马尔科夫决策过程（Markov Decision Process, MDP）

模型已知，环境完全可观察

Agent 感知环境状态 $S_t = s \in \mathcal{S}$, 决定做出行动 $A_t = a \in \mathcal{A}$, 得到奖赏 $R_t = r \in \mathcal{R}$

马尔可夫假设 (性质): 状态 S_{t+1} 和奖赏 R_t 仅依赖于当前状态 S_t 和行动A_t ，与更早的状态和行动无关:

$$P(S_{t+1} = s', R_t = r | S_0, A_0, \ldots, S_t, A_t) = P(S_{t+1} = s', R_t = r | S_t, A_t)$$

MDP:

1. 状态空间 $\mathcal{S}$
2. 行动空间 $\mathcal{A}$
3. 奖赏空间 $\mathcal{R}$
4. 动力函数 $P(S_{t+1}, R_t | S_t, A_t)$

得到

- 状态转移函数 $P(S_{t+1} | S_t, A_t) = \sum_{r \in \mathcal{R}} P(S_{t+1}, R_t = r | S_t, A_t)$
- 奖赏函数 $P(R_t | S_t, A_t) = \sum_{s' \in \mathcal{S}} P(S_{t+1} = s', R_t | S_t, A_t)$

稳态MDP: 动力函数不随时间变化, 即 $p(s', r | s, a) = P(S_{t+1} = s', R_t = r | S_t = s, A_t = a)$

重新定义

- 状态转移函数 $T(s' | s, a) = \sum_{r \in \mathcal{R}} p(s', r | s, a)$
- 奖赏函数 $p(r | s, a) = \sum_{s' \in \mathcal{S}} p(s', r | s, a)$

若已知当前状态 s 与决定做出的行动 a , 期望奖赏 (函数) 为

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

若问题步数无限, 为方便处理, 定义总奖赏 (效用/回报) 为 $\sum_{t=0}^{\infty} \gamma^t R_t$, 其中折扣因子 $\gamma \in [0, 1)$

eg. $R = R_0 + 0.9R_1 + 0.81R_2 + \ldots$

策略 π_t : 给定当前状态 s_t , 给出行动

- 随机性策略 $\pi(a | s) = P(A_t = a | S_t = s)$
- 确定性策略 $\pi(s) \in \mathcal{A}$

时刻 t 的折扣回报：从时刻 t 起，Agent将得到的折扣奖赏之和 (递归关系)

$$G_t = R_t + \gamma R_{t+1} + \gamma^2 R_{t+2} + \ldots = \sum_{k=0}^{\infty} \gamma^k R_{t+k} = R_t + \gamma G_{t+1}$$
$$eg. G_5 = R_5 + 0.9R_6 + 0.81R_7 + \ldots = R_5 + 0.9G_6$$

Bellman期望方程

状态值函数 $U^\pi(s)$ ： 已知当前状态 s ，执行策略 π 的期望总回报

行动值函数 $Q^\pi(s,a)$: 已知当前状态 s 与采取行动 a，执行策略 π 的期望总回报

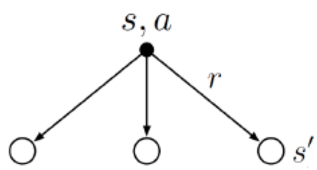
$$
\begin{gathered}
U^\pi(s) \\
= E_\pi[G_t|S_t = s] \\
= E_\pi[R_t + \gamma G_{t+1}|S_t = s] \\
= \sum_a [\pi(a|s)\; Q^\pi(s,a)] \\
= \sum_a \left[\pi(a|s) \sum_{s',r} p(s',r|s,a)\,(r + \gamma E_\pi[G_{t+1}|S_t = s'])\right] \\
= \sum_a \left[\pi(a|s) \sum_{s',r} p(s',r|s,a)\,(r + \gamma U^\pi(s'))\right]
\end{gathered}
$$

其中

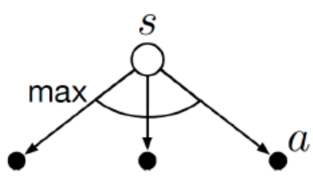
$$
\begin{gathered}
Q^\pi(s,a) = \sum_{s',r} p(s',r|s,a)\,(r + \gamma U^\pi(s')) \\
= \sum_{s',r} r \cdot p(s',r|s,a) + \gamma \sum_{s',r} p(s',r|s,a) U^\pi(s') \\
= R(s,a) + \gamma \sum_{s'} T(s'|s,a) U^\pi(s') \\
\Rightarrow U^\pi(s) = \sum_a \left[\pi(a|s)\left(R(s,a) + \gamma \sum_{s'} T(s'|s,a) U^\pi(s')\right)\right]
\end{gathered}
$$

可见 Q 包含 U, U 包含 Q

$$Q^*(s,a) = \sum_{s',r} p(s',r|s,a)\left[r + \gamma\, U^*(s')\right]$$



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



若为确定性策略, 则 $U^\pi(s) = R(s,\pi(s)) + \gamma \sum_{s'} T(s'|s,\pi(s)) U^\pi(s')$

策略的偏序关系: if $\forall s \in \mathcal{S},\ U^\pi(s) \le U^{\pi'}(s)$, then $\pi \le \pi'$

