混合动机马尔可夫博弈中的利他主义与公平目标

Yao-hua Franck Xu
Orange Labs拉尼永 拉尼永，
法国
franck.xu@orange.com

Tayeb Lemlouma
IRISA 拉尼永，法国
tayeb.le
mlouma@irisa.fr

Arnaud Braud Orange
Labs 拉尼永 拉尼永，法国
arnaud.braud
@orange.com

Jean-Marie Bonnin
IRISA 雷恩，法国
jean-
marie.bonnin@irisa.fr

arXiv:2602.08389v1 [cs.MA] 9 Feb 2026

摘要

合作是社会得以存续的基础，因为它能促使寻求集体福祉的异质群体形成结构。然而，个体往往会选择背叛，借助群体合作获益，却不承担相应成本，从而导致不公平的局面。在博弈论中，社会困境体现了个人利益与集体结果之间的这种矛盾。多智能体合作最主流的方法是功利主义福利，但它可能产生高效却极不公平的结果。

本文提出了一种新框架，以比例公平取代标准的功利主义目标，从而促进更加公平的合作。我们为每个智能体引入一种公平利他效用，并在个体对数收益空间中对其进行定义，同时推导出确保经典社会困境中合作所需的解析条件。随后，我们通过定义公平马尔可夫博弈，将该框架扩展至序贯场景，并推导出新型公平行动者-评论家算法，以学习公平策略。最后，我们在多种社会困境环境中评估了该方法。

关键词

社会困境；博弈论；马尔可夫博弈；多智能体强化学习；强化学习；公平性

1 引言

合作是支撑几乎所有多智能体系统稳定与成功的基本机制，涵盖从生物生态系统到复杂的人类社会和人工智能等领域。它是独立智能体协调动作以追求共同目标的过程，从而促成复杂结构的形成，并实现任何单个智能体独自行动都无法达到的集体福祉。有效合作的能力不仅有益，而且往往对生存、进步和资源的高效利用至关重要。

尽管持续合作至关重要，但要实现这一点却 notoriously 困难。其主要障碍在于个人私利与集体利益之间固有的张力——博弈论将这种冲突正式称为社会困境 [7]。在这些情境中，理性智能体往往受到激励，脱离群体协作以谋取个人利益。这种个体理性可能自相矛盾地导致集体失败，这种现象通常被称为公地悲剧，在这种情况下，对自利的追求会让所有人都陷入极不公平且并非最优的结果 [12]。

从概念上说，驱动这些困境的自私背叛，其对立面是利他主义——即为他人谋利的原则。在多智能体学习领域， [3, 26], 这种利他动机通常通过功利主义社会福利目标加以形式化，该目标旨在最大化所有智能体效用的（加权）总和 [20, 23, 25]。通过追求总效益最大化，这种方法鼓励智能体采取行动：这些行动虽然未必总是对自身最优，却能为整个群体带来最高效的结果，从而为解决这一困境提供直接途径。

然而，功利主义方法虽然能够带来高效的解决方案，却往往只关注总福利，而忽视奖励和牺牲是如何分配的，因此常常导致极不公平的结果。例如，严格的功利主义计算可能会得出结论：健康的人应该向陌生人捐肾，因为受赠者获得的巨大收益（挽救生命）超过了捐赠者的损失。尽管总福利实现了最大化，但这一结果却要求一个人做出极端且直觉上不公平的牺牲。这凸显了未经校准的纯粹利他主义如何与稳定合作所必需的公平原则相冲突。因此，真正可持续的合作不仅需要促进集体利益，还要求建立一种机制，确保所有参与者公平分担合作的成本与收益。

公平性最初是在资源分配问题中得到研究的 [4, 5, 17] 。在多智能体强化学习（MARL）中，公平性才刚刚成为一个活跃的研究领域，通常与合作一并探讨。在多智能体强化学习中促进公平性的常用方法是奖励塑造 [2, 8, 10, 14] 或基于角色的奖励塑造 [19]。然而，这些方法往往依赖于特定环境，只奖励即时的公平行为，并忽视完整轨迹上的公平性。尽管相关研究 [15, 21] 探讨了类似问题，但这些研究仅限于有限轨迹，且缺乏对策略梯度的深入分析。

为解决这些局限性并满足对更公平方法的需求，本文提出了一种新颖的框架，旨在明确促进异构智能体之间更加公平的合作。我们不再局限于简单的奖励最大化，而是提出了一种基于比例公平的新目标，这一概念能够有效平衡效率与公平性。

我们的贡献主要有三点。首先，我们为每个智能体引入一种公平利他效用，并在个体对数收益空间中对其进行定义，

并推导出确保经典社会困境中实现合作所需的解析条件。其次，我们通过定义具有无限时域的公平利他马尔可夫博弈，将这一框架扩展至序贯场景。最后，基于这一新的博弈形式，我们推导出新颖的公平行动者-评论家算法，旨在学习能够同时实现高集体奖励和公平结果的策略，并在各种社会困境环境中验证了我们的方法。

2 预备知识

2.1 符号

我们用 $\mathbb{R}$ 表示实数集，用 $\mathbb{R}_+$ 表示正实数集。我们将 d 维实向量空间记作 $\mathbb{R}^d$。对于集合 A，其幂集记作 $\mathrm{P}(A)$。对于任意概率测度 $\mathbb{P}$，我们将事件 $X = x$ 的概率记作 $\mathbb{P}(X = x)$。类似地，条件概率记作 $\mathbb{P}(X = x \mid Y = y)$。当所指的随机变量明确时，为简化记号，我们通常将其省略，即 $\mathbb{P}(X = x) = \mathbb{P}(x)$。

我们用 $X \sim p$ 表示随机变量 X 服从概率分布 p。对于任意随机变量 $X \sim p$，其期望记作 $\mathbb{E}_{X\sim p}[X]$。我们将可测空间 A 上的概率分布集合记作 $\Delta(\mathrm{A})$。有限集合 A 的基数记作 $|A|$。我们使用向量记号 $\vec{\ }$ 指代笛卡尔积中的任意元素；也就是说，对于任意集合 A_1、……、A_n，$\vec{a} = (a_1$、……、$a_n) \in \bigtimes_{i=1}^{n} A_i$。

2.2 正规形式博弈

我们将标准式博弈 G 定义为元组 $N, \{S_i\}_{i\in N}, \{u_i\}_{i\in N}$，其中 $N = \{1, 2, \ldots, n\}$ 是玩家集合，S_i 是玩家 i 可用的纯策略集合。令 $\mathrm{G}_{N,S}$ 为具有固定玩家集合 $N = \{1, \ldots, n\}$ 和固定纯策略集合 $S = (S_i)_{i\in N}$ 的所有标准式博弈的集合。$p_i\colon\ S \to \mathbb{R}$ 是玩家 i 的收益函数，它为每个策略组合 $s \in S$ 赋予一个实值效用。在本工作中，我们假设效用是有界的。

纳什均衡。 在博弈论中，一个基本的解概念是纳什均衡，即在其他玩家保持策略不变的情况下，没有任何玩家可以通过改变自己的策略而获益。形式上，策略组合 $s^* = (s_1^*, s_2^*, \ldots, s_n^*)$ 是纳什均衡，如果对每个玩家 $i \in N$ 以及每个替代策略 $s_i \in S_i$：

$$p_i(s_i^*, s_{-i}^*) \geq p_i(s_i, s_{-i}^*).$$

2.3 利他博弈

博弈论中的利他主义通常通过社会福利函数（SWF）来表述，它将个体效用汇总为集体财富。功利主义社会福利函数虽能最大化整体效率，却可能容许极端不平等。相比之下，“公平”的社会福利函数满足效率（帕累托改进）、公平性（庇古-道尔顿）和匿名性公理，以确保分布公平 [22]。

顾名思义，它是使社会最优结果成为稳定纳什均衡所需向每位玩家提供的最小社会福利比例。换句话说，它衡量的是需要多大的激励才能让自私者采取合作行为。这是一种衡量个人自利与集体利益之间张力的方式。这就是 α-自私博弈，也是研究最多的利他博弈 [23]。

在本文中，我们采用一种略有不同的定义 [6]。

定义 *2.1.*（α-标准式博弈的利他扩展）给定任意标准式博弈 $G = N, \{S_i\}_{i\in N}, \{p_i\}_{i\in N}$，我们可以导出一个利他博弈 $G(\alpha) \doteq \{N, \{S_i\}_{i\in N}, \{u_i\}_{i\in N}\}$，其中，$u_i(s) \doteq (1-\alpha)p_i(s) + \alpha SW(s)$。策略博弈 G 的利他主义水平定义如下：

$$\alpha_G = \inf_{\alpha}\{\alpha \in \mathbb{R}_+ | G(\alpha) \text{ is } \alpha\text{-altruistic}\} \tag{1}$$

其中，G 对于某个 $\alpha \geq 0$ 而言是 α-利他的，当且仅当 $G(\alpha)$ 的一个纯纳什均衡是 G 的社会最优。

2.4 社会困境

社会困境是一类标准式博弈：玩家在追求自身利益时，会导致整个群体的集体结果更糟。这凸显了个体偏好与集体利益之间的二元性。

更正式地说，社会困境是指收益矩阵与表1所述类似的对称双人标准式博弈。

表1：收益矩阵中的结果类别

	C	D
C	R, R	S, T
D	T, S	P, P

其中，R 是相互合作的奖励，P 是相互背叛的惩罚，S 是一方合作而另一方背叛时合作方的受骗者收益，T 是一方合作时另一方背叛所获得的背叛诱惑。

如果收益满足以下偏好不等式，则该情形被定义为社会困境。

表2：定义社会困境的一组不等式。

不等式	偏好
$R > P$	个体偏好相互合作（C, C）胜过相互背叛（D, D）。
$R > S$	个体偏好相互合作胜过单方面合作（C, D）。
$2R > T + S$	群体偏好相互合作胜过单方面相互合作/背叛。
$T > R$ or $P > S$	个体偏好单方面背叛（D, C）而非相互合作（又称贪婪），或者个体——偏好相互背叛，而非单方面合作（又称恐惧）。

2.5 比例公平

资源分配问题是指在最大化某个目标函数（例如性能、公平性或成本）的同时，将有限资源分配给竞争任务或个人的问题。比例公平（PF）目标旨在兼顾效率与某种公平性，最初是在电信领域为速率控制而提出的 [4, 9, 17]。

