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

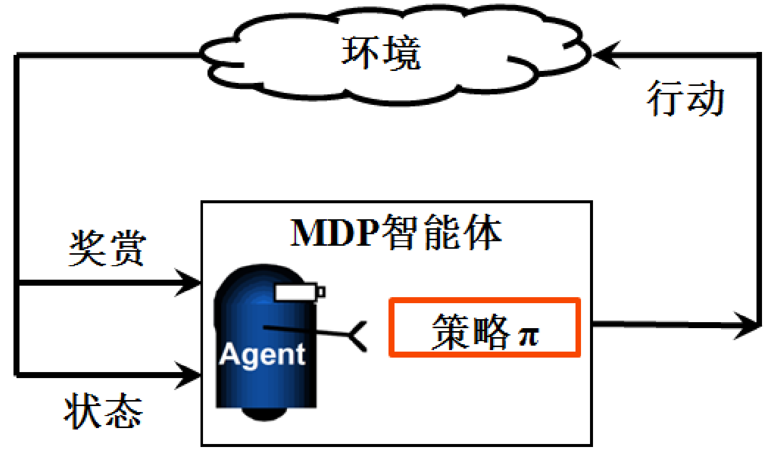
第六部分：部分可观察环境中的序贯决策系统

章宗长

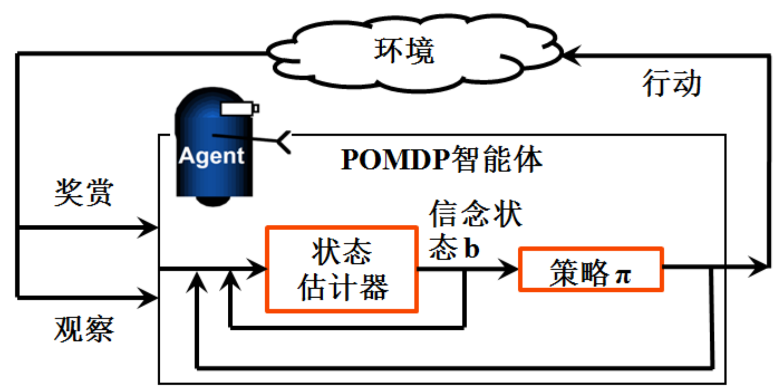
2021年6月9日

状态不确定性

- 前两部分讨论的是完全可观察MDP环境中的Agent
  - Agent准确知道当前状态



- 可能不能完美观察到状态
  - 传感器的限制或噪声
- 部分可观察的MDP（Partially Observable MDP, POMDP）
  - 有观察模型的MDP
  - 观察模型：采取行动a到达状态s'后得到观察o的概率
  - 计算（近似）最优策略的方法



内容安排

- 部分可观察的MDP
- 信念状态更新
- 精确求解方法
- 离线方法
- 在线方法
- 应用案例：飞行器避碰系统

部分可观察的MDP

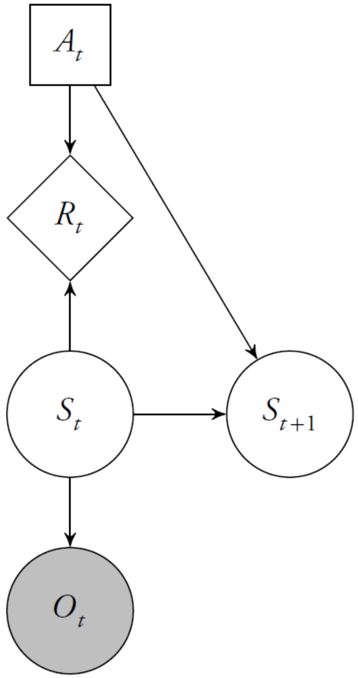
- 模型定义及示例
- 信念状态、策略、值函数

部分可观察的MDP（POMDP）

- POMDP
  - 状态空间$\mathcal{S}$
  - 行动空间$\mathcal{A}$
  - 状态转移函数$P(S_{t+1} \mid S_t, A_t)$
  - 奖赏函数$P(R_t \mid S_t, A_t)$

  MDP

  - 观察空间$\mathcal{O}$
  - 观察函数
    - 形式1： $P(O_{t+1} \mid S_{t+1}, A_t)$
    - 形式2： $P(O_t \mid S_t)$

