与纳什均衡相比，相关均衡具有更好的数学性质：式 (1) 中的约束在智能体的期望回报空间中定义了一个 N 维多面体，而纳什均衡集合由多面体中的孤立点[27] 组成。因此，由于式 (1) 是一个线性不等式组，可以通过线性规划求解标准形博弈（等价于单阶段马尔可夫博弈）的相关均衡集合。不过，由于两个原因，精确计算马尔可夫博弈的相关均衡通常难以处理：（i）由于式 (1) 中的 Q 和 $\boldsymbol{\pi}$ 均未知，相关均衡的约束会变为非线性不等式；（ii）随着时域增加，相关均衡多面体的顶点数量呈指数增长 [38]。

密度函数。 密度函数 [30] $\rho: \mathrm{S} \to \mathbb{R}_{\geq 0}$ 用于衡量使用策略在环境中导航时各状态的访问频率。形式上，对于初始分布为 η、折扣因子为 γ的无限时域马尔可夫博弈，联合策略 $\boldsymbol{\pi}$ 下的密度函数定义为

$$\rho^{\boldsymbol{\pi}}(s) \triangleq \sum_{t=0}^{\infty} \gamma^t \Pr(s^t = s|\boldsymbol{\pi}, s^0 \sim \eta).$$

请注意，密度函数也可以写成递归形式：

$$\rho^{\boldsymbol{\pi}}(s) = \eta(s) + \boldsymbol{\pi}(s,a)\gamma \sum_{s' \in \mathcal{S}} \sum_{\boldsymbol{a} \in \mathcal{A}} P(s|s', \boldsymbol{a}) \rho^{\boldsymbol{\pi}}(s').$$

占用度量。与密度函数类似，占用度量 $\rho(s, a) : \mathsf{S} \times \mathcal{A} \to \mathbb{R}_{\geq 0}$用于衡量给定平稳策略时状态-动作对的访问频率。形式上，策略 $\boldsymbol{\pi}$ 下的占用度量 $\rho^{\boldsymbol{\pi}}$ 定义为

$$\rho^{\boldsymbol{\pi}}(s, \boldsymbol{a}) \triangleq \sum_{t=0}^{\infty} \gamma^t \Pr(s^t = s, \boldsymbol{a}^t = \boldsymbol{a}|\boldsymbol{\pi}, s^0 \sim \eta).$$

我们可以通过方程 $\rho^{\boldsymbol{\pi}}(s) = \sum_{\boldsymbol{a} \in \mathcal{A}} \rho^{\boldsymbol{\pi}}(s, \boldsymbol{a})$ 计算状态的密度。占用度量还有一些有助于通过优化进行策略综合[33]的性质。首先，函数$f: \mathsf{S} \times \mathcal{A} \to \mathbb{R}$ 是某个平稳策略下的占用度量，当且仅当它满足以下贝尔曼流约束：

$$\mathrm{BFError}_f(s) = 0, \forall s \in \mathcal{S} \quad \text{AND} \quad f(s, \boldsymbol{a}) \geq 0, \forall s \in \mathcal{S}, \boldsymbol{a} \in \mathcal{A}, \tag{2}$$

其中，BFError$f(s)$表示相对于状态-动作访问频率的贝尔曼残差：

$$\mathrm{BFError}_f(s) = \sum_{\boldsymbol{a} \in \mathcal{A}} f(s, \boldsymbol{a}) - \eta(s) - \gamma \sum_{s' \in \mathcal{S}} \sum_{\boldsymbol{a} \in \mathcal{A}} P(s|s', \boldsymbol{a}) f(s', \boldsymbol{a}).$$

另一方面，对于满足 BF 约束的 fs，存在与 f 相关联的唯一平稳策略 $\boldsymbol{\pi} \in \boldsymbol{\Pi}$，使得 f 是策略 $\boldsymbol{\pi}$ 下的占用度量（即 $\rho^{\boldsymbol{\pi}} = f$）；此外，可以通过以下方式构造这样的策略：

$$\boldsymbol{\pi}(s, \boldsymbol{a}) = f(s, \boldsymbol{a}) \Big/ \sum_{\boldsymbol{a}' \in \mathcal{A}} f(s, \boldsymbol{a}'). \tag{3}$$

非奖励要求。除了均衡概念所涵盖的奖励要求外，非奖励要求也受到了关注。在此，我们考虑三种典型的非奖励要求：安全性、频率和公平性要求。这些要求可以通过博弈轨迹中某些状态出现的次数来衡量，博弈轨迹即由马尔可夫博弈中的策略生成的状态序列。长度为 $n+1$ 的有限轨迹表示为 $\tau \triangleq s_0, s_1, \ldots\ldots, s_n$。我们可以以轨迹为中心的方式形式化上述三类非奖励要求：

- 安全性：对于一组非期望状态 S^*，我们期望轨迹中非期望状态的计数等于零，$\sum_{i \in [0,n]} \mathbb{I}(s_i \in S^*) = 0$ 其中 $\mathbb{I}$ 为示性函数；
- 频率：对于一组特定状态 S^*，我们期望此类状态在轨迹中以一定比例出现 c，$\sum_{i \in [0,n]} \mathbb{I}(s_i \in S^*)/(n+1) = c$
- 公平性：对于两组状态 S_1^*，S_2^*，我们期望在轨迹中，这两组状态的出现次数相等，$\sum_{i \in [0,n]} \mathbb{I}(s_i \in S_1^*) = \sum_{i \in [0,n]} \mathbb{I}\ s_i \in S_2^*$ ()。直观地说，我们在图 2 中通过一个具有两种不同状态的马尔可夫博弈展示了三种非奖励要求。

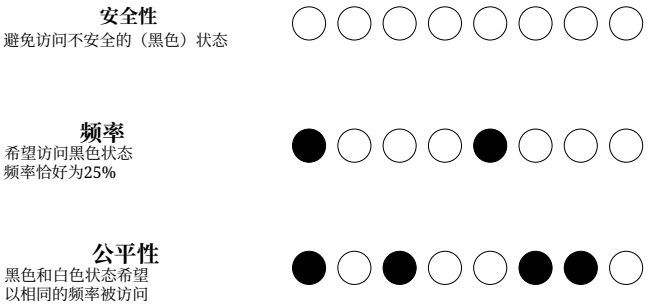


图2：三种非奖励要求的示例。圆圈序列表示轨迹，黑色和白色分别表示两种不同的状态。

4 基于密度的相关均衡

本节首先提出基于密度的相关均衡（DBCE）的一般定义。随后，我们将选择准则具体化为上述三种非奖励要求。

回顾一下，DBCE 的动机是寻找一种均衡，既能捕捉智能体之间的协调，也能捕捉策略的副作用。这些副作用无法简单地用奖励表示，但可以通过密度函数来解释。为此，我们将一个或一组密度函数作为选择准则，用以从相关均衡中筛选出满足预期非奖励要求的子集，从而对 DBCE 进行形式化。由此，DBCE 被定义为一个约束优化问题的解，其中密度函数充当目标函数，而约束则确保满足相关均衡的条件。

