标题 **title**

作者 *author*

2024 年 9 月 11 日

前言

目录

第一部分

AI 的逻辑

第二部分

信息与数据

第三部分

决策与优化

第四部分

逻辑与博弈

第九章　逻辑与博弈

2016 年 3 月，围棋界迎来了一场前所未有的挑战——谷歌 DeepMind 团队开发的人工智能 *AlphaGo* 挑战韩国围棋九段世界冠军李世乭. 这场比赛不仅引发了全球的关注，更成为了人工智能领域的里程碑.

围棋被认为是极其复杂的游戏，其复杂性远远超越国际象棋，因此，许多人曾认为围棋是人工智能无法攻克的“堡垒”. 然而，AlphaGo 改变了这一看法.

在首局比赛中，李世乭似乎还没有完全适应对手是一台超级计算机. 起初，李世乭运用了传统的围棋策略，期待通过人类棋手的经验与直觉来击败 AlphaGo. 然而，比赛后期，AlphaGo 展现出极其强大的计算能力，持续挖掘并利用局面中的潜在机会. 最终，李世乭被逼至绝境，AlphaGo 成功赢下了第一局.

第二局比赛成为整个系列赛的关键点，也正是在这一局中，AlphaGo 下出了它最令人惊叹的一步——第 37 手. 这一手棋打破了人们对围棋的传统理解，AlphaGo 将白子下在了一个似乎毫无意义的位置，许多围棋专家和职业棋手一度认为这是“臭棋”. 李世乭一度陷入沉思，走出赛场短暂休息.

然而，随着局面的展开，这步棋逐渐展示出了它的深远战略意图，它不仅打乱了李世乭的布局，还为 AlphaGo 赢得了巨大优势. 最终，李世乭输掉了第二局，这一局被认为是 AlphaGo 表现出超越人类直觉的关键胜利.

第三局中，李世乭试图改变策略，以更加复杂、创新且进攻的方式应对 AlphaGo. 然而，AlphaGo 表现得更加冷静和高效，它不仅成功化解了李世乭的进攻，还逐渐将局面转变为对自己有利的形式. 在对局的后期，李世乭再次被迫认输. 至此，AlphaGo 以 3:0 的比分提前赢得了这场五局比赛的胜利.

尽管前面三局失利，李世乭并没有放弃. 在第四局中，他展示了超凡的创造力和直觉，走出了被称为“神之一手”的第 78 手. 这一手棋打破了 AlphaGo 的计算预期，突然扭转了局面，让 AlphaGo 陷入困境. 尽管 AlphaGo 做出了顽强的抵抗，但李世乭凭借这一步棋最终赢得了这一局胜利. 这是人类在整个比赛中唯一的一胜.

在最后一局比赛中，李世乭保持了极高的斗志，但 AlphaGo 通过深度学习积累的经

验和计算能力再次发挥作用. 尽管李世乭尽力应对，但 AlphaGo 在关键时刻掌控了局面，最终赢得了第五局的胜利. 整个比赛以 4:1 的结果结束，AlphaGo 取得了压倒性的胜利.

李世乭与 AlphaGo 的第四局对决，不仅是那一次比赛的唯一一次胜利，也是此后人类与顶尖围棋人工智能较量中的最后一次胜利. 而第二局 AlphaGo 的神之一手，人类至今不能理解，只能效仿. AlphaGo 通过学习人类棋谱，再通过自我对弈，最终超越了人类的认知，成为了围棋的新王者.

毫无疑问，这一比赛彻底的改写了围棋的历史. 过去，围棋被视为一种具有智慧和创造力的艺术；但现在，围棋选手获胜唯一的出路是模仿人工智能的策略. 后来，*AlphaGo Zero* 横空出世，它完全不依赖人类知识，但是完胜 AlphaGo. 人类积累了几千年的围棋经验，在人工智能面前显得如此渺小.

围棋代表了一种特别的决策与优化问题：我们的决策依赖于对手，而对手的决策又依赖于我们. 这样的决策问题形成了*博弈论*的研究对象. 博弈是如此复杂，以至于如何恰当地描述博弈的过程都是一个巨大的挑战. 本章的目标是给出博弈论的通用语言和基本概念，以及一些经典的博弈模型和他们在人工智能中的应用.

§9.1 博弈的基本语言：以井字棋为例

大家都玩过井字棋，这是一个简单的博弈. 如**??**[lhy: 画个图] 所示，在棋局中，两名玩家轮流在一个 3×3 的棋盘上放置自己的标记（X 或 O），直到有一方连成一条线（横、竖、斜）或者棋盘填满，在前一种情况下，这个玩家获胜，否则平局.

上面的描述是自然语言，并不能被计算机直接理解. 我们需要将这个博弈的过程形式化，以便计算机能够理解和处理. 在井字棋中，有如下的基本概念：

- *玩家*：两名玩家，一个执 X 子，另一个执 O 子.
- *棋盘局面*：棋盘的当前状态，包括每个格子的占据情况（X，O 或空）.
- *行动*：每个玩家轮流在空格中放置自己的棋（X 或 O），直到出现胜负或棋盘填满.
- *收益*：游戏结束时，根据游戏的结果确定每个玩家的收益，胜者为 $+1$，平局为 0，败者为 -1.

以上概念足够描述*博弈*是什么了. 然而，它不足以描述*玩家*是如何下棋的. 为此，我们需要引入*策略*的概念. 我们将在本章中看到，如何定义策略是博弈论中最为复杂的问题之一. 此刻，我们只关注井字棋这一场景中的策略.

我们假定玩家都有充分大的计算能力和记忆力. 于是，玩家可以记住这一次游戏中所有的局面，以及每个轮次的自己和对手的行动. 玩家可以知道自己的内心活动（也就是有内省的能力），但是，玩家绝不可能知道对手的内心活动，更不知道他下一步会怎么走. 总而言之，玩家只能知道对大家都是公开的这些信息以及自己独有的信息.

在知道所有的信息之后，玩家需要决定每一轮的走法. 或许他会猜测对手的心理活动以及策略，并以此为根据做出自己的决策. 他也可能完全不管对手的行动，而是我行我素. 无论如何，玩家的决策都是基于他所知道的信息，因而我们可以认为玩家的决策是一个映射，将他知道的信息映射到他的行动空间.

因此，玩家的策略，就是一个映射，给定当前处于哪个轮次、所有历史局面和行动之后，它会输出下一步的行动.

需要注意的是，每个玩家在开局的时候就要选好自己的策略，此后只能遵循这个策略进行行动. 初看之下，这一定义是极强的，我们似乎无法在游戏中途做出调整. 然而，这一定义其实是合理的，因为“调整”本身也是策略的一部分.

例如，一个策略可以是“如果对手走了这一步，那么我就走这一步；否则，我就走那一步”. 这其实就是调整. 策略也可以包括自我反省和对对手的猜测. 例如，一个策略可以是“我刚刚下的这几步棋不是很好，我应该调整策略，尽量避免这种情况再次发生”. 另一个策略可以是“如果对手走了这一步，那么我就认为他是这样的人”.

到此，我们不仅定义了博弈的基本概念，还定义了玩家的策略. 有了这两个概念，我们就可以真正地让井字棋博弈进行起来了：两名玩家根据自己的策略产生行动，而棋盘则产生新的局面，直到游戏结束，然后获得收益.

接下来，我们讨论不同类型的博弈，以及他们对应的理论和应用.

§9.2 输赢博弈

输赢博弈指的是玩家的收益只能取两个值（输或赢，-1 或 1）的博弈. 输赢博弈中，我们通常会有多轮博弈，每轮博弈的结果会影响下一轮博弈的局面，通常，这种博弈被称为**扩展式博弈**. 围棋、象棋、斗地主都是输赢博弈.

§9.2.1 博弈的不同维度

输赢博弈有多种维度的分类方式，见表 9.1. 这些分类都是比较直观的. 但是，后面三个概念可能较为难以和形式化对应，我们这里加以解释.

- 完全信息与非完全信息：尽管这是一个直观的概念，但是如何在数学上区分完全信

二人	多人
输赢	输赢平
有限深	无穷深
完全信息	非完全信息
确定性	非确定性
非合作	合作

表 9.1: 输赢博弈的分类.

息与非完全信息确实极其困难的，我们这里给一种方法.

我们将博弈本身也看成一个玩家[1]，那么，完全信息意味着，任何玩家可以不依赖其他玩家，自己模拟出整个博弈的进行过程. 换句话说，他可以“扮演”其他任何角色. 反之，非完全信息意味着，玩家不能模拟博弈，这实际上意味着他无法获取所有需要的信息来进行模拟.

- 确定性与非确定性：确定性的意思是，给定当前格局和所有玩家的行动，可以唯一确定下一回合的格局. 例如，井字棋就是一个确定性博弈，因为每一步棋都会导致唯一的下一步棋局.