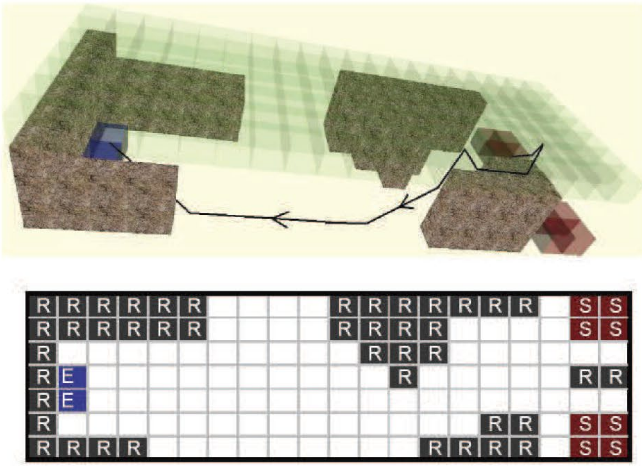


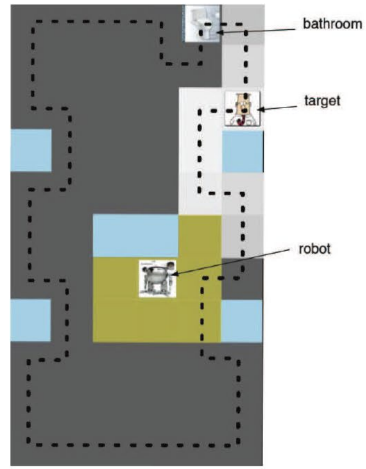
POMDP问题的结构

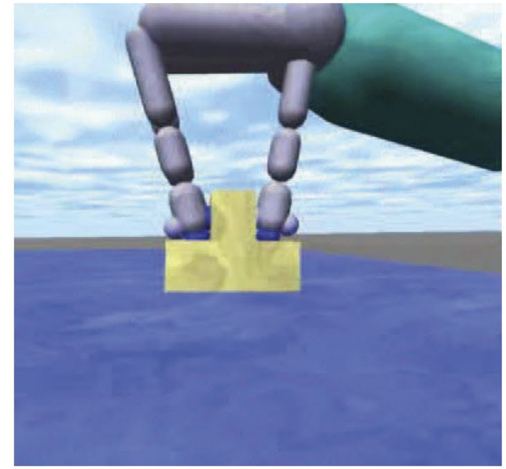
使用第2种形式的观察函数

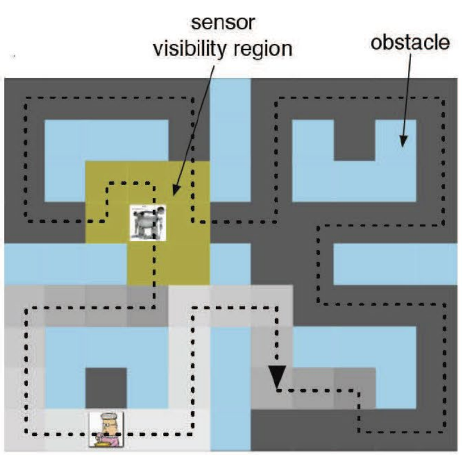
- 部分可观察性
  - 传感器仅能观察到环境的部分状态：多个不同状态的传感器数据相同
  - 传感器的缺陷：多次测量同一个状态获得的传感器数据也可能不同
  - 感知重名问题：不同的真实状态往往对应于同一个观察结果

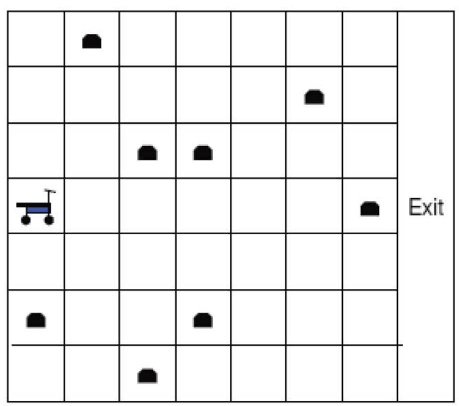
POMDP的应用

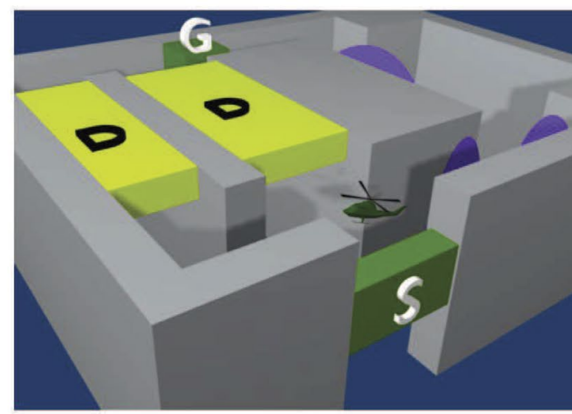



bathroom
target
robot




sensor
visibility region
obstacle


Exit


G
D
D
S

稳态POMDPs

- 状态转移函数$P(S_{t+1} \mid S_t, A_t)$、奖赏函数$P(R_t \mid S_t, A_t)$ 、观察函数$P(O_{t+1} \mid S_{t+1}, A_t)$（或者$P(O_t \mid S_t)$）不随时间发生变化
- 状态转移函数

$$T(s' \mid s, a) = p(s' \mid s, a) = P(S_{t+1} = s' \mid S_t = s, A_t = a)$$

- 奖赏函数

$$p(r \mid s, a) = P(R_t = r \mid S_t, = s\ A_t = a)$$

- 给定“状态-行动”的期望奖赏函数

$$R(s, a) = \sum_{r \in \mathcal{R}} r \cdot p(r \mid s, a)$$

稳态POMDPs（续）

- 观察函数

$$O(o \mid s', a) = P(O_{t+1} = o \mid S_{t+1} = s', A_t = a)$$

或者

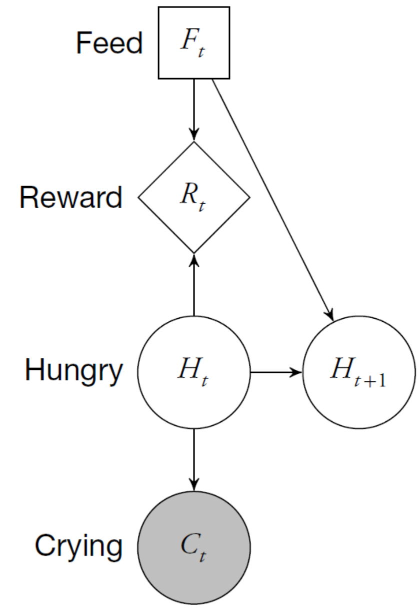
$$O(o \mid s) = P(O_t = o \mid S_t = s)$$

POMDP示例：啼哭婴儿

- 照看婴儿：基于婴儿是否在啼哭来决定什么时候给婴儿喂食物
  - 状态空间$\mathcal{S} = \{s_{饿}, s_{不饿}\}$
  - 行动空间$\mathcal{A} = \{a_{喂}, a_{不喂}\}$
  - 观察空间$\mathcal{O} = \{o_{哭}, o_{不哭}\}$
- 啼哭是“婴儿饿了”的有噪声的信号
  - 当婴儿不饿时，有10%的概率会啼哭
  - 当婴儿饿了时，有80%的概率会啼哭

观察函数

$$O\left(o_{哭} \mid s_{不饿}\right) = 0.1, O\left(o_{不哭} \mid s_{不饿}\right) = 0.9$$
$$O\left(o_{哭} \mid s_{饿}\right) = 0.8, O\left(o_{不哭} \mid s_{饿}\right) = 0.2$$



啼哭婴儿问题的结构

POMDP示例：啼哭婴儿（续）

- 状态转移规则
  - 给婴儿喂食物，在下一个时刻，婴儿会不饿
  - 如果婴儿不饿，且没给婴儿喂食物，则在下一个时刻，婴儿有10%的概率会饿
  - 一旦婴儿饿了，在喂食物之前，婴儿会一直饿