定义 4。设给定以下内容：

- 一个马尔可夫博弈 $(\mathcal{S}, \{\mathcal{A}_i\}_{i=1}^N, P, \{r_i\}_{i=1}^N, \eta, \gamma)$；
- 状态集合 $\mathcal{S}^* = \{s_1, \ldots\ldots, s_m\} \subseteq \mathcal{S}$的一个子集；
- 一个实值函数 $F: \mathbb{R}^m \to \mathbb{R}$；
- 一个函数 $\varphi(\boldsymbol{\pi}) = F(\rho^{\boldsymbol{\pi}}(s_1), \ldots\ldots, \rho^{\boldsymbol{\pi}}(s_m))$，我们称之为 $\boldsymbol{\pi}$的密度误差 。

若联合策略 $\boldsymbol{\pi}$ 是以下约束优化问题的解，则称其为(F-指定的)基于密度的相关均衡（*DBCE*）：

$$\min_{\boldsymbol{\pi} \in \boldsymbol{\Pi}} \varphi(\boldsymbol{\pi}) \quad \textit{subject to} \quad \mathrm{reg}_{\boldsymbol{\pi}}(s, i, a_i, a_i') \leq 0, \forall i \in [N], s \in \mathcal{S}, a_i, a_i' \in \mathcal{A}_i. \tag{4}$$

我们可以用密度误差 $\varphi(\boldsymbol{\pi})$ 的值来衡量 $\boldsymbol{\pi}$ 在状态密度方面的质量。通过选择合适的函数 F，我们可以构造出能够刻画特定非奖励要求的 DBCE。这里，我们介绍以下针对安全性、频率和公平性要求的特定 DBCE：若一个 DBCE 被称为

- 最小密度 *CE*（*MDCE*）当我们有φ（$\boldsymbol{\pi}$）= $\sum_{s \in \mathcal{S}^*} \rho^{\boldsymbol{\pi}}$（$s$）；
- 频率匹配 *CE*（*FMCE*）当我们有$\varphi(\boldsymbol{\pi}) = |\sum_{s \in \mathcal{S}^*} \rho^{\boldsymbol{\pi}}(s) - c|$对于某个 $c \in \mathbb{R}_{\geq 0}$；

- 最小密度差相关均衡 *(MDGCE)* 当我们有 $\varphi(\boldsymbol{\pi}) = |\sum_{s\in S_1}\rho^{\boldsymbol{\pi}}(s) - \sum_{s\in S_2}\rho^{\boldsymbol{\pi}}(s)|$对于 $S_1, S_2 \subseteq \mathcal{S}^*$。

直观而言，我们分别使用 MDCE、FMCE 和 MDGCE 来表示具有安全性、频率和公平性要求的 CE。这些具体实例表明，DBCE 具备刻画具有某些密度相关属性的均衡的一般能力，而这些属性尚无法由其他均衡概念表示。

5 基于密度的相关均衡求解

本节旨在介绍一种用于计算 DBCE 的策略迭代算法，并证明其在特定假设下的收敛性。为简便起见，我们的分析围绕最小密度 CE (MDCE) 展开；但该分析也适用于任何其他形式的 DBCE。

5.1 基于密度的相关策略迭代 tion

回顾一下，计算 MDCE 需要求解式 (4) 中定义的约束优化问题，其中 $\varphi(\boldsymbol{\pi}) = \sum_{s\in\mathcal{S}^*}\rho^{\boldsymbol{\pi}}(s)$。然而，直接求解该问题是难以处理的，因为目标函数中的密度函数与约束中的期望回报定义在两个不同的空间中；这使我们无法以统一的方式优化与密度相关的目标，同时满足与奖励相关的约束。因此，我们需要找到一种统一表示状态密度和期望回报的方法。幸运的是，我们发现，状态密度和期望回报都可以用第 3 节引入的占用度量来改写。因此，我们只需维护一个变量，而不必在两个独立的空间中分别维护变量，就能同时控制二者。

接下来，我们将展示如何推导出原始约束优化问题的一个等价且易于处理的形式。首先，我们根据占用度量重写公式 4 的约束，即 $\mathrm{reg}_\prime f(s, i, a_i, a_{\prime i})$，其定义如下：

$$\sum_{\boldsymbol{a}_{-i}} f(s, a_i, \boldsymbol{a}_{-i}) \left[Q_i^{\boldsymbol{\pi}}(s, a_i', \boldsymbol{a}_{-i}) - Q_i^{\boldsymbol{\pi}}(s, a_i, \boldsymbol{a}_{-i})\right].$$

根据公式 (3) 可知，这等价于 $\mathrm{reg}_\prime\rho^{\boldsymbol{\pi}}(s, i, a_i, a_{\prime i})$。

根据占用度量与密度函数之间的关系，目标函数 $\varphi(\boldsymbol{\pi})$ 也可以重写为 $\varphi_\prime(f)$，其中

$$\varphi'(f) \triangleq F\Big(\sum_{\boldsymbol{a}} f(s_1, \boldsymbol{a}), \ldots, \sum_{\boldsymbol{a}} f(s_m, \boldsymbol{a})\Big). \tag{5}$$

回顾一下，密度函数可以改写为单个状态内占用度量之和：$\rho^{\boldsymbol{\pi}}(s) = \sum_{\boldsymbol{a}\in\text{一个}}\rho^{\boldsymbol{\pi}}(s, \boldsymbol{a})$，新的目标函数就等价于原目标函数。至此，我们通过占用度量实现了非奖励要求与奖励要求之间的一致性。根据上一节讨论的占用度量的性质，我们将问题改写如下：

