高级机器学习

第十讲：强化学习简介

强化学习

从预测到决策：如何种出好瓜

种子 -> 💀

种子 -> 浇水 -> 浇水 -> 浇水 -> 💀

种子 -> 浇水 -> 施肥 -> 施肥 -> 施肥 -> 💀

种子 -> 浇水 -> 施肥 -> 浇水 -> 施肥 -> 施肥 -> 💀

种子 -> 浇水 -> 施肥 -> 浇水 -> 施肥 -> 浇水 -> 施肥 -> 💀

种子 -> 杀虫 -> 浇水 -> 施肥 -> 浇水 -> 施肥 -> 浇水 -> 施肥 -> 💀

种子 -> 浇水 -> 施肥 -> 浇水 -> 施肥 -> 杀虫 -> 浇水 -> 施肥 -> 杀虫 -> 💀

种子 -> 除草 -> 浇水 -> 施肥 -> 浇水 -> 施肥 -> 杀虫 -> 浇水 -> 施肥 -> 杀虫 -> 💀

种子 -> 浇水 -> 施肥 -> 浇水 -> 施肥 -> 除草 -> 杀虫 -> 浇水 -> 施肥 -> 杀虫 -> 除草 -> 浇水 -> 施肥 -> 浇水 -> 施肥 -> 浇水 -> 施肥 -> 除草 -> 杀虫 -> 🍉

强化学习

关键要素：<*A*, *X*, *R*, *P*>

action space: A
state space: X
reward: $R : X \times A \times X \to \mathbb{R}$
transition: $P : X \times A \times X \to \mathbb{R}$

策略： $a = \pi(x)$

$$P(a|x) = \pi(x,a) \qquad \sum_{a \in A} \pi(x,a) = 1 \qquad \forall a \in A, \pi(x,a) \geq 0$$