状态转移函数

$$T\left(s_{\text{不饿}} \mid *, a_{\text{喂}}\right) = 1.0$$

$$T\left(s_{\text{饿}} \mid s_{\text{不饿}}, a_{\text{不喂}}\right) = 0.1, \qquad T\left(s_{\text{不饿}} \mid s_{\text{不饿}}, a_{\text{不喂}}\right) = 0.9$$

$$T\left(s_{\text{饿}} \mid s_{\text{饿}}, a_{\text{不喂}}\right) = 1.0$$

POMDP示例：啼哭婴儿（续）

- 给婴儿喂食物的成本是5，婴儿饿了的成本是10
  - 成本可以相加：如果在婴儿饿了的时候喂食物，则成本是15

期望奖赏函数

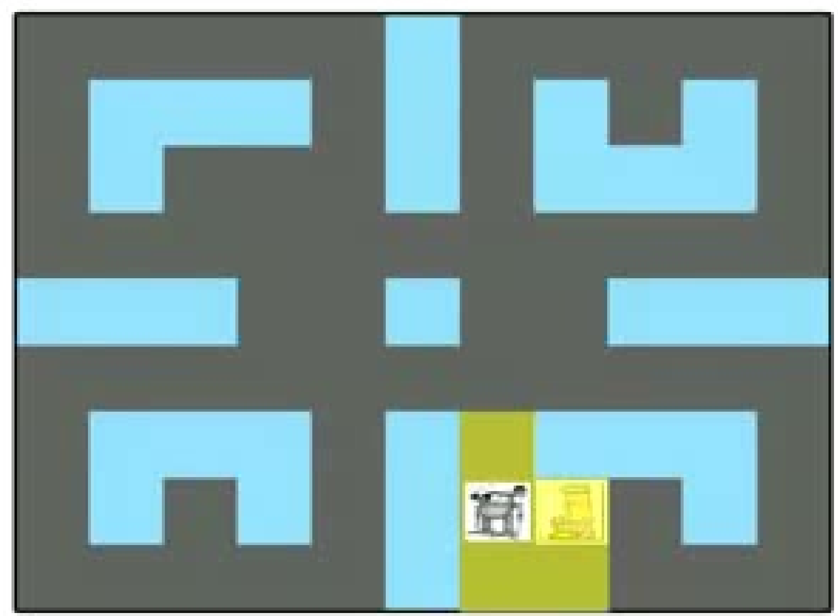
$$R\left(s_{\text{不饿}}, a_{\text{喂}}\right) = -5$$
$$R\left(s_{\text{饿}}, a_{\text{不喂}}\right) = -10$$
$$R\left(s_{\text{饿}}, a_{\text{喂}}\right) = -15$$

其他情况，期望奖赏为0

- 想找出一个折扣因子为0.9的无限步数的问题的最优策略

POMDP示例：目标跟踪

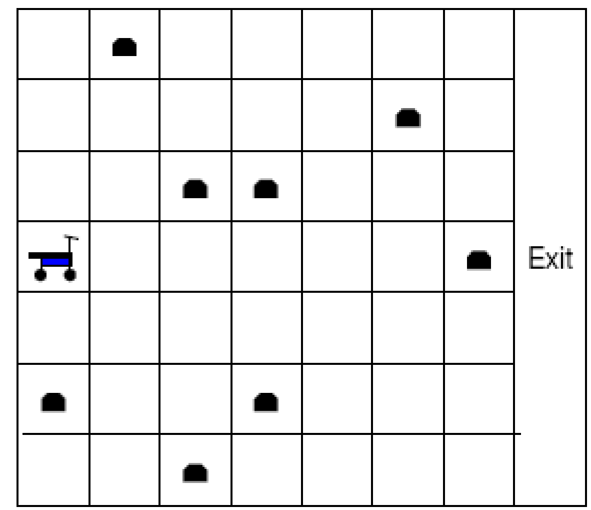
- 动机：用家庭服务机器人照看老人
- 老人有一个求助按钮，按下按钮后，求助信号会持续一段时间，然后关闭
- 机器人在有求助信号时到达老人旁边，会有一个正奖赏
- 机器人需要最小化移动以减少电池消耗



POMDP 通过最大化期望回报来帮助机器人获得有趣的行动

POMDP示例：岩石采样

- 状态（12545个） $7 \times 7 \times 2^8 + 1$（Exit）
  - 在7×7的栅格世界中，有8块岩石
  - 每块岩石有2种状态（有价值、没价值）
- 行动（13个）
  - 上下左右，观察第i块岩石，采样
- 观察（2个）
  - 被观察的岩石是否有价值（有噪声）
- 奖赏
  - +10：采样有价值的岩石，从Exit退出；−1：每步移动



RockSample(7, 8)

更多的POMDP例子
https://bigbird.comp.nus.edu.sg/pmwiki/farm/appl/index.php?n=Main.Repository

部分可观察的MDP

- 模型定义及示例
- **信念状态、策略、值函数**

信念状态

- Agent需要依赖过去行动和观察序列的完整历史信息来选择理想的行动

如何表示POMDP模型的最优策略？

- 方式一：显式地表示历史信息，构建历史到行动的映射

不实用，显式地保存历史信息需要大量存储空间

- 方式二：通过信念状态来表征与决策有关的、过去行动和观察序列的完整历史信息

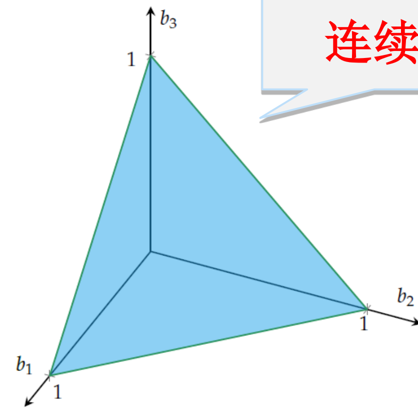
信念状态（续）

- 假设状态空间$\mathcal{S}$是离散的
- 信念状态b：定义在状态空间$\mathcal{S}$上的向量
  - $b_t(s)$：在t时刻，Agent在状态s的概率

$$\boxed{b_t(s) = P(S_t = s \mid O_t, A_{t-1}, O_{t-1}, \dots, A_0, b_0)}$$

  - 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$s \in \mathcal{S}$，均有$b(s) \in [0,1]$，且$\sum_{s \in S} b(s) = 1$

