南京大学
人工智能学院
SCHOOL OF ARTIFICIAL INTELLIGENCE, NANJING UNIVERSITY

第四部分： 多Agent决策

章宗长

2023年5月9日

内容安排

协商

- 概览
- 对资源分割的协商
- **对任务分配的协商**

面向任务领域的协商（Task-Oriented Domain, TOD）

- 考虑下面的例子：

假设你有三个孩子，每天早晨需要把每个孩子送到不同的学校，你的邻居有四个孩子，也需要把他们送到学校。

每送一个孩子可以建模成一个不能再分解的任务。你和你的邻居可以进行协商，并且达成一致，这样对双方都有利（例如捎上另一个人的孩子到同一个目的地，节省他的路程）。

最坏的情况是你和你的邻居不能就送孩子建立汽车统筹达成一致，那么对自己来说并没有损失。

假设，你的一个孩子和邻居的一个孩子上同一所学校。显然，同时送两个孩子是合理的（即只有你的邻居或者你去送孩子，来完成这两个任务）。

通过这个例子可以看到，通过重新分配任务，Agent能有比仅执行自己的任务做得更好的潜力

面向任务领域的协商（TOD）

- 一个TOD是一个三元组$< T, Ag, c >$
  - T：可能任务的（有限）集合
  - $Ag = \{1, \ldots, n\}$：参与协商的Agent的（有限）集合
  - $c: 2^T \rightarrow \mathbb{R}^+$：费用函数，定义了执行每个$T$的子集的费用
    - 满足单调性：如果$T_1, T_2 \subseteq T$，并且$T_1 \subseteq T_2$，则$c(T_1) \leq c(T_2)$
    - 满足什么都不做的费用为零：$c(\emptyset) = 0$
- TOD的一次相遇是一个任务集合$< T_1, T_2, \ldots, T_n >$
  - 对每个 $i \in Ag$，有$T_i \subseteq T$
  - Agent之间通过重新分配任务来达成交易（deal）

交易

- 假设参与协商的两个Agent是$\{1, 2\}$
  - 给定相遇$< T_1, T_2 >$，交易把任务$T_1 \cup T_2$分配给Agent 1和Agent 2
  - 一个纯交易是$< D_1, D_2 >$，其中$D_1 \cup D_2 = T_1 \cup T_2$
  - 交易$\delta =< D_1, D_2 >$的语义： Agent 1承诺执行任务D_1，Agent 2承诺执行任务D_2
- 假设参与协商的n个Agent是$\{1, 2, \ldots, n\}$
  - 给定相遇$< T_1, T_2, \ldots, T_n >$，交易把任务$T_1 \cup T_2 \ldots \cup T_n$分配给$n$个Agent
  - 交易$\delta =< D_1, D_2, \ldots, D_n >$，其中$D_1 \cup D_2 \cup \cdots \cup D_n = T_1 \cup T_2 \cup \cdots \cup T_n$，语义：Agent i承诺执行任务D_i

交易中的效用

- 对Agent i而言，交易δ的效用

$$\text{utility}_i(\delta) = c(T_i) - \text{cost}_i(\delta)$$

交易δ的费用$c(D_i)$

- 一项交易的效用表示了Agent从这个交易中获得收益的多少
  - 正效用表示Agent从交易中获益；负效用表示与Agent单独完成在相遇中最初分配的任务相比，它将遭受损失
- 如果没有达成交易，那么执行最初分配的任务
  - 冲突交易Θ：最初分配的任务$< T_1, T_2 >$
  - $\forall i \in Ag, \text{utility}_i(\Theta) = 0$

仅讨论有两个Agent的情形

交易中的优势（Dominance）

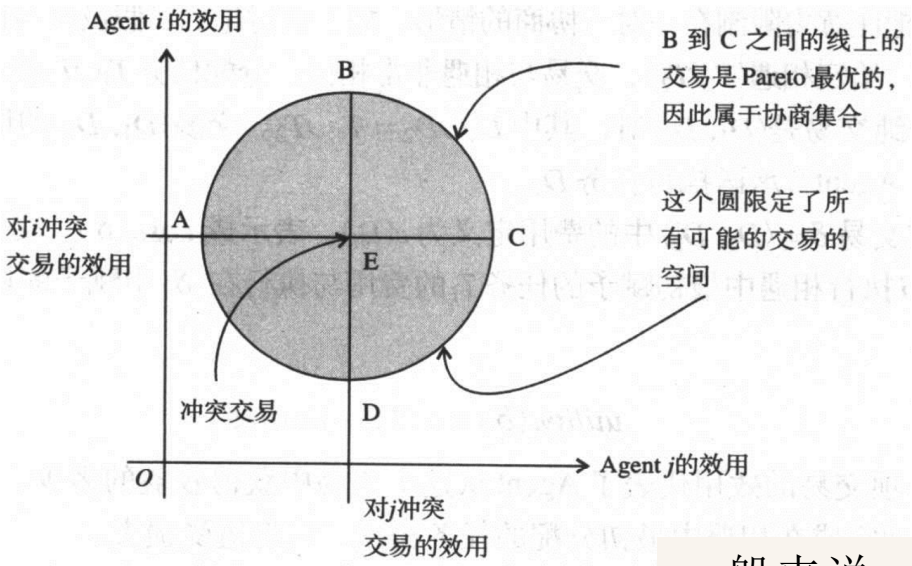
- 如果下列条件成立，则称δ_1优势于（dominate）δ_2 $(\delta_1 \succ \delta_2)$:
  - 对任意一个Agent，δ_1至少和δ_2一样好
    - $\forall i, \text{utility}_i(\delta_1) \geq \text{utility}_i(\delta_2)$
  - 至少存在一个Agent，使得δ_1比δ_2好
    - $\exists i, \text{utility}_i(\delta_1) > \text{utility}_i(\delta_2)$

如果至少第一个条件成立，则称δ_1弱优势于δ_2

- 当δ_1优势于δ_2，那么所有理性的Agent都会偏好δ_1
  - 如果没有任何一个交易优势于δ_1，则称δ_1是帕累托最优的
- 如果一个交易δ_1弱优势于冲突交易 $(\delta_1 \geq \Theta)$，那么这个交易是个体理性（individual rational）的

协商集合

- 协商集合由交易组成，这些交易具有以下性质：
  - 个体理性：Agent不会选择比冲突交易更差的交易
  - 帕累托最优：不存在另一个交易使得某个Agent在不损害别的Agent的情况下获得更高的效用



- 线段BD以左的交易对于Agent j来说不是个体理性的
- 线段AC以下的交易对于Agent i来说不是个体理性的
- 弧BC上的交易是个体理性且帕累托最优的，因此属于协商集合

一般来说，Agent i通过提出B点的交易开始协商，而Agent j通过提出C点的交易开始协商

单调让步协议（Monotonic Concession Protocol）

- 协商进行多轮
- 在第u轮协商中：
  - 两个Agent分别从协商集合中提出一项提议
  - 如果Agent发现另一个Agent提出的交易（弱）优势于他提出的交易，则达成一致
  - 如果没有达成一致，那么协商进入到下一轮
- 在第$u+1$轮协商中：
  - Agent不能提出比第u轮的提议对另一个Agent更差的提议
  - 如果没有Agent作出让步，则协商以交易冲突结束

