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

Enumerating the possible paths from square 2 for each value of r_s:
Longest path: $2 \rightarrow 10 \rightarrow 18 \rightarrow 24 \rightarrow 30 \rightarrow 31 \rightarrow 32$

For each of the possible values of r_s, the optimal policy would be:
$r_s = 1$ for the longest possible path to target square
$r_s = 0$ for the shortest possible path to target square
$r_s = -1$ for the shortest possible path to target square
$r_s = -4$ for the shortest possible path to a red square

In each case, the optimal policy is not unique and the optimal policy does depend on the value of the discount factor γ (it would definitely be affected if $\gamma = 0$). For our shortest path to the target square ($r_s = -1$), there are multiple equivalent optimal policies that exist.

1.b

The r_s values that would cause the optimal policy to return the shortest path to the green target square in all cases are $r_s = -1$ and $r_s = 0$

The optimal value function for $r_s = -1$

-5	-5	-5	-5	-5
-0.1585	-5.5	-5	2.15	2.15
-1.14265	0.935	-5	3.5	3.5
-0.1585	-0.1585	2.15	2.15	5
-1.14265	-5	0.935	3.5	-5.95
-5.5	-5	2.15	-5.5	-5.5
-5	-5	-5	-5	-5

The optimal action from square 27 is right and down

The optimal value function for $r_s = 0$

-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

The optimal action from square 27 is right and up

1.c

Possible values of r_s and the optimal policy for Flappy World 2:
$r_s = 1$ for an infinite loop
$r_s = 0$ for the shortest possible path to the target square
$r_s = -1$ for the shortest possible path to the target square
$r_s = -4$ for the shortest possible path to a red square

The value of r_s that would cause the optimal policy to return the shortest path to the green target square for all cases are: $r_s = -1$ nad $r_s = 0$

$r_s = -1$

-5	-5	-5	-5	-5
-0.1585	-5.5	-5	-2.0284	-4.7698
-4.0189	0.935	-5	3.5	-1.1427
-0.1585	-3.543	2.15	-2.0284	5
-4.1887	-5	-2.8256	3.5	-1.1427
-5.5	-5	2.15	-2.0284	-4.7698
-5	-5	-5	-5	-5

The optimal action from square 27 is right and up.

$r_s = 0$

-5	-5	-5	-5	-5
3.2805	-4.5	-5	2.6572	1.7434
1.9371	3.645	-5	4.5	2.9525
3.2805	2.15234	4.05	2.6572	5
1.9371	-5	2.3915	4.5	2.9525
-4.5	-5	4.05	2.6572	1.7434
-5	-5	-5	-5	-5

The optimal action from square 27 is right and up.

1.d

Given:

- General MDP $< S, A, P, R, \gamma >$
  - S is a (finite) set of Markov states s ∈ S
  - A is a (finite) set of actions a ∈ A
  - P is dynamics/transition model for each action, that specifies
  - P(st+1 = s'|st = s, at = a)
  - R is a reward function1
  - R(st = s, at = a) = E[rt|st = s, at = a]
- Horizon is infinite, no termination
- Policy π in this MDP induces a value function V^{π}_{old}
- Same MDP where all rewards have a constant c added and has been scaled by a constant a

My new expression for the new value function V^{π}_{old} induced by π in the second MDP in terms of V, c, a, and γ:

Start: $V^{\pi}_{old}(s) = E_{\pi}[G_{old,t} \mid x_t = s]$

Return is discounted sum of rewards: $V^{\pi}_{new}(s) = \mathbb{E}_{\pi}[G_{new,t} \mid x_t = s]$

$$V^{\pi}_{new}(s) = a * V^{\pi}_{old} + \frac{a * c}{1 - \gamma}$$

2.a

Given:

$$V_1^{\pi_1}(x_1) - V_1^{\pi_2}(x_1) = \sum_{t=1}^{H} \mathbb{E}_{x_t \sim \pi_2} \left(Q_t^{\pi_1}(x_t, \pi_1(x_t, t)) - Q_t^{\pi_1}(x_t, \pi_2(x_t, t)) \right)$$