Problem 1. $$\min_{f:\mathcal{S}\times\mathcal{A}\to\mathbb{R}} \sum_{s\in S^*}\sum_{\boldsymbol{a}\in\mathcal{A}} f(s, \boldsymbol{a}) \quad \text{subject to}$$

$$\mathrm{reg}'_f(s, i, a_i, a_i') \le 0, \qquad \forall i \in [N], s \in \mathcal{S}, a_i, a_i' \in \mathcal{A}_i; \tag{6}$$

$$\mathrm{BFError}_f(s) = 0, \qquad \forall s \in \mathcal{S}; \tag{7}$$

$$f(s, \boldsymbol{a}) \ge 0, \qquad \forall s \in \mathcal{S}, \boldsymbol{a} \in \mathcal{A}. \tag{8}$$

这里，贝尔曼流约束 (7) 和 (8) 确保 f 是某个 $\boldsymbol{\pi} \in \Pi$ 下的占用度量。对于这样的 f，新的目标函数 $\sum_{\boldsymbol{a}\in\text{一个}} f(s^*, \boldsymbol{a})$ 等于 $\rho^{\boldsymbol{\pi}}(s^*)$，而 (6) 确保 $\boldsymbol{\pi}$ 是相关均衡。由于占用度量与平稳策略之间存在一一对应关系，问题1的解就是原问题的某个解所对应的占用度量。

然而，问题1仍然难以直接求解，因为 f 和 Q（涉及 $\mathrm{reg}_\prime f$）均为未知量。我们引入一种迭代方法来处理这一问题，并将其称为基于密度的相关策略迭代 （DBCPI）。该方法在以下两步之间交替进行：(i) 策略评估：根据当前策略估计 Q 值；以及 (ii) 策略改进：根据当前 Q 函数计算一个 DBCE。更正式地，令 t 表示迭代次数的索引。在每次迭代中，$\mathbf{Q}^t = \{Q_i^t\}_{i\in[N]}$定义了一个阶段博弈 ，其中 Q值为常数。将 $\mathrm{reg}_f^t(s, i, a_i, a_{\prime i})$ 定义如下：

$$\sum_{\boldsymbol{a}_{-i}} f(s, a_i, \boldsymbol{a}_{-i}) \left[Q_i^t(s, a_i', \boldsymbol{a}_{-i}) - Q_i^t(s, a_i, \boldsymbol{a}_{-i})\right].$$

将问题1中的 $\mathrm{reg}_\prime f$替换为 reg_f^t后，阶段博弈 $\mathbf{Q}^t$ 即可通过线性规划求解。得出当前阶段博弈 $\mathbf{Q}^t$的 DBCE $\boldsymbol{\pi}^t$ 后，我们返回更新 Q 函数并得出 $\mathbf{Q}^{t+1}$。伪代码见算法 1。

算法1 基于密度的相关策略迭代

输入： 一个马尔可夫博弈 $(\mathrm{S}, \mathcal{A}, P, \{r_i\}_{i=1}^N, \eta, \gamma)$。
初始化： Q_i对每个 $i \in [N]$，学习率 α
$\boldsymbol{\pi}(s, \boldsymbol{a}) \leftarrow f(s, \boldsymbol{a})/\sum_{\boldsymbol{a}'\in\text{甲}} f(s, \boldsymbol{a}')$
对每次迭代**执行**
  $f \leftarrow$(问题1 的解，其中 $\{Q_i\}_{i\in[N]}$)
  $\boldsymbol{\pi}(s, \boldsymbol{a}) \leftarrow f(s, \boldsymbol{a})/\sum_{\boldsymbol{a}'\in\mathcal{A}} f(s, \boldsymbol{a}')$
  当 未收敛**时**
    初始化状态 $s \in \mathrm{S}$
    观测转移 $(s, \boldsymbol{a}, \boldsymbol{r}, s')$
    对每个 $i \in [N]$ **执行**
      $V_i(s') \leftarrow \sum_{\boldsymbol{a}'\in\mathrm{A}} \boldsymbol{\pi}(s'\boldsymbol{a}')Q_i(s'\boldsymbol{a}')$
      $Q_i(s, \boldsymbol{a}) \leftarrow (1-\alpha)Q_i(s, \boldsymbol{a}) + \alpha(r_i + \gamma V_i(s'))$
    结束循环
    衰减 α
  结束循环
结束循环
输出： 一个联合策略 $\boldsymbol{\pi}$，以及 $\varphi_\prime(f)$，作为 $\boldsymbol{\pi}$ 的误差。

5.2 收敛性分析

接下来，我们证明应用算法 1 时，$\mathbf{Q}^t$ 会收敛到 DBCE 下的 Q 值。首先，我们引入以下有用的技术假设。

假设 1。 每个状态 $s \in \mathcal{S}$ 和动作 $a_i \in \mathcal{A}_i$ 对于所有 $i \in [N]$ 都会被无限次访问。

假设2。 奖励由某个常数界定。

假设 3。 学习率 α_t 满足以下条件： $0 \le \alpha_t < 1\ \forall t$、$\sum_t \alpha_t = \infty$ 和 $\sum_t \alpha_t^2 < \infty$。

我们还需要以下引理，以保证在求解每个阶段博弈后，策略估计过程能够在 $\boldsymbol{\pi}^t$ 下收敛到 Qvalues。

引理1（[34]）。令 $\mathbb{Q}$ 为所有 Q 函数构成的空间。在假设*1-3*下，由以下公式定义的迭代以概率*1*收敛到$Q^{\boldsymbol{\pi}}$ ：

$$Q^{t+1}(s,\boldsymbol{a}) = (1-\alpha)Q^t(s,\boldsymbol{a}) + \alpha_t\Big(r(s',\boldsymbol{a}) + \gamma \sum_{\boldsymbol{a}} \boldsymbol{\pi}(\boldsymbol{a}|s')Q^t(s',\boldsymbol{a})\Big).$$

现在，我们可以陈述主要定理，该定理表明，在假设下，DBCPI收敛到一个DBCE Q 函数。

定理1。 在假设*1-3*下，如果对于所有 t、$s \in \mathcal{S}$和 $i \in [N]$，策略 $\boldsymbol{\pi}_t$被识别为以下式子所表示的全局最优解，则算法*1*中迭代更新的 Q 函数将收敛到*DBCE*下的函数：

$$\forall \boldsymbol{\pi}' \in \Pi, \quad \mathbb{E}_{\boldsymbol{a}\sim\boldsymbol{\pi}^t}[Q_i^t(s,\boldsymbol{a})] \ge \mathbb{E}_{\boldsymbol{a}\sim\boldsymbol{\pi}'}[Q_i^t(s,\boldsymbol{a})].$$