单调让步协议（续）

- 使用单调让步协议，在有限轮的协商之后，可以保证协商结束
  - 最后一轮中，两个Agent达成一致或互不让步（产生交易冲突）
  - 不保证快速达成一致
    - 可能的交易数量为$O(2^{|T|})$，协商可能会进行的轮数与分配的任务数量呈指数关系

Zeuthen策略

协商的参与者在使用单调让步协议的时候应该如何工作？

- Agent的第一个提议应该是什么？
  - Agent最偏好的交易
- 在给定的一轮协商中，谁应该让步？
  - 度量Agent冒冲突风险的意愿，应该是最不愿意冒冲突风险的Agent进行让步
- 如果一个Agent让步，它应该让步多少？
  - 足够改变风险平衡的让步

如何度量冒风险的意愿

- 如果一个Agent当前提议的效用和冲突交易的效用差别小，则它更愿意冒冲突的风险

- Agent i在第t轮冒冲突风险的意愿：

$$risk_i^t = \frac{\text{由于让步并接受}j\text{的提议导致}i\text{的效用损失}}{\text{由于没有让步并导致冲突致使}i\text{的效用损失}}$$

  - 值在0~1之间
  - 值越大（越接近1），表示Agent i由冲突遭受的损失越小，因此更愿意冒冲突风险
  - 反之亦然

- Agent i在第t轮冒冲突风险的意愿：

$$risk_i^t = \frac{\text{由于让步并接受}j\text{的提议导致}i\text{的效用损失}}{\text{由于没有让步并导致冲突致使}i\text{的效用损失}}$$

- 形式地，有：

$$risk_i^t = \begin{cases} 1 & \text{如果} utility_i(\delta_i^t) = 0 \\ \dfrac{utility_i(\delta_i^t) - utility_i(\delta_j^t)}{utility_i(\delta_i^t)} & \text{其他情况} \end{cases}$$

- 当$utility_i(\delta_i^t) = 0$时，$risk_i^t$的值为1：
  - i完全愿意冒冲突风险，而不愿意做出让步

Zeuthen策略达到纳什均衡

- 让步多少？
  - 做出足够的让步即可，即一个Agent应该做最小的必要的让步，来改变风险的平衡
- 如果遇到风险相同的情况
  - 可以通过一个Agent“投掷硬币”决定谁应该让步
- 使用Zeuthen策略能达到纳什均衡

这对于自动Agent的设计者是特别有意义的。它消除了所有程序员对于保密部分的要求。一个Agent的策略可以公开，并且其他Agent的设计者无法通过选择不同的策略利用这个信息。事实上，为了避免不小心引起的冲突，公开策略的行为对于设计者来说是所希望的。

协商中的欺骗（Deception）

- 在面向任务领域的协商中，Agent可能通过两种类型的欺骗行为来获利：
  - 谎报任务
  - 隐瞒任务
- 谎报任务
  - 假装已经被分配到了一个任务，而它并没有分配到这个任务
  - 一种解决办法：保证分配给Agent执行的任务是可验证的
- 隐瞒任务
  - 假装并没有分配到一个已经被分配到了的任务

小结

- 协商是就共同关心的问题达成一致的过程
  - 协商参数：协商集合、协议、策略、规则
  - 通常进行多轮，每个Agent每一轮都给出提议
- 对资源分割的协商
  - 轮流出价模型：一对一的协商协议
  - 切蛋糕的例子：固定/不固定轮数、玩家有耐心/耐心有限
  - 协商决策函数
- 对任务分配的协商
  - 形式化定义、交易、（弱）优势、协商集合
  - 单调让步协议、Zeuthen策略

内容安排

不一致性

- 在多Agent系统中，Agent应该如何就相信什么达成一致？
  - 在法庭上，律师的立场应该是理性且有根据的，能够通过论证（argument）得到
    - 如果所有的证据是一致的，那么不会有争论
    - 通常会有不一致

- 不同Agent的不一致观点
  - 显式的情况：Agent 1相信p， Agent 2相信$\neg p$
  - 隐式的情况： Agent 1相信p和$p \longmapsto q$，Agent 2相信$\neg q$

辩论

- 辩论提供了解决不一致的原则性技术
  - 至少，提供了在面对不一致的时候，如何选择观点的合乎情理的规则

- 在面对观点p和$\neg p$时，应该相信哪个观点？
  - 选择$\emptyset$作为观点
  - 选择接受其中一个观点，放弃另一个

可能有多个理性的观点，选择哪个最好呢？

辩论的种类

- 逻辑模式
  - 同数学证明类似，倾向于自然演绎
- 情感模式
  - 当呼吁带有情感和态度等情况时
  - 例子：如果这件事发生在你身上，你会感觉怎么样？

- 本能模式
  - 人类辩论的自然和社会的方面
  - 参与辩论的一方跺着脚表示其感觉的强烈程度

- 神秘模式
  - 借助于直觉、隐秘现象、宗教等

辩论的方法

- 抽象辩论（Abstract Argumentation）
  - 检查论证如何共存
- 演绎辩论（Deductive Argumentation）
  - 利用逻辑演绎推理

- 区别
  - 抽象辩论的论证具有原子性
    - 没有结构，是不可分割的实体
  - 演绎辩论的论证具有逻辑结构
    - 能通过逻辑演绎进行推理

抽象辩论

- 不关心每个论证的内部结构，关心其整体结构

几个论证的例子：

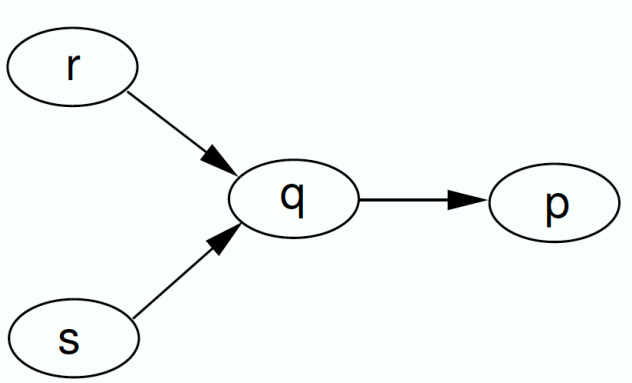
- p：由于天晴，我决定骑车出去
- q：由于今天是工作日并且我要工作，我不能骑车出去
- r：由于今天是节假日，我不用工作
- s：由于我今天请假了，我不用工作

- 抽象辩论由一些论证的集合和关系组成
  - 称这种形式的论证系统为Dung式抽象辩论系统
  - 论证由抽象符号表示
  - 不需要关注论证的具体意义

Dung式抽象辩论系统

- 用二元组$<\sum, \rhd>$表示Dung式抽象辩论系统
  - $\sum$：论证集合，由不同的论证组成
  - $\rhd$：论证集合$\sum$中论证之间攻击关系的集合
  - $(\varphi, \psi) \in \rhd$，读作：
    - 论证φ攻击论证ψ，φ是ψ的反例，φ是ψ的攻击者

- 例：$<\{p, q, r, s\}, \{(r, q), (s, q), (q, p)\}>$
  - 共有四个论证：p, q, r, s
  - 共有三个攻击
    - 论证r攻击论证q
    - 论证s攻击论证q
    - 论证q攻击论证p



