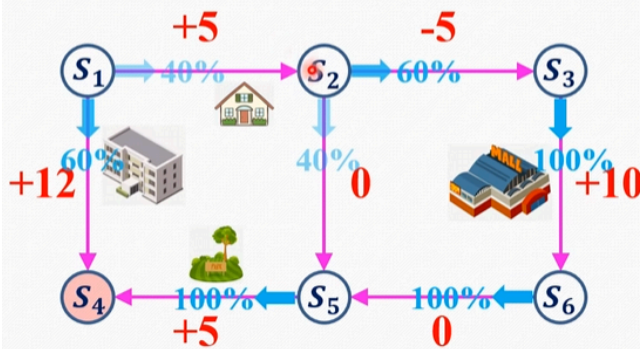


车辆盈利的例子（价值迭代算法）

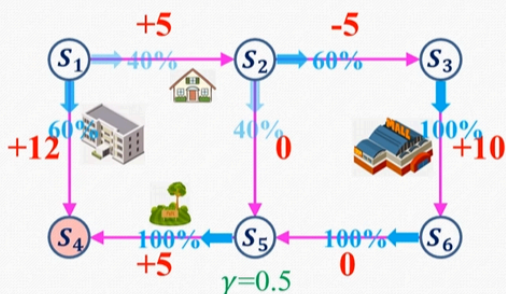
车辆的盈利 (More revenue)



- 车辆位于的结点为**状态**
- 所选择的道路为**动作**
- 每经过一段道路，**收益**已知
- 在每个状态，**决策**已知
- 折减系数 $\gamma=0.5$
- 到达 S_4 ，行驶结束
- 求每一结点的状态价值 $V(s)$
- 求每一结点的**最优**状态价值 $V_*(s)$
- 从哪里出发，如何行驶收益最大？

通过这样嵌套，达到终点后的状态价值应为0，然后再依次代回。这里没有涉及 $P_{ss'}^a$ 是因为每个动作 a 会导向确定的状态。

求每一状态的状态价值 $V(s)$

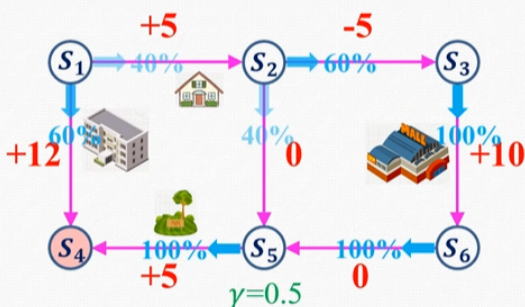


$$V_{\pi}(s) = \sum_{a \in A} \pi(a|s) [R_s^a + \gamma \sum_{s' \in S} P_{ss'}^a V_{\pi}(s')]$$

$$\begin{aligned} V_{\pi}(S_1) &= 40\%(5 + 0.5V_{\pi}(S_2)) + 60\%(12 + 0.5V_{\pi}(S_4)) & V_{\pi}(S_6) &= 100\%(0 + 0.5V_{\pi}(S_5)) \\ V_{\pi}(S_2) &= 60\%(-5 + 0.5V_{\pi}(S_3)) + 40\%(0 + 0.5V_{\pi}(S_5)) & V_{\pi}(S_5) &= 100\%(5 + 0.5V_{\pi}(S_4)) \\ V_{\pi}(S_3) &= 100\%(10 + 0.5V_{\pi}(S_6)) & V_{\pi}(S_4) &= 0 \end{aligned}$$

这样计算的是根据动作概率 $\pi(a|s)$ 计算出的每个状态的价值，但并非最优状态价值。

求每一状态**最优**状态价值 $V_*(s)$



0	0	0
0	0	0

第一次价值迭代

12	0	10
0	5	0

- 对于每个状态，初始化 $V(s) = 0$
- 重复循环取 \max $V_*(s) = \max_a [R_s^a + \gamma \sum_{s' \in S} P_{ss'}^a V_*(s')]$
- 至价值函数收敛，输出最优策略

求解**最优**状态价值就抛弃 $\pi(a|s)$ ，而是计算每个状态执行动作所带来的最高状态价值，不断迭代。