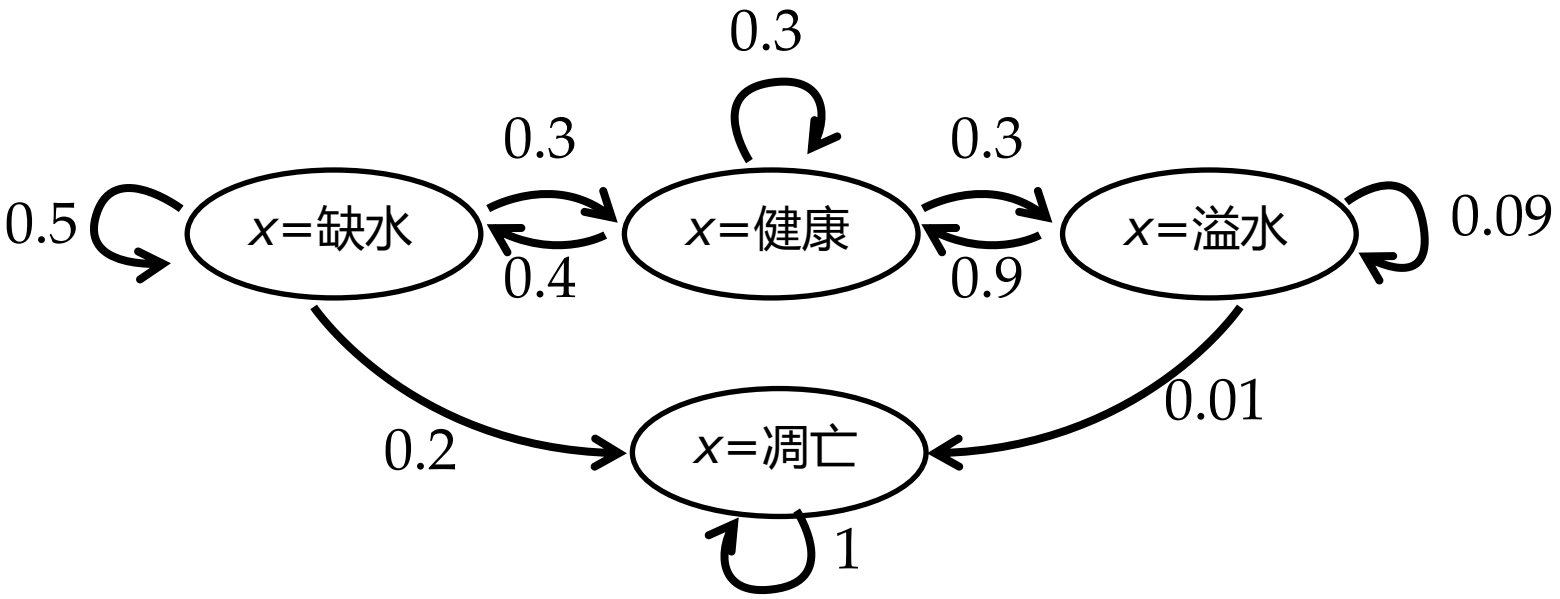
策略评价：累积回报

T-step: $\frac{1}{T}\sum_{t=1}^{T} r_t$ discounted: $\sum_{t=1}^{\infty} \gamma^t r_t$

