強化学習の基礎

1　実験の目的

強化学習の枠組について理解を深める．

まず，強化学習システム全体の構成と最適性について説明する．次に，システムを最適化するための 2 種類の代表的なアルゴリズムについて説明する．いくつかの応用問題への適用実験を行い，アルゴリズムの動作の観察を通じてそれぞれの特徴について考察を促す．

2　強化学習とは?

試行錯誤の経験を通じて，遅れのある報酬を手がかりにして未知なる環境に適応する学習制御の枠組

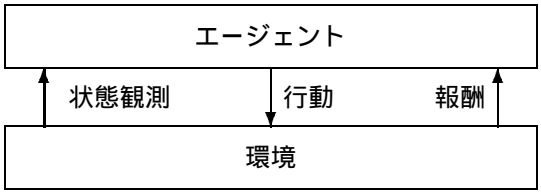

図 1: 強化学習の枠組

強化学習の枠組を図 1に示す．学習の主体『エー ジェント』と制御対象『環境』は，以下のやりとりを繰り返す．

1. エー ジェントは時刻 t において環境の状態観測 s_tに基づいて意志決定を行い，行動 a_tを出力
2. エー ジェントの行動により，環境は s_{t+1}へ状態遷移する．その状態遷移に応じた報酬 r_tをエー ジェントに与える．
3. 時刻 t を $t+1$ に進めてステップ 1 へ戻る

エー ジェントは，利得 (return: 最も単純な場合，報酬の総計) を最大化を目的として，状態観測から行動出力へのマッピング (政策 (policy) と呼ばれる) を獲得する．このとき，環境やエー ジェントには下記の条件が想定されるのが一般的である．

- エー ジェントはあらかじめ環境に関する知識を持たない
- 環境の状態遷移は確率的
- 報酬の与えられ方は確率的
- 状態遷移を繰り返した後，やっと報酬にたどり着くような，段取り的な行動を必要とする環境 (報酬の遅れ)

強化学習が注目を集める理由は，不確実性のある環境を扱っている点にある．多くの実世界の制御問題において不確実性の扱いが求められる．もう一つの理由は，報酬に遅れのある段取り的な行動の獲得を行う点にある．設計者がゴー

ル状態で報酬を与えるという形で，させたいタスクをエ- ジェントに指示しておけば，ゴ- ルへの到達方法はエ- ジェントの試行錯誤学習によって自動的に獲得される．つまり，設計者が「何をするか」をエ- ジェントに報酬という形で指示しておくだけで「どのように実現するか」を学習によってエ- ジェントが自動的に獲得する枠組となっている．

3 環境の数理モデルと最適性

強化学習で扱う環境のダイナミクスは，マルコフ決定過程 (Markov decision process: MDP) としてモデル化できることを仮定するのが一般的である．その理由は，マルコフ決定過程が不確実性を含む実世界の多くの制御問題を精度よく近似可能だからである．一般に，状態および行動の空間は，離散あるいは連続なユ- クリッド空間等さまざまな空間を想定可能である．強化学習では，解析の都合上，状態および行動の空間は離散的かつ有限な集合で，状態遷移にエルゴ- ト性を有するマルコフ決定過程の環境を仮定することが多い．

3.1 マルコフ決定過程 (MDP) とは?

- 環境のとりうる状態の集合 $\mathcal{S} = \{s_1, s_2, \cdots, s_n\}$
- エージェントが選択可能な行動の集合 $\mathcal{A} = \{a_1, a_2, \cdots, a_l\}$
- 状態遷移のダイナミクス：
  状態 s において行動 a を実行したときの状態 s'へ遷移する確率 $P^a(s, s')$ により記述．
  状態遷移は離散時間ステップで発生．
- 報酬のダイナミクス：
  状態 s において行動 a を実行した後，状態 s'へ遷移した場合の報酬の期待値 $R^a(s, s')$ により記述．

