重庆邮电大学

计算机科学与技术学院

人工智能原理

博弈论

博弈论的诞生：中国古代博弈思想

- 子曰：饱食终日，无所用心，难矣哉！不有博弈者乎？为之，犹贤乎已。

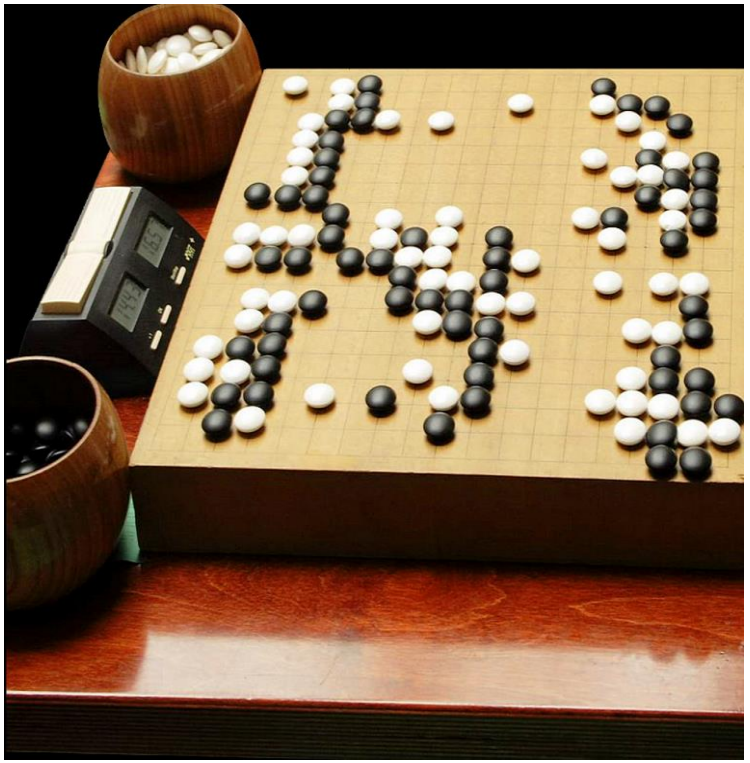
——《论语•阳货》

- 朱熹集注曰：“博，局戏；弈，围棋也。”；颜师古注：“博，六博；弈，围碁也。”
- 古语博弈所指下围棋，围棋之道蕴含古人谋划策略的智慧。

- 略观围棋，法于用兵，怯者无功，贪者先亡。

——《围棋赋》

- 《孙子兵法》等讲述兵书战法的古代典籍更是凸显了古人对策略的重视。



博弈论的诞生：田忌赛马

- ……齐将田忌善而客待之。忌数与齐诸公子驰逐重射。孙子见其马足不甚相远，马有上、中、下辈。于是孙子谓田忌曰：“君弟重射，臣能令君胜。”田忌信然之，与王及诸公子逐射千金。及临质，孙子曰：“今以君之下驷与彼上驷，取君上驷与彼中驷，取君中驷与彼下驷。”既驰三辈毕，而田忌一不胜而再胜，卒得王千金。于是忌进孙子于威王。威王问兵法，遂以为师。

——《史记•孙子吴起列传》

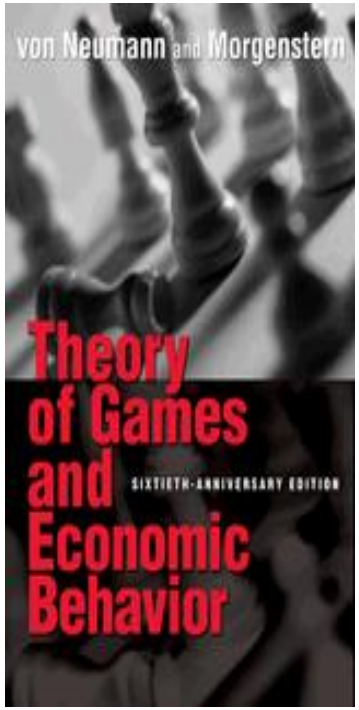
对局	齐王马	田忌马	结果	
1	A+	A-	齐王胜	3:0
2	B+	B-	齐王胜	
3	C+	C-	齐王胜	