学习目标：学习最大回报策略

马尔可夫过程 $< X, P >$

状态图



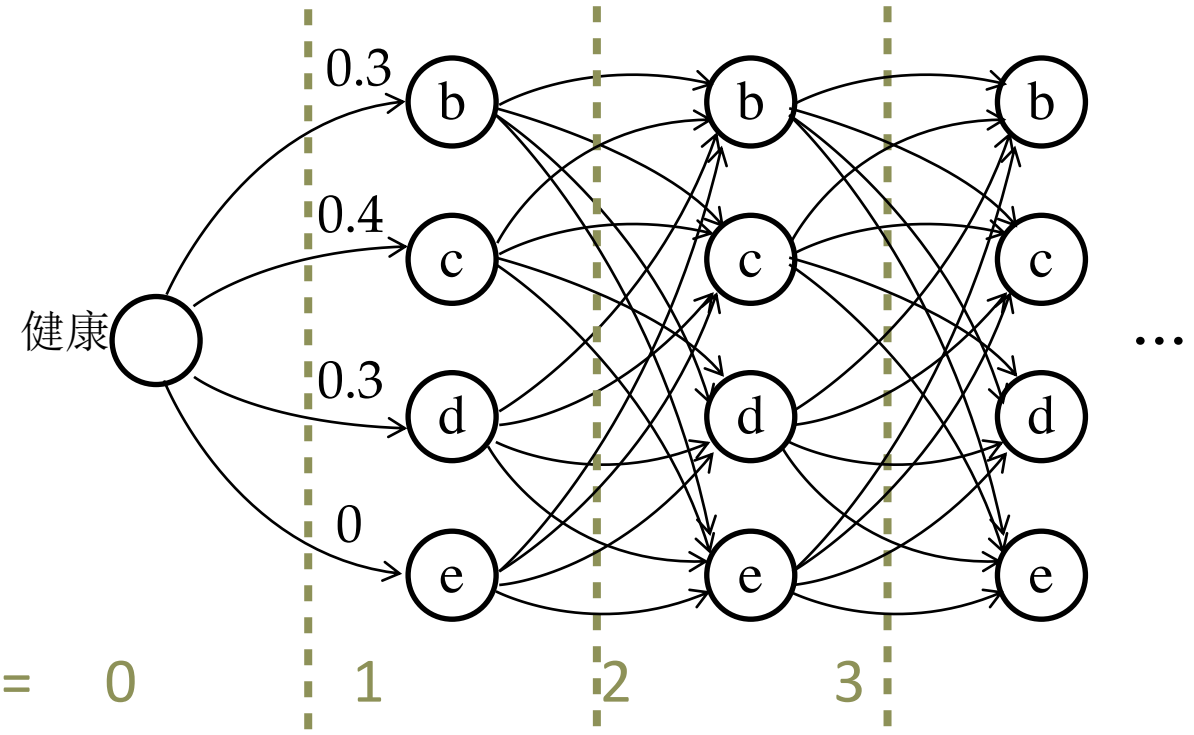
没有记忆的过程

$$P(x_{t+1}|x_t, \ldots, x_0) = P(x_{t+1}|x_t)$$

稳态分布 $\lim_{t\to\infty} P(x_{t+1}) - P(x_t) = 0$

马尔可夫过程 $< X, P >$

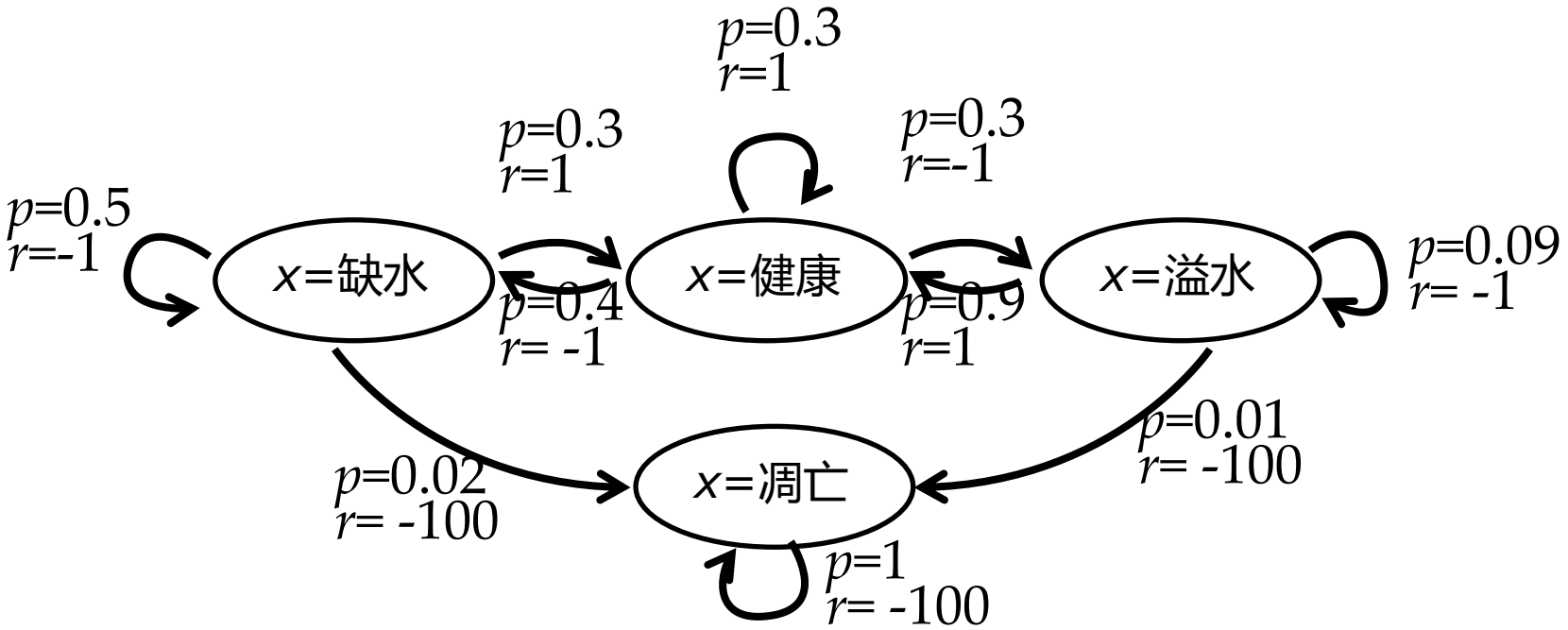