  与之相对的概念是*非确定性*，比如，考虑一个非常简单的博弈. 两名玩家轮流掷硬币，如果都是正面朝上，那么第一名玩家获胜，否则第二名玩家获胜. 这个博弈是非确定性的，因为玩家的行动（掷硬币）会导致多种可能的结果.

- 非合作与合作：在非合作博弈中，每个玩家的决策不会被其他玩家的影响，每个玩家都是在为自己的利益而行动. 在合作博弈中，玩家之间可以合作，共同制定策略，共同获得收益. 因此，合作博弈中的收益和策略都依赖于哪些玩家进行了合作.

注. 我们这里给出的关于完全信息的定义其实借鉴了密码学中的*零知识证明*的概念. 我们这里只给一个例子说明这个概念. 假设有甲乙两人，甲宣称自己是一个硬币鉴定大师，给任意两个硬币，他可以判断出这两个硬币是不是一样的. 乙不确定甲是不是骗子，所以想要验证这一能力. 而甲并不希望乙通过验证的过程学到他的鉴定方法.

于是，我们可以这样做：乙秘密随机准备两枚硬币，一样或者不一样，然后把这两枚硬币交给甲，甲进行鉴定，然后把硬币还给乙. 如此进行多次，如果甲能够正确判断每一次，那么乙就可以相信甲的能力.

[1]通常，在扩展式博弈中，我们将它称之为“天”（nature）. 这里借用了中国传统文化的观念，“天”常被视为一种至高无上的力量或存在，例如“天命”和“无法无天”.

图 9.1: 斗地主.

如何判断是零知识？直观上，乙不知道除了硬币之外的任何信息，所以他无法模拟出整个过程. 我们可以如下定义：如果乙只知道甲有这个能力，但是不知道甲的鉴定方法，他依然可以把整个过程模拟出来，那么这个过程就是零知识的.

我们给一个具体的例子.

例 9.1 斗地主是一个多人有限轮非完全信息合作输赢博弈. 这个博弈有三个人，两个农民和一个地主，农民和地主是两个阵营. 三个人轮流出牌，如果不能出牌，要摸牌，直到有一个人出完牌. 先出完牌的阵营获胜.

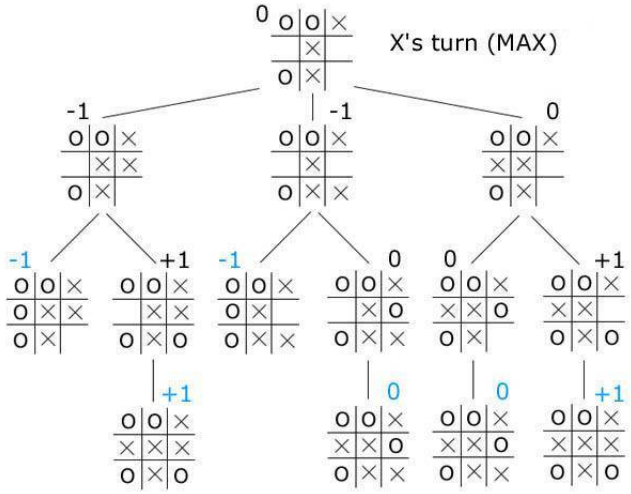
“多人”是显然的，有限轮是因为牌是有限多的，非完全信息是因为有摸牌，因此每个玩家只知道自己的牌，不知道其他玩家的牌. 合作是因为农民之间可以合作，地主是一个人. 输赢是因为有且只有一个阵营先出完牌.

我们在本部分主要关注最简单的一种博弈，即*完全信息确定性回合制博弈*. 这样的博弈可以用博弈树表示出来，例如，井字棋的博弈树可以画作图 9.2.

§9.2.2 Zermelo 定理与 AlphaGo Zero

输赢博弈一个自然的问题是：玩家是否总可以获胜？这就涉及到*必胜策略*的概念：无论对手如何进行行动，玩家都可以取得胜利的策略. 必胜策略是一种*解概念*，即给定一个博弈，求解具有一定性质的玩家策略. 如果某个玩家具有必胜策略，那么我们就说这个博弈是*被决定的*.

什么博弈是被决定的？这一问题的答案由 Zermelo 定理给出.



[lhy: 重画]

图 9.2: 井字棋的博弈树

定理 9.1 (Zermelo 定理, Von Neumann) 如果一个博弈是双人的、有限深的、确定的、完全信息的、输赢的，那么这个博弈是被决定的.

以上限定词缺一不可，缺少了任何一个都可能导致结论不成立.

证明. (证明一：逻辑证明) 设 W_i 表示“玩家 i 获胜”, $i = 1, 2$. 于是 $x \in W_1 \iff x \notin W_2$.

先手玩家有必胜策略当且仅当

$$\exists a_0 \forall b_0 \exists a_1 \forall b_1 \ldots \exists a_n \forall b_n : (a_0 b_0 \ldots a_n b_n) \in W_1.$$

后手玩家有必胜策略当且仅当

$$\forall a_0 \exists b_0 \forall a_1 \exists b_1 \ldots \forall a_n \exists b_n : (a_0 b_0 \ldots a_n b_n) \in W_2.$$

两个命题互为否定，因此二者恰有一个成立！ □

证明. (证明二：后向归纳法) 从博弈树的叶节点往根节点推理，见图**??**[lhy: 画个图].

如果此节点是玩家 i 的回合，那么往后一轮的局面已经完全确定.

- 如果有一种走法使得玩家 i 必胜，那么玩家 i 选择这种走法即可.
- 否则，玩家 i 无论如何也不可能获胜.

当到达到根节点的时候，有一方有必胜策略，另一方必输.

这种证明方式被称为*后向归纳法*：从最后一期开始往前推理，最终确定策略. □

如果博弈的结局还有平局，我们有如下 Zermelo 定理：

定理 9.2 (有平局的 Zermelo 定理) 如果一个博弈是双人的、有限深的、确定的、完全信息的，博弈的结果有输赢平局三种，那么下面三条有且仅有一条成立：

- 第一个玩家有必胜策略.
- 第二个玩家有必胜策略.
- 双方都有不败策略.

证明见习题[lhy: 出一下].

尽管 Zermelo 定理的第二个证明构造出了必胜策略，但是后向归纳法的搜索空间过于庞大. 例如，充分大但有限的棋盘上，五子棋先手玩家存在不败策略（见习题[lhy: 出一下]），但是没有经过训练的人类或者简单的算法先手不一定会胜利.

究其原因，人的思考以及机器搜索的过程实际上是前向探索的过程. 如何进行（启发式）搜索是取得胜利重要的因素. 在本章开头，我们讲述了 AlphaGo 的故事，这是正是一个很好的例子. 下面我们就如何对围棋进行建模进行讨论.

由 Zermelo 定理可知，围棋也存在必胜策略. 然而标准围棋棋盘大小为 19×19，状态空间量级为 10^{170}，过大的状态空间使得我们无法使用后向归纳法求解出必胜策略. 以下我们探讨 AlphaGo Zero（下面简称 Zero）如何通过神经网络建模博弈的过程.

首先，我们假定 Zero 可以记住的是从当前局开始往前 k 步的棋局信息（即落子方式）. 我们假定这样的 k 步棋形成的棋局序列集合为 C. 于是，Zero 的策略是一个随机映射

$$\pi : C \to \Delta(\mathcal{A}),$$

其中 $\mathcal{A}$ 是所有可能的落子方式的集合，而 $\Delta(\mathcal{A})$ 是 $\mathcal{A}$ 上的概率分布. 这里，我们假定 Zero 的策略是一个随机策略，而非确定性策略. 此时，概率分布 $\pi(s)$ 表示在状态 s 下，Zero 选择对应的落子方式的似然（或者对胜利的自信程度）.

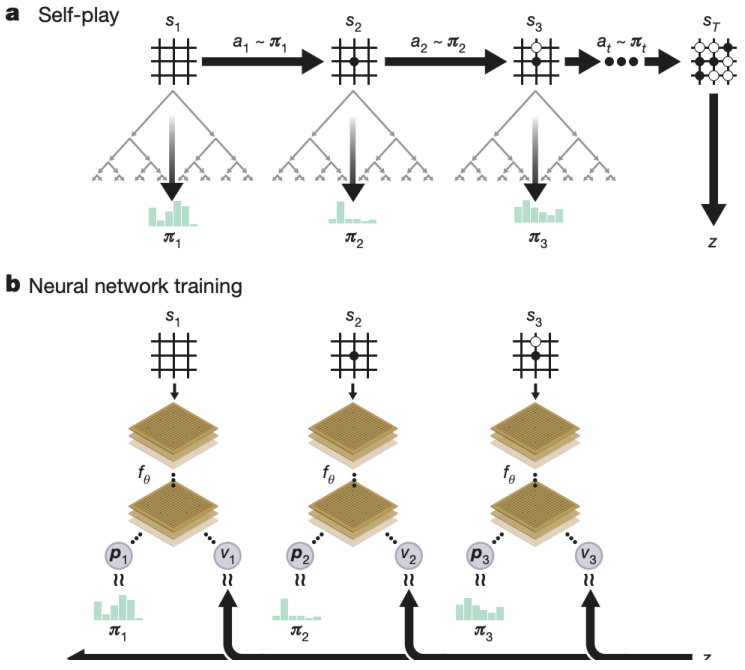
最后，当博弈结束时，Zero 会获得一定的收益，我们假定 Zero 赢的时候收益为 $+1$，输的时候收益为 -1.