立场（Position）

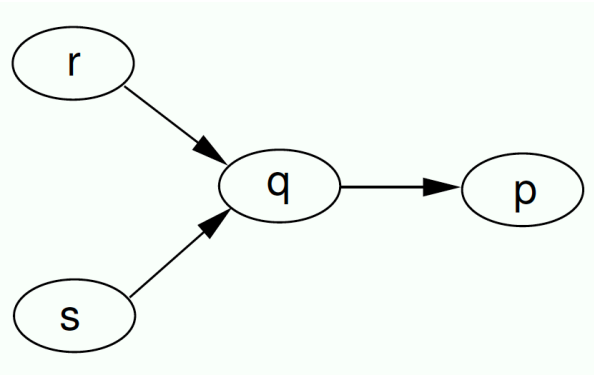
- 例：$<\sum, \triangleright>=<\{p, q, r, s\}, \{(r, q), (s, q), (q, p)\}>$

p：由于天晴，我决定骑车出去

q：由于今天是工作日并且我要工作，我不能骑车出去

r：由于今天是节假日，我不用工作

s：由于我今天请假了，我不用工作



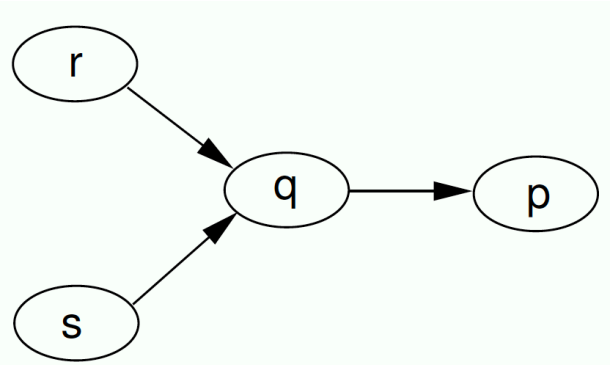
问题：给定$<\sum, \triangleright>$，我应该相信哪些论证呢？

- 立场$S \subseteq \sum$为一些论证构成的集合
  - 立场可以是不一致的，它只是选择了一些论证

无冲突（Conflict Free）的立场

- 如果立场S中没有论证攻击S中的其他论证，那么S是无冲突的
  - 如果论证a被论证a'攻击了，那么如果存在a''攻击a'，则a获得了a''的辩护

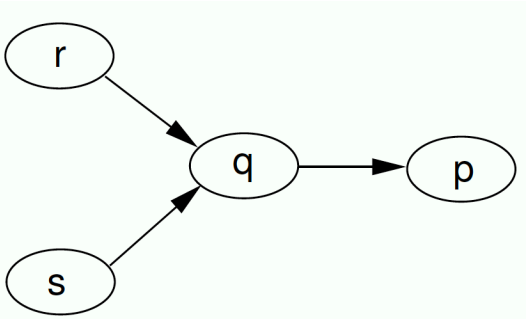
一个无冲突的立场中各个论证之间没有不一致



- 右图：无冲突的立场为
  - $\emptyset, \{q\}, \{p\}, \{r\}, \{s\}, \{r, s\}, \{p, r\}, \{p, s\}, \{r, s, p\}$
  - p获得了r和s的辩护

互相辩护（Mutually Defensive）的立场

- 如果S的每一个被攻击的论证被S中的一些论证辩护，那么S是互相辩护的
  - 论证可以自己辩护自己



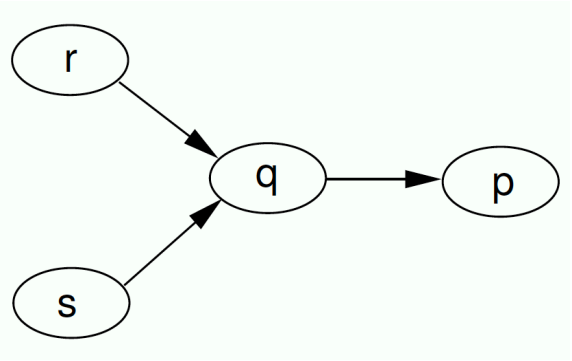
- 右图：互相辩护的立场为
  - $\emptyset, \{r\}, \{s\}, \{r, s\}, \{p, r\}, \{p, s\}, \{r, s, p\}$
  - 例子：$\{p, r\}$是互相辩护的，因为如果加入论证p的攻击论证q，那么论证r会为它辩护
- 下列立场不是互相辩护的
  - $\{p\}, \{q\}$

例子：$\{p\}$不是互相辩护的，因为如果加入论证p的攻击论证q，那么没有论证会为它辩护

可采纳的（Admissible）立场

- 如果一个立场是无冲突并且互相辩护的，那么这个立场是可采纳的
  - 无冲突：没有一个论证攻击另一个论证
  - 互相辩护：如果一个论证被攻击，那么这个论证被另一个论证辩护

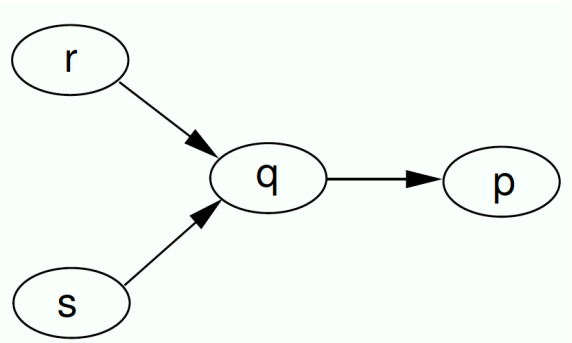
- 下列立场是可采纳的
  - $\emptyset, \{r\}, \{s\}, \{r, s\}, \{p, r\}, \{p, s\}, \{r, s, p\}$



- 可采纳性是合理立场的最小限度
  - 内部一致，并且自己能对攻击进行辩护

偏好拓展（Preferred Extension）

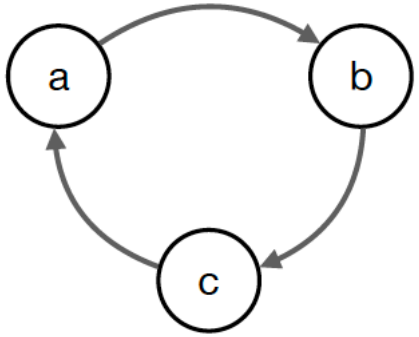
- 偏好拓展是最大的可采纳集合
  - 加入另一个论证会使得这个立场变得不可采纳
- 如果立场S是可采纳的，并且任何S的超集都不是可采纳的，那么S是一个偏好拓展
  - $\{p, r, s\}$是偏好拓展
    - 加入q会使得立场变得不可采纳



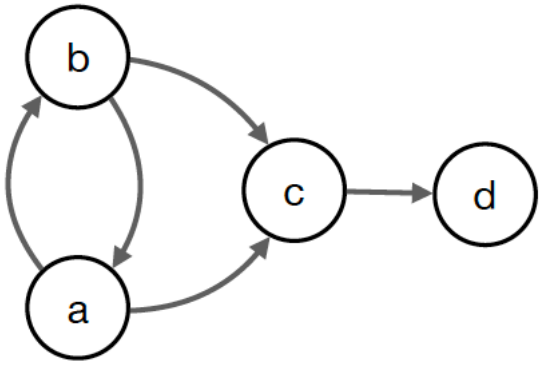
- 一个论证集合必定有一个偏好拓展
  - $\emptyset$是一个可采纳的立场
  - 如果没有别的可采纳的立场，那么$\emptyset$将是最大的可采纳集合

