人工智能博弈

教学课程组

2022年

- 参考教材：吴飞，《人工智能导论：模型与算法》，高等教育出版社
- 在线课程(MOOC)：https://www.icourse163.org/course/ZJU-1003377027
- 在线实训平台（智海-Mo）：https://mo.zju.edu.cn/classroom/class/zju_ai_2022
- 系列科普读物《走进人工智能》 https://www.ximalaya.com/album/56494803

提纲

一、博弈论的相关概念

二、博弈策略求解

三、博弈规则设计

四、非完全信息博弈的实际应用

博弈论的诞生：中国古代博弈思想

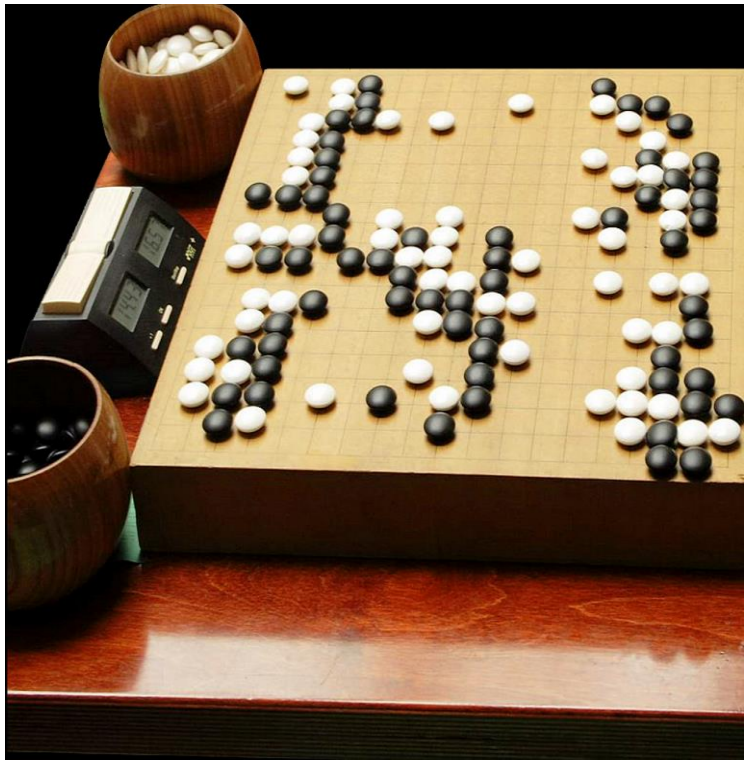
- 子曰：饱食终日，无所用心，难矣哉！不有博弈者乎？为之，犹贤乎已。

——《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》

表8.1 齐威王与田忌进行赛马中所采取的不同对局

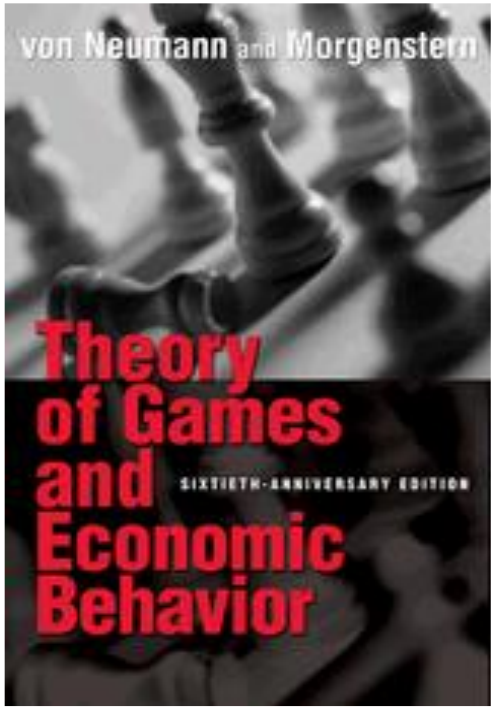
表8.1　齐威王与田忌进行赛马中所采取的不同对局

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

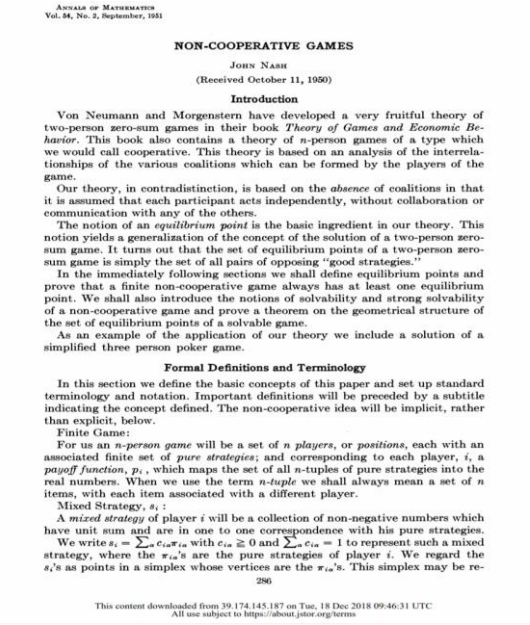
- 合作博弈与非合作博弈
  - 合作博弈（cooperative game）：部分参与者可以组成联盟以获得更大的收益
  - 非合作博弈（non-cooperative game）：参与者在决策中都彼此独立，不事先达成合作意向
- 静态博弈与动态博弈
  - 静态博弈（static game）：所有参与者同时决策，或参与者互相不知道对方的决策
  - 动态博弈（dynamic game）：参与者所采取行为的先后顺序由规则决定，且后行动者知道先行动者所采取的行为
- 完全信息博弈与不完全信息博弈
  - 完全信息（complete information）：所有参与者均了解其他参与者的策略集、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

表8.4 雇主-雇员每次采取对应行动后的收益

表8.4　雇主－雇员每次采取对应行动后的收益

		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

- 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$H < W < V$，$W > C$

若雇主检查的概率为α，
雇员偷懒的概率为β。

表8.4 雇主-雇员每次采取对应行动后的收益

表8.4　雇主－雇员每次采取对应行动后的收益

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

表8.5 雇主-雇员之间博弈的期望收益

表8.5　雇主－雇员之间博弈的期望收益

参与者	采取策略	期望收益
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	$T_4=\alpha(W-H)+(1-\alpha)(W-H)=W-H$

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

表8.5 雇主-雇员之间博弈的期望收益

表8.5 雇主－雇员之间博弈的期望收益

参与者	采取策略	期望收益
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，雇主的收益都一样；无论雇员是否偷懒，雇员的收益都一样
  - 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$T_{max}=V-2\sqrt{CV}+F$

博弈论的相关概念：策梅洛定理

策梅洛定理（Zermelo's theorem）：对于任意一个有限步的双人完全信息零和动态博弈，一定存在先手必胜策略或后手必胜策略或双方保平策略

假设双人博弈过程结果只有胜、负、和三种结果，博弈过程最长持续N步，下面通过数学归纳法来如下证明策梅洛定理：

1) 当$N=1$时，先手玩家必选择自己收益最大的策略，若存在获胜的行动，则存在先手必胜策略。若不存在使先手玩家获胜的行动，则先手玩家会选择平局的行动，于是存在双方保平策略。若仅存在后手玩家获胜的行动，则存在后手必胜策略。
2) 假设当$N=1,2\ldots,k$时，均存在先手必胜策略、后手必胜策略或双方保平策略。
3) 当$N=k+1$时，将先手玩家行动一次后的局势视为开启了一个新的博弈。在这开启的新博弈中，先手玩家为原博弈的后手玩家，新博弈中的后手玩家为原博弈的先手，应用2)中的假设，新开启的博弈必存在先手必胜策略或后手必胜策略或双方保平策略，则原博弈亦必存在后手必胜策略或先手必胜策略或双方保平策略。
4) 综上所述，对于任意一个有限步的双人完全信息零和动态博弈，一定存在先手必胜策略或后手必胜策略或双方保平策略。