对于人类来说，我们的任务是让 Zero 的策略 π 尽可能地接近必胜策略，为此，我们需要用一个神经网络来拟合这个策略. 此外，我们通常需要告诉 Zero 每一步棋获胜的概率（或者说期望收益），这也需要一个神经网络. 具体来说，AlphaGo Zero 算法包含策略网络，价值网络和 *Monte-Carlo* 树搜索（MCTS）.

- 策略网络 p 和价值网络 v 的输入为当前状态 $s \in C$，即 $(P(s,\cdot), V(s)) = f_\theta(s)$.

- 策略网络 $P(s,\cdot)$ 的输出为下一步落子位置 $a \in \mathcal{A}$ 的概率分布.
- 价值网络 $V(s)$ 的输出为该状态的价值评估（期望收益、胜率）.
- MCTS 利用策略网络进行扩展，使用价值网络进行评估，利用 UCB 公式返回最优的搜索结果作为落子决策.

Zero 使用强化学习（自博弈，策略梯度）的方式训练策略网络，使用自我博弈过程中的数据监督训练价值网络. 这个过程如图 9.3 所示.



[lhy: 重画]

图 9.3: AlphaGo Zero 的训练过程

在开始细节之前，我们先给一个比喻，看看整个流程是如何模拟人类学习的. 假设 Zero 是一个人.

- 他可以按照他的技术水平（即启发式搜索函数）在脑中模拟一场围棋比赛，并且假定对手和自己一样聪明. 最终，这一模拟会有一个结果（输或者赢），这一结果反映了他的策略的好坏. 这就是用 MTCS 自博弈的过程.
- 假设他在脑中模拟了很多场比赛，他可以把这些比赛记录下来，然后根据这些记录来调整自己的策略，并调整自己对于胜率的评判. 这就是策略网络和价值网络的训练过程.

- 他可以不断重复上面两个过程，来精进自己的技术水平.

下面，我们逐步解释 Zero 的训练过程.

- 自博弈过程 $s_1, \cdots, s_T$: 在每个状态 s_t, 使用最近一次的网络 f_θ，执行一次 MCTS α_θ（具体过程见后面）. 下法根据 MCTS 计算的搜索概率 π_t 来选择，$a_t \sim \pi_t$. 最后，依据围棋规则，对终止状态 s_T 打分，来计算胜利者 z.

- 神经网络训练：使用原始的棋盘状态 s_t 作为输入，输出 $(p_t, v_t) = f_\theta(s)$，表示当前玩家在 s_t 的策略和胜率. 训练时更新网络参数 θ，以最大化策略 p_t 和搜索概率 π_t 的相似性，并最小化预测赢家 v_t 与实际赢家 z 的误差. 新参数将应用于下一次自博弈 a 的迭代.

MTCS 的过程较为复杂，我们单独介绍. 树的组成如下：搜索节点是状态 s，边是状态-行动对 (s, a). 每条边需要存储以下信息：

- $N(s, a)$：边的访问次数.

- $P(s, a)$：策略网络在状态 s 中选择行动 a 的概率.

- $Q(s, a)$：动作价值，$Q(s, a) = \frac{1}{N(s,a)} \sum_{s':s,a\to s'} V(s')$，其中 V 是价值网络. 这一值反映了在状态 s 选择行动 a 的平均收益.

接下来，MTCS 要做如下迭代（见图 9.4）：

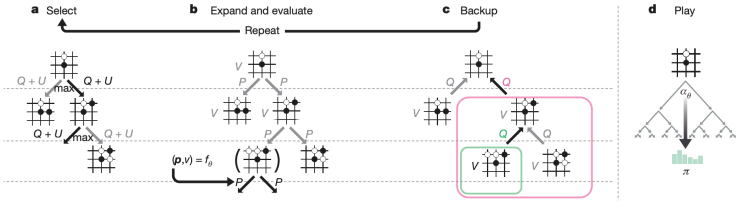
- 从根节点开始，状态 s 固定，选择具有最大的 $Q(s, a) + U(s, a)$ 的分支.
  - U 是上限置信度，$U(s, a) \propto P(s, a)/(1 + N(s, a))$.
  - $Q + U$ 是置信区间的上界，称为 UCB 公式.

  这一过程其实就是在模拟自己和对手的多轮行动，其中选择 $Q + U$ 最大的分支即是启发式搜索的形式.

- 当选到叶节点时，扩展叶节点. 使用神经网络 $f_\theta(s)$ 来计算新的 $P(s, a)$ 和 $V(s)$，并把 P 存储到对应的边上. 只要还可以扩展，就说明还有一方玩家可以继续行动，所以这一过程可以持续到有一方获胜或者到达最大深度.

- 根据 V 更新动作价值 Q，反映所有该动作的子树的平均值. 这反应了此次模拟的结果如何影响这一动作的评价：输了的话，这一动作的评价会降低，赢了的话，这一动作的评价会提高.

- 一旦搜索结束，返回搜索概率 π，$\pi(a)$ 正比于 $N(s,a)^{1/\tau}$，τ 是一个参数，控制着温度. 温度反映了 π 允许的随机性程度，当 τ 趋于正无穷的时候，π 趋于均匀分布，当 τ 趋于零的时候，π 趋于一个退化分布，以概率 1 取最大 $N(s,a)$ 对应的 a.



[lhy: 重画]

图 9.4: MCTS 的过程

§9.3 正则形式博弈

输赢博弈被建模为*扩展形式博弈*，它代表了多轮博弈的过程. 与之相对的是*正则形式博弈*，即玩家只有一次行动的机会，所有玩家同时操作. 正则形式博弈通常要求信息是完全的. 那么，如何定义正则形式博弈和对应的解概念呢?

§9.3.1 定义

我们直接给出一个很直观的的模型.

定义 9.1 (正则形式博弈) 一个**正则形式博弈**由以下要素构成:

- 玩家集合: I，我们总是假设这是一个有限集合.
- 玩家的行动空间: A_i，$i \in I$.
- 玩家的收益: $u_i : \prod_j A_j \to \mathbb{R}$. □

有以下特殊的正则形式博弈:

- 当 A_i 有限，我们称之为*矩阵博弈*. 之所以称为矩阵博弈，是因为我们可以用一个矩阵来表示玩家的收益.

  比如，考虑 $|I| = 2$，即有两个玩家，$|A_1| = m, |A_2| = n$，那么要确定 u_i 只需要确定 $u_i(a_1, a_2)$ 的值，这恰好就是一个 $m \times n$ 的矩阵. 此时，我们将第一个玩家称之为*行玩家*，第二个玩家称之为*列玩家*.

对于更多玩家的博弈，尽管 u_i 不再可以被矩阵表示，但上面的这一表示的思路依然适用，所得到的结果在线性代数上称之为*张量*. 但是遵循传统习惯，我们依然把这种博弈称之为矩阵博弈.

- 当 A_i 是 $\mathbb{R}^n$ 的连通子集且和 u_i 都是连续的，我们称之为*连续博弈*.
- 当 $\sum_i u_i = 0$，我们称之为*零和博弈*. 一般地，如果存在常数 c 使得 $\sum_i u_i = c$，我们称之为*常和博弈*. 常和博弈和零和博弈通常具有一致的性质，所以我们也把常和博弈视为零和博弈. 如果 u_i 只有两个取值（例如 $-1, 1$），我们称之为*输赢博弈*.

下面，我们看两个典型的例子.

例 9.2 (囚徒困境) 在囚徒困境中，一共有两个玩家，行玩家和列玩家. 玩家的第一个选择是保持沉默，第二个选择是认罪并检举对方. 它有如下收益矩阵：

$$\begin{pmatrix} -1,-1 & -10,0 \\ 0,-10 & -5,-5 \end{pmatrix}.$$

矩阵每一项第一个元素是行玩家的收益，第二个是列玩家的收益. 显然，这是一个非零和的矩阵博弈.

在这个例子中，博弈的属性非常鲜明：每个人的收益不仅仅取决于自己的选择，还取决于对方的选择. 因此，玩家在做决策时，必须要考虑对方玩家可能会如何行动. □

例 9.3 (猜硬币游戏) 在猜硬币游戏中，行列玩家分别有一枚硬币，他们秘密地抛掷. 如果两个玩家的硬币上面相同，行玩家获胜；否则列玩家获胜. 收益矩阵为：

$$\begin{pmatrix} 1,0 & 0,1 \\ 0,1 & 1,0 \end{pmatrix}.$$

容易验证，这个博弈是一个零和博弈，进一步，是一个输赢博弈. □

至于连续博弈的例子，这里我们先略过. 我们会在后面详细讨论*生成对抗网络*，它是一个连续博弈的例子.