对局	齐王马	田忌马	结果	
1	A+	C-	齐王胜	
2	B+	A-	田忌胜	1:2
3	C+	B-	田忌胜	

以己之长 攻彼之短

博弈论的诞生：现代博弈论的建立

- 博弈论（game theory），又称对策论。
- 博弈行为：带有相互竞争性质的主体，为了达到各自目标和利益，采取的带有对抗性质的行为。
- 博弈论主要研究博弈行为中最优的对抗策略及其稳定局势，协助人们在一定规则范围内寻求最合理的行为方式。
- 1944年冯·诺伊曼与奥斯卡·摩根斯特恩合著《博弈论与经济行为》，以数学形式来阐述博弈论及其应用，标志着现代系统博弈理论的初步形成，冯·诺伊曼被称为现代博弈论之父。



John von Neumann(1903-1957), Oskar Morgenstern(1902-1977), *Theory of Games and Economic Behavior*, Princeton University Press, 1944

博弈论的相关概念：博弈的要素

- 参与者或玩家（player）：参与博弈的决策主体
- 策略（strategy）：参与者可以采取的行动方案，是一整套在采取行动之前就已经准备好的完整方案。
  - 某个参与者可采纳策略的全体组合形成了策略集（strategy set）
  - 所有参与者各自采取行动后形成的状态被称为局势（outcome）
  - 如果参与者可以通过一定概率分布来选择若干个不同的策略，这样的策略称为混合策略（mixed strategy）。 若参与者每次行动都选择某个确定的策略，这样的策略称为纯策略（pure strategy）
- 收益（payoff）：各个参与者在不同局势下得到的利益
  - 混合策略意义下的收益应为期望收益（expected payoff）
- 规则（rule）：对参与者行动的先后顺序、参与者获得信息多少等内容的规定

博弈论的相关概念：研究范式

建模者对参与者（player）规定可采取的策略集(strategy sets)和取得的收益，观察当参与者选择若干策略以最大化其收益时会产生什么结果

两害相权取其轻，两利相权取其重

博弈论的相关概念：囚徒困境（prisoner's dilemma）

- 1950年，兰德公司的梅里尔·弗勒德和梅尔文·德雷希尔拟定了相关困境理论，后来美国普林斯顿大学数学家阿尔伯特·塔克以“囚徒方式”阐述：
  - 警方逮捕了共同犯罪的甲、乙两人，由于警方没有掌握充分的证据，所以将两人分开审讯：
  - 若一人认罪并指证对方，而另一方保持沉默，则此人会被当即释放，沉默者会被判监禁10年
  - 若两人都保持沉默，则根据已有的犯罪事实（无充分证据）两人各判半年
  - 若两人都认罪并相互指证，则两人各判5年

	乙沉默（合作）	乙认罪（背叛）
甲沉默（合作）	二人各服刑半年	乙被释放，甲服刑10年
甲认罪（背叛）	甲被释放，乙服刑10年	二人各服刑5年

- **参与者**：甲、乙
- **规则**：甲、乙两人分别决策，无法得知对方的选择
- **策略集**：认罪、沉默（纯策略）
- **局势及对应收益（年）**
  - 甲认罪：0　　乙沉默：-10
  - 甲认罪：-5　　乙认罪：-5
  - 甲沉默：-10　　乙认罪：0
  - 甲沉默：-0.5　　乙沉默：-0.5
- 在囚徒困境中，最优解为两人同时沉默，但是两人实际倾向于选择同时认罪（均衡解）

