中国科学: 信息科学 2023 年 第 53 卷 第 4 期: 623–646
SCIENTIA SINICA Informationis
《中国科学》杂志社
SCIENCE CHINA PRESS
评述

博弈收益控制研究进展

王龙[1*], 陈芳[1], 陈星如[2]

1. 北京大学系统与控制研究中心, 北京 100871
2. 北京邮电大学理学院, 北京 100876
* 通信作者. E-mail: longwang@pku.edu.cn

收稿日期: 2022–06–30; 修回日期: 2022–08–27; 接受日期: 2022–10–05; 网络出版日期: 2023–04–12

国家自然科学基金 (批准号: 62036002) 资助项目

摘要 在博弈论中, 单个个体控制全部个体的收益通常被认为是不可能的. 一个例外是 20 世纪末在重复囚徒困境中提出的均衡器策略: 使用这种策略的个体可以将对手的收益设置为由收益函数所决定的某个区间内的任意值. 十余年后发现的零行列式策略通过单方面设置个体收益的线性关系, 将该结果一般化. 在此基础上, 关于博弈收益控制的研究取得了一系列成果. 本文概述了博弈收益控制的研究现状; 介绍了单次博弈和重复博弈中的收益控制技术; 从收益控制的基本概念、能控制的收益关系、收益控制策略的形式和收益控制策略的演化特性等方面总结了博弈中收益控制的主要进展和成果; 并讨论了博弈收益控制的未来发展趋势.

关键词 博弈论, 收益控制, 零行列式策略, 演化博弈论, 策略设计

1 引言

博弈论是一门运用数学方法研究利益冲突的个体在竞争性活动中的最优化策略的科学理论[1]. 著名数学家 von Neumann 和经济学家 Morganstern 在 1944 年合著的 "*Theory of Games and Economic Behavior*" 一书[1] 标志着博弈论的诞生. 经典的博弈论立足于两个关键因素: (1) 个体之间存在利益冲突; (2) 参与博弈的个体都是理性的[2]. 任何具有这两个要素的问题都可以被抽象为博弈问题, 进而可以利用博弈论的分析方法来研究[3~20].

在博弈论的研究中, 最主要的目的是: (1) 寻找最大化个体收益的策略; (2) 确定能在博弈中获胜的个体. Nash 在 1950 年和 1951 年发表的两篇文章[21,22] 创造性地在非合作博弈中引入了混合策略纳什均衡 (Nash equilibrium) 的概念, 极大地推动了博弈论的发展. 任何有限博弈中至少存在一个纳什均衡. 当一个博弈处于纳什均衡时, 任何个体都不能通过单方面地改变自己的策略而提高自己的收益. 找到这个博弈的纳什均衡, 便可以确定使用哪种策略的个体可以在博弈中获胜. 在不同博弈场景

引用格式: 王龙, 陈芳, 陈星如. 博弈收益控制研究进展. 中国科学: 信息科学, 2023, 53: 623–646, doi: 10.1360/SSI-2022-0263
Wang L, Chen F, Chen X R. Payoff control in game theory (in Chinese). Sci Sin Inform, 2023, 53: 623–646, doi: 10.1360/SSI-2022-0263

下, 纳什均衡有不同的变形, 如子博弈精炼纳什均衡 (subgame perfect Nash equilibrium)、贝叶斯纳什均衡 (Bayesian Nash equilibrium)、精炼贝叶斯纳什均衡 (perfect Bayesian Nash equilibrium) 等.

一个博弈往往包含多个纳什均衡. 例如, 最简洁的博弈范式 —— 雪堆博弈 —— 有三个纳什均衡; 更复杂的重复博弈由于纳什均衡太多涉及均衡精炼问题. 尽管近些年机器学习和人工智能的发展为寻找博弈的最优策略和均衡提供了便利 [23~26], 但是数学上确定一个博弈所有的纳什均衡并判断博弈最终处于哪个均衡状态仍是非常困难的. 如果存在一种策略, 可以让使用该策略的个体固定对手的收益, 或者获得比对手更高的收益, 那么也可以解决博弈中谁能赢的问题. 然而, 单个个体的收益会受到博弈的收益函数和每个参与博弈的个体策略的影响. 想要摆脱对手策略的影响去控制所有个体的收益通常被认为是不可能的 [27, 28].

1997 年, Boerlijst 等 [29] 发现一类均衡器策略 (equalizer strategy). 无论对手选择什么策略, 使用均衡器策略的个体都可以单方面地将对手的收益固定为由收益函数所决定的某个区间内的任意值. 这是关于囚徒困境收益控制的早期研究, 学术界并未给予足够的重视. 2012 年, Press 和 Dyson [30] 发现了一类更一般的策略 —— 零行列式策略 (zero-determinant strategy, ZD strategy). 和均衡器策略类似, 无论对手使用何种策略, 这类策略可以使自己的收益和对手的收益满足一个线性关系, 从而单方面实现对博弈收益的控制. 通过调整该线性关系的参数, 使用零行列式策略的个体可以使对手收益固定或获得比对手更高的收益. 零行列式策略前所未有的控制能力引起了学者们的广泛关注 [31~40]. 进一步的研究发现, 可以控制收益的策略不仅在重复博弈中存在, 而且在单次博弈中也存在; 通过调整单个个体的策略, 除了能够让所有个体的收益满足线性关系, 也可以使其满足非线性关系.

本文总结了自 Press 和 Dyson 提出零行列式策略以来关于收益控制的研究进展: 包括收益控制策略基本的数学表达, 将收益控制为满足特定关系式的收益控制策略的形式以及这些策略的演化特性. 第 2 节回顾博弈和演化博弈的基本模型与相关概念. 第 3 节介绍在单次博弈中能够实现收益控制的策略及寻找这类策略的一般方法. 第 4 节梳理重复博弈中零行列式策略的研究内容, 包括零行列式策略的发现过程、存在性、判别以及演化特性, 并介绍了可以反制敲诈的不屈策略. 第 5 节呈现随机博弈中的收益控制. 最后, 第 6 节总结收益控制的研究现状, 展望收益控制的研究前景, 并从学科交叉融合的角度讨论收益控制的研究成果对其他领域的影响和启发.

2 基本模型与相关概念

本节介绍博弈的基本模型, 从单次到重复博弈的过渡及演化博弈中的一些重要概念.

2.1 基本模型

一般来说, 一个博弈由 3 个基本要素构成 [2, 41]:

(1) 参与博弈的个体, 通常用 $\mathcal{N}=\{1,\ldots,n\}$ 表示所有参与者组成的集合;

(2) 行动集, 通常用 $\mathcal{A}_i=\{a_i^1,\ldots,a_i^{m_i}\}$ 表示个体 i 在每次博弈中可以选择的行动的集合, 其中 m_i 表示个体 i 的总行动数目;

(3) 收益函数, 通常用 $u_i(\boldsymbol{a})$ 表示个体 i 的收益. 这里, $\boldsymbol{a}=(a_1,\ldots,a_n)$ 表示博弈的结果, a_i 表示个体 i 所选的行动. 收益函数决定了每次博弈后个体能获得的收益.

综合以上 3 个要素, 一个由 n 个个体参与的博弈的标准式可以写为

$$\mathcal{G}=\{\mathcal{A}_1,\ldots,\mathcal{A}_n;u_1,\ldots,u_n\}. \tag{1}$$

当 $n=2$ 时, 对应的博弈为两人博弈, 经典的两人博弈包括囚徒困境 (prisoner's dilemma), 斗鸡博弈 (chicken) 和猎鹿博弈 (stag hunt) 等, 它们从贪婪 (greed) 和 (或) 恐惧 (fear) 的角度刻画了社会困境里个人和集体利益之间的冲突 [42]. 其中, 同时存在贪婪和恐惧的囚徒困境是研究两人博弈的基本模型 [43~45]. 在囚徒困境模型中, 每个个体的行动集为 $\mathcal{A}_i=\{C,D\}$; 这里 $i\in\{1,2\}$, C 和 D 分别表示合作 (cooperation) 和背叛 (defection). 如果两个个体都选择合作, 则两人都将获得 R (相互合作的奖励); 如果一个个体合作, 另一个个体背叛, 则合作者获得 S (上当的代价), 背叛者获得 T (背叛的诱惑); 如果两个个体都背叛, 则两人分别获得 P (相互背叛的惩罚). 用收益函数表示囚徒困境的收益, 有 $u_i(C,C)=R, u_i(C,D)=S, u_i(D,C)=T, u_i(D,D)=P$. 两人博弈的收益还可以用矩阵的形式更直观地表示, 其中第 (i,j) 个元素代表行个体选择行动 a^i, 列个体选择行动 a^j $(a^i,a^j\in\{C,D\})$ 时, 行个体的收益. 于是, 囚徒困境的收益矩阵为

$$\begin{array}{c} \\ C \\ D \end{array}\begin{array}{c} \begin{array}{cc} C & D \end{array} \\ \begin{pmatrix} R & S \\ T & P \end{pmatrix} \end{array}. \tag{2}$$

在囚徒困境中, 个体的收益函数满足 $T>R>P>S$ 且 $2R>T+S$. 可以发现, 由于个体的贪婪 $(T>R)$ 和恐惧 $(P>S)$, 无论对手选择合作还是背叛, 每个个体最大化收益的行动总是背叛 [42]. 但是对于整体而言, 相互合作的收益大于相互背叛的收益 $(R>P)$. 个体利益和集体利益之间的冲突导致困境产生.

除基本的三要素之外, 一个博弈还可以包括其他要素, 比如, 行动次序、信息等. 行动次序一般有两种: 同时行动 (simultaneous game) 和交替行动 (alternating game) [46~49]. 同时行动并不意味着所有个体在同一时刻行动, 只要满足每个个体行动之前不知道其他个体的行动便可以称为同时行动. 在一个博弈中, 个体依据已知的信息采取行动. 在博弈重复进行时, 个体的一个重要信息是过往博弈的结果. 个体依赖的信息也可能受到环境的影响. 比如, 在直接互惠和间接互惠模型中, 环境可能使个体无法正确观察和感知其他个体的行为 [50~56]. 这些由环境产生的观测误差 (observation error) 和感知误差 (perception error) 在收益控制中也受到了许多关注.

基于行动集 $\mathcal{A}_i$, 个体 i 可以设置不同的策略. 考虑单次博弈, 对于博弈 $\mathcal{G}$, 个体 i 的策略 $\boldsymbol{A}_i$ 描述了该个体如何从其行动集 $\mathcal{A}_i$ 中选择行动. 个体 i 的策略可以由一个 m_i 维向量 $\boldsymbol{A}_i=(A_i^j)$ 表示, 其中, 每一个元素 A_i^j 表示个体 i 选择行动 a_i^j 的概率, 满足 $A_i^j\in[0,1]$ 且 $\sum_{j=1}^{m_i}A_i^j=1$. 若存在 $A_i^j=1, A_i^k=0$ $(k\neq j)$, 则称策略 $\boldsymbol{A}_i$ 为纯策略; 否则, 称该策略为混合策略. 一次博弈的结果由每个个体选择的行动组成, 记为 $\boldsymbol{a}=(a_1,\ldots,a_n)$; a_i 表示个体 i 选择的行动. 所有可能的博弈结果为个体行动集的笛卡尔积 $\mathcal{A}=\mathcal{A}_1\times\cdots\times\mathcal{A}_n$. 给定所有个体的策略 $\boldsymbol{A}_1,\ldots,\boldsymbol{A}_n$, 可以得到每个博弈结果 $\boldsymbol{a}$ 出现的概率:

$$\pi_{\boldsymbol{a}=(a_1,\ldots,a_n)}=\prod_{i=1}^{n}A_i^{a_i}.$$

用 U_i 表示个体 i 的收益, 则个体 i 的期望收益为

$$\mathbb{E}[U_i]=\sum_{\boldsymbol{a}\in\mathcal{A}}\pi_{\boldsymbol{a}}u_i(\boldsymbol{a}).$$

个体的期望收益衡量了该个体在博弈中的表现, 也衡量了该个体所采用的策略的优劣. 在收益控制中, 通常考虑控制个体的期望收益.

2.2 重复博弈

顾名思义, 重复博弈指重复地进行某个博弈 $\mathcal{G}$. 在合作的演化机制中, 重复博弈用于刻画相同的个体之间重复的相遇, 这也是直接互惠 (direct reciprocity) 的基本前提 [57]. 一般地, 重复博弈的模型可以表示为

$$\mathcal{G}_{\mathcal{R}} = \{\mathcal{A}_1, \ldots, \mathcal{A}_n; u_1, \ldots, u_n; c\},$$

其中, $c = c(t)$ 表示第 $t-1$ 轮博弈结束后, 第 t 轮博弈发生的概率. 通常假设在一轮博弈发生后, 下一轮博弈以一个固定的概率 δ $(0 < \delta \leqslant 1)$ 发生 $(c(t) = \delta)$. 当 $\delta \in (0,1)$ 时, 博弈平均进行 $1/(1-\delta)$ 轮. 在经济学中, δ 也被称为折现因子, 表示未来的收益折算到现在的比例 [7, 15]. 当 $\delta = 1$ 时, 下一轮博弈一定会发生, 此时重复博弈将进行无穷多轮, 故称 $c(t) = 1$ 的重复博弈为无穷轮重复博弈.

当参与博弈的个体重复交互时, 每个个体可以根据之前的博弈结果选择行动. 在重复博弈中, 一个个体的策略明确了该个体如何根据过去的博弈结果选择当前的行动. 随着博弈轮数的增加, 可能出现的博弈结果数呈指数增长, 个体可以采用的策略数也随之不断 "翻番". 目前发现的能控制收益的策略大多属于一步记忆策略 (memory-one strategy). 本文主要介绍一步记忆策略空间中能控制收益的策略. 除了通过不同记忆深度的观察表格选择行动的观察者策略 (looker-up strategy) 以外 [25], 目前还没有一般性的多步记忆策略 (memory-n strategy) 的相关成果.