现在，我们有了博弈，接下来的任务就是定义*策略*. 假设所有人之间是不能交流的，每个人独立做决策，因此玩家之间不能协调彼此的决策. 所以，我们可以很简单地定义玩家 i 的策略 s_i 为行动空间 A_i 的一个元素. 将所有玩家的策略收集起来，我们就得到了*策略组合*：

$$s = (s_i)_{i \in I}.$$

§9.3.2 理性与均衡

接下来，我们要定义正则形式博弈中的解概念，换言之，玩家最终会如何进行这场博弈并获得对应的收益. 这里，我们将引入博弈论中两个重要的概念：*理性*和*均衡*.

尽管理性这一词已经被广泛使用，但是，究竟什么是理性依然是一个有巨大争议的问题. 比如，我们考虑这些定义，以及让他们变得很微妙的场景：

- 理性指的是无穷的信息获取能力. 我们把理性玩家处在一个到处都是抛硬币的世界，他怎么做才算理性?
- 理性是无穷的计算资源. 我们知道，这个世界上存在一些数学命题（记为 p）虽然有真假，但是我们既不能证明它，也不能证伪它（Gödel 第二不完备性定理）. 现在，我们给理性玩家设计一个博弈，如果证明了 p，他就赢，否则他就输. 这个玩家怎么做才算理性?

以上两种对于理性的理解只是冰山一角. 这里通过这些例子想说明，理性是一个非常复杂的概念，我们很难达成一个共识. 因此，更加务实的做法是，我们不去讨论理性的定义，而是直接给出一个理性的定义，然后看看这个定义下会发生什么.

*均衡*这一概念正是诞生于这样的背景下. 所谓均衡，指的是在某种理性的前提下，面对其他玩家，每个人都做出了他能做出的最优选择，这种最优选择被称为*最优反应*. 因此，所有人互相都是最优反应的时候，就构成了一个均衡.

上面的讨论总结起来就是如下的对应关系：

$$\text{理性} \iff \text{最优反应} \iff \text{均衡}.$$

博弈论的核心就在于研究以上三个概念之间的关系.

以上讨论比较抽象和一般，我们现在回到正则形式博弈，看看在这个背景下，这三个概念是如何定义的.

首先，理性的定义是非常简单的：在玩家知道其他玩家的信息（即策略）之后，他会最大化自己的收益. 然后，在这一定义下，最优反应可以自然定义如下：

定义 9.2 (最优反应) 给定对手的策略组合 s_{-i}，玩家 i 的**最优反应**是一个策略 s_i，满足对任意 $a_i \in A_i$，

$$u_i(s_i, s_{-i}) \geq u_i(a_i, s_{-i}),$$

即玩家 i 的收益最大化. 最优反应对应的策略集合记为 $BR_i(s_{-i})$. □

这里，s_{-i} 表示除了玩家 i 之外的所有玩家的策略组合，我们将会频繁使用这个记号.

最后，我们定义均衡的概念，我们很自然有如下的定义：

定义 9.3 (纯策略 Nash 均衡) **纯策略 Nash 均衡**指的是策略组合 s，对任意玩家 i 和任意 $a_i \in A_i$，有

$$u_i(s_i, s_{-i}) \geq u_i(a_i, s_{-i}).$$ □

我们也可以用不动点（见第八章）来理解纯策略 Nash 均衡. 纯策略 Nash 均衡的等价定义是每个人都达到了自己的最优反应，即最优反应的不动点. 更一般地，任何一个均衡的概念都可以理解为最优反应对应的不动点.

作为一个例子，我们继续考虑囚徒困境.

例 9.4 (囚徒困境的纯策略 Nash 均衡) 回忆例 9.2，它有如下收益矩阵：

$$\begin{pmatrix} -1,-1 & -10,0 \\ 0,-10 & -5,-5 \end{pmatrix}.$$

矩阵每一项第一个元素是行玩家的收益，第二个是列玩家的收益.

这个博弈有唯一的纯策略 Nash 均衡：每个人都认罪（第二行），此时大家都会获得 -5 的收益. 比如，行玩家选择第一行，那么，无论列玩家选了第一列还是第二列，行玩家都会选择第二行，因为这样他的收益更高. 同理，列玩家也是如此. 因此，这个策略组合是一个唯一的纯策略 Nash 均衡.

注意，如果两个选择都是保持沉默，那么他们的收益会更高（-1），但他们却无法达到这个收益. 这正是我们定义的理性所蕴含的意义：在博弈中，每个人都是自私的，如果知道了对方的选择，他们会选择最优的策略，而不会稍微放弃一点自己的利益，以换取博弈双方更大的利益. □

然而，我们继续猜硬币游戏，我们会发现这个博弈没有纯策略 Nash 均衡. 回忆，这一博弈的收益矩阵为：

$$\begin{pmatrix} 1,0 & 0,1 \\ 0,1 & 1,0 \end{pmatrix}.$$

我们现在验证，这个博弈没有纯策略 Nash 均衡. 如果行玩家选择第一行，那么列玩家的最优反应是第二列，然而，此时行玩家的最优反应是第二行；而行玩家选择第二行的情况类似. 因此，无论如何选，最优反应都不可能形成一个不动点（即均衡）.

更一般地，二人正则形式输赢博弈中纯策略 Nash 均衡往往不存在. 我们有如下定理：

定理 9.3 *设 $G = (I, \{A_i\}_{i\in I}, \{u_i\}_{i\in I})$ 是一个二人正则形式输赢博弈，其中 $I = \{1,2\}$. 那么，G 存在纯策略 Nash 均衡当且仅当其中一个玩家存在必胜策略.*

证明见习题[lhy: 出一下].

对比动态博弈中的 Zermelo 定理，静态的二人完全信息输赢博弈已经不能够保证必胜策略的存在性. 因此，静态输赢博弈的结局往往比动态输赢博弈更加不确定.

§9.3.3 生成对抗网络

接下来，我们给一个连续博弈的例子，即*生成对抗网络*（GAN）. 生成对抗网络由两个子模型组成，一个被称为*生成模型*，一个被称为*判别模型*. 生成模型的任务是生成看似真实的数据，而判别模型的任务是识别给定的数据是真实的还是伪造的. 这一过程的示意见图**??**[lhy: 画一下].

假设真实数据的分布为 F_{data}.

- 生成模型为 $G(x;\theta_g)$，参数为 θ_g，输入随机向量 x，输出数据向量 z. 当 x 服从分布 F_x，G 的输出会形成一个分布 F_g.
- 判别模型为 $D(z;\theta_d)$，参数为 θ_d，接受一个数据向量 z，输出一个 $[0,1]$ 中的实数，表示 z 来自分布 F_{data} 的概率（似然）.

我们假设 F_{data} 和 F_x 都是连续型分布，有密度函数 p_{data} 和 p_x. 我们再假设 D 和 G 都是连续的.

将 G 和 D 看成两个玩家，于是 GAN 可以被看成一个二人零和博弈，收益函数为：

$$V(G,D)=\mathbb{E}_{z\sim F_{data}}(\log D(z))+\mathbb{E}_{x\sim F_x}(\log(1-D(G(x)))).$$

D 最大化 V，G 最小化 V.

接下来，我们要解释，为什么这个收益函数能够达到我们的目标. 我们有以下三点讨论：

- 一方面，如果 D 越厉害，那么 D 会尽可能地把真实数据和生成数据区分开，$1-D(G(x))$ 也会更大，因此 V 会更大. 另一方面，如果 G 越厉害，那么 D 会更容易把真实数据和生成数据混淆，于是 $1-D(G(x))$ 会更小，V 也会更小. 因此，D 和 G 之间是一个对抗的关系.
- 如果收益函数只有第二项，判别器 D 可以“作弊”，即无论输入什么都判定为假，这样他总是能得到最大的收益. 为了避免这种情况，我们引入了第一项，即真实数据的似然. 这样，D 还需要努力判断真实数据，而不是只判断生成数据.

- 为什么 V 是对数的形式？在第三章中我们解释过，这样形式的收益函数是交叉上损失函数的形式. 因此这一损失函数可以衡量两个分布之间的相似性，对于随机的数据来说，这是一个比较好的衡量方式.

从博弈论角度出发，一个基本的问题是 Nash 均衡是否存在？假设 D 和 G 都可以任意选择连续函数. 我们将展示一种通用的方式求解连续博弈的 Nash 均衡. 注意到 $G(x)$ 形成了一个连续分布，密度记为 p_g. 首先证明密度函数存在性定理：

定理 9.4 *设 $X \sim \mathcal{U}(0,1)$. 对于任意密度函数 p，存在一个连续函数 F 使得 $F(X)$ 具有密度 p.*

证明. 设 F_p 是 p 对应的分布函数，它是一个单调的连续函数. 取 $F(x) = \inf\{y \in \mathbb{R} : F_p(y) \geq x\}$ 即可. □

因此，G 的行动等价于选择 p_g.