形式化定义：设 N 为智能体集合， $X \subseteq \mathbb{R}_+^N$为所有可行分配的集合。记 u_i： $X \to \mathbb{R}_+$为智能体 i的效用。若对于任意 $x \in X$，比例变化之和均为负，则解 $x^* \in X$是比例公平的：

$$\sum_{i=1}^{N} \frac{u_i(x) - u_i(x^*)}{u_i(x^*)} \leq 0$$

比例公平已在相关文献中得到广泛研究。特别是，已有证明表明，比例公平分配 x^*是以下优化问题的一个解：

$$\max_{x \in X} \sum_{i=1}^{N} \log u_i(x)$$

比例公平与纳什福利函数 [5, 16], 直接相关，该函数旨在最大化智能体效用的乘积；从某种意义上说，最大化效用乘积等价于最大化其对数之和。

2.6 马尔可夫博弈

在多智能体环境中对序贯决策进行建模时，标准框架是马尔可夫博弈（也称为随机博弈）。

马尔可夫博弈 M 由以下元组定义：$(N,\ \mathrm{S},\ \{\mathrm{A}_i\}_{i\in N},\ P,\ \rho_0,\ \{r_i\}_{i\in N},\ \gamma)$，其中 N 是有限玩家集合，S 是环境所有可能状态构成的空间，$\{\mathrm{A}_i\}_{i\in N}$是每个玩家各自的动作集， $P(s',\ \vec{a},\ s) = \mathbb{P}(s' \mid s,\ \vec{a})$ 是转移概率核， $\rho_0 \in \Delta(\mathrm{S})$ 是初始状态分布，$\{r_i\}_{i\in N}$是每个玩家各自的奖励， $\gamma \in [0, 1)$ 是折扣因子。在本文中，我们假设状态空间和动作空间是有限的，且奖励函数为正并有界于 $R_{\max}$。

3 方法论

在本节中，我们通过重新缩放收益函数，使其采用比例公平函数，从而扩展 α-利他博弈。随后，我们研究这种变换对简单博弈的影响，尤其是对双人社会困境的影响。

3.1 重新定义利他博弈

先前对 α-利他博弈的定义无法直接与比例公平函数兼容，因为将个体奖励与对数奖励相加并非齐次运算。为解决这一问题，我们对自利奖励应用函数 F_i，使其与社会福利函数处于同一尺度。例如，对于比例公平，我们使用 F_i =log。此外，对数函数将所考虑的数值限制为严格正实数值。因此，我们给出如下替代定义，

定义*3.1*。（标准式博弈的 α-利他扩展）给定任意标准式博弈 $G = N, \{S_i\}_{i\in N}, \{p_i\}_{i\in N}$，我们可以导出一个公平利他博弈 $G(\alpha) \doteq \{N, \{S_i\}_{i\in N}, \{u_i\}_{i\in N}\}$，其中，$u_i(s) \doteq (1-\alpha)F_i(p_i(s) - m_p) + \alpha SW(s)$，其中 $m_p = \inf_{i,s}[u_i(s)]$是玩家在所有可能策略下可获得的最低收益（假定该值存在），而 F_i： $\mathbb{R} \to \mathbb{R}$ 是作用于玩家收益的递增函数。策略博弈 G 的利他主义水平定义如下：

$$\alpha_G = \inf_{\alpha} \{\alpha \in \mathbb{R}_+ | G \text{ is } \alpha\text{-altruism}\} \tag{2}$$

其中， G是 α-利他主义，当且仅当存在某个 $\alpha \geq 0$，使得 $G(\alpha)$的一个纯纳什均衡是 G 的社会最优。

在本文余下部分中，我们对所有玩家均设定 $F_i = \log i$。

3.2 社会困境分析

令 $G \in \mathrm{G}_{N,\ S}$ 为一个标准式博弈，其中 N 是玩家集合，S是策略组合集合， R是收益函数，其中$R_i(s)$是玩家 i的收益。我们假设所有收益均严格为正，即对于所有 $i,\ s,\ R_i(s) > 0$。

我们定义一个变换后的博弈 $G'(\alpha) = N,\ S,\ R'$，其变换后的收益函数 R' 定义如下：

$$R'_i = (1-\alpha)\log R_i + \alpha \sum_{j\in N} \log R_j = \log R_i + \alpha \sum_{j\neq i} \log R_j$$

现在，我们运用这一框架，推导囚徒困境、猎鹿博弈和胆小鬼博弈中 α 的阈值。

对于双人博弈（$N = 2$），玩家 1 和玩家 2 的规则可简化为

$$R'_1 = \log R_1 + \alpha \log R_2,$$

$$R'_2 = \log R_2 + \alpha \log R_1.$$

参数 α 表示收益相互依赖的程度。如果 $\alpha = 0$，博弈保持不变（$R'_i = R_i$）。随着 α 增大，一名玩家的变换后收益会越来越多地受到另一名玩家原始收益的影响。

我们的分析将确定 α 的条件，使相互合作结果（$C,\ C$）在 $G'(\alpha)$ 中成为均衡。要实现这一点，任何玩家都不能通过单边偏离获利。对于对称双人博弈，我们只需确保在玩家 2 合作的情况下，玩家 1 不会因背叛而获益。该条件为

$$R'_1(C, C) \geq R'_1(D, C).$$

定理 3.2。 社会困境的利他主义水平为

$$\alpha_G = \begin{cases} 0 & if\ T \leq R, \\ \frac{\log T - \log R}{\log R - \log S} & if\ T > R. \end{cases} \tag{3}$$

证明。在变换后的博弈中，收益变为

$$R'_2(C, C) = R'_1(C, C) = (1+\alpha)\log R$$
$$R'_1(D, C) = \log T + \alpha \log S$$
$$R'_2(D, C) = \log S + \alpha \log T$$

要使（$C,\ C$）成为纳什均衡，必须满足条件 $R'_1(C,\ C) \geq R'_1(D,\ C)$。

囚徒困境。 囚徒困境由偏好顺序 $T > R > P > S$定义。该困境在于，尽管相互合作（$D,\ D$）是帕累托占优结果，但相互背叛（$C,\ C$）却是唯一的纳什均衡。

该条件变为 $(1+\alpha)\log\ R \geq \log T + \alpha \log S$。对该不平等求解 α，可得到以下条件

$$\alpha > \frac{\log T - \log R}{\log R - \log S}$$

猎鹿博弈。 猎鹿博弈由偏好顺序$R > P > S$和$T \leq R$定义。该博弈有两个纯策略均衡：(C, C) 和 (D, D)。合作结果已经是一个均衡。因此，分析考察该变换是否保持这一均衡。

该条件与上文相同：$R_1'(C,C) \geq R_1'(D,C)$。由此得到的不等式也为 $(1+\alpha)\log R \geq \log T + \alpha \log S$。然而，其含义有所不同。在猎鹿博弈中，$T \leq R$，因此 $\log T - \log R$为负，而 $\log R - \log S$为正。比率$\frac{\log T - \log R}{\log R - \log S}$因此为负。由于 α 始终为正，因此对任意 $\alpha \in [0,1]$，该条件都成立。这证实了这种变换永远不会改变合作均衡。

胆小鬼博弈。胆小鬼博弈由偏好顺序 $T > R > S \geq P$ 定义。相互合作（C,C）不是均衡，因为每个玩家都受到诱惑，想要背叛以获得最高收益。

要使（C,C）成为纳什均衡，必须满足以下条件

$$\alpha > \frac{\log T - \log R}{\log R - \log S}$$

因此，只要满足以下条件，合作就会成为纳什均衡

$$\alpha = \begin{cases} 0 & \text{if } T \leq R, \\ \frac{\log T - \log R}{\log R - \log S} & \text{if } T > R. \end{cases}$$

为确保与利他主义水平 $\alpha \leq 1$ 的定义一致，我们要求 $TS \leq R^2$，这对标准收益矩阵成立。

□

在合作最初并不稳定的各个情形（囚徒困境、斗鸡博弈）中，α 所需的阈值是对数收益的比率。分子 $\log(T/R)$ 表示背离合作状态而选择背叛的“诱惑”。分母则反映了背叛后果。这种变换实际上使玩家能够将其动作带来的外部性内化，而 α 的阈值则量化了使合作符合个体理性所需的确切内化程度。

在传统的静态社会困境中，比例公平与利他主义的结合显而易见。然而，标准式博弈的简单性也构成了局限。事实上，现实世界的现象很少只在单一步骤中发生，而是随着时间推移不断演变。要准确建模此类动态系统，分析必须超越这种形式体系，转向马尔可夫博弈这样的框架。

4 公平利他主义马尔可夫博弈

本节提出并分析一种新颖的目标函数，在马尔可夫博弈框架中融合比例公平与利他主义，并推导策略梯度定理[1]。

4.1 价值函数

为了将我们的方法扩展到马尔可夫博弈，我们必须针对给定的策略组合 $\vec{\pi}$，建立与玩家静态收益函数相对应的动态类比。这可以通过单个玩家的状态价值函数来实现，该函数表示该玩家在所有可能的无限轨迹上获得的折扣奖励累积期望。

[1]我们在附录A中给出了更详细的证明。

定义状态价值函数的一项挑战，是要在整个无限轨迹集合 $\mathrm{T} = (\mathrm{S} \times \mathrm{A})^{\infty}$ 上构造一个统一一致的概率测度 $\mathbb{P}_{\vec{\pi}}$。我们通过定义任意有限序列的概率，再利用 Ionescu-Tulcea定理将其扩展到无限情形。

首先，为每个智能体 $i \in N$以及给定轨迹 $\tau = (s_0, \vec{a}_0, s_1, \vec{a}_1, \ldots \ldots) \in \mathrm{T}$ 定义折扣回报 R_i

$$R_i(\tau) = \sum_{t=0}^{\infty} \gamma^t \, r_i(s_t, \vec{a}_t).$$

由于奖励函数有界，折扣回报绝对收敛。

智能体 i 的状态价值函数是给定初始状态 $s_0 \in \mathrm{S}$ 时关于 $\mathbb{P}_{\vec{\pi}}$ 的期望回报，

$$V_i^{\vec{\pi}}(s) = \mathop{\mathbb{E}}_{\tau \sim \mathbb{P}_{\vec{\pi}}} \left[R_i(\tau) \mid s_0 = s \right].$$

类似地，智能体 i 的动作价值函数定义为给定初始状态和第一个联合动作时的期望折扣回报，

$$Q_i^{\vec{\pi}}(s, \vec{a}) = \mathop{\mathbb{E}}_{\tau \sim \mathbb{P}_{\vec{\pi}}} \left[R_i(\tau) \mid s_0 = s, \, \vec{a}_0 = \vec{a} \right]$$

在马尔可夫博弈中，贝尔曼算子是多智能体强化学习中的一个基本概念，用于定义特定智能体在某个状态下的价值。

定义 *4.1.* 在具有智能体集合 N 和联合平稳策略 $\vec{\pi}$ 的马尔可夫博弈中，智能体 i的贝尔曼算子记作T_i^{π}，作用于任意状态价值函数 $V_i(s)$，其定义如下：

$$(T_i^{\vec{\pi}} V_i)(s) = \sum_{\vec{a} \in \mathcal{A}} \pi(\vec{a} \mid s) \left(r_i(s, \vec{a}) + \gamma \sum_{s' \in \mathcal{S}} \mathbb{P}(s' \mid s, \vec{a}) V_i(s') \right).$$