示例：偏好拓展

两个论证互相攻击
有两个偏好拓展，分别为$\{a\},\{b\}$



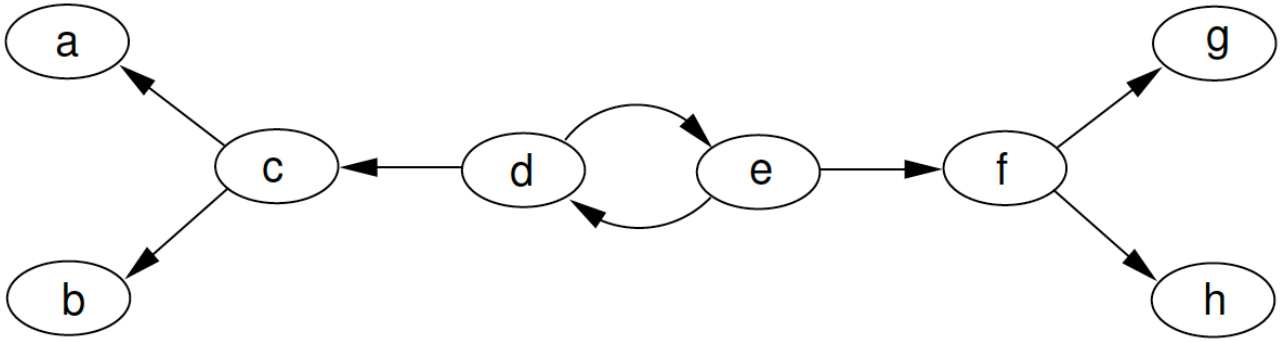
奇数个论证，呈环形互相攻击
偏好拓展为$\emptyset$



a和b互相攻击，由于他们都攻击c，d受到了辩护
有两个偏好拓展，分别为$\{a,d\},\{b,d\}$

示例：偏好拓展

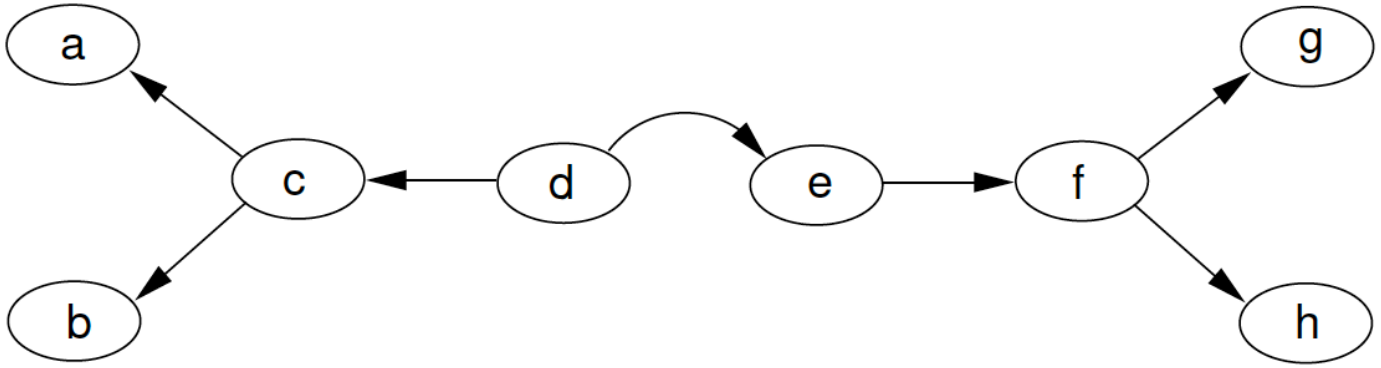
- 随着论证数量的增加，找到偏好拓展的难度呈指数增长
  - 大小为n的论证集合拥有2^n个可能的立场



- $\{a, b, d, f\}$和$\{c, e, g, h\}$为偏好拓展
  - d和e互相攻击，因此最多有两个偏好拓展，由d和e的攻击关系决定

示例：偏好拓展

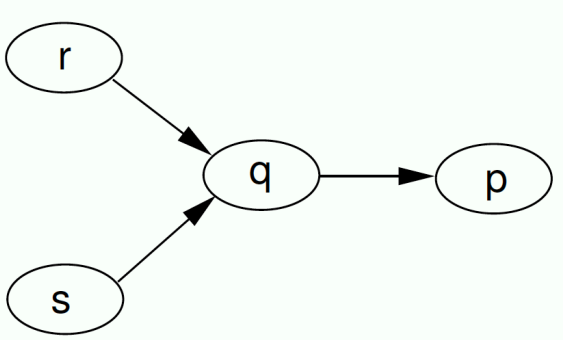
- e不攻击d的情况



- 仅有$\{a, b, d, f\}$为偏好拓展
  - 当c和e受到d攻击时，没有论证为它们辩护
  - c和e都不会出现在可采纳的集合中

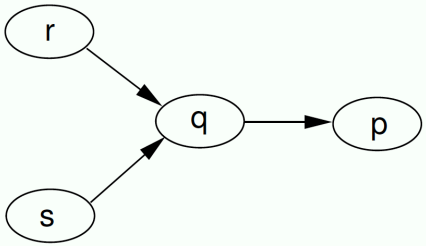
轻信的（Credulous）和怀疑的（Sceptical）接受

- 如果一个论证是每一个偏好拓展的元素，那么这个论证是被怀疑接受的
- 如果一个论证是至少一个偏好拓展的元素，那么这个论证是被轻信接受的
- 任何一个被怀疑接受的论证都是被轻信接受的
  - $\{p, r, s\}$是偏好拓展
  - p、r、s是被怀疑接受的
  - q既不被轻信接受也不被怀疑接受