给定 G 的选择 p_g，我们来求 D 的最优反应 D^*.

$$V(G,D) = \int (p_{data}(x)\log D(x) + p_g(x)\log(1-D(x)))\mathrm{d}x.$$

函数 $a\log x + b\log(1-x)$ 最大值在 $x = a/(a+b)$ 的时候取得. 因此，

$$D^*(x) = \frac{p_{data}(x)}{p_{data}(x) + p_g(x)}.$$

现在，给定最优反应 $D^* = p_{data}(x)/(p_{data}(x) + p_g(x))$，我们来求 G 的最优反应. 直观上，G 能做到的最好选择就是 $p_g = p_{data}$. 此时，$D^*(x) = 1/2$，因此对任意 G，$V(G,D^*) = -\log 4$. G 选任何策略都是一样的收益，因此这是一个 Nash 均衡. 我们证明了：

定理 9.5 (GAN 的 Nash 均衡存在性) *在 GAN 的博弈中，G 选择 p_{data}，D 选择 $1/2$ 是一个 Nash 均衡.*

这一证明见习题[lhy: 出一下].

§9.3.4 混合策略

我们已经看到，在相当普遍的情况下，纯策略 Nash 均衡并不存在. 所以我们需要允许玩家进行随机行动，这就是**混合策略**.

定义 9.4 (混合策略) 给定一个正则形式博弈 $G = (I, \{A_i\}_{i\in I}, \{u_i\}_{i\in I})$，其中 I 是玩家集合，A_i 是玩家 i 的行动空间，u_i 是玩家 i 的收益函数. 玩家 i 的行动空间上的概率分布集合为 $\Delta(A_i)$. 那么，玩家 i 的**混合策略**是一个概率分布 $\sigma_i \in \Delta(A_i)$. □

当 A_i 有 n 个元素（有限），$\Delta(A_i)$ 可以被表示为标准的 n-单纯形：

$$\Delta(A_i) = \left\{ x \in \mathbb{R}^n : \sum_{i=1}^{n} x_i = 1, \forall j, x_j \geq 0 \right\}.$$

特别地，纯策略可以被看成是一个特殊的混合策略，即退化的概率分布.

有了混合策略，我们还需要重新定义玩家的理性. 一个非常标准的回答是*期望效用理论*，它由 Von Neumann 和 Morgenstern 提出. 该理论认为，在面对不确定性时，人按照期望效用进行决策.

因此，我们需要计算玩家的期望效用. 为此，引入混合*策略组合*：$\sigma = (\sigma_i)_{i\in I}$，其中 $\sigma_i \in \Delta(A_i)$. σ 是一个 $(A_i)_{i\in I}$ 上的概率分布，每一维相互独立. 当所有玩家选定策略之后，玩家 i 的期望收益是：

$$u_i(\sigma) = \mathbb{E}_{a\sim\sigma} u_i(a).$$

有了期望效用，我们可以重新定义最优反应：

定义 9.5 (最优反应) 给定对手的策略组合 σ_{-i}，玩家 i 的**最优反应**是一个策略 σ_i，满足对任意 $\sigma_i' \in \Delta(A_i)$，

$$u_i(\sigma_i, \sigma_{-i}) \geq u_i(\sigma_i', \sigma_{-i}),$$

即玩家 i 的期望收益最大化. 最优反应对应的策略集合记为 $BR_i(\sigma_{-i})$. □

最后，我们可以重新定义均衡的概念：

定义 9.6 (Nash 均衡) 对于一个博弈 $G = (I, \{A_i\}_{i\in I}, \{u_i\}_{i\in I})$，如果混合策略组合 σ 满足对于任意玩家 i 和任意 $\sigma_i' \in \Delta(A_i)$，都有

$$u_i(\sigma_i, \sigma_{-i}) \geq u_i(\sigma_i', \sigma_{-i}),$$

那么，σ 是一个 **Nash 均衡**. □

同样，Nash 均衡也可以理解为最优反应的不动点. 利用这一观点，Nash 得以使用不动点定理证明了 Nash 均衡的存在性：

定理 9.6 (Nash 均衡存在性定理) *任意有限正则形式博弈都存在一个 Nash 均衡.*

注. 目前为止，我们一共定义了两种均衡：纯策略 Nash 均衡和 Nash 均衡. 后一种均衡实际上应该被称作*混合策略 Nash 均衡*，但是为了和文献统一，我们直接称之为 Nash 均衡，而忽略前缀“混合策略”.

实际上，按照我们理性-最优反应-均衡的思路，Nash 均衡和纯策略 Nash 均衡是本质不同的两种均衡：前者引入了期望效用理论来体现玩家面对不确定性时候的行为，而后者则是直接最大化，没有不确定性. 因此，我们并不能简单说前者是后者的推广.

我们来看一个例子.

例 9.5 (猜硬币游戏的 Nash 均衡) 继续考虑猜硬币游戏，收益矩阵为

$$\begin{pmatrix} 1,0 & 0,1 \\ 0,1 & 1,0 \end{pmatrix}.$$

我们接下来证明，唯一的 Nash 均衡是两个玩家都选择 $(1/2,1/2)^\mathsf{T}$，即两个玩家都以等概率来猜硬币的正反面. 我们有如下分类讨论：

- 两个玩家都是纯策略：我们之前已经说明，这个博弈没有纯策略 Nash 均衡.
- 行玩家是纯策略而列玩家的策略是 $(p,1-p)^\mathsf{T}$，其中 $p\in(0,1)$. 考虑行玩家的期望收益：

  $$u_1(\sigma_1,\sigma_2)=\begin{cases} p, & \sigma_1 \text{ 是第一行,} \\ 1-p, & \sigma_1 \text{ 是第二行.} \end{cases}$$

  因此，如果 $p>1/2$，行玩家会选择第一行，但此时列玩家的最优反应是选择第二列，因此这不是一个 Nash 均衡. 同理，如果 $p\le 1/2$，也不是一个 Nash 均衡. 因此，这种情况也不是一个 Nash 均衡.
- 列玩家是纯策略而行玩家的策略是 $(p,1-p)^\mathsf{T}$，其中 $p\in(0,1)$. 同理，这种情况也不是一个 Nash 均衡.
- 行玩家的策略是 $\sigma_1^*=(p,1-p)^\mathsf{T}$，列玩家的策略是 $\sigma_2^*=(q,1-q)^\mathsf{T}$，其中 $p,q\in(0,1)$. 这种情况下，要确定 p,q 的值计算会很复杂. 所以我们这里介绍一个技巧.

  注意到，行玩家的任意一个混合策略可以写作

  $$\sigma_1=p_1a_1+p_2a_2,$$

  其中 $p_1+p_2=1$，a_1 和 a_2 是两个纯策略. 因此，在 Nash 均衡下，对任意 p_1,p_2，

  $$\begin{aligned} u_1(\sigma_1^*,\sigma_2^*) &= pu_1(a_1,\sigma_2^*)+(1-p)u_1(a_2,\sigma_2^*) \\ &\ge p_1u_1(a_1,\sigma_2^*)+p_2u_1(a_2,\sigma_2^*). \end{aligned}$$

取 $p_1 = 1$，$p_2 = 0$，我们有

$$pu_1(a_1, \sigma_2^*) + (1-p)u_1(a_2, \sigma_2^*) \geq u_1(a_1, \sigma_2^*).$$

同理，取 $p_1 = 0$，$p_2 = 1$，我们有

$$pu_1(a_1, \sigma_2^*) + (1-p)u_1(a_2, \sigma_2^*) \geq u_1(a_2, \sigma_2^*).$$

注意到，两个不等式的左边其实是右边的加权平均，平均值不小于任何一方，因此这两个不等式实际上是等号. 因此，我们有

$$\begin{aligned} & u_1(a_1, \sigma_2^*) = u_1(a_2, \sigma_2^*) \\ \iff & q \cdot 1 + (1-q) \cdot 0 = q \cdot 0 + (1-q) \cdot 1 \\ \iff & q = \frac{1}{2}. \end{aligned}$$

同理，我们有 $p = 1/2$. 因此，双方都选择 $(1/2, 1/2)^\mathsf{T}$ 是唯一的 Nash 均衡.

如此，我们证明了猜硬币游戏的唯一 Nash 均衡是双方都选择 $(1/2, 1/2)^\mathsf{T}$. □

在上面的例子中，我们实际上得到了一个非常重要的结论：Nash 均衡具有*无差别原理*，它说明，Nash 均衡中出现的那些行动一定都是取到了最大收益的行动. 这一原理可以被用来求解任意二人博弈的 Nash 均衡（见习题[lhy: 出一下]）.