接下来，我们可以写出贝尔曼定理，该定理给出了一个关键方程，由 V_i^{π} 验证：

定理 4.2. 对于任意联合策略 $\vec{\pi}$ 以及与 $\vec{\pi}$ 相关联的任意状态价值函数，以下称为贝尔曼方程的等式成立

$$T_i^{\vec{\pi}} V_i^{\vec{\pi}} = V_i^{\vec{\pi}}$$

4.2 公平目标

基于传统的马尔可夫博弈框架，我们可以引入一个兼顾公平性与效率的新目标，并将利他型正则博弈扩展到马尔可夫博弈。

定义 *4.3.* 对于每个智能体 i，我们定义比例公平状态价值函数：

$$V^{\vec{\pi},\,\mathrm{Prop}}(s) = \sum_{j=1}^{n} \log V_j^{\vec{\pi}}(s).$$

状态价值函数发挥收益函数的作用，而比例公平函数则应用于状态价值函数。因此，我们可以根据利他主义水平为每个智能体定义一个新目标：

定义 *4.4*。 每个智能体的目标是最大化

$$J_i(\vec{\pi}) = \mathop{\mathbb{E}}_{s_0\sim\rho_0}\left[(1-\alpha)\log V_i^{\vec{\pi}}(s_0) + \alpha V^{\vec{\pi},\,\mathrm{Prop}}(s_0)\right]$$
$$= \mathop{\mathbb{E}}_{s_0\sim\rho_0}\left[\sum_{j=1}^{N} c_i(j)\log V_j^{\vec{\pi}}(s_0)\right]$$

其中

$$c_i(j) = \begin{cases} 1 & \text{if } j = i, \\ \alpha & \text{if } j \neq i. \end{cases}$$

4.3 策略梯度

策略梯度方法是一类强化学习算法，它们直接学习并优化参数化策略 $\pi_\theta(a\,|\,s)$，而无需借助价值函数。其核心原则是调整策略的参数 θ，提高能够带来高累积奖励的动作的概率，并降低导致低奖励的动作的概率。

假设智能体 i 的策略 π_i 由 θ_i 参数化，使得 π_{θ_i} 关于 θ_i 可微。

因此，智能体 i 的目标是最大化目标 $J_i(\theta) = \mathbb{E}_{s_0\sim\rho_0}[\sum_{j=1}^{N} c_i(j)\log V_j^{\vec{\theta}}(s_0)]$。我们可以证明该目标函数关于任意参数 θ 的可微性，具体如下述定理所述。

定理 4.5。 （策略梯度定理。）设 $\mathcal{M}$ 为一个马尔可夫博弈，且 $J_i(\vec{\theta}) = \mathbb{E}_{s_0\sim\rho_0}\left[\sum_{j=1}^{N} c_i(j)V_j^{\vec{\theta}}(s_0)\right]$ 为前文定义的目标；对于所有智能体 i，假设在温和条件下，Q_i^{π} 关于 θ_i 可微。J_i 关于 θ_i 的梯度为

$$\nabla_{\theta_i} J_i(\vec{\theta}) = \mathop{\mathbb{E}}_{\tau\sim\mathbb{P}^{\vec{\theta}}}\left[\sum_{t=0}^{\infty}\sum_{j=1}^{N} c_i(j)\frac{Q_j^{\vec{\theta}}(s_t,\,\vec{a}_t)}{V_j^{\vec{\theta}}(s_0)}\nabla_{\theta_i}\log\pi_{\theta_i}(a_{i,t}\mid s_t)\right]$$

证明。让我们使用链式法则推导智能体 i 的目标，

$$\nabla_{\theta_i} J_i(\vec{\theta}) = \mathop{\mathbb{E}}_{s_0\sim\rho_0}\left[\sum_{j=1}^{N} c_i(j)\frac{\nabla_{\theta_i}V_j^{\vec{\theta}}(s_0)}{V_j^{\vec{\theta}}(s_0)}\right]$$

接下来，我们需要求出梯度 $\nabla_{\theta_i}V_j^{\vec{\theta}}(s_0)$。为此，我们从贝尔曼方程入手，对其求导，并在状态价值函数梯度空间上定义一个压缩映射，然后应用巴拿赫-皮卡不动点定理，得到梯度的公式。

1. 对贝尔曼方程求导。r_i, P 和策略 π_{θ_j} for $j \neq i$ do not depend on θ_i。因此，可以使用乘积法则对贝尔曼方程求导，得到一个递归方程：

$$\nabla_{\theta_i}V_j^{\vec{\theta}}(s) = \sum_{\vec{a}\in\mathcal{A}}\nabla_{\theta_i}\pi_{\vec{\theta}}(\vec{a}\,|\,s)\cdot\left(r_j(s,\vec{a}) + \gamma\sum_{s'\in\mathcal{S}}\mathbb{P}(s'|s,\vec{a})V_j^{\vec{\theta}}(s')\right)$$
$$+\ \pi_{\vec{\theta}}(\vec{a}\,|\,s)\cdot\gamma\sum_{s'\in\mathcal{S}}\mathbb{P}(s'|s,\vec{a})\nabla_{\theta_i}V_j^{\vec{\theta}}(s')$$

我们定义函数 $G_{i,j}^{\theta}(s)$，它表示状态 s 处的即时梯度分量

$$G_{i,j}^{\vec{\theta}}(s) := \sum_{\vec{a}\in\mathcal{A}}\pi_{\vec{\theta}}(\vec{a}|s)(\nabla_{\theta_i}\log\pi_{\theta_i}(a_i|s))Q_j^{\vec{\theta}}(s,\vec{a})$$
$$= \mathbb{E}_{\vec{a}\sim\vec{\pi}_\theta}[\nabla_{\theta_i}\log\pi_{\theta_i}(a_i|s)Q_j^{\vec{\theta}}(s,\vec{a})]$$

由此，我们可以将递归方程简化为

$$\nabla_{\theta_i}V_j^{\vec{\theta}}(s) = G_{i,j}^{\vec{\theta}}(s) + \gamma\sum_{\vec{a}\in\mathcal{A}}\pi_{\vec{\theta}}(\vec{a}|s)\sum_{s'}\mathbb{P}(s'|s,\vec{a})\nabla_{\theta_i}V_j^{\vec{\theta}}(s')$$

2. 定义价值函数梯度的贝尔曼算子

该方程定义了一个不动点关系。令

$g_j \quad S \to$: 为一个梯度

$\mathbb{R}^{|\theta_j|}$向量场。我们可以将价值函数梯度的贝尔曼算子定义为 $\mathfrak{T}_{i,\,j}^{\theta}$，如下所示：

$$(\mathfrak{T}_{i,j}^{\vec{\pi}}g_j)(s) := G_{i,j}^{\vec{\theta}}(s) + \gamma\mathbb{E}_{\vec{a}\sim\pi_{\vec{\theta}}(\cdot|s),\,s'\sim P(\cdot|s,\vec{a})}[g_j(s')]$$

我们要寻找的梯度 $\nabla_{\theta_i}V_j^{\theta}$ 是该算子的不动点

$$\nabla_{\theta_i}V_j^{\vec{\theta}} = \mathfrak{T}_{i,j}^{\vec{\theta}}(\nabla_{\theta_i}V_j^{\vec{\theta}})$$

3. 马尔可夫博弈策略梯度定理

之前的算子 $\mathfrak{T}_{i,\,j}^{\vec{\theta}}$ 在状态价值函数梯度空间上定义了一个压缩映射。根据**巴拿赫-皮卡不动点定理**，它具有唯一不动点。这证明了梯度向量场 $\nabla_{\theta_i}V_j^{\vec{\theta}}$ 是唯一的。此外，**巴拿赫-皮卡不动点定理**还指出，对于任意序列 $(g_n)_{n\in\mathbb{N}}$，其中 $g_{n+1} = \mathfrak{T}_{i,\,j}^{\theta}(g_n)$ 以及 $g_0 : S \to \mathbb{R}^{|\theta_j|}$，$(g_n)_{n\in\mathbb{N}}$ 均收敛 $\nabla_{\theta_i}V_j^{\vec{\theta}}$。

我们可以通过展开递归并从 $g_0 = 0$ 开始来求解。由于 g_n 收敛到 $\nabla_{\theta_i}V_j^{\vec{\theta}}$，以下等式成立

$$\nabla_{\theta_i}V_j^{\vec{\theta}}(s_0) = g_\infty(s_0) = \mathop{\mathbb{E}}_{\tau\sim\mathbb{P}^{\vec{\theta}}}\left[\sum_{t=0}^{\infty}\gamma^t G_{i,j}^{\vec{\theta}}(s_t)\,\middle|\,s_0\right]$$

代入 $G_{i\,\text{的定义}\,j}^{\theta}(s_t)$：

$$\nabla_{\theta_i}V_j^{\vec{\pi}}(s_0) = \mathop{\mathbb{E}}_{\tau\sim\mathbb{P}^{\vec{\theta}}}\left[\sum_{t=0}^{\infty}\gamma^t\nabla_{\theta_i}\log\pi_{\theta_i}(a_{i,t}|s_t)Q_j^{\vec{\theta}}(s_t,\vec{a}_t)\,\middle|\,s_0\right]$$

我们可以简化这个方程，并将 $\nabla_{\theta_i}V_j^{\pi}(s_0)$ 的表达式代入前一个梯度，得到**公平策略梯度定理**

$$\nabla_{\theta_i}J_i(\vec{\theta}) = \mathop{\mathbb{E}}_{\tau\sim\mathbb{P}^{\vec{\theta}}}\left[\sum_{t=0}^{\infty}\gamma^t\nabla_{\theta_i}\log\pi_{\theta_i}(a_{i,t}|s_t)\left(\sum_{j=1}^{N}c_i(j)\frac{Q_j^{\vec{\theta}}(s_t,\vec{a}_t)}{V_j^{\vec{\theta}}(s_0)}\right)\right].$$

□

在强化学习中，一种常见的稳定学习的方法是在梯度中加入基线。同样的方法也可以应用于我们的情况。

定理 4.6。（公平利他优势策略梯度）。梯度可以改写为

$$\nabla_{\theta_i} J_i(\vec{\theta}) = \mathop{\mathbb{E}}_{\tau\sim\mathbb{P}^{\vec{\pi}_\theta}}\left[\sum_{t=0}^{\infty} A_i^{F,\vec{\theta}}(s_t,\vec{a}_t,s_0)\nabla_{\theta_i}\log\pi_{\theta_i}(a_{i,t}\mid s_t)\right]$$

where $A_i^{F,\vec{\theta}}(s,\vec{a},s') = \sum_{j=1}^{N} c_i(j)\frac{A_j^{\vec{\theta}}(s,\vec{a})}{V_j^{\vec{\theta}}(s')}$ *is the fair altruistic advantage function of agent i.*

证明。证明的关键在于说明，与 s_t 的任意基线函数相关的梯度项的期望值为零。考虑给定状态 s_t 下，在单个时间步 t 处该基线项的期望。此期望是对从联合策略 $\vec{\pi}(\cdot|s_t)$ 中抽取的联合动作 $\vec{a}_t$ 求得的。

我们要证明，对于任意 $f\colon \mathrm{S} \to \mathbb{R}$,

$$\mathbb{E}_{\vec{a}_t\sim\pi_{\vec{\theta}}(\cdot|s_t)}\left[\nabla_{\theta_i}\log\pi_{\theta_i}(a_{i,t}|s_t)f(s_t)\right] = 0$$