水平视角



采样：运行过程，得到过程采样

马尔可夫回报过程 MRP < X, R, P >

状态图

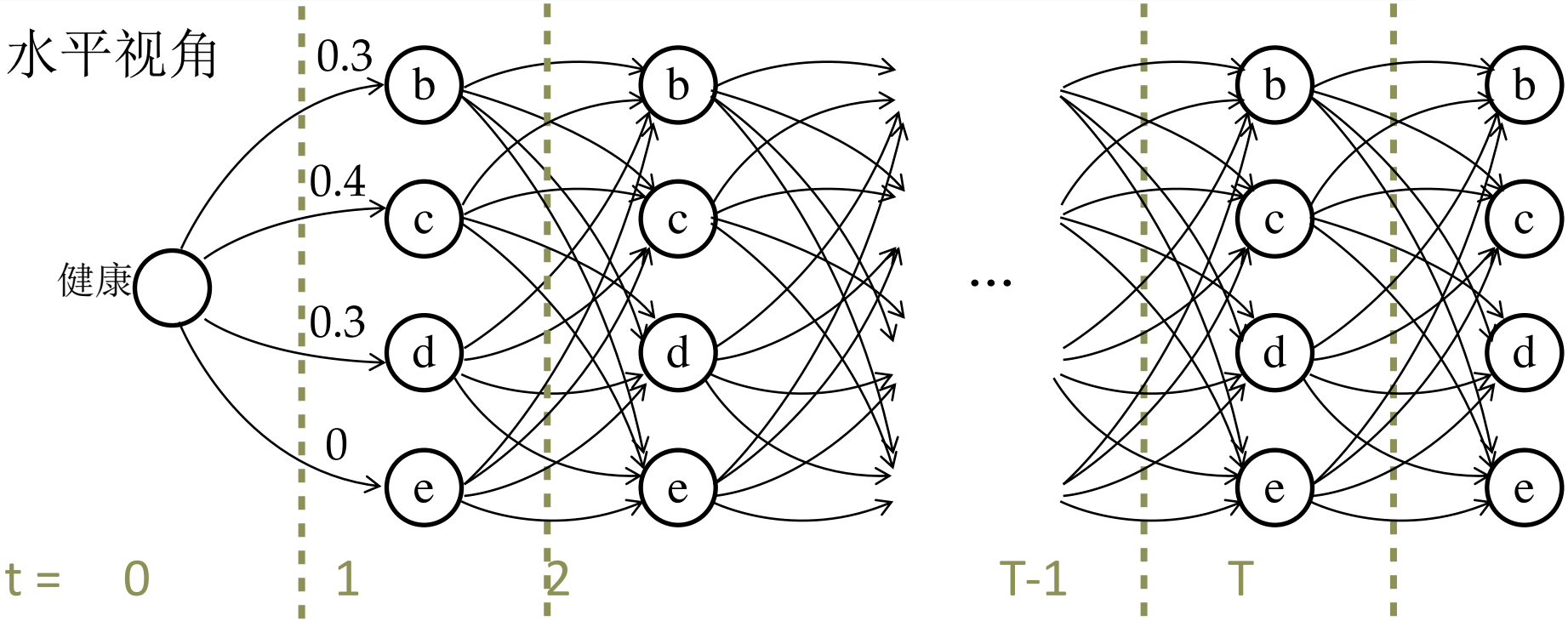


累积回报是多少？

$$V(b) = E[\sum_{t=1}^{T} r_t | x_0 = b]$$

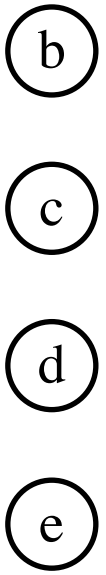
$$V(b) = E[\sum_{t=1}^{T} \gamma r_t | x_0 = b]$$