采取一步记忆策略的个体只根据上一轮博弈的结果决定当前轮采取的行动. 对于重复博弈 $\mathcal{G} = \{\mathcal{A}_1, \ldots, \mathcal{A}_n; u_1, \ldots, u_n, c\}$, 个体 i 的一步记忆策略可以由一个 $(m_i - 1) \times (1 + \prod_{j=1}^{n} m_j)$ 维的矩阵 $\boldsymbol{P} = (\boldsymbol{p}_0; \boldsymbol{P}_{\mathcal{A}})$ 表示. 该矩阵的第一列为 $\boldsymbol{p}_0 = (p_{0,a_i^j})$, 每个元素 p_{0,a_i^j} 表示第一轮博弈中个体采取行动 a_i^j 的概率; 剩余部分为 $\boldsymbol{P}_{\mathcal{A}} = (p_{\boldsymbol{a},a_i^j})$, 每个元素 $p_{\boldsymbol{a},a_i^j}$ 表示当上一轮结果为 $\boldsymbol{a}$ 时个体 i 采取行动 a_i^j 的概率; 考虑到个体选择不同行动的概率和为 1, 我们省去第 m_i 个行动的概率. 如果一步记忆策略 $\boldsymbol{P}$ 在每种博弈结果下都确定性地选择某种行动 (即 $\forall \boldsymbol{a}, \exists a_i^j, \text{s.t.}\, p_{\boldsymbol{a},a_i^j} = 1$), 称该策略为纯一步记忆策略 (pure/deterministic memory-one strategy); 否则, 称其为随机一步记忆策略 (stochastic memory-one strategy). 特别地, 当个体 i 只有两个可能的行动时, 对应的一步记忆策略退化为向量形式. 例如, 在囚徒困境博弈中, 一步记忆策略由一个 5 维向量 $\boldsymbol{p} = (p_0; p_{CC}, p_{CD}, p_{DC}, p_{DD})$ 表示: p_0 表示该一步记忆策略在第一轮合作的概率; $p_{a_1a_2}$ $(a_1, a_2 \in \{C, D\})$ 表示该策略在上一轮选择行动 a_1 且对手策略选择行动 a_2 时合作的概率.

在重复博弈中, 个体 i 的收益可以视为整个博弈过程所得的平均收益 (即期望收益), 这一收益也被称为长期收益 (long-term payoff). 用 $U_i(t)$ 表示个体 i 在第 t 轮获得的收益, 其在重复博弈中获得的平均收益为

$$\mathbb{E}(U_i) = \lim_{t \to +\infty} \frac{C(1)U_i(1) + \cdots + C(t)U_i(t)}{C(1) + \cdots + C(t)}, \tag{3}$$

其中, $C(t) = \prod_{\tau=1}^{t} c(\tau)$ 为博弈至少发生 t 轮的概率. 当 $c(t) = \delta$ 且所有个体都采用随机策略时, 上述极限总是存在. 此时, 不需要通过求极限的方式计算个体的长期收益, 可以利用马尔可夫链 (Markov chain) 描述重复博弈的过程, 从而求解个体的长期收益. 注意到, 所有个体都采取随机策略是一个相对容易满足的假设: 在个体无法总是正确执行自己的策略 $\boldsymbol{P}$, 而是存在执行误差 ε 时 (即个体以概率 $1-\varepsilon$ 按照 $\boldsymbol{P}$ 采取行动, 以概率 $\varepsilon/(m_i-1)$ 采取其他行动), 个体的有效策略等价为 $(1-\varepsilon)\boldsymbol{P} + \varepsilon(1-\boldsymbol{P})/(m_i-1)$, 该有效策略总为随机策略. 可以证明, 当 $c(t) = \delta = 1$ 时, 个体获得的长期收益与第一轮选择各行动的概率 $\boldsymbol{p}_0$ 无关. 此时, 可以将一步记忆策略简化为 $\boldsymbol{P} = (\boldsymbol{P}_{\mathcal{A}})$.

2.3 策略的演化稳定性

1973 年, Maynard Smith 结合博弈论和进化论研究种群中个体行为 (策略) 的变化 [58], 这形成了一门交叉学科 —— 演化博弈论. 演化博弈论与经典博弈论的区别在于: (1) 演化博弈论不要求个体是完全理性的. 在演化过程中, 个体可以通过模仿更高收益的个体来更新自己的策略, 甚至通过自然选择实现策略的优胜劣汰, 而不是通过复杂的分析寻找最优策略 [59~72]. (2) 个体的收益受种群结构的影响. 演化博弈论的研究对象通常为一个种群, 其研究目标是刻画一个种群的整体特征. 在演化过程中, 我们从种群中随机选择 n 个个体进行博弈, 每个个体的交互对象是随机的, 收益也随之受到影响 [73~85]. 利用演化博弈的相关理论, 可以研究一种策略以突变的方式出现在种群后能否通过自然选择在种群中广泛存在; 以及种群是否会被其他外来策略入侵. 许多学者结合演化博弈论中的相关概念探索收益控制策略的演化特性, 进而从理论上分析这类策略在现实中是否可以获得自然选择的 "青睐". 本小节介绍演化博弈论中的 3 个重要概念: 演化稳定策略, 固定概率和演化鲁棒策略.

演化稳定策略 (evolutionarily stable strategy, ESS) 是纳什均衡策略的 "精细化表达" (refinement), 最早由 Maynard Smith 提出. 一个策略 $\boldsymbol{P}$ 仅在两种情况下为演化稳定策略: 对于任何突变策略, (1) 当与 $\boldsymbol{P}$ 博弈时, $\boldsymbol{P}$ 的收益严格大于突变策略的收益; (2) 若与 $\boldsymbol{P}$ 策略博弈时, 两策略的收益相等, 则还需要满足 $\boldsymbol{P}$ 与突变策略博弈的收益严格大于突变策略之间博弈的收益. 如果一个种群中所有个体采取同一演化稳定策略, 则即使少量个体发生突变, 种群也会恢复成原来的结构. 从系统的角度看, 由演化稳定策略组成的系统是抗干扰的.

演化稳定策略是种群无限大 (infinite population) 时衡量策略优劣的概念, 此时可以用复制动力学方程 (replicator equation) 来分析种群的稳态. 当种群中个体数目有限 (finite population) 时, 往往用固定概率 (fixation probability) 来刻画一个策略是否占优 [86]. 固定概率指单个突变个体成功占领整个种群的概率. 当种群中的个体等概率交互, 且以对比较 (pair comparison) 规则更新策略时, 一个突变策略 $\boldsymbol{P}$ 入侵由策略 $\boldsymbol{Q}$ 构成的种群的固定概率为

$$\rho_{\boldsymbol{P},\boldsymbol{Q}}=\frac{1}{1+\sum_{i=1}^{N-1}\prod_{j=1}^{i}\exp[-\beta(U_{\boldsymbol{P}}(j)-U_{\boldsymbol{Q}}(j))]},$$

其中, N 为种群规模; $U_{\boldsymbol{P}}(j)$ 和 $U_{\boldsymbol{Q}}(j)$ 分别表示当种群中有 j 个 $\boldsymbol{P}$ 个体时, $\boldsymbol{P}$ 和 $\boldsymbol{Q}$ 个体的平均收益. 对比较规则指更新策略的个体 X 以概率 $(1+\exp[-\beta(U_Y(j)-U_X(j))])^{-1}$ 学习个体 Y 的策略. β 是选择强度 (selection strength), 用来衡量博弈对策略更新的贡献 [87~89]. 当 $\beta=0$ 时, 该博弈对个体的策略更新没有影响, 对应的固定概率 ρ 为 $1/N$, 这种情况被称为中性漂移 (neutral drift). 通常用固定概率 ρ 与 $1/N$ 的相对大小衡量策略是否占优. 如果 $\rho_{\boldsymbol{P},\boldsymbol{Q}}>1/N$, 说明演化支持 $\boldsymbol{P}$ 策略入侵 $\boldsymbol{Q}$ 策略; 否则, $\boldsymbol{Q}$ 策略可以抵御 $\boldsymbol{P}$ 策略的入侵.

演化鲁棒策略 (evolutionarily robust strategy) 最初由 Stewart 和 Plotkin 提出 [90], 是衡量有限种群中一个策略是否稳定的概念. 如果策略 $\boldsymbol{P}$ 可以抵御任意突变个体的入侵 (即 $\forall\boldsymbol{Q},\rho_{\boldsymbol{Q},\boldsymbol{P}}\leqslant 1/N$), 则称策略 $\boldsymbol{P}$ 为演化鲁棒策略. 演化鲁棒策略对策略抵御入侵的能力做了严格限制, 但对该策略入侵其他策略的能力不做要求. 当种群大小 N 趋于无穷大时, 条件 $\forall\boldsymbol{Q},\rho_{\boldsymbol{Q},\boldsymbol{P}}<1/N$ 退化为演化稳定策略的条件 [90].

3 单次博弈中的收益控制与约束策略

本节介绍在博弈 $\mathcal{G}=\{\mathcal{A}_1,\ldots,\mathcal{A}_n;u_1,\ldots,u_n\}$ 中能控制收益的策略. Tan [37] 发现在博弈 $\mathcal{G}$ 中不仅存在可以使所有个体的收益满足线性关系的策略, 还存在使收益满足非线性关系的策略. 他将这类可以控制收益的策略命名为约束策略 (constraint strategy).

每个约束策略与它施加的约束条件相关联. 考虑任意的 n 元实函数 $f:\mathbb{R}^n\to\mathbb{R}$. 将所有参与博弈的个体的收益 $u_1(\boldsymbol{a}),\ldots,u_n(\boldsymbol{a})$ 作为 f 的输入, 则 f 将这些收益映射为一个实数. 函数 $f(\boldsymbol{U})=f(u_1(\boldsymbol{a}),\ldots,u_n(\boldsymbol{a}))$ 的值受博弈结果 $\boldsymbol{a}$ 的影响. 给定每个个体的策略 $\boldsymbol{A}_1,\ldots,\boldsymbol{A}_n$, 根据所有可能的博弈结果 $\mathcal{A}$, 可以得到 $f(\boldsymbol{U})$ 的期望:

$$\mathbb{E}[f(\boldsymbol{U})]=\sum_{\boldsymbol{a}\in\mathcal{A}}f(u_1(\boldsymbol{a}),\ldots,u_n(\boldsymbol{a}))\pi_{\boldsymbol{a}}. \tag{4}$$

称 $\mathbb{E}[f(\boldsymbol{U})]$ 为一个约束条件. 注意到, 任意个体策略的改变都会引起约束条件的改变. 下面给出与约束条件相关联的约束策略的定义.

定义1 (约束策略) 在博弈 $\mathcal{G}=\{\mathcal{A}_1,\ldots,\mathcal{A}_n;u_1,\ldots,u_n\}$ 中, 如果一个个体通过使用策略 $\boldsymbol{A}^*$ 可以单方面地使所有个体的收益 $\boldsymbol{U}=(U_1,\ldots,U_n)$ 满足约束条件 $\mathbb{E}[f(\boldsymbol{U})]=0$, 则称策略 $\boldsymbol{A}^*$ 为对应于约束条件 f 的约束策略.

给定博弈的基本模型 $\mathcal{G}$, 约束条件 $f(u_1(\boldsymbol{a}),\ldots,u_n(\boldsymbol{a}))$ 可以视为关于 $\boldsymbol{a}$ 的一元函数. 进而, $\mathbb{E}[f(\boldsymbol{U})]$ 可以写作内积的形式:

$$\begin{aligned}\mathbb{E}[f(\boldsymbol{U})]&=\sum_{\boldsymbol{a}\in\mathcal{A}}f(u_1(\boldsymbol{a}),\ldots,u_n(\boldsymbol{a}))\pi_{\boldsymbol{a}}\\&=\sum_{\boldsymbol{a}\in\mathcal{A}}f\circ(u_1,\ldots,u_n)(\boldsymbol{a})\pi_{\boldsymbol{a}}\\&=\sum_{\boldsymbol{a}\in\mathcal{A}}f\circ\boldsymbol{u}(\boldsymbol{a})\pi_{\boldsymbol{a}}\\&=(f\circ\boldsymbol{u})\cdot\pi,\end{aligned}$$

其中, $f\circ\boldsymbol{u}$ 是函数 f 与函数 $\boldsymbol{u}$ 的复合函数, π 是博弈结果的稳态概率分布. 由于博弈结果 $\boldsymbol{a}$ 的个数是有限的, $f\circ\boldsymbol{u}$ 可以通过向量的形式表示, 其元素与集合 $\mathcal{A}$ 中的元素是一一映射的. 上式说明, 对应于约束条件 f 的约束策略 $\boldsymbol{A}^*$ 可以使函数 $f\circ\boldsymbol{u}$ 始终与 π 正交 ($(f\circ\boldsymbol{u})\cdot\pi=0$). 当某个个体使用约束策略 $\boldsymbol{A}^*$ 时, 无论其他个体选择什么策略, 向量 $f\circ\boldsymbol{u}$ 始终位于 π 的正交子空间内. 如果我们固定个体 i 的策略 $\boldsymbol{A}^*$, 而让其他个体任意选择自己的策略, 得到的所有可能的 π 所形成的 "共有正交子空间" 便被称为个体 i 的约束策略 $\boldsymbol{A}^*$ 的约束空间, 记作 $V(\boldsymbol{A}_i^*)$. $V(\boldsymbol{A}_i^*)$ 满足

$$V(\boldsymbol{A}_i^*)=\{\boldsymbol{v}|\boldsymbol{v}\cdot\pi=0,\forall\pi,\boldsymbol{A}_i=\boldsymbol{A}^*\}. \tag{5}$$

根据上述定义, 可以得到对于个体 i, 约束策略 $\boldsymbol{A}^*$ 存在的充要条件.