注意，$f(s_t)$ 不依赖于 $\vec{a}_t$，因此可以将其移到期望符号之外，

$$f(s_t)\cdot\mathbb{E}_{\vec{a}_t\sim\pi_{\vec{\theta}}(\cdot|s_t)}\left[\nabla_{\theta_i}\log\pi_{\theta_i}(a_{i,t}|s_t)\right].$$

期望中的内层函数仅取决于智能体 i 的动作。因此，我们可以从期望中消去其他动作的影响，

$$f(s_t)\cdot\mathbb{E}_{a_{i,t}\sim\pi_{\theta_i}(\cdot|s_t)}\left[\nabla_{\theta_i}\log\pi_{\theta_i}(a_{i,t}|s_t)\right].$$

展开期望

$$f(s_t)\sum_{a_{i,t}\in\mathcal{A}_i}\pi_{\theta_i}(a_{i,t}|s_t)\nabla_{\theta_i}\log\pi_{\theta_i}(a_{i,t}|s_t).$$

使用对数导数技巧

$$\begin{aligned}\sum_{a_{i,t}}\pi_{\theta_i}(a_{i,t}|s_t)\nabla_{\theta_i}\log\pi_{\theta_i}(a_{i,t}|s_t) &= \sum_{a_{i,t}}\nabla_{\theta_i}\pi_{\theta_i}(a_{i,t}|s_t)\\ &= \nabla_{\theta_i}\sum_{a_{i,t}}\pi_{\theta_i}(a_{i,t}|s_t)\\ &= \nabla_{\theta_i}(1)\\ &= 0.\end{aligned}$$

由于基线项的期望值为零，将其从原始策略梯度定理中期望内的项中减去，不会改变总期望。在本工作中，我们使用基线 $f(s_t) = \sum_{j=1}^{N} c_i(j)V_j^{\vec{\theta}}(s_t)$，然后我们得到公平策略梯度的优势形式

$$\nabla_{\theta_i} J_i(\vec{\theta}) = \mathop{\mathbb{E}}_{\tau\sim\mathbb{P}^{\vec{\theta}}}\left[\sum_{t=0}^{\infty} A_i^{F,\vec{\theta}}(s_t,\vec{a}_t,s_0)\nabla_{\theta_i}\log\pi_{\theta_i}(a_{i,t}\mid s_t)\right]$$

□

特别是，当 $\alpha = 1$ 时，所有智能体都试图最大化同一个目标

$$J_i(\vec{\theta}) = \mathop{\mathbb{E}}_{s_0\sim\rho_0}\left[V^{\vec{\theta},\,\mathrm{Prop}}(s_0)\right]$$

而策略梯度变为

$$\nabla_{\theta_i} J_i(\vec{\theta}) = \mathop{\mathbb{E}}_{\tau\sim\mathbb{P}^{\vec{\pi}_\theta}}\left[\sum_{t=0}^{\infty} A^{F,\vec{\theta}}(s_t,\vec{a}_t,s_0)\nabla_{\theta_i}\log\pi_{\theta_i}(a_{i,t}\mid s_t)\right]$$

其中 $A^{F,\vec{\theta}}(s,\vec{a},s') = \sum_{j=1}^{N} A_j^{\theta}(s,\vec{a})V_j^{\vec{\theta}}(s')$ 是公平优势函数。

根据该策略梯度定理，我们可以推导出学习公平策略的多智能体 Actor-Critic 算法。我们试图通过比较多智能体系统在一个著名环境中测得的公平性与效率来验证我们的方法。

5 实验

本节介绍实验。简而言之，我们研究在不同利他主义水平下，公平目标对智能体性能和群体公平性的影响。实验采用一种广受欢迎的混合动机博弈，名为“CleanUp”。

5.1 环境

“CleanUp”环境首次于 [13] 提出，如今已成为 Melting Pot 2.0 [1] 的一部分。它是一种常用的多智能体基底，旨在探索复杂的社会困境。该环境构建为一款资源管理游戏，多个智能体必须合作，维护一个可再生资源池。在“CleanUp”中，智能体必须尽可能多地采摘苹果。智能体只有在采摘一个苹果时才能立即获得正奖励。

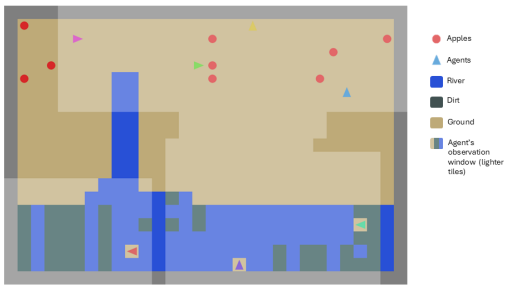

图 1：“CleanUp”环境。较浅色的方格表示智能体的观测窗口。

然而，苹果的生长率取决于河流的洁净程度。污染水平过高时，生长率会降至零，无法再长出苹果。为防止这种情况，智能体必须清理河流、减少污染，从而使所有玩家受益。因此，智能体必须牺牲短期收益以换取长期奖励，因为它们无法同时采摘苹果和清理河流。这一设置迫使智能体应对典型的社会困境：在个体奖励最大化（采摘苹果）的自利动机与长期可持续性（清理河流）的集体需求之间进行权衡，从而检验其合作能力和亲社会行为。

5.2 设置

环境采用部分可观测性设置。智能体只能看到世界的一部分，即以自身位置为中心、大小为

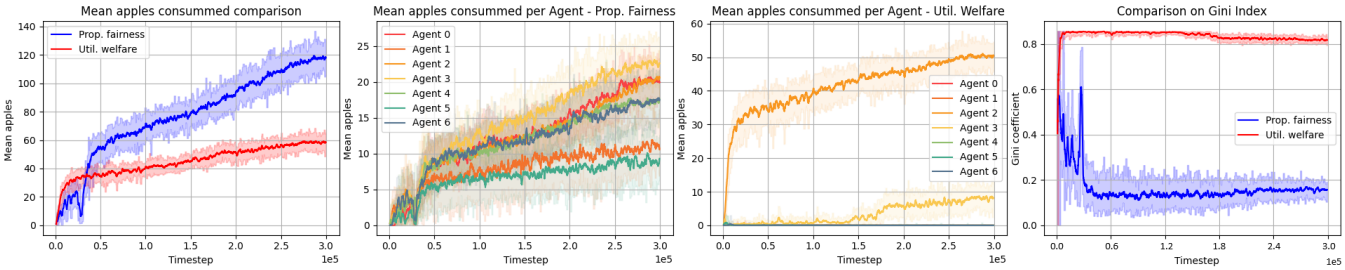

图2：在完全合作设置下，使用多智能体近端策略优化训练的比例公平性与效用福利目标对比（$\alpha = 1$）。实线表示对滚动窗口内50 次运行的指标取平均所得。阴影区域表示同一滚动窗口内相应指标的最小值和最大值。

11×11像素的正方形窗口。在本研究中，我们为每个环境配置了 7 个智能体，并为它们随机设置初始位置。

我们的代码库[2] 采用了 SocialJax 框架提供的基于 Jax 的“CleanUp”环境 [11]。我们使用分散式行动者-评论家算法，每个智能体都由各自的一组神经网络建模，以确保参数不共享（即独立学习）。这种非共享参数化使我们能够模拟更广泛的策略组合，更充分地刻画 CleanUp 社会困境的复杂性。具体而言，我们使用两个卷积层处理观测的空间特性，随后接入两个全连接层。对于每个智能体，行动者网络表示参数化策略，而评论家网络则估计其个体状态价值函数。

我们在 10 个并行环境中运行训练算法。回合长度设为 100 个时间步。我们特意将回合长度设得较短，以研究初始位置对学习过程的影响。即使采用随机策略，出生在苹果区域附近的玩家也比靠近河流的玩家拥有初始优势。总时间步数设为 3×10^5。PPO 类算法的学习率设为 0.001 [27]，A2C 类算法的学习率设为 0.01；两者都采用衰减计划，使学习率最终降至 0.00001。对于 PPO 类算法，我们利用熵正则化来改善探索效果。

在实验中，我们研究了以下算法的性能。

公平 *MAA2C*。 这是对策略梯度定理的一种直接应用。每个智能体的行动者都根据其个体_行动者损失进行更新，该损失由 $\mathbb{E}_t[A^{F,\theta}_{i,t} \log \pi_{\theta_i}(a_{i,\ t} \mid s_t)]$ 给出。

公平 *MAPPO*。 它基于 MAPPO 构建。我们保留了 MAPPO 的核心思想，并在近端策略优化损失中以公平优势函数取代经典优势函数，尽管这在数学上并不严谨。

为了评估我们算法的性能，我们采用两项指标。第一项用于评估效率，定义为一个回合内消耗的苹果总数。第二项用于评估智能体之间的公平性，定义为所消耗苹果的基尼系数 $Gini_t$，其中

$$Gini_t(c_0, \ldots, c_n) = \frac{\sum_{i \in N} \sum_{j \in N} |c_i - c_j|}{2N \sum_{i \in N} c_i}$$

其中，c_i 表示智能体 i 在一个回合内消耗的苹果数量。基尼系数介于 0.0 和 1.0 之间。数值越低，分布越公平；当分布均匀时，基尼系数为 0.0。反之，数值越高，分布越不平等。

我们还深入研究了完全合作设置，并比较比例公平（PF）目标与功利主义福利（UW）目标。

5.3 结果

图2展示了完全合作设置（$\alpha = 1$）下 PF 目标与 UW 目标的比较。令人惊讶的是，比例公平性方法展现出更高的整体效率。该群体消耗的苹果总数稳步增加，在训练结束时达到约 120 个。相反，以功利主义为导向的方法所产生的苹果总消耗量显著较低，并稳定在 40 个左右。

细分总体消费情况后可以发现，在 PF 设置中，七个智能体消耗的苹果数量都比较接近。尽管存在细微差异，但所有智能体的趋势线都聚集在一起。没有任何一个智能体掉队，所有智能体似乎都为这项公平的任务作出贡献并从中受益，这一点体现为较低的基尼系数。与此形成鲜明对比的是，在 UW 设置中，只有两个智能体（2 和 3）学会了采摘苹果，其余智能体则在清理河流或无所事事。因此，这种结果造成了极不平等的局面，其特点是基尼指数非常高，接近 0。8。

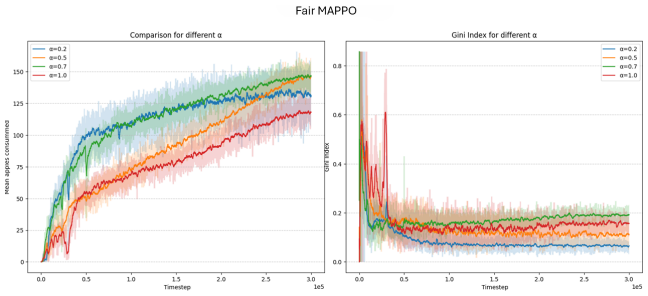

图 3：在 CleanUp 环境中使用 Fair-MAPPO 训练的不同 α 的性能。

[2]代码可在 https://github.com/AkuBrains/altruistic-fair-MARL/ 获取。

在功利主义设置中，每个智能体都能独立于自身采取的动作，从其他智能体的奖励信号中获益。因此，智能体获得的奖励与其动作不一定相关。智能体陷入次优策略，为每个智能体形成了专门化角色。当智能体共享相同参数时，它们基本上会在假设存在自然合作和同质行为的情况下学习单一策略。在 PF 设置中，PF 目标带来的自然公平性会在出现不平衡时鼓励智能体进一步探索。在公平优势函数中，经典优势函数由状态价值函数进行归一化，从而显著降低了其他智能体的奖励信号在 PF 目标中的权重。

