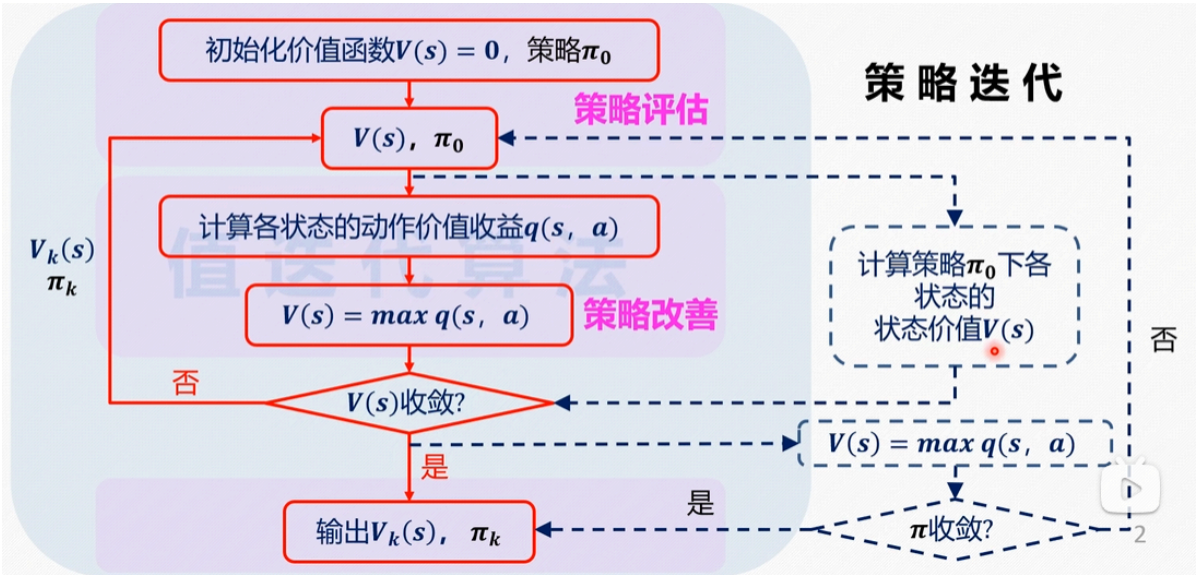





策略迭代

价值迭代是计算每个状态的状态价值后立刻计算出当前状态的最优价值，然后迭代。而策略迭代是先进行状态价值的迭代，当他收敛后再计算最优策略，然后用新的策略再迭代状态价值，直至最优策略收敛。




对于一个方格，初始时的策略就是上下左右各25%的可能性，对于边界的方格，仍使用四个方向计算，但在超过边界的方向执行动作后仍保持在原方格不动。

- 初始化 $V(s)$ 、 $\pi(s)$
- 进行策略评估，计算 $V(s)$ 至收敛
- 进行策略提升，贪婪思想 $\max q(s, a)$
- 至 $\pi(s)$ 收敛，输出 最优策略 π^* ，最优状态值 V^*

 0	 0	0
0	0	0
0	 0	0

策略评估



$$v_{\pi}(s) = 25\% \times (-1 + 0) + 25\% \times (-1 + 0) + 25\% \times (-1 + 0) + 25\% \times (-1 + 0) = -1$$

当第一次价值收敛时，计算最优策略，然后第二次的时候每个方格的状态价值就只计算刚才的最优动作产生的关联状态，即只计算一个动作而非四个动作，然后再迭代状态价值至收敛。