信念状态（续）

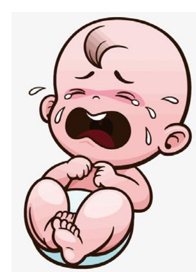


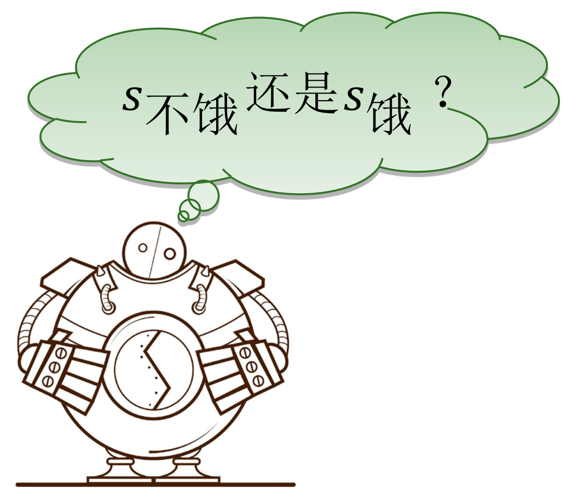
连续！

- 信念状态空间$\mathcal{B}$
  - 所有信念状态构成的空间

示例：3个状态的POMDP问题的信念状态空间

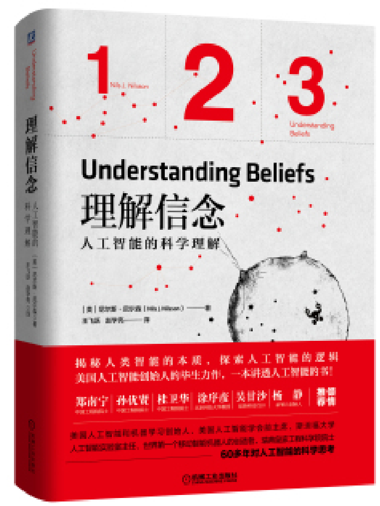
- 啼哭婴儿问题
  - $s_{不饿}$：婴儿不饿（not-hungry ）
  - $s_{饿}$：婴儿饿了（hungry）





信念状态（续）

我越来越不相信绝对真理的存在，真理不过是非常合理而且十分可靠的信念而已！
–尼尔斯·尼尔森（Nils J. Nilsson）



信念状态MDPs

- POMDP：状态为信念状态的MDP，即信念状态MDP

	标准的MDP	信念状态MDP
状态	$s \in \mathcal{S}$	$b \in \mathcal{B}$
行动	$a \in \mathcal{A}$	$a \in \mathcal{A}$
状态转移函数	$T(s' \mid s, a)$	$\tau(b' \mid b, a)$
期望奖赏函数	$R(s, a)$	$R(b, a)$

$$R(b, a) = \sum_{s} R(s, a) b(s)$$

信念状态MDPs（续）

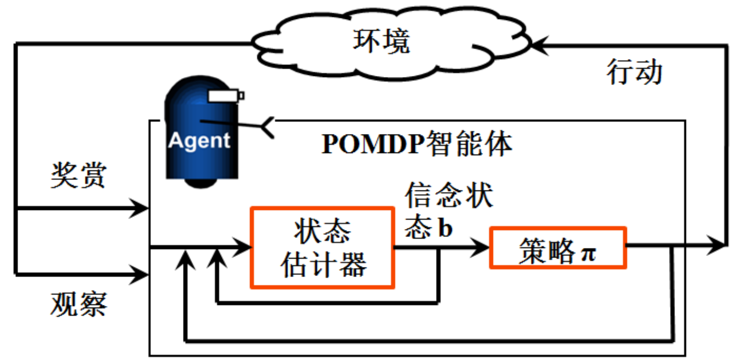
- 状态转移函数$\tau(b'|b,a)$的计算过程如下：

$$
\begin{aligned}
\tau(b' \mid b,a) &= P(b' \mid b,a) \\
&= \sum_{o} P(b' \mid b,a,o)P(o \mid b,a) \\
&= \sum_{o} P(b' \mid b,a,o)\sum_{s} P(o \mid b,a,s)P(s \mid b,a) \\
&= \sum_{o} P(b' \mid b,a,o)\sum_{s'} O(o \mid s',a)\sum_{s} P(s' \mid b,a,s)P(s \mid b,a) \\
&= \sum_{o} P(b' \mid b,a,o)\sum_{s'} O(o \mid s',a)\sum_{s} T(s' \mid s,a)b(s)
\end{aligned}
$$

- 求解信念状态MDPs是有挑战的，因为信念状态空间是连续的
  - 能使用近似规划技术来求解
  - 更好的方法：利用信念状态MDPs的结构

策略

- 策略：信念状态$b \in \mathcal{B}$到行动$a \in \mathcal{A}$的映射



- 策略执行算法

Algorithm 6.1 POMDP policy execution

function POMDPPOLICYEXECUTION(π)
    $b \leftarrow$ initial belief state
    loop
        Execute action $a = \pi(b)$
        Observe o and reward r
        $b \leftarrow$ UPDATEBELIEF(b, a, o)

值函数

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

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

- POMDP中的最优值函数U^{*}也满足Bellman最优方程：

$$U^{*}(b)=\max_{a\in\mathcal{A}}Q^{*}(b,a)$$

$$Q^{*}(b,a)=R(b,a)+\gamma\sum_{o\in\mathcal{O}}P(o\mid b,a)\ U^{*}\big(\text{UPDATEBELIEF}(b,a,o)\big)$$

- 通过下式获得信念状态b处的最优行动：

$$\pi^{*}(b)\leftarrow\operatorname{argmax}_{a\in\mathcal{A}}Q^{*}(b,a)$$

小结：部分可观察的MDP

- POMDP
  - 有观察模型的MDP、部分可观察性
  - 示例：啼哭婴儿问题、目标跟踪、岩石采样
- 信念状态：Agent在各个状态的概率分布
- 信念状态MDP
  - 状态为信念状态的MDP
  - 求解有挑战：信念状态空间连续
- 策略：信念状态到行动的映射
- 值函数：定义在信念状态空间上的期望折扣回报

内容安排

- 部分可观察的MDP
- **信念状态更新**
- 精确求解方法
- 离线方法
- 在线方法
- 应用案例：飞行器避碰系统

信念状态更新

- 给定一个信念状态，在执行一个行动a，获得一个观察o之后，可以使用**递归贝叶斯估计**来更新信念状态

Algorithm 6.1 POMDP policy execution

function POMDPPolicyExecution(π)
    $b \leftarrow$ initial belief state
    loop
        Execute action $a = \pi(b)$
        Observe o and reward r
        $b \leftarrow$ UpdateBelief(b, a, o)

