“囚徒困境”的前世今生 (上)

郝继仁　北京大学

一、引言

大到国家之间的力量制衡，小到蚁群内部的协作分工，博弈思想在现实世界中无处不在，它与个体的决策行为息息相关。追溯到中国古代，《孙子兵法》中就蕴含了博弈思想的精髓。例如，“知己知彼，百战不殆”就充分说明了制定一个好策略离不开对于双方信息的掌握。但是直到 20 世纪，学者们才开始系统地研究博弈。

Zermelo 和 Borel 分别于 1913 年和 1921 年最先对棋类游戏进行探索。1928 年 von Neumann 发表了一篇名为 Zur Theorie der Gesellschaftsspiele（即“社交游戏理论”）的论文。在这篇论文中他证明了两人零和博弈（一类“我赢即你输”的博弈，没有双赢的可能，棋类游戏就是其中一种）的极小极大定理，该定理成为博弈论中的一个基本定理。

在此基础上，von Neumann 和 Morgenstern 于 1944 年合著了 Theory of Games and Economic Behavior，详尽讨论了两人零和博弈，并首次对合作博弈进行了系统化和形式化的研究，由此奠定了博弈论的基础和理论体系。该书的出版也标志着博弈论的正式诞生。此后，合作博弈论在 20 世纪 50 年代得到了空前的发展。

与此同时，非合作博弈论也开始得到关注。Nash 在 1950 年和 1951 年分别发表了两篇论文 Equilibrium Points in N-person Games 和 Non-cooperative Games，创新性地提出了策略均衡（被后人称为纳什均衡）的概念，并利用不动点定理证明了该均衡点的存在性，奠定了非合作博弈论的基石。

不仅如此，Selten、Harsanyi 和其他学者后续的相关研究也极大地推动了博弈论的发展。如今，博弈论已经广为人知，成为经济学及其它社会科学的标准分析工具之一。

博弈论中有不少经典的博弈模型，例如囚徒困境博弈、雪堆博弈、猎鹿博弈和智猪博弈。在众多的模型中，最为经典的莫过于 1950 年被提出的囚徒困境模型。囚徒困境模型是两人非零和博弈中最具代表性的例子：在该博弈中，个体单方面最优的决策是不合作；但相互合作却能给双方都带来更高的收益。

这充分反映了个体利益与集体利益之间的冲突。有鉴于此，囚徒困境模型是研究合作进化的典型博弈范式。

二、囚徒困境模型及其起源

囚徒困境模型起源于 1950 年。当时，Flood 和 Dresher 在兰德公司任职时提出了初步的理论模型。之后，普林斯顿大学的数学家 Tucker 在此基础上引入了收益的概念，并用囚徒的故事来解释该模型，于是就成为了广为人知的囚徒困境博弈模型。

想象这样一个场景：

两个人 A 和 B 计划去抢劫银行，带着作案工具在银行门前徘徊准备动手时被警察抓住。警察怀疑他们意图抢劫，但由于证据不足只能暂时将他们拘留。为了防止嫌疑犯之间沟通和串供，警察将他们分别关押在两个房间并进行单独审讯（见图 2.1）。在审讯中，警察分别给 A 和 B 提供了两个选择：坦白二人抢银行的计划，或者保持沉默。如果一人坦白而另一人保持沉默，那么坦白的一方有戴罪立功表现被立即释放，而保持沉默的一方将被监禁 10 个月；如果双方都坦白，则二人合谋抢银行的罪名成立，A 和 B 都将被监禁 8 个月；如果他们都保持沉默，则因仍有犯罪嫌疑，则各自被拘留 1 个月（见表 2.1）。

图 2.1 囚徒困境场景图

如果你是 A 或 B，你会作何选择呢？

		囚徒 B	
		保持沉默	坦白罪行
囚徒 A	保持沉默	A 被拘留 1 个月， B 被拘留 1 个月	A 被监禁 10 个月， B 获释
	坦白罪行	A 获释， B 被监禁 10 个月	A 被监禁 8 个月， B 被监禁 8 个月

