一. 策略评估：

指计算给定策略下状态价值函数的过程。从任意一个状态价值函数开始，依据给定的策略，结合 Bellman 方程、状态转移概率和奖励同步迭代更新状态价值函数，直至其收敛，得到该策略下最终的状态价值函数。

$$V_{k+1}(s) = \sum_{a \in A} \pi(a|s)\left(R_s^a + \gamma \sum_{s' \in S} P_{ss'}^a V_k(s')\right)$$

根据上述 Bellman 期望方程，我们可以对得到的新的状态价值函数再次进行迭代，直到状态价值函数收敛。

*例：

0	1	2	3
4	5	6	7
8	9	10	11
12	13	14	15

从 1→14 中任意 State 到达{0, 15}.
1~14 状态 reward 为 -1
0, 15 ... ... 0.

策略评估过程：

0.0	0.0	0.0	0.0
0.0	0.0	0.0	0.0
0.0	0.0	0.0	0.0
0.0	0.0	0.0	0.0

(a) k=0

0.0	-1.0	-1.0	-1.0
-1.0	-1.0	-1.0	-1.0
-1.0	-1.0	-1.0	-1.0
-1.0	-1.0	-1.0	0.0

(b) k=1

0.0	-1.7	-2.0	-2.0
-1.7	-2.0	-2.0	-2.0
-2.0	-2.0	-2.0	-1.7
-2.0	-2.0	-1.7	0.0

(c) k=2

0.0	-2.4	-2.9	-3.0
-2.4	-2.9	-3.0	-2.9
-2.9	-3.0	-2.9	-2.4
-3.0	-2.9	-2.4	0.0

(d) k=3

0.0	-6.1	-8.4	-9.0
-6.1	-7.7	-8.4	-8.4
-8.4	-8.4	-7.7	-6.1
-9.0	-8.4	-6.1	0.0

(e) k=10

0.0	-14	-20	-22
-14	-18	-20	-20
-20	-20	-18	-14
-22	-20	-14	0.0

(f) k=∞

只看 $V_k(1)$ 的更新过程：(其它 State 同 $V_k(1)$).
由于使用均一概率随机策略，且对于 State1 ~ State16 Value function 初始化为 0.0，即 $V_0(1) \sim V_0(16)$ 均为 0，且 $P_{ss'}^a = 0.25$ (对于任意 [State, action])

由于 $q_\pi(s,a) = R_s^a + \gamma \sum_{s' \in S} P_{ss'}^a \cdot V_\pi(s')$

$\Rightarrow$ 第 k 次的 Q 价值为:

$$q_k(s,a) = R_s^a + \gamma \sum_{s' \in S} P_{ss'}^a \cdot V_k(s')$$

又因为 $V_\pi(s) = \sum_{a \in A} \pi(a|s) \cdot q_\pi(s,a)$

$\Rightarrow$ 第 k+1 次的 Value Function 为:

$$V_{k+1}(s) = \sum_{a \in A} \pi(a|s) \cdot q_k(s,a)$$

$q_k(s,a)$ ↓

第 k 次的 q value 作为第 k+1 次的 q value，从而更新 k+1 次的 V-function.

则 $V_1(1) = 0.25 \cdot q_0(1, up) + 0.25 \cdot q_0(1, down) + 0.25 \cdot q_0(1, left) + 0.25 \cdot q_0(1, right)$.

又因为 $q_0(1, up) = R_1^{up} + 1 \times \sum_{s' \in S} P_{ss'}^{up} V_0(s')$

$= -1 + 1 \times 100\% \times 0 = -1$

-1 ↓ 此问题下取 1

100% ↓ 由于 State 1 执行 up 动作只会回到 1，则 100% 的概率转移到 1

0 ↓ $V_0(1)$ 为 0.

State-1 往上走回到 1，reward 为 -1

同理: $q_0(1, left) = -1 + 1 \times 100\% \times 0 = -1$

$q_0(1, right) = -1 + 1 \times 100\% \times 0 = -1$

$q_0(1, down) = -1 + 1 \times 100\% \times 0 = -1$

则 $V_1(1) = 0.25 \times -1 + 0.25 \times -1 + 0.25 \times -1 + 0.25 \times -1$

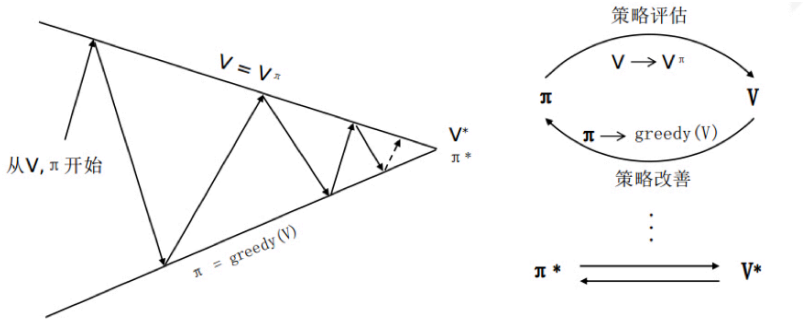
$= -1$

从 $k = 1 \sim +\infty$ (直到收敛) 依次更新 $V_k(1) \sim V_k(16)$ 即可得到最终策略评估.