- 将讨论三类POMDP问题
  - 有离散状态的问题
  - 有线性高斯状态转移和观察的问题
  - 有连续状态空间的一般问题

离散状态滤波器

- 用下式计算一个新的信念状态：

$$\boxed{b'(s') = \frac{O(o \mid s', a)}{P(o \mid b, a)} \sum_{s} T(s' \mid s, a) b(s)}$$

- 推导过程：

$$\begin{aligned}
b'(s') &= P(s' \mid o, a, b) \\
&\propto P(o \mid s', a, b) P(s' \mid a, b) \\
&= O(o \mid s', a) P(s' \mid a, b) \\
&= O(o \mid s', a) \sum_{s} P(s' \mid a, b, s) P(s \mid a, b) \\
&= \underline{O(o \mid s', a)} \sum_{s} T(s' \mid a, s) b(s)
\end{aligned}$$

观察空间可以是连续的，
这时表示概率密度

离散状态滤波器（续）

- 用啼哭婴儿问题来解释信念状态更新

假设：初始信念状态$\left(b\left(s_{\text{不饿}}\right), b(s_{\text{饿}})\right) = (0.5, 0.5)$

行动	观察	新的信念状态
不给婴儿喂食物	婴儿啼哭	$(0.0928, 0.9072)$
给婴儿喂食物	婴儿不啼哭	$(1, 0)$
不给婴儿喂食物	婴儿不啼哭	$(0.9759, 0.0241)$
不给婴儿喂食物	婴儿不啼哭	$(0.9701, 0.0299)$
不给婴儿喂食物	婴儿啼哭	$(0.4624, 0.5376)$

线性高斯滤波器

- 连续状态空间中的信念状态更新公式：

$$b'(s') \propto O(o \mid a, s') \int T(s' \mid s, a) b(s)\, ds$$

- 连续的（状态转移和观察）模型有如下线性高斯形式：

$$T(\mathbf{s}' \mid \mathbf{s}, \mathbf{a}) = \mathcal{N}(\mathbf{s}' \mid \mathbf{T}_s \mathbf{s} + \mathbf{T}_a \mathbf{a}, \Sigma_s)$$

$$O(\mathbf{o} \mid \mathbf{s}') = \mathcal{N}(\mathbf{o} \mid \mathbf{O}_s \mathbf{s}', \Sigma_o)$$

- 假设初始信念状态服从高斯分布：$b(\mathbf{s}) = \mathcal{N}(\mathbf{s} \mid \boldsymbol{\mu}_b, \boldsymbol{\Sigma}_b)$

- 更新信念状态的公式如下：

卡尔曼（Kalman）滤波器

K：卡尔曼增益

$$\mathbf{K} \leftarrow \boldsymbol{\Sigma}_p \mathbf{O}_s^\top \left(\mathbf{O}_s \boldsymbol{\Sigma}_p \mathbf{O}_s^\top + \boldsymbol{\Sigma}_o \right)^{-1}$$

$$\boldsymbol{\mu}_b \leftarrow \boldsymbol{\mu}_p + \mathbf{K}\left(\mathbf{o} - \mathbf{O}_s \boldsymbol{\mu}_p \right)$$

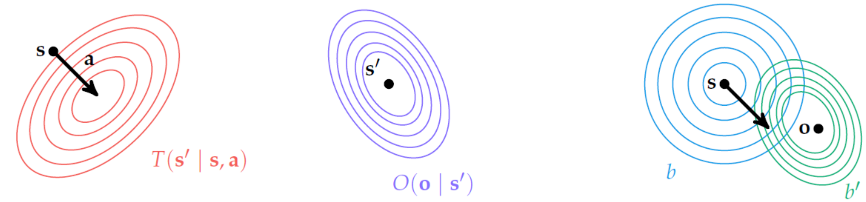
$$\boldsymbol{\Sigma}_b \leftarrow (\mathbf{I} - \mathbf{K}\mathbf{O}_s)\boldsymbol{\Sigma}_p$$

其中，$\boldsymbol{\mu}_p \leftarrow \mathbf{T}_s \boldsymbol{\mu}_b + \mathbf{T}_a \mathbf{a}$

$$\boldsymbol{\Sigma}_p \leftarrow \mathbf{T}_s \boldsymbol{\Sigma}_b \mathbf{T}_s^\top + \boldsymbol{\Sigma}_s$$

分别为得到一个观察前预测的均值和方差

线性高斯滤波器（续）



- 考虑如下线性高斯转移和观察函数：

$$T(\mathbf{s}' \mid \mathbf{s}, \mathbf{a}) = \mathcal{N}\left(\mathbf{s}' \mid \mathbf{s} + \mathbf{a}, \frac{1}{10}\begin{bmatrix} 1 & 1/2 \\ 1/2 & 1 \end{bmatrix}\right)$$

$$O(\mathbf{o} \mid \mathbf{s}') = \mathcal{N}\left(\mathbf{o} \mid \mathbf{s}', \frac{1}{20}\begin{bmatrix} 1 & -1/2 \\ -1/2 & 2 \end{bmatrix}\right)$$

- 假设信念状态 $\mathbf{b} = \mathcal{N}([-0.75, 1], \mathbf{I})$

行动 $a = [0.5, -0.5]$

观察 $\mathbf{o} = [0.3, 0.5]$

- 使用卡尔曼滤波器的信念状态更新公式

$$\mathbf{b}' = \mathcal{N}\left(\begin{bmatrix} 0.184 \\ 0.571 \end{bmatrix}, \begin{bmatrix} 0.037 & -0.011 \\ -0.011 & 0.050 \end{bmatrix}\right)$$