图 3 展示了公平 MAPPO 的结果。所有智能体群体都成功学会了随着时间推移采集更多苹果，这一点从所有曲线的上升趋势中可见一斑。此外，所有实验的基尼指数都相当低，所有运行结果均集中在 0.05 到 0.2 之间。 $\alpha = 0.7$ 所在的群体总体性能最佳，在训练结束时消耗的苹果最多。出乎意料的是，表现最差的是完全合作群体，其总体性能和基尼指数的得分都显著较低。

直观来看，在完全合作设置中，存在与 UW 设置相同的现象，但 PF 目标所体现的公平意愿削弱了这一现象。相反，当 $\alpha < 1$时，智能体更倾向于遵循自身目标，并且更可能自主采摘苹果。当 α过低（$\alpha = 0.2$）时，整体采摘性能会下降。在这种情况下，智能体只是自私自利，不太愿意相互协调来清理河流。

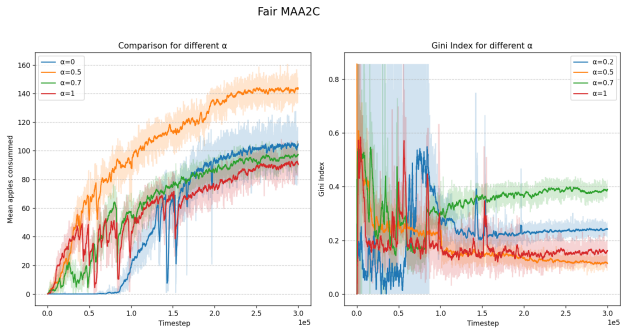

图 4：使用公平 MAA2C 在 CleanUp 环境中训练时，不同 α 的性能。

图 4 展示了公平 MAA2C 的结果。从所有曲线的上升趋势可以看出，所有智能体群体都成功学会了随着时间推移收集更多苹果。然而，训练过程中的噪声很大，训练期间苹果平均消耗量多次急剧下降便说明了这一点。众所周知，MAA2C 训练过程非常不稳定，而我们的公平 MAA2C 也呈现出同样的情况，因此结果难以解读。尽管如此，$\alpha = 0.5$ 对应的结果与公平 MAPPO 的结果相符，其功利主义表现和基尼指数均为最佳。此外，公平 MAA2C 的苹果平均消耗量似乎低于其近端策略优化版本， $\alpha = 0.5$ 除外，这可能是学习过程中的不稳定性所致。

简而言之，自私行为可能带来更平等的结果，但这只是因为每个人都在争夺相同的个人奖励，而这可能会牺牲群体的整体成功。这凸显了个体自私与集体目标之间根本性的权衡。我们还证明，PF 目标所促进的公平性能够实现可持续合作，而功利主义方法则无法做到这一点，尤其是在异质群体中。

5.4 讨论与后续步骤

我们的算法在理论证明方面仍有待改进。特别是，即使均衡存在，我们也尚未证明算法会收敛到任何类型的均衡。此外，用公平优势函数替代传统优势函数与理论并不一致，因为我们在这种情况下并未证明信赖域策略优化的单调改进定理 [24, 29]。

在实验方面，值得在更多环境中测试我们的算法 [1, 11]，或在更贴近现实的真实世界模拟中进行测试 [28]，并为 CleanUp 创建更多样化的场景。当苹果再生率为 0.05 时，仅有 7 个智能体所施加的采收压力太低，无法造成供需之间的显著失衡。一个恰当的场景是增加智能体数量并降低再生率，从而形成资源稀缺、需求大于供给的局面。

在功利主义设定中改进我们实验的另一种方式，是使用 HAPPO 或 HAA2C 等最新的最先进算法 [18]。

6 结论

本文提出了一个新颖的框架，将比例公平性与利他主义相结合，以促进多智能体系统中更加公平的合作。我们提出了公平利他效用，并通过定义公平利他马尔可夫博弈，将其扩展到序贯场景，由此推导出新颖的公平策略梯度算法。

我们在“CleanUp”社会困境中的实验表明，与采用异构智能体的传统功利主义福利目标相比，该方法不仅实现了显著更加公平的结果（以基尼指数衡量），还带来了更高的整体效率。通过有效平衡个体激励与集体福祉，我们的方法避免了智能体采取高度专业化且不公平的角色，从而为更加稳定和可持续的合作铺平了道路。

参考文献

[1] Agapiou、Alexander Sasha Vezhnevets、Edgar A. Duéñez-Guzmán、Jayd Matyas、Yiran Mao、Peter Sunehag、Raphael Köster、Udari Madhushani、Kavya Kopparapu、Ramona Comanescu、D. J. Strouse、Michael B. Johanson、Sukhdeep Singh、Julia Haas、Igor Mordatch、Dean Mobbs 和 Joel Z. Leibo。2023。《Melting Pot 2.0》。https://doi.org/10.48550/arXiv.2211.13746 arXiv:2211.13746 [cs]。[2] Farinaz Alamiyan-Harandi 和 Pouria Ramazi。2023。《基于环境影响的多智能体强化学习》。https://doi.org/10.48550/arXiv。2311.04240 arXiv:2311.04240 [cs]。[3] Stefano V. Albrecht、Filippos Christianos 和 Lukas Schäfer。2024。多智能体强化学习：基础与现代方法。麻省理工学院出版社。https: //www.marl-book.com[4] Siddhartha Banerjee、Vasilis Gkatzelis、Safwan Hossain、Billy Jin、Evi Micha 和 Nisarg Shah。2022。《带预测的公共物品比例公平在线分配》。https://doi.org/10.48550/arXiv.2209.15305 arXiv:2209.15305 [cs]。[5] Ioannis Caragiannis、David Kurokawa、Hervé Moulin、Ariel D. Procaccia、Nisarg Shah 和 Junxing Wang。2019。《最大纳什的非理性公平性》

福利》。《ACM经济学与计算汇刊》7，3（2019年8月），1–32。https://doi.org/ 10 1145 3355902 。/ Po-An Chen、Bart De Keijzer、David Kempe 和 [6] Guido Schäfer。2011。《利他博弈的稳健无政府代价》。载于《互联网与网络经济学》，Ning Chen、Edith Elkind 和 Elias Koutsoupias（编）。第7090卷。Springer Berlin Heidelberg，柏林，海德堡，383–390。https://doi.org/10.1007/978-3-642- 25510-6_33 丛书名称：计算机科学讲义。[7] Robyn M. Dawes 和 David M. Messick。2000。《社会困境》。《国际心理学杂志》35，2（2000），111–116。 https://doi.org/10.1080/002075900399402 arXiv:https://onlinelibrary.wiley.com/doi/pdf/ 10 1080 002075900399402 。/ Ishan Durugkar、 [8] Elad Liebman 和 Peter Stone。2020。《多智能体强化学习中平衡个体偏好与共同目标》。载于《第二十九届国际人工智能联合会议论文集》。国际人工智能联合会议组织，横滨，日本， 2505 2511 10 24963 2020 347 –。https://doi.org/。/ [9] ijcai。/ Soroush Ebadian、Rupert Freeman 和 Nisarg Shah。2024。《伤害比率：一种新颖且通用的公平性准则》。载于《第四届ACM算法、机制与优化公平及可及性会议论文集》。ACM，圣路易斯波托西，墨西哥，1–14。https://doi.org/10.1145/3689904.3694701[10] Promise Osaine Ekpo、Brian La、Thomas Wiener、Saesha Agarwal、Arshia Agrawal、Gonzalo Gonzalez-Pumariega、Lekan P. Molu 和 Angelique Tay-lor。2025。《医疗协作多智能体学习中的技能对齐公平性》。 https://doi.org/10.48550/arXiv.2508.18708 arXiv:2508.18708 [cs]。[11] Zihao Guo、Shuqing Shi、Richard Willis、Tristan Tomilin、Joel Z. Leibo 和 Yali Du。2025。《 SocialJax：序贯社会困境中多智能体强化学习评测套件》。 https://doi.org/10.48550/arXiv.2503. 14576 arXiv:2503.14576 [cs]。[12] Garrett Hardin。1968。《公地悲剧：人口问题没有技术解决方案；它要求在道德观念上进行根本性扩展》。《科学》162，3859（1968），1243–1248。[13] Edward Hughes、Joel Z. Leibo、Matthew G. Phillips、Karl Tuyls、Edgar A. Duéñez-Guzmán、Antonio García Castañeda、Iain Dunning、Tina Zhu、Kevin R. McKee、Raphael Koster、Heather Roff 和 Thore Graepel。2018。《不平等厌恶改善了跨期社会困境中的合作》。 arXiv:1803.08884 [cs.NE] https://arxiv.org/abs/ 1803 08884 。 [14] Jiechuan Jiang 和 Zongqing Lu。2019。《多智能体系统中的公平性学习》。载于《神经信息处理系统进展》第32卷。Curran Associates, Inc. https://proceedings.neurips.cc/paper_files/paper/2019/hash/ 10493aa88605cad5ab4752b04a63d172-Abstract.html[15] Peizhong Ju、Arnob Ghosh 和 Ness B. Shroff。2023。《使用强化学习实现多智能体马尔可夫决策过程中的公平性》。 https: //doi.org/10.48550/arXiv.2306.00324 arXiv:2306.00324 [cs]。[16] Mamoru Kaneko 和 Kenjiro Nakamura。1979。《纳什社会福利函数》。《计量经济学》 47, 2 1979 423 10 2307 1914191 （3月），。 [17] https://doi.org/。/ FP Kelly、AK Maulloo 和 DKH Tan。未知年份。《通信网络速率控制：影子价格、比例公平与稳定性》。（[未知年份]）。[18] Jakub Grudzien Kuba、Xidong Feng、Shiyao Ding、Hao Dong、Jun Wang 和 Yaodong Yang。2022。《异质智能体镜像学习：合作式多智能体强化学习的一系列解》。 https://doi.org/10.48550/arXiv.2208.01682 arXiv:2208.01682 [cs]。[19] Weifan Long、Wen Wen、Peng Zhai 和 Lihua Zhang。2024。《角色扮演：在多智能体交互中学习自适应角色专属策略》。https://doi.org/ 10.48550/arXiv.2411.01166 arXiv:2411.01166 [cs]。 [20] Ryan Lowe、Yi Wu、Aviv Tamar、Jean Harb、Pieter Abbeel 和 Igor Mordatch。2020。《混合合作竞争环境中的多智能体行动者-评论家》。 https://doi.org/10.48550/arXiv.1706.02275 arXiv:1706.02275 [cs]。[21] Debmalya Mandal 和 Jiarui Gan。2023。《社会公平强化学习》。 https://doi.org/10.48550/arXiv.2208.12584 arXiv:2208.12584 [cs]。[22] Hervé Moulin。2003。《公平分配与集体福利》。麻省理工学院出版社。https: //doi.org/ 10 7551 2954 001 0001 。/mitpress/。。Stefan Roesch、Stefanos [23] Leonardos 和 Yali Du。[未知年份]。《自私程度促成序贯社会困境中的合作》。（[未知年份]）。[24] John Schulman、Sergey Levine、Philipp Moritz、Michael I. Jordan 和 Pieter Abbeel。2017。《信赖域策略优化》。arXiv:1502.05477 [cs.LG] https://arxiv.org/abs/ 1502 05477 。Umer Siddique、Peilang Li 和 Yongcan Cao。2024。[25]《迈向合作式多智能体强化学习中的公平公正策略学习》。（2024）。 [26] Richard S. Sutton 和 Andrew Barto。2014。《强化学习：导论》（再版）。麻省理工学院出版社，剑桥，马萨诸塞州。[27] Chao Yu、Akash Velu、Eugene Vinitsky、Ji axuan Gao、Yu Wang、Alexandre Bayen 和 Yi Wu。2022。《PPO在合作式多智能体博弈中的惊人有效性》。https://doi.org/ 10 48550 2103 01955 2103 01955 。/arXiv。。 [ ] arXiv：。cs。[28] Tianyu Zhang、Andrew Williams、Soham Phade、Sunil Srinivasa、Yang Zhang、Prateek Gupta、Yoshua Bengio 和 Stephan Zheng。2022。《全球气候合作人工智能：在RICE-N中建模全球气候谈判、协议与长期合作》。《SSRN电子期刊》（ 2022）。https://doi.org/ 10.2139/ssrn.4189735[29] Yifan Zhong、JakubGrudzien Kuba、Xidong Feng、SiyiHu、Jiaming Ji 和 Yaodong Yang。 [n.d.]。异构智能体强化学习。（[n. d.]）。

