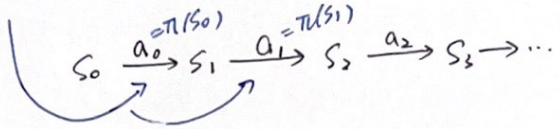
4. Optimal Policy / optimal Value Function

Fixed point iteration 不动点迭代

(1). Given s_0, π. $p(s_{t+1} \mid s_t, a_t)$. there exists a distribution over state trajectories



(2). Value Function:

$V^\pi(s) \triangleq \mathbb{E}$[total payoff from starting in s and using π].

$= \mathbb{E}_{\pi, p(s'|s,a)}[R(s_0,a_0) + \gamma R_1(s_1,a_1) + \gamma^2 R(s_2,a_2) + \cdots \mid s_0 = s]$.

Break joint prob. to conditional prob.

$\mathbb{E}_{x,y\sim p(x,y)}[f(x)+g(y)] = R(s_0,a_0) + \gamma \mathbb{E}_{\pi, p(s'|s,a)}[R(s_1,a_1) + \gamma R(s_2,a_2) + \gamma^2 R(s_3,a_3) + \cdots \mid s_0 = s]$.

$= \sum_x p(x)[f(x) + \mathbb{E}_{y\sim p(y|x)}[g(y)]]$.

$= R(s_0,a_0) + \gamma \sum_{s_1 \in S} p(s_1 \mid s_0, a_0)[\underbrace{R(s_1,a_1) + \gamma \mathbb{E}_{\pi, p(s'|s,a)}[R(s_2,a_2) + \cdots \mid s_1]}_{V^\pi(s_1)}]$.

递归式子.

(3). Bellman equation:

$$V^\pi(s) = R(s_0,a_0) + \gamma \sum_{s_1 \in S} p(s_1 \mid s_0, a_0) V^\pi(s_1).$$

for fixed π, system of $|S|$ equations and $|S|$ variables.

optimal policy. $\pi^* \triangleq \arg\max_\pi V^\pi(s)$. $\forall s$.

optimal value function $V^* \triangleq V^{\pi^*}$ (By def.).

Given V^*, $R(s,a)$, $p(s'|s,a)$. we can compute π^*.

$$\pi^*(s) = \arg\max_{a \in A} \underbrace{R(s,a)}_{\text{immediate reward}} + \gamma \underbrace{\sum_{s' \in S} p(s'|s,a) V^*(s')}_{\text{discounted future reward}}.$$

$$V^*(s) = \max_{a \in A} R(s,a) + \gamma \sum_{s' \in S} p(s'|s,a) V^*(s').$$