定理1 (约束策略的充要条件) 在博弈 $\mathcal{G}=\{\mathcal{A}_1,\ldots,\mathcal{A}_n;u_1,\ldots,u_n\}$ 中, 对应于约束条件 f 的策略 $\boldsymbol{A}^*$ 为个体 i 的约束策略, 当且仅当函数 $f\circ\boldsymbol{u}$ 属于与策略 $\boldsymbol{A}^*$ 对应的约束空间 $V(\boldsymbol{A}_i^*)$ 且 $f\circ\boldsymbol{u}$ 不为零函数, 即

$$f\circ\boldsymbol{u}\in V(\boldsymbol{A}_i^*)\text{ 且 }f\circ\boldsymbol{u}\neq 0.$$

在此基础上, Tan [37] 给出了约束空间的一组基向量.

定理2 (约束空间的基向量) 考虑博弈 $\mathcal{G}=\{\mathcal{A}_1,\ldots,\mathcal{A}_n;u_1,\ldots,u_n\}$, 对应于策略 $\boldsymbol{A}_i$ 的约束空间的一组基为 $\{\boldsymbol{v}^{a_i^j}|a_i^j\in\mathcal{A}\backslash\{a_i^{j^*}\}\}$, 其中,

$$v_{\boldsymbol{a}}^{a_i^j}=\begin{cases}A_i^{j*}, & \text{当 } \boldsymbol{a}=(\ldots,a_i^j,\ldots),\\ -A_i^j, & \text{当 } \boldsymbol{a}=(\ldots,a_i^{j^*},\ldots),\\ 0, & \text{其他,}\end{cases}$$

$a_i^{j^*}$ 为个体 i 可能选择的某一行动 (即 $A_i^{j^*}\neq 0$), 在不失普遍性的情况下, 可以令 $j^*=1$ (即 $A_i^1\neq 0$).

定理 1 与 2 明确了约束策略的形式. 利用定理 1 和 2 可以判断一个策略 $\boldsymbol{A}$ 是否为对应于约束条件 f 的约束策略; 也可以去确定一个策略 $\boldsymbol{A}$ 能够实现的收益关系 f; 还可以去搜索能够施加收益关系 f 的策略 $\boldsymbol{A}$.

约束策略的发现表明在单次博弈中存在能单方面将收益约束为 $\mathbb{E}[f(U)]=0$ 的策略. 这里对 f 的具体形式没有限制. 个体既可以将收益约束为线性形式, 也可以将其约束为非线性形式. 约束空间是一个线性空间. 两个约束策略数乘和加减运算后得到的策略仍为约束策略. 从几何角度看, 所有的约束策略构成一个凸集. 这为寻找特定的约束策略提供了便利.

4 重复博弈中的收益控制与零行列式策略

对于重复博弈, Boerlijst 等于 1997 年构造了一种可以令对手的期望收益等于惩罚 P 和奖励 R 之间的任意值的策略, 由于这种特殊的性质 —— 等于 (equal to), 该策略被称为均衡器策略. 更一般的收益控制策略由 Press 和 Dyson 在 2012 年提出. 他们发现存在一种策略, 可以单方面地令自己和对手的收益满足一个线性关系式. 他们将这类策略命名为零行列式策略. 这一发现颠覆了人们对单个策略控制能力的认识, 因此备受关注.

4.1 零行列式策略的发现

研究者们最初在无穷轮重复囚徒困境中研究零行列式策略, 其发现过程与个体长期收益的计算方法密切相关. 如果我们用马尔可夫链表示重复囚徒困境的博弈过程, 每一轮可能出现的结果 $\mathcal{A}=\{CC,CD,DC,DD\}$, 即为马尔可夫链的状态空间. 假设参与博弈的两个个体 X 和 Y 分别选择一步记忆策略 $\boldsymbol{p}=(p_{CC},p_{CD},p_{DC},p_{DD})$ 和 $\boldsymbol{q}=(q_{CC},q_{CD},q_{DC},q_{DD})$, 那么该马尔可夫链的转移矩阵为

$$\boldsymbol{M}=\begin{array}{c}\\ CC\\ CD\\ DC\\ DD\end{array}\begin{array}{c}\begin{array}{cccc}CC & CD & DC & DD\end{array}\\ \begin{pmatrix}p_{CC}q_{CC} & p_{CC}(1-q_{CC}) & (1-p_{CC})q_{CC} & (1-p_{CC})(1-q_{CC})\\ p_{CD}q_{DC} & p_{CD}(1-q_{DC}) & (1-p_{CD})q_{DC} & (1-p_{CD})(1-q_{DC})\\ p_{DC}q_{CD} & p_{DC}(1-q_{CD}) & (1-p_{DC})q_{CD} & (1-p_{DC})(1-q_{CD})\\ p_{DD}q_{DD} & p_{DD}(1-q_{DD}) & (1-p_{DD})q_{DD} & (1-p_{DD})(1-q_{DD})\end{pmatrix}\end{array}. \tag{6}$$

当个体 X 和 Y 采取的不是纯策略时, 该马尔可夫链是不可约的. 概率转移矩阵 $\boldsymbol{M}$ 有对应于特征值 1 唯一的左特征向量 $\boldsymbol{v}$ (即 $\boldsymbol{v}\boldsymbol{M}=\boldsymbol{v}$). 将该特征向量归一化为分布 $\boldsymbol{v}=(v_{CC},v_{CD},v_{DC},v_{DD})$ 后, 我们便得到无穷轮重复博弈中每种博弈结果的平均分布. 用 $\boldsymbol{S}_X=(R,S,T,P)$ 和 $\boldsymbol{S}_Y=(R,T,S,P)$ 分别表示个体 X 和 Y 的收益向量, 这两个个体的长期收益分别为 $\mathbb{E}[U_X]=\boldsymbol{v}\cdot\boldsymbol{S}_X$ 和 $\mathbb{E}[U_Y]=\boldsymbol{v}\cdot\boldsymbol{S}_Y$.

以往的大多数文献通过直接计算平均分布 $\boldsymbol{v}$ 来得到个体的长期收益, 然而在很多情况下 $\boldsymbol{v}$ 的表达是极其繁琐的. Press 和 Dyson 提出了一种更为巧妙的方法, 不需要计算平均分布 $\boldsymbol{v}$ 便可以得到个体的长期收益. 这种方法是发现零行列式策略的关键. 我们对其进行详细介绍.

记 $\boldsymbol{M}' = \boldsymbol{M} - \boldsymbol{I}$, 有

$$\boldsymbol{v}\boldsymbol{M}' = \boldsymbol{0}. \tag{7}$$

根据上式, 平均分布 $\boldsymbol{v}$ 是 $\boldsymbol{M}'$ 对应于特征值 0 的非平凡的特征向量, 因此 $\det(\boldsymbol{M}') = 0$. 结合克莱姆法则 (Cramer's rule), 得到

$$\mathrm{Adj}(\boldsymbol{M}')\boldsymbol{M}' = \det(\boldsymbol{M}')\boldsymbol{I} = \boldsymbol{0}, \tag{8}$$

其中, $\det(\boldsymbol{M}')$ 表示矩阵 $\boldsymbol{M}'$ 的行列式; $\mathrm{Adj}(\boldsymbol{M}')$ 表示矩阵 $\boldsymbol{M}'$ 的伴随矩阵; $\boldsymbol{I}$ 和 $\boldsymbol{0}$ 分别为对应维度的单位矩阵和零矩阵. 对比式 (7) 与 (8), 可以发现 $\boldsymbol{v}$ 应该与 $\mathrm{Adj}(\boldsymbol{M}')$ 的每一行都成比例. 不失一般性, 我们假设 $\boldsymbol{v}$ 为 $\mathrm{Adj}(\boldsymbol{M}')$ 的第 4 行. 对于任意的向量 $\boldsymbol{f} = (f_1, f_2, f_3, f_4) \in \mathbb{R}^{1\times 4}$, 有

$$\boldsymbol{v}\cdot\boldsymbol{f} = (M'_{1,4}, M'_{2,4}, M'_{3,4}, M'_{4,4})\cdot\boldsymbol{f}. \tag{9}$$

此处, $M'_{i,j}$ 表示 $\boldsymbol{M}'$ 第 i 行第 j 列元素的代数余子式. 将式 (9) 写为矩阵行列式的形式并将第 1 列分别加到第 2 列和第 3 列, 可得

$$\boldsymbol{v}\cdot\boldsymbol{f} = \det\begin{pmatrix} -1+p_{CC}q_{CC} & -1+p_{CC} & -1+q_{CC} & f_1 \\ p_{CD}q_{DC} & -1+p_{CD} & q_{DC} & f_2 \\ p_{DC}q_{CD} & p_{CD} & -1+q_{CD} & f_3 \\ p_{DD}q_{DD} & p_{DD} & q_{DD} & f_4 \end{pmatrix}. \tag{10}$$

令 $D(\boldsymbol{p},\boldsymbol{q},\boldsymbol{f}) \triangleq \boldsymbol{v}\cdot\boldsymbol{f}$. 个体 X 和 Y 的长期收益可以表示为

$$\begin{aligned} \mathbb{E}(U_X) &= \frac{\boldsymbol{v}\cdot\boldsymbol{S}_X}{\boldsymbol{v}\cdot\boldsymbol{1}} = \frac{D(\boldsymbol{p},\boldsymbol{q},\boldsymbol{S}_X)}{D(\boldsymbol{p},\boldsymbol{q},\boldsymbol{1})}, \\ \mathbb{E}(U_Y) &= \frac{\boldsymbol{v}\cdot\boldsymbol{S}_Y}{\boldsymbol{v}\cdot\boldsymbol{1}} = \frac{D(\boldsymbol{p},\boldsymbol{q},\boldsymbol{S}_Y)}{D(\boldsymbol{p},\boldsymbol{q},\boldsymbol{1})}, \end{aligned} \tag{11}$$

其中, $\boldsymbol{1}$ 是所有元素都为 1 的向量; 分母部分起到将 $\boldsymbol{v}$ 归一化为分布的作用. 式 (11) 说明个体 X 和 Y 的收益线性依赖于它们的收益向量 $\boldsymbol{S}_X$ 和 $\boldsymbol{S}_Y$. 因此任何关于 $\mathbb{E}(U_X)$ 和 $\mathbb{E}(U_Y)$ 的线性运算都可以通过矩阵行列式的形式表示:

$$\alpha\mathbb{E}(U_X) + \beta\mathbb{E}(U_Y) + \gamma = \frac{D(\boldsymbol{p},\boldsymbol{q},\alpha\boldsymbol{S}_X + \beta\boldsymbol{S}_Y + \gamma\boldsymbol{1})}{D(\boldsymbol{p},\boldsymbol{q},\boldsymbol{1})}. \tag{12}$$

注意到, 式 (10) 中行列式 $D(\boldsymbol{p},\boldsymbol{q},\boldsymbol{f})$ 的第 2 列元素完全由个体 X 决定; 第 3 列元素则完全由个体 Y 决定. 记 $\boldsymbol{p}^{\mathrm{Rep}} = (1,1,0,0)$ 为重复策略 (repeat) [91]. 当 $\boldsymbol{f} = \alpha\boldsymbol{S}_X + \beta\boldsymbol{S}_Y + \gamma\boldsymbol{1}$ 时, 如果个体 X 或个体 Y 采取策略 $\boldsymbol{p}^{\mathrm{Rep}} + \alpha\boldsymbol{S}_X + \beta\boldsymbol{S}_Y + \gamma\boldsymbol{1}$, $D(\boldsymbol{p},\boldsymbol{q},\boldsymbol{f})$ 的第 2 列或第 3 列将正比于最后一列, 因此该行列式为零, 进一步有

$$\alpha\mathbb{E}(U_X) + \beta\mathbb{E}(U_Y) + \gamma = 0. \tag{13}$$

因为 $D(\boldsymbol{p},\boldsymbol{q},\boldsymbol{f}) = 0$ 的出现, Press 和 Dyson 便将这类策略命名为零行列式策略.

定义2 (零行列式策略) 在无穷轮重复囚徒困境博弈中, 形如

$$\boldsymbol{p} = \boldsymbol{p}^{\mathrm{Rep}} + \alpha \boldsymbol{S}_X + \beta \boldsymbol{S}_Y + \gamma \boldsymbol{1} \tag{14}$$

的策略被称为零行列式策略. 其中, $\boldsymbol{S}_X$ 和 $\boldsymbol{S}_Y$ 表示个体 X 和 Y 的收益向量; $\boldsymbol{p}^{\mathrm{Rep}} = (1,1,0,0)$ 表示重复策略.

代入任意的零行列式策略 $\boldsymbol{p}$, 矩阵行列式 $D(\boldsymbol{p},\boldsymbol{q},\boldsymbol{f})$ 都为零, 从而 $\boldsymbol{f}$ 对应的线性表达式的值 (式 (12) 等号左边) 也为零. 于是有如下定理.

定理3 如果一个个体采取零行列式策略 $\boldsymbol{p} = \boldsymbol{p}^{\mathrm{Rep}} + \alpha \boldsymbol{S}_X + \beta \boldsymbol{S}_Y + \gamma \boldsymbol{1}$, 那么它可以单方面地使自己的长期收益 $\mathbb{E}(U_X)$ 和对手的长期收益 $\mathbb{E}(U_Y)$ 满足如下的线性关系:

$$\alpha \mathbb{E}(U_X) + \beta \mathbb{E}(U_Y) + \gamma = 0. \tag{15}$$

值得注意的是, 策略 $\boldsymbol{p}$ 的每个元素 $p_{a_1 a_2}$ 都表示一个概率, 因此都被限制在单位区间 $[0,1]$ 内. 这导致了零行列式策略个体能够实现的线形收益关系是有限制的. 特别地, 个体无法利用零行列式策略设置自己的长期收益 (此时 $\beta = 0$) [30].