二. 策略迭代：（通过 Value Function 的策略评估来调整策略）

在上述"策略评估"过程中，我们策略从开始更新到收敛一直使用均一随机策略，这样虽然可以求得最优解，但由于 π 在 Value 更新过程中始终维持初始策略，并未随着 Value 的改变而改变，因而收敛较慢。策略迭代可采用贪婪策略，即根据 Value 的改变，个体在处在在一状态时，将比较所有可能后续状态的价值，从中选择最大价值的状态，再选择能到该状态的行为，若最大价值状态有多个，则从多个最大价值状态中随机选择一个对应的行为。

当给定一个 π 时，可基于该 π 得到价值函数 V_π，基于 V_π 可得到贪婪策略 $\pi' = greedy(\pi)$，依据 π' 会得到 $V_{\pi'}$ 并产生 $\pi'' = greedy(\pi')$ 如此循环可求得 V^* 与 π^*，如下图：



且，$\pi'(s) = \arg\max_a q_\pi(s,a)$

* π 迭代过程中收敛后为什么就得到了 π^* 与 V^*？

假如个体与 Env 交互仅下一步采取该贪婪策略产生的行为，而后续步仍采用原策略行为，则下不等式成立：

$$q_\pi(s, \pi'(s)) = \max_{a \in A} q_\pi(s,a) \geq q_\pi(s, \pi(s)) = V_\pi(s)$$

若上式对 $s \in S$ 均成立，那么 s 后所有状态均用贪婪策略产生行为，不等式 $V_{\pi'}(s) \geq V_\pi(s)$ 成立，推导如下：

$$
\begin{aligned}
v_\pi(s) &\le q_\pi(s, \pi'(s)) = \mathbb{E}_{\pi'}\left[R_{t+1} + \gamma v_\pi(S_{t+1}) \mid S_t = s\right] \\
&\le \mathbb{E}_{\pi'}\left[R_{t+1} + \gamma q_\pi(S_{t+1}, \pi'(S_{t+1})) \mid S_t = s\right] \\
&\le \mathbb{E}_{\pi'}\left[R_{t+1} + \gamma R_{t+2} + \gamma^2 q_\pi(S_{t+2}, \pi'(S_{t+2})) \mid S_t = s\right] \\
&\le \mathbb{E}_{\pi'}\left[R_{t+1} + \gamma R_{t+2} + \ldots \mid S_t = s\right] = v_{\pi'}(s)
\end{aligned}
$$

如果在某一个迭代周期内，状态价值函数不再改善，即：

$$
q_\pi(s, \pi'(s)) = \max_{a \in A} q_\pi(s, a) = q_\pi(s, \pi(s)) = v_\pi(s)
$$

那么就满足了贝尔曼最优方程的描述：

$$
v_\pi = \max_{a \in A} q_\pi(s, a)
$$

此时，对于所有状态集内的状态 $s \in S$，满足：$v_\pi(s) = v_*(s)$，这表明此时的策略 π 即为最优策略。证明完成。

三. 价值迭代：

状态价值的最优化告诉我们：一个 State 的最优价值可由其后续 State 的最优价值通过 Bellman 方程计算：

$$
V_*(s) = \max_{a \in A}\left(R_s^a + \gamma \sum_{s' \in S} P_{ss'}^a V_*(s')\right)
$$

① 已知终态与终态价值，根据上述公式迭代求解所有状态最优价值：见 P37

② 个体不知道终态，通过迭代求解所有状态最优价值：P37

两种情形的相同点都是根据后续状态的价值，利用贝尔曼最优方程来更新得到前接状态的价值。两者的差别体现在：前者每次迭代仅计算相关的状态的价值，而且一次计算即得到最优状态价值，后者在每次迭代时要更新所有状态的价值。

迭代过程中更新公式为：

$$
V_{k+1}(s) = \max_{a \in A}\left(R_s^a + \gamma \sum_{s' \in S} P_{ss'}^a V_k(s')\right)
$$