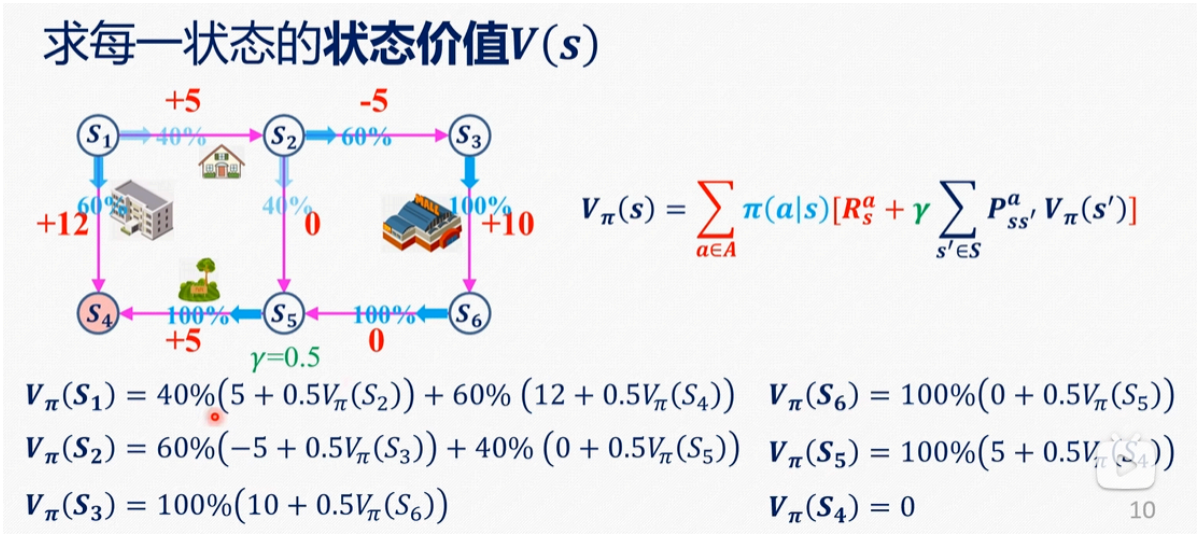
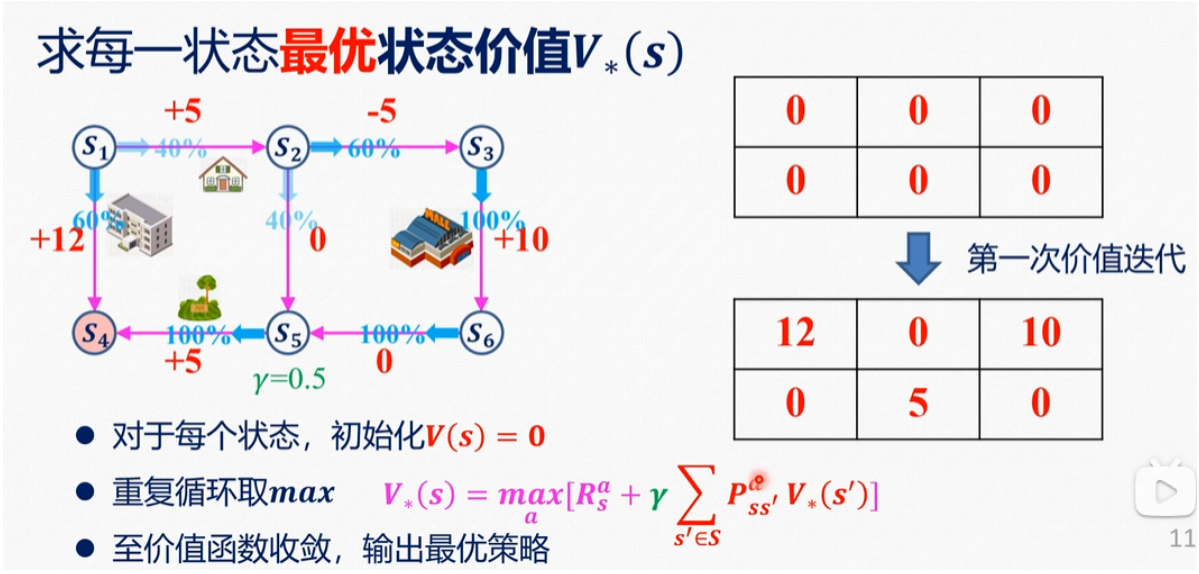
车辆盈利的例子（价值迭代算法）

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$P^a_{ss'}$是因为每个动作a会导向确定的状态。



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。