最优策略 π^* ： $\forall \pi,\ \pi^* \ge \pi$

- 随机

$$
\begin{gathered}
\pi^*(a|s) = 1,\ if\ a \in \arg\max_{a' \in \mathcal{A}} Q^*(s,a') \\
= 0, otherwise
\end{gathered}
$$

- 确定 $$\pi^*(s) \in \arg\max_{a' \in \mathcal{A}} Q^*(s,a')$$

Bellman最优方程

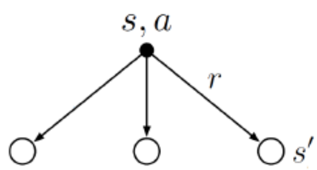
最优状态值函数 $U^*(s)$ ： 已知当前状态 s ，执行最优策略 π^* 的期望总回报

最优行动值函数 $Q^\pi(s,a)$: 已知当前状态 s 与采取行动 a，执行策略 π 的期望总回报

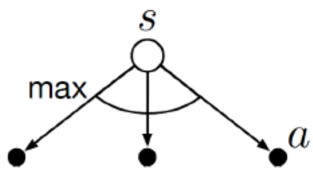
$$
\begin{aligned}
U^*(s) &= \max_\pi U^\pi(s) \\
&= \max_{a\in\mathcal{A}(s)} Q^*(s,a) \\
&= \max_{a\in\mathcal{A}(s)} E_{\pi^*}[G_t|S_t=s,A_t=a] \\
&= \max_{a\in\mathcal{A}(s)} E[R_t+\gamma U^*(S_{t+1})|S_t=s,A_t=a] \\
&= \max_{a\in\mathcal{A}(s)} \sum_{s',r} p(s',r|s,a)[r+\gamma U^*(s')] \\
&= \max_{a\in\mathcal{A}(s)} [R(s,a)+\gamma\sum_{s'} T(s'|s,a)U^*(s')] \\
Q^*(s,a) &= \max_\pi Q^\pi(s,a) = \sum_{s',r} p(s',r|s,a)[r+\gamma U^*(s')]
\end{aligned}
$$

可见 Q^* 包含 U^*, U^* 包含 Q^*

$$Q^*(s,a) = \sum_{s',r} p(s',r|s,a)[r+\gamma U^*(s')]$$



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



精确动态规划

- Bellman期望方程 $U^\pi(s) = R(s,\pi(s)) + \gamma\sum_{s'} T(s'|s,\pi(s))U^\pi(s')$
- Bellman最优方程 $U^*(s) = \max_{a\in\mathcal{A}(s)}[R(s,a)+\gamma\sum_{s'} T(s'|s,a)U^*(s')]$

$\Rightarrow$ 可以用动态规划

策略迭代

策略评价

计算一个策略的期望回报

Algorithm 4.1 Iterative policy evaluation

逐次逼近（successive approximation）

function IterativePolicyEvaluation(π, n)
　$U_0^\pi(s) \leftarrow 0$ for all s
　for $t \leftarrow 1$ **to** n
　　$U_t^\pi(s) \leftarrow R(s,\pi(s)) + \gamma\sum_{s'} T(s' \mid s,\pi(s))U_{t-1}^\pi(s')$ for all s
　return U_n^π

迭代地计算一个策略的n步回报

- 法一: 当 $\gamma < 1$且 n 足够大, 用 U_n^π 近似 U^π
- 法二: $U^\pi = R^\pi + \gamma T^\pi U^\pi \Rightarrow U^\pi = (I - \gamma T^\pi)^{-1} R^\pi$

策略改进

令 $\pi_{k+1}(s) = \arg\max_a Q^{\pi_k}(s, a)$, 则 $U^{\pi_k}(s) \le U^{\pi_{k+1}}(s)$

当 $\pi_{k+1} = \pi_k$, 则 $U^{\pi_k}(s) = \max_a Q^{\pi_k}(s, a)$, 满足Bellman最优方程, π_k 是最优策略

$$\pi_{k+1}(s) = \arg\max_a \left(R(s,a) + \gamma \sum_{s'} T(s' \mid s, a) U^{\pi_k}(s')\right) \text{ for all states } s$$

策略迭代

$$\pi_0 \xrightarrow{\text{E}} U^{\pi_0} \xrightarrow{\text{I}} \pi_1 \xrightarrow{\text{E}} U^{\pi_1} \xrightarrow{\text{I}} \pi_2 \xrightarrow{\text{E}} \cdots \xrightarrow{\text{I}} \pi^* \xrightarrow{\text{E}} U^*$$

E = Evaluate, I = Improve

值迭代

$$U_{k+1}(s) \leftarrow \max_a \left[R(s,a) + \gamma \sum_{s'} T(s' \mid s, a) U_k(s')\right] \text{ for all states } s$$

$$\pi(s) \leftarrow \arg\max_a \left(R(s,a) + \gamma \sum_{s'} T(s' \mid s, a) U^*(s')\right)$$

高斯-赛德尔（Gauss-Seidel）值迭代

$$U(s) \leftarrow \max_a \left(R(s,a) + \gamma \sum_{s'} T(s' \mid s, a) U(s')\right)$$

U 没有下标, 是一个 U 内部更新