表 2.1 囚徒困境的表格描述

这个场景其实是一个两人博弈，具体来讲，这个博弈是在两位参与者之间展开的。每位参与者都有两个策略：一，选择保持沉默，即包庇同伙，称之为合作策略（“Cooperate”，以下简称 C 策略）；二，选择坦白二人合谋抢银行的计划，即背叛同伙，称之为背叛策略（“Defect”，以下简称 D 策略）。不同的策略组合对应不同的收益：在对方采取 C 策略的情况下，个体选择 C 策略的收益为 R，选择 D 策略的收益为 T；在对方采取 D 策略的情况下，个体选择 C 策略的收益为 S，选择 D 策略的收益为 P。当这四种收益满足 T > R > P > S 时，我们称这种两人博弈模型为囚徒困境博弈。如果用简单的数学语言来描述表 2.1 中的模型，我们便有了如下的收益矩阵：

$$\begin{matrix} & \begin{matrix} C & D \end{matrix} \\ \begin{matrix} C \\ D \end{matrix} & \begin{pmatrix} R & S \\ T & P \end{pmatrix} \end{matrix} = \begin{matrix} & \begin{matrix} C & D \end{matrix} \\ \begin{matrix} C \\ D \end{matrix} & \begin{pmatrix} -1 & -10 \\ 0 & -8 \end{pmatrix} \end{matrix}$$

由上面的收益矩阵我们可以看出：当对方选择 C 策略时，个体选择 D 策略的收益（获释）要大于选择 C 策略的收益（被拘留 1 个月）；当对方选择 D 策略时，个体选择 D 策略的收益（被监禁 8 个月）要大于选择 C 策略的收益（被监禁 10 个月）。因此不管对方作何选择，使自己收益最大化的策略总是 D 策略。从这个角度讲，在不知道对方会做出什么选择时，试图最大化自己收益的个体总会选择 D 策略，这导致的结果就是双方都选择 D 策略，最终每人都会被监禁 8 个月。

有意思的是，这种策略组合对应的集体收益却是最低的。相反，如果两个个体都选择 C 策略，那么两人的总刑期是最短的，也就是说这种策略组合对应的集体收益最大。如果使用更加准确的博弈论语言来描述，这就是纳什均衡（D,D）的整体收益小于帕雷托最优（C,C）的整体收益。其中纳什均衡指的是这样一个状态，在此状态下个体无法通过单方面改变策略来提高自己的收益；而在帕雷托最优对应的状态下，所有参与者的整体收益最高。很显然，囚徒困境博弈中的纳什均衡和帕雷托最优并不相同，这表明了在囚徒困境中个体利益最优和集体利益最优是相冲突的，这正是“困境”的由来。

基于经济学中理性人的假设，博弈论研究理性人如何在互动的情况下做决策。囚徒困境是博弈论中非常经典的博弈模型，刻画了为什么两个理性的个体在选择合作能够得到最优收益的情况下依然有可能选择不合作。在实际生活中，我们可以遇到很多囚徒困境的实际案例，例如国家间的军备竞赛（见图 2.2），两个生产商的产品价格竞争（见图 2.3）等等。这些案例有很多不同的形式，但是其核心思想是一致的：即人类生活中个体利益与集体利益经常是互相冲突的。由此可见，虽然囚徒困境模型非常简单，但是却形象地描述了生活中普遍存在的困境。

虽然上面所述的囚徒困境模型能够预测完全理性个体的行为，但是其预测和真实的人类行为相比依然存在系统性的偏差。因为上述博弈只进行一次，所以暗含了一定的前提假设：参与者只关注当前利益而不考虑长远利益。例如，对方出狱后不会对自己进行报复，或者自己做出背叛对方的行为不会影响到自己今后在朋友中的声誉等。与之不同的是，在现实中，囚徒们选择策略时的衡量标准并非只有刑期的长短，他们可能还会考虑刑期之外的某些因素，例如出狱后会不会被同伙报复等，这导致实际中的情况

图 2.2 美苏军备竞赛

图 2.3 价格战

与之对应，重复博弈模型能够更好地描述现实中发生的情况。在重复的囚徒困境博弈中，双方会反复进行囚徒困境博弈。在这种情况下，参与者可能不会像单次博弈时只考虑眼前利益，还会关心长远收益。同时，每个参与者都有机会在下一回合“报复”对手在上一回合的不合作行为，在博弈论语言中，这种行为被称为“惩罚”。如果博弈重复的次数足够多且未知，合作策略对理性个体来说可能是最好的选择。不仅如此，根据博弈论中的无名氏定理，在无限期的重复博弈中，如果参与人足够有耐心且对未来足够重视，那么任何程度的合作都是一个子博弈精炼纳什均衡。