证明概要。基本思路是证明由DBCPI引导的策略在奖励方面单调改进。根据引理1，对于所有 t，经过充分多轮更新后，我们在 $\boldsymbol{\pi}^t$ 下得到 Q函数$\mathbf{Q}^{t+1} = \{Q_i^{t+1}\}_{i\in[N]}$。根据假设，对于每个遇到的阶段博弈$\mathbf{Q}^t$，始终存在一个全局最优策略。因此，随着迭代的进行，每次迭代中的策略都会单调改进：

$$\mathbb{E}_{\boldsymbol{a}\sim\boldsymbol{\pi}^{t+1}}[Q_i^{t+1}(s,\boldsymbol{a})] \ge \mathbb{E}_{\boldsymbol{a}\sim\boldsymbol{\pi}^{t}}[Q_i^{t+1}(s,\boldsymbol{a})],$$

对所有 $t, s \in \mathrm{S}$，以及 $i \in [N]$ 这意味着，经过足够多次迭代后，策略将收敛到全局最优策略， Q 函数也会如此。通过使用线性规划求解每个阶段博弈 $\mathbf{Q}^t$，即可满足问题1中的所有约束。收敛时，该策略因此是一个可行的 DBCE。□

尽管定理1指出，在对每个阶段博弈施加严格约束的条件下，收敛性能够成立，但我们在实验中发现，这些约束并非 DBCPI 收敛的必要条件。这一事实与 [20, 36] 中的实证分析一致。

6 实验

我们通过实验试图回答以下问题：

问题*1.* 我们提出的算法能否找到优于其他方法的相关均衡？我们通过检查算法 1 终止时的以下数值来进行评估，终止于经过 K 次迭代后：

$$\mathrm{MaxReg} \triangleq \max_{s,i,a_i,a_i'} \mathrm{reg}_f^{K+1}(s,i,a_i,a_i'),$$

$$\mathrm{MaxBF} \triangleq \max_{s} |\mathrm{BFError}_f(s)|.$$

MaxReg 可以看作 $\boldsymbol{\pi}$ 与相关均衡集合之间的“距离”：该值越大，某个智能体偏离 $\boldsymbol{\pi}$ 的激励就越大。特别地，当该值不大于零时， $\boldsymbol{\pi}$ 是相关均衡。MaxBF 用于评估第 5 行计算的可靠性：该值越大， f 偏离 $\rho^{\boldsymbol{\pi}}$ 的程度就越大，即



(a) FairGamble (b) 猎人 (c) CaE

图 3：博弈截图。

$\boldsymbol{\pi}$ 的占用度量。这种偏离 f 意味着，值 $\varphi_f(f)$ 作为 $\boldsymbol{\pi}$ 的误差是不可靠的。

为评估此问题中的算法，我们重点关注所有情形下的 MaxReg 和 MaxBF 值。值越小，算法表现越好。我们额外开展了 9 项 MDCE 任务，以比较风险敏感奖励修改下修改后的博弈与原始博弈中的 MaxReg。

问题 *2.* 我们的策略是否会生成期望的轨迹？我们通过考察在所有三项要求下，由算法 1 计算出的策略所生成的各条轨迹模式来评估这一点。

问题 *3.* 与现有方法相比，我们的 *DBCPI* 准确度如何？为比较准确度，我们比较误差（即算法 1 输出中的值 $\varphi_f(f)$），并从数据中选取若干实例，通过绘图加以说明。

Q4. 我们的 *DBCPI* 在学习中的收敛表现如何？我们通过在实验中绘制 DBCPI 的迭代误差图来对此进行评估。

6.1 实验设置

6.1.1 博弈环境与任务。 我们考虑三种博弈模型（公平赌博、猎人、收集与探索）；对于每种模型，我们分别施加三种不同的状态分布要求，因此总共产生了九个作为算法 1 输入的实例。我们已匿名发布了算法在这些博弈中的动画演示，可通过 https://github.com/nanaralala/Density-based-Correlated-Equilibrium/ 获取。截图见图 3。以下介绍各个博弈。

(1) 公平赌博。在这个博弈中，两名赌徒相互对弈，并从三种不同的游戏中进行选择。这三种游戏都完全公平，因此无论他们如何选择，奖励都将随机发放。游戏 1 公平地发放 0 奖励；游戏 2 向一名赌徒发放 0.5，向另一名赌徒发放 -0.5；游戏 3 向一名赌徒发放 1，向另一名赌徒发放 -1。每回合，赌徒从 0 、 1 、 2 中选择一个数字，我们将这些数字进行比较，以确定他们进行哪种游戏。说明见图 3a。我们考虑三项与状态相关的要求，即：(a) 一项安全要求，要求赌徒避开游戏 3；(b) 一项频率匹配要求，要求赌徒选择游戏 3 的频率为 10%；(c) 一项公平性要求，要求游戏 1 和游戏 2 的频率相同。

(2) 猎人 在这个非合作博弈中，3 名猎人住在同一个村庄里。每回合，他们都位于村庄内或村庄外，并可以选择去打猎或守卫

村庄，抵御动物的袭击，见图 3b。如果一名猎人从村庄出发去打猎，该猎人将获得高奖励（1），其余猎人将获得低奖励（0.1）。如果一名猎人守卫村庄，所有猎人都将获得相同的中等奖励（0.5）。如果一名猎人在村庄外并继续在外打猎，我们认为这种行为不够安全，因此该猎人获得的奖励会减少（0.5），而其他猎人将获得惩罚奖励（-0.5）。此外，如果守卫村庄的猎人少于或等于 1 名，他们将受到高惩罚奖励（-3）。

我们考虑三项与状态相关的要求，即：(a) 安全要求，要求至少有 2 名猎人留在村庄。(b) 频率匹配要求，要求 10% 的时间里，在村庄守卫的智能体数量不超过 1 名。(c) 公平性要求，要求猎人 1 外出狩猎的频率等于猎人 2 和猎人 3 外出狩猎频率之和。