提醒注意的是，策梅洛定理仅对两人博弈有效，如果博弈竞技者超过了2人，如对于三人博弈，策梅洛定理无法保证三方中一定有一方获胜、其他两方必败或者三方和局的策略。

提纲

一、博弈论的相关概念

二、博弈策略求解

三、博弈规则设计

四、非完全信息博弈的实际应用

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

虚拟遗憾最小化算法（Regret Minimization）：若干定义

- 对于一个有N个玩家参加的博弈，玩家i在博弈中采取的策略记为σ_i。
- 对于所有玩家来说，他们的所有策略构成了一个策略组合，记作$\sigma = \{\sigma_1, \sigma_2, \dots, \sigma_N\}$。策略组中，除玩家$i$外，其他玩家的策略组合记作$\sigma_{-i} = \{\sigma_1, \sigma_2, \dots, \sigma_{i-1}, \sigma_{i+1}, \dots, \sigma_N\}$。

虚拟遗憾最小化算法：最优反应策略

- 给定策略组合σ，玩家i在终结局势下的收益记作$u_i(\sigma)$。

- 在给定其他玩家的策略组合σ_{-i}的情况下，对玩家i而言的**最优反应策略**σ_i^*满足如下条件：

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

Σ_i是玩家i可以选择的所有策略，如上条件表示当玩家i采用最优反应策略时，玩家i能够获得最大收益。

虚拟遗憾最小化算法：纳什均衡

- 在策略组合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，那么策略组合σ^*就是一个**纳什均衡（Nash equilibrium）策略**。

策略组$\sigma^* = \{\sigma_1^*, \sigma_2^*, \dots, \sigma_N^*\}$对任意玩家$i = 1, .., N$，满足如下条件：

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

- 在博弈策略求解的过程中，希望求解得到每个玩家最优反应策略，若所有玩家都是理性的，则算法求解最优反应策略就是一个纳什均衡。考虑到计算资源有限这一前提，难以通过遍历博弈中所有策略组合来找到一个最优反应策略，因此需要找到一种能快速发现近似纳什均衡的方法。

遗憾最小化算法

遗憾最小化算法是一种根据以往博弈过程中所得遗憾程度来选择未来行为的方法。

玩家i在过去T轮中采取策略σ_i的累加遗憾值定义如下：

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

其中σ^t和σ_{-i}^t分别表示第t轮中所有玩家的策略组合和除了玩家i以外的策略组合。简单地说，累加遗憾值代表着在过去T轮中，玩家i在每一轮中选择策略σ_i所得收益与采取其他策略所得收益之差的累加。

遗憾最小化算法：有效遗憾值

在得到玩家i的所有可选策略的遗憾值后，可以根据遗憾值的大小来选择后续第$T+1$轮博弈的策略，这种选择方式被称为**遗憾匹配**[Greenwald 2006]。当然，通常遗憾值为负数的策略被认为不能提升下一时刻收益，所以如下定义有效遗憾值：

$$Regret_i^{T,+}(\sigma_i) = \max\left(Regret_i^T(\sigma_i), 0\right)$$

遗憾最小化算法：有效遗憾值

利用有效遗憾值的遗憾匹配可得到玩家i在T轮后第$T+1$轮选择策略σ_i的概率$P(\sigma_i^{T+1})$为：

$$P(\sigma_i^{T+1}) = \begin{cases} \dfrac{Regret_i^{T,+}(\sigma_i)}{\sum_{\sigma_i' \in \Sigma_i} Regret_i^{T,+}(\sigma_i')} & if \sum_{\sigma_i' \in \Sigma_i} Regret_i^{T,+}(\sigma_i') > 0 \\ \dfrac{1}{|\Sigma_i|} & \text{otherwise} \end{cases}$$

$|\Sigma_i|$表示玩家i所有策略的总数。显然，如果在过往T轮中策略σ_i所带来的遗憾值大、其他策略σ_i'所带来的遗憾值小，则在第$T+1$轮选择策略σ_i的概率值$P(\sigma_i^{T+1})$就大。也就是说，带来越大遗憾值的策略具有更高的价值，因此其在后续被选择的概率就应该越大。如果没有一个能够提升前T轮收益的策略，则在后续轮次中随机选择一种策略。依照一定的概率选择行动是为了防止对手发现自己所采取的策略（如采取遗憾值最大的策略）。

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：$\mu_i(\sigma_i, \sigma_{-i}^t) - \mu_i(\sigma^t)$

- 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：

表8.7 玩家A在两轮后所得到的遗憾值

表8.7　玩家A在两轮后所得到的遗憾值

遗憾值/策略	石头	剪刀	布
第一轮	0	2	1
第二轮	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
- 但是当博弈状态空间呈指数增长时，对一个规模巨大的博弈树无法采用最小遗憾算法，因此需要采取虚拟遗憾最小化算法（counterfactual regret minimization)

遗憾最小化算法：计算理论

对于任何序贯决策的博弈对抗，可将博弈过程表示成一棵博弈树，博弈树中的每一个中间节点都是一个信息集I，信息集中包含了博弈中当前的状态。给定博弈树的每一个节点，玩家都可以从一系列的动作中选择一个，然后状态发生转换，如此周而复始，直到终局（博弈树的叶子节点）。玩家在当前状态下可采取的策略就是当前状态下所有可能动作的一个概率分布。

具体而言，在信息集I下，玩家可以采取的行动集合记作$A(I)$。玩家i所采取的行动$a_i \in A(I)$可认为是其采取的策略σ_i的一部分。在信息集I下采取行动a所代表的策略记为$\sigma_{I \to a}$。这样，要计算虚拟遗憾值的对象就是博弈树中每个中间节点在信息集下所采取的行动，并根据遗憾值匹配得到该节点在信息集下应该采取的策略$\sigma_{I \to a}$。

遗憾最小化算法：计算理论

- 在一次博弈中，所有玩家交替采取的行动序列记为h（从根节点到当前节点的路径），对于所有玩家的策略组合σ，行动序列h出现的概率记为$\pi^{\sigma}(h)$，不同的行动序列可以从根节点到达当前节点的信息集I（即不同决策路径可到达博弈树中同一个中间节点）。在策略组合σ下，所有能够到达该信息集的行动序列的概率累加就是该信息集的出现概率，即$\pi^{\sigma}(I)=\sum_{h\in I}\pi^{\sigma}(h)$。
- 博弈的终结局势集合也就是博弈树中叶子节点的集合，记为Z。对于任意一个终结局势$z\in Z$，玩家i在此终点局势下的收益记作$u_i(z)$。给定行动序列h，依照策略组合σ最终到达终结局势z的概率记作$\pi^{\sigma}(h,z)$。

遗憾最小化算法：计算理论

- 在策略组合σ下，对玩家i而言，如下计算从根节点到当前节点的行动序列路径h的虚拟价值：

$$v_i(\sigma, h$$

$$= \sum_{z \in Z} \underbrace{\pi^{\sigma}_{-i}(h)}_{\text{不考虑玩家}i\text{的策略到达当前节点概率}} \times \underbrace{\pi^{\sigma}(h,z)}_{\text{从当前节点到叶子结点概率}}$$

$$\times \underbrace{u_i(z)}_{\text{叶子结点}z\text{收益}}$$

在上式中，$\pi^{\sigma}_{-i}(h)$表示从根节点出发，不考虑玩家i的策略，仅考虑其他玩家策略而经过路径h到达当前节点的概率。也就是说，即使玩家i有其他策略，总是要求玩家i在每次选择时都选择路径h中对应的动作，以保证从根节点出发能够到达当前节点。可见，行动序列路径h的虚拟价值等于如下三项结果的乘积：不考虑玩家i的策略（仅考虑其他玩家策略）经过路径h到达当前节点的概率、从当前节点走到叶子结点（博弈结束）的概率、所到达叶子节点的收益。