三、重复囚徒困境研究

在现实世界中，人与人之间、动物与动物之间一生可能会多次相遇而非只相遇一次，例如同一个工作单位的同事、同一片森林生活的动物。多次相遇使得个体与个体之间的重复博弈成为可能，也使得“未来的阴影”开始起作用。“未来的阴影”指的是当博弈被重复进行时，个体在决策的时候不仅会考虑当前决策带来的收益，同时也会考虑当前的决策对未来交互的影响。例如，单次囚徒困境博弈下，背叛可能是使得自己利益最大化的策略；但在重复博弈中，当前的背叛可能会招致对方日后报复性的背叛，从而使得自己长期的收益受损。既然如此，个体在“未来的阴影”的笼罩下是否会采取合作性的策略、最优的策略是什么，这些问题引起了相关学者的研究兴趣。

对重复囚徒博弈的研究中最著名的是密歇根大学的政治学教授 Axelrod 发起的计算机锦标赛。20 世纪 70 年代后期，Axelrod 在经典囚徒困境的基础上进行了拓展探索，想要确定重复囚徒困境中合作能够出现的条件，于是他开展了一个计算机锦标赛并面向全球的博弈论专家们征集参赛的计算机策略。不同的策略之间进行多回合的博弈，然后每一个策略的所有收益被累加起来，最后累计收益最多的策略被选为最优策略，其收益矩阵如下图所示。

$$\begin{matrix} & C & D \\ C & R & S \\ D & T & P \end{matrix} = \begin{matrix} & C & D \\ C & 3 & 0 \\ D & 5 & 1 \end{matrix}$$

该比赛一共进行了两次。在第一次比赛中，提交的策略一共有 14 种，再加上随机策略（即以相同的概率在每一轮使用合作或背叛）一共有 15 种策略。每种策略都与其它所有策略以及它自身（即博弈双方使用相同的策略）进行配对，每对策略之间的博弈进行 200 回合，整个循环赛事重复 5 次以提高得分的可靠性。在进行了 14 万次对局，以及 24 万种不同选择后，最终获胜的是最简单的以牙还牙策略（“Tit for tat”，以下简称为 TFT 策略），它是参赛者 Rapoport 提供的。TFT 策略具体是指在第一回合选择合作，而后在每一回合采用对手在上一回合使用的策略。例如，如果对手在上一回合采用了合作（或背叛），那么按照 TFT 策略，参与者在本回合就相应地采用合作（或背叛）。第一次比赛结束后，Axelrod 公布了比赛结果并进行了全面分析。然后，他又在全球范围内发起了第二次比赛，这次比赛的目的和赛制与第一次相同，一共收到了来自世界各地的 62 种策略。很多提交的策略为了取得优势，专门针对对手的策略设计了非常复杂的规则。但是结果出乎人们的意料，最后取胜的依然是 TFT 策略。在两轮锦标赛之后，Axelrod 总结了几条重复囚徒困境中所使用的策略获得较高分数的建议：

- 不要嫉妒
- 不要首先背叛
- 合作和背叛都要回报
- 不要耍小聪明

值得注意的是，TFT 策略最终能赢得比赛不是因为 TFT 在对战所有其它策略时都能打败对手，而是因为它跟所有策略对战时都能取得很不错的分数。

虽然 TFT 策略赢了两次比赛，但是该策略也存在一个明显的缺点，即 TFT 无法更正错误。这在 Axelrod 发起的锦标赛中没有体现出来，因为锦标赛中策略之间的博弈是按照写好的程序运行的，因此不会出现错误，然而真实世界充满各种噪声，例如，人类在博弈中很容易犯错误，他们可能想要合作但却错误地选择了背叛。一旦存在噪声，TFT 策略很容易陷入交替背叛，甚至完全背叛的境地。图 3.1 描述了这样的情形：最初，采用 TFT 策略的 A 个体和 B 个体都选择合作；之后，A 个体不小心犯错、选择了背叛，接下来的几个回合 A 个体和 B 个体陷入了交替背叛；随后，B 个体也犯了错，于是，A 个体和 B 个体进入了完全背叛的境地。

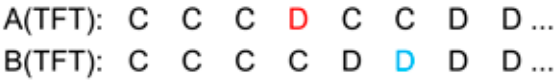

图 3.1　TFT 出现两次错误陷入完全背叛

Axelrod 发起的比赛及其结果引起了牛津大学 Nowak（现为哈佛大学数学和生物学教授）的兴趣。基于噪声对于合作进化产生的影响，Nowak 从现实世界的不确定性中找到了比 TFT 策略更优的策略。从上文我们可以知道由于重复囚徒困境是两人多次博弈的过程，所以它的策略空间非常庞大，如果可以将所

需计算的策略空间减小，那么就更容易找到好的策略。根据这个思路，Nowak 提出了反应策略的概念，和之前的（确定性）策略不同的是，反应策略是概率性的策略。并且反应策略只考虑上一回合对手的行为，根据对手的行为再选择自己这一回合的行为。这种反应策略（如图 3.2 所示）由两个参数给出，即 s=（p, q）：p 和 q 分别代表对手在上一回合选择合作和背叛的情况下，自己在这一回合选择合作的概率。

