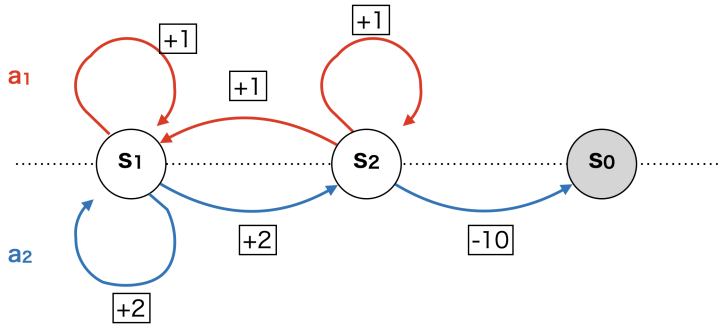
2023 年度インテリジェントシステム
レポート課題＃ 4（MDP・強化学習：提出締切 7 月 24 日）

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

(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。

(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）

(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。

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。

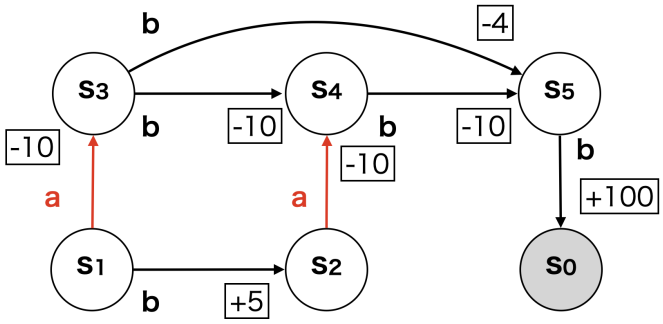
状態遷移確率 $P(s'|s, a)$ や報酬 $R(s, a, s')$ 詳細は図の左側に示す。

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。

(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。

(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。