# 标题 **title**

作者 *author*

2024 年 9 月 9 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 $3 \times 3$ 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.

输赢博弈有多种分类方式，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息与非完全信息确实极其困难的，我们这里给一种方法。

| 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.

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.

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

[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.

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.

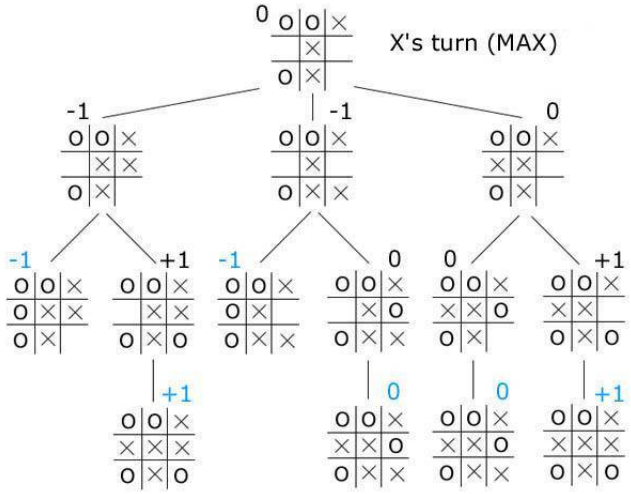
我们在本部分主要关注最简单的一种博弈，即*完全信息确定性回合制博弈*。这样的博弈可以用博弈树表示出来，例如，井字棋的博弈树可以画作图 9.2.

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.

**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.




[lhy: 重画]

图 9.2: 井字棋的博弈树

**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，因此完全理性的玩家在博弈中必然平局.

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 $19 \times 19$，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）.

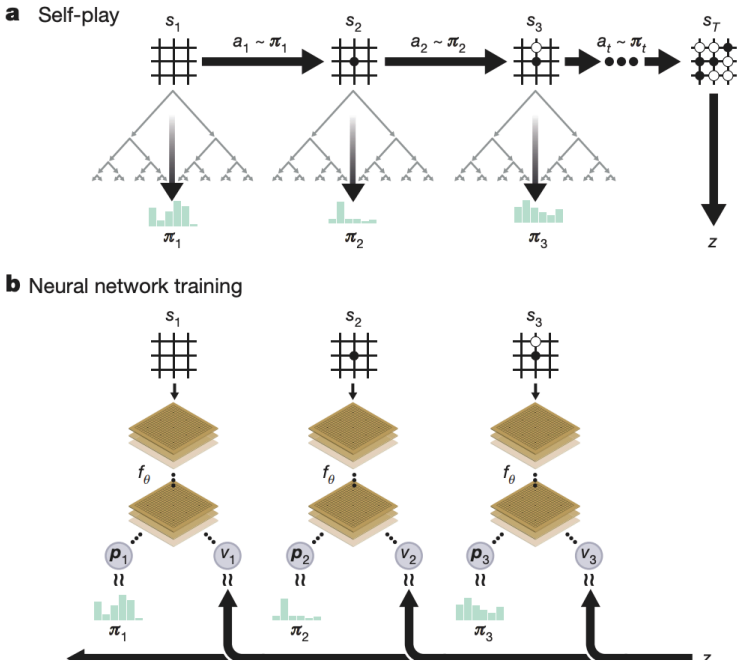
最后，当博弈结束时，Zero 会获得一定的收益，我们假定 Zero 赢的时候收益为 $+1$，输的时候收益为 $-1$.

对于人类来说，我们的任务是让 Zero 的策略 $\pi$ 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 $f_\theta$，执行一次 MCTS $\alpha_\theta$（具体过程见后面）. 下法根据 MCTS 计算的搜索概率 $\pi_t$ 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 $\theta$，以最大化策略 $p_t$ 和搜索概率 $\pi_t$ 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）：

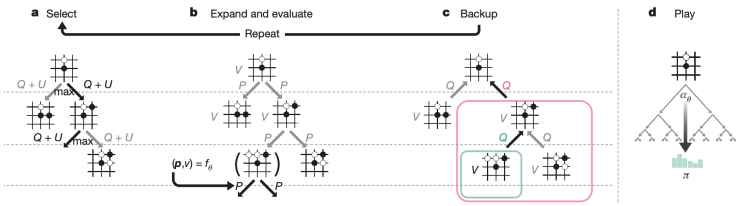
- 从根节点开始，状态 $s$ 固定，选择具有最大的 $Q(s, a) + U(s, a)$ 的分支。
  - $U$ 是上限置信度，$U(s, a) \propto P(s, a)/(1 + N(s, a))$.
  - $Q + U$ 是置信区间的上界，称为 UCB 公式。