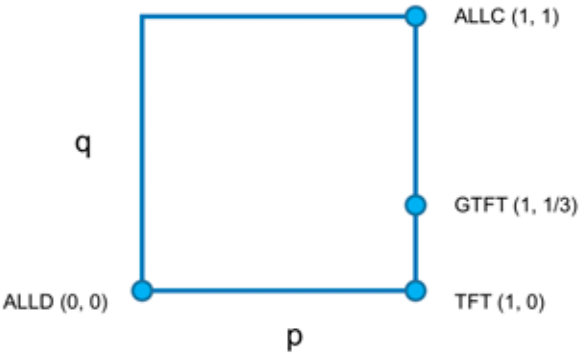

图 3.2 反应策略的示意图

在 Axelrod 所举办的比赛思路下，Nowak 和 Sigmund 设计了一个计算机仿真实验来模拟生物种群中反应策略的进化动态，仿真思路如表 3.1 所示。此时，p 和 q 分别属于[0,1]区间，他在此区间内随机生成了服从均匀分布的 100 个反应策略，代入收益矩阵和复制方程，得到了如图 3.3 所示的结果。我们可以看到：起初，100 个反应策略随机均匀分布在[0,1]x[0,1]区间内；20 代之后，最具合作性的策略趋于消失，整个群体朝着始终背叛策略（“Always defect”，以下简称为 ALLD 策略）进化；之后，剩下少数 TFT 策略和 ALLD 策略进行竞争，此时合作者的数量便突然变得多了起来；接着，TFT 策略被大度的 TFT 策略 GTFT 策略（“Generous tit-for-tat”，以下简称为 GTFT 策略）所取代。所谓 GTFT 策略，就是 p=1 且 q=1/3 时的策略，即 s=（1,1/3），如图 3.2 所示。为什么会出现这种现象呢？这是因为 GTFT 策略与 TFT 策略相比起来更加宽容对手的背叛行为：它在对方选择合作时总是合作；在对方采取 D 策略时以 1/3 的概率宽容对方的背叛，伸出合作的橄榄枝。因为 GTFT 策略比 TFT 策略更宽容，所以在背叛行为占主导的时候，TFT 策略得到的收益比 GTFT 策略高，于是 TFT 策略能存活下来；一旦 TFT 策略的使用者达到一定规模，由于 TFT 策略无法纠正偶然的背叛行为，TFT 策略就会被更加宽容的 GTFT 策略所取代。在这个过程中，TFT 策略的角色更像是催化剂，在初始阶段促进合作的产生，但是后期又被 GTFT 策略所取代。

然而，该结果并不是演化的终点，因为上述过程只考虑了自然选择（即优胜劣汰的选择过程）。而对于生物进化而言，变异也是非常重要的。当考虑策略的变异时，一个比较有意思的现象是合作策略和背叛策略之间的振荡转化现象。首先，背叛策略会获得优势，ALLD 策略成为第一个获胜策略；之后 TFT 策略最能激发合作，合作策略数量增多，一旦 TFT 策略大量存在，之后又会被 GTFT 策略所取代；GTFT 策略会演化到更为宽容的合作性策略上；这个时候由于过度宽容，背叛性策略又会通过剥削获得优势，从而不断地循环往复。

Nowak 和 Sigmund 的计算机仿真实验思路
1. 生成初始策略：在 p 和 q 属于 [0,1]区间内随机生成服从均匀分布的 100 个反应策略（因此，初始种群不是同质的，由 100 种不同类型的个体组成） 2. 一代更新： a) 策略之间两两进行博弈（包括和自己进行博弈），即共进行 10000 次博弈，每个策略获得累计收益 b) 按照离散时间的复制方程更新策略频率：策略的频率和在步骤 a)获得的收益成比例，收益越大更新后的频率越大 c) 记录更新后的每种策略的比例和分布 3. 重复步骤 2，记录每一代的策略变化情况 4. 画出初始种群，第 20 代，第 100 代，第 150 代，第 200 代和第 1000 代的策略分布图（见图 3.3）