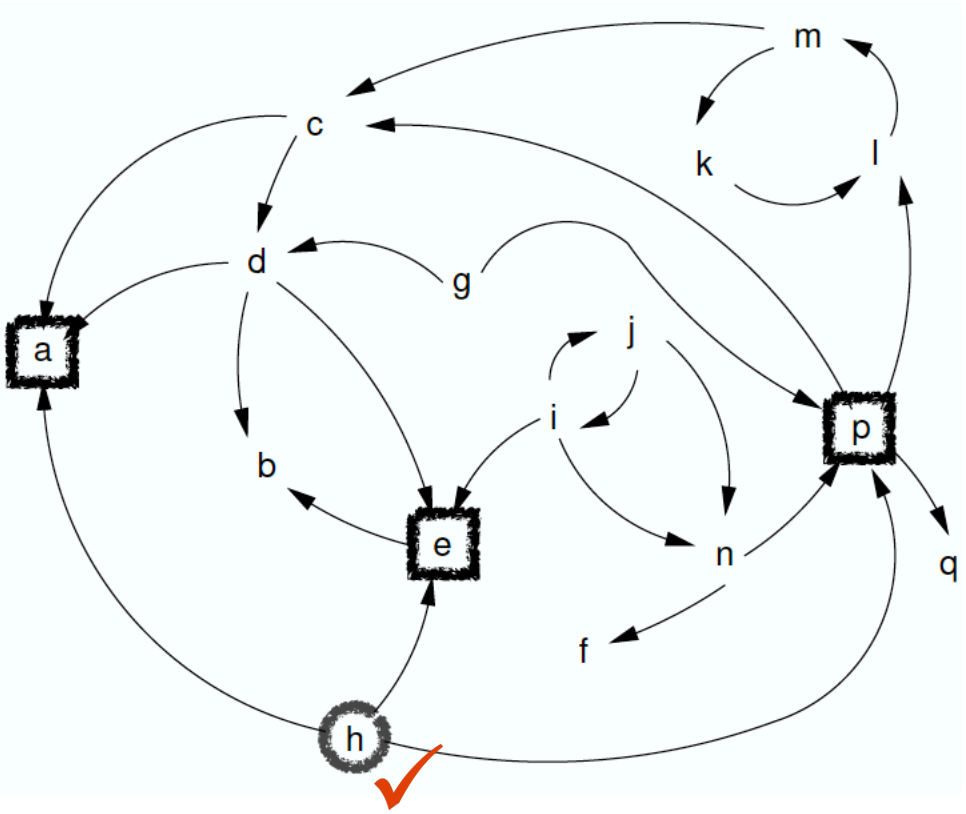
理性拓展（Grounded Extension）

- 理性拓展是最不可能被质疑（Least Questionable）的论证的集合
  - 只接受无法避免接受的论证
  - 只拒绝无法避免拒绝的论证
- 理性拓展的构造
  - 如果论证没有被攻击，那么它会被接受
  - 如果论证被接受的论证攻击，那么它不能够被接受
  - 删除不被接受的论证，直到接受的论证集合不再变化
- 理性拓展是最终被接受的论证集合
  - 右图：理性拓展为$\{r, s, p\}$



示例：理性拓展

- 求右图所示的Dung式抽象辩论系统的理性拓展

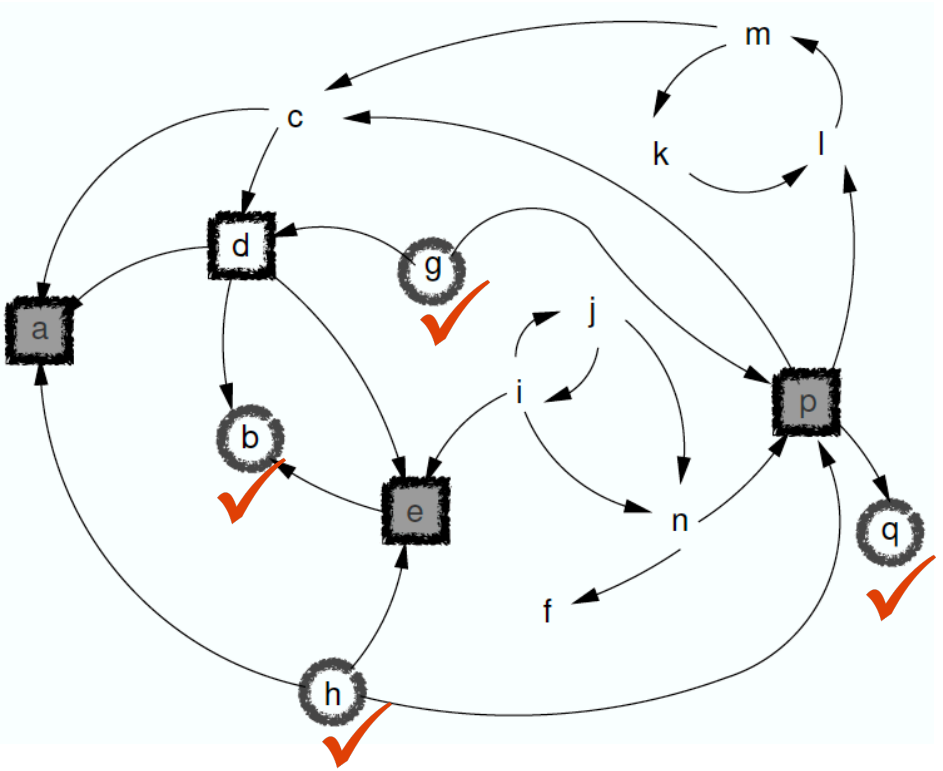


- h没有被攻击，因此h会被接受
  - h攻击了a，a不被接受
  - h攻击了p，p不被接受
  - h攻击了e，e不被接受

示例：理性拓展

- p没有被接受，并且q只受到了p的攻击，所以q会被接受

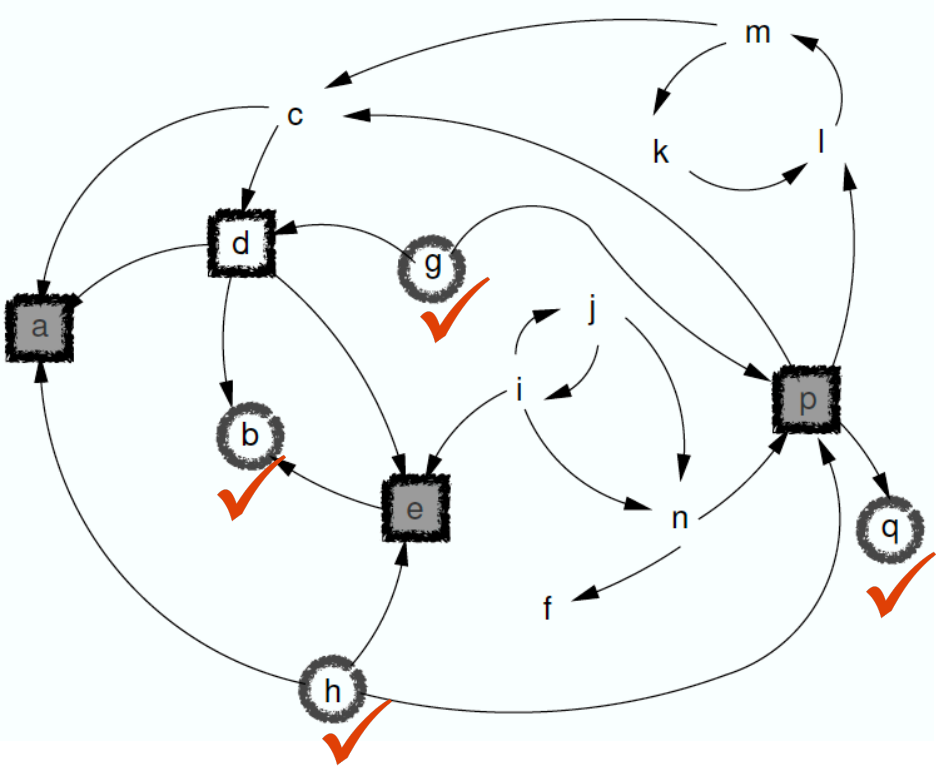
- g没有被攻击，所以g会被接受
  - g攻击d，d不被接受
  - g攻击p（p也被h攻击了），所以p不被接受



- b不再被攻击，因此b会被接受

示例：理性拓展

- 不能确定是否接受以下论证
  - m、k、l：它们以环形互相攻击对方
  - i、j：它们互相攻击
  - n：不确定i、j的状态
  - f：不确定n的状态



