#1

Since no reward is before the terminal state of the episode, and no penalty is assigned for taking unnecessary steps in progress to the terminal state, discounting or otherwise, the robot has no incentive to minimize the steps taken to the terminal state. In fact, a random walk will eventually arrive at the same score as a direct path, so the robot never develops any path improvements.

$G_t = 1 \; \forall \; t$, regardless of action

#2

$$V_\pi(s) = E_\pi\left[G_t \mid S_t = s\right]$$

$$= \sum_a \pi(a|s) \sum_{s',r} p(s', r|s, a)\left[r + \gamma V_\pi(s')\right]$$

$$V_\pi(\text{center}) = (0.25)(1)\left[0 + 0.9(-0.4)\right]$$

$$+(0.25)(1)\left[0.9(0.7)\right]$$

$$+0.25\left[0.9(2.3)\right]$$

$$+0.25\left[0.9(0.4)\right]$$

$$= -0.09 + 0.1575 + 0.5175 + 0.09$$

$$= \underline{0.675 \approx 0.7}$$

#3

$$q_\pi(s,a) = \mathbb{E}_\pi[G_t \mid S_t = s, A_t = a]$$

$$= \mathbb{E}_\pi[R_{t+1} + \gamma G_{t+1} \mid S_t = s, A_t = a]$$

$$= \sum_{s'} \sum_{r} p(s', r \mid s, a)\left[r + \gamma \mathbb{E}\left[G_{t+1} \mid S_{t+1}, A_{t+1}\right]\right]$$

$$= \sum_{s'} \sum_{r} p(s', r \mid s, a)\left[r + \gamma q_\pi(s', a')\right]$$

#4

$$V_\pi = R^\pi + \gamma P^\pi V_\pi$$

$$V_\pi - \gamma P^\pi V_\pi = R^\pi$$

$$(I - \gamma P^\pi) V_\pi = R^\pi$$

$$V_\pi = (I - \gamma P^\pi)^{-1} R^\pi$$

$$V_{\pi c} = (I - \gamma P^\pi)^{-1}(R^\pi + c\bar{J})$$

$$V_{\pi c} - V_\pi = V_c$$

$$V_c = (I - \gamma P^\pi)^{-1}(R^\pi + c\bar{J}) - (I - \gamma P^\pi)^{-1} R^\pi$$

$$= (I - \gamma P^\pi)^{-1}\left[(R^\pi + c\bar{J}) - R^\pi\right]$$

$$V_c = (I - \gamma P^\pi)^{-1}(c\bar{J})$$

γ, P^π, c are constant under any given policy π

$\Rightarrow V_c$ is constant under any given policy π.

$$\bar{J} = \begin{bmatrix} 1 & \cdots & 1 \\ \vdots & \ddots & \vdots \\ 1 & \cdots & 1 \end{bmatrix}$$

□

#5

Consider example 3.4, pole-balancing.

Suppose a reward of +1 is given at every time interval without failure, and a reward of 0 for failure and termination of the episode.

Also consider a reward of 0 at every time interval and -1 at failure with discount factor $\gamma < 1$.

In either case, maximization of G_t would involve keeping the pole balanced as long as possible.

The result from the previous exercise holds.

#6

$$q_* = \mathbb{E}\left[R_{t+1} + \gamma \max_{a'} q_*(S_{t+1}, a') \mid S_t = s, A_t = a\right]$$

$$= \sum_{s', r} p(s', r \mid s, a)\left[r + \gamma \max_{a'} q_*(s', a')\right]$$

$$\max_{a'} q_*(s', a') = v_*(s')$$

$$\therefore$$

$$q_*(h, s) = p(h, r_{search} \mid h, s)\left[r_{search} + \gamma v_*(h)\right] + p(l, r_{search} \mid h, s)\left[r_{search} + \gamma v_*(l)\right]$$

$$= \alpha\left[r_{search} + \gamma v_*(h)\right] + (1-\alpha)\left[r_{search} + \gamma v_*(l)\right]$$

$$q_*(h, w) = p(h, r_{wait} \mid h, w)\left[r_{wait} + \gamma v_*(h)\right] + p(l, r_{wait} \mid h, w)\left[r_{wait} + \gamma v_*(l)\right]$$

$$= r_{wait} + \gamma v_*(h)$$

$$q_*(l, w) = p(h, r_{wait} \mid l, w)\left[r_{wait} + \gamma v_*(h)\right] + p(l, r_{wait} \mid l, w)\left[r_{wait} + \gamma v_*(l)\right]$$

$$= r_{wait} + \gamma v_*(l)$$

$$q_*(l, s) = p(h, r_{search} \mid l, s)\left[r_{search} + \gamma v_*(h)\right] + p(l, r_{search} \mid l, s)\left[r_{search} + \gamma v_*(l)\right]$$

$$= (1-\beta)\left[-3 + \gamma v_*(h)\right] + \beta\left[r_{search} + \gamma v_*(l)\right]$$

$$q_*(l, re) = p(h, r_{recharge} \mid l, re)\left[r_{recharge} + \gamma v_*(h)\right] + p(l, r_{recharge} \mid l, re)\left[r_{recharge} + \gamma v_*(l)\right]$$

$$= \gamma v_*(h)$$

where $v_*(h)$, $v_*(l)$ are defined in example 3.9, page 65.