# CS 234 H.W #1 

Adhitya Venkatesh

1. To show this expression holds $\forall i \in \{1, \dots, H\}$, we proceed by mathematical induction:

<u>Base Case ($i = H$)</u>: As $\hat{V}_{H+1}(s) = 0 \ \forall s \in S$,

$$\hat{V}_H(s) - V_H(s) = \sum_{t=H}^{H} E\left[\hat{r}(s_t, a_t) - r(s_t, a_t) \,\middle|\, s_H = s\right]$$

$$= E[\hat{r}(s_H, a_H)] - E[r(s_H, a_H)]$$

Hence, this checks out trivially.

<u>Inductive step</u>: Suppose the expression holds for $i = T+1$ (inductive hypothesis). We now show it holds for $i = T$:

$$\hat{V}_T(s) - V_T(s) = \sum_{t=T}^{H} E\left[(\hat{r}(s_t, a_t) - r(s_t, a_t)) \,\middle|\, s_T = s\right]$$

$$= E[(\hat{r}(s_T, a_T) - r(s_T, a_T))] + \sum_{t=T+1}^{H} E\left[\hat{r}(s_t, a_t) - r(s_t, a_t) \,\middle|\, s_{T+1} = s\right]$$

Using the Bellman equation, this becomes

$$E\left[\hat{r}(s_T, a_T) - r(s_T, a_T)\right] + \sum_{s'} \Big(\hat{P}(s'|s_T, a_T)\hat{V}_{T+1}(s') - P(s'|s_T, a_T)V_{T+1}(s')\Big)$$

$2^{nd}$ term can be re-written as

$$\sum_{s'} (\hat{P} - P)\hat{V}_{T+1}(s') + \sum_{s'} P\Big(\hat{V}_{T+1}(s') - V_{T+1}(s')\Big)$$

Now making the substitution via inductive hypothesis, the $2^{nd}$ term above becomes

$$\sum_{s'_{T+1}} P(s'_{T+1}|s, a) \sum_{t=T+1}^{H} E\Big[\hat{r}(s_t, a_t) - r(s_t, a_t) + \sum_{s'} (\hat{P} - P)\hat{V}_{t+1}(s) \Big| s = s'_{T+1}\Big]$$

$$= \sum_{t=T+1}^{H} E\Big[\hat{r}(s_t, a_t) - r(s_t, a_t) + \sum_{s'} (\hat{P} - P)\hat{V}_{t+1}(s) \Big| s = s_T\Big]$$

Adding this to $\sum_{s'} (\hat{P} - P)\hat{V}_{T+1}(s') + E\left[\hat{r}(s_T, a_T) - r(s_T, a_T)\right]$

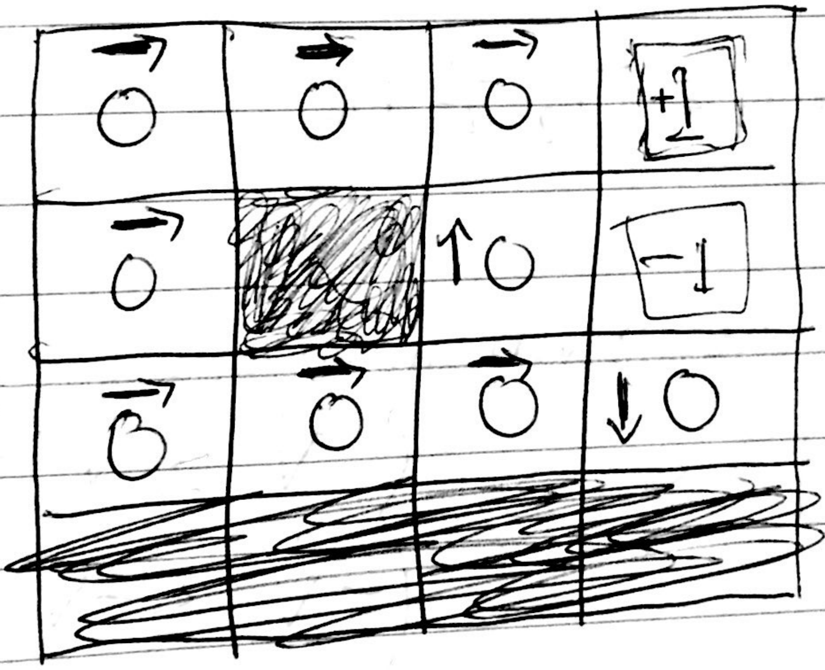
we obtain the desired result

$\therefore$ By principle of mathematical induction, we are done.

2. a) This situation is certainly a possibility if the value function hasn't converged and extraction of the current policy yields a suboptimal policy that is stuck for the moment.

Ex) Consider the following gridworld setup:

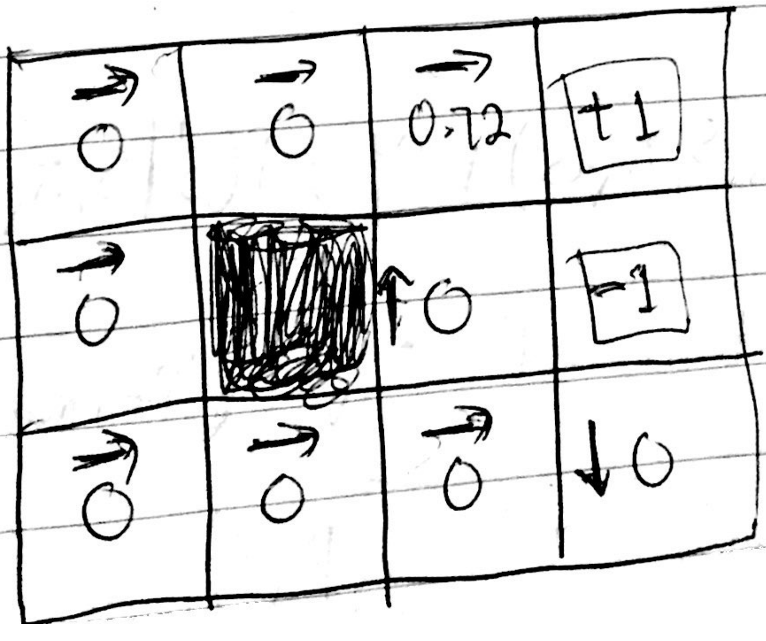
i=1:




Terminal states & corresponding rewards

- Arrows indicate policy.
- Shaded squares are walls, disregard last row

i=2:



$i = 3$:

| | | | |
|---|---|---|---|
| $\overrightarrow{0}$ | $\overrightarrow{0.52}$ | $\overrightarrow{0.78}$ | $\boxed{1}$ |
| $\overrightarrow{0}$ | | $\uparrow 0.43$ | $\boxed{-1}$ |
| $\overrightarrow{0}$ | $\overrightarrow{0}$ | $\uparrow 0$ | $\downarrow 0$ |

Clearly $\pi_1 = \pi_2 \neq \pi_3$

**b)** From the given information,

$$\| V^* - \tilde{V} \| \leq \varepsilon \Rightarrow \begin{cases} \tilde{V} \leq V^* + \varepsilon & -① \\ -\tilde{V} \leq -V^* + \varepsilon \end{cases}$$

So,

$$V_{\pi_{\tilde V}}(s) = R(s, \pi_{\tilde V}(s)) + \gamma \sum_{s'} P(s' | s, \pi_{\tilde V}(s)) \, V_{\pi_{\tilde V}}(s)$$

By definition of $\pi_{\tilde V}(s)$ (Bellman update of $\tilde V$)

$$= \left\{ \begin{aligned} &\max_a \left[ R(s,a) + \gamma \sum_{s'} P(s'|s,a) \cdot \tilde{V}(s') \right] \\ &\quad + \gamma \sum P(s'|s,a) \left( V_{\pi_{\tilde V}}(s') - \tilde{V}(s') \right) \end{aligned} \right\} -②$$

Substituting the terms from ① into ②, we have

$$= \max_a \left[ R(s,a) + \gamma \sum_{s'} P(s'|s,a) \cdot (V^*(s') + \varepsilon) \right]$$
$$+ \gamma \sum_{s'} P(s'|s, \pi_{\tilde{V}}(s)) (V_{\pi_{\tilde{V}}}(s') - V^*(s') + \varepsilon)$$

$$= \max_a \left[ R(s,a) + \gamma \sum_{s'} p(s'|s,a) V^*(s') \right] + \gamma \varepsilon \sum_{s'} P(s'|s,a)$$
$$+ \gamma \sum_{s'} P(s'|s, \pi_{\tilde{V}}(s)) (V_{\pi_{\tilde{V}}}(s') - V^*(s')) + \gamma \varepsilon \sum_{s'} P(s'|s,a)$$

As $V^*$ is optimal, $BV^* = V^*$ and
$\sum_{s'} P(s'|s,a) = 1$ ($\because$ probability distribution)

$$= V^*(s) + 2\gamma\varepsilon + \gamma \sum_{s'} P(s'|s, \pi_{\tilde{V}}(s)) (V_{\pi_{\tilde{V}}}(s') - V^*(s'))$$

$$\leq V^*(s) + 2\gamma\varepsilon + \gamma (V_{\pi_{\tilde{V}}}(s) - V^*(s))$$

This is because ~~the~~ the contraction property of the Bellman operator i.e

$$B(V_{\pi_{\tilde{V}}}(s) - V^*(s)) \leq \gamma (V_{\pi_{\tilde{V}}}(s) - V^*(s))$$

Solving for $L_{\tilde{V}}(s)$ after re-arranging terms, we obtain

$$\boxed{L_{\tilde{V}}(s) \leq \frac{2\gamma\varepsilon}{1-\gamma}}$$

3. a.) In order to ensure the optimal policy leads to shortest path, we define $R(s) = -1$ $\forall s \in S'$ where $S' = \{1, \dots, 24\} \setminus \{5, 11\}$. This is forced because the agent is penalized for every move taken to reach $s = 5$.

b.) As $\gamma = 1$, $V_i(S) = E_\pi [R(s) + V_{i-1}(s')]$ backpropagating from $s = 5$ yields

$$V(s) = \begin{cases} 5 & s = 5 \\ -5 & s = 11 \\ 4 & s = 0, 6, 10 \\ 3 & s = 1, 7, 15 \\ 2 & s = 2, 8, 12, 16, 20 \\ 1 & \\ 0 & s = 4, 14, 18, 22 \\ -1 & s = 19, 23 \\ -2 & s = 24 \end{cases}$$

c.) This changes state values but not the optimal policy because discounting only adds further incentive to reduce number of moves required as rewards will still be negative $\forall s \neq 5$. Thus, the best strategy is to still take the shortest path.

d.) Now we have $R(s)=0 \ \forall s \in S$. So ~~the~~ the agent is no longer penalized for moving to any $s \neq 5, 11$. However, as the discount factor is no longer 1, the agent is still incentivized to reach state 5 in the fewest number of moves. Hence, once again shortest path is the optimal strategy to maximize reward. The value of each state will therefore be equal to that of the previous part for the same corresponding state plus an increment of

$$\sum_{i=0}^{L-1} \gamma^i$$

where $L=$ length of shortest path to state 5 $\forall s \in S$

4 c.)

Stochasticity in the second environment increases number of required iterations for convergence. The policies in both environments are identical.