博弈论的相关概念：囚徒困境（prisoner's dilemma）

- 1950年，兰德公司的梅里尔·弗勒德和梅尔文·德雷希尔拟定了相关困境理论，后来美国普林斯顿大学数学家阿尔伯特·塔克以“囚徒方式”阐述：
  - 警方逮捕了共同犯罪的甲、乙两人，由于警方没有掌握充分的证据，所以将两人分开审讯：
  - 若一人认罪并指证对方，而另一方保持沉默，则此人会被当即释放，沉默者会被判监禁10年
  - 若两人都保持沉默，则根据已有的犯罪事实（无充分证据）两人各判半年
  - 若两人都认罪并相互指证，则两人各判5年

	乙沉默（合作）	乙认罪（背叛）
甲沉默（合作）	二人各服刑半年	乙被释放，甲服刑10年
甲认罪（背叛）	甲被释放，乙服刑10年	二人各服刑5年

- **参与者**：甲、乙
- **规则**：甲、乙两人分别决策，无法得知对方的选择
- **策略集**：认罪、沉默（纯策略）
- **局势及对应收益（年）**
  - 甲认罪：0　　乙沉默：-10
  - 甲认罪：-5　　乙认罪：-5
  - 甲沉默：-10　　乙认罪：0
  - 甲沉默：-0.5　　乙沉默：-0.5

- 在囚徒困境中，最优解为两人同时沉默，但是两人实际倾向于选择同时认罪（均衡解）

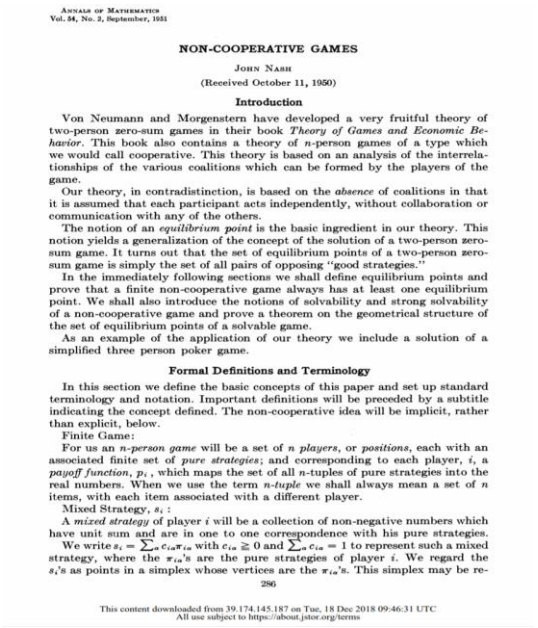
博弈论的相关概念：博弈的分类

- 合作博弈与非合作博弈
  - 合作博弈（cooperative game）：部分参与者可以组成联盟以获得更大的收益
  - 非合作博弈（non-cooperative game）：参与者在决策中都彼此独立，不事先达成合作意向
- 静态博弈与动态博弈
  - 静态博弈（static game）：所有参与者同时决策，或参与者互相不知道对方的决策
  - 动态博弈（dynamic game）：参与者所采取行为的先后顺序由规则决定，且后行动者知道先行动者所采取的行为
- 完全信息博弈与不完全信息博弈
  - 完全信息（complete information）：所有参与者均了解其他参与者的策略集、收益等信息
  - 不完全信息（incomplete information）：并非所有参与者均掌握了所有信息
- **囚徒困境是一种非合作、不完全信息的静态博弈**

博弈论的相关概念：纳什均衡

- 博弈的稳定局势即为纳什均衡（Nash equilibrium）：指的是参与者所作出的这样一种策略组合，在该策略组合上，任何参与者单独改变策略都不会得到好处。换句话说，如果在一个策略组合上，当所有其他人都不改变策略时，没有人会改变自己的策略，则该策略组合就是一个纳什均衡。
- Nash定理：若参与者有限，每位参与者的策略集有限，收益函数为实值函数，则博弈必存在混合策略意义下的纳什均衡。
- 囚徒困境中两人同时认罪就是这一问题的纳什均衡。

纳什均衡的本质
不后悔



ANNALS OF MATHEMATICS
Vol. 54, No. 2, September, 1951

NON-COOPERATIVE GAMES

JOHN NASH

(Received October 11, 1950)

Introduction

Von Neumann and Morgenstern have developed a very fruitful theory of two-person zero-sum games in their book *Theory of Games and Economic Behavior*. This book also contains a theory of n-person games of a type which we would call cooperative. This theory is based on an analysis of the interrelationships of the various coalitions which can be formed by the players of the game.