4.2 几类具体的零行列式策略

零行列式策略是一类一步记忆策略, 它在一步记忆策略的表达上额外添加了一个约束, 因此会减少一个自由度. 在无穷轮重复囚徒困境博弈中, 和一般的一步记忆策略比, 零行列式策略的自由度从 4 降为 3, 对应有 3 个可以调节的参数 α, β, γ. 通过改变这些参数, 个体可以对收益施加不同类型的线性关系 [92]. 本小节将介绍几类具体的零行列式策略, 包括均衡器策略, 敲诈策略 (extortionate ZD strategy) 和慷慨策略 (generous ZD strategy).

均衡器策略最初由 Boerlijst 等 [29] 发现. 对于任意的收益水平 u ($P \leqslant u \leqslant R$) 和归一化因子 a ($0 \leqslant a \leqslant 1/\max\{T-u, u-S\}$), 只要个体 X 的策略 $\boldsymbol{p}$ 满足

$$p_{CC} = 1-(R-u)a,\ \ p_{CD} = 1-(T-u)a,\ \ p_{DC} = (u-S)a,\ \ p_{DD} = (u-P)a,$$

无论个体 Y 选择何种策略, 其收益都会被固定为 u. 等价地, Press 和 Dyson 发现通过设置 $\alpha = 0$ 和 $\beta \neq 0$, 均衡器策略可以单方面地将对手的收益固定为 $-\gamma/\beta$. 只要个体 X 采取均衡器策略

$$\boldsymbol{p} = \boldsymbol{p}^{\mathrm{Rep}} + \beta \boldsymbol{S}_Y + \gamma \boldsymbol{1},$$

个体 Y 始终无法摆脱收益为 $-\gamma/\beta$ 的命运 (图 1(b)). 值得注意的是, 虽然控制者 X 可以单方面地决定被控制者 Y 的收益, 但是却无法阻止 Y 反过来对 X 自己的收益产生影响. 当个体 Y 采取不同策略时, 个体 X 的收益也会不同. 因此, 如果个体 Y 发现自己的收益被个体 X 控制, 它可以通过调整自己的策略 "反抗" 对手, 使其获得较低的收益, 从而令个体 X 不得不放弃控制收益的想法.

敲诈策略是 Press 和 Dyson 提出的一种最贪婪的零行列式策略 [30]. 在讨论的大多数重复囚徒困境博弈中, 采用敲诈策略的个体总能够获得比对手更高的收益 (图 1(c)). 令 $\alpha = \phi, \beta = -\phi\chi, \gamma = \phi(\chi-1)O$, 零行列式策略可以等价地表示为

$$\boldsymbol{p} = \boldsymbol{p}^{\mathrm{Rep}} + \phi[(\boldsymbol{S}_X - O\boldsymbol{1}) - \chi(\boldsymbol{S}_Y - O\boldsymbol{1})], \tag{16}$$

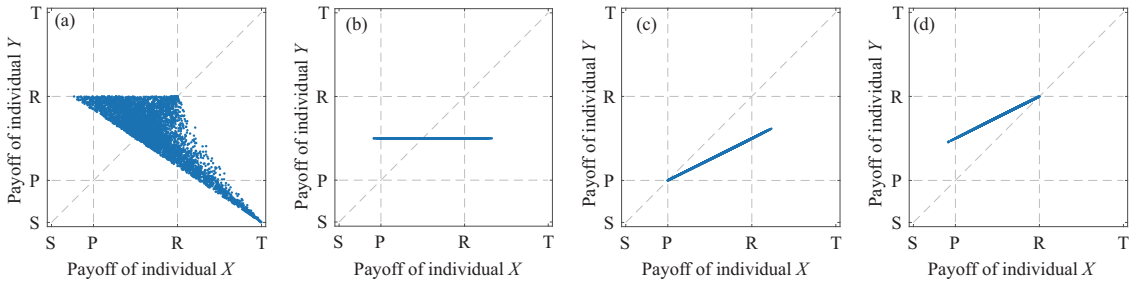

图 1 (网络版彩图) 几类零行列式策略

Figure 1 (Color online) Examples of zero-determinant strategies. (a) WSLS; (b) equalizer; (c) extortionate ZD strategy; (d) generous ZD strategy

其中, O 和 χ 分别被称为基准收益 (baseline payoff) 和敲诈系数 (extortion factor). 当 $O=P$ 且 $\chi>1$ 时, 可以得到敲诈策略的一般表示:

$$\boldsymbol{p}=\boldsymbol{p}^{\mathrm{Rep}}+\phi[(\boldsymbol{S}_X-P\boldsymbol{1})-\chi(\boldsymbol{S}_Y-P\boldsymbol{1})]. \tag{17}$$

根据定理 3, 当个体 X 使用敲诈策略时, 它可以施加线性关系

$$\mathbb{E}(U_X)-P=\chi(\mathbb{E}(U_Y)-P). \tag{18}$$

如果 $T+S>2P$ (大多数囚徒困境的收益矩阵满足该不等式), 由上式可以得到 $\mathbb{E}(U_Y)\geqslant P$. 又因为 $\chi>1$, 个体 X 的收益高于 P 的部分总是大于个体 Y 的收益高于 P 的部分, 所以个体 X 获得的收益总是多于个体 Y 获得的收益. 这也是用 "敲诈" 命名这类策略的原因. 有趣的是, 如果个体 Y 也采用敲诈策略, 那么无论二者的敲诈系数 χ 孰高孰低, 个体 X 和 Y 的收益都将是相互背叛的收益 P. 这表明当采取敲诈策略的个体处于一个与自己的策略相同或相似的群体中时, 会出现 "同室操戈" 的现象, 个体获得的收益都很不理想. 有另外一类零行列式策略 —— 慷慨策略 —— 能够避免这类情况出现.

受到 Press 和 Dyson 的启发 [30], Stewart 和 Plotkin 在零行列式策略发现的第二年提出了慷慨策略 [90], 其形式为

$$\boldsymbol{p}=\boldsymbol{p}^{\mathrm{Rep}}+\phi[(\boldsymbol{S}_X-R\boldsymbol{1})-\chi(\boldsymbol{S}_Y-R\boldsymbol{1})]. \tag{19}$$

采用慷慨策略的个体能使自己和对手的收益满足

$$\mathbb{E}(U_X)-R=\chi(\mathbb{E}(U_Y)-R). \tag{20}$$

与敲诈策略不同, 慷慨策略选择的基准收益 $O=R$, 采用该策略的个体不再把收益和相互背叛的惩罚 P 作比较, 而是和相互合作的奖励 R 作比较. 同时, 该策略依旧要求系数 $\chi>1$. 鉴于 $T+S<2R$ (相互合作的收益和大于合作与背叛交替出现的收益和), 由式 (20) 可得 $\mathbb{E}(U_Y)\leqslant R$, 因此采用慷慨策略的个体总是获得比对手更少的收益 (图 1(d)). 但是当个体 Y 的策略也为慷慨策略时, 两个个体获得的收益均为 R. 于是慷慨策略在同类群体中总能够获得较高的收益.

图 1 给出了这几类具体的零行列式策略与任意一步记忆策略博弈的收益关系图. 每一个子图中的个体 X 固定采取某一类具体的零行列式策略, 个体 Y 随机采取一个一步记忆策略. 作为对照, 图 1(a) 中的个体 X 采取一个在演化中表现非常出众的非零行列式策略 —— 赢留输变策略 (win-stay lose-shift, WSLS, $\boldsymbol{p}=(1,0,0,1)$) [93]. 此时, 个体 X 的收益与个体 Y 的收益会形成一个三角形区域. 图

1(b) 中个体 X 采取均衡器策略 $\boldsymbol{p}=(0.8,0.4,0.4,0.2)$, 个体 Y 的收益被固定为 $(R+P)/2$. 图 1(c) 中个体 X 采取敲诈策略 $\boldsymbol{p}=(0.8,0.1,0.6,0)$, 个体 X 和 Y 的收益满足 $\mathbb{E}(U_X)-P=2(\mathbb{E}(U_Y)-P)$. 图 1(d) 中个体 X 采取慷慨策略 $\boldsymbol{p}=(1,0.3,0.8,0.2)$, 两个个体的收益满足 $\mathbb{E}(U_X)-R=2(\mathbb{E}(U_Y)-R)$. 所使用的收益矩阵和 Axelrod 计算机锦标赛 [94~96] 保持一致: $T=5,R=3,P=1,S=0$.

除了均衡器策略、敲诈策略和慷慨策略以外, 在无穷轮重复囚徒困境中, 著名的以牙还牙策略 (tit-for-tat) [97] 也是一种零行列式策略 [98]; 此时, 经过基准收益 O 调整后的收益比 $\chi=1$, 采取以牙还牙策略的个体总是获得和对手一样高的收益 ("the white sheep of the family"[1)]), 以牙还牙策略是贪婪的敲诈策略中的一个友好的策略). 具有这一特性的策略也被称为公平策略 (fair strategy). 对于更一般的重复博弈, 其他重要的策略在某些博弈中也为零行列式策略. 比如无条件合作策略 (always cooperate, AllC) 和无条件背叛策略 (always defect, AllD) 在无穷轮重复公共品博弈中可以被看作边界上的零行列式策略 [98]: 无条件合作策略是一种收益比最大的慷慨策略; 无条件背叛策略则是一种收益比最大的敲诈策略.

4.3 零行列式策略的存在性

Press 和 Dyson 在无穷轮重复囚徒困境中发现的零行列式策略为收益控制提供了一个可以借鉴的思路. 本小节进一步介绍这种能够将收益控制为线性形式的策略在其他类型重复博弈中的存在性.

将收益控制策略从无穷轮重复囚徒困境推广到收益恒定的重复博弈的过程中, Akin 发现的一步记忆策略和平均分布之间的潜在关系起到了关键作用 [91].

定理4 (Akin) 对于重复博弈 $\mathcal{G}_{\mathcal{R}}=\{\mathcal{A}_1,\ldots,\mathcal{A}_n;u_1,\ldots,u_n;c\}$, 如果每个个体只有两个可以选择的行动 C 和 D 且博弈进行无穷多轮 (即 $\mathcal{A}_i=\{C,D\}$ 且 $c(t)=1$), 那么一步记忆策略 $\boldsymbol{p}$ 与博弈结果的平均分布 $\boldsymbol{v}$ 之间存在如下关系:

$$\boldsymbol{v}\cdot(\boldsymbol{p}-\boldsymbol{p}^{\mathrm{Rep}})=0, \tag{21}$$

其中, $\boldsymbol{v}$ 为 $\boldsymbol{p}$ 与其他策略博弈所产生的平均分布; $\boldsymbol{p}^{\mathrm{Rep}}$ 为重复策略.

证明定理 4 的过程中, Akin 不再像 Press 和 Dyson 一样列出概率转移矩阵的具体形式, 而是把合作率作为桥梁将个体策略和博弈的平均分布建立联系, 并通过极限运算来证明. 这一证明方法更加简洁, 也更容易推广到重复多人博弈和多行为博弈中.

利用定理 4, 可以证明无穷轮重复多人博弈中也存在零行列式策略 [27,35,99]. 对于个体 i, 收益控制策略的形式为

$$\boldsymbol{p}=\boldsymbol{p}^{\mathrm{Rep}}+\phi\left[s\boldsymbol{S}_i-\sum_{j\neq i}w_j\boldsymbol{S}_j+(1-s)l\boldsymbol{1}\right], \tag{22}$$

其中, $\boldsymbol{S}_i$ 为个体 i 的收益向量; s,w_j,l,ϕ 为可以调节的参数且 $\sum_{j\neq i}w_j=1$. 在重复多人博弈中, 采取零行列式策略的个体 i 能够使参与博弈的全部个体的收益满足

$$s(\mathbb{E}(U_i)-l)=\sum_{j\neq i}w_j\mathbb{E}(U_j)-l, \tag{23}$$

其中, l 被称为基准收益, 因为如果所有个体都采取零行列式策略, 每个个体获得的收益均为 l. 此外, s 被称为零行列式策略的斜率, 因为采取此策略的个体的收益与其他个体加权平均 (权重为 w_j) 后的收

1) On "iterated prisoner's dilemma contains strategies that dominate any evolutionary opponent". https://www.edge.org/conversation/william_h_press-freeman_dyson-on-iterated-prisoners-dilemma-contains-strategies-that. Accessed: 2022-06-21.

益所形成的直线的斜率为 s. 将 $\sum_{j\neq i} w_j \mathbb{E}(U_j)$ 视为一个整体, s 和 l 分别与式 (16) 中的 χ 和 O 相对应, 其中 $l = O$ 表示基准收益, $\chi = 1/s$ 表示敲诈系数. 在无穷轮重复多人博弈中, 同样存在均衡器策略、敲诈策略和慷慨策略. 但此时使用零行列式策略的个体不能控制每一个个体的收益, 而只能对某一个个体的收益或者其他个体收益的加权平均进行控制. 另外, 随着多人博弈中参与博弈人数的增加, 零行列式策略在一步记忆策略空间中的占比也不断减小. 这两点都说明在人数越多的博弈中, 越难进行收益控制.

当下一轮博弈以固定的概率 δ $(0 < \delta < 1)$ 发生时, 定理 4 中的式 (21) 变为

$$\boldsymbol{v} \cdot (\delta \boldsymbol{p} - \boldsymbol{p}^{\mathrm{Rep}}) = -(1-\delta)p_0,$$

其中, p_0 表示在第一轮合作的概率. 根据上式, 此时零行列式策略具有形式

$$\delta \boldsymbol{p} = \boldsymbol{p}^{\mathrm{Rep}} + \phi \left[s\boldsymbol{S}_i - \sum_{j\neq i} w_j \boldsymbol{S}_j + (1-s)l\mathbf{1} \right] - (1-\delta)p_0\mathbf{1}.$$

在 δ 严格小于 1 的博弈中, 以牙还牙策略不再是零行列式策略 [28]. 并且, 这样的博弈中也不存在其他的公平策略. 除上述两类重复博弈类型, 学者们发现在行动集为连续空间的重复博弈 [100, 101]、个体交替行动的重复博弈 [49]、非对称博弈 [56] 以及有误差 [55, 102~104] 的重复博弈中也存在能将收益控制为线性关系的策略.