遗憾最小化算法：计算理论

- 在定义了行动序列路径h的虚拟价值之后，就可如下计算玩家i在基于路径h到达当前节点采取行动a的遗憾值：

$$r_i(h,a) = v_i(\sigma_{I\to a}, h) - v_i(\sigma, h)$$

- 对能够到达同一个信息集I（即博弈树中同一个中间节点）的所有行动序列的遗憾值进行累加，即可得到信息集I的遗憾值：

$$r_i(I,a) = \sum_{h\in I} r_i(h,a)$$

- 类似于遗憾最小化算法，虚拟遗憾最小化的遗憾值是T轮重复博弈后的累加值：

$$Regret_i^T(I,a) = \sum_{t=1}^{T} r_i^t(I,a)$$

- $r_i^t(I,a)$表示玩家i在第t轮中于当前节点选择行动a的遗憾值。

遗憾最小化算法：计算理论

- 进一步可以定义有效虚拟遗憾值：

$$Regret_i^{T,+}(I,a) = \max\left(R_i^T(I,a), 0\right)$$

- 根据有效虚拟遗憾值进行遗憾匹配以计算经过T轮博弈后，玩家i在信息集I情况下于后续$T+1$轮选择行动a的概率：

$$\sigma_i^{T+1}(I,a) = \begin{cases} \dfrac{Regret_i^{T,+}(I,a)}{\sum_{a\in A(I)} Regret_i^{T,+}(I,a)} & if \sum_{a\in A(I)} Regret_i^{T,+}(I,a) > 0 \\ \dfrac{1}{|A(I)|} & \text{otherwise} \end{cases}$$

遗憾最小化算法

在虚拟最小化算法的求解过程中，同样需要反复模拟多轮博弈来拟合最佳反应策略，算法步骤如下：

1) 初始化遗憾值和累加策略表为0

2) 采用随机选择的方法来决定策略

3) 利用当前策略与对手进行博弈

4) 计算每个玩家采取每次行为后的遗憾值

5) 根据博弈结果计算每个行动的累加遗憾值大小来更新策略

6) 重复3)到5)步若干次，不断的优化策略

7) 根据重复博弈最终的策略，完成最终的动作选择

表8.8 双人库恩扑克游戏规则

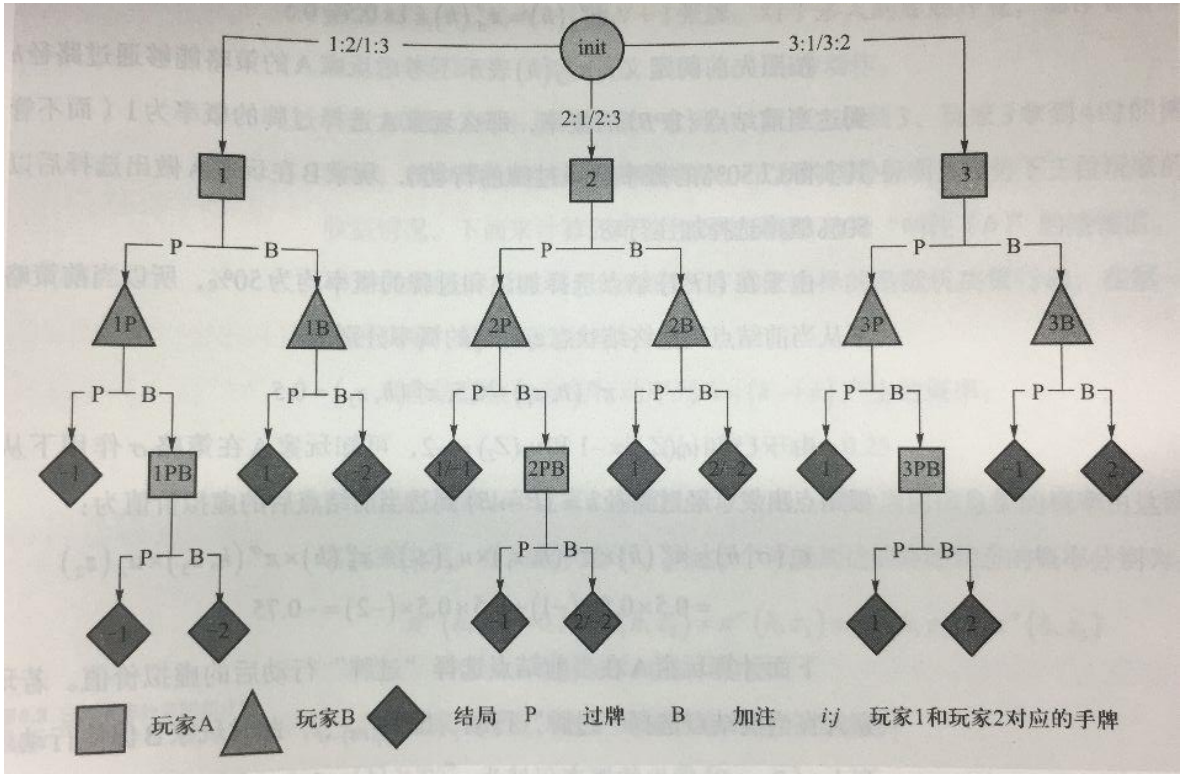
表8.8 双人库恩扑克游戏规则

玩家A	玩家B	玩家A	结果
过牌	过牌	游戏结束	牌值大的玩家 +1
加注	加注	游戏结束	牌值大的玩家 +2
过牌	加注	过牌	玩家B +1
过牌	加注	加注	牌值大的玩家 +2
加注	过牌	游戏结束	玩家A +1

库恩扑克[Kuhn 1950]是一种简单的有限注扑克游戏，由两名玩家进行游戏博弈，游戏中仅提供牌值为1、2和3的三张纸牌。每轮中每位玩家各持一张纸牌，每位玩家根据各自判断来决定是否追加定额赌注。摊牌阶段时来比较未弃牌玩家的底牌大小，底牌值最大的玩家即为胜者。

遗憾最小化算法

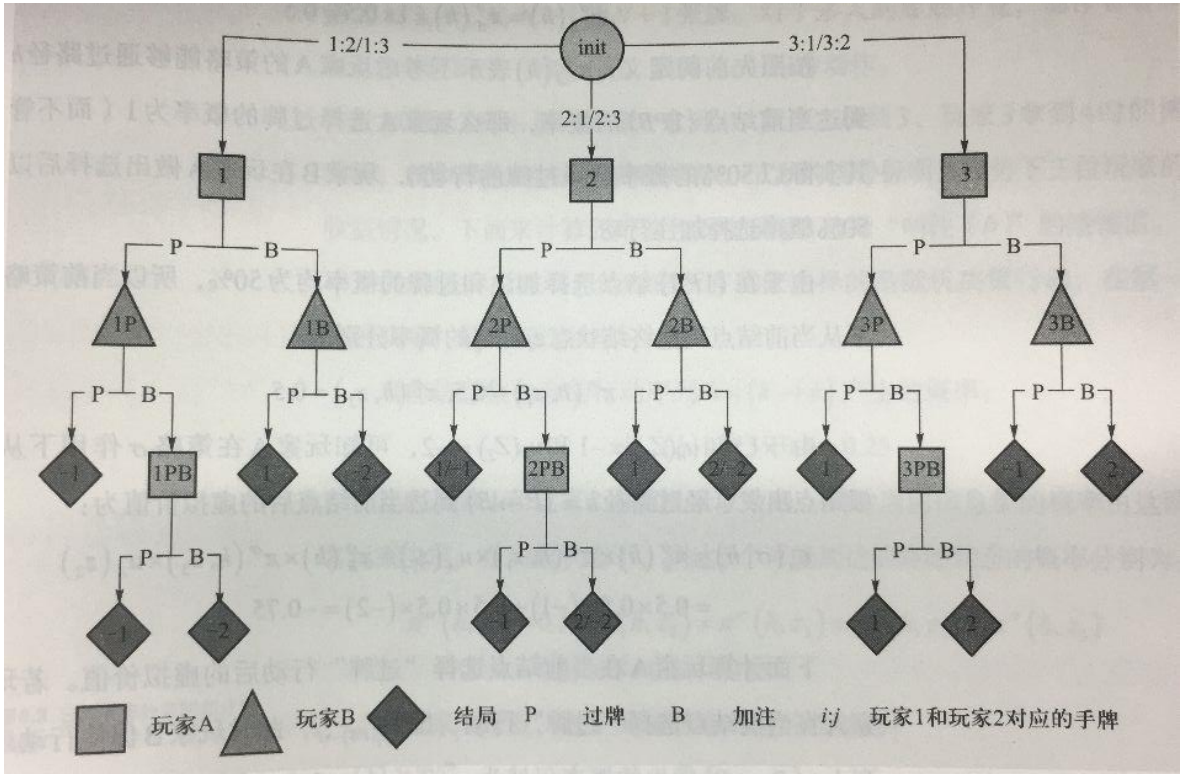
图8.1 先手玩家A对应的库恩扑克博弈树



- 在初始根节点下，三条边分别代表玩家A拿到三张不同手牌的情况。库恩扑克的信息集（即博弈树的中间节点）有12个：$\{1,1P,1B,1BP,2,2P,2B,2PB,3,3P,3B,3PB\}$。在信息集中，$P$表示过牌、$B$表示下注。