Our theory, in contradistinction, is based on the *absence* of coalitions in that it is assumed that each participant acts independently, without collaboration or communication with any of the others.

The notion of an *equilibrium point* is the basic ingredient in our theory. This notion yields a generalization of the concept of the solution of a two-person zero-sum game. It turns out that the set of equilibrium points of a two-person zero-sum game is simply the set of all pairs of opposing "good strategies."

In the immediately following sections we shall define equilibrium points and prove that a finite non-cooperative game always has at least one equilibrium point. We shall also introduce the notions of solvability and strong solvability of a non-cooperative game and prove a theorem on the geometrical structure of the set of equilibrium points of a solvable game.

As an example of the application of our theory we include a solution of a simplified three person poker game.

Formal Definitions and Terminology

In this section we define the basic concepts of this paper and set up standard terminology and notation. Important definitions will be preceded by a subtitle indicating the concept defined. The non-cooperative idea will be implicit, rather than explicit, below.

Finite Game:

For us an *n-person game* will be a set of n *players*, or *positions*, each with an associated finite set of *pure strategies*; and corresponding to each player, i, a *payoff function*, p_i, which maps the set of all n-tuples of pure strategies into the real numbers. When we use the term *n-tuple* we shall always mean a set of n items, with each item associated with a different player.

Mixed Strategy, s_i:

A *mixed strategy* of player i will be a collection of non-negative numbers which have unit sum and are in one to one correspondence with his pure strategies. We write $s_i = \sum_\alpha c_{i\alpha}\pi_{i\alpha}$ with $c_{i\alpha} \geqq 0$ and $\sum_\alpha c_{i\alpha} = 1$ to represent such a mixed strategy, where the $\pi_{i\alpha}$'s are the pure strategies of player i. We regard the s_i's as points in a simplex whose vertices are the $\pi_{i\alpha}$'s. This simplex may be re-

286

Nash, J, Non-Cooperative Games. *The Annals of Mathematics*. 54, 2 (1951), 286.

博弈论的相关概念：混合策略下纳什均衡的例子

- 例子：公司的雇主是否检查工作与雇员是否偷懒
- V是雇员的贡献，W是雇员的工资，H是雇员的付出，C是检查的成本，F是雇主发现雇员偷懒对雇员的惩罚（没收抵押金）。
- 假定$H < W < V$，$W > C$

		雇员	
		偷懒	不偷懒
雇主	检查	$-C+F,-F$	$V-W-C,W-H$
	不检查	$-W,W$	$V-W,W-H$

- 参与者：
  - 雇员、雇主
- 规则：
  - 雇员与雇主两人分别决策，事先无法得知对方的选择
- 混合策略集：
  - 雇员：偷懒、不偷懒
  - 雇主：检查、不检查
- 局势及对应收益
  - 雇主采取检查策略时雇员工作与偷懒对应的结果
  - 雇主采取不检查策略时雇员工作与偷懒对应的结果

博弈论的相关概念：混合策略下纳什均衡的例子

若雇主检查的概率为α，雇员偷懒的概率为β

- V是雇员的贡献，W是雇员的工资，H是雇员的付出，C是检查的成本，F是雇主发现雇员偷懒而对雇员的惩罚（没收抵押金）。
- 假定$H < W < V$，$W > C$