附录

A 方法论

定理 A.1. 社会困境的利他主义水平为

$$\alpha_G = \begin{cases} 0 & if\, T \le R, \\ \frac{\log T - \log R}{\log R - \log S} & if\, T > R. \end{cases} \tag{4}$$

在证明的最后，我们需要确保其与利他主义水平的定义 α 一致。也就是说，比率 $\frac{\log T - \log R}{\log R - \log S}$应小于 1

证明。要求比率 $\frac{\log T - \log R}{\log R - \log S} < 1$ 等价于约束 $TS \le R^2$。从社会困境的第三个不等式出发，我们有

$$\begin{aligned} T + S &< 2R, \\ T^2 + 2TS + S^2 &< 4R^2, \\ \frac{T^2 + S^2}{4} + \frac{1}{2}TS &< R^2. \end{aligned}$$

对于任意实数值 T和 S，一个众所周知的不等式为

$$T^2 + S^2 \ge 2TS.$$

于是，我们得到以下不等式

$$\begin{aligned} \frac{T^2 + S^2}{4} + \frac{1}{2}TS &< R^2, \\ \frac{2 \cdot TS}{4} + \frac{1}{2}TS &\le R^2, \\ TS &\le R^2. \end{aligned}$$

□

B 公平利他马尔可夫博弈

在此，我们为策略梯度定理和优势策略梯度定理提供更完整的证明。

证明。在此，假设 $V_j^{\vec{\theta}}$关于任意 θ_i均可微。

让我们推导智能体 j的目标，

$$\begin{aligned} \nabla_{\theta_i} J_i(\vec{\theta}) &= \nabla_{\theta_i} \mathop{\mathbb{E}}_{s_0 \sim \rho_0} \left[\sum_{j=1}^{N} c_i(j) \log V_j^{\vec{\theta}}(s_0) \right] \\ &= \mathop{\mathbb{E}}_{s_0 \sim \rho_0} \left[\sum_{j=1}^{N} c_i(j) \nabla_{\theta_i} \left[\log V_j^{\vec{\theta}}(s_0) \right] \right] \\ &= \mathop{\mathbb{E}}_{s_0 \sim \rho_0} \left[\sum_{j=1}^{N} c_i(j) \frac{\nabla_{\theta_i} V_j^{\vec{\theta}}(s_0)}{V_j^{\vec{\theta}}(s_0)} \right] \end{aligned}$$

（使用链式法则）。

然后，我们需要求出梯度 $\nabla_{\theta_i} V_j^{\vec{\theta}}(s_0)$。为此，我们从贝尔曼方程出发，对其求导。接着，我们在状态价值函数空间上定义一个压缩映射，并应用巴拿赫-皮卡不动点定理，得到梯度的公式。**1. 对贝尔曼方程求导**。

让我们回顾一下状态价值函数的贝尔曼方程:

$$V_j^{\vec{\theta}}(s) = \sum_{\vec{a} \in \mathcal{A}} \pi_{\vec{\theta}}(\vec{a} \mid s) \left(r_j(s, \vec{a}) + \gamma \sum_{s' \in \mathcal{S}} \mathbb{P}(s' \mid s, \vec{a}) V_j^{\vec{\theta}}(s') \right)$$

请注意， r_i、 P以及用于 $j \neq i$的策略π_{θ_j}均不依赖于 θ_i。然后，可以使用乘积法则对贝尔曼方程求导：

$$\begin{aligned}\nabla_{\theta_i} V_j^{\vec{\theta}}(s) &= \sum_{\vec{a}\in\mathcal{A}} \nabla_{\theta_i}\pi_{\vec{\theta}}(\vec{a}\,|\,s)\cdot\left(r_j(s,\vec{a}) + \gamma\sum_{s'\in\mathcal{S}}\mathbb{P}(s'|s,\vec{a})V_j^{\vec{\theta}}(s')\right)\\ &\quad + \pi_{\vec{\theta}}(\vec{a}\,|\,s)\cdot\nabla_{\theta_i}\left(r_j(s,\vec{a}) + \gamma\sum_{s'\in\mathcal{S}}\mathbb{P}(s'|s,\vec{a})V_j^{\vec{\theta}}(s')\right)\\ &= \sum_{\vec{a}\in\mathcal{A}} \nabla_{\theta_i}\pi_{\vec{\theta}}(\vec{a}\,|\,s)\cdot\left(r_j(s,\vec{a}) + \gamma\sum_{s'\in\mathcal{S}}\mathbb{P}(s'|s,\vec{a})V_j^{\vec{\theta}}(s')\right)\\ &\quad + \pi_{\vec{\theta}}(\vec{a}\,|\,s)\cdot\gamma\sum_{s'\in\mathcal{S}}\mathbb{P}(s'|s,\vec{a})\nabla_{\theta_i}V_j^{\vec{\theta}}(s')\end{aligned}$$

让我们分析求和中的第一项。

第 1 项：展开联合策略

$$\nabla_{\theta_i}\pi_{\vec{\theta}}(\vec{a}|s) = \nabla_{\theta_i}\left(\pi_{\theta_i}(a_i|s)\prod_{k\neq i}\pi_{\theta_k}(a_k|s)\right) = \left(\nabla_{\theta_i}\pi_{\theta_i}(a_i|s)\right)\prod_{k\neq i}\pi_k(a_k|s)$$

然后，使用对数导数技巧， $\nabla f = f\cdot\nabla\log f$：

$$\nabla_{\theta_i}\pi_{\vec{\theta}}(\vec{a}|s) = \left(\pi_{\theta_i}(a_i|s)\nabla_{\theta_i}\log\pi_{\theta_i}(a_i|s)\right)\prod_{k\neq i}\pi_{\theta_k}(a_k|s) = \pi_{\vec{\theta}}(\vec{a}|s)\cdot\nabla_{\theta_i}\log\pi_{\theta_i}(a_i|s)$$

此外，回顾一下动作状态价值函数的定义：

$$Q_j^{\vec{\theta}}(s,\,\vec{a}) = r_j(s,\vec{a}) + \gamma\sum_{s'\in\mathcal{S}}\mathbb{P}(s'|s,\vec{a})V_j^{\vec{\theta}}(s').$$

因此，第一项可以改写为

$$\pi_{\vec{\theta}}(\vec{a}|s)\cdot\nabla_{\theta_i}\log\pi_{\theta_i}(a_i|s)\cdot Q_j^{\vec{\theta}}(s,\,\vec{a})$$

现在，将其代回 $\nabla_{\theta_i}V_j^{\theta}(s)$ 的表达式中：

$$\nabla_{\theta_i}V_j^{\vec{\theta}}(s) = \sum_{\vec{a}}\pi_{\vec{\theta}}(\vec{a}|s)\left[\left(\nabla_{\theta_i}\log\pi_{\theta_i}(a_i|s)\right)Q_j^{\vec{\theta}}(s,\vec{a}) + \gamma\sum_{s'}\mathbb{P}(s'|s,\vec{a})\nabla_{\theta_i}V_j^{\vec{\theta}}(s')\right]$$

我们定义函数 $G_{i,j}^{\theta}(s)$，它表示状态 s 处的即时梯度分量：

$$G_{i,j}^{\vec{\theta}}(s) := \sum_{\vec{a}\in\mathcal{A}}\pi_{\vec{\theta}}(\vec{a}|s)\left(\nabla_{\theta_i}\log\pi_{\theta_i}(a_i|s)\right)Q_j^{\vec{\theta}}(s,\vec{a}) = \mathbb{E}_{\vec{a}\sim\vec{\pi}_\theta}\left[\nabla_{\theta_i}\log\pi_{\theta_i}(a_i|s)Q_j^{\vec{\theta}}(s,\vec{a})\right]$$

由此，我们的递归方程变为：

$$\nabla_{\theta_i}V_j^{\vec{\theta}}(s) = G_{i,j}^{\vec{\theta}}(s) + \gamma\sum_{\vec{a}\in\mathcal{A}}\pi_{\vec{\theta}}(\vec{a}|s)\sum_{s'}\mathbb{P}(s'|s,\vec{a})\nabla_{\theta_i}V_j^{\vec{\theta}}(s')$$

2. 定义价值函数梯度的贝尔曼算子

此方程定义了一个不动点关系。设 $g_j\colon S\to\mathbb{R}^{|\theta_j|}$ 为一个梯度向量场。我们可以将价值函数梯度 $\mathfrak{T}_{i,j}^{\theta}$ 的贝尔曼算子定义为：

$$(\mathfrak{T}_{i,j}^{\vec{\pi}}g_i)(s) := G_{i,j}^{\vec{\theta}}(s) + \gamma\mathbb{E}_{\vec{a}\sim\pi_{\vec{\theta}}(\cdot|s),s'\sim P(\cdot|s,\vec{a})}\left[g_j(s')\right]$$

我们要寻找的梯度 $\nabla_{\theta_i}V_j^{\theta}$ 是该算子的不动点：

$$\nabla_{\theta_i}V_j^{\vec{\theta}} = \mathfrak{T}_{i,j}^{\vec{\theta}}(\nabla_{\theta_i}V_j^{\vec{\theta}})$$