遗憾最小化算法

图8.1 先手玩家A对应的库恩扑克博弈树



- 要注意的是从根节点到达每个信息集的路径只有一条，该路径所对应的行动序列也只有一种，例如信息集$1PB$（博弈树中的一个中间节点）可通过如下路径到达：

$$1(\text{玩家A拿到大小为1纸牌}) \xrightarrow{P} 1P(\text{玩家A选择过牌}) \xrightarrow{B} 1PB(\text{玩家B选择加注})$$

遗憾最小化算法

- 这里以信息集$\{1PB\}$为例，来计算玩家A通过路径$1 \xrightarrow{P} 1P \xrightarrow{B} 1PB$到达当前节点$\{1PB\}$后选择“过牌”行动的遗憾值。
- 在当前策略下，行动序列路径$h = \{P \to B\}$产生的概率：

$$\pi_{-A}^{\sigma}(h) = \pi_{B}^{\sigma}(h) = 1 \times 0.5 = 0.5$$

- 按照先前的定义，$\pi_{-A}^{\sigma}(h)$表示不考虑玩家A的策略能够通过路径h到达当前节点$\{1PB\}$的概率，那么玩家A选择过牌的概率为1（而不管其实际以50%的概率选择过牌的行动）、玩家B在玩家A做出选择后以50%概率选择加注行动。
- 由于在$\{1PB\}$节点选择加注和过牌的概率均为50%，所以当前策略下从当前节点到达终结状态z_1和z_2的概率分别为：

$$\pi^{\sigma}(h, z_1) = 0.5, \pi^{\sigma}(h, z_2) = 0.5$$

- 由于已知$u_A(z_1) = -1$和$u_A(z_2) = -2$，可知玩家A在策略σ作用下从根节点出发、经过路径$h = \{P \to B\}$到达当前节点后的虚拟价值：

$$v_A(\sigma, h) = \pi_B^{\sigma}(h) \times \pi^{\sigma}(h, z_1) \times u_A(z_1) + \pi_B^{\sigma}(h) \times \pi^{\sigma}(h, z_2) \times u_A(z_2)$$

$$= 0.5 \times 0.5 \times (-1) + 0.5 \times 0.5 \times (-2) = -0.75$$

遗憾最小化算法

- 下面计算玩家A在当前节点选择“过牌”行动后的虚拟价值。若玩家A在当前节点选择“过牌”行动，即$\sigma_{\{1PB\}\to P}$，此时玩家B促使行动序列$h=\{P\to B\}$发生的概率仍然为$\pi_B^{\sigma_{\{1PB\}\to P}}(h)=0.5$。由于从当前节点出发抵达的终结状态只有$z_1$，所以$\pi^{\sigma_{\{1PB\}\to P}}(h,z_1)=1$
- 则玩家A在当前节点选择“过牌”行动后的虚拟价值为：

$$v_A\left(\sigma_{\{1PB\}\to P},h\right)=\pi_B^{\sigma_{\{1PB\}\to P}}(h)\times\pi^{\sigma_{\{1PB\}\to P}}(h,z_1)\times u_A(z_1)$$
$$=0.5\times 1\times(-1)=-0.5$$

遗憾最小化算法

- 最终，在信息集$\{1PB\}$上采取“**过牌**”的虚拟遗憾值如下计算：

$$\begin{aligned} r_A(\{1PB\}, P) &= r_A(h, P) \\ &= v_A\left(\sigma_{\{1PB\}\to P}, h\right) - v_A(\sigma, h) \\ &= (-0.5) - (-0.75) = 0.25 \end{aligned}$$

- 在第一轮博弈中，累加虚拟遗憾值与行为遗憾值相同：

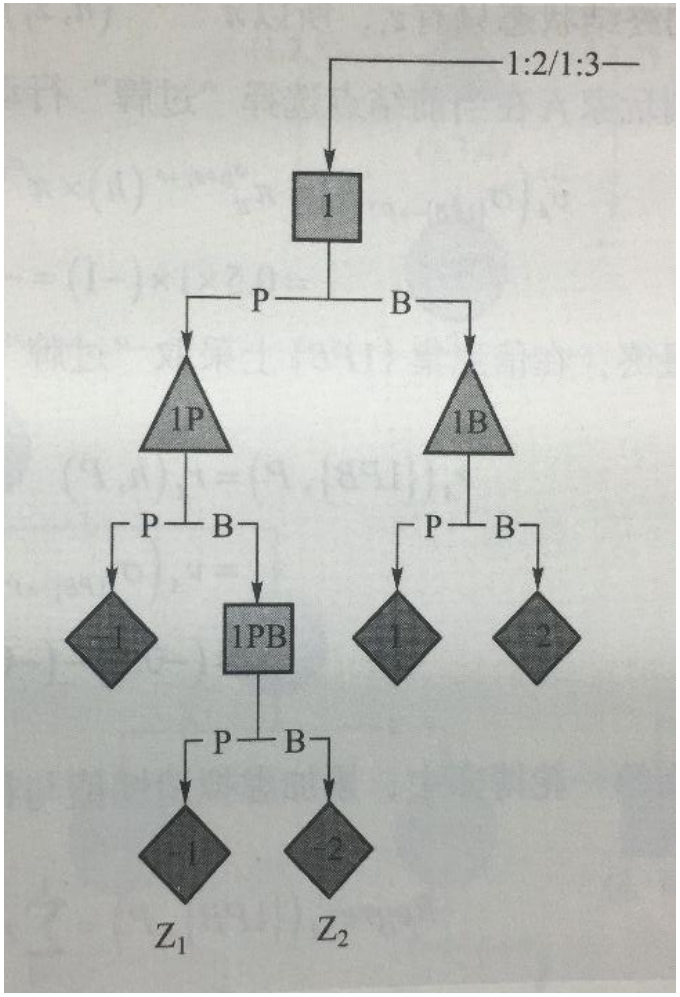
$$Regret_A^1(\{1PB\}, P) = \sum_{t=1}^{1} r_A^t(\{1PB\}, P)$$

