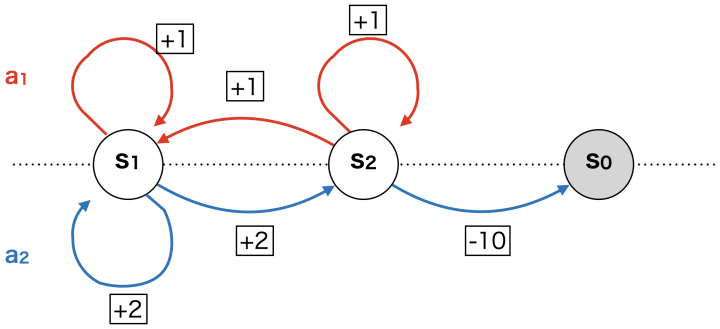
2023 年度インテリジェントシステム
レポート課題＃ 4（MDP・強化学習：解答例）

以下の問 1, 問 2 に対する解答をレポートにまとめて（文書ファイルを）eALPS から提出せよ。提出するファイルは pdf であること。文書作成には latex, MS-Office などを用いることが望ましいが、手書きのレポートをスキャンして pdf に変換後提出してもよい。

1. 下図に示す MDP に関する問 (a)∼ (d) に解答せよ。割引率は $\gamma = \frac{1}{2}$ とする。下図に示す MDP においては状態 3 種類（s_0, s_1, s_2）であり、s_0 は終端状態である。各状態（終端状態は除く）において可能な行動は a_1, a_2 の 2 種類である。図において四角い枠で囲まれた数値は報酬を示している。状態遷移確率 $P(s'|s, a)$ や報酬 $R(s, a, s')$ 詳細は図の左側に示す。

s	a	s'	P(s'|s,a)	R(s,a,s')
s_1	a_1	s_1	1.0	1
s_1	a_2	s_1	0.5	2
s_1	a_2	s_2	0.5	2
s_2	a_1	s_1	0.5	1
s_2	a_1	s_2	0.5	1
s_2	a_2	s_0	1.0	-10



(a) この MDP に関する状態価値関数 $U(s)$ を価値反復法で得ることを考える。以下に示すような初期値 $U_0(s)$ から開始し、１回 Bellman update を適用して得られる価値関数 $U_1(s)$ を求めよ。結果だけでなく計算の過程も示すこと。

	s_1	s_2	s_0
U_0	0	0	0

Bellman update は

$$U_1(s) \leftarrow \max_a \sum_{s'} P(s'|s, a)[R(s, a, s') + \gamma U_0(s')]$$

で与えられるから、これに従って計算すればよい。上の表より $U_0(s) = 0$ であるから

$$U_1(s) \leftarrow \max_a \sum_{s'} P(s'|s, a)R(s, a, s')$$

ということになる。各状態について計算してみると

$$\begin{cases} U_1(s_1) = \max\{1 \times 1, \frac{1}{2} \times 2 + \frac{1}{2} \times 2\} = \max\{1, 2\} = 2 \\ U_1(s_2) = \max\{\frac{1}{2} \times 1 + \frac{1}{2} \times 1, 1 \times (-10)\} = \max\{1, -10\} = 1 \\ U_1(s_0) = 0 \end{cases}$$

表にまとめると

	s_1	s_2	s_0
U_1	2	1	0

(b) 上の問 (b) からさらにもう 1 回 Bellman update を適用して得られる価値関数 $U_2(s)$ を求めよ。結果だけでなく計算の過程も示すこと。