对于一般的收益恒定的重复博弈 $\mathcal{G}_{\mathcal{R}}$, Tan 等 [36] 发现当且仅当下一轮博弈发生的概率保持不变时, 才存在零行列式策略.

定理5 (收益控制策略的存在性) 对于重复博弈 $\mathcal{G}_{\mathcal{R}} = \{\mathcal{A}_1, \ldots, \mathcal{A}_n; u_1, \ldots, u_n; c\}$, 假设 $n = 2$ 且 $\mathcal{A}_i = \{a_i^1, a_i^2\}$. 零行列式策略只在以下两种情况中存在:

(1) 下一轮博弈必然发生, 博弈重复进行无限轮, 即 $c(t) = 1, \forall t$;

(2) 下一轮博弈以固定的概率 $\delta \in (0,1)$ 发生, 即 $c(t) = \delta$ $(0 < \delta < 1), \forall t$.

当 $n = 2$ 时, 定理 5 是定理 4 的拓展, 说明零行列式策略的存在性仅由重复方式 $c(t)$ 决定. 只有当博弈发生无穷多轮或者下一轮博弈以固定概率 δ $(0 < \delta < 1)$ 发生时, 才存在零行列式策略. 这为寻找零行列式策略确定了范围.

4.4 零行列式策略的判别

在 4.3 小节末尾对博弈类型做出的限制下, 本小节进一步介绍如何判断一个一步记忆策略是否为零行列式策略. 为此, 首先引入约束向量 (ruling vector) 的概念.

定义3 (约束向量) 对于重复博弈 $\mathcal{G}_{\mathcal{R}} = \{\mathcal{A}_1, \ldots, \mathcal{A}_n; u_1, \ldots, u_n; c\}$, 如果一个向量 $\boldsymbol{r}$ 由个体 i 的策略 $\boldsymbol{P}_i$ 单独决定, 且 $\boldsymbol{v} \cdot \boldsymbol{r} = 0$, 则称向量 $\boldsymbol{r}$ 为一个约束向量. 这里, $\boldsymbol{v}$ 为 $\boldsymbol{P}_i$ 与其他策略博弈的结果的平均分布.

约束向量是寻找零行列式策略的中间量. 特别地, 在无穷轮重复囚徒困境中, $\boldsymbol{p} - \boldsymbol{p}^{\mathrm{Rep}}$ 为零行列式策略 $\boldsymbol{p}$ 的约束向量; 在下一轮博弈以概率 δ $(0 < \delta < 1)$ 发生的重复博弈中, $\delta\boldsymbol{p} - \boldsymbol{p}^{\mathrm{Rep}} + (1-\delta)p_0\mathbf{1}$ 为零行列式策略 $\boldsymbol{p}$ 的约束向量. 对于一般的重复博弈 $\mathcal{G}_{\mathcal{R}} = \{\mathcal{A}_1, \ldots, \mathcal{A}_n; u_1, \ldots, u_n; c\}$ 和策略 $\boldsymbol{P}$, Tan 等 [36] 有如下发现:

(1) 当博弈重复进行无限轮时 $(c(t) = 1, \forall t)$, 向量 $\boldsymbol{p}_{a_i^j} - \boldsymbol{p}^{\mathrm{Rep}}_{a_i^j}$ 为约束向量;

(2) 当下一轮博弈以固定概率 $\delta \in (0,1)$ 发生时 $(c(t)=\delta, \forall t)$, 向量 $\delta \boldsymbol{p}_{a_i^j} - \boldsymbol{p}_{a_i^j}^{\mathrm{Rep}} + (1-\delta) p_{0,a_i^j} \mathbf{1}$ 为约束向量.

其中, $\boldsymbol{p}_{a_i^j}$ 表示博弈 $\mathcal{G}_{\mathcal{R}}$ 中策略 $\boldsymbol{P}$ 关于行动 a_i^j 的分量 (即个体在每一种可能的博弈结果下采取行动 a_i^j 的概率所构成的向量); $\boldsymbol{p}_{a_i^j}^{\mathrm{Rep}}$ 表示重复策略 $\boldsymbol{P}^{\mathrm{Rep}}$ 关于行动 a_i^j 的分量, 使用该策略的个体总是重复上一轮的行动; p_{0,a_i^j} 表示第一轮博弈中个体采取行动 a_i^j 的概率. 可以发现, 当个体有两个以上可以选择的行动时, 一个策略 $\boldsymbol{P}$ 有可能单独决定不止一个线性无关的约束向量. 由于内积运算对加减法和数乘运算封闭, 约束向量的线性组合也为约束向量. 这些约束策略因而会生成一个线性子空间, 利用这个子空间, 可以判断一个一步记忆策略是否为零行列式策略 [36].

定理6 (零行列式策略的判别) 在一般的重复博弈 $\mathcal{G}_{\mathcal{R}}$ 中, 一个一步记忆策略 $\boldsymbol{P}$ 为零行列式策略的充要条件为

$$\mathrm{span}\{\boldsymbol{S}_1, \ldots, \boldsymbol{S}_n, \mathbf{1}\} \cap \mathrm{span}\{\boldsymbol{r}_1, \ldots, \boldsymbol{r}_m\} \neq \emptyset, \tag{24}$$

其中, $\boldsymbol{S}_i$ 为个体 i 的收益向量, $\boldsymbol{r}_j$ 为策略 $\boldsymbol{P}$ 可以单独控制的第 j 个约束向量; $\mathrm{span}\{\cdot\}$ 为 $\cdot$ 生成的子空间; $\emptyset$ 表示空集.

对于一个一步记忆策略 $\boldsymbol{P}$, 如果式 (24) 成立, 则存在向量 $\alpha = (\alpha_1, \ldots, \alpha_n, \gamma)$, 使得线性方程组

$$(\boldsymbol{S}_1, \ldots, \boldsymbol{S}_n, \mathbf{1})\alpha^{\mathrm{T}} = (\boldsymbol{r}_1, \ldots, \boldsymbol{r}_m)(y_1, \ldots, y_m)^{\mathrm{T}}$$

成立. 通过求解 α, 可以得到策略 $\boldsymbol{P}$ 能够控制的线性关系的具体形式. 另一方面, 定理 6 也可以用来寻找能够施加某个具体的线性关系的策略. 为此, 需要从反方向处理上面的线性方程组, 即把向量 α 作为已知量, 而把 $\boldsymbol{r}_i$ 作为需要求解的未知量. 确定了 $\boldsymbol{r}_i$ 后, 根据约束向量的形式, 可以得到对应的零行列式策略. 上述结论的推导过程基于 Akin 证明定理 4 的思路进行. 按照 Press 和 Dyson 的思路, Cheng 等 [105,106] 则用半张量积的方法直接给出了能够实现具体的线性关系的策略形式 [107].

4.5 零行列式策略的演化特性

本小节介绍零行列式策略在种群中的演化特性. 通过分析零行列式策略的演化表现, 可以回答如下的问题:

(1) 零行列式策略这类收益控制策略是否可以在自然界广泛存在?

(2) 零行列式策略对种群中的个体行为有什么影响?

(3) 什么因素影响了零行列式策略的演化?

在大部分重复囚徒困境中, 一个理智的, 追求收益最大化的个体与采取敲诈策略的个体交互的最佳选择是无条件合作. 此时, 原始博弈等价于一个最后通牒博弈 (ultimatum game): 采取敲诈策略的个体为提议者 (proposer), 其对手为回应者 (responder), 对敲诈者的敲诈行为做出响应. 如果对手选择反抗敲诈策略 (即拒绝这个不公平的提议), 那么两个人都会获得很少的收益; 相反, 如果对手选择屈服于敲诈策略 (即接受提议), 那么两个人的收益都会增加, 但是提议者的收益更高.

在 Press 和 Dyson 的工作中, 他们把上述以收益最大化为目标而调整策略的个体称为 “进化” 个体 (evolutionary players). 当与固定的敲诈策略交互时, 这些个体可以对自己的策略进行连续细致的调整, 从而使得自己的收益沿着梯度不断上升. 不过, Press 和 Dyson 也意识到, 这种假设下的演化并不是真正的演化, 在他们的研究中没有出现对策略的自然选择 [30]. 随后, Sigmund 和 Nowak 进一步指出, 生物和文化进化不是个人层面的现象, 而是在种群中发生的; Press 和 Dyson 的 “进化” 个体所经历的也不是进化, 而是适应.

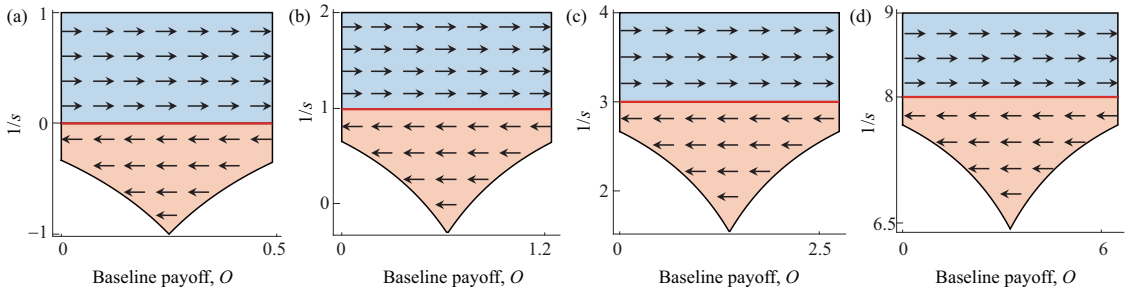

图 2 (网络版彩图) 零行列式策略的演化动力学

Figure 2 (Color online) Evolutionary dynamics of zero-determinant strategies. (a) $n = 2$; (b) $n = 3$; (c) $n = 5$; (d) $n = 10$

基于这样的考虑, 同时也为了探索在一对一交互中所向披靡的敲诈策略是否在群体的演化过程中也可以立于不败之地, 学者们首先研究了均衡器策略和敲诈策略的演化特性. 在无限种群中, Adami 等 [108] 证明了这两类策略都不是演化稳定策略, 因为赢留输变策略相互博弈的收益总是高于这两类策略与赢留输变策略博弈的收益. Noordman 等 [109] 进一步确认了这一结论. 与此同时, 在有限种群中, Hilbe 等 [34] 发现只有当种群规模非常小时, 敲诈策略才可以抵御其他策略的入侵并在种群中广泛存在; 当种群规模变大时, 敲诈策略无法在种群中稳定存在, 但是它对合作的出现可以起到催化剂的作用 —— 有了敲诈策略, 支持合作的赢留输变策略可以通过入侵敲诈策略而颠覆由无条件背叛策略所构成的种群.

"成也萧何, 败也萧何", 敲诈策略的贪婪属性虽然可以令它尽情地剥削对手, 在一对一交互中获得更高的收益, 却也造成了它在种群里的凋敝. 究其原因, 在于敲诈策略是一个对 "自己人" 不友好的策略. 一如前文所述, 当采取敲诈策略的个体身处同类之中时, 它们相互博弈的收益都将为基准收益, 即相互背叛的惩罚 P. 因此, 如果没有像绿胡子 (green beards) 这样的识别机制 [108] (长着绿胡子的个体会友好地对待其他有绿胡子的个体, 绿胡子是它们的一个特征标志), 除非它们进化得更加慷慨, 否则敲诈策略与同类个体的相互敲诈只会导致它们被自然选择淘汰. 而与敲诈策略相互映照的慷慨策略在与同类进行博弈时, 也将获得基准收益. 注意到, 这种情况下的基准收益为相互合作的奖励 R. 对于慷慨策略, Stewart 和 Plotkin 发现慷慨程度比较小 $(1 < \chi < (2N+1)/(N+1))$ 的慷慨策略是一类演化鲁棒策略 [90].

在最新的研究中, Chen 等 [35] 更加全面地分析了重复公共品博弈中零行列式策略的演化特性, 并明确了零行列式策略可以在种群中广泛存在的条件. 他们首先通过适应性动力学 (adaptive dynamics) 研究了全体零行列式策略的演化特性. 当所有个体都选择零行列式策略时, 种群要么只由采取敲诈策略的个体组成, 要么只由采取慷慨策略的个体组成. 其他类型的零行列式策略都不能在种群中稳定地存在. 至于种群最终会演化为哪种策略则取决于斜率 s 和参与博弈的人数 n. 当 $1/s < n-2$ 时, 种群由敲诈策略组成; 当 $1/s > n-2$ 时, 种群由慷慨策略组成 (图 2). 可以发现, 随着参与博弈的人数增加, 慷慨策略的吸引域 (图 2 中蓝色部分) 越来越小; 敲诈策略的吸引域 (图 2 中橙色部分) 越来越大. 这表明博弈人数的增加利于敲诈策略而不是慷慨策略的演化.

博弈人数的这种影响也体现在敲诈策略和慷慨策略分别与所有一步记忆策略的演化竞争中. Chen 等 [35] 发现, 当敲诈策略和慷慨策略分别与所有一步记忆策略竞争时, 存在一个临界值 N_1, 当种群规模 $N > N_1$ 时, 慷慨策略可以抵御其他策略的入侵; 也存在另一个临界值 N_2, 当 $N > N_2$ 时, 自然选择支持慷慨策略入侵其他策略. 在参与人数更多的博弈中, 慷慨策略抵御入侵和入侵其他策略的能力

更弱. 这表现在随着博弈人数的增多, 临界值 N_1 和 N_2 不断增加. 敲诈策略能够在种群中演化的条件与慷慨策略相反: 更小的种群和更多的参与者利于敲诈策略的演化. 这种现象出现的主要原因是在规模越大的种群中, 与同类策略博弈的收益的影响越大; 在参与者更多的博弈中, 慷慨策略会受到更多倾向背叛的策略的剥削, 而敲诈策略可以同时敲诈更多对手, 于是敲诈策略更容易在种群中占优.