  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 $\pi$，$\pi(a)$ 正比于 $N(s, a)^{1/\tau}$，$\tau$ 是一个参数，控制着温度. 温度反映了 $\pi$ 允许的随机性程度，当 $\tau$ 趋于正无穷的时候，$\pi$ 趋于均匀分布，当 $\tau$ 趋于零的时候，$\pi$ 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 随机博弈（Markov 博弈）

现在我们考虑无穷博弈中最简单的一种：**随机博弈**，也叫 **Markov 博弈**. 在随机博弈中，玩家的行动是随机的，但是玩家的行动空间是有限的. 为了简化问题，我们考虑两人随机博弈，即两个玩家轮流进行随机行动. 其相关概念如下：

- 有限局面：$C = \{s_1, s_2, \cdots, s_N\}$.
- 有限策略：$\mathcal{A} = \mathcal{A}_1 \times \mathcal{A}_2$.
  - 每个局面具有自己的行动空间：$\mathcal{A}_p = \{\mathcal{A}_{p,1}, \mathcal{A}_{p,2}, \cdots, \mathcal{A}_{p,N}\}, (p = 1, 2)$.
  - 每个行动空间有限：$|\mathcal{A}_{p,k}| = n_{p,k}, (p = 1, 2; k = 1, 2, \cdots, N)$.
- 在局面 $s_k$，若玩家 1 选择第 $i$ 个行动 $a_{1,k,i} (1 \leq i \leq n_{1,k})$，玩家 2 选择第 $j$ 个行动 $a_{2,k,j} (1 \leq j \leq n_{2,k})$，则
  - 局面之间有转移概率：
    - 该博弈以概率 $s_{ij}^k$ 停止.
    - 该博弈以概率 $p_{ij}^{kl}$ 转移到状态 $s_l$.
  - 收益：玩家 1 收获 $Q(a_{1,k,i}, a_{2,k,j}; s_k)$，玩家 2 收获 $-Q(a_{1,k,i}, a_{2,k,j}; s_k)$. 假设 $Q$ 有界.

随机博弈的过程如下. 首先，博弈从某一个局面状态 $s^0$ 开始，$s^0 \in C$. 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 $s^t$，给予每个玩家对应的收益 $q(\boldsymbol{a}^t, s^t)$，并转移到新的状态 $s^{t+1} \in C$.

假设在阶段 $T$，所有玩家可以观察到所有历史动作 $\{\boldsymbol{a}^t\}_{t \leq T}$. 和一般的动态博弈一样，我们可以定义每个玩家的策略 $\pi$——基于历史信息（状态、行动）到当前局面的行动的映射. 玩家在博弈的过程中，其实就是按照某个策略 $\pi$ 进行行动的. 求解一个博弈也是求解最优策略 $\pi_*$. 下面我们定义什么是“最优”.

依赖历史信息的策略 $\pi$ 一般很复杂. 由于收益只与当前局面、当前玩家的行动有关，我们可以缩小策略空间. 考虑第 $p$ 个玩家 $(p=1,2)$，定义*平稳策略*为 $N$ 个概率分布：$\vec{\pi}_p=(\pi_p^1,\pi_p^2,\cdots,\pi_p^N)$，分别对应 $N$ 个状态；每个概率分布 $\pi_p^k=(\pi_{p,1}^k,\pi_{p,2}^k,\cdots,\pi_{p,n_{p,k}}^k)$，分别对应 $|n_{p,k}|$ 个行动. 使用平稳策略时，无论博弈的历史轨迹如何，玩家 $p$ 在状态 $s_k$ 采取行动 $a_{p,k,i}$ 的概率为 $\pi_{p,i}^k$.