- マルコフ性：
状態 s'への遷移が，そのときの状態 s と行動 a にのみ依存し，それ以前の状態や行動には関係ないこと．

- エルゴ- ト性：
任意の状態 s からスタ- トし，無限時間経過した後の状態分布確率は最初の状態とは無関係になること．

3.2 最適性：割引報酬による評価

ある時間ステップで実行した行動が，その後の報酬獲得にどの程度貢献したのかを評価するため，その後得られる報酬の時系列を考える．報酬の時系列評価は利得 (return) と呼ばれ，最も単純な場合，時系列中の報酬の総計で表される．強化学習では，割引報酬合計による評価を利得として用いる場合が多い．これは，時間の経過とともに報酬を割引率γ $(0 \leq \gamma < 1)$ で割引いて合計する．ある時刻 t で実行した行動の利得 R_tを以下で定義する．

$$R_t = r_t + \gamma r_{t+1} + \gamma^2 r_{t+2} + \cdots = \sum_{k=0}^{\infty} \gamma^k r_{t+k} \tag{1}$$

ただし r_tは時刻 t でエ- ジェントが受け取る報酬である．未来の報酬を割引く理由は以下による．

1. 実環境では，時間の経過とともに環境が変化する可能性があるため，時系列上の全ての報酬を同じ重みで考慮するのは妥当ではない
2. 無限時間ステップの時系列の利得を有限の値として扱えるため．

- 政策πは状態集合から行動集合への写像関数 (確率分布関数でも良い)
- 定常政策とは，時間とともに変化することのない政策である．

エルゴード性マルコフ決定過程においてエージェントが定常政策πをとるとき，利得の期待値は，時間には関係なく状態sだけに依存する性質を持つ．このとき定常政策πにおいて状態sに依存する利得の期待値を状態sのValueと呼び，sの関数として表されるValueをState-Value関数$V^\pi(s)$と表す．
この$V^\pi(s)$は以下のBellman方程式の解である．

$$\begin{aligned}
V^\pi(s) &= E_\pi\{R_t|s_t=s\} \\
&= E_\pi\left\{\sum_{k=0}^{\infty}\gamma^k r_{t+k}\middle| s_t=s\right\} \\
&= E_\pi\left\{r_t+\gamma\sum_{k=0}^{\infty}\gamma^k r_{t+k+1}\middle| s_t=s\right\} \\
&= \sum_a \pi(s,a)\sum_{s'}P^a(ss')\left[R^a(s,s')+\gamma E_\pi\left\{\sum_{k=0}^{\infty}\gamma^k r_{t+k+1}\middle| s_{t+1}=s'\right\}\right] \\
&= \sum_a \pi(s,a)\sum_{s'}P^a(s,s')\left[R^a(s,s')+\gamma V^\pi(s')\right]
\end{aligned}$$

ただし，$E_\pi\{\cdot\}$はエージェントが政策πに基づいて行動選択するときの期待値，$\pi(s,a)$は政策πのもとで状態sにおいて行動aを選択する確率である．

- 最適なState-Value関数：全ての状態sにおいて$V^\pi(s)\geq V^{\pi'}(s)$となるとき，政策πはπ'より優れているという．上記のマルコフ決定過程では，他のどんな政策よりも優れた，あるいは同等な政策が少なくとも1つ存在する．これを最適政策π^*という．最適政策は複数存在することもあるが，全ての最適政策は同じState-Value関数を共有する．これは最適なState-Value関数V^*と呼ばれ，以下のように定義される．

$$V^*(s)=\max_\pi V^\pi(s) \qquad \text{for all } s\in S.$$

- 最適なAction-Value関数：最適な政策はまた以下に示す同一のAction-Value関数を共有する．

$$Q^*(s,a)=\max_\pi Q^\pi(s,a) \qquad \text{for all } s\in S \text{ and } a\in A.$$