此外, Chen 等 [35] 还研究了一个容易被忽略的因素 —— 敲诈系数 χ 对零行列式策略演化特性的影响. 对于慷慨策略来说, 敲诈系数 χ 也被称为慷慨系数, 来表示慷慨策略的慷慨程度; 慷慨系数越大表示慷慨程度越高. Chen 等发现慷慨系数越小的慷慨策略和敲诈系数越大的敲诈策略越容易在种群中占优. 这一结果也是符合直觉的.

4.6 不屈策略

参考 4.5 小节中对 "进化" 个体的理解, Press 和 Dyson 的工作以及后续的研究几乎无一例外地假设面对一个固定的敲诈策略, 理性的个体试图通过调整自己的策略以适应对方对收益的单边控制, 从而达到利益最大化的目的. 然而在现实中, 过分的剥削往往会遭遇阻力: 出于对公平的追求, 个体有时候宁愿牺牲自己的利益也要反抗剥削者的压迫. 以 Milinski 等 [40] 的人机交互实验为例, 当计算机采取敲诈策略时, 人类玩家的合作意愿并不强烈. 一个追求公平的个体可以通过拒绝无条件合作同时减少博弈双方的收益, 从而使敲诈策略无法达成其敲诈目的.

另一方面, 注意到在式 (16) 中, 零行列式策略的具体形式不仅依赖于基准收益 O, 敲诈系数 χ 和归一化因子 ϕ 这 3 个参数, 还受到收益矩阵的制约. 在囚徒困境的具体研究中, 大部分工作都沿用了 Axelrod 计算机锦标赛的收益矩阵, 即

$$\begin{array}{c} \\ C \\ D \end{array} \begin{array}{c} \begin{array}{cc} C & D \end{array} \\ \begin{pmatrix} 3 & 0 \\ 5 & 1 \end{pmatrix} \end{array}. \tag{25}$$

但是, 只要收益满足 $T > R > P > S$ 和 $T + S < 2R$ 的两人重复博弈都可以被视为重复囚徒困境, 所以即使对于同类博弈也会存在各种各样的收益矩阵. 一个最直接的例子就是比较囚徒困境中 $T + S$ (合作与背叛交替出现的收益和) 和 $2P$ (相互背叛的收益和) 的大小关系: $T + S > 2P$ 表示社会对抗程度低, 而 $T + S < 2P$ 表示社会对抗程度高 [110]. 因此, 收益的变化 (这种变化可能来自于博弈环境在演化过程中的不确定性 [111]) 如何影响零行列式策略对收益关系的控制能力, 特别是对对手的剥削程度, 成为了以往研究遗留的一个问题.

基于这两方面的考虑, Chen 和 Fu [112] 在最近的研究中进一步揭示了敲诈策略除了演化不稳定以外的其他弱点. 这些缺陷的发现甚至不需要将敲诈策略放置于种群环境中, 只需要讨论它们和对手策略的一对一交互.

Chen 和 Fu 首先细致地分析了零行列式策略在收益控制和敲诈对手方面的有效性, 以及该有效性如何受到参数 O, χ 和收益矩阵的影响. 在零行列式策略的 3 个参数中, 基准收益 O 确定了策略的慷慨程度: O 可以选择的范围介于相互背叛的收益 P 和相互合作的收益 R 之间 (包括两个端点), 策略的慷慨程度随着 O 的增大逐渐上升. 敲诈系数 χ 进一步影响了策略的剥削程度: χ 越大, 零行列式策略和对手策略经过基准收益调整后的收益比越大. 这两个参数和博弈的收益共同给出了归一化因子 ϕ 的范围. 值得注意的是, 参数 ϕ 的上界被 $T + S - 2P$ 的符号所决定. 这个过去被忽略的囚徒困境收益的关系, 却令人惊讶地成为了控制零行列式策略敲诈能力的一个重要条件.

和大部分关于收益控制的研究所使用的二元法 (即定性地判断个体是否获得比对手更高的收益: 要么是, 要么否) 不同的是, Chen 和 Fu 将零行列式策略敲诈对手的能力视为一个关于基准收益 O 的

连续的谱系, 并通过获得比对手更高收益的可能性 (概率) 来衡量这种敲诈能力. 对该数值进行定量的分析和比较, 就可以观察策略将如何依赖于上述控制参数; 更重要地, 可以明确策略的敲诈能力将如何受到 $T+S-2P$ 的符号的影响. 他们发现, 这一概率会随着 O 的增大而降低, 一旦 O 超过 P, 零行列式策略就有可能失去对其对手的支配地位. 比起基准收益, 博弈的收益关系对收益控制能力起着更显著的作用: 当 $T+S>2P$ 时, 零行列式策略获得更高收益的概率关于 O 的曲线是上凸的 (concave), 即使在 $O>P$ 的情况下, 该策略也很有可能保持支配地位; 而当 $T+S<2P$ 时, 曲线是下凸的 (convex), 在任何 $P<O\leqslant R$ 的情况下, 该策略都更有可能失去支配地位 (曲线的形状见文献 [112] 的图 1). 特别地, 对于敲诈策略 (即 $O=P$ 下的零行列式策略), 如果 $T+S>2P$, 敲诈者总可以成功地敲诈对手, 但是当 $T+S<2P$, 结果却是对半分的 (在敲诈能力的概率定义下, 敲诈成功的可能性只有二分之一).

另外, Chen 和 Fu 注意到, ϕ 作为一个隐藏的参数, 尽管没有出现在收益控制的线性关系式里, 却可以作为一只 "看不见的手" 影响博弈双方得到的期望收益. 根据前文, 零行列式策略的收益可以由两个矩阵的行列式之比计算 (式 (11)), 它其实是一个关于敲诈系数 χ 和归一化因子 ϕ 的有理函数. Chen 和 Fu 证明了该收益随着 ϕ 单调变化, 但是关于 χ 却可以有严格的非单调性, 具体地, 期望收益为 χ 的单峰函数. 这一结果进一步说明零行列式策略也可以单方面地调整自己的控制参数, 特别是以前被忽视的参数 ϕ (调整后的 ϕ 将是它可以接受的区间的边界值), 使得收益分配对自己更加有利.

在充分了解了零行列式策略的上述局限性之后, 一个自然的问题是: 面对一个试图通过施加不对等的线性关系以敲诈对手的策略, 个体除了屈服之外, 是否有其他的选择? 如果博弈双方都可以选择或保持策略不变或对策略进行调整, 一个更直接的问题是: 是否存在这样的策略, 在与其博弈时, 一个敲诈者不得不通过减少敲诈系数的方式减轻甚至取消压迫, 才能够实现收益最大化的目标? 注意到, 敲诈系数 χ 逐渐降低的过程也是收益分配不断向公平靠近的过程, 当 $\chi\to 1$ 时, 博弈双方的收益相等, 敲诈策略自己会变为公平策略. 在 Press 和 Dyson 的工作中, 他们假设的是零行列式策略固定而对手策略 "进化"; Chen 和 Fu 通过讨论与之对立场景, 即零行列式策略 "进化" 而对手策略固定, 发现了一类特殊的一步记忆策略. 在面对这类策略时, 任何敲诈策略都会受到收益降低的约束 — 敲诈程度越高, 期望收益越少. 因此不公平的收益分配关系会对敲诈者本身产生反作用 (backfire). 由于它们拒绝顺从敲诈策略的剥削, 这类策略被称为不屈策略 (unbending strategy).

具体来说, 不屈策略需要 (1) 将参数 ϕ 对收益的影响消除, 即 $\partial U/\partial\phi=0$; (2) 使敲诈策略的收益随着 χ 的增大而减少, 即 $\partial U/\partial\chi<0$. 根据这两点, Chen 和 Fu 发现了四类不屈策略 $\boldsymbol{q}=[q_{CC},q_{CD},q_{DC},q_{DD}]$ (对 $\boldsymbol{q}$ 的各个分量的部分约束关系见表 1). 这里, $h_D=[T-R-P+S-(T+S-2P)q_{CC}+(R-P)(q_{CD}+q_{DC})]/(2R-T-S)$, 这也是线性收益关系的等价条件, 因此 D 类不屈策略包含了所有满足 $P<O\leqslant R$ 的零行列式策略. 另外, B 类策略只存在于 $T+S<2P$ 的情况下, 博弈双方能够得到的最大收益为 R (在两者收益相同的情况下, 即 $\chi=1$); C 类策略的边界上包含 "愿意" 策略 (willing) $[1,1,1,0]$ [113], 对于无限接近它的策略 $[1-\delta,1-\delta,1-\delta,\varepsilon]$ ($\delta\to 0,\varepsilon\to 0$), 博弈双方能够得到的最大收益无限趋于 R (敲诈者不得不尽量保证公平, 即 $\chi\to 1$).

所有不屈策略形成的整个策略空间可以通过要求导数 $\partial U/\partial\chi<0$ 来描述 (图 3). 特别地, $T+S-2P$ 的符号决定了策略空间的几何形状. 对于 A 类策略, 当 $T+S>2P$ 时, 一个有趣的例子是 PSO 赌徒策略 (PSO gambler) $[1,0.522,0,0.121]$, 它是一个在重复囚徒困境中通过粒子群算法优化得到的策略; 当 $T+S<2P$ 时, 著名的赢留输变策略甚至是一个可以打败敲诈策略的不屈策略 (这部分收益高于对手收益的不屈策略见图 3(b) 的阴影区域). 而对于代表基准收益 O 严格大于惩罚 P 的不屈策略, 值得注意的是, 其策略空间的边界平面具有十分特殊的意义. 如图 3(c) 所示, 三角形 ADE 代表 $O=P$ 且 $\chi>1$ 的敲诈策略集, 四边形 $BCDE$ 代表均衡器策略集, D 类中的所有不屈策略都在这两

表 1 四类不屈策略

Table 1 Four classes of unbending strategies

Class	Expression	Class	Expression
A	$q_{CC}=1$ and $q_{DC}=0$	B	$q_{CD}=q_{DC}=0$
C	$q_{CC}=q_{CD}=q_{DC}$	D	$q_{DD}=h_D(q_{CC},q_{CD},q_{DC})$

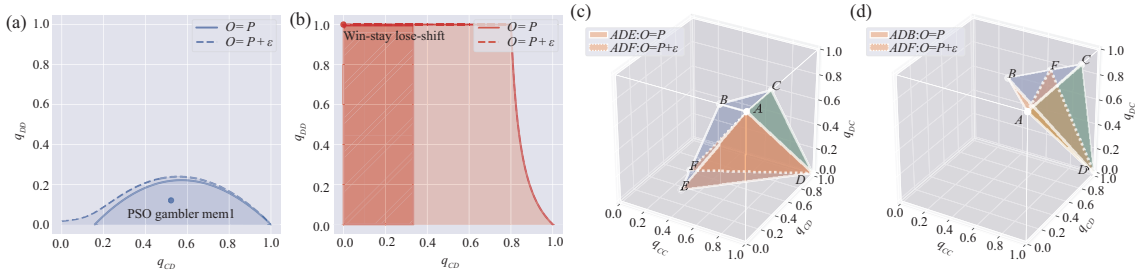

图 3 (网络版彩图) 不屈策略的几何空间

Figure 3 (Color online) Geometric space of unbending strategies. (a) Class A, $(R,S,T,P)=(3,0,5,1)$, PSO gambler mem1, $\boldsymbol{p}=(1,0.5217,0,0.1205)$; (b) Class A, $(R,S,T,P)=(1,-3,2,0)$; (c) Class D, $(R,S,T,P)=(3,0,5,1)$; (d) Class D, $(R,S,T,P)=(1,-3,2,0)$

个平面之间, 同时以单位立方体为界. 此外, 三角形 ACD 代表 $O=R$ 的慷慨策略集, 三角形 ABD 代表 $O=(T+S)/2$ 的零行列式策略集. 当 $T+S<2P$ (图 3(d)) 时, 策略空间退化为三角形 ABD (敲诈策略) 和三角形 BCD (均衡器策略) 之间的区域.

为了进一步理解固定的不屈策略在推动公平和合作方面前所未有的引导作用, Chen 和 Fu 还从两个角度分析了不屈策略在面对更一般的零行列式策略时的表现. 对于基准收益 $O=P+\varepsilon$ 的零行列式策略, 如图 3 所示 (虚线部分), 它们对不屈策略 (满足 $\partial U/\partial\phi=0$ 和 $\partial U/\partial\chi<0$) 的分类依旧是非常稳健的. 而对于捐赠博弈 (donation game) 这样一个特殊的囚徒困境博弈和该博弈中零行列式策略的一个特殊的子集 —— 反应策略 (reactive strategy, 满足 $p_{CC}=p_{DC},p_{CD}=p_{DD}$), 其受不屈策略驱动的适应性学习 (adaptive learning) 结果也是趋于合作的.

5 随机博弈中的收益控制和福利时间策略

上述博弈的收益函数在博弈过程中是恒定的. 近几年, 环境和个体行为的相互影响引起了广泛的关注 [72,111,114~118]. 个体的收益函数 $u_1,\ldots,u_n$ 与环境相关联, 环境的变化导致个体收益函数也随之改变. 那么在个体收益函数不断变化的博弈中存在收益控制策略吗? 最近, Liu 和 Wu [119] 证明在这样的博弈中也存在收益控制策略.