$$V_1^{\pi_1}(x_1) \geq V_1^{\pi_2}(x_1)$$

$$S = S^+ \cup \overline{S^+}$$

$$S^+ \cap \overline{S^+} = 0$$

$$p\left(s_{t+1} = s \mid s_t = s, a_t = a^+\right) = 1$$

$$p\left(s_{t+1} \neq s \mid s_t = s, a_t = a^+\right) = 0$$

$$r(s, a^+) = 1$$

$$r(s, a) = H$$

$$a \neq a^+$$

$$\pi^+(s) = a^+ \text{ if } s \in S^+$$

$$\pi^+(s) = \pi(s) \text{ if } s \notin S^+$$

$$V_1^{\pi}(s_0) \geq V_1^{\pi^+}(s_0)$$

For any $t = 1, 2, \ldots H$

$$\begin{aligned}
\left(V_t^{\pi^+} - V_t^{\pi}\right)(s_0) &= \mathbb{E}_{(s,a)\sim\pi}\left[\left(Q_t^{\pi^+}(s, \pi^+(s)) - Q_t^{\pi^+}(s, a)\right) \mathbb{I}\{s \in S^+ \cap a \neq a+\}\right] \\
&+ \mathbb{E}_{(s,a)\sim\pi}\left[\left(Q_t^{\pi^+}(s, \pi^+(s)) - Q_t^{\pi^+}(s, a)\right) \mathbb{I}\{s \notin S^+ \cap a \neq a+\}\right] \\
&= \mathbb{E}_{(s,a)\sim\pi}\left[\left(Q_t^{\pi^+}(s, \pi^+(s)) - Q_t^{\pi^+}(s, a)\right) \mathbb{I}\{s \in S^+ \cap a \neq a+\}\right] \\
&\leq \mathbb{E}_{(s,a)\sim\pi}\left[\left(H - H - \mathbb{E}_{s'\sim p(s,a)} V_{t+1}^{\pi^+}(s')\right) \mathbb{I}\{s \in S^+ \cap a \neq a+\}\right] \leq 0
\end{aligned}$$

3.a

$$\left\|B_k V - B_k V'\right\|_\infty \leq \gamma_k \|V - V'\|_\infty$$

$$\left\|B_k V - B_k V'\right\|_\infty = \left\| max\left[R(s,a) + \gamma_k \sum_{s' \in S} p(s'|s,a)\, V_k(s')\right] - max\left[R(s,a) + \gamma_k \sum_{s' \in S} p(s'|s,a)\, V_k'(s')\right]\right\|_\infty$$

$$\leq max\left\|\gamma_k \sum_{s' \in S} p(s' \mid s,a)\left((V_k(s') - V_k'(s')\right)\right\|_\infty$$

$$\leq max\left\|\gamma_k\right\| V_k - V_k' \left\|\sum_{s' \in S} p(s'|s,a\right\|_\infty$$

$$= \gamma_k \|V - V'\|_\infty$$

3.b

$$\left\|B_1 B_2 \ldots B_k V_k - B_1 B_2 \ldots B_k V_k'\right\|_\infty \leq \gamma_1 \gamma_2 \ldots \gamma_k \left\|V_k - V_k'\right\|_\infty$$

$$\begin{aligned}\left\|B_1 (B_2 \ldots B_k V_k) - B_1 \left(B_2 \ldots B_k V_k'\right)\right\|_\infty &\leq \gamma_1 \left\|B_2 (B_3 \ldots B_k V_k) - B_2 \left(B_3 \ldots B_k V_k'\right)\right\|_\infty \\ &\leq \gamma_1 \gamma_2 \left\|B_3 (B_4 \ldots B_k V_k) - B_3 \left(B_4 \ldots B_k V_k'\right)\right\|_\infty \ \ldots \\ &\leq \gamma_1 \gamma_2 \ldots \gamma_k \left\|V_k - V_k'\right\|_\infty\end{aligned}$$

3.c

$\gamma_k \approx 1$ when k is large

$\gamma_1\gamma_2\cdots\gamma_k = \frac{1}{K+1}$

$$\left\|B_1B_2\ldots B_kV_k - B_1B_2\ldots B_kV'_k\right\|_\infty \leq \frac{1}{K+1}\left\|V_k - V'_k\right\|_\infty$$

$$\begin{aligned}\gamma_1\gamma_2\cdots\gamma_k &= \left(1-\tfrac{1}{2}\right)\left(1-\tfrac{1}{3}\right)\ldots\left(1-\tfrac{1}{k+1}\right)\\ &= \left(\tfrac{1}{2}\right)\left(\tfrac{2}{3}\right)\ldots\left(\tfrac{k}{k+1}\right)\\ &= \frac{1}{K+1}\end{aligned}$$