$Q^*(s,a)$はQ値と呼ばれ，状態sで行動aを選択後，ずっと最適政策をとりつづけるときの利得の期待値を表す．

- 最適なValue関数のBellman方程式：最適なValue関数は以下の方程式の解である．

$$\begin{aligned}
V^*(s) &= \max_{a\in A} Q^{\pi^*}(s,a) \\
&= \max_a E_{\pi^*}\{R_t|s_t=s,a_t=a\} \\
&= \max_a E_{\pi^*}\left\{\sum_{k=0}^{\infty}\gamma^k r_{t+k}\middle| s_t=s,a_t=a\right\} \\
&= \max_a E_{\pi^*}\left\{r_t+\gamma\sum_{k=0}^{\infty}\gamma^k r_{t+k+1}\middle| s_t=s,a_t=a\right\} \\
&= \max_a E\left\{R_t+\gamma V^*(s_{t+1})\middle| s_t=s,a_t=a\right\} \\
&= \max_a \sum_{s'}P^a(s,s')\left[R^a(s,s')+\gamma V^*(s')\right] \quad , \\
Q^*(s,a) &= E\left\{r_{t+1}+\gamma\max_{a'}Q^*(s_{t+1},a')\middle| s_t=s,a_t=a\right\} \\
&= \sum_{s'}P^a(ss')\left[R^a(s,s')+\gamma\max_{a'}Q^*(s',a')\right]
\end{aligned}$$

$Q^*(s,a)$が得られたとき，状態sにおいて最大のQ値を持つ行動aが最適な行動である．

4　マルコフ決定過程下での強化学習

前章で示したマルコフ決定過程下での強化学習問題は，以下のように定式化できる．

- エージェントは予め環境の状態遷移確率 $P^a(s,s')$ や報酬の与えられ方 $R^a(s,s')$ についての知識を持たない
- エージェントは環境との試行錯誤的な相互作用を繰り返して，最適な政策を学習する．

4.1　代表的な強化学習アルゴリズム：Q-learning

前章で示したように，$Q^*(s,a)$ が得られれば，最適な政策は簡単に得られる．Q-learning[Watkins et.al 92] は環境との試行錯誤的な相互作用の繰り返しを通じて $Q^*(s,a)$ を推定するアルゴリズムである．以下にその概要を示す．

1. エージェントは環境の状態観測 s_t を受けとる．
2. エージェントは任意の行動選択方法 (探査戦略) に従って行動 a_t を実行する．
3. 環境から報酬 r_t を受け取る．
4. 状態遷移後の状態観測 s_{t+1} を受け取る．
5. 以下の更新式により Q 値を更新

$$Q(s_t,a_t) \leftarrow Q(s_t,a_t) + \alpha \left[r_t + \gamma \max_a Q(s_{t+1},a) - Q(s_t,a_t) \right]$$

   ただしαは学習率，γは割引率 $(0 \leq \gamma < 1)$ である．
6. 時間ステップ t を $t+1$ へ進めて手順 1 へ戻る．

図 2: Watkins の One-step Q-learning アルゴリズムの概要

- Watkins の収束定理：

エージェントの行動選択において，全ての行動を十分な回数選択し，かつ学習率αが$\sum_{t=0}^{\infty} \alpha(t) \rightarrow \infty$ かつ $\sum_{t=0}^{\infty} \alpha(t)^2 < \infty$ を満たす時間 t の関数となっているとき，Q-learning のアルゴリズムで得る Q 値は確率 1 で最適な行動 Value に収束する (概収束)．ただし環境はエルゴート性を有する離散有限マルコフ決定過程であることを仮定する．

- 行動選択方法 (探査戦略)：

Watkins の収束定理は，全ての行動を十分な回数選択しさえすれば行動選択方法 (探査戦略) には依存せずに成り立つ．よって行動選択はランダムでもよい．しかし，強化学習ではまだ Q 値が収束していない学習の途中においてもなるべく多くの報酬を得るような行動選択を求められることが多い．学習に応じて序々に挙動を改善していくような行動選択方法として，

1) ϵ-greedy 選択： ϵの確率でランダム，それ以外は最大の Q 値を持つ行動を選択

2) ボルツマン選択：$exp(Q(s,a)/T)$ に比例した割合で行動選択，ただし T は時間とともにゼロに近付く

などの方法が提案されている．

- Q-learning はその更新式の形より，確率近似法による非同期ダイナミックプログラミング (動的計画法) と呼ばれる．

4.2 政策を逐次改善する強化学習アルゴリズム：確率的傾斜法

Q-learning はある条件下では最適性が保証される強力なアルゴリズムだが，状態遷移のマルコフ性の仮定が少しでも満たされなかったり状態空間や行動空間を連続値に拡張すると致命的な性能低下を生じる．ここで紹介する確率的傾斜法は，完全な離散マルコフ決定過程の環境では Q-learning の性能に及ばないが，比較的幅広いクラスの環境へ適用可能であるという特徴を持つ．