- 类似的，可以计算信息集$\{1PB\}$上采取“**加注**”策略的遗憾值：

$$Regret_A^1(\{1PB\}, B) = r_A(\{1PB\}, B) = r_A(h, B) = -0.25$$

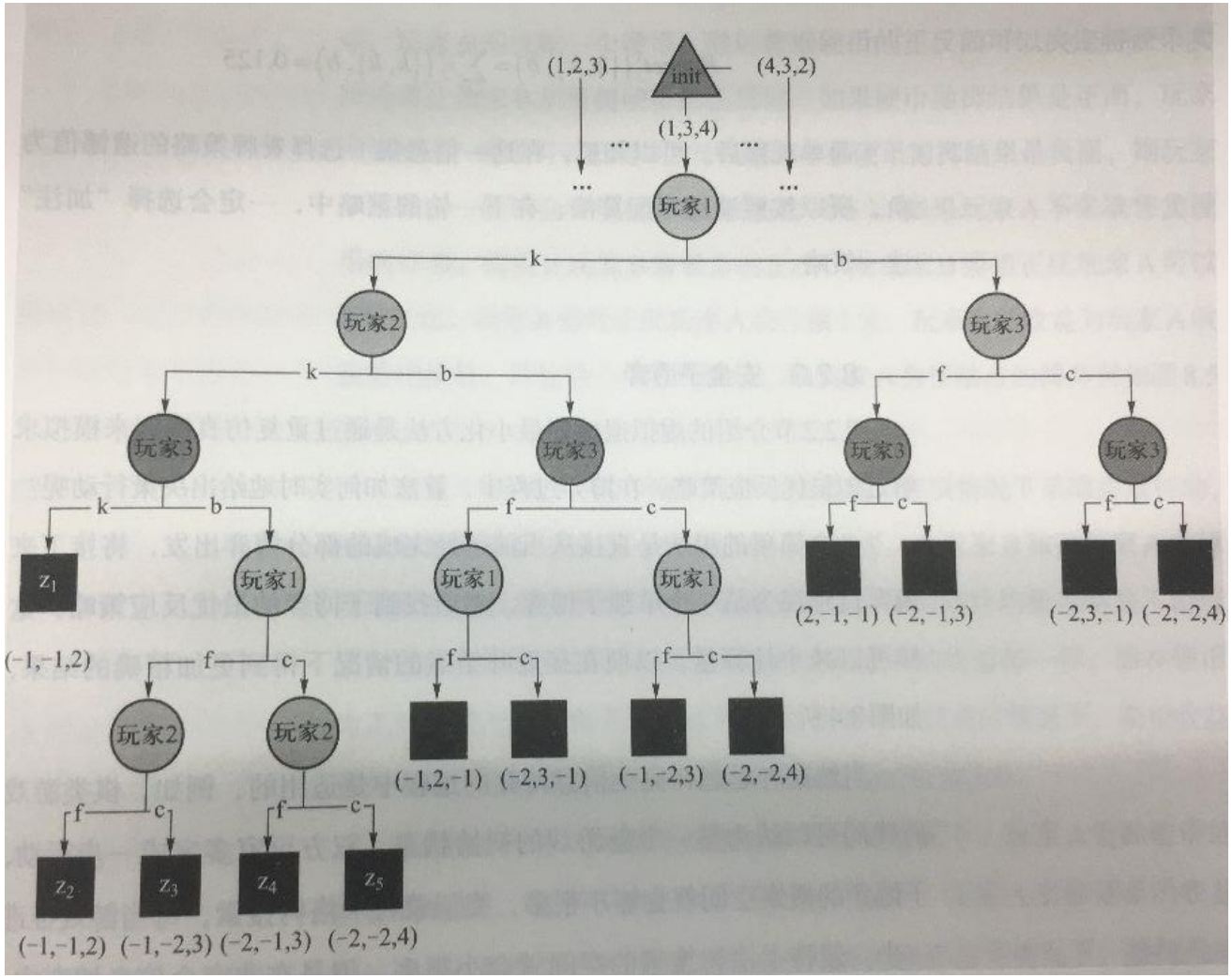
于是，根据遗憾值匹配算法计算结果，在下一轮的策略中会选择“过牌”策略。

图8.2 先手玩家手牌为1的库恩扑克博弈树



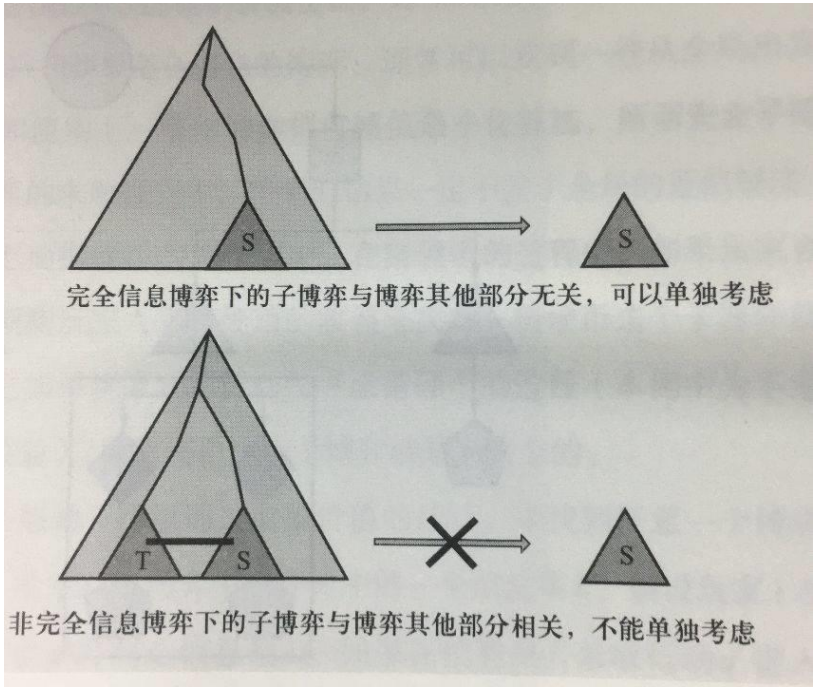
遗憾最小化算法

图8.3 三人库恩扑克的部分博弈树



安全子博弈

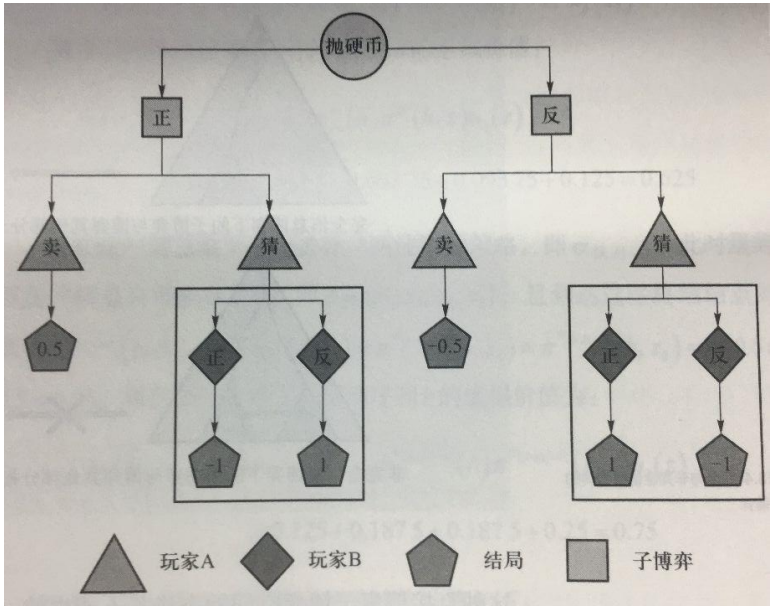
图8.4 完全与非完全信息博弈的子博弈



从当前已经完成的部分博弈出发，将接下来博弈过程视为是一个单独子博弈，然后找到子博弈的最优反应策略，这样可以减小计算量，以便在接近叶节点的情况下得到更加精确的结果

安全子博弈

图8.5 抛硬币游戏博弈树



- 玩家A随机抛一个硬币，然后根据硬币的正反面可以决定将硬币卖掉或者让玩家B来猜测硬币的正反面。如果硬币抛掷结果是正面，玩家A卖掉这一硬币可以获得0.5元收益，如果硬币抛掷结果是负面，则玩家A卖掉这一硬币会亏损0.5元（收益为-0.5）。
- 如果玩家A不采取售卖硬币的行动，而是让玩家B猜硬币的正反，则玩家B猜错正反前提下玩家A可以获得1元或者玩家B猜对正反前提下玩家A会亏损1元。

安全子博弈

- 在硬币投掷结果为正面情况下，玩家A卖掉硬币的收益$T_1 = 0.5$，在硬币投掷结果为反面情况下，玩家A卖掉硬币的收益$T_2 = -0.5$。在硬币投掷结果为正面或反面的概率相等情况下，要使玩家A选择卖硬币和猜硬币的收益一样，则玩家B的策略要使得玩家A在硬币投掷结果为正面情况下选择猜硬币这一行动的期望收益$T_3 = 0.5$，在硬币投掷结果为反面情况下选择猜硬币这一行动的期望收益$T_4 = -0.5$。

- 假设玩家B猜硬币投掷结果为正面的概率为a，对应的其猜硬币投掷结果为反面的概率为$(1-a)$，则有：

$$T_3 = -a + (1-a) = 0.5, T_4 = a - (1-a) = -0.5$$

- 从而可求得玩家B猜测硬币投掷结果为正面的概率为0.25、猜测硬币投掷结果为反面的概率为0.75。在玩家B按照这样的概率来猜测硬币投掷结果为正面或反面的情况下，将使得玩家A的最终期望收益为0。
- 在这场博弈中，即使玩家A选择了让玩家B来猜测硬币投掷结果的正反，玩家B也必须根据玩家A选择卖掉硬币的期望收益来计算自己纳什均衡的策略。