3. 巴拿赫-皮卡不动点定理的应用

为了证明该算子具有唯一不动点，我们需要证明它在完备度量空间中是压缩映射。考虑所有梯度向量场 $g_j\colon S\to\mathbb{R}^{|\theta_i|}$ 组成的空间，并为其配备无穷范数：

$$\|g_j\|_\infty = \max_{s\in S}\|g_j(s)\|_2$$

其中 $\|\cdot\|_2$ 是标准欧几里得范数。

设 g_1 和 g_2 是该空间中的两个梯度向量场。我们考察它们在该算子下映射之间的距离：

$$(\mathfrak{T}_{i,j}^{\vec{\theta}} g_1)(s) - (\mathfrak{T}_{i,j}^{\vec{\theta}} g_2)(s) = \left(G_{i,j}^{\vec{\theta}}(s) + \gamma \mathbb{E}_{s',\vec{a}}[g_1(s')]\right) - \left(G_{i,j}^{\vec{\theta}}(s) + \gamma \mathbb{E}_{s',\vec{a}}[g_2(s')]\right)$$
$$= \gamma \mathbb{E}_{\vec{a}\sim\vec{\pi}(\cdot|s), s'\sim P(\cdot|s,\vec{a})}[g_1(s') - g_2(s')]$$

现在，我们取范数：

$$\|(\mathfrak{T}_{i,j}^{\vec{\theta}} g_1)(s) - (\mathfrak{T}_{i,j}^{\vec{\theta}} g_2)(s)\|_2 = \gamma \left\| \sum_{s'\in\mathcal{S}} \left(\sum_{\vec{a}\in\mathcal{A}} \pi_{\vec{\theta}}(\vec{a}|s) \mathbb{P}(s'|s,\vec{a}) \right) (g_1(s') - g_2(s')) \right\|_2$$

利用三角不等式和范数的性质：

$$\leq \gamma \sum_{s'\in\mathcal{S}} \left(\sum_{\vec{a}\in\mathcal{A}} \pi_{\vec{\theta}}(\vec{a}|s) \mathbb{P}(s'|s,\vec{a}) \right) \|g_1(s') - g_2(s')\|_2$$

括号中的项表示在联合策略 $\pi_{\vec{\theta}}$ 下从 s 转移到 s' 的总概率，我们将其记为 $p_{\vec{\theta}}(s'|s)$。

$$\leq \gamma \sum_{s'} p_{\vec{\pi}}(s'|s) \|g_1(s') - g_2(s')\|_2$$

根据无穷范数的定义，对于任意 s'， $\|g_1(s') - g_2(s')\|_2 \leq \|g_1 - g_2\|_\infty$。

$$\leq \gamma \sum_{s'} P^{\vec{\pi}}(s'|s) \|g_1 - g_2\|_\infty$$

由于概率之和为 1，$\sum_{s'} p_{\vec{\theta}}(s'|s) = 1$：

$$\|(\mathfrak{T}_{i,j}^{\vec{\theta}} g_1)(s) - (\mathfrak{T}_{i,j}^{\vec{\theta}} g_2)(s)\|_2 \leq \gamma \|g_1 - g_2\|_\infty.$$

这个不平等对所有状态 $s \in \mathrm{S}$ 都成立。因此，对所有 s取最大值：

$$\|\mathfrak{T}_{i,j}^{\vec{\theta}} g_1 - \mathfrak{T}_{i,j}^{\vec{\theta}} g_2\|_\infty \leq \gamma \|g_1 - g_2\|_\infty$$

由于 $\gamma \in [0,1)$，算子 $\mathfrak{T}_{i,j}^{\vec{\theta}}$是压缩映射。根据**巴拿赫-皮卡不动点定理**，它有唯一不动点。这证明了梯度向量场 $\nabla_{\theta_i} V_j^{\vec{\theta}}$是唯一的。

4. 马尔可夫博弈策略梯度定理

不动点方程 $\nabla_{\theta_i} V_j^{\vec{\theta}} = \mathfrak{T}_{i,j}^{\vec{\theta}}(\nabla_{\theta_i} V_j^{\vec{\theta}})$ 得出了最终定理。**巴拿赫-皮卡不动点定理** 还指出，对于任意序列 $g_{n+1} = \mathfrak{T}_{i,j}^{\vec{\theta}} g_n$和 $g_0 : \mathrm{S} \to \mathbb{R}^{|\theta_j|}$，$(g_n)_{n\in\mathbb{N}}$ 都会收敛 $\nabla_{\theta_i} V_j^{\vec{\theta}}$。

我们可以从 $g_0 = 0$开始展开递归来求解：

$$g_1(s_0) = (\mathfrak{T}_{i,j}^{\vec{\theta}} g_0)(s_0) = G_{i,j}^{\vec{\theta}}(s_0)$$

$$g_2(s_0) = (\mathfrak{T}_{i,j}^{\vec{\theta}} g_1)(s_0) = G_{i,j}^{\vec{\theta}}(s_0) + \gamma \mathop{\mathbb{E}}_{\substack{\vec{a}_0\sim\pi_{\vec{\theta}}(.|s_0) \\ s_1\sim P(.|s_0,\vec{a}_0)}} \left[G_{i,j}^{\vec{\theta}}(s_1) \right]$$

$$g_3(s_0) = (\mathfrak{T}_{i,j}^{\vec{\theta}} g_2)(s_0) = G_{i,j}^{\vec{\theta}}(s_0) + \gamma \mathop{\mathbb{E}}_{\substack{\vec{a}_0\sim\pi_{\vec{\theta}}(.|s_0) \\ s_1\sim P(.|s_0,\vec{a}_0)}} \left[G_{i,j}^{\vec{\theta}}(s_1) + \gamma \mathop{\mathbb{E}}_{\substack{\vec{a}_1\sim\pi_{\vec{\theta}}(.|s_1) \\ s_2\sim P(.|s_1,\vec{a}_1)}} \left[G_{i,j}^{\vec{\theta}}(s_2) \right] \right]$$

$$\vdots$$

$$g_n(s_0) = (\mathfrak{T}_{i,j}^{\vec{\theta}} g_{n-1})(s_0) = \mathop{\mathbb{E}}_{\tau\sim\mathbb{P}^{\vec{\theta}}} \left[\sum_{t=0}^{n-1} \gamma^t G_{i,j}^{\vec{\theta}}(s_t) \,\middle|\, s_0 \right].$$

因此，当 g_n 收敛到 $\nabla_{\theta_i} V_j^{\theta}$ 时，以下等式成立

$$\nabla_{\theta_i} V_j^{\vec{\theta}}(s_0) = g_\infty(s_0) = \mathop{\mathbb{E}}_{\tau\sim\mathbb{P}^{\vec{\theta}}} \left[\sum_{t=0}^{\infty} \gamma^t G_{i,j}^{\vec{\theta}}(s_t) \,\middle|\, s_0 \right]$$

代入$G_{i,j}^{\theta}(s_t)$的定义：

$$\nabla_{\theta_i} V_j^{\vec{\theta}}(s_0) = \mathop{\mathbb{E}}_{\tau\sim\mathbb{P}^{\vec{\theta}}} \left[\sum_{t=0}^{\infty} \gamma^t \mathop{\mathbb{E}}_{\vec{a}_t\sim\pi_{\vec{\theta}}(\cdot|s_t)} \left[\nabla_{\theta_i} \log \pi_{\theta_i}(a_{i,t}|s_t) Q_j^{\vec{\theta}}(s_t,\vec{a}_t) \right] \,\middle|\, s_0 \right]$$

外层期望已经考虑了动作采样，因此我们可以将其简化为：

$$\nabla_{\theta_i} V_j^{\vec{\pi}}(s_0) = \mathop{\mathbb{E}}_{\tau\sim\mathbb{P}^{\vec{\theta}}}\left[\sum_{t=0}^{\infty}\gamma^t \nabla_{\theta_i}\log\pi_{\theta_i}(a_{i,t}|s_t) Q_j^{\vec{\theta}}(s_t,\vec{a}_t)\,\middle|\, s_0\right]$$

通过注 入 $\nabla_{\theta_i} V_j^{\pi}(s_0)$ 的表达式到前一个梯度中，我们得到**公平策略梯度定理** **m**

$$\begin{aligned}
\nabla_{\theta_i} J_i(\vec{\theta}) &= \mathop{\mathbb{E}}_{s_0\sim\rho_0}\left[\sum_{j=1}^{N} c_i(j)\frac{\nabla_{\theta_i} V_j^{\vec{\theta}}(s_0)}{V_j^{\vec{\theta}}(s_0)}\right] \\
&= \mathop{\mathbb{E}}_{s_0\sim\rho_0}\left[\sum_{j=1}^{N} c_i(j)\frac{\mathop{\mathbb{E}}_{\tau\sim\mathbb{P}^{\vec{\theta}}}\left[\sum_{t=0}^{\infty}\gamma^t\nabla_{\theta_i}\log\pi_{\theta_i}(a_{i,t}|s_t)Q_j^{\vec{\theta}}(s_t,\vec{a}_t)\,\middle|\, s_0\right]}{V_j^{\vec{\theta}}(s_0)}\right] \\
&= \mathop{\mathbb{E}}_{s_0\sim\rho_0}\left[\sum_{j=1}^{N} c_i(j)\mathop{\mathbb{E}}_{\tau\sim\mathbb{P}^{\vec{\theta}}}\left[\sum_{t=0}^{\infty}\gamma^t\nabla_{\theta_i}\log\pi_{\theta_i}(a_{i,t}|s_t)\frac{Q_j^{\vec{\theta}}(s_t,\vec{a}_t)}{V_j^{\vec{\theta}}(s_0)}\,\middle|\, s_0\right]\right] \\
&= \mathop{\mathbb{E}}_{s_0\sim\rho_0}\left[\mathop{\mathbb{E}}_{\tau\sim\mathbb{P}^{\vec{\theta}}}\left[\sum_{t=0}^{\infty}\gamma^t\nabla_{\theta_i}\log\pi_{\theta_i}(a_{i,t}|s_t)\left(\sum_{j=1}^{N} c_i(j)\frac{Q_j^{\vec{\theta}}(s_t,\vec{a}_t)}{V_j^{\vec{\theta}}(s_0)}\right)\,\middle|\, s_0\right]\right] \\
&= \mathop{\mathbb{E}}_{\tau\sim\mathbb{P}^{\vec{\theta}}}\left[\sum_{t=0}^{\infty}\gamma^t\nabla_{\theta_i}\log\pi_{\theta_i}(a_{i,t}|s_t)\left(\sum_{j=1}^{N} c_i(j)\frac{Q_j^{\vec{\theta}}(s_t,\vec{a}_t)}{V_j^{\vec{\theta}}(s_0)}\right)\right].
\end{aligned}$$

□

证明。证明的关键在于说明，与基线 $V_i^{\vec{\pi}}(s_t)$ 相关的梯度项的期望值为零。考虑给定状态 s_t 时，单个时间步 t 上该基线项的期望。此处的期望是针对从联合策略 $\vec{\pi}(\cdot|s_t)$ 中抽取的联合动作 $\vec{a}_t$ 求取的。