假设两个玩家分别用平稳策略 $\pi_1,\pi_2$ 进行博弈，则从局面 $s^0$ 开始的随机博弈中，第 $p$ 个玩家（$p=1,2$）的*远期收益*：

$$\Pi_p(\pi_1,\pi_2;s^0)=\mathbb{E}\left[\sum_{t=0}^{\infty}\gamma^t(-1)^{p-1}Q(\pi_1(s^t),\pi_2(s^t);s^t)\right].$$

以上定义容易扩展为一般随机博弈（多人、非零和）. 随机博弈可以看做 MDP 的多人扩展 $(N,C,\mathcal{A},\mathcal{P},\mathcal{Q},\gamma)$：

- $N$: 玩家的数量，$N=1$ 退化为 MDP.
- $C$: 局面的集合.
- $\mathcal{A}$: 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.
- $\mathcal{P}:C\times\mathcal{A}\times C\to[0,1]$: 给定玩家的联合动作 $\boldsymbol{a}\in\mathcal{A}$, 局面从状态 $s\in C$ 转移到 $s'\in C$ 的概率 $P(s'|s,\boldsymbol{a})$.
- $\mathcal{Q}:C\times\mathcal{A}\to\mathbb{R}$: 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）.
- $\gamma\in[0,1]$ 表示折扣系数，用于计算远期收益.

*Markov 完美均衡*（MPE）是一种解概念. 求解博弈的过程中，我们限制所有玩家使用平稳策略. 此时，面对对手，玩家的最优策略被称为 *Markov 最优反应*：对每个状态 $s$，给定其他玩家的平稳策略 $\pi_{-i}$，玩家 $i$ 的行动 $a_i\in\mathcal{A}_i(s)$ 最大化它的远期收益 $\Pi_i(s;\pi_{-i})$：

$$\Pi_i(s;\pi_{-i})=\mathbb{E}\left[Q_i(a_i,\pi_{-i}(s);s)+\gamma\sum_{s'\in C}P(s'|a_i,\pi_{-i}(s),s)\Pi_i(s';\pi_{-i})\right].$$

MPE 被定义为：所有玩家的平稳策略组合，其中每个玩家的行动都是 Markov 最优反应.

我们可以类比 MDP 中求解最优价值的 Bellman 方程（动态规划）的形式. 当假设其他玩家都使用平稳策略，对每个状态 $s$，存在一个价值函数 $V_i(s;\pi_{-i})$ 取得玩家 $i$ 从 $s$ 出发的最高远期收益，满足：

$$V_i(s;\pi_{-i}) = \max_{a_i\in\mathcal{A}_i(s)} \mathbb{E}[Q_i(a_i,\pi_{-i}(s);s) \\ + \gamma \sum_{s'\in C} P(s'|a_i,\pi_{-i}(s),s)V_i(s';\pi_{-i})].$$

**定理 9.3** *对于 $N$ 个玩家、有限局面状态、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. 对于一般的情况，我们留做习题.

首先介绍一下*矩阵博弈*的概念. 假设 $P$ 是一个 $m\times n$ 的矩阵，玩家 1 有 $m$ 种动作 (动作集合 $\mathcal{A}_1$)，玩家 2 有 $n$ 种动作 (动作集合 $\mathcal{A}_2$)，元素 $P_{ij}$ 表示双方采取动作 $(i,j)$ 时玩家 1 的收益，玩家 2 的收益为 $-P_{ij}$.

回忆在不动点课程中的 minimax 定理**??**：

$$\mathsf{val}(P) = \max_{s_1\in\Delta(\mathcal{A}_1)}\min_{s_2\in\Delta(\mathcal{A}_2)} s_1^\top P s_2 = \min_{s_2\in\Delta(\mathcal{A}_2)}\max_{s_1\in\Delta(\mathcal{A}_1)} s_1^\top P s_2.$$

$s_i$ 是玩家 $i$ 的混合策略，$\Delta(\mathcal{A}_i)$ 表示玩家 $i$ 所有混合策略的集合. $\mathsf{val}(P)$ 为矩阵 $P$ 定义的*矩阵博弈的值*. 在任意 Nash 均衡中，玩家 1 的期望收益即为 $\mathsf{val}(P)$. 下面我们将看到：双人零和随机博弈也存在值（即可定义均衡收益）.