马尔可夫回报过程 MRP < X, R, P >



马尔可夫回报过程 MRP < X, R, P >

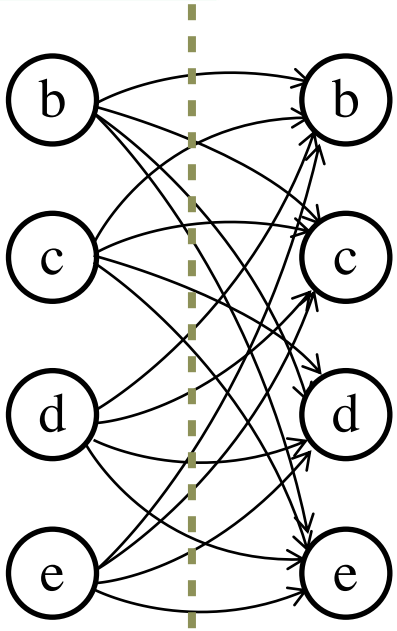
水平视角



$$V(x) = 0$$

马尔可夫回报过程 MRP < X, R, P >

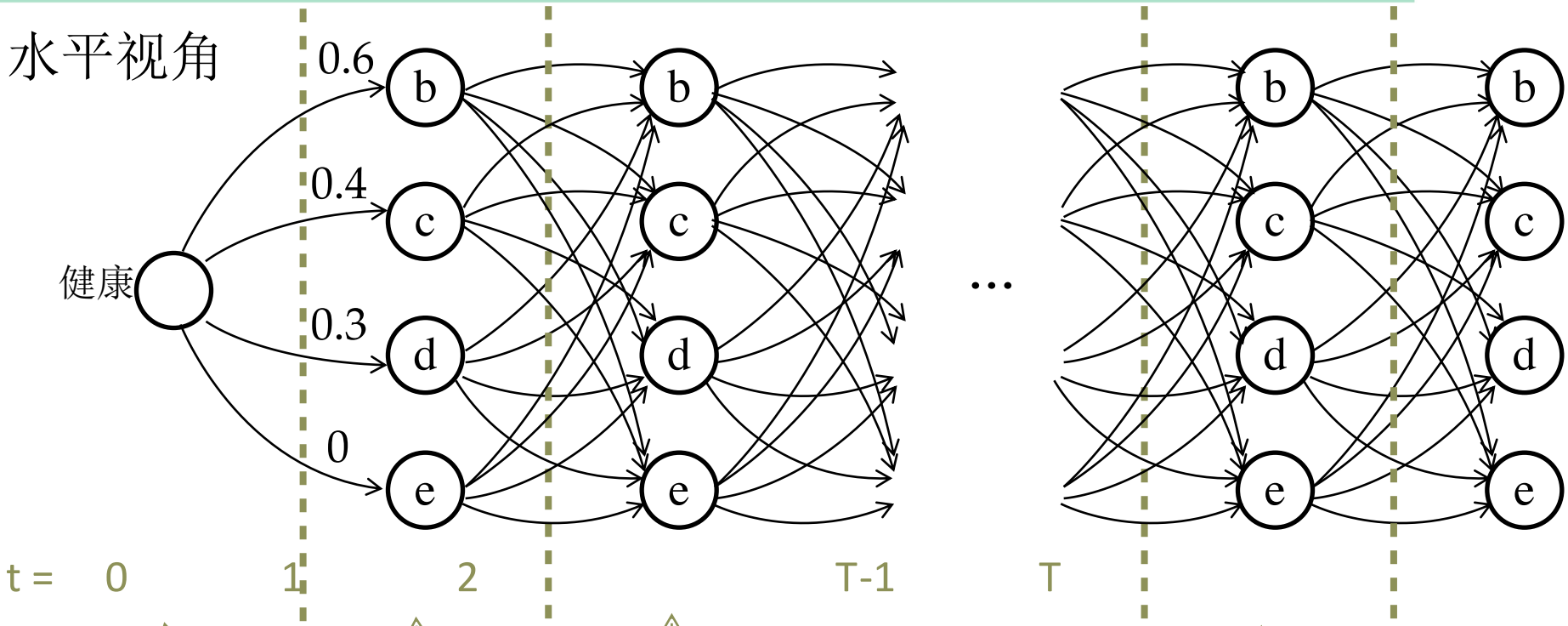
水平视角



$$V(x) = \sum_{x' \in X} P(x'|x) R(x', x) \quad V(x) = 0$$

$$V(x) = \sum_{x'} P(x'|x)(R(x', x) + V(x'))$$

马尔可夫回报过程 MRP < X, R, P >



$$V(x) = 0$$

迭代计算

