演習：　強化学習の基礎

科学技術振興事業団　鮫島和行

1. 山賊問題

山間の道を越えて積荷を運ばなければならないが、山賊が山間の道に出没する。道はｎ本あり、山賊にであうと積荷が盗まれてしまう。山賊が出没する確率は各道によって違うが、年間を通じていつも同じ（定常）である。このときどの道を選択すべきであろうか？

この問題は、道がｎ本ある場合にはｎ本の腕を持つスロットマシーンと同じである。このスロットマシーンでは各腕に対して定常な確率で当たりが出る。山賊問題との違いはなるべく盗まれないように道を選択する問題であるのに対して、スロットマシーンでは当たり回数をできるだけ多くする問題となる。すなわち、山賊の出現しない確率がスロットマシーンの当たり確率ということになる。

最適な行動は最も当たり確率の高いレバーを引きつづけること（山賊の出没しない道を選択すること）であるが、その確率を知るためにはほかのレバーを引いて（道を通ってみて）確かめなければならない。このように、強化学習では常に自らの行動に乱雑性を持たせて探索し環境を同定しながら未来にわたる報酬を最大化するような最適な行動戦略を獲得しなければならない。

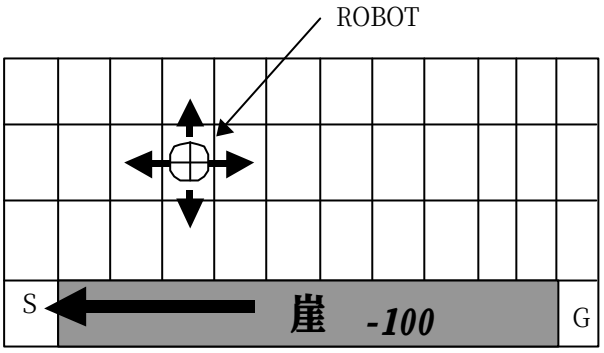
行動の乱雑性導入の仕方にはいくつかの方法があるが、行動の何回かに１回は必ずランダムな行動をし、それ以外は現在近似している評価がもっとも高くなるような行動を選択する貪欲戦略をとる手法、ε-greedy という手法がある。このランダムな行動をとる確率εを変化させた場合、１試行に得られる報酬の平均は学習とともにどのように変化するであろうか？

ε-greedy によって山賊問題を解く Matlab プログラムを bandit.m に書いてある。
これを用いて各確率ε を変化させた場合の学習曲線をプロットせよ。

また、もし1000回しか道が通れないとした場合に、もっとも多くの荷物を運ぶにはεをどのように設定するのが良いだろうか？道が10本であり、積荷は１回に定量、山賊の出る確率が各道で[0,1]の一様乱数で与えられる場合に数値的に解いてみよ、また解析的にこの問題を解く方法はあるだろうか？

2. 落とし穴のある迷路の問題

図に示すような２次元平面内を移動するロボットがある。ロボットは周囲の４方向に動作することができるが、行動には不確実性があり 10%の確率でランダムな方向に動いてしまう。また、ロボットが環境から出る方向に行動しようとする場合には、その状態にとどまる。このロボットの目的は出発地点から目的地へとたどりつくことである。環境内には「崖」が存在し、落ちてしまうと出発地点に引き戻されてペナルティとして-100 を得る。そのほかの状態では１step の動作ごとに-1 の報酬を得るものとし、ゴールにたどりついた場合には報酬０とする。



Q-learning の学習プログラムの例を Qlearning.m に書いてある。
Matlab 上では、このような迷路問題では、状態を表すために２次元のベクトル state を用いる。この場合スタート点は(4,1), ゴールは(4,12)という状態になる。各４種類の行動を action とする。このとき Q 関数は state 成分を最初の２次元にとり、それぞれの行動を最後の次元にとった３次元マトリックス Q(state(2),state(1),action)で表すことにする。このときの state の次元を逆にしたのは、行列の列番号が x 方向、行番号が y 方向の値にして直感と合わせるためである（matlab では Q(i,j)は行列 Q の i 行 j 列成分)。環境を定義する構造体 env は、壁の位置 protect、崖の位置 cliff をそれぞれ表す行列、スタート状態 start、ゴール状態 goal の情報を含んでいる。状態の遷移、行動の決定については、それぞれ別の matlab プログラムに gridmove.m、egreedy. m によってあらかじめ与えてある。

この迷路課題を off policy 強化学習である Q-learning を用いて解く場合の学習アルゴリズムをインプリメントしてみよう。Q-learning では一歩先の評価値として行動評価値の最大値をもちいている（テキスト 4.1 の式)。この行動評価値を学習するための学習則は プログラム中% Q-learning learning rule　以下の部分にかかれている。

このプログラムを走らせて、ロボットの行動評価関数がどのように学習されるのかを確認せよ。

では、実際の行動を Actor によって決定し、その行動確率に従った状態評価関数を Critic で学習する on policy 型の強化学習である Actor-Critic アルゴリズムの場合はどのようになるだろうか？ここでは、Actor-Critic の特別な場合として、行動評価関数に相当する行動決定関数 $p(s_t, a_t)$ を Actor で学習則として、単純に状態評価関数の TD 誤差 δ_t のみを用いるのではなく、一歩先の Actor の行動決定関数 $p(s_{t+1}, a_{t+1})$ を考慮に入れた学習則

$$p(s_t, a_t) \leftarrow p(s_t, a_t) + \beta[\delta_t + \gamma p(s_{t+1}, a_{t+1}) - p(s_t, a_t)]$$

によって学習するとしよう。このとき、off policy 強化学習である Q-learning と on policy 強化学習である Actor-Critic では、学習曲線、最適行動、状態もしくは行動評価関数にどのような違いが見られるだろうか？また、そのような違いはなぜ生じるのだろうか？