4.2.1 パラメ− タ表現された確率的政策

エージェントの学習目標は，報酬獲得の評価関数を最大化するように観測入力に対する行動出力を獲得することである．確率的傾斜法では図 3に示すように観測入力に対する行動出力の確率分布関数として定義される確率的政策を扱う．マルコフ決定過程の環境では，エージェントの状態観測 Xは環境の状態 s と等しいが，ここでは分けて記述する．ある政策πのもとで観測 Xにおいて行動 a を選択する確率 $Pr(a|\pi, X)$ は，エージェントの内部変数ベクトル W を用いて式 (2) の確率分布関数で表す．

$$Pr(a|\pi, X) \;=\; \pi(a, W, X) \tag{2}$$

エージェントは内部変数 Wを調節することにより政策πを変えることができる．$\pi(a, W, X)$ の関数形については，エージェントに実装できる計算資源の制限など，一般に個別の問題ごとに制約が存在する．すなわちエージェントの構造および制約条件は$\pi(a, W, X)$ の関数形で規定される．例えば行動選択確率を表す機構がニューラルネットで構成されているならば，内部変数 Wはリンクの重み変数に相当し，中間層の個数や演算素子の特性等はπ関数形として完全に記述される．Wをパラメータとした単純な関数として政策πを記述することの利点は，上記のように様々な構造を持つエージェント全てに対して共通の理論的基礎を与えられることである．

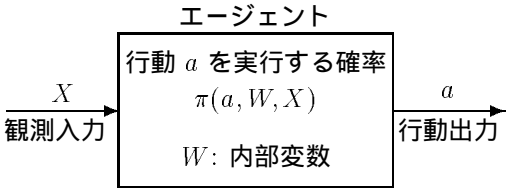

図 3: 確率的政策：政策パラメータベクトル Wを調節して政策を変える．

$\forall X \in \mathcal{X}$, $\forall a \in \mathcal{A}$ において政策$\pi(a, W, X)$ はベクトル Wの全ての要素で偏微分可能であると仮定する．Wの任意の i 番目の要素を w_iと表す．ある時間ステップ t においてエージェントが行動 a_tを実行するとき，エ− ジェントは行動選択確率πの計算過程において式 (3) で定義される $e_i(t)$ を容易に計算できる．

$$\begin{aligned} e_i(t) &= \frac{\partial}{\partial w_i} \ln\Big(\pi(a_t, W, X_t)\Big) \\ &= \frac{1}{\pi(a_t, W, X_t)} \frac{\partial}{\partial w_i} \pi(a_t, W, X_t) \end{aligned} \tag{3}$$

式 (3) の $e_i(t)$ は適正度 (characteristic eligibility) [Williams 92] と呼ばれ，(観測入力-行動出力) という確率事象の情報論的な価値における変数 w_iの相関であると考えられる．このとき以下の定理が成り立つ [Williams 92]．

定理 1 環境からの報酬 r_tの期待値がエージェントの内部変数 Wの関数となる場合，以下が成り立つ．

$$E\big\{(r_t - b)\, e_i(t)\big\} \;=\; \frac{\partial}{\partial w_i} E_\pi \{r_t\} \tag{4}$$

ただし b は報酬基底 *(reinforcement baseline)* と呼ばれる定数，r_tは時間ステップ t においてエージェントが受けとる報酬である．

定理 1は，強化学習エージェントの政策改善方法に関して重要な手がかりを与える．エージェントは毎ステップにおいて容易に計算できる式 (3) と報酬 r_tだけから，報酬の期待値を最も大きくする方向へ政策πを確率的に改善可能である．

