This handout includes space for every question that requires a written response. Please feel free to use it to handwrite your solutions (legibly, please). If you choose to typeset your solutions, the —README.md— for this assignment includes instructions to regenerate this handout with your typeset LaTeX solutions.

1.a

$r_s = 1$ Take longest possible path (2,10,18,24,30,31,32) to target square.
$r_s = 0$ Take shortest possible path to target square
$r_s = -1$ Take shortest possible path to target square
$r_s = -4$ Take shortest possible path to red square

In general, the optimal policy depends on γ (think what would happen if $\gamma = 0$).

In general the optimal policy is not unique.

Given the additional restrictions we apply in this question the following answers should also recieve marks where the student considers policies that are only defined for the states that are reachable from the starting state (Note: in general a policy is defined for all states in the state space irrespective of the starting state nonetheless the following will also be accepted given assumptions made are outlined by the student):

In the case, for $r_s = -1, r_s = 0$ multiple equivalent optimal deterministic policies exist and hence the optimal deterministic policy is not unique (not counting actions in the terminal state or states where an action doesn't influence the next state e.g. states with a wall directly to the right. Also assuming a deterministic policy).

For the other cases the optimal deterministic policy is unique (not counting actions in the terminal state or states where an action doesn't influence the next state e.g. states with a wall directly to the right. Also assuming a deterministic policy).

1.b

$r_s = 0$

-5	-5	-5	-5	-5
3.2805	-4.5	-5	4.05	4.05
2.95245	3.645	-5	4.5	4.5
3.2805	3.2805	4.05	4.05	5
2.95245	-5	3.645	4.5	-4.05
-4.5	-5	4.05	-3.645	-4.5
-5	-5	-5	-5	-5

Right and Up is the optimal action sequence from square 27.

1.c

$r_s = 1$ Loop infintely.
$r_s = 0$ Take shortest possible path to target square.
$r_s = -1$ Take shortest possible path to target square.
$r_s = -4$ Take shortest possible path to red square.

$r_s = 0$

-5	-5	-5	-5	-5
3.2805	-4.5	-5	2.657205	1.7433922
1.93710245	3.645	-5	4.5	2.95245
3.2805	2.15233605	4.05	2.657205	5
1.93710245	-5	2.3914845	4.5	2.95245
-4.5	-5	4.05	2.657205	1.7433922
-5	-5	-5	-5	-5

Right and Up is the optimal actions sequence from square 27.

1.d

The value of a state induced by following policy π is defined as the expected sum of discounted rewards through following π, which may be represented as follows:

$V^{\pi}_{\text{old}}(s) = \mathbb{E}_{\pi}[G_{\text{old},t}|x_t = s]$

$G_{\text{old},t} = r_{\text{old},t+1} + \gamma r_{\text{old},t+2} + \gamma^2 r_{\text{old},t+3}... = \sum_{k=0}^{\infty} \gamma^k r_{\text{old},t+k+1}$

Given constants $a, c \in \mathbb{R}$ we translate and scale the rewards as follows:

$r_{\text{new}} = a(c + r_{\text{old}})$

We can define the updates discounted sum of rewards for the new MDP as $G_{\text{new},t}$ yielding the following expressions,

$V^{\pi}_{\text{new}}(s) = \mathbb{E}_{\pi}[G_{\text{new},t}|x_t = s]$

$G_{\text{new},t} = a(c + r_{\text{old},t+1}) + \gamma a(c + r_{\text{old},t+2}) + \gamma^2 a(c + r_{\text{old},t+3})... = \sum_{k=0}^{\infty} \gamma^k a(c + r_{\text{old},t+k+1})$

$= ac \sum_{k=0}^{\infty} \gamma^k + a \sum_{k=0}^{\infty} \gamma^k r_{\text{old},k+t+1} = \frac{ac}{1-\gamma} + a \sum_{k=0}^{\infty} \gamma^k r_{\text{old},k+t+1}$

Given we wish to relate $V^{\pi}_{\text{new}}(s)$ to $V^{\pi}_{\text{old}}(s)$ we can consider plugging our expression for the updated discounted sum of rewards into our expression for $V^{\pi}_{\text{new}}(s)$:

$V^{\pi}_{\text{new}}(s) = \mathbb{E}_{\pi}[G_{\text{new},t}|x_t = s]$

$= \mathbb{E}_{\pi}[\frac{ac}{1-\gamma} + a \sum_{k=0}^{\infty} \gamma^k r_{\text{old}k+t+1}|x_t = s]$