	采取策略	收益
雇主	检查	$T_1 = \beta(-C+F)+(1-\beta)(V-W-C)$
	不检查	$T_2 = -\beta W+(1-\beta)(V-W)$
雇员	偷懒	$T_3 = -\alpha F+(1-\alpha)W$
	不偷懒	$T_4 = (W-H)+(1-\alpha)(W-H)$=$(W-H)$

		雇员	
		偷懒	不偷懒
雇主	检查	$-C$ $+F,-F$	$V-W-C,W$ $-H$
	不检查	$-W,W$	$V-W,W-H$

博弈论的相关概念：混合策略下纳什均衡的例子

若雇主检查的概率为α，雇员偷懒的概率为β

	采取策略	收益
雇主	检查	$T_1 = \beta(-C+F)+(1-\beta)(V-W-C)$
	不检查	$T_2 = -\beta W+(1-\beta)(V-W)$
雇员	偷懒	$T_3 = -\alpha F+(1-\alpha)W$
	不偷懒	$T_4 = \alpha(W-H)+(1-\alpha)(W-H)=(W-H)$

混合策略纳什均衡：博弈过程中，博弈方通过概率形式随机从可选策略中选择一个策略而达到的纳什均衡被称为混合策略纳什均衡。

- 纳什均衡：其他参与者策略不变的情况下，某个参与者单独采取其他策略都不会使得收益增加⇔无论雇主是否检查，雇员的收益都一样；无论雇员是否偷懒，雇主的收益都一样
  - 于是有$T_1 = T_2$ 以及$T_3 = T_4$
  - 在纳什均衡下，由于$T_3 = T_4$，可知雇主采取检查策略的概率（雇主趋向于用这个概率去检查）：
$$\alpha = \frac{H}{W+F}$$
  - 在纳什均衡下，由于$T_1 = T_2$，可知雇员采取偷懒策略的概率（雇员趋向于用这个概率去偷懒）：
$$\beta = \frac{C}{W+F}$$
- 在检查概率为α之下，雇主的收益：
$$T_1 = T_2 = V - W - \frac{CV}{W+F}$$
  - 对上式中W求导，则当$W = \sqrt{CV} - F$时，雇主的收益最大，其值为$T_{max} = V - 2\sqrt{CV} + F$

博弈论与计算机科学

- 冯·诺依曼：现代计算机之父+现代博弈论之父

- 博弈论与计算机科学的交叉领域非常多
  - 理论计算机科学：算法博弈论
  - **人工智能：**多智能体系统、AI游戏玩家、人机交互、机器学习、广告推荐
  - 互联网：互联网经济、共享经济
  - 分布式系统：区块链

- 人工智能与博弈论相互结合，形成了两个主要研究方向
  - **博弈策略的求解**
  - **博弈规则的设计（例子：推恩令）**

博弈策略求解

- 动机
  - 博弈论提供了许多问题的数学模型
  - 纳什定理确定了博弈过程问题存在解
  - 人工智能的方法可用来求解均衡局面或者最优策略

- 主要问题
  - 如何高效求解博弈参与者的策略以及博弈的均衡局势？

- 应用领域
  - 大规模搜索空间的问题求解：围棋
  - 非完全信息博弈问题求解：德州扑克
  - 网络对战游戏智能：Dota、星球大战
  - 动态博弈的均衡解：厂家竞争、信息安全

遗憾最小化算法（Regret Minimization）：若干定义

- 假设一共有N个玩家。玩家i所采用的策略表示为σ_i。
- 玩家i的策略空间用Σ_i表示。
- 一个策略组包含所有玩家策略，用$\sigma = (\sigma_1, \sigma_2, \ldots, \sigma_{|N|})$。
- σ_{-i} 表示σ中除了σ_i之外的策略（即除去玩家i所采用的策略)
- 对于每个玩家$i \in N$，$u_i: Z \to R$ 表示玩家i的收益函数，即在到达终止序列集合Z中某个终止序列时，玩家i所得到的收益。
- 玩家i在给定策略σ下所能得到的期望收益表示为$u_i(\sigma)$