4.2.2 確率的傾斜法の一般形

1. 環境の観測 X_tを受けとる．
2. $\pi(a_t, W, X_t)$ の確率で行動 a_tを実行する．
3. 環境から報酬 r_tを受け取る．
4. 内部変数 W の全ての要素 w_iについて以下の $e_i(t)$ と $D_i(t)$ を求める．ただしγは割引率 $(0 \leq \gamma < 1)$ である．

$$\begin{aligned} e_i(t) &= \frac{\partial}{\partial w_i} \ln\Big(\pi(a_t, W, X_t)\Big) \,, \\ D_i(t) &= e_i(t) + \gamma D_i(t-1) \,, \end{aligned}$$

5. 以下の式を用いて$\Delta w_i(t)$ を求める．

$$\Delta w_i(t) = (r_t - b)\, D_i(t) \,,$$

ただし b は報酬基底と呼ばれる定数である．

6. 政策の改善: 以下の式で Wを更新

$$\begin{aligned} \Delta W(t) &= (\Delta w_1(t), \Delta w_2(t) \cdots \Delta w_i(t) \cdots) \,, \\ W &\leftarrow W + \alpha(1-\gamma)\, \Delta W(t) \,, \end{aligned}$$

ただしαは非負の学習定数である．

7. 時間ステップ t を $t+1$ へ進めて (1) へ戻る．

図 4: 確率的傾斜法による強化学習法の一般形

確率的傾斜法による強化学習アルゴリズムの一般形を図 4に示す [Kimura 95], [木村 96]．アルゴリズム中の手順 4 の $e_i(t)$ は式 (3) の適正度である．$D_i(t)$ は適正度の履歴 (eligibility trace) と呼ばれ，$D_i(t-1)$ を割引率γで割引きつつ適正度 $e_i(t)$ を足し合わせることにより，過去の行動ほど減衰 / 忘却しながら今までとった行動の履歴を記憶している．エージェントは報酬を受けとると，手順 5,6 にて履歴に示された行動の確率を高めるように Wを更新する．このような処理を多数の時間ステップにわたり繰り返すと，報酬獲得に関係ない行動は打ち消され，報酬獲得に関係する行動だけが強化される．行動の履歴を強化するため，報酬の獲得に遅れのある行動も強化される．本手法は報酬を受けとった時点で今までの経験を強化するため，経験強化型の学習アルゴリズムに分類できる．本アルゴリズムは以下の性質を持つため，定理 1 のアルゴリズムを時系列の利得へ拡張したものと考えられる．

$$\begin{aligned} \sum_{t=0}^{\infty} \Delta w_i(t) &= \sum_{t=0}^{\infty} \left[(r_t + \gamma r_{t+1} + \gamma^2 r_{t+2} + \cdots)\, e_i(t) \;-\; (b + \gamma b + \gamma^2 b + \cdots)\, e_i(t) \right] \\ &= \sum_{t=0}^{\infty} \left(R_t - \frac{b}{1-\gamma} \right) e_i(t) \end{aligned}$$

5　具体例 1：タクシーマネージメントの強化学習

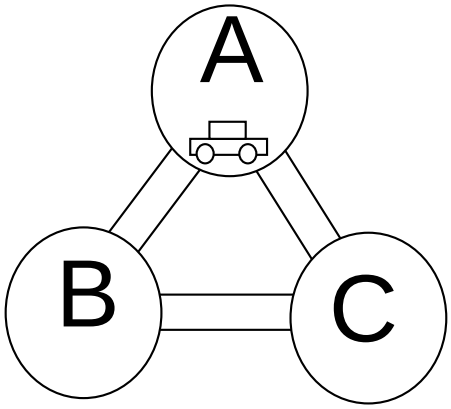

図 5: タクシーマネージメントの問題

A, B, C の 3 つの町を運転して回るタクシーの運転手をエージェントと考える．
それぞれの町において，以下の (Act0), (Act1), (Act2) の 3 つの行動のうち一つを選択して実行できる．
　(Act0) 町中を流して客を拾う．
　(Act1) タクシースタンドへ行って客を拾う．
　(Act2) 無線連絡による呼出しで客を拾う．
行動を実行して客を拾うと，町を移動するという状態遷移が発生し，報酬が得られる．

　この問題では，町 A, B, C が環境の状態に相当する．(Act0), (Act1), (Act2) が行動である．状態遷移や報酬は，状態と行動の関数となっているある確率分布に従うが，この確率分布は最初はエージェントにとって未知である．このとき，短期的にみて利得を最大とするためには，各都市においてどの行動を選択すべきか? また，長期的にみて利得を最大とするためには，どのような行動をとるべきか?