Bellman update を実行すると

$$\begin{cases} U_2(s_1) = \max\{1 \times (1 + \frac{1}{2}U_1(s_1)), \frac{1}{2} \times (2 + \frac{1}{2}U_1(s_1)) + \frac{1}{2} \times (2 + \frac{1}{2}U_1(s_2))\} \\ \qquad = \max\{1 + \frac{1}{2} \times 2, \frac{1}{2} \times (2 + \frac{1}{2} \times 2) + \frac{1}{2} \times (2 + \frac{1}{2} \times 1)\} = \max\{2, \frac{11}{4}\} \\ \qquad = \frac{11}{4} \\ U_2(s_2) = \max\{\frac{1}{2} \times (1 + \frac{1}{2}U_1(s_1)) + \frac{1}{2} \times (1 + \frac{1}{2}U_1(s_2)), 1 \times (-10 + \frac{1}{2}U_1(s_0))\} \\ \qquad = \frac{1}{2} \times (1 + \frac{1}{2} \times 2) + \frac{1}{2} \times (1 + \frac{1}{2} \times 1) \\ \qquad = 1 + \frac{3}{4} = \frac{7}{4} \\ U_2(s_0) = 0 \end{cases}$$

(c) この MDP から最適方策を求めるために、方策反復を適用することを考える。
初期方策として以下のような π_0 を用いたとき、方策評価（policy evaluation）によって得られる価値関数 $U^{\pi_0}(s)$ を求めよ。

	s_1	s_2
π_0	a_1	a_2

（注：$U^{\pi_0}(s)$ に関する線形方程式が得られるがこれは容易に手で解くことができるはず。Bellman update による値更新で求める必要はない）

policy evaluation による線形方程式は

$$U^{\pi}(s) = \sum_{s'} P(s'|s, \pi(s))[R(s, \pi(s), s') + \frac{1}{2}U^{\pi}(s')]$$

となる。まず $\pi_0(s_1) = a_1$ であり、s_1 で行動 a_1 を行うと次の状態は確率 1 で s_1 だから

$$U^{\pi_0}(s_1) = 1 \times (1 + \frac{1}{2}U^{\pi_0}(s_1))$$

となる。これを解くと $U^{\pi_0}(s_1) = 2$ である。
次に $U^{\pi_0}(s_2)$ に関する方程式は $\pi_0(s_2) = a_2$ であり、s_2 で行動 a_2 を行うと次の状態は確率 1 で s_0 だから

$$U^{\pi_0}(s_2) = -10$$

である。
以上をまとめると

	s_1	s_2	s_0
$U^{\pi_0}(s)$	2	-10	0

(d) 上の問 (c) の結果から新たな方策 $\pi_1(s)$ が得られる。$\pi_1(s)$ を求めよ。結果だけでなく計算の過程も示すこと。

方策 $\pi_1(s)$ は

$$\pi_1(s) = \arg\max_a \sum_{s'} P(s'|s,a)[R(s,a,s') + \frac{1}{2}U^{\pi_0}(s')]$$

によって求めればよい。
$\pi_1(s_1)$ は、行動 a_1 を選択したとき

$$1 \times (1 + \frac{1}{2}U^{\pi_0}(s_1)) = 1 + 1 = 2$$

となる。一方行動 a_2 の場合は

$$\frac{1}{2} \times (2 + \frac{1}{2}U^{\pi_0}(s_1)) + \frac{1}{2} \times (2 + \frac{1}{2}U^{\pi_0}(s_2)) = \frac{3}{2} - \frac{3}{2} = 0$$

であるから、$\pi_1(s_1) = a_1$ となる。
$\pi_1(s_2)$ は、行動 a_1 を選択したとき

$$\frac{1}{2} \times (1 + \frac{1}{2}U^{\pi_0}(s_1)) + \frac{1}{2} \times (1 + \frac{1}{2}U^{\pi_0}(s_2)) = 1 - 2 = -1$$

となる。一方、行動 a_2 の場合は

$$1 \times (-10 + 0) = -10$$

であるから、$\pi_1(s_2) = a_1$ である。
まとめると

	s_1	s_2
π_1	a_1	a_1

2. 下図に示す MDP に関する問 (a)∼ (c) に解答せよ。割引率は $\gamma = 1$ とする。下図に示す MDP においては状態 3 種類（$s_0, s_1, s_2, \cdots, s_5$）であり、$s_0$ は終端状態である。可能な行動は a, b の 2 種類である。図において四角い枠で囲まれた数値は報酬を示している。