遗憾最小化算法：最佳反应策略与纳什均衡

- 玩家i对于所有其他玩家的策略组σ_{-i}的**最佳反应策略**σ_i^*满足如下条件：

$$u_i(\sigma_i^*, \sigma_{-i}) \geq \max_{\sigma_i' \in \Sigma_i} u_i(\sigma_i', \sigma_{-i})$$

在策略组 σ 中，如果每个玩家的策略相对于其他玩家的策略而言都是最佳反应策略，那么策略组σ就是一个**纳什均衡**（Nash equilibrium）策略。

纳什均衡：策略组$\sigma = (\sigma_1^*, \sigma_2^*, \dots, \sigma_{|N|}^*)$是纳什均衡当且仅当对每个玩家$i \in N$，满足如下条件：

$$u_i(\sigma^*) \geq \max_{\sigma_i' \in \Sigma_i} u_i(\sigma_1^*, \sigma_2^*, \dots, \sigma_i', \dots, \sigma_{|N|}^*)$$

遗憾最小化算法：策略选择

- 遗憾最小化算法是一种根据过去博弈中的遗憾程度来决定将来动作选择的方法
- 在博弈中，玩家i在第T轮次（每一轮表示一次博弈完成）采取策略σ_i的遗憾值定义如下（累加遗憾）：

$$Regret_i^T(\sigma_i) = \sum_{t=1}^{T}(\mu_i(\sigma_i, \sigma_{-i}^t) - \mu_i(\sigma^t))$$

- 通常遗憾值为负数的策略被认为不能提升下一时刻收益，所以这里考虑的遗憾值均为正数或0
- 计算得到玩家i在第T轮次采取策略σ_i的遗憾值后，在第$T+1$轮次玩家i选择策略a的概率如下（悔值越大、越选择，即亡羊补牢）

$$P(a) = \frac{Regret_i^T(a)}{\sum_{b\in\{\text{所有可选择策略}\}} Regret_i^T(b)}$$

遗憾最小化算法：石头-剪刀-布的例子

- 假设两个玩家A和B进行石头-剪刀-布（Rock-Paper-Scissors，RPS）的游戏，获胜玩家收益为1分，失败玩家收益为-1分，平局则两个玩家收益均为零分
- 第一局时，若玩家A出石头（R），玩家B出布（P），则此时玩家A的收益$\mu_A(R,P)=-1$，玩家B的收益为$\mu_B(P,R)=1$
- 对于玩家A来说，在玩家B出布（P）这个策略情况下，如果玩家A选择出布（P）或者剪刀（S），则玩家A对应的收益值$\mu_A(P,P)=0$或者$\mu_A(S,P)=1$
- 所以第一局之后，玩家A没有出布的遗憾值为$\mu_A(P,P)-\mu_A(R,P)=0-(-1)=1$，没有出剪刀的遗憾值为$\mu_A(S,P)-\mu_A(R,P)=1-(-1)=2$
- 所以在第二局中，玩家A选择石头、剪刀和布这三个策略的概率分别为0、2/3、1/3。**因此，玩家A趋向于在第二局中选择出剪刀这个策略**

遗憾最小化算法：石头-剪刀-布的例子

玩家i每一轮悔值计算公式：$\mu_i\left(\sigma_i, \sigma_{-i}^t\right) - \mu_i(\sigma^t)$

- 在第一轮中玩家A选择石头和玩家B选择布、在第二局中玩家A选择剪刀和玩家B选择石头情况下，则玩家A每一轮遗憾值及第二轮后的累加遗憾取值如下：

每轮悔值\策略	石头	剪刀	布
第一轮悔值	0	2	1
第二轮悔值	1	0	2
$Regret_A^2$	1	2	3

- 从上表可知，在第三局时，玩家A选择石头、剪刀和布的概率分别为1/6、2/6、3/6
- 在实际使用中，可以通过多次模拟迭代累加遗憾值找到每个玩家在每一轮次的最优策略