23210980044 加兴华 PJ1计算题.

Q1. $V_{k+1}(s) \leftarrow \max_a \sum_{s'} T(s,a,s')[R(s,a,s') + \gamma V_k(s')]$ $R(\cdot) = 0$.

[step 0]

0 0 0 0

0 x 0 0

0 0 0 0

[step 1]

0 0 0 1

0 x 0 -1

0 0 0 0

[step 2]

0 0 0 / 0.8γ / 0 → $V_{[1,3]} = 0.8\gamma$

0 x 0 / -0.8γ / -1 → $V_{[2,3]} = 0$.

0 0 0 / -0.8γ / -0.8γ → $V_{[3,4]} = 0$.

运行HLN代码得 $V_{[1,3]} = 0.72$, $V_{other} = 0$. 对于 $\gamma = 0.9$ 符合计算结果.

Q2.1 记 A_1 一 A_6 为对应(1)-(6)对应最优轨迹

R为生存奖励 γ为衰减率

$R(A_1) = (1+\gamma+\cdots\gamma^2)\cdot R + \gamma^3$ (冒险+1路线)

$R(A_2) = (1+\gamma+\cdots\gamma^4)\cdot R + 10\cdot\gamma^5$ (冒险+10路线).

$R(A_3) = (1+\gamma+\cdots\gamma^6)\cdot R + \gamma^7$ (稳妥+1).

$R(A_4) = (1+\gamma+\cdots\gamma^8)\cdot R + 10\cdot\gamma^9$ (稳妥+10).

$R(A_5) = (1+\cdots\gamma^\infty)\cdot R = \frac{1}{1-\gamma}\cdot R$ (永久路线)

$R(A_6) = R - 10\gamma$ (放弃路线)

$$\begin{cases} R(A_2) - R(A_1) = \gamma^3[(1+\gamma)R + 10\gamma^2 - 1] = \frac{1}{\gamma^3}[R(A_4) - R(A_3)] \\ R(A_3) - R(A_1) = \gamma^3[(1+\gamma^3)R + \gamma^6 - 1] = \gamma^3(1+\gamma^3)(R+\gamma-1) \\ R(A_4) - R(A_2) = \gamma^5[(1+\gamma^3)R + (\gamma^6-1)10] = \gamma^5(1+\gamma^3)[R - 10(1-R)] \\ R(A_7) - R(A_3) = \gamma^7[\frac{1}{1-\gamma}R - 1] = \frac{\gamma^7}{1-\gamma}(R+\gamma-1) \\ R(A_8) - R(A_4) = \gamma^9[\frac{1}{1-\gamma}R - 10] = \frac{\gamma^9}{1-\gamma}[R - 10(1-\gamma)] \\ R(A_1) - R(A_6) = \gamma[(1+\gamma)R + \gamma^2 + 10] \end{cases}$$

R 存在两个零点 $-\frac{\gamma^2+10}{1+\gamma} < \frac{1-10\gamma^2}{1+\gamma} < 1-\gamma < 10(1-\gamma)$

Note: A_3 不可能最优. 因为当 $R(A_3) - R(A_1) > 0$, 必有 $R(A_5) - R(A_3) > 0$.

A_4 同理不可能最优.

当 $R < -\frac{\gamma^2+10}{1+\gamma}$ 时, $R(A_6)$ 最大. $\Rightarrow A_6$ 最优.

当 $R > -\frac{\gamma^2+10}{1+\gamma}$ 时, $\begin{cases} R < \frac{1-10\gamma^2}{1+\gamma} \Leftrightarrow A_1 \text{ 最优} \\ \frac{1-10\gamma^2}{1+\gamma} < R < 10(1-\gamma) \Leftrightarrow A_2 \text{ 最优} \\ R > 10(1-\gamma) \Leftrightarrow A_5 \text{ 最优} \end{cases}$

Q2.3

[$\gamma=1$]

当 $R\geq 0$. 最优策略为无尽路线 因为无终止轨迹回报→∞

当 $-1.2<R<0$. 最优策略为稳妥+10路线 { 此时 1步生存收益 < +10奖励收益 ∧ 4步风险

当 $-3.2<R<1.2$ 最优策略为风险+1路线 { 此时 2步风险 < +1奖励收益 < 4步风险 ∨ 1步生存收益

当 $R<-3.2$ 最优策略为放弃-10路线 { 此时生存奖励过低，策略希望尽快结束

[$\gamma=0.8$].