$= \mathbb{E}_{\pi}[\frac{ac}{1-\gamma}|x_t = s] + a\mathbb{E}_{\pi}[\sum_{k=0}^{\infty} \gamma^k r_{\text{old},k+t+1}|x_t = s]$

$= \frac{ac}{1-\gamma} + aV^{\pi}_{\text{old}}(s)$

$\therefore V^{\pi}_{\text{new}}(s) = \frac{ac}{1-\gamma} + aV^{\pi}_{\text{old}}(s)$

2.a

Using the performance difference lemma we may write:

$$(V_t^{\pi^+} - V_t^{\pi})(s_{\text{start}}) = \sum_{t=1}^{H} \mathbb{E}_{(s,a)\sim\pi}\Big[(Q_t^{\pi^+}(s,\pi^+(s)) - Q_t^{\pi^+}(s,a))\mathbb{1}\{s \in \mathcal{S}^+ \cap a \neq a^+\}\Big]$$
$$+ \mathbb{E}_{(s,a)\sim\pi}\Big[(Q_t^{\pi^+}(s,\pi^+(s)) - Q_t^{\pi^+}(s,a))\mathbb{1}\{s \notin \mathcal{S}^+\}\Big]$$

For any $t \in \{1, 2, ..., H\}$ we can write,

$$\mathbb{E}_{(s,a)\sim\pi}\Big[(Q_t^{\pi^+}(s,\pi^+(s)) - Q_t^{\pi^+}(s,a))\mathbb{1}\{s \in \mathcal{S}^+ \cap a \neq a^+\}\Big]$$
$$+ \mathbb{E}_{(s,a)\sim\pi}\Big[(Q_t^{\pi^+}(s,\pi^+(s)) - Q_t^{\pi^+}(s,a))\mathbb{1}\{s \notin \mathcal{S}^+\}\Big]$$
$$= \mathbb{E}_{(s,a)\sim\pi}\Big[(Q_t^{\pi^+}(s,\pi^+(s)) - Q_t^{\pi^+}(s,a))\mathbb{1}\{s \in \mathcal{S}^+ \cap a \neq a^+\}\Big]$$
$$= \mathbb{E}_{(s,a)\sim\pi}\Big[(H - t - H - \mathbb{E}_{s'\sim p(s,a)} V_{t+1}^{\pi^+}(s'))\mathbb{1}\{s \in \mathcal{S}^+ \cap a \neq a^+\}\Big] \leq 0$$

From here we have the claim since each term is our sum is less than or equal to zero.

3.a

$$
\begin{aligned}
||B_k V - B_k V'||_\infty &= ||max_a[R(s,a) + \gamma_k \sum_{s' \in \mathcal{S}} p(s'|s.a) V_k(s') \\
&\quad - max_a[R(s,a) + \gamma_k \sum_{s' \in \mathcal{S}} p(s'|s,a) V'_k(s')]||_\infty \\
&\leq max_a||\gamma_k \sum_{s' \in \mathcal{S}} p(s'|s,a)(V_k(s') - V'_k(s'))||_\infty \\
&\leq max_a||\gamma_k||V_k - V'_k|| \sum_{s \in \mathcal{S}} p(s'|s,a)||_\infty \\
&= \gamma_k||V - V'||_\infty
\end{aligned}
$$

3.b

$$
\begin{aligned}
||B_1(B_2...B_K V_K) - B_1(B_2...B_k V'_K)||_\infty &\leq \gamma_1 ||B_2(B_3...B_k V_K) - B_2(B_3...B_K V'_K)||_\infty \\
&\leq \gamma_1\gamma_2 ||B_3(B_4...B_K V_K) - B_3(B_4...B_K V'_K)||_\infty \\
&... \\
&\leq \gamma_1\gamma_2...\gamma_K ||V_K - V'_K||_\infty
\end{aligned}
$$

3.c

$$\begin{aligned}\gamma_1\gamma_2...\gamma_K &= (1-\frac{1}{2})(1-\frac{1}{3})(1-\frac{1}{4})...(1-\frac{1}{K+1}) \\ &= (\frac{1}{2})(\frac{2}{3})(\frac{3}{4})...(\frac{K}{K+1}) \\ &= \frac{1}{K+1}\end{aligned}$$