首先证明一个引理：

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

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

**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：$s_1^\top B\bar{s}_2 \ge s_1^\top B s_2$，且 $\bar{s}_1^\top C\bar{s}_2 \ge s_1^\top C\bar{s}_2$，因此

$$s_1^\top B s_2 - \bar{s}_1^\top C\bar{s}_2 \le s_1^\top B\bar{s}_2 - s_1^\top C\bar{s}_2 \le \max_{i,j}|B_{ij}-C_{ij}|.$$

□

下面，我们将矩阵博弈的概念迁移到随机博弈. 在双人零和的语境下，我们去掉收益函数 $Q$ 的下标 $i$. 定义*值迭代*为以下过程：

- 首先，我们选择一个任意的函数 $\alpha: C\to\mathbb{R}$，其中 C 是局面的状态空间，称 $\alpha$ 为*值函数*（value function）.

- 对任意 $s \in C$，定义矩阵 $R_s(\alpha)$ 为

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

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

- 值函数从 $\alpha_0$ 开始迭代，记 $\alpha_k(s) = \mathsf{val}(R_s(\alpha_{k-1}))$.

如何理解 $\alpha_k(s)$? 假设选取 $\alpha_0(s) \equiv 0$，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. $\alpha_1(s) = \mathsf{val}(R_s(\alpha_0)) = \mathsf{val}(Q(\cdot, \cdot; s))$.

再看 $R_s(\alpha_1)(a_1, a_2) = Q(a_1, a_2; s) + \gamma \sum_{s' \in C} P(s'|a_1, a_2, s)\alpha_1(s')$. 我们可以假想有一个被截断的两阶段随机博弈：

- 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，对于每个可能的状态 $s' \in C$，玩家采用矩阵博弈 $R_{s'}(\alpha_0)$ 的 Nash 均衡的行动.

博弈在第二阶段终止，远期收益累积的折扣部分为 $\gamma \sum_{s' \in C} P(s'|a_1, a_2, s)\mathsf{val}(R_{s'}(\alpha_0))$. 从而，矩阵博弈 $R_s(\alpha_1)$ 的值也是这个两阶段随机博弈的值. 更一般地，$\alpha_k(s)$ 是一个被截断的 $k$ 阶段随机博弈的值.

为方便，我们定义迭代算子 $(T\alpha)(s) = \mathsf{val}(R_s(\alpha))$.

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

第一个不等式的成立使用了矩阵博弈值的不等式. 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，$\alpha_k \to \alpha^*$ 满足 $T\alpha^* = \alpha^*$.

考虑任意一个从 $s$ 出发的双人零和随机博弈，在前 $k$ 局的博弈中，玩家 1 采用最优策略，后续局面可选择任意动作. 由之前的分析可知，前 $k$ 局构成的截断随机博弈远期收益为 $\alpha_k(s)$. 而对于之后的博弈，玩家 1 损失的累积收益最差不超过 $\gamma^k/(1-\gamma) \cdot \sup |Q|$. 因此，当 $k \to \infty$ 时，玩家 1 的收益至少是 $\alpha^*(s)$. 注意，这个下界是无视玩家 2 的行动得出来的.

另一方面，玩家 2 也可以确保自己的收益至少是 $-\alpha^*(s)$. 由零和，因此均衡时玩家 1 的收益必定是 $\alpha^*(s)$. 因此双人零和随机博弈的均衡收益（值）为 $\alpha^*(s), s \in C$.

这样，我们就证明了 MPE 的存在性.

以上我们只说明了双人零和随机博弈的值存在，还没有指明最优策略如何取得. 类比矩阵博弈，我们有：

**定理 9.4** *$R_s(\alpha^*)$ 定义的矩阵博弈的最优策略 $(\pi_1, \pi_2)$ 是随机博弈的 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）. 首先考虑一个 $k$ 阶段截断博弈，我们定义 $\alpha_0 = \alpha^*$，可理解为，原博弈前 $k$ 步动作待定，后面使用策略取得 $\alpha^*$ 的远期收益.