我们要证明，对于任意 $f\colon \mathsf{S}\to\mathbb{R}$

$$\mathbb{E}_{\vec{a}_t\sim\pi_{\vec{\theta}}(\cdot|s_t)}\left[\nabla_{\theta_i}\log\pi_i(a_{i,t}|s_t)f(s_t)\right] = 0$$

展开该期望：

$$\sum_{\vec{a}_t\in\mathcal{A}}\pi_{\vec{\theta}}(\vec{a}_t|s_t)\nabla_{\theta_i}\log\pi_i(a_{i,t}|s_t)f(s_t)$$

函数 $f(s_t)$ 取决于状态 s_t，而不取决于该状态下抽取的具体动作 $\vec{a}_t$，因此我们可以将它从对动作的求和中提出，因为它在求和过程中保持不变。更正式地说，我们可以将对智能体 i 的动作的求和与对其他智能体动作的求和分开。令 $\vec{a}_t = (a_{i,\,t},\ a_{-i,\,t})$，其中 $a_{-i,\,t}$ 是除 i 之外所有智能体的动作。

$$\sum_{a_{-i,t}}\prod_{j\neq i}\pi_j(a_{j,t}|s_t)\sum_{a_{i,t}}\pi_i(a_{i,t}|s_t)\nabla_{\theta_i}\log\pi_i(a_{i,t}|s_t)f(s_t)$$

由于 $f(s_t)$ 不依赖于 $a_{i,\,t}$，我们可以将其从内层求和中提取出来：

$$= \left(\sum_{a_{-i,t}}\prod_{j\neq i}\pi_{\theta_j}(a_{j,t}|s_t)\right) f(s_t)\left(\sum_{a_{i,t}}\pi_{\theta_i}(a_{i,t}|s_t)\nabla_{\theta_i}\log\pi_{\theta_i}(a_{i,t}|s_t)\right)$$

第一项是对所有动作的概率乘积求和，其结果为 1。我们来关注最后一项。使用对数导数技巧

$$\begin{aligned}
\sum_{a_{i,t}}\pi_{\theta_i}(a_{i,t}|s_t)\nabla_{\theta_i}\log\pi_{\theta_i}(a_{i,t}|s_t) &= \sum_{a_{i,t}}\nabla_{\theta_i}\pi_{\theta_i}(a_{i,t}|s_t) \\
&= \nabla_{\theta_i}\sum_{a_{i,t}}\pi_{\theta_i}(a_{i,t}|s_t) \qquad \text{（所有可能动作的概率之和等于 1）} \\
&= \nabla_{\theta_i}(1) \\
&= 0
\end{aligned}$$

由于基线项的期望值为零，因此从原始策略梯度_定理中期望内的项中减去该基线项，并不会改变总期望。在本文中，我们使用基线 $f(s_t) = \sum_{j=1}^{N} c_i(j) V_j^{\theta}(s_t) V_j^{\vec{\theta}}(s_0)$

由于基线项的期望为零，我们可以在外层期望中将其减去：

$$
\begin{aligned}
\nabla_{\theta_i} J_i(\vec{\theta}) &= \mathop{\mathbb{E}}_{\tau\sim\mathbb{P}^{\vec{\theta}}}\left[\sum_{t=0}^{\infty}\sum_{j=1}^{N} c_i(j)\frac{Q_j^{\vec{\theta}}(s_t,\vec{a}_t)}{V_j^{\vec{\theta}}(s_0)}\nabla_{\theta_i}\log\pi_{\theta_i}(a_{i,t}\mid s_t)\right] \\
&= \mathop{\mathbb{E}}_{\tau\sim\mathbb{P}^{\vec{\theta}}}\left[\sum_{t=0}^{\infty}\sum_{j=1}^{N} c_i(j)\frac{Q_j^{\vec{\theta}}(s_t,\vec{a}_t)}{V_j^{\vec{\theta}}(s_0)}\nabla_{\theta_i}\log\pi_{\theta_i}(a_{i,t}\mid s_t)\right] \\
&\quad - \mathop{\mathbb{E}}_{\tau\sim\mathbb{P}^{\vec{\theta}}}\left[\sum_{t=0}^{\infty}\left(\sum_{j=1}^{N} c_i(j)\frac{V_j^{\vec{\theta}}(s_t)}{V_j^{\vec{\theta}}(s_0)}\right)\nabla_{\theta_i}\log\pi_{\theta_i}(a_{i,t}\mid s_t)\right] \\
&= \mathop{\mathbb{E}}_{\tau\sim\mathbb{P}^{\vec{\theta}}}\left[\sum_{t=0}^{\infty}\sum_{j=1}^{N} c_i(j)\frac{Q_j^{\vec{\theta}}(s_t,\vec{a}_t)-V_j^{\vec{\theta}}(s_t)}{V_j^{\vec{\theta}}(s_0)}\nabla_{\theta_i}\log\pi_{\theta_i}(a_{i,t}\mid s_t)\right] \\
&= \mathop{\mathbb{E}}_{\tau\sim\mathbb{P}^{\vec{\theta}}}\left[\sum_{t=0}^{\infty} A_i^{F,\vec{\theta}}(s_t,\vec{a}_t,s_0)\nabla_{\theta_i}\log\pi_{\theta_i}(a_{i,t}\mid s_t)\right]
\end{aligned}
$$

□

C 伪代码示例

算法 1公平 MAA2C

要求：全局学习率 lr**、折扣因子** γ**、智能体数量** N**、步数** T_{max}**、利他主义水平** α**。确保：对于每个智能体** $i \in \{1, ..., N\}$**，都有一个参数为随机值** ϕ_i **的评论家网络** $V(\mathbf{s}; \phi_i)$**。确保：对于每个智能体** $i \in \{1, ..., N\}$**，都有一个参数为随机值** θ_i **的行动者网络** $\pi(a_i|o_i; \theta_i)$**。1：初始化全局步数计数器** $T \leftarrow 0$**。2：当** $T < T_{max}$ **时执行 3：重置梯度：** $d\phi_i \leftarrow 0$**，且对于所有** $i = 1, ..., N$**，令** $d\theta_i \leftarrow 0$**。4：初始化环境并获取初始联合观测 o** $= \{o_1, ..., o_N\}$ **和全局状态 s。5：初始化空的回合缓冲区** D**。6：对于从** $t = 1$ **到终止状态或最大回合长度的每一步，执行 7：对于每个智能体** i**，选择动作** $a_i \sim \pi(a_i|o_i; \theta_i)$**。8：执行联合动作 a** $= \{a_1, ..., a_N\}$**，观测奖励 r** $= \{r_1, ..., r_N\}$**、下一个联合观测 o′ 和下一个全局状态 s′。9：将 (s, a, r, s′) 存入回合缓冲区** D**。10：令 s** $\leftarrow$ **s′，o** $\leftarrow$ **o′。11：**$T \leftarrow T + 1$**。12：结束循环 13：对于每个智能体** i**，使用时序差分更新估计回报** $R_{i,t}$**。14：对于每个智能体** i**，估计初始状态的状态价值函数：** $V_{i,0} \leftarrow V(s_0; \phi_i)$**。15：对于每个智能体** i**及每个时间步** t**，估计智能体** i **的优势函数（使用广义优势估计）：** $A_{i,t}$**。16：对于每个智能体** i**及每个时间步** t**，估计智能体** i **的公平优势函数：** $A_{i,t}^{F} \leftarrow \sum_j c_i(j) A_{j,t}/V_{j,0}$**。17：对于每个智能体** i**，计算行动者损失：** $L_{\text{actor},i}(\theta_i) = \mathbb{E}_t[A_{i,t}^{F} \log \pi(a_i|o_i; \theta_i)]$**。18：对于每个智能体，计算评论家损失：, ( ) ((;))。19：对于每个智能体，更新评论家参数：** $\phi_i \leftarrow \phi_i - lr \cdot d\phi_i$**。20：对于每个智能体** i**，更新行动者参数：** $\theta_i \leftarrow \theta_i - \alpha \cdot d\theta_i$**。21：结束循环**

算法 2公平 MAPPO

要求： 全局学习率 lr、折扣因子 γ、智能体数量 N、步数 T_{max}、利他主义水平 α、熵正则化 β，近端策略优化更新次数 T_{PPO}。

确保： 对于每个智能体 $i \in \{1, ..., N\}$，一个具有随机参数 ϕ_i 的评论家网络 $V(\mathbf{s}; \phi_i)$。

确保： 对于每个智能体 $i \in \{1, ..., N\}$，一个具有随机参数 θ_i 的行动者网络 $\pi(a_i|o_i; \theta_i)$。

初始化全局步数计数器 $T \leftarrow 0$。

当 $T < T_{max}$ **do**

  重置梯度：$d\phi_i \leftarrow 0$，并对所有 $i = 1, ..., N$执行此操作。

  初始化环境并获取初始联合观测$\mathbf{o} = \{o_1, ..., o_N\}$以及全局状态$\mathbf{s}$。

  初始化空的回合缓冲区 D。

  对于 $t = 1$，直到达到终止状态或最大回合长度**执行**

    对于每个智能体 i，选择动作 $a_i \sim \pi(a_i|o_i; \theta_i)$。

    执行联合动作 $\mathbf{a} = \{a_1, ..., a_N\}$，观测奖励 $\mathbf{r} = \{r_1, ..., r_N\}$、下一个联合观测 $\mathbf{o}'$ 和下一个全局状态 $\mathbf{s}'$。

    将 $(\mathbf{s}, \mathbf{a}, \mathbf{r}, \mathbf{s}')$ 存入回合缓冲区 D。

    $\text{状态} \leftarrow \text{状态}'$, $\text{观测} \leftarrow \text{观测}'$.

    $T \leftarrow T + 1$。

  结束循环

  对于每个智能体 $i \in \{1、.....、 N\}$，使用时序差分更新估计回报 $R_{i,\ t}$

  对于每个智能体 $i \in \{1、.....、 N\}$，估计初始状态的状态价值函数：$V_{i,\ 0} \leftarrow V(s_0;\ \phi_i)$。

  对于每个智能体 $i \in \{1、.....、 N\}$ 和每个时间步 t，估计智能体 i 的优势函数（使用广义优势估计）：$A_{i,\ t}$。

  对于每个智能体 $i \in \{1、.....、 N\}$ 和每个时间步 t，估计智能体 i 的公平优势函数：$A_{i,\ t} \leftarrow \sum_j c_i(j) A_{j,\ t} / V_{j,\ 0}$。

  对于 $b = 1$ 到 T_{PPO}，**执行**

    对于每个智能体 $i \in \{1、.....、 N\}$，计算演员PPO损失：$L_{\text{actor},\ i}(\theta_i)$。

    对于每个智能体 $i \in \{1，.....，N\}$，计算评论家PPO损失：$L_{\text{critic},\ i}$

    对于每个智能体 i，更新评论家参数：$\phi_i \leftarrow \phi_i - lr \cdot d\phi_i$。

    对于每个智能体 i，更新演员参数：$\theta_i \leftarrow \theta_i - \alpha \cdot d\theta_i$。

  结束 for

结束 while