- 理性拓展为$\{b, g, h, q\}$

小结

- 辩论提供了解决不一致的原则性技术
  - 种类：逻辑模式、情感模式、本能模式、神秘模式
  - 方法：抽象辩论、演绎辩论
    - 抽象辩论的论证具有原子性
    - 演绎辩论的论证具有逻辑结构
- 抽象辩论
  - 不关心每个论证的内部结构，关心其整体结构
  - Dung式抽象辩论系统：二元组$<\sum, \triangleright>$
    - $\sum$：论证集合，由不同的论证组成
    - $\triangleright$：论证集合$\sum$中论证之间攻击关系的集合
    - 立场：无冲突的立场、互相辩护的立场、可采纳的立场
    - 拓展：偏好拓展、理性拓展

内容安排

分布式规划

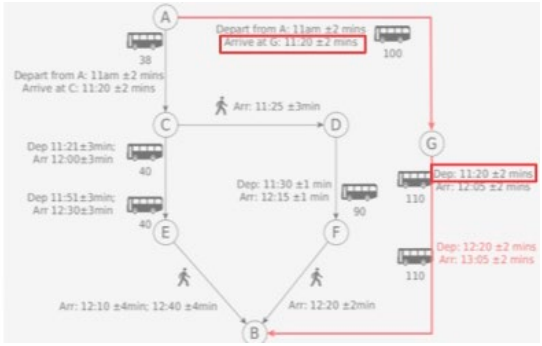
- **规划的基础知识**
- 分布式规划的决策模型
- 分布式规划的离线算法
- 分布式规划的在线算法

规划

- 研究源于20世纪60年代前后，是人工智能的一个重要领域

- 两大任务
  - 问题描述：如何方便地表示规划问题
  - 问题求解：如何高效地求解规划问题

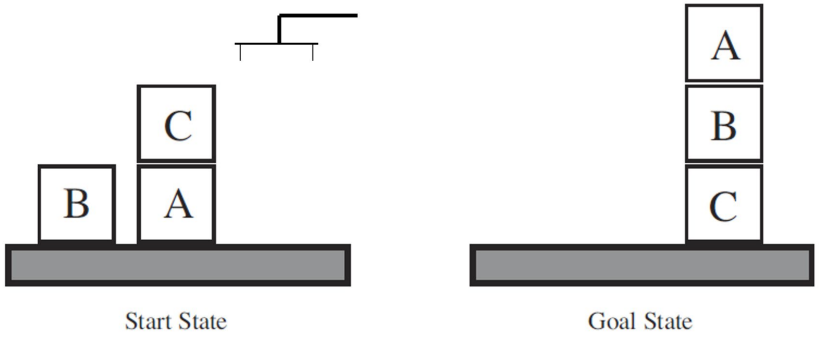
- 应用：智能机器人、后勤调度、自动驾驶等领域



经典规划

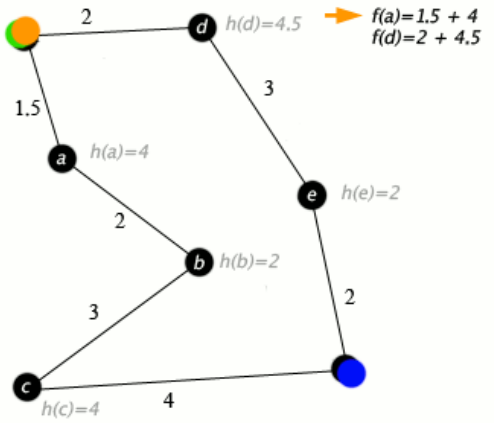
- 经典规划的基本假设
  - （A0）有限系统：问题只涉及有限的状态、动作、事件等
  - （A1）完全可观察：总知道当前所在的状态
  - （A2）确定性：每个动作只会导致一种确定的影响
  - （A3）静态性：不存在外部动作，环境所有的改变都来自Agent的动作
  - （A4）状态目标：目标是一些需要达到的目标状态
  - （A5）序列规划：规划结果是一个线性动作序列
  - （A6）隐含时间：不考虑时间连续性
  - （A7）离线规划：规划求解器不考虑执行时的状态

经典规划



- 典型的问题：积木世界
- 问题描述
  - 集合描述：使用有限的命题符号集合
  - 经典描述：使用一阶逻辑符号
- 求解方法分为状态空间的求解和规划空间的求解
- 状态空间搜索
  - 在状态转移图中搜索从初始状态到目标状态的一条路径
  - 前向搜索、后向搜索、启发式搜索
- 规划空间搜索
  - 用找缺陷的方法对规划求精，直到规划可执行
  - 偏序规划

A*算法



- 一种启发式搜索方法
- 存储两张表
  - OPEN表：保存所有已生成而未访问过的节点
  - CLOSE表：记录已访问过的节点
- 特点：OPEN表中的每个节点n都有一个优先值$f(n)$

越小优先级越高

$$f(n) = g(n) + h(n)$$

起始节点到节点n的成本函数

节点n到目标节点的启发式函数

- 可容纳最优：启发式函数满足$0 \le h(n) \le h^*(n)$

节点n到目标节点的最优值

概率规划

- 基于概率模型和效用函数，制定一系列的理性决策
  - 使用最大化期望效用原则
  - 多步决策：在计算理性决策时，要求推理未来的动作和观察序列

- 问题描述
  - 马尔可夫决策过程（Markov Decision Process, MDP）
  - 部分可观察的MDP（Partially Observable MDP, POMDP）
  - 分布式POMDP（Decentralized POMDP, Dec-POMDP）
  - ……

- MDP/POMDP/Dec-POMDP规划问题的求解方法
  - 离线规划：动态规划
  - 在线规划：蒙特卡洛树搜索

MDP模型

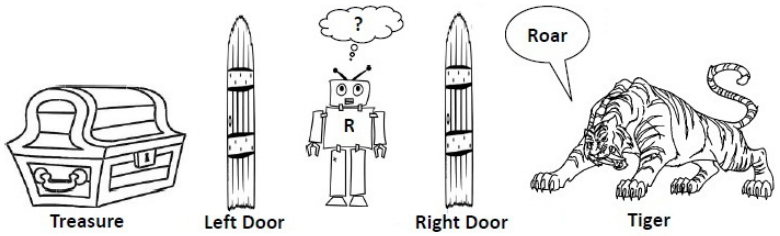
- 一个MDP问题可以形式化地建模为
  - 有限的状态集合$\mathcal{S}$
  - 有限的动作集合$\mathcal{A}$
  - 状态转移函数$T(s'|s,a)$
    - Agent在状态s执行动作a转移到新的状态s'的概率
  - 奖励函数$R(s,a)$
    - Agent在状态s执行动作a所能得到的即时奖励
- MDP模型的特点
  - 考虑了状态转移的不确定性
  - Agent可以直接获得环境的状态信息

POMDP模型

- 一个POMDP问题可以形式化地建模为
  - 有限的状态集合$\mathcal{S}$ MDP
  - 有限的动作集合$\mathcal{A}$
  - 状态转移函数$T(s'|s,a)$
  - 奖励函数$R(s,a)$
  - 有限的观察集合$\mathcal{O}$
  - 观察函数$\Omega(o|s',a)$
    - Agent执行动作a转移到新的状态s'后，获得观察o的概率
- POMDP模型的特点
  - Agent不能直接获得环境的状态信息，其对于状态的观察来源于传感器收集到的带噪声的局部信息