定理 9.7 (无差别原理) *考虑一个正则形式博弈 $G = (I, \{A_i\}_{i\in I}, \{u_i\}_{i\in I})$，其中 I 是玩家集合，A_i 是玩家 i 的行动空间，u_i 是玩家 i 的收益函数. 如果 σ^* 是一个 Nash 均衡，那么对于任意玩家 i 和任意 a_i, a_i' 满足*

$$\sigma_i^*(a_i) > 0, \quad \sigma_i^*(a_i') > 0,$$

有

$$u_i(a_i, \sigma_{-i}^*) = u_i(a_i', \sigma_{-i}^*) = u_i(\sigma^*) = \max_{\sigma_i \in \Delta(A_i)} u_i(\sigma_i, \sigma_{-i}^*).$$

上面我们说明，有限正则形式博弈一定有 Nash 均衡，实际上，类似的不动点技术可以被用来证明更一般的均衡存在定理：

定理 9.8 (Debreu-Glicksberg-Fan 定理) *一个正则形式博弈满足如下所有条件，就会存在一个纯策略 Nash 均衡：*

- *对任意玩家 i，行动空间 A_i 是 Euclid 空间中的非空紧凸子集，*

- 对任意玩家 i，收益函数 u_i 连续，对第 i 个分量上凹.

定理 9.9 (Glicksberg 定理) 一个正则形式博弈满足如下所有条件，就会存在一个 *Nash* 均衡：

- 对任意玩家 i，行动空间 A_i 是度量空间中的非空紧子集，
- 对任意玩家 i，收益函数 u_i 连续.

尽管在数学上，期望效用理论导出了漂亮的结果，即 Nash 均衡的存在性，但是（期望）效用理论非常不符合实验的观察. 比如说，下面两个选项，大部分人会更倾向于选择第一个：

- 你得到 1000 元.
- 你有 1% 的概率得到 100000 元，99% 的概率得到 0 元.

然而，如果我们套用期望效用理论，两个选项的期望效用是一样的！

此外，从某种角度来看，价值也是相对的. 如果我们有 50% 的机会赢 100 块钱或 10 块钱，拿到 10 块钱时可能会感到失望. 但如果改成我们只能赢得 10 块钱或 1 块钱，那得到 10 块钱时我们就会感到更开心. 因此，即便效用理论是对的，在现实中我们完全不能知道真正的效用函数是什么，也无法在现实中真的用来建模.

那么，为什么我们还要研究（期望）效用理论，甚至整个经济学和博弈论的体系都是基于（期望）效用理论的？它之所以如此重要，就是因为它很好地平衡了“可以写理论结果”和“可以解释现实现象”之间的关系. 正如 Robert Aumann 所说，

> “另一种强调这一点的原因——即理论不应简单地被视为真或假——是为了避免过分抠字眼带来的问题……有人反对效用最大化的概念，认为人并不真的在最大化效用. 对此，有人提出了诸如‘满意准则’这样的替代方案……然而，效用最大化的有效性并不在于它能否精确描述个体行为，而在于它作为经济理论的基础假设，能够整合大量经济学理论……
>
> “像‘满意准则’这样的替代方案虽然看起来有吸引力，但几乎没有什么实际作用，它们很少带来有趣的结果. 在评价效用最大化时，我们不应问‘它是否合理？’，而是应该问‘它能整合什么？它能引导我们走向哪里？’”

在任何时候，作为一门语言，博弈论都应该具备 Aumann 所说的这种特性. 我们不能为了数学的方便而过分简化理论，但更不能过分拟合现实，变得毫无指导和应用价值.

§9.4 随机博弈（Markov 博弈）

本节我们讲讲述一种动态博弈和正则形式博弈的结合：随机博弈. 为了引入随机博弈，我们需要 Markov 链和 Markov 决策过程（MDP）相关的知识，更详细的讨论请参阅第二章. 一个 MDP 有如下的组成：

- 有限状态集合：$\mathcal{S} = \{s_1, s_2, \cdots, s_N\}$.
- 有限动作集合：$\mathcal{A}$.
- 每个状态具有自己的动作空间：$\mathcal{A}_s = \{a_{s,1}, a_{s,2}, \cdots, a_{s,N}\}, s \in S$.
- 每个动作空间有限：$|\mathcal{A}_{s,k}| = n_{s,k}, s \in S, k = 1, 2, \ldots, N$.
- 在状态 s_k, 若选择第 i 个动作 $a_{k,i}$ $(1 \leq i \leq n_k)$，则可以定义
  - 状态转移概率：$\mathcal{P}(s'|s, a)$，表示在状态 s 选择动作 a 后，转移到状态 s' 的概率.
  - 即时奖励：$R(s, a, s')$，表示在状态 s 选择动作 a 后，转移到状态 s' 时获得的即时奖励.
- 折扣因子 $\gamma \in [0, 1]$，用于计算远期收益.

MDP 的目标是找到一个策略 π，使得在该策略下可以获得最大的期望累积奖励.

随机博弈可以看做 MDP 的多人扩展.

定义 9.7 (随机博弈，Markov 博弈) **随机博弈**（又称为 **Markov 博弈**）有如下组成：

- N: 玩家的数量，$N = 1$ 退化为 MDP. 用 $1, \ldots, N$ 表示玩家的编号.
- $\mathcal{S}$: 状态的集合.
- $\mathcal{A}$: 玩家的行动集合. $\mathcal{A} = \mathcal{A}^1 \times \cdots \mathcal{A}^N$. 设 $\mathcal{A}_i(s)$ 表示第 i 个玩家在状态 s 的行动空间.
- $\mathcal{P} : \mathcal{S} \times \mathcal{A} \times \mathcal{S} \to [0, 1]$: 给定玩家的联合动作 $\boldsymbol{a} \in \mathcal{A}$, 从状态 $s \in \mathcal{S}$ 转移到 $s' \in \mathcal{S}$ 的概率 $\mathcal{P}(s'|s, \boldsymbol{a})$.
- $Q : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$: 在状态 s，当玩家的联合动作为 $\boldsymbol{a}$ 时，玩家 i 的奖励值 $Q_i(\boldsymbol{a}; s)$（有界）.
- $\gamma \in [0, 1]$ 表示折扣因子. □

有了博弈的框架，我们可以讨论随机博弈的过程.

- 首先，博弈从某一个状态 s^0 开始，$s^0 \in \mathcal{S}$.
- 在每个阶段 t，所有玩家同时选择自己的动作 a^t.
- 环境根据所有玩家的动作 $\boldsymbol{a}^t$ 和状态 s^t，进行如下操作：
  - 给予每个玩家对应的收益 $q(\boldsymbol{a}^t, s^t)$.
  - 转移到新的状态 $s^{t+1} \in \mathcal{S}$.

接下来，我们可以定义玩家的策略. 假设在阶段 T，所有玩家可以观察到所有历史动作 $\{\boldsymbol{a}^t\}_{t \le T}$. 和井字棋一样，我们可以定义每个玩家的策略 π——基于历史信息（状态、行动）到当前状态的行动的映射. 玩家在博弈的过程中，其实就是按照某个策略 π 进行行动的.

然而，每个玩家的策略可以依赖于历史信息，但这种策略通常很复杂，为了简化，我们考虑一些更简单的策略. 简化的关键在于，这是一个无穷轮的博弈，所以玩家需要有无穷大的记忆来存储历史信息.

如果我们让玩家只有固定大小的记忆，那么玩家的策略就只能依赖于有限的历史信息. 在最简化（健忘）的情况下，玩家的行动选择仅依赖于当前状态，而与之前的历史无关. 这种策略被称为平稳策略.

平稳策略的正式定义如下：

定义 9.8 (平稳策略) 对于玩家 i，**平稳策略** π_i 是一个映射，从当前状态 $s \in \mathcal{S}$ 到行动空间 $\mathcal{A}_p(s)$ 的概率分布.

因而，平稳策略可以表示为

$$\pi_i : \mathcal{S} \to \Delta(\mathcal{A}_i),$$

其中 $\Delta(\mathcal{A}_i)$ 表示行动空间 $\mathcal{A}_i$ 上的所有概率分布.

在每个状态 s 下，玩家 i 选择每个可能行动 $a \in \mathcal{A}_i(s)$ 的概率由 $\pi_i(s, a)$ 给出.

假设每个玩家 i 都采用平稳策略 π_i，那么整个策略组合 $\boldsymbol{\pi} = (\pi_1, \ldots, \pi_N)$ 也被称为**平稳策略组合**. □

平稳这一词在概率论中有明确的含义. 考虑一列随机变量 $X_1, X_2, \ldots$，如果对于任意 n 和 k，

$$(X_1, X_2, \ldots, X_n) \overset{d}{=} (X_{k+1}, X_{k+2}, \ldots, X_{k+n}),$$

这个等号的意思是两边的联合分布相同. 这时我们称 $X_1, X_2, \dots$ 是一个*平稳过程*.