安全子博弈

- 假设玩家A在硬币投掷结果为正面情况下卖掉硬币的收益$T_1 = -0.5$，在硬币投掷结果为反面情况下卖掉硬币的收益$T_2 = 0.5$，那么类似地可以求出玩家B的纳什均衡反应策略应该是以0.25概率来猜测硬币投掷结果是反面，以0.75概率来猜测硬币投掷结果是正面。如果单独考虑猜硬币的情况，而不考虑卖硬币的情况，在正面和反面出现概率相同的情况下，玩家B应该以0.5的概率猜测正面，以0.5的概率猜测反面。
- 对于一个非完全信息的博弈，通常可以找到一些从全局出发的近似解法，例如使用上一节中的虚拟遗憾值最小化算法，所谓**安全子博弈**是指在子博弈的求解过程中，得到的结果一定不差于全局的近似解法。
- 在上面抛硬币的例子当中，在猜硬币的过程中，如果玩家B的行动是通过预测玩家A卖掉硬币的收益来决定，猜硬币这个子博弈就是安全的。反之如果玩家B的行动仅考虑猜硬币的过程（本例中为不考虑A买硬币的收益），则猜硬币这个子博弈就是不安全的。

提纲

一、博弈论的相关概念

二、博弈策略求解

三、博弈规则设计

四、非完全信息博弈的实际应用

博弈规则设计：研究问题

- 在现实生活中，如果所有博弈者都追求自己利益最大化，很可能会导致两败俱伤的下场，那么应该如何设计博弈的规则使得博弈的最终局势能尽可能达到整体利益的最大化呢?

- 例如许多城市为了避免机动车过快增长造成城市拥堵，因此限制了车牌的发放量。这样，车牌的发放方式就是需要决策者经过周密的考虑后决定的。通常而言，车牌的发放既要能满足有紧迫需求的人，又要满足普通家庭的日常需求，所以很多城市都采取了车牌竞价和随机摇号两种方式发放车牌。

博弈规则设计：双边匹配算法

- 在生活中，人们常常会碰到与资源匹配相关的决策问题(如求职就业、报考录取等)，这些需要双向选择的情况被称为是**双边匹配问题**。在双边匹配问题中，需要双方互相满足对方的需求才会达成匹配。

稳定婚姻问题（stable marriage problem）

就是一个典型的双边匹配问题，这个问题是指在给定成员偏好顺序的情况下，为两组成员寻找稳定的匹配。假设有n个单身男性构成的集合$M = \{m_1, m_2, \dots, m_n\}$，以及$n$个单身女性构成的集合$F = \{f_1, f_2, \dots, f_n\}$。对于任意一名单身男性$m_i$，都有自己爱慕的单身女性的顺序$s_{m_i} := f_{m_{i,1}} \succ f_{m_{i,2}} \succ \cdots \succ f_{m_{i,n}}$，这里$f_{m_{i,j}}$表示第$i$名男性所喜欢单身女性中排在第$j$位的单身女性，同理对于任意一名单身女性$f_i$也有其爱慕的单身男性顺序$s_{f_i} := m_{f_{i,1}} \succ m_{f_{i,2}} \succ \cdots \succ m_{f_{i,n}}$，$m_{f_{i,j}}$表示第$i$名女性所喜欢单身男性中排在第$j$位的单身男性。算法的最终目标是为这$2n$个男士和女士匹配得到$n$对伴侣，每一对伴侣可以表示为$(m_i, f_j)$。

博弈规则设计：双边匹配算法

- 假设有4名单身男性$\{1,2,3,4\}$和4名单身女性$\{A,B,C,D\}$，他（她）们的爱慕序列如表8.9所示

表8.9 稳定婚姻匹配偏好序列

表8.9 稳定婚姻匹配偏好序列

男性	偏好	女性	偏好
1	$A \succ D \succ C \succ B$	A	$3 \succ 4 \succ 2 \succ 1$
2	$A \succ B \succ C \succ D$	B	$3 \succ 2 \succ 4 \succ 1$
3	$A \succ C \succ D \succ B$	C	$1 \succ 3 \succ 4 \succ 2$
4	$B \succ A \succ D \succ C$	D	$2 \succ 4 \succ 3 \succ 1$

博弈规则设计：双边匹配算法

- 按照“修补”策略，匹配和修补过程如下

表8.10 修补策略下的稳定婚姻匹配过程

表8.10 修补策略下的稳定婚姻匹配过程

匹配1	修补	匹配2	修补	匹配3	修补	匹配4	修补	匹配5
(1, *A*)	(2, *A*)	(1, *B*)	(4, *A*)	(1, *B*)	(4, *B*)	(1, *A*)	(2, *B*)	(1, *A*)
(2, *B*)		(2, *A*)		(2, *D*)		(2, *D*)		(2, *B*)
(3, *C*)		(3, *C*)		(3, *C*)		(3, *C*)		(3, *C*)
(4, *D*)		(4, *D*)		(4, *A*)		(4, *B*)		(4, *D*)

博弈规则设计：双边匹配算法

1962年，美国数学家大卫·盖尔和博弈论学家沙普利提出了针对双边稳定匹配问题的解决算法（也被称为Gale- Shapely算法或G-S算法），并将其应用于稳定婚姻问题的求解[Gale 1962]，算法过程如下：

1) 单身男性向最喜欢的女性表白

2) 所有收到表白的女性从向其表白男性中选择最喜欢的男性，暂时匹配

3) 未匹配的男性继续向没有拒绝过他的女性表白。收到表白的女性如果没有完成匹配，则从这一批表白者中选择最喜欢男性。即使收到表白的女性已经完成匹配，但是如果她认为有她更喜欢的男性，则可以拒绝之前的匹配者，重新匹配

4) 如此循环迭代，直到所有人都成功匹配为止

表8.11 G-S算法下的稳定婚姻匹配过程

表8.11 G-S算法下的稳定婚姻匹配过程

第一轮		第二轮		第三轮		第四轮		第五轮	
表白	选择	表白	选择	表白	选择	表白	选择	表白	选择
(1, *A*)	\	(1, *D*)	(1, *D*)	\	(1, *D*)	\	\	(1, *C*)	(1, *C*)
(2, *A*)	\	(2, *B*)	(2, *B*)	\	(2, *B*)	\	(2, *B*)	\	(2, *B*)
(3, *A*)	(3, *A*)	\	(3, *A*)	\	(3, *A*)	\	(3, *A*)	\	(3, *A*)
(4, *B*)	(4, *B*)	\	\	(4, *A*)	\	(4, *D*)	(4, *D*)	\	(4, *D*)

在第一轮中，4名男性分别向自己最喜欢的女性告白，而收到3人告白的女性A选择了自己最喜欢的男性3，另一个收到告白的女性B选择了男性4；在第二轮中，尚未匹配的男性1和男性2继续向自己第二喜欢的对象告白，收到告白的女性B选择了自己更喜欢的男性2而放弃了男性4，同理继续三轮告白和选择，所有人都找到了自己的伴侣，且所有匹配都是稳定的。可以看出，使用G-S算法得到了稳定匹配的结果。

博弈规则设计：单边匹配算法

在匹配问题中，除了需要双向选择的双边匹配问题，还有一种类似于以物易物方式的交换匹配问题，被称为**单边匹配问题**，例如室友的匹配或者是座位的分配。这些问题中分配的对象都是不可分的标的物，他们只能属于一个所有者，且可以属于任何一个所有者。

博弈规则设计：单边匹配算法

对于单边匹配问题，1974年，沙普利和斯卡夫提出了针对单边匹配问题的稳定匹配算法“最大交易圈算法（Top-trading cycle，TTC）”[Shapley 1974]，算法过程如下：