通常用随机博弈 (stochastic game) 来刻画收益随时间变化的重复博弈 [120]. 考虑两个个体 X 和 Y 参与的重复囚徒困境博弈. 假设个体可能处于两个收益不同的囚徒困境博弈中:

$$\text{囚徒困境 1}\quad \begin{array}{c} \\ C \\ D \end{array}\!\!\begin{array}{c} \begin{array}{cc} C & D \end{array} \\ \begin{pmatrix} a & -c_1 \\ a+c_1 & 0 \end{pmatrix} \end{array}, \qquad \text{囚徒困境 2}\quad \begin{array}{c} \\ C \\ D \end{array}\!\!\begin{array}{c} \begin{array}{cc} C & D \end{array} \\ \begin{pmatrix} a & -c_2 \\ a+c_2 & 0 \end{pmatrix} \end{array}, \tag{26}$$

其中, $a,c_1,c_2\geqslant 0$; c_1 和 c_2 表示囚徒困境 1 和囚徒困境 2 中个体采取背叛行为的诱惑. 不失一般性, 假设 $c_1<c_2$, 即囚徒困境 1 的困境弱于囚徒困境 2. 个体处于囚徒困境 1 还是囚徒困境 2 由当前所处

的囚徒困境和个体所采取的行动共同决定. 设在囚徒困境 i ($i \in \{1,2\}$) 中有 j ($j \in \{0,1,2\}$) 个个体选择合作, 那么下一轮玩囚徒困境 1 的概率为 f_{ij}. 在随机博弈中, 个体的策略不仅与每个个体上一轮采取的行动相关, 还与当前所处的博弈状态相关. 记 $\boldsymbol{p} = (p_{a_1a_2i})$ ($a_1, a_2 \in \{C, D\}, i \in \{1,2\}$) 为处于囚徒困境 i 且个体 X 选择行动 a_1 个体 Y 选择行动 a_2 时个体 X 下一轮合作的概率. Liu 和 Wu 发现策略

$$\begin{aligned}\boldsymbol{p} &= (p_{CC1}, p_{CD1}, p_{DC1}, p_{DD1}, p_{CC2}, p_{CD2}, p_{DC2}, p_{DD2}) \\ &= (1 + f_{12} - f_{11} + ha, 1, 0, f_{10} - f_{11} - ha, 1 + f_{22} - f_{21} + ha, 1, 0, f_{20} - f_{21} - ha)\end{aligned} \tag{27}$$

可以将两个体的收益之和 W (也称社会福利 (welfare)) 和处于囚徒困境 1 的时间 Q 控制为一个线性关系:

$$Q + hW = ah + f_{11}. \tag{28}$$

Liu 和 Wu 称这类策略为福利时间策略 (welfare-time strategy). 当个体 X 采取福利时间策略时, 参与博弈的两个体处于囚徒困境 1 的时间是其收益总和的单调函数, 具体单调递增还是单调递减由 h 的符号决定. Liu 和 Wu 证明 h 的取值既可以为正, 也可以为负. 因此, 收益和 (W) 既可以与处于囚徒困境 1 的时间 (Q) 正相关, 也可以与其负相关. 这表明即使两个体能更长时间处于困境较小的囚徒困境 1, 也无法保证它们的收益和变大. 于是, 当处于囚徒困境 1 的时间比例变大, 单个个体就可以决定收益和随之增加还是减少. 特别地, 若参与博弈的两个体都采取福利时间策略且选择的控制系数 h 不相同, 那么两个体的收益和 W 将被固定为 a, 处于囚徒困境 1 的时间比例将被固定为 f_{11}. 另一方面, Liu 和 Wu 的发现也表明当存在博弈切换时, 个体不仅可以单方面地控制收益, 还可以控制处于某一博弈的时间.

6 总结与展望

本文从 4 个方面介绍了当前博弈收益控制的研究进展: 收益控制策略的形式、能控制的收益关系、收益控制策略的判别和收益控制策略的演化特性. 首先介绍了单次博弈中能控制收益的策略 —— 约束策略 —— 的定义、性质和充要条件. 然后介绍了重复博弈中的收益控制策略 —— 零行列式策略, 包括零行列式策略的发现过程、存在性、判别和演化特性. 最后介绍了随机博弈中的收益控制策略 —— 福利时间策略. 在博弈中能够单方面地控制收益是一个颠覆认知的重大发现. 从发现收益控制策略的过程来看, Press 和 Dyson [30] 以及 Akin [91] 从不变分布与其他向量点乘的结果出发来寻找收益控制策略; 而 Tan 等从代数和几何的角度分析了收益控制策略组成的空间, 从而得到更一般的博弈中的收益控制策略 [36], 这也为单次博弈中的收益控制策略 —— 约束策略的发现奠定了基础 [37]. 自 Press 和 Dyson 提出零行列式策略以来, 博弈中的收益控制已经得到了很大的发展. 但是仍然有许多方面亟待研究.

在控制的收益形式方面, 目前大多数研究关注能将收益控制为线性形式的策略, 将收益形式控制为非线性形式的策略的研究还很匮乏. Chen 和 Fu [112] 的发现表明如果仅将收益控制为线性形式, 当 $2P > T + S$ 时不屈策略可以让敲诈策略无法达成敲诈的目的. 如果能将收益形式控制为非线性关系, 有可能避免这种情况的出现. 最近在单次博弈中已经发现能够将所有个体的收益控制为非线性形式的策略 [37]. 把相应的结果推广到重复博弈中是非常有意义的. 另一方面, 虽然目前已经证明了将收益控制为非线性形式的策略的存在性, 并明确了这类策略的形式, 但是还未提出能实现某一具体功能的例子, 相关的研究还需要进行.

在记忆长度方面, 目前针对重复博弈中收益控制的研究大部分关注以零行列式策略为代表的一步记忆策略. 特别地, Press 和 Dyson 已经证明了, 在无穷轮重复囚徒困境中, 采取多步记忆策略的个体在面对一步记忆策略的个体时, 相当于采取了一个一步记忆策略. 事实上, 只要博弈个体没有物理或信息上的交互, 博弈记忆的有效长度取决于该博弈中记忆长度最短的个体. 在这一基础上, 近来也有一些学者开始探索拥有更长记忆的控制策略, 比如两步记忆的零行列式策略, 并得到了一些有趣的结论 [39]. 不过, 记忆长度的这种 "木桶效应" 仅仅在纯粹的重复博弈 ("pure" repeated games, 即博弈个体唯一的接触是博弈交互, 唯一的信息是博弈结果) 中成立, 对于更一般的情况, 记忆长度对收益控制的影响还有待进一步讨论.

另外, 随着博弈人数增多, 单个策略的控制能力不断下降. Hilbe 等 [27] 提出通过联盟的方式可以提高策略收益控制的能力. 其他能够提高单个策略的控制能力的机制和可以同时控制多人博弈中多个个体的收益的策略也待研究. 一个个体可以控制其他个体的收益, 也可能受到其他个体施加的收益控制. 个体如何避免、判断和摆脱收益控制都是非常值得研究的问题. 此外, 人在决策时存在各种不确定性, 无论是 "颤抖的双手" (trembling hands) 还是 "模糊的大脑" (fuzzy minds [121]) 都会增加博弈的噪声. 研究个体决策的不确定性对收益控制的影响是十分有意义的课题. 个体决策的不确定性也体现在随着博弈的进行, 它们的收益控制目标可能发生变化. 目前的研究大多讨论对期望收益 (即长期收益) 的控制, 如何控制短期收益甚至在博弈过程中调整控制策略也非常值得探究. 人工智能在学习和决策中具有良好的 "灵活性", 是解决这个问题的一个可能的切入点.

博弈收益控制策略的发现为解决博弈基本问题 —— 谁能在博弈中获胜 —— 提供了新的思路与方法. 以零行列式策略为代表的收益控制策略可以在二人和多人博弈中实现对集体收益的不平等的分割. 理解收益控制问题不仅能够推动博弈论的研究发展, 也对抵制剥削、促进公平以及维护社会的和谐具有重要的参考意义.

参考文献

1 von Neumann J, Morgenstern O. Theory of Games and Economic Behavior. Princeton: Princeton University Press, 1944

2 Binmore K. Game Theory. Oxford: Oxford University Press, 2007

3 Nowak M A. Five rules for the evolution of cooperation. Science, 2006, 314: 1560–1563

4 Zhang J F. Preface to special topic on games in control systems. Natl Sci Rev, 2020, 7: 1115

5 Wang L, Fu F, Chen X J, et al. Collective decision-making over complex networks. CAAI Trans Intell Syst, 2008, 3: 14 [王龙, 伏锋, 陈小杰, 等. 复杂网络上的群体决策. 智能系统学报, 2008, 3: 14]

6 Shamma J S. Game theory, learning, and control systems. Natl Sci Rev, 2020, 7: 1118–1119

7 Hilbe C, Martinez-Vaquero L A, Chatterjee K, et al. Memory-n strategies of direct reciprocity. Proc Natl Acad Sci USA, 2017, 114: 4715–4720

8 Li A M, Zhou L, Su Q, et al. Evolution of cooperation on temporal networks. Nat Commun, 2020, 11: 2259

9 Santos F C, Santos M D, Pacheco J M. Social diversity promotes the emergence of cooperation in public goods games. Nature, 2008, 454: 213–216

10 Santos F P, Santos F C, Pacheco J M. Social norm complexity and past reputations in the evolution of cooperation. Nature, 2018, 555: 242–245

11 Nowak M A, Sigmund K. Evolution of indirect reciprocity. Nature, 2005, 437: 1291–1298

12 Zhou L, Wu B, Du J M, et al. Aspiration dynamics generate robust predictions in heterogeneous populations. Nat Commun, 2021, 12: 3250

13 Cao M. Merging game theory and control theory in the era of AI and autonomy. Natl Sci Rev, 2020, 7: 1122–1124

14 Wang L, Wu B, Du J M, et al. Spreading dynamics on complex dynamical networks. Sci Sin Inform, 2020, 50:

1714–1731 [王龙, 武斌, 杜金铭, 等. 复杂动态网络上的传播行为分析. 中国科学: 信息科学, 2020, 50: 1714–1731]
15 Hilbe C, Šimsa Š, Chatterjee K, et al. Evolution of cooperation in stochastic games. Nature, 2018, 559: 246–249
16 Wu T, Fu F, Wang L. Phenotype affinity mediated interactions can facilitate the evolution of cooperation. J Theor Biol, 2019, 462: 361–369
17 Chakra M A, Traulsen A. Evolutionary dynamics of strategic behavior in a collective-risk dilemma. Plos Comput Biol, 2012, 8: e1002652
18 Helbing D, Szolnoki A, Perc M, et al. Evolutionary establishment of moral and double moral standards through spatial interactions. Plos Comput Biol, 2010, 6: e1000758
19 Su Q, Li A M, Wang L, et al. Spatial reciprocity in the evolution of cooperation. Proc R Soc B, 2019, 286: 20190041
20 Barfuss W, Donges J F, Vasconcelos V V, et al. Caring for the future can turn tragedy into comedy for long-term collective action under risk of collapse. Proc Natl Acad Sci USA, 2020, 117: 12915–12922
21 Nash J F. Equilibrium points in n-person games. Proc Natl Acad Sci USA, 1950, 36: 48–49
22 Nash J F. Non-cooperative games. Ann Math, 1951, 54: 286
23 Golman R, Page S E. General Blotto: games of allocative strategic mismatch. Public Choice, 2009, 138: 279–299
24 Huang F, Cao M, Wang L. Learning enables adaptation in cooperation for multi-player stochastic games. J R Soc Interface, 2020, 17: 20200639
25 Harper M, Knight V, Jones M, et al. Reinforcement learning produces dominant strategies for the Iterated Prisoner's Dilemma. Plos One, 2017, 12: e0188046
26 Behnezhad S, Dehghani S, Derakhshan M, et al. Faster and simpler algorithm for optimal strategies of blotto game. In: Proceedings of the 31st AAAI Conference on Artificial Intelligence, 2017. 1921: 369–375
27 Hilbe C, Wu B, Traulsen A, et al. Cooperation and control in multiplayer social dilemmas. Proc Natl Acad Sci USA, 2014, 111: 16425–16430
28 Govaert A, Cao M. Zero-determinant strategies in repeated multiplayer social dilemmas with discounted payoffs. IEEE Trans Automat Contr, 2021, 66: 4575–4588
29 Boerlijst M C, Nowak M A, Sigmund K. Equal pay for all prisoners. Am Math Mon, 1997, 104: 303–305
30 Press W H, Dyson F J. Iterated Prisoner's Dilemma contains strategies that dominate any evolutionary opponent. Proc Natl Acad Sci USA, 2012, 109: 10409–10413
31 Hilbe C, Traulsen A, Sigmund K. Partners or rivals? Strategies for the iterated prisoner's dilemma. Games Economic Behav, 2015, 92: 41–52
32 Hilbe C, Chatterjee K, Nowak M A. Partners and rivals in direct reciprocity. Nat Hum Behav, 2018, 2: 469–477
33 Hilbe C, Nowak M A, Traulsen A. Adaptive dynamics of extortion and compliance. Plos One, 2013, 8: e77886
34 Hilbe C, Nowak M A, Sigmund K. Evolution of extortion in Iterated Prisoner's Dilemma games. Proc Natl Acad Sci USA, 2013, 110: 6913–6918
35 Chen F, Wu T, Wang L. Evolutionary dynamics of zero-determinant strategies in repeated multiplayer games. J Theor Biol, 2022, 549: 111209
36 Tan R F, Su Q, Wu B, et al. Payoff control in repeated games. In: Proceedings of the 33rd IEEE Chinese Control and Decision Conference (CCDC), 2021. 997–1005
37 Tan R F. Payoff Control in Non-cooperative Games. Technical Report. Beijing: Peking University, 2022
38 Ueda M. Tit-for-tat strategy as a deformed zero-determinant strategy in repeated games. J Phys Soc Jpn, 2021, 90: 025002
39 Ueda M. Memory-two zero-determinant strategies in repeated games. Royal Soc Open Sci, 2021, 8: 202186
40 Hilbe C, Röhl T, Milinski M. Extortion subdues human players but is finally punished in the prisoner's dilemma. Nat Commun, 2014, 5: 3976
41 Boyd R, Richerson P J. The evolution of reciprocity in sizable groups. J Theor Biol, 1988, 132: 337–356
42 Macy M W, Flache A. Learning dynamics in social dilemmas. Proc Natl Acad Sci USA, 2002, 99: 7229–7236
43 Poundstone W. Prisoner's Dilemma. New York: Doubleday, 1993
44 Nowak M A. Evolutionary Dynamics: Exploring the Equations of Life. Cambridge: Harvard University Press, 2006