【デモプログラム】
ブラウザ上で `TaxiDemo.html` を開くと，自分がエージェントとして行動選択 / 状態遷移するデモプログラムが走る．
または，コマンドライン上で
```
appletviewer TaxiDemo.html
```
を実行しても同様である．この場合はコンソールへ状態遷移や報酬の履歴が出力される．

5.1　Q-learning を用いた学習

状態も行動も離散なので，(3 状態) × (3 行動) の Q 値のテーブルを用意し，図 2のアルゴリズムをそのまま適用

【デモプログラム】 コマンドライン上で
```
appletviewer TaxiQLDemo.html
```
を実行すると，本例題を Q-learning で学習する．状態遷移や報酬の履歴はコンソールへ出力される．
Q-learning アルゴリズムのパラメータ (学習率や割引率等) は `qagent1.cnf` の中に記述する．

5.2 確率的傾斜法を用いた学習

確率的傾斜法を用いる場合，政策の表現には任意性があるため，まずどのような政策表現を用いるかを決めなければならない．重み付きルールベースシステム，あるいは Classifier system [Goldberg 89][宮崎 94] と呼ばれる強化学習システムは計算機へのプログラムが容易なためによく用いられる．重み付きルールベースシステムを用いた確率的傾斜法を以下に示す．

ルール 1 個は条件部 (if...) と行動部 (then...) および重み W と適正度の履歴 Dから構成される．つまりルール 1 個あたり 2 個のパラメータ (Wと D) を持つ．学習前，全ルールの Wは正の値，Dは 0 に初期化しておき，6に示す手順で学習を行う．図 6の手順 (6) で重みを更新する際，Wが負の値になると政策πの計算がおかしくなってしまうため，Wが負の値にならないように配慮する必要がある．

1. 観測入力 X_tを受取り，条件部に適合するルールを全て選び出す．ここでは説明のため 3 個のルールが適合したものとし，ルール 1,2,3 のそれぞれの行動を a_1, a_2, a_3と表し，それぞれに対応する重みを w_1, w_2, w_3，適正度を e_1, e_2, e_3，適正度の履歴を D_1, D_2, D_3とする．
2. 条件部に適合したルールの重みの大きさに比例した確率で行動を選択し，実行する．ここでは説明のためルール 1 を実行したと仮定すると，実行した行動に関する政策πは以下の式で表される．
$$\pi(a_t = a_1, W, X_t) = \frac{w_1}{w_1 + w_2 + w_3}$$
3. 環境から報酬 r_tを受け取る．
4. まず条件部に適合したルールについて適正度と適正度の履歴を計算する．行動 a_1を実行したから，手順 (2) で示したπおよび式 (3) より
$$\begin{aligned} e_1 &= \frac{\partial}{\partial w_1} \ln \frac{w_1}{w_1 + w_2 + w_3} = \frac{(w_1 + w_2 + w_3) - w_1}{w_1(w_1 + w_2 + w_3)} \\ e_2, e_3 &= \frac{\partial}{\partial w_i} \ln \frac{w_1}{w_1 + w_2 + w_3} = \frac{-1}{w_1 + w_2 + w_3} \quad , \text{where } w_i \neq w_1 \end{aligned}$$
これらを用いて条件部に適合したルール 1,2,3 の適正度の履歴を以下のように計算する．
$$\begin{aligned} D_1 &\leftarrow e_1 + \gamma D_1 \\ D_2 &\leftarrow e_2 + \gamma D_2 \\ D_3 &\leftarrow e_3 + \gamma D_3 \end{aligned}$$
ただしγは割引率，$\leftarrow$は代入演算子を表す．
5. 条件部に適合しなかったルールについて適正度と適正度の履歴を計算する．条件部に適合しなかったルールは，政策πを重みで微分すると 0 だから，適正度も 0 である．よって条件部に適合しなかったルール全てについて適正度の履歴を以下のように更新する．
$$D \leftarrow \gamma D$$
6. 全てのルールについて以下のように重みを更新し，手順 1 へもどる．
$$w \leftarrow w + \alpha (r_t - b) D$$
ただしαと b はそれぞれ学習係数，報酬基底である．