1) 首先记录每个标的物的初始占有者，或者对物品进行随机分配。
2) 每个交易者连接一条指向他最喜欢的标的物的边，并从每一个标的物连接到其占有者或是具有高优先权的交易者。
3) 此时形成一张有向图，且必存在环，这种环被称为“交易圈”，对于交易圈中的交易者，将每人指向节点所代表的标的物赋予交易者，同时交易者放弃原先占有的标的物，占有者和匹配成功的标的物离开匹配市场。
4) 接着从剩余的交易者和标的物之间重复进行交易圈匹配，直到无法形成交易圈，算法停止。

博弈规则设计：单边匹配算法

稳定室友匹配问题就是一个典型的单边匹配问题。假设某寝室有A、B、C、D四位同学和1、2、3、4四个床位，当前给A、B、C、D四位同学随机分配4、3、2、1四个床位。

表8.12 床位偏好顺序

表8.12 床位偏好顺序

同学	偏好
A	$1 \succ 2 \succ 3 \succ 4$
B	$2 \succ 1 \succ 4 \succ 3$
C	$1 \succ 2 \succ 4 \succ 3$
D	$4 \succ 3 \succ 1 \succ 2$

博弈规则设计：单边匹配算法

图8.6 第一轮单边匹配

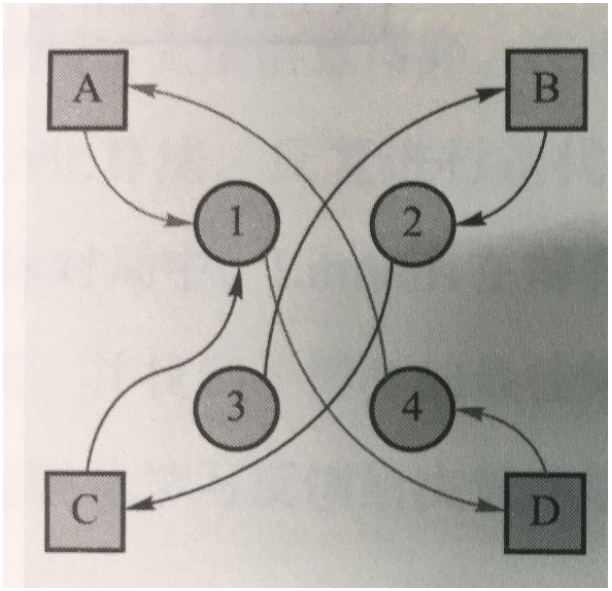
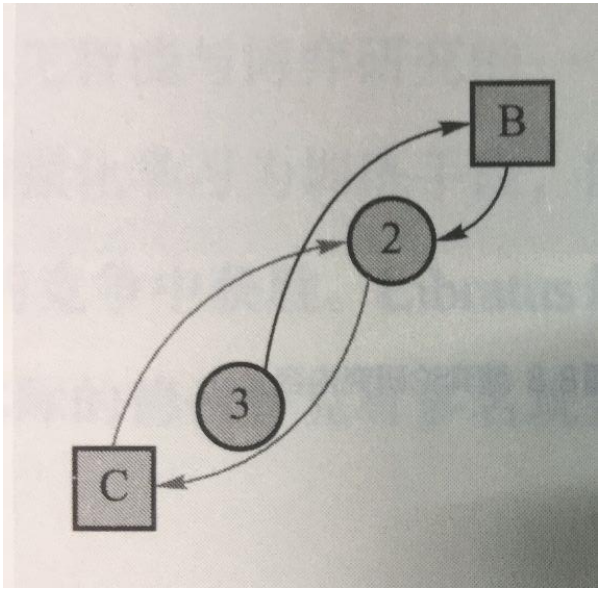


图8.7 第二轮单边匹配



- 在图8.6可以看出:，A和D之间构成一个交易圈，可达成交易，所以A得到床位1，D得到床位4，之后将A和D以及1和4从匹配图中移除。
- 从图8.7可以看出，B和C都希望得到床位2，无法再构成交易圈，但是由于C是床位的本身拥有者，所以C仍然得到床位2，B只能选择床位3 。

博弈规则设计：单边匹配算法

图8.6 第一轮单边匹配

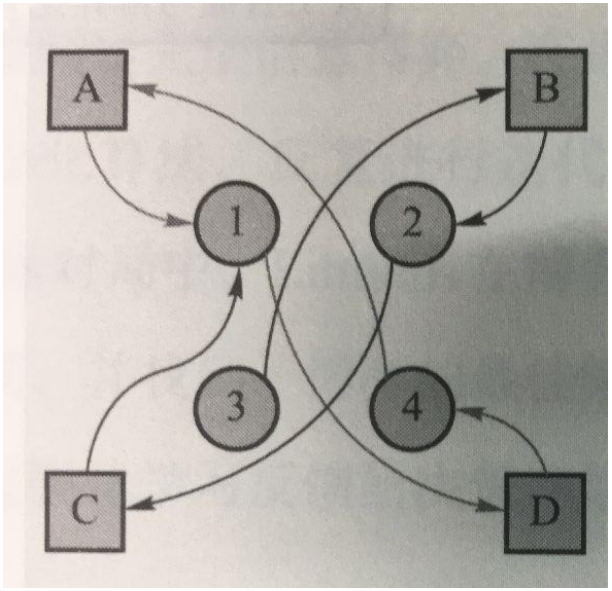
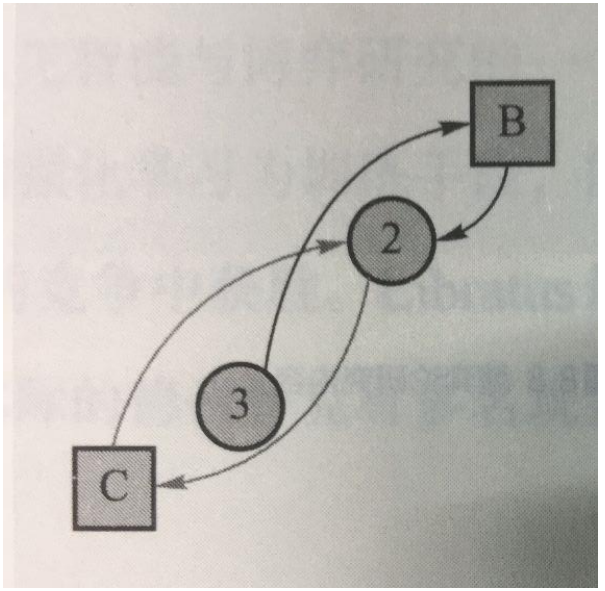


图8.7 第二轮单边匹配



- 在图8.6可以看出:，A和D之间构成一个交易圈，可达成交易，所以A得到床位1，D得到床位4，之后将A和D以及1和4从匹配图中移除。
- 从图8.7可以看出，B和C都希望得到床位2，无法再构成交易圈，但是由于C是床位的本身拥有者，所以C仍然得到床位2，B只能选择床位3 。