其中，卡尔曼增益

$$\mathbf{K} = \begin{bmatrix} 0.789 & 0.110 \\ 0.128 & 0.716 \end{bmatrix}$$

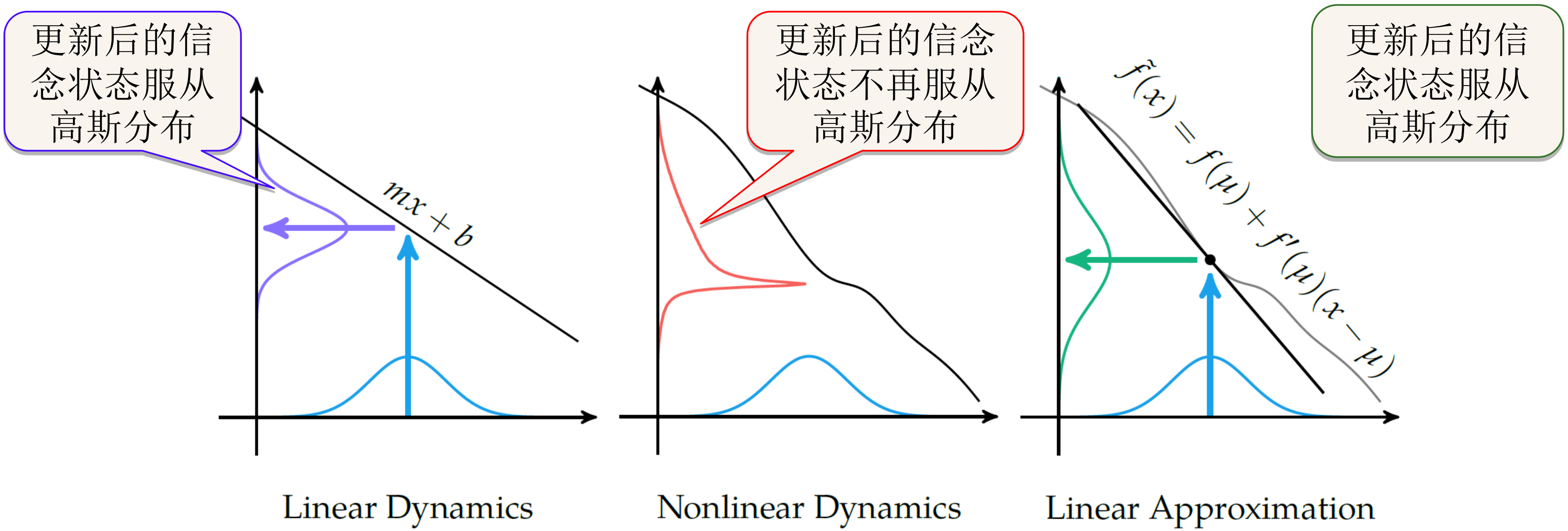
扩展的卡尔曼滤波器

- 考虑如下非线性高斯转移和观察函数：

$$T(\mathbf{s}' \mid \mathbf{s}, \mathbf{a}) = \mathcal{N}(\mathbf{s}' \mid \underline{\mathbf{f}_T(\mathbf{s}, \mathbf{a})}, \boldsymbol{\Sigma}_s)$$
$$O(\mathbf{o} \mid \mathbf{s}') = \mathcal{N}(\mathbf{o} \mid \underline{\mathbf{f}_O(\mathbf{s}')}, \boldsymbol{\Sigma}_o)$$

可微分的函数

- 使用一阶泰勒展开式来局部线性近似非线性函数，然后用卡尔曼滤波器更新信念状态



粒子滤波器

- 使用基于采样的方法来更新信念状态
  - 状态空间很大或连续
  - 动力系统不能用线性高斯模型很好近似
- 粒子滤波器
  - 信念状态：粒子的集合
  - 粒子：状态空间中的样本
  - 随着粒子数的增加，用粒子集合表示的信念状态会接近真实的信念状态
- 基于一个产生式模型G来更新b
  - 黑盒仿真器
  - 不需要转移或观察概率的显式知识
  - 粒子损失问题：由于随机性，有可能产生的样本并不在真实状态附近
    - 缓解的方法：给粒子添加些额外的噪声

带拒绝的粒子滤波器

Algorithm 6.2 Particle filter with rejection

```
function UpdateBelief(b, a, o)
    b′ ← ∅
    for i ← 1 to |b|
        repeat
            s ← random state in b
            (s′, o′) ~ G(s, a)
        until o′ = o
        Add s′ to b′
    return b′
```

从b中随机采样一个样本s，然后从$G(s,a)$中抽取样本(s',o')

重复这一过程，直至抽到的o'与观察到的o相同

把抽取的s'添加到新的信念状态b'中

返回b'，其中存放了$|b|$个新粒子

- 问题：要求从$G(s,a)$中抽取很多样本，直至抽到的观察与实际的观察相同
  - 在观察空间很大或连续时，这个问题就会凸显出来
  - 与第二部分的直接采样推理类似

不带拒绝的粒子滤波器

Algorithm 6.3 Particle filter without rejection

```
function UpdateBelief(b, a, o)
    b′ ← ∅
    for i ← 1 to |b|
        s_i ← random state in b
        s′_i ∼ G(s_i, a)
        w_i ← O(o | s′_i, a)
    for i ← 1 to |b|
        Randomly select k with probability proportional to w_k
        Add s′_k to b′
    return b′
```

阶段1：
从b中随机采样一个样本s_i
用$G(s_i, a)$返回下一个状态s_i'
对每个新样本s_i'，基于观察模型来计算其权重w_i

阶段2：
从阶段1得到的$|b|$个带权重的新样本中，依权重概率来采样$|b|$个样本，即为b'

- 要求有一个观察模型，用它来定义样本的权重，基于这些权重来得到b'

小结：信念状态更新

- 离散状态滤波器
  - 离散状态的POMDPs
- 卡尔曼滤波器及扩展

连续POMDPs

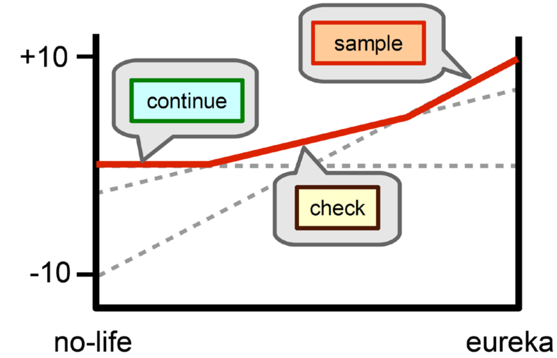
  - 具有线性高斯转移和观察模型
  - 具有非线性（含高斯噪声）高斯转移和观察模型
- 粒子滤波器
  - 连续POMDPs
  - 信念状态：粒子集合
  - 带拒绝的粒子滤波器：在观察空间大的问题中失效
  - 不带拒绝的粒子滤波器：要求有一个观察模型