(3) 收集与探索(CaE)。在这个合作博弈中，3 名智能体被困在森林里，见图 3c。他们可以选择探索环境，或在住所附近收集食物。如果多名智能体选择外出，我们会随机选择其中一名外出，其他智能体将得不到奖励。每回合，如果现有智能体外出探索，我们就将奖励增加 1；如果有智能体在附近收集食物，我们就将奖励增加 0.3。由于这是一个合作环境，我们为所有智能体设置相同的奖励。

我们考虑三项与状态相关的要求，即：(a) 一项安全要求，要求智能体 1 不要外出；(b) 一项频率匹配要求，将智能体 1 外出的频率限制为 10%；(c) 一项公平性要求，要求智能体 1 探索的频率等于智能体 2 和智能体 3 探索频率之和。

6.1.2 基线方法。 我们采用带有风险敏感奖励修改的 功利主义 *CE-Q* [14] 以及约束方法。详情如下：

(1) 风险敏感奖励修改（*RM*）。 在 $s \in S^*$处的奖励中加上一个负常数 $p < 0$。我们用 RM-p 表示采用特定 p 的算法。(2) 约束方法（*CM*）。 添加一项额外约束 $\varphi_r(f) \leq b$（参见公式 (5)）。我们用 CM-b 表示采用特定 b 的算法。

RM 仅作为安全要求的基线使用；据我们所知，对于频率匹配要求和公平性要求，尚不存在一种规范的方法。对于实验中考虑的所有状态要求，我们都使用 CM。

6.1.3 实现细节。算法 1 的迭代次数设为 250。算法 1 中的参数设为 $\gamma = 0.99$， α 从 0.3 衰减至 0.001。对于每个实验环境，我们运行算法 3 次，并将 3 次运行的结果取均值。在程序中，算法 1 第 5 行的优化问题使用 [35] 中的优化器求解。

表 1：所找到的相关均衡能力比较。

Game	度量指标	方法	要求		
			安全性	公平性	频率-10
FairGamble	误差	DBCE	**1.225**	32.967	8.683
		CM-0.05	8.331	**0.466**	**6.053**
		CM-5	9.65	8.814	5.668
		RM-1.5	17.725	—	—
	MaxBF	DBCE	0.464	**0.13**	**0.521**
		CM-0.05	16.496	8.16	11.612
		CM-5	18.054	7.781	6.623
		RM-1.5	**0.032**	—	—
	MaxReg	DBCE	0.164	**0.034**	**0.08**
		CM-0.05	0.174	0.172	0.22
		CM-5	0.391	0.107	0.231
		RM-1.5	**0.11**	—	—
猎人	误差	DBCE	14.129	4.442	3.643
		CM-0.05	2.313	2.283	**0.05**
		CM-5	2.124	**2.174**	5
		RM-1.5	**0.828**	—	—
	MaxBF	DBCE	**0**	**0**	0.035
		CM-0.05	0.005	0.001	**0**
		CM-5	0.004	0.032	**0**
		RM-1.5	0.001	—	—
	MaxReg	DBCE	**0.044**	**0.061**	**0.037**
		CM-0.05	9.171	0.52	0.225
		CM-5	0.479	1.18	0.18
		RM-1.5	0.842	—	—
CaE	误差	DBCE	0.419	**0.002**	7.608
		CM-0.05	**0.242**	0.05	**1.174**
		CM-25	16.802	9.688	11.854
		RM-0.5	23.838	—	—
	MaxBF	DBCE	**0**	**0**	**0**
		CM-0.05	0.023	**0**	**0**
		CM-25	0.004	1.151	**0**
		RM-0.5	0.023	—	—
	MaxReg	DBCE	**0.002**	**0.001**	0.003
		CM-0.05	1.266	0.002	0.174
		CM-25	0.419	6.695	**0.001**
		RM-0.5	1.434	—	—

6.2 实验结果

以下结果和讨论回答了本节开头提出的问题。

Q1. 这个问题的结果见表1中的 MaxBF 和 MaxReg 两行。在 9 种情况中的 7 种情况下，DBCE 的 MaxBF 值最小，这意味着 DBCE 在寻找占用度量并将其映射为策略方面表现更好。在 9 种情况中的 7 种情况下，DBCE 的 MaxReg 值也最小，这意味着 DBCE 在相关均衡集合中寻找策略方面表现更好。根据这一观察，我们的选择标准在这些实验中寻找相关均衡策略的表现更好。在额外进行的 9 次 MDCE 运行中，有 7 次显示，原始博弈中到相关均衡集合的距离大于

修改后的博弈中到相关均衡集合的距离，这表明 RM 方法存在劣势。

问题*2*。关于该问题的结果见图4a（MDCE）、图4b（FMCE）和图4c（MGDCE）。每幅图都包含若干条由我们的算法生成的轨迹。我们估算轨迹与期望访问次数之间的差距，因此数值越接近0，轨迹表现越好。在图4a中，我们观察到这些轨迹具有预期的性质，即访问非期望状态的次数较少。在图4b中，我们观察到算法生成的轨迹逐渐收敛到期望的访问比例。少数轨迹与目标值的偏差较大，但其他轨迹表现良好。在图4c中，所有轨迹都能在30步内很好地平衡访问情况（其余220步也呈现出类似的模式）。

问题*3*。该问题所选结果如图2所示。我们选取了 MaxReg 和 MaxBF 数值相近的案例，也就是说，剔除了 MaxReg 或 MaxBF 数值较大的案例；请注意，要声称算法以计算所得的误差找到了一个相关均衡，这些数值应当（接近）为零。完整结果见表1。

具体而言，DBCE 在大多数情况下的 MaxReg 和 MaxBF 值最小。在 CaE-MinGap 情况下，DBCE 和 Cons005 在 MaxReg 和 MaxBF 上的表现几乎相同，但 DBCE 的误差更小。

Q4. 这个问题的结果见图5。我们可以观察到，误差-步数曲线逐渐趋于稳定。随着学习的进行，早期阶段的波动逐渐消失。我们认为，这种早期波动可能是由 Q 函数的更新引起的；在 Q 函数稳定后，风险也趋于稳定。这是合理的，因为这是一个策略迭代过程。