在这个博弈中，玩家 1 可以无视玩家 2 的策略，确保至少取得 $\alpha^*$ 的远期收益（已证明），因此若采用 $\pi_1$ 也能取得:

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

化简可得

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

因此

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

同样地，令 $k \to \infty$ 可得，上式 R.H.S. 趋于 $\alpha^*(s)$. 对于玩家 2 的 $\pi_2$ 证明对称，因此 $(s_1, s_2)$ 是一个 MPE. □

随机博弈在机器学习中对应着*多智能体强化学习*（MARL），正如 Markov 决策过程对应着（单智能体）强化学习. MARL 是多智能体系统（MAS）研究领域中的一个重要分支，它将强化学习技术、博弈论等应用到多智能体系统，使得多个智能体能在更高维且动态的真实场景中通过交互和决策完成更错综复杂的任务. 解 MDP 的过程是根据环境信息，优化决策，最大化远期收益，这是单智能体强化学习. 解随机博弈的过程，是在 MDP 的基础上引入多玩家，玩家有自己的效用函数，最大化自己的远期收益，这是多智能体强化学习.

本章我们讨论静态博弈的基本概念和分析方法，并以此为基础，讨论博弈中认知相关的问题.

## §9.4 正则形式博弈

动态博弈通常被建模为*扩展形式博弈*. 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. 正则博弈通常要求信息是完全的. 这种博弈的过程与

时间无关，属于静态博弈.

一个正则形式博弈有如下构成要素

- 玩家集合：$I$，我们总是假设这是一个有限集合.
- 玩家的行动集（纯策略集）：$A_i$，$i \in I$.
- 玩家的收益：$u_i : \prod_j A_j \to \mathbb{R}$.
- 完全信息：以上内容是所有玩家的共同知识.

所有人的策略拼在一起，即 $s = (s_i)_{i \in I}$，构成博弈的策略组合. 有以下特殊的正则博弈：

- 当 $A_i$ 有限，我们称之为矩阵博弈.
- 当 $A_i$ 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，当所有策略组合，收益和都是常数时，解概念的分析可以保持一致，我们也可以按零和处理.

如何定义正则博弈的均衡？首先要明确均衡的概念. 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. 因为只能行动一次，所以所谓均衡，指的是没有人对自己的决策感到后悔的状态，没有人可以通过改变自己现在的策略来获得更多的收益. 因此我们有如下定义：

**定义 9.1 (Nash 均衡)**（纯策略）**Nash 均衡**指的是策略组合 $s$，满足

$$\forall i \in I \, \forall a_i \in A_i : u_i(s_i, s_{-i}) \geq u_i(a_i, s_{-i}).$$

我们也可以用不动点来理解 Nash 均衡. 首先定义最优反应：给定对手的策略 $s_{-i}$，玩家 $i$ 选择的最大化自己收益的策略 $s_i$. 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.

**例 9.2 (囚徒困境)** 考虑一个经典的非合作博弈，囚徒困境. 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. 这个博弈有唯一的 Nash 均衡：每个人都认罪. 思考：打破 Nash 均衡的假设，有没有可能得到更好的结果？ □

然而，纯策略 Nash 均衡并不一定存在. 考虑如下的输赢（零和）博弈：猜硬币游戏. 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，这个博弈没有纯策略 Nash 均衡. 更一般地，二人正则输赢博弈中纯策略 Nash 均衡往往不存在. 我们有如下定理：

**定理 9.5** *设 $G = (I, \{A_i\}_{i\in I}, \{u_i\}_{i\in I})$ 是一个二人正则输赢博弈，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.*

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. 我们可以利用这一事实去理解生成对抗网络模型的不稳定性.

### §9.4.1 生成对抗网络

*生成对抗网络*（GAN）有两个子模型组成，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，二判别模型的任务是识别给定的数据是真实的还是伪造的.

假设真实数据的分布为 $F_{data}$. 生成模型为 $G(x;\theta_g)$，参数为 $\theta_g$，输入向量 $x$，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 $\theta_d$，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. 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 $V$.

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 9.6** *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 9.7 (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.*

我们刚刚的分析过于理想化，需要考虑一些问题. 首先，神经网络的大小是有限的，因此 $G$ 不能选择任何 $p_g$. 因此，我们刚刚找到的 Nash 均衡可能不存在. 其次，$p_{data}$ 是一个未知的量，我们只有一些样本. 因此，$G$ 和 $D$ 都需要一个算法来找到它们的最优策略. 这就是训练 GAN 的过程.

我们接下来给出一种更符合实际的均衡概念.