内容安排

- 部分可观察的MDP
- 信念状态更新
- **精确求解方法**
- 离线方法
- 在线方法
- 应用案例：飞行器避碰系统

阿尔法向量

- 一步离散状态POMDP的最优值函数：



值函数的示例

$$U^*(b) = \max_a \sum_s b(s) R(s,a)$$

$$U^*(\mathbf{b}) = \max_a \boldsymbol{\alpha}_a^\top \mathbf{b}$$

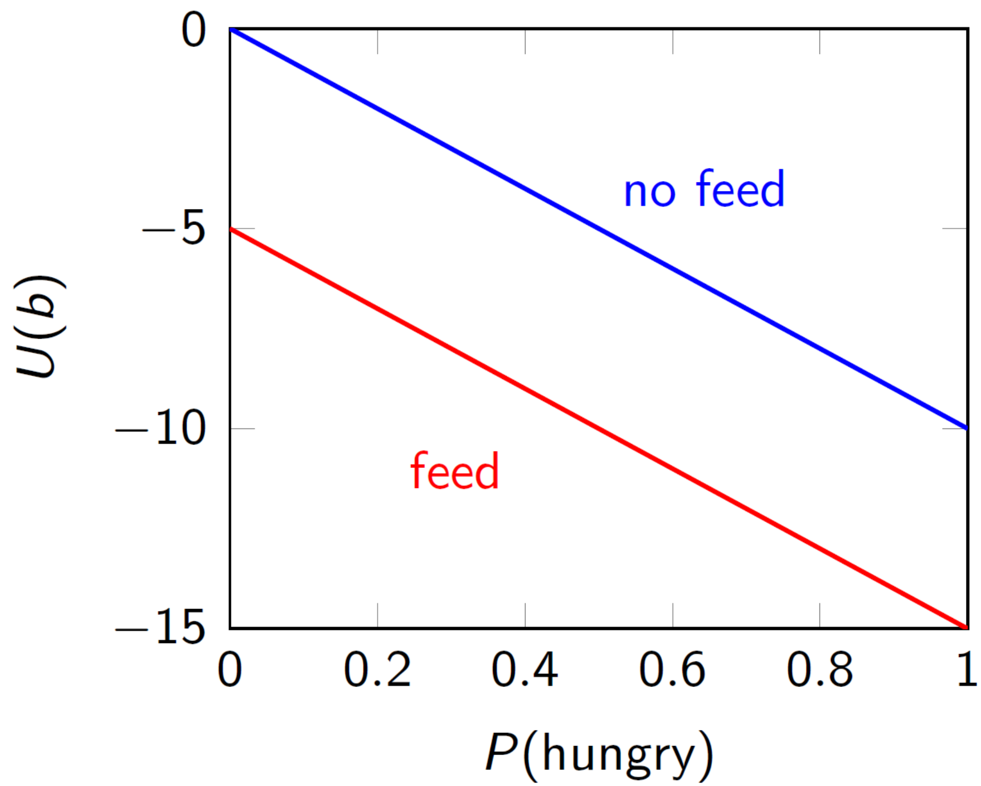
$\boldsymbol{\alpha}_a$：$R(\cdot, a)$的向量表示
b：信念状态的向量表示

阿尔法向量：表示值函数的向量

- 在一步POMDP中，每个行动都有一个阿尔法向量
- 每个阿尔法向量对应信念状态空间的一个超平面
- 值函数：分段线性凸函数

阿尔法向量（续）

- 一步婴儿啼哭问题的阿尔法向量：



信念状态：

$$(b(\text{not-hungry}), b(\text{hungry}))^\top$$

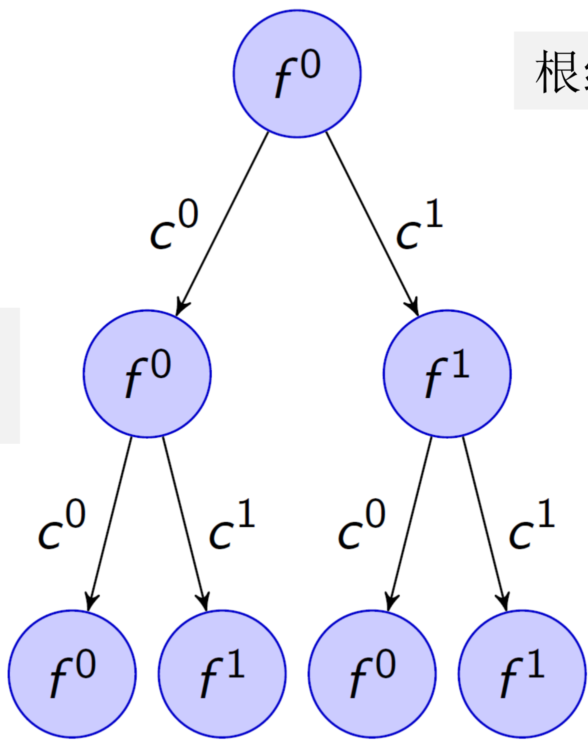
两个阿尔法向量：

$$\alpha_{\text{not-feed}} = (0, -10)^\top$$
$$\alpha_{\text{feed}} = (-5, -15)^\top$$

不论当前信念状态是什么，一步最优策略都是不给婴儿喂食物

条件规划

- 在多步POMDP中，一个策略是一个条件规划，表示为一棵树



根结点：不给婴儿喂食物

观察到婴儿不哭，不给婴儿喂食物

观察到婴儿啼哭，给婴儿喂食物

不啼哭，不喂食物
啼哭，喂食物

婴儿啼哭问题的一个3步规划的例子

条件规划（续）

- 递归地计算$U^p(s)$：在s使用条件规划p的期望回报

$$U^p(s) = R(s,a) + \sum_{s'} T(s' \mid s,a) \sum_{o} O(o \mid s',a) U^{p(o)}(s')$$

a：与p的根结点关联的行动
$p(o)$：与观察o关联的子规划

- 计算与一个信念状态关联的期望回报：$U^p(b) = \sum_s U^p(s) b(s)$
- 则：$U^p(\mathbf{b}) = \boldsymbol{\alpha}_p^\top \mathbf{b}$

