Ex 10-1

(a)

	roof	sleigh	terminal
roof	0.5	0.4	0.1
sleigh	0.4	0.5	0.1
terminal	0	0	0

$$P = \begin{pmatrix} 0.5 & 0.4 & 0.1 \\ 0.4 & 0.5 & 0.1 \\ 0 & 0 & 0 \end{pmatrix}$$

(b)

$$U = R + \gamma P U$$

$$\Rightarrow (I - \gamma P)\, U = R$$

expensive

$$\Rightarrow U = (I - \gamma P)^{-1} \cdot R \qquad O(n^3)$$

$$\Rightarrow \left[\begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix} - 1 \cdot \begin{pmatrix} 0.5 & 0.4 & 0.1 \\ 0.4 & 0.5 & 0.1 \\ 0 & 0 & 0 \end{pmatrix}\right] \begin{bmatrix} u_r \\ u_s \\ u_t \end{bmatrix} = \begin{bmatrix} -2 \\ -1 \\ 0 \end{bmatrix}$$

$$\Rightarrow \begin{pmatrix} 0.5 & -0.4 & -0.1 \\ -0.4 & 0.5 & -0.1 \\ 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} u_r \\ u_s \\ u_t \end{pmatrix} = \begin{pmatrix} -2 \\ -1 \\ 0 \end{pmatrix}$$

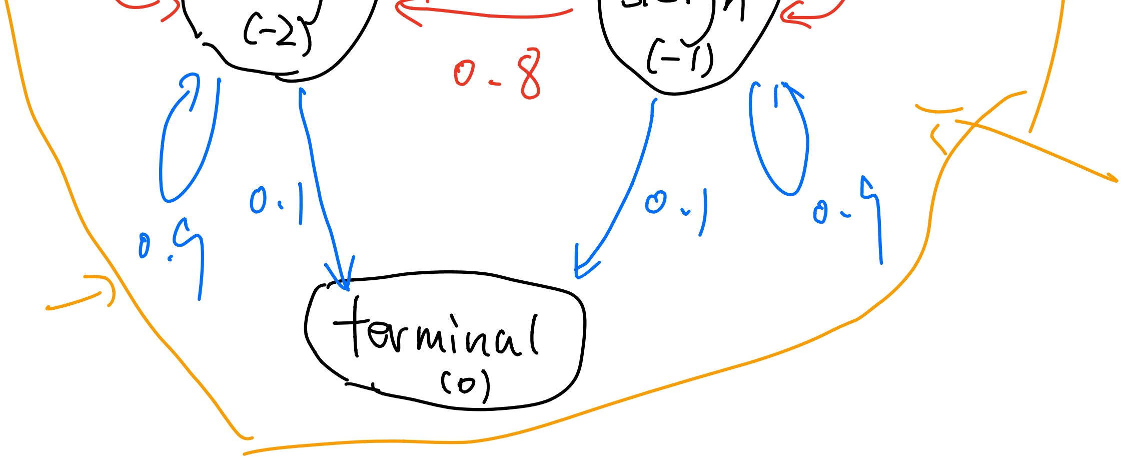
$$\Rightarrow \begin{cases} u_s = -\frac{130}{9} \\ u_r = -\frac{140}{9} \\ u_t = 0 \end{cases}$$

Ex 10-2

action: c (change)

action: t (throw)

a)



$$p(x_s \mid x_r, c) = 0.8 \qquad p(x_t \mid x_r, t) = 0.1$$

$$p(x_s \mid x_s, c) = 0.2 \qquad p(x_r \mid x_r, t) = 0.9$$

$$p(x_r \mid x_s, c) = 0.8 \qquad p(x_t \mid x_s, t) = 0.1$$

$$p(x_r \mid x_r, c) = 0.2 \qquad p(x_s \mid x_s, t) = 0.9$$

---

c) $\boxed{\begin{aligned} \pi_0(x_s) &= t \\ \pi_0(x_r) &= t \end{aligned}}$ c c (initial)

① policy evaluation:

$$U_s = R_s + \gamma \sum_{s' \in S} p(s' \mid s, a)\, U(s')$$

$$= -1 + 0.9 \cdot U_s + 0.1 \cdot U_t$$

similarly:

$$\left.\begin{aligned} U_r &= -2 + \gamma(0.9\, U_r + 0.1 \cdot U_t) \\ U_t &= 0 \end{aligned}\right\} \text{unsolved.}$$

$$\Rightarrow \begin{cases} u_s = -10 \\ u_r = -20 \\ u_t = 0 \end{cases}$$

$$\arg\max_x f(x)$$

② policy improvement.

$$U = R + \arg\max_a \left(r \underbrace{\sum_{s'} p(s'|s,a)\, U}_{Q_t} \right)$$

$$\Rightarrow \quad T(x_i, a) = \sum p(s'|s,a)\, U$$

$$T(x_s, c) = 0.8 \times (-20) + 0.2 \times (-10) = -18$$

$$T(x_s, t) = 0.9 \times (-10) + 0.1 \times (0) = -9$$

$$T(x_r, c) = 0.8 \times (-10) + 0.2 \times (-20) = -12$$

$$T(x_r, t) = 0.9 \times (-20) + 0.1 \times (0) = -18$$

$$T(x_s, c) < T(x_s, t) \Rightarrow \pi_1(x_s) = t$$

$$T(x_r, c) > T(x_r, t) \Rightarrow \pi_1(x_r) = c$$

③ policy evaluation

$$\begin{cases} u_s = -1 + 0.9 u_s + 0.1 u_t \\ u_r = -2 + 0.8 u_r + 0.2 u_s \end{cases} \Rightarrow \begin{cases} u_s = -10 \\ u_r = -12.5 \end{cases}$$

$$\begin{cases} u_r = -2 + 0.8\, u_s + 0.2\, u_r \\ u_t = 0 \end{cases} \qquad u_t = 0$$

④ policy inprovment.

$$T(x_r, c) = 0.8 \times (-10) + 0.2 \times (-12.5) = -10.5$$

$$T(x_r, t) = 0.9 \times (-12.5) + 0.1 \times 0 = -11.25$$

$$T(x_s, c) = 0.8 \times (-12.5) + 0.2 \times (-10) = -12$$

$$T(x_s, t) = 0.9 \times (-10) + 0.1 \times 0 = -9$$

$$T(x_s, c) < T(x_s, t) \qquad \boxed{\begin{aligned} \pi_2(x_s) &= t \\ \pi_2(x_r) &= c \end{aligned}}$$

$$T(x_r, c) > T(x_r, t)$$

terminate. π_2 is optimal