*局部 Nash 均衡* $(G^*, D^*)$ 是指在 $G^*$ 和 $D^*$ 的一个邻域内 $(G^*, D^*)$ 形成了一个 Nash 均衡. *稳定局部 Nash 均衡* $(G^*, D^*)$ 是指 $(G^*, D^*)$ 是一个局部 Nash 均衡，并且在 $(G^*, D^*)$ 的一个邻域内，对任意 $(G, D)$ 都有 $V(G, D^*) \geq V(G, D)$ 和 $V(G^*, D) \leq V(G, D)$. GAN 的训练实际上就是在寻找稳定局部 Nash 均衡的过程.

稳定局部 Nash 均衡表明了，即便对手的策略具有（很小的）不确定性，玩家的策略依然是最优反应. 在训练过程中，这样的不确定性很可能出现，源自精度或者误差. 因此，稳定局部 Nash 均衡是一个更有可能被找到的解，不稳定局部 Nash 均衡则很容易偏离. 然而，我们刚刚在理想条件下找到的 Nash 均衡其实也是不稳定的. 实际上，GAN 的训练是一个非常不稳定的过程. 我们有如下结果：

**定理 9.8** *设 GAN 博弈的收益函数 $V$ 是解析的，$(0,0)$ 是稳定局部 Nash 均衡，在 $(0,0)$ 的一个邻域内，$V(G,D) = C + V^2 f(V) + D^2 g(D) + V^2 D^2 h(G,D)$，其中 $f, g, h$ 都是解析函数，满足 $f(0), g(0) \geq 0$，$C$ 是常数.*

$V$ 要具备这种形式才可能有稳定局部 Nash 均衡. 然而一般的神经网络并不能具备这样的形式，所以很多情况下根本不存在稳定局部 Nash 均衡！

### §9.4.2 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**. 混合策略就是建立在纯策略空间 $S$ 上的一个概率分布. 混合策略空间记为 $\Delta(S)$. 当 $S$ 有 $n$ 个元素（有限），$\Delta(S)$ 可以被表示为标准的 $n$-单纯形：

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

那么，有了混合策略，玩家的决策思考过程是怎么样的？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 $\sigma_i \in \Delta(A_i)$. $\sigma$ 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：

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

**定义 9.2 (Nash 均衡)** 对于一个博弈 $G = (I, \{A_i\}_{i \in I}, \{u_i\}_{i \in I})$，混合策略 Nash 均衡 $\sigma$ 满足对于任意玩家 $i$ 和任意 $\sigma_i' \in \Delta(A_i)$，都有

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

Nash 著名的定理是：

**定理 9.9 (Nash 均衡存在性定理)** *对于任意有限正则形式博弈，都存在一个混合策略 Nash 均衡.*

我们来看一个例子.

**例 9.3** 继续考虑猜硬币游戏，收益矩阵为

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

容易证明，唯一的均衡是两个玩家都选择 $(1/2, 1/2)$. □

尽管在数学上，混合策略是导出了漂亮的结果，但是混合策略并不是一个非常合理的概念. 如何理解混合策略？我们将在后面通过似然、知识论等方式来解释混合策略.

## §9.5 习题

[lhy: 仔细讨论这部分关于输赢博弈的讨论

- 什么叫完全理性的玩家？
- 表示完整的策略需要多少比特？
- 是否有高效的算法计算必胜策略？
- 如果博弈多方不是完全对抗的（即零和），那么是否还有必胜策略？是否有其他合理的解概念？

]

我们再看一个有趣的例子：*对话博弈*. 在对话博弈中，我们可以把对命题 $\phi$ 的辩论过程形式化为一个博弈. 博弈中有两个玩家，一个需要证明 $\phi$ 是真的，被称为*正方* $P$，一个需要说明 $P$ 的论据是有矛盾的，它被称为*反方* $O$. 两个玩家可以对命题的某个部分发起*质疑*，或者对某个质疑作出*辩护*. 当正方成功辩护了所有的质疑，而反方已经无法再发出新的质疑，正方获胜；否则反方获胜.