实际上，平稳策略这一词正来源于此. 我们现在说明更强的结论，即平稳策略组合其实诱导了一个 Markov 链. 假设玩家采用平稳策略 $\boldsymbol{\pi}$，$s^0, s^1, \dots$ 是一个随机博弈的状态序列，那么容易证明 $\{s^t\}_{t=0}^{\infty}$ 是一个时齐 Markov 链（见习题[lhy: 出一下]），转移概率可以表示为

$$\mathcal{P}^{(\boldsymbol{\pi})}(s'|s) = \mathbb{E}_{\boldsymbol{a}\sim\boldsymbol{\pi}(s)}[\mathcal{P}(s'|s,\boldsymbol{a})].$$

因此，从这个视角来看，随机博弈就是在 Markov 链上进行的博弈，因此有时被称为 *Markov 博弈*.

由于 Markov 博弈兼具正则形式博弈和 MDP 的特点，所以我们接下来会将这两部分对应的概念和性质都引入到随机博弈中.

从博弈论的角度，我们要讨论玩家的理性. 首先，我们可以按照期望效用理论，扩展收益函数 $Q_i(\boldsymbol{a}; s)$ 为 $Q_i(\boldsymbol{\pi}; s)$：

$$Q_i(\boldsymbol{\pi}; s) = \mathbb{E}_{\boldsymbol{a}\sim\boldsymbol{\pi}(s)}[Q_i(\boldsymbol{a}; s)].$$

接下来，我们仿照 MDP，定义玩家在整个博弈中的收益：

定义 9.9 (价值函数) 对于一个随机博弈，玩家 i 的**价值函数** $V_i^{\boldsymbol{\pi}}(s)$ 定义为

$$V_i^{\boldsymbol{\pi}}(s) = \mathbb{E}\left[\sum_{t=0}^{\infty}\gamma^t Q_i(\boldsymbol{a}; s^t)\middle| s^0 = s, \boldsymbol{a}\sim\boldsymbol{\pi}(s^t), s^{t+1}\sim\mathcal{P}^{(\boldsymbol{\pi})}(\cdot|s^t)\right],$$

即从初始状态 s^0 开始，玩家 i 对每一期折现收益的期望. □

在随机博弈中，玩家的理性即是在给定其他玩家的策略 $\boldsymbol{\pi}_{-i}$ 的情况下，最大化其价值函数.

与 MDP 类似，价值函数也满足 Bellman 方程：

定理 9.10 (Bellman 方程)

$$V_i^{\boldsymbol{\pi}}(s) = Q_i(\boldsymbol{\pi}(s); s) + \gamma\sum_{s'\in\mathcal{S}}\mathcal{P}^{\boldsymbol{\pi}}(s'|s)V_i^{\boldsymbol{\pi}}(s').$$

证明见习题[lhy: 出一下].

有了理性的定义，我们可以自然地引入最优反应的概念：

定义 9.10 (Markov 最优反应) 对于一个随机博弈，给定其他玩家的平稳策略 $\boldsymbol{\pi}_{-i}$，玩家 i 的 **Markov 最优反应**是一个策略 $\boldsymbol{\pi}_i$，使得对于任意状态 s 和任意平稳策略 $\boldsymbol{\pi}_i'$，

$$V_i^{\boldsymbol{\pi}_i,\boldsymbol{\pi}_{-i}}(s) \geq V_i^{\boldsymbol{\pi}_i',\boldsymbol{\pi}_{-i}}(s).$$

最优反应对应的策略集合记为 $BR_i(\boldsymbol{\pi}_{-i})$. □

最后，我们可以用最优反应来定义均衡：

定义 9.11 (Markov 完美均衡，MPE) 在所有玩家的平稳策略组合中，每一个玩家的策略都是 Markov 最优反应，那么这个策略组合被称为 **Markov 完美均衡**（MPE）. □

同样，MPE 也可以被看作是最优反应的不动点. 类似 Nash 均衡的存在性定理，我们有如下的定理：

定理 9.11 *对于 N 个玩家、有限状态、有限动作空间的随机博弈，MPE 存在.*

下面我们介绍 Shapley 关于双人零和随机博弈情形的证明，这一证明基于 Banach 不动点定理（定理 8.2）. 对于一般的情况，可以类似 Nash 均衡存在性，利用 Brouwer 不动点定理证明，见习题[lhy: 出一下].

在双人零和的语境下，我们去掉收益函数 Q 的下标 i，玩家 2 的收益函数是 $-Q$. 注意，价值函数满足 Bellman 方程：

$$V^{\pi}(s) = Q(\boldsymbol{\pi}(s); s) + \gamma \sum_{s' \in \mathcal{S}} \mathcal{P}^{\pi}(s'|s) V^{\pi}(s').$$

我们的证明策略是用迭代的方法逼近这个方程的解 V. 为了利用 Banach 不动点定理，我们需要定义一个迭代算子，它是压缩映射，并且迭代收敛到 V.

Bellman 方程的左右是同一个 V，此时是不动点方程，因此，只要把右边看成迭代的过程，左边看成迭代的结果，我们就自然得到了一个迭代算子. 具体过程如下：

- 首先，我们选择一个任意的函数 $\alpha : \mathcal{S} \to \mathbb{R}$，称 α *为值函数*. 这个函数与 V 属于同样的空间，这是迭代的初始值 α_0.
- 对任意 $s \in \mathcal{S}$，定义矩阵 $R_s(\alpha)$ 为

  $$R_s(\alpha)(a_1, a_2) = Q(a_1, a_2; s) + \gamma \sum_{s' \in \mathcal{S}} \mathcal{P}(s'|a_1, a_2, s) \alpha(s').$$

  其中 $a_1 \in \mathcal{A}_1, a_2 \in \mathcal{A}_2$.

  如此，我们形成了一个矩阵零和博弈，玩家 1 的收益矩阵是 $R_s(\alpha)$，玩家 2 的收益矩阵是 $-R_s(\alpha)$. 注意，根据习题[lhy: 出一下]，玩家 1 在 Nash 均衡时候的收益是确定的一个值，记为 $\mathsf{val}(R_s(\alpha))$.
- 进行下一步迭代 $\alpha_k(s) = \mathsf{val}(R_s(\alpha_{k-1}))$.

我们中间之所以先定义了一个矩阵 $R_s(\alpha)$，是因为 Bellman 方程中 $\boldsymbol{\pi}(s)$ 是不知道的，因此 Q 是无法计算的. 根据 Q 的含义，这其实是一期的收益，所以我们可以用一个矩阵博弈来代替. 价值函数 V 满足

$$V(s) = \mathsf{val}(R_s(V)),$$

因此 V 是 $\mathsf{val}(R_s(\cdot))$ 的不动点，这说明我们的矩阵博弈选择是合理的. 接下来，只要证明 $\mathsf{val}(R_s(\cdot))$ 是一个压缩映射，因此根据 Banach 不动点定理，α_k 就会收敛到 V.

为方便，我们定义迭代算子 $(T\alpha)(s) = \mathsf{val}(R_s(\alpha))$. 我们需要有一个方法来衡量 $T\alpha$ 和 $T\alpha'$ 的差距，因此，我们给出如下引理：

引理 9.1 *对任意 $m \times n$ 的矩阵 B, C，成立：*

$$|\mathsf{val}(B) - \mathsf{val}(C)| \leq \max_{i,j} |B_{ij} - C_{ij}|.$$

证明. 设 (s_1, s_2) 为矩阵博弈 B 的 Nash 均衡，$(\bar{s}_1, \bar{s}_2)$ 为矩阵博弈 C 的 Nash 均衡. 于是由定义有：$s_1^\top B\bar{s}_2 \geq s_1^\top Bs_2$，且 $\bar{s}_1^\top C\bar{s}_2 \geq s_1^\top C\bar{s}_2$，因此

$$\begin{aligned}
\mathsf{val}(B) - \mathsf{val}(C) &= s_1^\top B\bar{s}_2 - s_1^\top C\bar{s}_2 \\
&\leq s_1^\top B\bar{s}_2 - s_1^\top Bs_2 \\
&\leq \max_{i,j} |B_{ij} - C_{ij}|.
\end{aligned}$$

根据 B 和 C 的对称性，引理得证. □

根据这个引理，我们可以证明 T 是一个压缩映射：

引理 9.2 *如果 $\gamma \in (0,1)$，那么 T 是一个压缩系数为 γ 的压缩映射.*

证明.

$$\begin{aligned}
\left\|T\alpha - T\alpha'\right\|_\infty &= \max_{s\in\mathcal{S}} |\mathsf{val}(R_s(\alpha)) - \mathsf{val}(R_s(\alpha'))| \\
&\leq \gamma \max_{s\in\mathcal{S}} \max_{a_1,a_2} |\sum_{s'\in\mathcal{S}} \mathcal{P}(s'|a_1,a_2,s)(\alpha(s') - \alpha'(s'))| \\
&\leq \gamma \max_{s'\in\mathcal{S}} |\alpha(s') - \alpha'(s')| \\
&= \gamma \left\|\alpha - \alpha'\right\|_\infty.
\end{aligned}$$

第一个不等式的成立使用了引理 9.1，第二个不等式的成立使用了 $\mathcal{P} \in [0,1]$ 的性质. □

对于有折扣的博弈，$\gamma \in (0,1)$，因此 T 是一个压缩映射，由 Banach 不动点定理可知，α_k 会收敛到 α^* 满足 $T\alpha^* = \alpha^*$，这就是我们要求的 V.