表 3.1 Nowak 计算机仿真实验思路

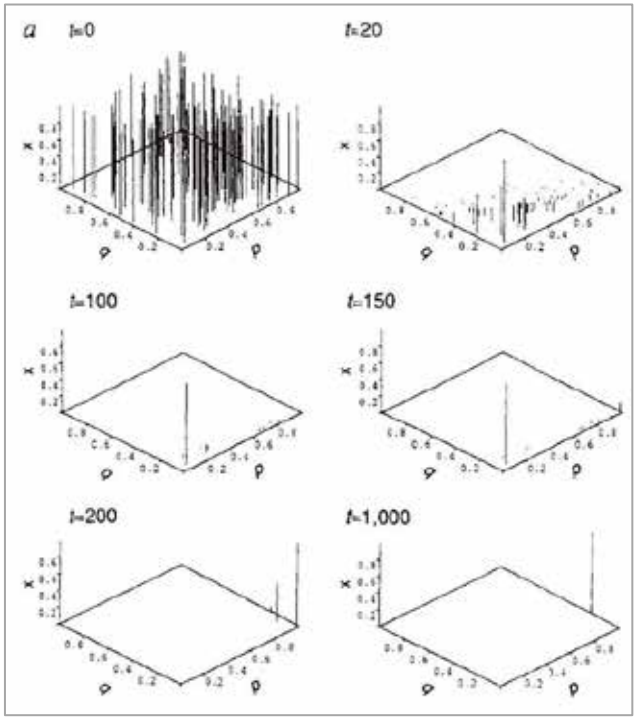

图 3.3 具有随机策略(p,q)的有限种群策略进化图

（源自 Nowak M A, SIGMUND K. Tit for Tat in Heterogenous Populations [J]. Nature, 1992, 355(6357):250-253）

以上介绍的是反应策略，更一般的策略是一步记忆策略。这种策略综合考虑上一回合对手和自己的行为，然后再选择自己这一回合的行为。一步记忆策略由四个参数给出 s=（p1,p2,p3,p4）：p1/p2/p3/p4 分别代表在上一回合自己和对方策略为 CC/CD/DC/DD 组合的情况下，自己在这一回合选择合作的概率。此时，TFT 策略可以写成 s=（1,0,1,0），GTFT 策略为 s=（1,1/3,1,1/3）。在一步记忆策略下，Nowak 和 Sigmund 在随后的实验中发现了一种新的策略，s=（1,0,0,1）。这种策略在第一回合会合作；随后，当上一回合自己和对手的策略为 CC 或者 DD 时，自己在这一回合会选择合作，反之则选择背叛。这个策略总是遵循简单的“赢留输变”（“Win-stay, lose-shift”，以下简称 WSLS 策略）原则。在有噪声的环境下，WSLS 策略相比 TFT 策略更有优势，它能纠正偶然出现的错误（如图 3.4 所示），并且它能无情地剥削更为合作的策略，例如始终合作策略（“Always cooperate”，以下简称为 ALLC 策略）（如图 3.5 所示），这使得更为宽容的策略无法取代 WSLS 策略，因此 WSLS 策略是演化稳定的。

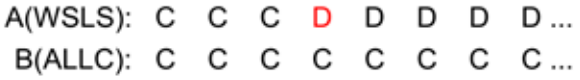

图 3.4 WSLS 策略出现噪声时能够修正错误

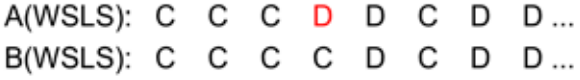

图 3.5 WSLS 策略相对 ALLC 策略占优

参考文献

[1] von NEUMANN J. Zur Theorie der Gesellschaftsspiele[J]. Mathematische Annalen, 1928, 100(1): 295-320.

[2] von NEUMANN J, MORGENSTERN O. Theory of Games and Economic Behavior[M]. Princeton University Press, 1944.

[3] NASH J F. Equilibrium points in n-person games[J]. Proceedings of the National Academy of Sciences, 1950, 36(1): 48-49.

[4] NASH J F. Non-cooperative games[J]. Annals of Mathematics, 1951: 286-295.

[5] 程代展. 矩阵代数、控制与博弈[M]. 北京理工大学出版社, 2017.

[6] TADELIS S. Game Theory: an Introduction[M]. Princeton University Press, 2013.

[7] AXELROD R, HAMILTON W D. The evolution of cooperation[J]. Science, 1981, 211(4489): 1390-1396.

[8] AXELROD R. The evolution of cooperation[M]. Basic Books, 1984.

[9] NOWAK M A, SIGMUND K. Tit for tat in heterogeneous populations[J]. Nature, 1992, 355(6357): 250-253.

[10] NOWAK M A, SIGMUND K. A strategy of win-stay, lose-shift that outperforms tit-for-tat in the Prisoner's Dilemma game[J]. Nature, 1993, 364(6432): 56-58.