$$V(x) = \sum_{x'} P(x'|x)(R(x', x) + V(x'))$$

马尔可夫回报过程 MRP < X, R, P >

迭代计算

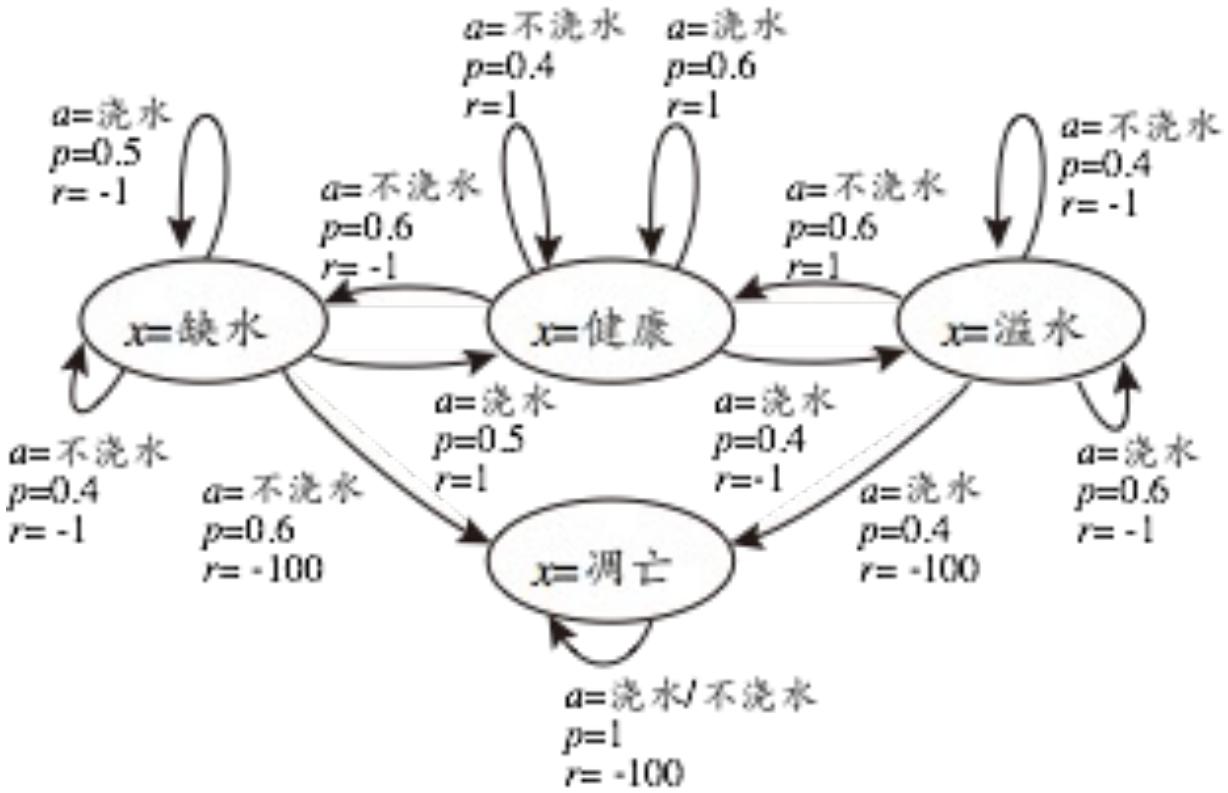
$V(x) = 0$ for all x
loop
  for each x: $V'(x) = \sum_{x'} P(x'|x)(R(x', x) + V(x'))$
  V=V'
until T iterations

折扣回报

$V(x) = 0$ for all x
while true
  for each x: $V'(x) = \sum_{x'} P(x'|x)(R(x', x) + \gamma V(x'))$
  break if $||V - V'|| < \theta$
  V=V'