表2：可比实例说明

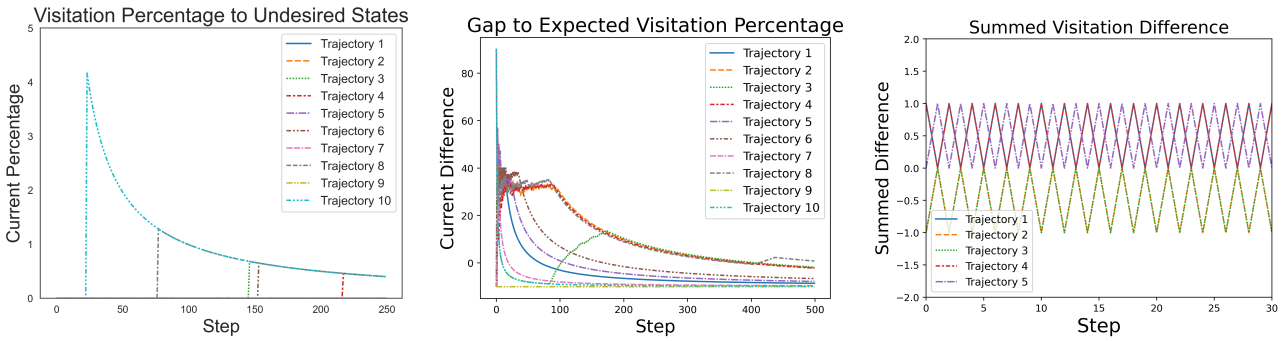
Task	方法	MaxReg	MaxBF	误差
Hunters-MinGap	DBCE	**0.061**	**0**	4.442
	Cons005	0.52	0.001	2.283
	Cons5	1.18	0.032	2.174
FairGamble-MDCE	DBCE	0.164	0.464	1.225
	RewMod	**0.11**	**0.032**	17.725
CaE-MinGap	DBCE	**0.001**	**0**	0.002
	cons005	0.002	**0**	0.05
	cons5	6.695	1.151	9.688
CaE-MDCE	DBCE	**0.002**	**0**	0.419
	cons005	1.266	0.023	0.242
	cons25	0.419	0.004	16.802
	RewMod	1.434	0.023	23.838

7 结论与未来工作 S

本文提出了一种新的相关均衡概念——基于密度的相关均衡（DBCE）。它使我们能够找到同时满足奖励要求（即均衡）和由状态密度函数刻画的非奖励要求的联合策略。与现有方法不同，DBCE 既不会改变博弈中可行相关均衡集合的形状或大小，也不存在参数调优问题。我们通过占用度量关联密度与奖励，并设计了基于密度的相关策略迭代（DBCPI）来计算 DBCE。在各种博弈上的实验验证了我们的方法在寻找期望的相关均衡方面的优势。未来，可以考虑实现参数化版本的 DBCPI，以求解状态—动作空间连续且更复杂的博弈。此外，基于密度的目标可能会在相关均衡空间中产生多个点，因此，进一步从这些候选点中进行选择也可以作为下一步工作。

参考文献

[1] E. Altaian、K. Avrachenkov、Nicolas Bonneau、mérouane Debbah、Rachid El- Azouzi和Daniel Menasché。2007年。无线网络中的受约束随机博弈。GLOBECOM - IEEE全球电信会议，315–320。https://doi.org/ 10 1109 2007 66 。/GLOCOM。. Eitan Altman。. 受约 [2] 1993 束马尔可夫决策过程的渐近性质。运筹学研究杂志37，第2期（1993年），151–170。[3] Eitan Altman和Adam Shwartz。2000年。受约束马尔可夫博弈：纳什均-衡。在《动态博弈与应用进展》中。施普林格，213–221。[4] Robert J Aumann。1987年。作为贝叶斯理性表达的相关均衡。《计量经济学》：计量经济学会杂志（1987年），1–18。[5] Vivek S Borkar。2002年。用于风险敏感控制的Q学习。《运筹学数学》27，第2期（2002年），294–311。[6] Kun-Jen Chung和Matthew J Sobel。1987年。折扣马尔可夫决策过程：分布函数与指数效用最大化。《SIAM控制与优化杂志》25，第1期（1987年），49–62。[7] Liam Dermed和Charles Isbell。2009年。求解随机博弈。神经信息处理系统进展 22——2009年会议论文集，1186–1194。[8] Arlington M Fink。1964年。随机 *n*人博弈中的均衡。《广岛大学科学杂志》，AI系列（数学）28，第1期（1964年），89–93。[9] Javier Garcıa和Fernando Fernández。2015年。安全强化学习的综合综述。《机器学习研究杂志》16，第1期（2015年），1437–1480。[10] Ather Gattami、Qinbo Bai和Vaneet Agarwal。2019年。用于多目标与受约束马尔可夫决策过程的强化学习。arXiv预印本arXiv:1901.08978（2019年）。[11] Yangyang Ge、Fei Zhu、Wei Huang、Peiyao Zhao和Quan Liu。2020年。基于受约束马尔可夫博弈的多智能体协作Q学习算法。《计算机科学与信息系统》17，第2期（2020年），647–664。[12] Peter Geibel和Fritz Wysotzki。2005年。应用于受约束控制的风险敏感强化学习。《人工智能研究杂志》24（2005年），81–108。[13] Geoffrey J Gordon、Amy Greenwald和Casey Marks。2008年。凸博弈中的无悔学习。载于第25届国际机器学习会议论文集。360–367。[14] Amy Greenwald、Keith Hall、Roberto Serrano等。2003年。相关Q学习。载于ICML，第3卷。242–249。[15] Zhu Han、Charles Pandana和KJ Ray Liu。2007年。利用相关均衡和无悔学习实现认知无线电的分布式机会频谱接入。载于2007年IEEE无线通信与网络会议。IEEE，11–15。[16] Sergiu Hart和Andreu Mas-Colell。2000年。导向相关均衡的简单自适应过程。《计量经济学》68，第5期（2000年），1127–1150。[17] Sergiu Hart和Andreu Mas-Colell。2001年。导向相关均衡的强化过程。载于《经济学论文集》。施普林格，181–200。[18] Mohammadhosein Hasanbeig、Alessandro Abate和Daniel Kroening。2018年。逻辑约束强化学习。arXiv预印本arXiv:1801.08099（2018年）。[19] Charles A Holt和Alvin E Roth。2004年。纳什均衡：一种视角。《美国国家科学院院刊》101，第12期（2004年），3999–4002。[20] Junling Hu和Michael P Wellman。2003年。一般和随机博弈的纳什Q学习。《机器学习研究杂志》4，11月（2003年），1039–1069。[21] Xiaofeng Jiang、Shuangwu Chen、Jian Yang、Han Hu和Zhenliang Zhang。2020年。平均准则下连续受约束马尔可夫博弈的均衡求解。IEEE自动控制汇刊65，第12期（2020年），5399–5406。[22] Haiming Jin、Hongpeng Guo、Lu Su、Klara Nahrstedt和Xinbing Wang。2019年。具有马尔可夫相关均衡的多请求方移动群智感知中的动态任务定价。载于IEEE INFOCOM 2019-IEEE计算机会议