一个简单的POMDP问题：老虎问题

- 老虎问题：通过听老虎的叫声判断它的位置，尽可能打开没有关老虎的那扇门
  - 状态集合$\mathcal{S} = \{\text{Tiger}_{\text{Left}}, \text{Tiger}_{\text{Right}}\}$
  - 动作集合$\mathcal{A} = \{\text{Open}_{\text{Left}}, \text{Open}_{\text{Right}}, \text{Listen}\}$
  - 观察集合$\mathcal{O} = \{\text{Roar}_{\text{Left}}, \text{Roar}_{\text{Right}}\}$



- 状态转移规则
  - Agent执行动作Listen后，老虎的位置不变
  - Agent执行其他动作后，老虎会等概率地放置在两扇门后

状态转移函数

$$T(\text{Tiger}_{\text{Left}} \mid \text{Tiger}_{\text{Left}}, \text{Listen}) = 1.0, T(\text{Tiger}_{\text{Right}} \mid \text{Tiger}_{\text{Right}}, \text{Listen}) = 1.0$$
$$T(* \mid *, \text{Open}_{\text{Left}}) = 0.5, T(* \mid *, \text{Open}_{\text{Right}}) = 0.5$$

- 叫声是老虎在哪扇门后的有噪声的信号
  - 当老虎在左边门后时，Agent会有85%的概率观察到$\mathrm{Roar}_{\mathrm{Left}}$，还有15%的概率会观察到$\mathrm{Roar}_{\mathrm{Right}}$；反之亦然

观察函数

$$\Omega(\mathrm{Roar}_{\mathrm{Left}} \mid \mathrm{Tiger}_{\mathrm{Left}}, \mathrm{Listen}) = 0.85, \Omega(\mathrm{Roar}_{\mathrm{Right}} \mid \mathrm{Tiger}_{\mathrm{Left}}, \mathrm{Listen}) = 0.15$$
$$\Omega(\mathrm{Roar}_{\mathrm{Right}} \mid \mathrm{Tiger}_{\mathrm{Right}}, \mathrm{Listen}) = 0.85, \Omega(\mathrm{Roar}_{\mathrm{Left}} \mid \mathrm{Tiger}_{\mathrm{Right}}, \mathrm{Listen}) = 0.15$$
$$\Omega(* \mid *, \mathrm{Open}_{\mathrm{Left}}) = 0.5, \Omega(* \mid *, \mathrm{Open}_{\mathrm{Right}}) = 0.5$$

- 听一次的惩罚为-1
- 打开有财宝的门的奖励为10，打开有老虎的门的惩罚为-100

奖励函数

$$R(*, \mathrm{Listen}) = -1$$
$$R(\mathrm{Tiger}_{\mathrm{Left}}, \mathrm{Open}_{\mathrm{Right}}) = 10, R(\mathrm{Tiger}_{\mathrm{Right}}, \mathrm{Open}_{\mathrm{Left}}) = 10$$
$$R(\mathrm{Tiger}_{\mathrm{Left}}, \mathrm{Open}_{\mathrm{Left}}) = -100, R(\mathrm{Tiger}_{\mathrm{Right}}, \mathrm{Open}_{\mathrm{Right}}) = -100$$

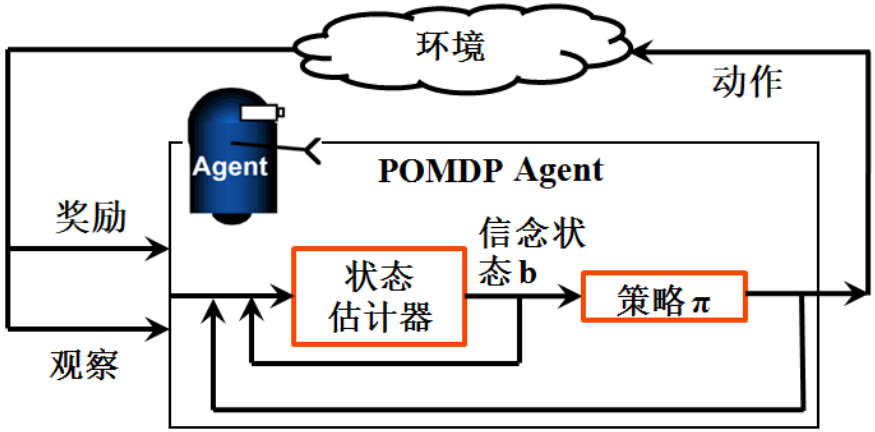
信念状态

- Agent需要依赖过去动作和观察序列的完整历史信息来选择理想的动作
- 信念状态：表征与决策有关的、过去动作和观察序列的完整历史信息
- 信念状态b：定义在状态集合S上的向量
  - $b_t(s)$：在t时刻，Agent在状态s的概率

$$b_t(s) = P(s_t = s \mid o_t, a_{t-1}, o_{t-1}, \dots, a_0, b_0)$$

  - 初始信念状态b_0：Agent在时刻$t = 0$的初始状态概率分布
- 对所有状态$s \in S$，均有$b(s) \in [0,1]$，且$\sum_{s \in S} b(s) = 1$

MDP vs. POMDP



策略：状态s到动作a的映射

值函数$V^{\pi}(s)$：由状态s开始，执行策略π所能获得的期望折扣回报

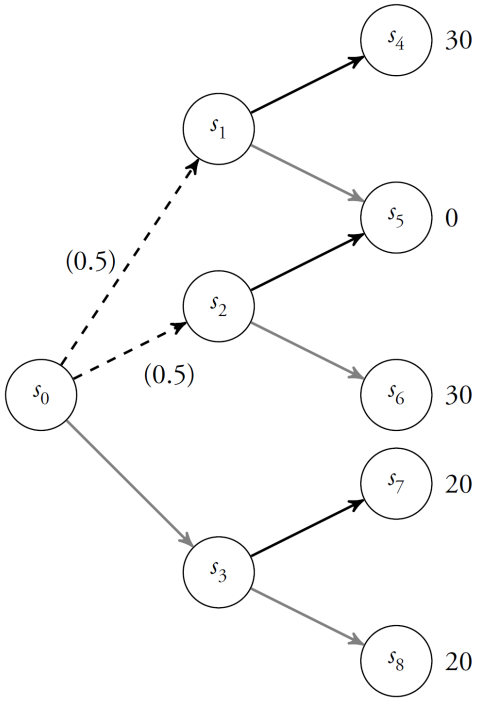
$$\mathbb{E}_{\pi}\left[\sum_{t=0}^{T-1}\gamma^{t}R(s_t,\pi(s_t))\mid s_0=s\right]$$

策略：信念状态b到动作a的映射

值函数$V^{\pi}(b)$：由信念状态b开始，执行策略π所能获得的期望折扣回报

$$\mathbb{E}_{\pi}\left[\sum_{t=0}^{T-1}\gamma^{t}R(b_t,\pi(b_t))\mid b_0=b\right]$$