马尔可夫决策过程 MDP < *A*, *X*, *R*, *P* >

状态图

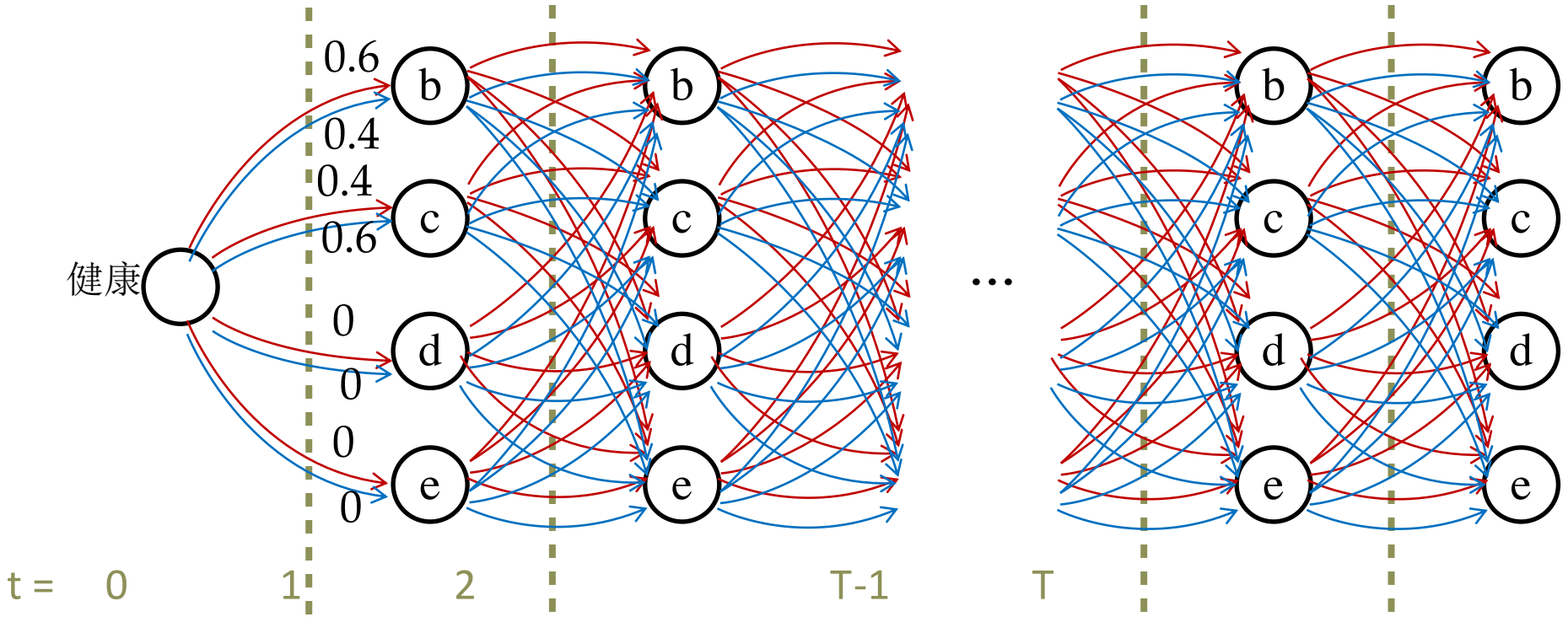


表格策略表达

$\pi =$

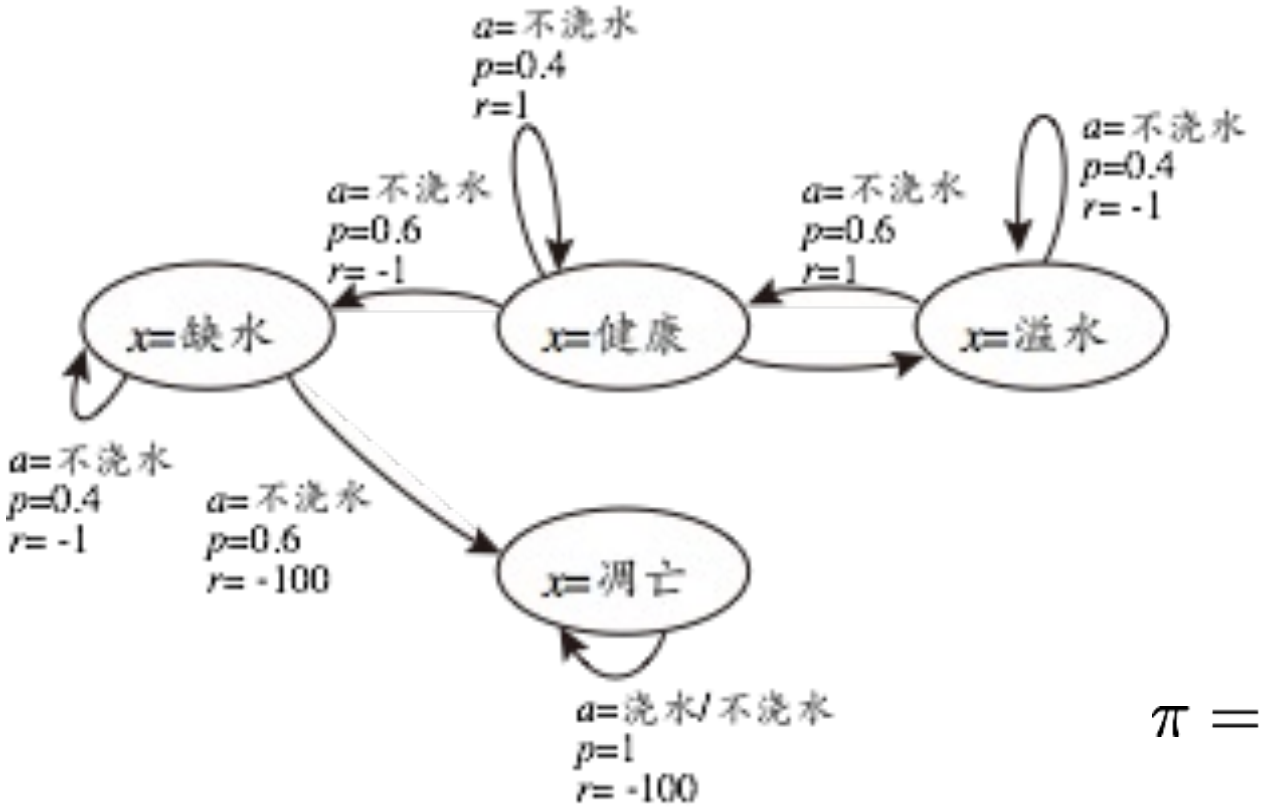
b	浇	0.3
	不	0.7
c	浇	0.6
	不	0.4
d	浇	0.1
	不	0.9
e	浇	0.5
	不	0.5