$\boldsymbol{\alpha}_p$：U^p的向量表示

- 如果**最大化**所有可能给定步数的规划，则有

有限步数的最优值函数是分段线性凸函数

$$U^*(\mathbf{b}) = \max_p \boldsymbol{\alpha}_p^\top \mathbf{b}$$

最优行动是规划 $\arg\max_p \boldsymbol{\alpha}_p^\top \mathbf{b}$ 在根结点中的行动

值迭代

- 通常，枚举所有可能的h步规划，从中找到 $\arg\max_{p} \boldsymbol{\alpha}_p^\top \mathbf{b}$ 不可行
- 一个h步规划的结点数：

$$\frac{|\mathcal{O}|^h - 1}{|\mathcal{O}| - 1}$$

- 每个结点与$|\mathcal{A}|$个行动关联，因此有

$$|\mathcal{A}|^{\frac{|\mathcal{O}|^h - 1}{|\mathcal{O}| - 1}}$$

个可能的h步规划

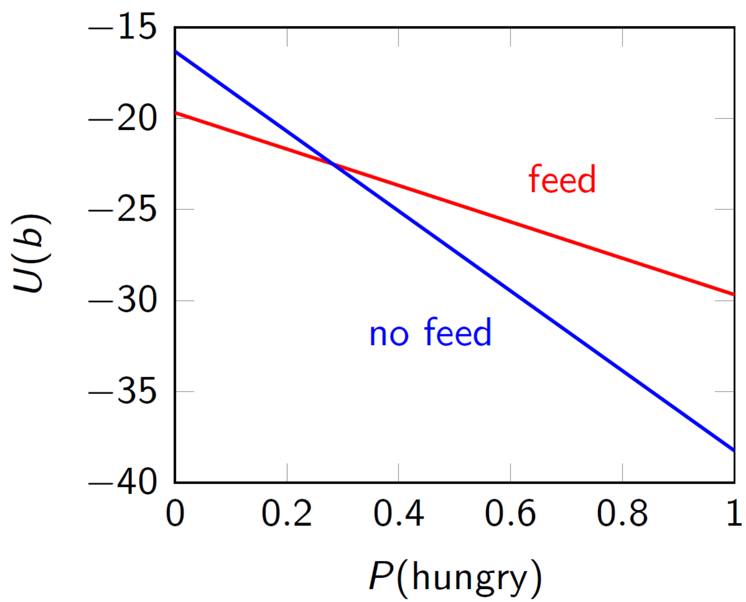
- 有两个行动、两个观察的啼哭问题，有2^{63}个6步条件规划
- 在最坏情况下，精确求解有限步数的POMDP问题是PSPACE-完全的（难度≥NP-完全类）

值迭代（续）

- 在POMDP中值迭代的想法：
  - 遍历1步规划，丢掉在任意初始信念状态处都不是最优的1步规划
  - 用剩下的1步规划来产生可能最优的2步规划，丢掉次优的2步规划
  - 重复这一过程，直至到达指定步数
- 可以用线性规划来鉴别规划在某些信念状态处是否被占优

两个不被占优的阿尔法向量，交点为$P(hungry) = 0.28206$

如果$P(hungry) > 0.28206$，则给婴儿喂食物



婴儿啼哭问题的最优策略（折扣因子为0.9）

小结：精确求解方法

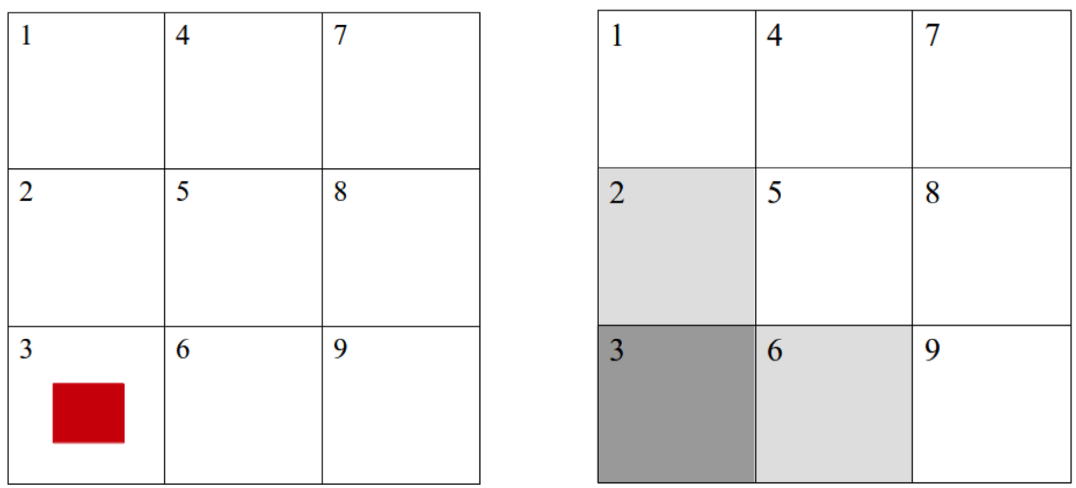
- 阿尔法向量
  - 表示值函数的向量
  - $\boldsymbol{\alpha}_p$：U^p的向量表示，其中p是条件规划
- 有限步数的POMDP问题
  - 最优值函数：分段线性凸函数
  - 时间复杂度：PSPACE-完全
- 值迭代
  - 遍历所有k步规划，丢掉次优的规划，产生$k+1$步规划

课后练习6.1

- POMDP是什么的缩写？它与MDP有何不同？画出POMDP的决策网络结构，它与MDP的决策网络结构有何不同？

课后练习6.2

- 有如下两个栅格世界。在左侧的栅格世界中，已知Agent的位置（表示为红色方块），在右侧的栅格世界中，只有可能状态的一个概率分布。对每种情况，如何表示状态？请用这个例子来解释为什么称有些POMDPs为信念状态MDPs？为什么求解信念状态MDPs是困难的？



课后练习6.3

- 可以用哪些方法来更新POMDP中的信念状态？使用时怎样在这些方法中选择？

课后练习6.4

- 对一个离散状态的滤波器，从信念状态更新的定义

$$b'(s') = P(s' \mid o, a, b)$$

出发，推导出如下方程：

$$b'(s') \propto O(o \mid s', a) \sum_{s} T(s' \mid s, a) b(s)$$

课后练习6.5

- 假想你已经解出了一个3个状态的POMDP问题的策略，可以表示为如下阿尔法向量：

$$\begin{pmatrix} 300 \\ 100 \\ 0 \end{pmatrix}, \begin{pmatrix} 167 \\ 10 \\ 100 \end{pmatrix}, \begin{pmatrix} 27 \\ 50 \\ 50 \end{pmatrix}$$

  第1、3个阿尔法向量对应的行动为1，第2个阿尔法向量对应的行动为2。这是一个有效的策略吗？能每个行动有多个阿尔法向量吗？如果该策略有效，请确定在信念状态

$$b = \begin{bmatrix} 0 \\ 0.7 \\ 0.3 \end{bmatrix}$$

  应采取何种行动？