为了证明定理，我们还需要说明：

1. α^* 一定是均衡时候的玩家 1 的收益.
2. 存在一个策略组合达到均衡. 我们将要证明更强的结论：玩家 1 有一个策略，不论玩家 2 如何选择，玩家 1 的收益至少是 α^*.

要想证明这两点，我们需要理解 α 这一含义的直观. 选取 $\alpha_0(s) \equiv 0$，则 $R_s(\alpha_0) = Q(a_1, a_2; s)$ 是从 s 出发、由 Q 定义的矩阵博弈. 于是，$\alpha_1(s)$ 就是这一矩阵博弈玩家 1 的收益.

$$\alpha_1(s) = \mathsf{val}(R_s(\alpha_0)) = \mathsf{val}(Q(\cdot, \cdot; s)).$$

为了方便起见，我们总将“玩家 1 的收益”称之为“值”. 再看下一轮迭代，

$$R_s(\alpha_1)(a_1, a_2) = Q(a_1, a_2; s) + \gamma \sum_{s' \in \mathcal{S}} \mathcal{P}(s' | a_1, a_2, s) \alpha_1(s').$$

这个公式可以被这样理解，假想有一个被截断的两阶段随机博弈，玩家的策略如下：

- 玩家在第一阶段从状态 s 出发，行动 (a_1, a_2) 待定；
- 在第二阶段，对于每个可能到达的状态 $s' \in \mathcal{S}$，玩家采用矩阵博弈 $R_{s'}(\alpha_0)$ 的 Nash 均衡的行动.

注意，此时我们没有要求玩家策略是平稳的，也就是说，第二阶段和第一阶段的策略不需要是一样的. 根据定义，我们上面描述的这种策略达到了矩阵博弈 $R_s(\alpha_1)$ 的值，因此这个两阶段随机博弈的值不小于 $R_s(\alpha_1)$ 的值. 更一般地，被截断的 k 阶段随机博弈的值不小于 $\alpha_k(s)$.

接下来，我们证明 α^* 一定是均衡时候的值.

引理 9.3 *α^* 是一个 MPE 时候玩家 1 的收益.*

证明. 考虑从 s 出发的双人零和随机博弈，在前 k 局的博弈中，玩家 1 采用前 k 局截断随机博弈的最优策略，后续状态可选择任意动作. 由之前的分析可知，前 k 局截断的随机博弈的值不小于 $\alpha_k(s)$. 而对于之后的博弈，玩家 1 损失的累积收益最差不超过

$$(\gamma^k + \gamma^{k+1} + \dots) \sup |Q| = \frac{\gamma^k}{1-\gamma} \cdot \sup |Q|.$$

□

因此，令 k 趋于无穷，我们得到玩家 1 的收益至少是 $\alpha^*(s)$. 另一方面，根据同样的论证，玩家 2 也可以确保自己的收益至少是 $-\alpha^*(s)$. 由零和的定义，均衡时玩家 1 的收益必定是 $\alpha^*(s)$. □

最后，我们说明存在一个策略组合达到均衡. 如果让 R_s 作用在不动点 α^* 上，玩家 1 最大化 V 就是在选择 $R_s(\alpha^*)$ 的 Nash 均衡. 设 $R_s(\alpha^*)$ 的 Nash 均衡为 $(\pi_1(s), \pi_2(s))$，我们证明 (π_1, π_2) 是一个 MPE.

引理 9.4 *设 $R_s(\alpha^*)$ 的 Nash 均衡为 $(\pi_1(s), \pi_2(s))$，则 (π_1, π_2) 是一个 MPE.*

证明. 固定玩家 2 的一个任意策略 $\hat{\pi}_2$（不一定是平稳策略）. 玩家 1 在前 k 步选择 π_1，因为 π_1 是 $R_s(\alpha^*)$ 的 Nash 均衡，它也是前 k 步截断随机博弈的最优策略. 因此，根据之前的论证，无论玩家 2 选择何种行动，玩家 1 都能够至少拿到 α^* 的收益，即

$$\mathbb{E}\left[\sum_{t=0}^{k-1}\gamma^t Q(\pi_1(s^t), \hat{\pi}_2(s^t); s^t) + \gamma^k \alpha^*(s^k)\middle| s^0 = s\right] \geq \alpha^*(s).$$

整理得

$$\mathbb{E}\left[\sum_{t=0}^{k-1}\gamma^t Q(\pi_1(s^t), \hat{\pi}_2(s^t); s^t)\middle| s^0 = s\right] \geq \alpha^*(s) - \gamma^k \|\alpha^*\|_\infty.$$

因此，

$$V^{\pi_1, \hat{\pi}_2}(s) \geq \alpha^*(s) - \gamma^k \|\alpha^*\|_\infty - \frac{\gamma^k}{1-\gamma}\sup|Q|.$$

令 $k \to \infty$ 可得，上式右边趋于 $\alpha^*(s)$，因此我们为玩家 1 选择的策略达到了 α^* 的收益.

玩家 2 的 π_2 证明对称，因此 (π_1, π_2) 是一个 MPE. □

最后，我们指出随机博弈和人工智能的关系. 正如 MDP 是强化学习的基础，随机博弈是多智能体强化学习的基础. 在多智能体强化学习中，每个智能体都是一个独立的决策者，按照一个 MDP 来决策. 但是，由于环境中有多个智能体，每个智能体的奖励函数都会受到其他智能体的影响，这正是随机博弈的情形.

利用多智能体强化学习，我们可以训练出极其强大的人工智能，例如，DeepMind 在星际争霸 2 中训练出的 AlphaStar 就是一个例子. 通过人类数据与多智能体强化学习的自我训练，AlphaStar 在星际争霸 2 的比赛中击败了世界冠军. 它展示了博弈论如何为人工智能的发展提供语言和训练方法.

§9.5 习题

[lhy: TODO]

§9.6 章末注记

[lhy: TODO]

第五部分

认知逻辑

第六部分

附录：预备知识

参考文献

[Bre57] Leo Breiman. The Individual Ergodic Theorem of Information Theory. *The Annals of Mathematical Statistics*, 28(3):809–811, 1957.

[CT12] Thomas M. Cover and Joy A. Thomas. *Elements of Information Theory*. John Wiley & Sons, 2012.

[Huf52] David A. Huffman. A Method for the Construction of Minimum-Redundancy Codes. *Proceedings of the IRE*, 40(9):1098–1101, September 1952.

[Inf] Information | Etymology, origin and meaning of information by etymonline. https://www.etymonline.com/word/information.

[Jay02] Edwin T. Jaynes. *Probability Theory: The Logic of Science*. Cambridge University Press, 2002.

[KL51] S. Kullback and R. A. Leibler. On Information and Sufficiency. *The Annals of Mathematical Statistics*, 22(1):79–86, 1951.

[LLG+19] Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Ves Stoyanov, and Luke Zettlemoyer. BART: Denoising Sequence-to-Sequence Pre-training for Natural Language Generation, Translation, and Comprehension, October 2019.

[McM53] Brockway McMillan. The Basic Theorems of Information Theory. *The Annals of Mathematical Statistics*, 24(2):196–219, June 1953.

[RHW86] D. E. Rumelhart, G. E. Hinton, and R. J. Williams. Learning internal representations by error propagation. In *Parallel Distributed Processing: Explorations in the Microstructure of Cognition, Vol. 1: Foundations*, pages 318–362. MIT Press, Cambridge, MA, USA, January 1986.

[Rob49] Robert M. Fano. *The Transmission of Information*. March 1949.

[Sha48] C. E. Shannon. A mathematical theory of communication. *The Bell System Technical Journal*, 27(3):379–423, July 1948.

[Shi96] A. N. Shiryaev. *Probability*, volume 95 of *Graduate Texts in Mathematics*. Springer, New York, NY, 1996.

[Tin62] Hu Kuo Ting. On the Amount of Information. *Theory of Probability & Its Applications*, 7(4):439–447, January 1962.

[Uff22] Jos Uffink. Boltzmann's Work in Statistical Physics. In Edward N. Zalta, editor, *The Stanford Encyclopedia of Philosophy*. Metaphysics Research Lab, Stanford University, summer 2022 edition, 2022.

[李 10] 李贤平. 概率论基础. 高等教育出版社, 2010.

索引