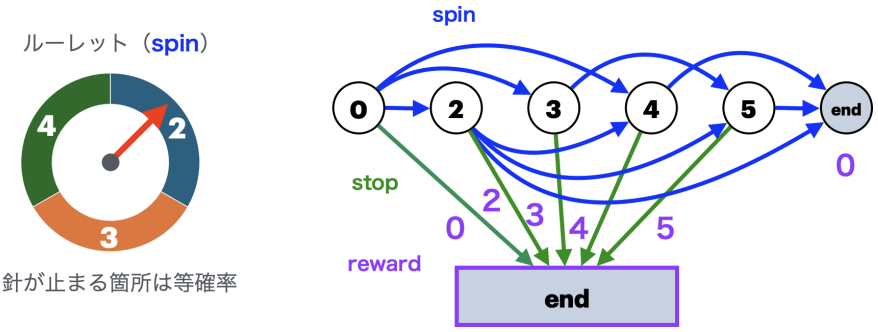
2024年度インテリジェントシステム
レポート課題＃4（MDP：提出締切7月8日）

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. 下図は持ち点0から開始し、ルーレットを回した結果針が止まった箇所の数字（2,3,4のいずれか）を得点として加算していくゲームを示している。ゲームの得点が6以上になった場合はゲームオーバー（ゲーム終了）であり、得られる報酬は0である。ゲームにおいてはゲームオーバーになるかゲーム終了を選択するまで、ルーレットを回す（spin）かゲームを終える（stop）かいずれかの行動を選択できる。

   持ち点6点未満の状態でゲーム終了を選択した場合は、その時点の持ち点が報酬として与えられる。ルーレットを回したとき得られる数値は等確率（すなわち、2,3,4の数字が得られる確率は1/3）である。

   このゲームを0,2,3,4,5,endの6状態を持ち、0,2,3,4,5の状態で取りうる行動$\{spin, stop\}$を持つMDPと考える。このとき、以下問(a)~(e)に解答せよ。なお、報酬の割引率は$\gamma = 1$とする。



   (a) このMDPに関する状態価値関数$U(s)$を価値反復法（value iteration）によって求めることを考える。初期値$U_0(s) = 0$としたとき、1回Bellman updateを適用して得られる価値関数$U_1(s)$, $(s \in \{0, 2, 3, 4, 5\})$を求めよ。

   (b) 上で求めた$U_1(s)$からさらにBellman updateを繰り返したとき$U_2(s), U_3(s), U_4(s)$, $(s \in \{0, 2, 3, 4, 5\})$はどのような値となるか示せ。

   (c) $U_4(s)$を用いて、状態$s \in \{0, 2, 3, 4, 5\}$において取るべき最適な行動$\pi(s) \in \{spin, stop\}$を求めよ。

   (d) 初期方策として以下のような方策π_0から開始し、方策反復（policy iteration）を1ステップ適用してみる。このため、まずπ_0の下で方策評価（policy evaluation）によって得られる価値関数$U^{\pi_0}(s)$を求めよ。

	$s=0$	$s=2$	$s=3$	$s=4$	$s=5$
π_0	spin	stop	spin	stop	spin

   （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）

   (e) 次いで上の問題によって得られたU^{π}から得られる方策（方策改善, policy emprovement の結果）を示せ。

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。

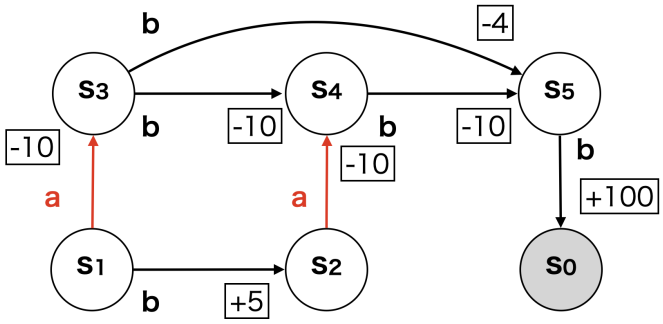
状態遷移確率 $P(s'|s, a)$ や報酬 $R(s, a, s')$ 詳細は図の左側に示す。

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。