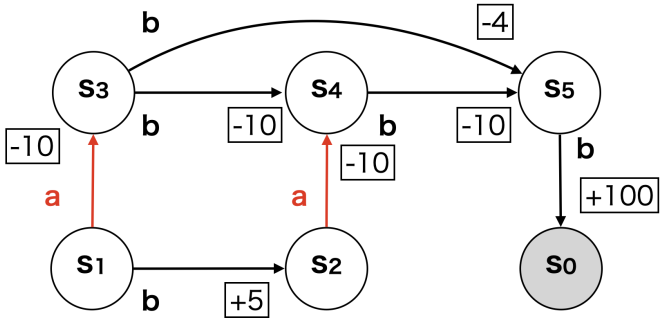
状態遷移確率 $P(s'|s,a)$ や報酬 $R(s,a,s')$ 詳細は図の左側に示す。

p, q は $p, q > 0, p + q = 1$ を満たす実数である。

（注：余計なお世話ですが... 状態 s の価値はその状態から開始して以降に最適な行動を取ったときの報酬の和の期待値なので、終端状態では、価値は 0）

s	a	s'	P(s'|s,a)	R(s,a,s')
s_1	a	s_3	1.0	-10
s_1	b	s_2	1.0	5
s_2	a	s_4	1.0	-10
s_3	b	s_4	p	-10
s_3	b	s_5	q	-4
s_4	b	s_5	1.0	-10
s_5	b	s_0	1.0	100

$p + q = 1,\ p, q > 0$



(a) 状態 s_2, s_4, s_5 の価値 $U(s_2), U(s_4), U(s_5)$ を求めよ。

Bellman 方程式を確認しておくと（$\gamma = 1$）：

$$U(s) = \max_a \sum_{s'} P(s'|s,a)[R(s,a,s') + \gamma U(s')]$$

まずは $U(s_5)$ について考える。s_5 からは行動は b 以外はなくこれにより確率 1 で終端状態 s_0 に至り（$U(s_0) = 0$ である）報酬 100 を得るから、$U(5) = 100$
s_4 については、$U(s_4) = -10 + U(s_5) = -10 + 100 = 90$、同様にして $U(s_2) = -10 + U(4) = -10 + 90 = 80$
以上、まとめると

$$U(s_2) = 80,\ U(s_4) = 90,\ U(s_5) = 100$$

(b) 状態 s_3 の価値 $U(s_3)$ を p の関数として示せ。また同じ値を q の関数として示せ。

s_3 の場合、可能な行動は b のみであるが、それにより得られる次の状態が 2 通り存在する。Bellman 方程式より

$$U(s_3) = p \cdot (-10 + U(s_4)) + q \cdot (-4 + U(5)) = 80p + 96q$$

となる。$p + q = 1$ であるから、p を用いて表すと $U(s_3) = 80p + 96(1 - p) = 96 - 16p$
同様にして q を用いて表すと $U(s_3) = 80(1 - q) + 96q = 80 + 16q$

(c) 状態 s_1 の価値 $U(s_1)$ は p の値の変化とともにどのように変動するか示せ。

Bellman 方程式より、$U(s_1)$ は行動 a をとった場合の価値と行動 b による価値のうちの大きな方となる。行動 a の場合

$$U(s_1) = -10 + U(s_3) = -10 + 96 - 16p = 86 - 16p$$

である。一方、行動 b の場合は

$$U(s_1) = 5 + U(s_2) = 5 + 80 = 85$$

従って、

$$U(s_1) = \begin{cases} 85 & p > \frac{1}{16} \\ 86 - 16p & \text{otherwise} \end{cases}$$