当 $R\geq 2$. 最优策略为无尽路线

当 $-1.2\leq R<2$. 最优策略为稳妥+10路线

当 $-4.2<R<-1.2$. 最优策略为风险+1路线

当 $R<-4.2$. 最优策略为放弃-10路线

} 同 $\gamma=1$ 时.

[$\gamma=0.6$]

当 $R\geq 4$ ⇒ 无尽路线

$-2.4\leq R<4$ ⇒ 稳妥+10路线

$-5.2\leq R<-2.4$ ⇒ 风险+1路线

$R<-5.2$ ⇒ 放弃-10路线

[$\gamma=0.9$]

当 $R \geq 6$ ⇒ 无尽路线

$-3.2 \leq R < 6$ ⇒ 稳妥+1路线

$-4.6 \leq R < -3.3$ ⇒ 无尽路线，

$-6.6 \leq R \leq -4.7$ ⇒ 风险+1路线

$R \leq -6.7$ ⇒ 放弃-1路线

此时
- $S(0,1)$ 希望稳妥走向终点 T，
- $(0,2)$ 由于 γ 衰减希望走风险路线↓
- ∴ 策略为 up, down 循环
- 而不是真的无尽收益 > 终止收益

[$\gamma=0.2$]

当 $R \geq 8$ ⇒ 无尽路线

$0.8 \leq R < 8$ ⇒ 稳妥+1路线

$-7.9 \leq R \leq 0.1$ ⇒ 无尽路线. 原因同 $\gamma=0.9$ 时

$-8.1 \leq R \leq -7.5$ ⇒ 风险+1路线

$R \leq -8.2$ ⇒ 放弃路线

下证：对于任一策略，其效用值随生存奖励线性变化. (已知 γ、π、terminal R)

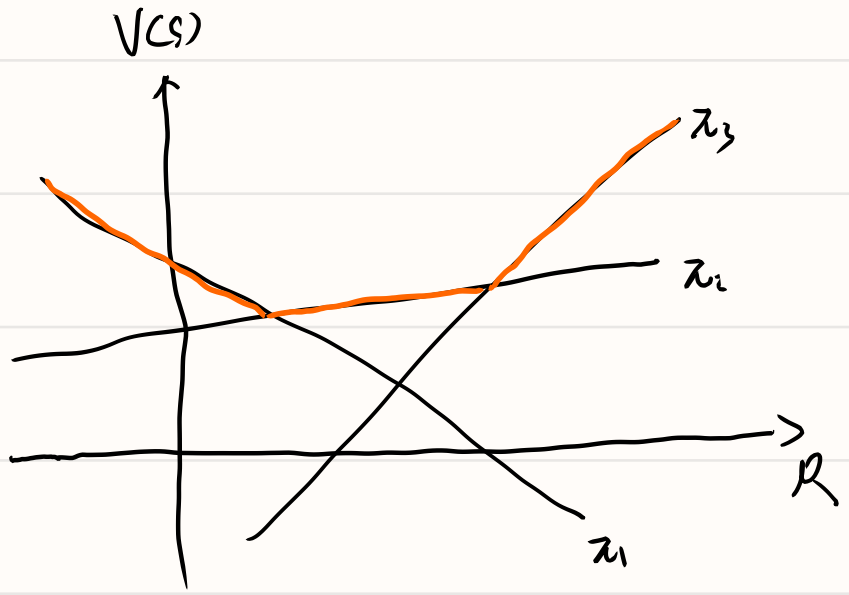
$$V^{\pi}(s) = \sum p_i \cdot V^{\pi}(\text{trajectory}_i) \quad p_i \text{ 为 trajectory}_i \text{ 的概率}$$

$$V^{\pi}(\text{trajectory}_i) = \begin{cases} 1 + \gamma R + \gamma^2 R + \cdots \gamma^n \cdot \text{terminal } R & \text{有限轨迹} \\ 1 + \gamma R + \gamma^2 R + \cdots \gamma^{\infty} R = \frac{1}{1-\gamma} R & \text{无限轨迹} \end{cases}$$

易见 R 系数总 $\leq \frac{1}{1-\gamma}$.

∴ $V^{\pi}(s) = c_1 \cdot R + \sum_{n=1} c_n \cdot \text{terminal } R_n$. $c_1 < \sum p_i \frac{1}{1-\gamma} = \frac{1}{1-\gamma}$ 有界.

∴ $V(s) = \max_{\pi} V(s)$ 可知 $V(s)$ 为 R 的分段线性函数. 形如下：



∴对于上述讨论，只需找策略分界点即可.