最后交易结果A→1，B →3，C →2，D →4

提纲

一、博弈论的相关概念

二、博弈策略求解

三、博弈规则设计

四、非完全信息博弈的实际应用

非完全信息博弈的实际应用

图8.8 博弈论研究内容

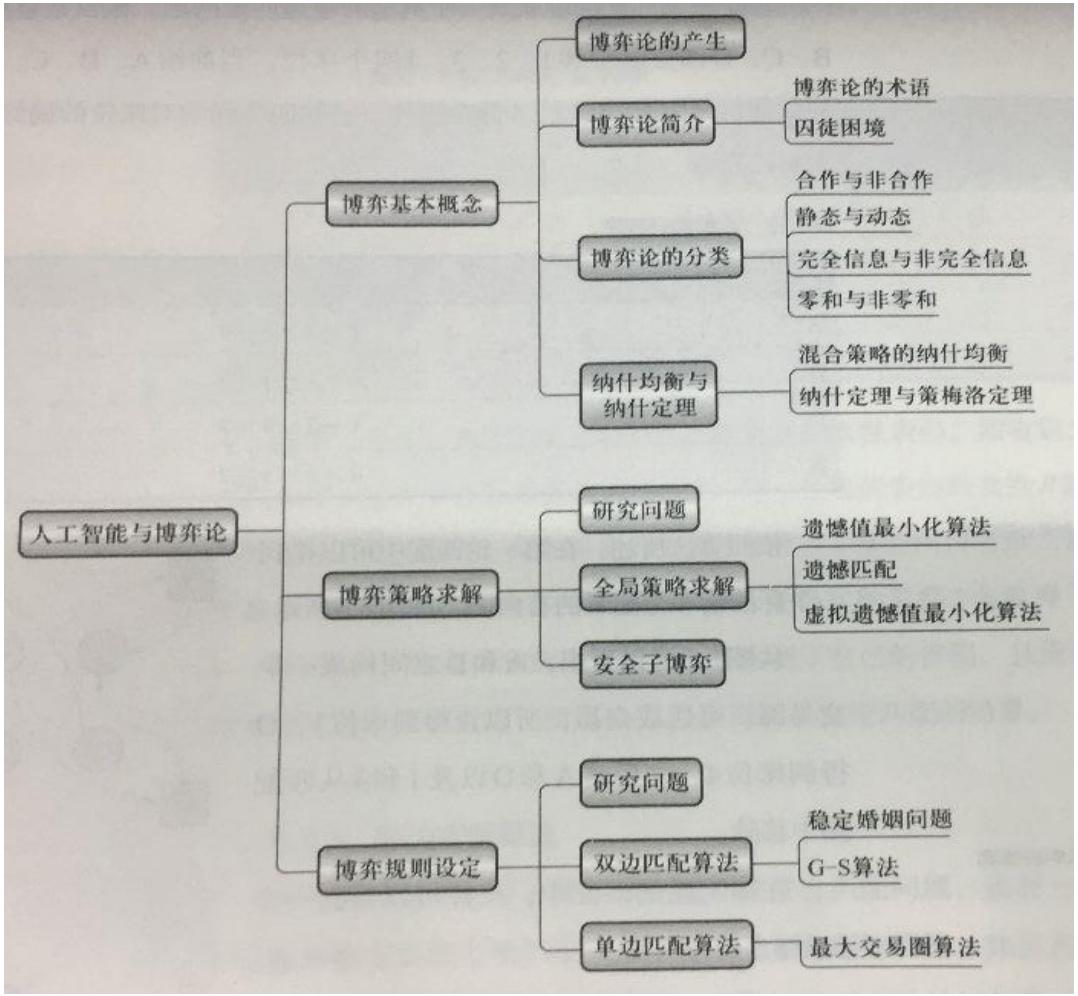
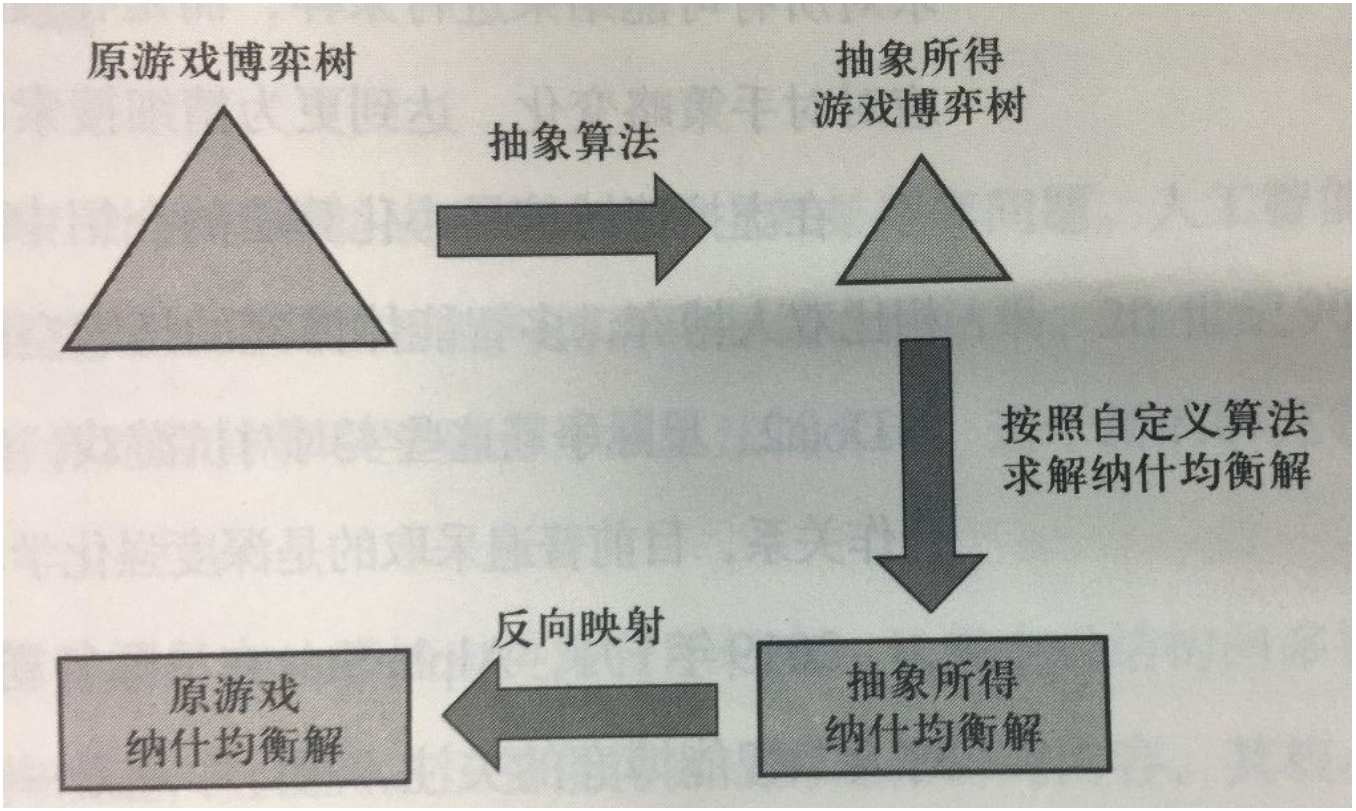


图8.9 求解非完全信息博弈纳什均衡的一般方法



非完全信息博弈的实际应用

- 深度强化学习越来越多的被应用于博弈的策略求解。强化学习的主要思想是通过试错和搜索来优化自身的行动[Sutton 2018]，这一过程可以直接被应用于博弈的决策中。在连续动态的环境中，通过强化学习的训练，可以让参与博弈的AI智能体在不同情况下能够找到可获得最大收益的最佳策略[Racanière 2017]，在许多游戏AI中，都运用了这种方法来进行对战，并最终战胜人类玩家[Jaderberg 2018] [Arulkumaran 2019]。
- 近年来，非完全信息下的德州扑克智能系统Libratus在与人类选手对弈比赛中屡战屡捷，德州扑克就是一种典型的非完全信息博弈。在进行游戏前，Libratus首先通过虚拟遗憾值最小化算法，反复进行迭代，进行自我博弈，锻炼进行游戏的能力。在实际对局中，Libratus在博弈中不断寻找安全子博弈，进一步缩小搜索空间，并找到子博弈的最佳策略。在每局比赛结束后，将游戏最终结果通过强化学习反馈给决策网络以进一步优化参数[Brown 2018]。

非完全信息博弈的实际应用

最近，Libratus被拓展到德州扑克多人博弈系统Pluribus，Pluribus在六人无限注德州扑克的较为简单场景下击败人类专业选手[Brown 2019]。在训练中，Pluribus 采用了蒙特卡洛虚拟遗憾最小化算法（Monte Carlo counterfactual regret minimization, MCCFR）。MCCFR随机考虑一部分行动，来选择应该采取的决定。在每一次迭代中，Pluribus根据在场玩家已经展示出来策略来模拟一盘游戏，然后在模拟游戏中寻找最适合自己的最优策略。每一回合，Pluribus都会加入一个虚拟遗憾值，使它会后悔上次没有用其他更好的策略，以保证Pluribus在下一轮倾向于选择上次后悔没选的策略。Pluribus就这样在“吾日三省吾身”中不断提高自己的水平。在实际对决了，为了应对对手可能会改变策略这一情况，Pluribus采用了有限前瞻搜索（Depth-limited search）方法，不苛求对所有可能结果进行采样，而是在搜索中可“截断搜索”、以自适应来应对对手策略变化，达到更为精细搜索的目的。