我们以命题逻辑为例，考虑合取和蕴含. 感叹号 ! 表示陈述，问号 ? 表示询问. 当一个玩家 $X$（$P$ 或者 $O$）说了一个合取式，另一个玩家 $Y$ 如果想质疑，需要询问合取中的某个部分（$L^\wedge$ 或 $R^\wedge$）. $X$ 需要陈述该部分对应的命题作为辩护. 这一规则可以总结为表 9.2.

| 陈述 | 质疑 | 辩护 |
|---|---|---|
| $X!\varphi \wedge \psi$ | $Y?L^\wedge$ 或 $Y?R^\wedge$ | 对应地，$X!\varphi$ 或 $X!\psi$ |

表 9.2: 对话博弈的规则：合取

当一个玩家 $X$（$P$ 或者 $O$）说了一个蕴含式，另一个玩家 $Y$ 如果想质疑，需要陈述前提. $X$ 则需要陈述结论作为辩护. 这一规则可以总结为表 9.3.

| 陈述 | 质疑 | 辩护 |
|---|---|---|
| $X!\varphi \to \psi$ | $Y!\varphi$ | $X!\psi$ |

表 9.3: 对话博弈的规则：蕴含

我们考虑一个具体的例子 $(p \wedge q) \to p$. 用一个表格来表示辩论的过程，这个表格有两列，分别表示玩家 $O$ 和 $P$. 每个玩家分别有三列，$A$ 表示当前操作是第几步（两个玩

家统一计数），$B$ 表示当前操作质疑的是哪一步，中间一列表示当前的操作（陈述或者询问）.

| $O$ | | | $P$ | | |
|---|---|---|---|---|---|
| $A$ | | $B$ | $B$ | | $A$ |

玩家 $P$ 陈述要辩论的命题.

| $O$ | | | | $P$ | |
|---|---|---|---|---|---|
| | | | | $!p \wedge q \to p$ | 0 |

玩家 $O$ 质疑这一陈述.

| $O$ | | | | $P$ | |
|---|---|---|---|---|---|
| | | | | $!p \wedge q \to p$ | 0 |
| 1 | $!p \wedge q$ | (0) | | | |

现在又一次轮到了玩家 $P$，他可以选择为蕴含式辩护，或者质疑这个合取式. 我们假设他这次选择质疑合取式，那么他需要询问左边或者右边，假设他询问了左边：

| $O$ | | | $P$ | | |
|---|---|---|---|---|---|
| | | | | $!p \wedge q \to p$ | 0 |
| 1 | $!p \wedge q$ | (0) | | | |
| | | | (1) | $?L^{\wedge}$ | 2 |

现在轮到了玩家 $O$. 他已经没有别的可以进行的操作了，只能对操作 2 进行辩护.

| $O$ | | | $P$ | | |
|---|---|---|---|---|---|
| | | | | $!p \wedge q \to p$ | 0 |
| 1 | $!p \wedge q$ | (0) | | | |
| 3 | $!p$ | | (1) | $?L^{\wedge}$ | 2 |

现在轮到了玩家 $P$. 他已经没有别的可以进行的操作了，只能对操作 1 进行辩护. 因为玩家 $O$ 已经陈述了 $p$，所以他可以用这个陈述来辩护.

| $O$ | | | $P$ | | |
|---|---|---|---|---|---|
| | | | | $!p \wedge q \to p$ | 0 |
| 1 | $!p \wedge q$ | (0) | | $!p$ | 4 |
| 3 | $!p$ | | (1) | $?L^{\wedge}$ | 2 |

现在轮到了玩家 $O$. 他已经不能操作了（没有可以质疑的，也没有可以辩护的），所以玩家 $P$ 获胜.

| $O$ | | | $P$ | | |
|---|---|---|---|---|---|
| | | | | $!p \wedge q \to p$ | 0 |
| 1 | $!p \wedge q$ | (0) | | $!p$ | 4 |
| 3 | $!p$ | | (1) | $?L^{\wedge}$ | 2 |

我们还可以定义否定 $\neg p$ 相关的辩论规则，留作练习.

根据 Zermelo 定理，对话博弈一定有一人有必胜策略，我们还有更精细的定理：

**定理 9.10** *考虑对命题 $\phi$ 的对话博弈，$\phi$ 是重言式当且仅当正方玩家 $P$ 有必胜策略.*

证明只需要考虑对 $\phi$ 做归纳法. 实际上，我们如此规定对话博弈的规则，就是为了保证这一定理成立. 我们可以将一个命题真值的判定问题转化为玩家 $P$ 博弈必胜策略的存在性问题，这对于一阶逻辑来说非常有用.

## §9.6 章末注记

# 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