45 郝继仁. “囚徒困境” 的前世今生. 系统与控制纵横, 2019, 2: 45–52
46 Fudenberg D, Tirole J. Game Theory. Cambridge: MIT Press, 1991
47 Wedekind C, Milinski M. Human cooperation in the simultaneous and the alternating Prisoner's Dilemma: Pavlov versus Generous Tit-for-Tat. Proc Natl Acad Sci USA, 1996, 93: 2686–2689
48 Nowak M A, Sigmund K. The alternating Prisoner's Dilemma. J Theor Biol, 1994, 168: 219–226
49 McAvoy A, Hauert C. Autocratic strategies for alternating games. Theor Population Biol, 2017, 113: 13–22
50 Mailath G J, Olszewski W. Folk theorems with bounded recall under (almost) perfect monitoring. Games Economic Behav, 2011, 71: 174–192
51 Mailath G J, Morris S. Repeated games with almost-public monitoring. J Economic Theor, 2002, 102: 189–228
52 Barlo M, Carmona G, Sabourian H. Repeated games with one-memory. J Economic Theor, 2009, 144: 312–336
53 Nowak M A, Sigmund K, El-Sedy E. Automata, repeated games and noise. J Math Biol, 1995, 33: 703–722
54 Kraines D, Kraines V. Learning to cooperate with Pavlov an adaptive strategy for the iterated Prisoner's Dilemma with noise. Theor Decis, 1993, 35: 107–150
55 Rand D G, Fudenberg D, Dreber A. It's the thought that counts: the role of intentions in noisy repeated games. J Economic Behav Organization, 2015, 116: 481–499
56 Taha M A, Ghoneim A. Zero-determinant strategies in repeated asymmetric games. Appl Math Computation, 2020, 369: 124862
57 Trivers R L. The evolution of reciprocal altruism. Quart Rev Biol, 1971, 46: 35–57
58 Smith J M, Price G R. The logic of animal conflict. Nature, 1973, 246: 15–18
59 Boyd R, Lorberbaum J P. No pure strategy is evolutionarily stable in the repeated Prisoner's Dilemma game. Nature, 1987, 327: 58–59
60 Akçay E. Collapse and rescue of cooperation in evolving dynamic networks. Nat Commun, 2018, 9: 2692
61 Feng X, Zhang Y L, Wang L. Evolution of stinginess and generosity in finite populations. J Theor Biol, 2017, 421: 71–80
62 Su Q, Li A M, Wang L. Evolution of cooperation with interactive identity and diversity. J Theor Biol, 2018, 442: 149–157
63 Traulsen A, Nowak M A. Evolution of cooperation by multilevel selection. Proc Natl Acad Sci USA, 2006, 103: 10952–10955
64 Traulsen A, Nowak M A, Pacheco J M. Stochastic payoff evaluation increases the temperature of selection. J Theor Biol, 2007, 244: 349–356
65 Antal T, Ohtsuki H, Wakeley J, et al. Evolution of cooperation by phenotypic similarity. Proc Natl Acad Sci USA, 2009, 106: 8597–8600
66 Ohtsuki H, Hauert C, Lieberman E, et al. A simple rule for the evolution of cooperation on graphs and social networks. Nature, 2006, 441: 502–505
67 Huang F, Chen X J, Wang L. Evolution of cooperation in a hierarchical society with corruption control. J Theor Biol, 2018, 449: 60–72
68 Zhou L, Wu B, Vasconcelos V V, et al. Simple property of heterogeneous aspiration dynamics: beyond weak selection. Phys Rev E, 2018, 98: 062124
69 Ohtsuki H, Nowak M A. The replicator equation on graphs. J Theor Biol, 2006, 243: 86–97
70 Fu F, Tarnita C E, Christakis N A, et al. Evolution of in-group favoritism. Sci Rep, 2012, 2: 460
71 Wu T, Fu F, Wang L. Coevolutionary dynamics of aspiration and strategy in spatial repeated public goods games. New J Phys, 2018, 20: 063007
72 Hilbe C, Schmid L, Tkadlec J, et al. Indirect reciprocity with private, noisy, and incomplete information. Proc Natl Acad Sci USA, 2018, 115: 12241–12246
73 Zhou L, Li A M, Wang L. Evolution of cooperation on complex networks with synergistic and discounted group interactions. Europhys Lett, 2015, 110: 60006
74 Boyd R, Gintis H, Bowles S, et al. The evolution of altruistic punishment. Proc Natl Acad Sci USA, 2003, 100:

3531–3535
75 Wang L, Cong R, Li K. Feedback mechanism in cooperation evolving. Sci Sin Inform, 2014, 44: 1495–1514 [王龙, 丛睿, 李昆. 合作演化中的反馈机制. 中国科学: 信息科学, 2014, 44: 1495–1514]
76 Su Q, Li A M, Zhou L, et al. Interactive diversity promotes the evolution of cooperation in structured populations. New J Phys, 2016, 18: 103007
77 Donahue K, Hauser O P, Nowak M A, et al. Evolving cooperation in multichannel games. Nat Commun, 2020, 11: 3885
78 Imhof L A, Nowak M A. Stochastic evolutionary dynamics of direct reciprocity. Proc R Soc B, 2010, 277: 463–468
79 Gross J, de Dreu C K W. The rise and fall of cooperation through reputation and group polarization. Nat Commun, 2019, 10: 776
80 Huang F, Chen X J, Wang L. Conditional punishment is a double-edged sword in promoting cooperation. Sci Rep, 2018, 8: 528
81 Henrich J, McElreath R, Barr A, et al. Costly punishment across human societies. Science, 2006, 312: 1767–1770
82 Wang L, Du J M. Evolutionary game theoretic approach to coordinated control of multi-agent systems. J Syst Sci Math Sci, 2016, 36: 302–318 [王龙, 杜金铭. 多智能体协调控制的演化博弈方法. 系统科学与数学, 2016, 36: 302–318]
83 Hauert C, de Monte S, Hofbauer J, et al. Volunteering as red queen mechanism for cooperation in public goods games. Science, 2002, 296: 1129–1132
84 Rand D G, Nowak M A. The evolution of antisocial punishment in optional public goods games. Nat Commun, 2011, 2: 434
85 Huang F, Chen X J, Wang L. Evolutionary dynamics of networked multi-person games: mixing opponent-aware and opponent-independent strategy decisions. New J Phys, 2019, 21: 063013
86 Nowak M A, Sasaki A, Taylor C, et al. Emergence of cooperation and evolutionary stability in finite populations. Nature, 2004, 428: 646–650
87 Wu B, Altrock P M, Wang L, et al. Universality of weak selection. Phys Rev E, 2010, 82: 046106
88 Wu B, García J, Hauert C, et al. Extrapolating weak selection in evolutionary games. Plos Comput Biol, 2013, 9: e1003381
89 Wu B, Bauer B, Galla T, et al. Fitness-based models and pairwise comparison models of evolutionary games are typically different-even in unstructured populations. New J Phys, 2015, 17: 023043
90 Stewart A J, Plotkin J B. From extortion to generosity, evolution in the Iterated Prisoner's Dilemma. Proc Natl Acad Sci USA, 2013, 110: 15348–15353
91 Akin E. The iterated Prisoner's Dilemma: good strategies and their dynamics. In: Proceedings of Ergodic Theory, 2016. 77–107
92 Chen X R, Wang L, Fu F. The intricate geometry of zero-determinant strategies underlying evolutionary adaptation from extortion to generosity. New J Phys, 2022, 24: 103001
93 Nowak M, Sigmund K. A strategy of win-stay, lose-shift that outperforms tit-for-tat in the Prisoner's Dilemma game. Nature, 1993, 364: 56–58
94 Axelrod R. Launching "the evolution of cooperation". J Theor Biol, 2012, 299: 21–24
95 Axelrod R, Hamilton W D. The evolution of cooperation. Science, 1981, 211: 1390–1396
96 Axelrod R. The emergence of cooperation among egoists. Am Polit Sci Rev, 1981, 75: 306–318
97 Nowak M A, Sigmund K. Tit for tat in heterogeneous populations. Nature, 1992, 355: 250–253
98 Hilbe C, Wu B, Traulsen A, et al. Evolutionary performance of zero-determinant strategies in multiplayer games. J Theor Biol, 2015, 374: 115–124
99 Pan L M, Hao D, Rong Z H, et al. Zero-determinant strategies in iterated public goods game. Sci Rep, 2015, 5: 13096
100 McAvoy A, Hauert C. Autocratic strategies for iterated games with arbitrary action spaces. Proc Natl Acad Sci USA, 2016, 113: 3573–3578

101 Stewart A J, Parsons T L, Plotkin J B. Evolutionary consequences of behavioral diversity. Proc Natl Acad Sci USA, 2016, 113: E7003–E7009

102 Cheng Z Y, Chen G P, Hong Y G. Misperception influence on zero-determinant strategies in iterated Prisoner's Dilemma. Sci Rep, 2022, 12: 5174

103 Mamiya A, Ichinose G. Strategies that enforce linear payoff relationships under observation errors in Repeated Prisoner's Dilemma game. J Theor Biol, 2019, 477: 63–76

104 Mamiya A, Ichinose G. Zero-determinant strategies under observation errors in repeated games. Phys Rev E, 2020, 102: 032115

105 Cheng D Z, Qi H S, Li Z Q. Analysis and control of Boolean networks: a semi-tensor product approach. Acta Autom Sin, 2011, 37: 529–540

106 Cheng D Z, Qi H S, Zhao Y. An Introduction to Semi-Tensor Product of Matrices and Its Applications. Singapore: World Scientific, 2012

107 Cheng D Z. A formula for designing zero-determinant strategies. 2021. ArXiv:2107.03255

108 Adami C, Hintze A. Evolutionary instability of zero-determinant strategies demonstrates that winning is not everything. Nat Commun, 2013, 4: 2193

109 Noordman C R, Vreeswijk G A W. Evolving novelty strategies for the Iterated Prisoner's Dilemma in deceptive tournaments. Theor Comput Sci, 2019, 785: 1–16

110 D'Orsogna M R, Perc M. Statistical physics of crime: a review. Phys Life Rev, 2015, 12: 1–21

111 Weitz J S, Eksin C, Paarporn K, et al. An oscillating tragedy of the commons in replicator dynamics with game-environment feedback. Proc Natl Acad Sci USA, 2016, 113: E7518–E7525

112 Chen X R, Fu F. Outlearning extortioners by fair-minded unbending strategies. 2022. ArXiv:2201.04198

113 van den Berg P, Weissing F J. The importance of mechanisms for the evolution of cooperation. Proc R Soc B, 2015, 282: 20151382

114 Su Q, McAvoy A, Wang L, et al. Evolutionary dynamics with game transitions. Proc Natl Acad Sci USA, 2019, 116: 25398–25404

115 Shao Y X, Wang X, Fu F. Evolutionary dynamics of group cooperation with asymmetrical environmental feedback. Europhys Lett, 2019, 126: 40005

116 Wang X, Fu F. Eco-evolutionary dynamics with environmental feedback: cooperation in a changing world. Europhys Lett, 2020, 132: 10001

117 Tilman A R, Plotkin J B, Akçay E. Evolutionary games with environmental feedbacks. Nat Commun, 2020, 11: 915

118 Attia L, Oliu-Barton M. A formula for the value of a stochastic game. Proc Natl Acad Sci USA, 2019, 116: 26435–26443

119 Liu F L, Wu B. Environmental quality and population welfare in Markovian eco-evolutionary dynamics. Appl Math Computation, 2022, 431: 127309

120 Shapley L S. Stochastic games. Proc Natl Acad Sci USA, 1953, 39: 1095–1100

121 Ohtsuki H, Iwasa Y, Nowak M A. Indirect reciprocity provides only a narrow margin of efficiency for costly punishment. Nature, 2009, 457: 79–82

Payoff control in game theory

Long WANG[1*], Fang CHEN[1] & Xingru CHEN[2]

1. *Center for Systems and Control, Peking University, Beijing* 100871*, China*;
2. *School of Sciences, Beijing University of Posts and Telecommunications, Beijing* 100876*, China*
* Corresponding author. E-mail: longwang@pku.edu.cn

Abstract In game theory, a single player usually cannot control the payoffs of all players in a game. An exception is the equalizer strategy proposed at the end of the last century for prisoner's dilemma, with which a player can set their opponent's payoff to be any designated value in a certain interval, regardless of which strategy the opponent uses. This result was further generalized with the discovery of the zero-determinant (ZD) strategies, which allow a player to unilaterally enforce a linear relationship between his own payoff and that of the opponent. The question of how payoff control can be established has attracted significant attention from computer scientists, control theorists, and evolutionary biologists, and many new results have been subsequently derived. This paper discusses the latest advances in payoff control, enforcing either a linear or nonlinear relation in one-shot or repeated games. In particular, we highlight the above question from four aspects: the concept of payoff control, forms of payoff relation that can be established, strategies that can control payoffs, and the evolutionary behavior of these strategies. We also provide an outlook on the directions for future research.

Keywords game theory, payoff control, zero-determinant strategy, evolutionary game theory, strategy design