(a) FairGamble-MDCE **(b) Hunters-Freq-10** **(c) CaE-Fairness**

图 4：轨迹性能结果。接近 0 表示期望的轨迹。

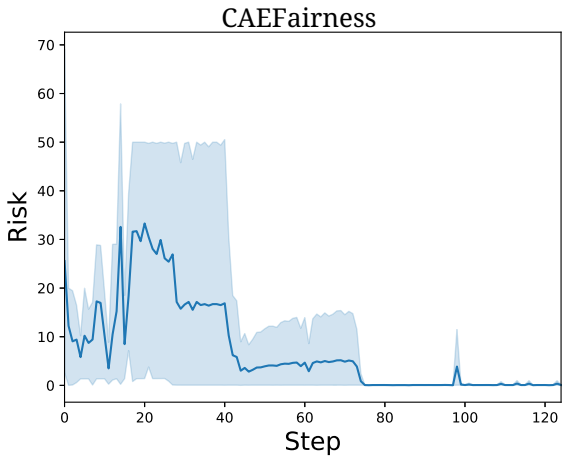


图 5：学习过程中我们 DBCE 搜索算法的误差图。以 CAE-Fairness 为例。误差图展示了算法逐渐趋于稳定的过程。

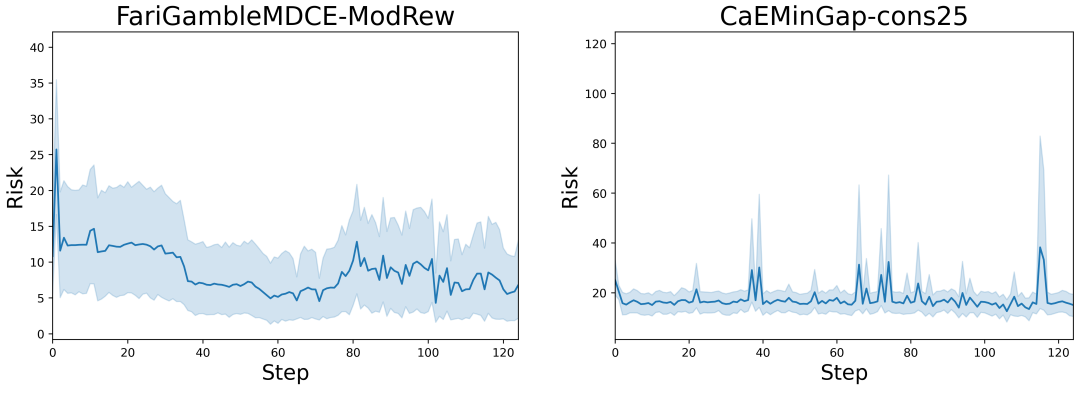
通信。IEEE，1063–1071。[23] Toryn Q Klassen、Sheila A McIlraith和Christian Muise。2022年。学习型规划模型带来的人工智能安全威胁。（2022年）。[24] Harry Markowitz。1952年。投资组合选择*。《金融杂志》7，第1期（1952年），77–91。https://doi.org/10.1111/j.1540- 6261.1952.tb01525.x arXiv:https://onlinelibrary.wiley.com/doi/pdf/10.1111/j.1540-6261.1952.tb01525.x[25] Oliver Mihatsch和Ralph Neuneier。2002年。风险敏感强化学习。机器学习49，第2期（2002年），267–290。[26] Chris Murray和Geoff Gordon。2007年。在一般和随机博弈中寻找相关均衡。卡内基梅隆大学，计算机科学学院，机器学习.....[27] Abraham Neyman。1997年。相关均衡与势博弈。《国际博弈论杂志》26，第2期（1997年），223–227。[28] Luis E Ortiz、Robert E Schapire和Sham M Kakade。2007年。最大熵相关均衡。载于人工智能与统计学。PMLR，347–354。[29] Zengyi Qin、Yuxiao Chen和Chuchu Fan。2021年。密度约束强化学习。载于国际机器学习会议。PMLR，8682–8692。[30] Anders Rantzer。2001年。李雅普诺夫稳定性定理的对偶。《系统与控制快报》42，第3期（2001年），161–168。[31] Daniel Schiff、Justin Biddle、Jason Borenstein和Kelly Laas。2020年。人工智能伦理、政策和治理的下一步是什么？全球概览。载于《.....论文集》

*AAAI/ACM*人工智能、伦理与社会会议。153–158。[32] Yun Shen、Michael J Tobia、Tobias Sommer 和 Klaus Obermayer。2014年。风险敏感型强化学习。神经计算 26, 7 (2014), 1298–1328。[33] Umar Syed、Michael Bowling 和 Robert E Schapire。2008年。利用线性规划进行学徒式学习。载于第*25*届国际机器学习会议论文集。1032–1039。[34] Csaba Szepesvári 和 Michael L Littman。1999年。基于价值函数的强化学习算法的统一分析。神经计算 11, 8 (1999), 2017–2060。[35] Pauli Virtanen、Ralf Gommers、Travis E. Oliphant、Matt Haberland、Tyler Reddy、David Cournapeau、Evgeni Burovski、Pearu Peterson、Warren Weckesser、Jonathan Bright、Stéfan J. van der Walt、Matthew Brett、Joshua Wilson、K. Jarrod Millman、Nikolay Mayorov、Andrew R. J. Nelson、Eric Jones、Robert Kern、Eric Larson、C J Carey、İlhan Polat、Yu Feng、Eric W. Moore、Jake VanderPlas、Denis Laxalde、Josef Perktold、Robert Cimrman、Ian Henriksen、E. A. Quintero、Charles R. Harris、Anne M. Archibald、Antônio H. Ribeiro、Fabian Pedregosa、Paul van Mulbregt 和 SciPy 1.0 贡献者。2020年。SciPy 1.0: Python科学计算的基础算法。《自然方法》17 (2020), 261–272。https://doi.org/10 1038 41592 019 0686 2 。/s - - - Yaodong Yang、Rui Luo、[36] Minne Li、Ming Zhou、Weinan Zhang 和 Jun Wang。2018年。平均场多智能体强化学习。载于国际机器学习会议。PMLR，5571–5580。[37] Tao Yu、HZ Wang、Bin Zhou、Ka Wing Chan 和 J Tang。2014年。面向互联电网协调智能发电控制的多智能体相关均衡 Q (λ) 学习。《*IEEE*电力系统汇刊》30, 4 (2014), 1669–1679。[38] Brian D Ziebart、Drew Bagnell 和 Anind K Dey。2010年。马尔可夫博弈的最大因果熵相关均衡。载于第二十四届*AAAI*人工智能会议研讨会。