开环规划



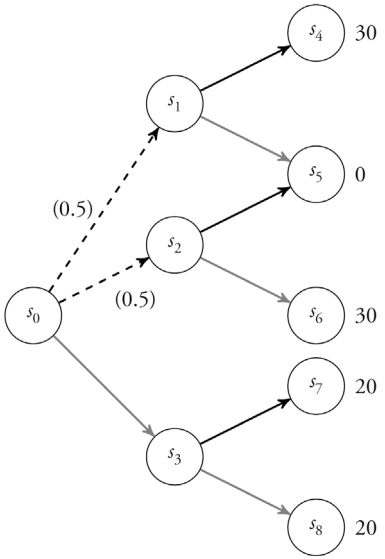
- 开环规划：不考虑未来状态信息
  - 如：很多路径规划算法
  - 得到静态的动作序列
  - 计算开销较小，仅能获得次优解
- 示例：开环规划的次优性
  - 9个状态，起始状态s_0
  - 两个决策步，每步决定向上走（up）还是向下走（down）
  - 有4个开环序列：
    - (up, up), (up, down), (down, up), (down, down)
    - 期望效用：
    - 在s_0处的最优动作是down

- $U(\text{up, up}) = 0.5 \times 30 + 0.5 \times 0 = 15$
- $U(\text{up, down}) = 0.5 \times 0 + 0.5 \times 30 = 15$
- $U(\text{down, up}) = 20$
- $U(\text{down, down}) = 20$

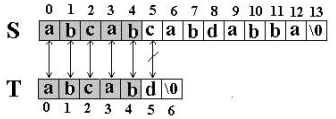
闭环规划

- 闭环规划：考虑未来状态信息
  - 如：**动态规划**
  - 得到反应式的策略，能对动作的不同结果做出不同反应
  - 计算开销较大，能获得近似最优解
  - 在行动效果不确定的序贯决策问题中，闭环规划更有优势

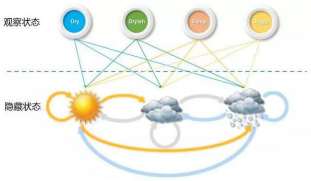
- 示例：闭环规划的最优性
  - 根据执行第一个动作后所观察到的结果来选择下一个动作
  - 在s_0处往上走，根据是到了s_1还是s_2来选择向上还是向下，从而保证得到30的奖励

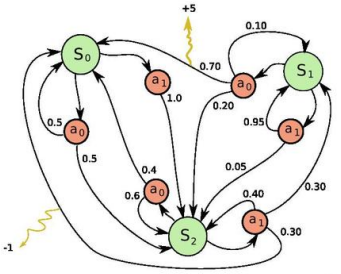


动态规划



- 动态规划是一种通用的技术
  - 计算斐波那契数列
  - 计算两个字符串的最长子串匹配
  - 计算隐马尔可夫模型的最可能状态序列
  - 求解MDP/POMDP/Dec-POMDP规划问题的最优策略





- 要素
  - 最优子结构：将原问题分解成多个子问题，如果知道了子问题的解，就很容易知道原问题的解
  - 重叠子问题：分解得到的多个子问题中，有很多子问题是相同的，不需要重复计算

课后作业4-12

- 假设一个资源的价值为1，两个Agent通过轮流出价的协商协议把它分成两份，每份的价值在0到1之间，这两份的价值总和为1。如果协商的轮数不固定，那么Agent 1在第0轮应该如何出价？并解释为什么。请分两个Agent都是有耐心的玩家和耐心有限的玩家这样两种情况分别讨论。

课后作业4-13

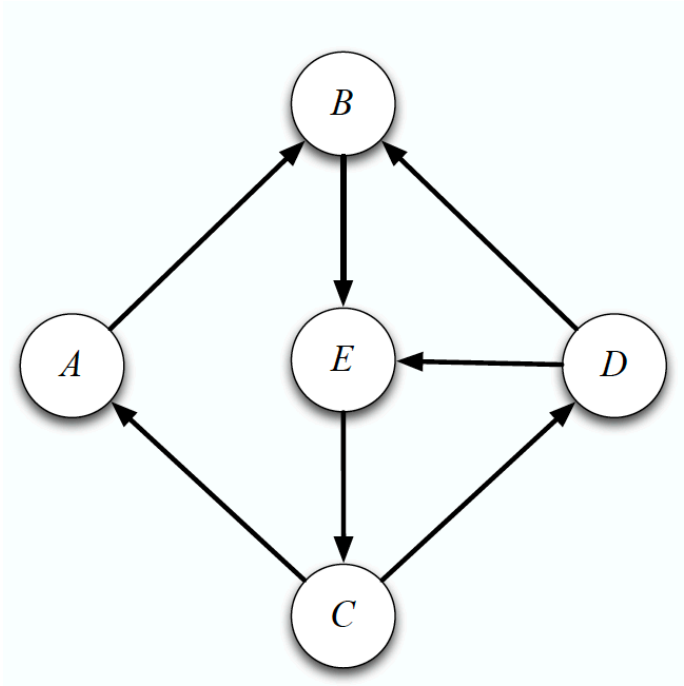
- 简述轮流出价协议的规则，单调让步协议的规则。

课后作业4-14

- 简述在使用单调让步协议进行协商时，使用Zeuthen策略的协商参与者是如何解决下面三个问题的：
  （1）Agent的第一个提议应该是什么？
  （2）在给定的一轮协商中，谁应该让步？
  （3）如果一个Agent让步，它应该让步多少？

课后作业4-15

- 给定如下图所示的Dung式抽象辩论系统。

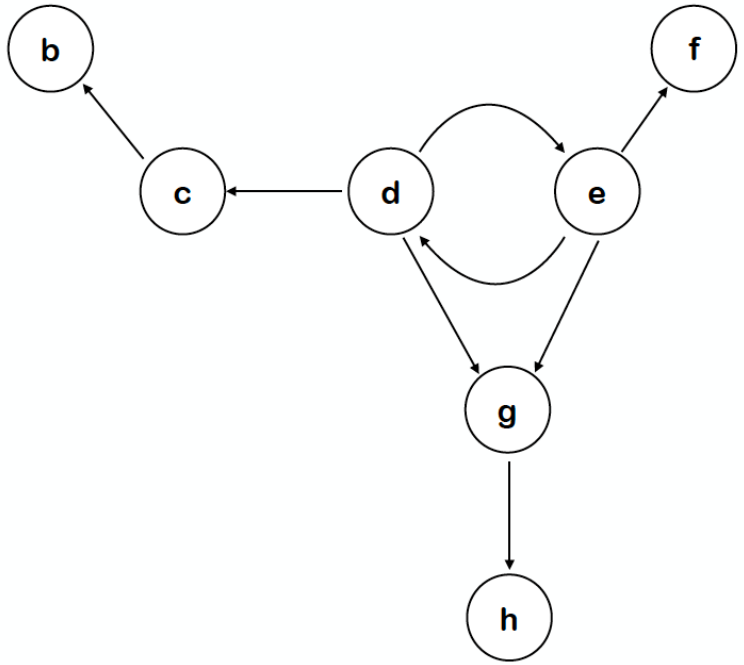


请写出以下内容：

- 无冲突的立场
- 互相辩护的立场
- 可采纳的立场
- 偏好拓展
- 轻信接受的论证集合
- 怀疑接受的论证集合
- 理性拓展

课后作业4-16

- 给定如下图所示的Dung式抽象辩论系统。



请写出以下内容：

- 可采纳的立场
- 偏好拓展
- 轻信接受的论证集合
- 怀疑接受的论证集合
- 理性拓展