马尔可夫决策过程 MDP < *A*, *X*, *R*, *P* >



马尔可夫决策过程 MDP < *A*, *X*, *R*, *P* >

状态图



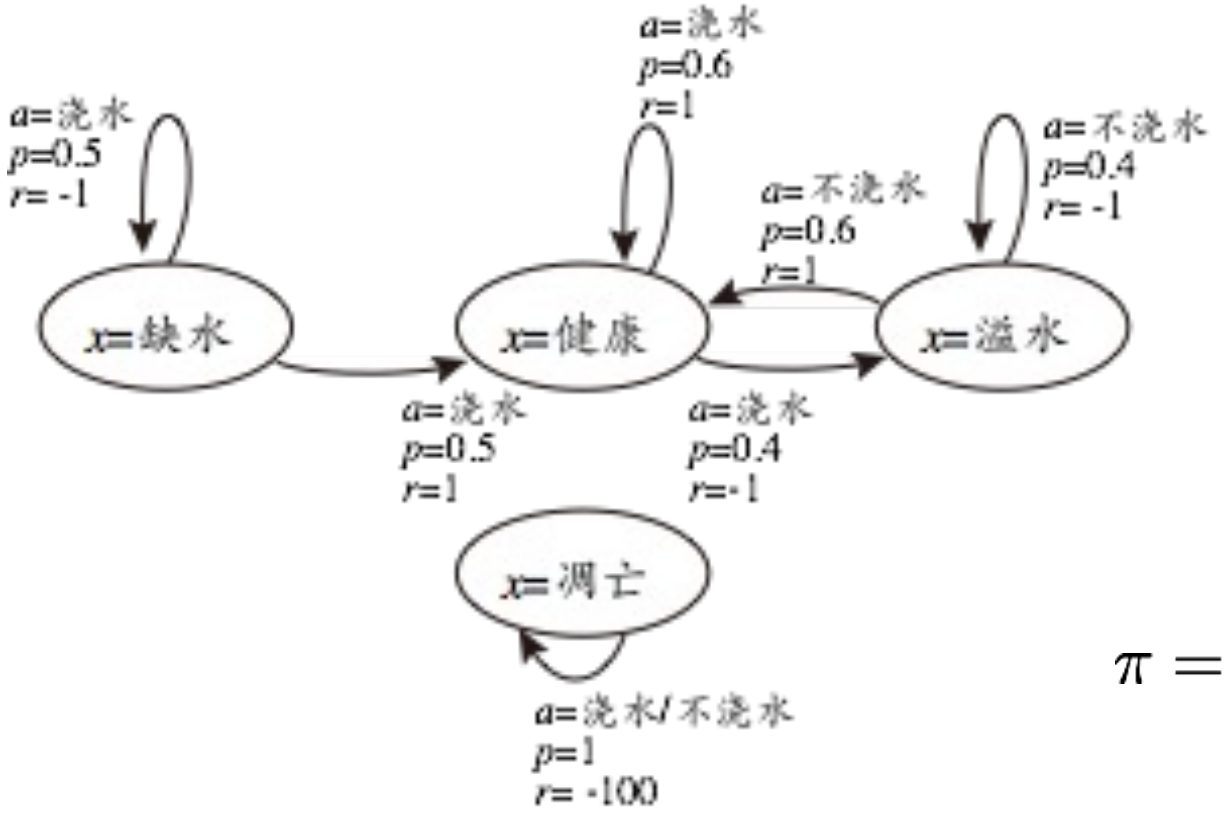
表格策略表达

$\pi =$

b	浇	0
	不	1
c	浇	0
	不	1
d	浇	0
	不	1
e	浇	0.5
	不	0.5

马尔可夫决策过程 MDP < *A*, *X*, *R*, *P* >

状态图



表格策略表达

$\pi =$

b	浇	1
	不	0
c	浇	0
	不	1
d	浇	1
	不	0
e	浇	0.5
	不	0.5