附录*7.0.1* 更多实验结果。这里展示了 20 次运行的实验结果，以提供更全面的统计信息。

方法	Task	误差 Mean	误差 Std	MaxBF Mean	MaxBF Std	MaxReg Mean	MaxReg Std	运行时间 Mean	运行时间 Std
cons005	CaEMinGap	0.07	0.06	0.00	0.01	0.27	0.81	161.20	28.49
cons25	CaEMinGap	12.04	10.22	0.00	0.01	1.31	3.37	194.49	37.35
DBCE	CaEMinGap	4.27	8.39	0.05	0.13	0.31	0.68	255.07	42.78
cons005	CaEMDCE	0.48	0.66	0.05	0.08	3.49	7.28	256.63	33.06
cons25	CaEMDCE	13.93	10.78	0.70	3.02	0.58	1.09	171.41	27.09
DBCE	CaEMDCE	2.23	5.99	0.00	0.01	0.22	0.67	166.01	23.90
ModRew	CaEMDCE	20.70	16.81	0.42	1.75	0.07	0.13	208.10	21.92
cons005	CaEFreq-10	1.30	2.50	0.03	0.10	0.26	0.47	265.13	88.26
cons25	CaEFreq-10	10.84	5.91	0.00	0.01	0.07	0.21	211.06	33.89
DBCE	CaEFreq-10	11.35	4.78	0.04	0.14	0.03	0.05	203.83	44.49
cons005	公平赌博最小间隙	3.07	4.34	10.84	8.03	0.18	0.17	150.41	20.72
cons5	公平赌博最小间隙	8.50	6.00	6.75	3.85	0.15	0.07	177.02	53.26
DBCE	公平赌博最小间隙	23.47	16.63	0.46	0.78	0.12	0.12	266.40	31.82
cons005	FairGambleMDCE	5.69	5.93	11.25	11.79	0.19	0.21	347.57	182.65
cons5	FairGambleMDCE	6.54	5.86	9.73	12.22	0.20	0.21	297.36	128.42
DBCE	FairGambleMDCE	8.05	14.31	0.19	0.34	0.15	0.34	370.30	36.37
ModRew	FairGambleMDCE	8.72	14.34	0.24	0.32	0.18	0.35	338.67	46.01
cons005	FairGambleFreq-10	5.68	1.57	13.47	10.93	0.34	0.30	130.20	38.12
cons5	FairGambleFreq-10	6.17	1.52	8.04	9.80	0.22	0.19	195.69	67.60
DBCE	FairGambleFreq-10	9.63	7.32	0.42	0.94	0.07	0.07	322.61	43.02
cons005	HuntMinGap	1.30	1.89	0.00	0.01	1.69	1.07	153.83	8.26
cons5	HuntMinGap	1.04	1.57	0.01	0.04	1.32	1.31	154.30	12.01
DBCE	HuntMinGap	2.76	4.35	0.90	2.94	0.14	0.35	397.15	666.03
cons005	HuntMDCE	1.20	0.73	0.04	0.13	4.33	1.78	525.69	131.39
cons5	HuntMDCE	2.29	2.44	0.02	0.07	1.05	1.50	162.71	8.36
DBCE	HuntMDCE	11.49	7.07	0.01	0.03	0.03	0.05	206.20	40.72
ModRew	HuntMDCE	1.67	1.62	4.31	18.43	2.81	1.75	200.18	14.53
cons005	HuntFreq-10	2.59	4.67	0.02	0.05	0.91	1.20	435.08	37.31
cons5	HuntFreq-10	5.41	1.24	0.14	0.47	0.70	1.40	257.46	27.31
DBCE	HuntFreq-10	7.69	5.26	0.00	0.00	0.07	0.09	298.59	39.70

Bel下面我们展示图6中两种基线方法的误差图，其解读方式与图5类似。



(a) FairGamble-MDCE **(b) CaE-MinGap**

图6：constrained-25 和 ModRew 的误差图

8 博弈模型详情

模型的详情如下。

(1) 公平赌博。在这个博弈中，两名赌徒相互进行博弈，并从 3 种不同的博弈中选择一种。这 3 种博弈都极其公平，因此无论他们做什么，奖励都会随机分配。博弈 1 公平地给予 0 奖励；博弈 2 给一名赌徒 0.5，给另一名赌徒 -0.5；博弈 3 给一名赌徒 1，给另一名赌徒 -1。每回合，赌徒从 0、1、2 中选择一个数字，我们根据所选数字决定他们进行哪种博弈。解释见图 3a。

(2) 猎人。在这个博弈中，3 名猎人住在一个村庄里。每回合，他们可以待在村庄内或村庄外，并可以选择去打猎，或者守卫村庄以抵御动物，见图 3b。如果一名猎人从村庄出发去打猎，他将获得高奖励（1），其余猎人将获得低奖励（0.1）。如果一名猎人守卫村庄，所有猎人都将获得相同的中等奖励（0.5）。如果一名猎人在村庄外，并仍在外面打猎，这种行为不够安全，因此他获得的奖励会更少（0.5），而其他人将获得惩罚奖励（-0.5）。此外，如果守卫村庄的猎人不超过 1 名，他们将受到高额惩罚奖励（-3）。(3) 收集与探索。这是一个合作博弈。在这个环境中，3 个智能体被困在一片森林里，见图 3c。他们可以选择探索环境，或在住所附近收集食物。每回合只有一名智能体可以外出探索环境。如果有多名智能体选择外出，我们会随机选出其中一名外出，其余智能体将得不到奖励。每回合，如果现有智能体进行探索，我们就将奖励增加 1；如果任何智能体在附近收集食物，我们就将奖励增加 0.3。由于这是一个合作环境，我们为所有智能体设定相同的奖励。