図 6: 確率的傾斜法を用いた重み付きルールベースシステム

【デモプログラム】 コマンドライン上で

```
appletviewer TaxiSGADemo.html
```

を実行すると，本例題を確率的傾斜法で学習する．状態遷移や報酬の履歴はコンソールへ出力される．
確率的傾斜法のパラメータ (学習率や割引率等) は `sgaagent1.cnf` の中に記述する．

6　具体例 2：サーバアクセスコントロールの強化学習

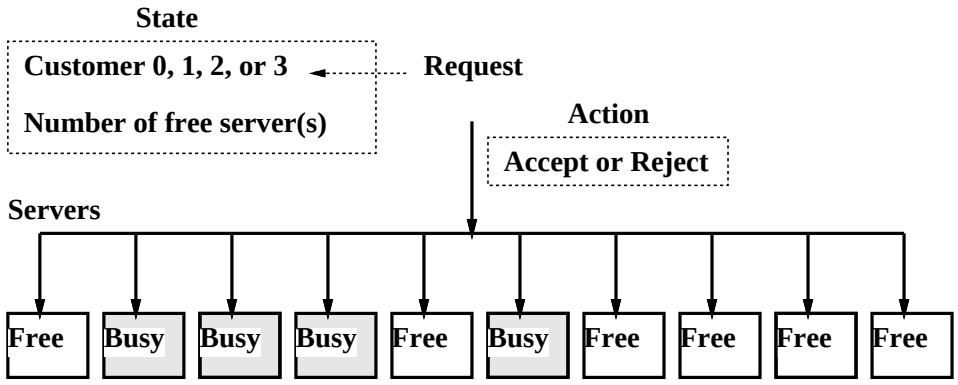

図 7: サーバアクセスコントロールの問題

図に示すような 10 台のサーバを持つシステムへ，0 から 3 までの優先度を持つ顧客がアクセス要求を行う．エージェントはアクセス要求を受け入れる (accept) するか，拒絶する (reject) かのどちらかの行動を選択することにより，サーバシステムをコントロールしている．サーバに空きがあるときに顧客のアクセス要求を受け入れると，顧客の優先度に応じてエージェントに報酬が与えられ，顧客のアクセス要求を受け入れたサーバは Busy 状態になる．エージェントは離散時間ステップでサーバシステムのコントロールを行う．Busy 状態になったサーバはその時間ステップ毎にある確率 p で空き状態へ遷移する．エージェントは，利得を最大化するために，サーバの空き状況と到着した顧客の優先度に応じ，適切にアクセスをコントロールする必要がある．

この問題では，アクセス要求を行っている顧客の優先度とサーバの空き台数が環境の状態に相当する．(accept), (reject) が行動である．状態遷移や報酬は，状態と行動の関数となっているある確率分布に従うが，この確率分布は最初はエージェントにとって未知である．このとき，短期的にみて利得を最大とするためには，それぞれの状態においてどの行動を選択すべきか? また，長期的にみて利得を最大とするためには，どのような行動をとるべきか?

【デモプログラム】
ブラウザ上で `AccesCtrlDemo.html` を開くと，自分がエージェントとして行動選択 / 状態遷移するデモプログラムが走る．または，コマンドライン上で

```
appletviewer AccesCtrlDemo.html
```

を実行しても同様である．この場合はコンソールへ状態遷移や報酬の履歴が出力される．

6.1　Q-learning を用いた学習

状態も行動も離散なので，(4 × 10 状態) × (2 行動) の Q 値のテーブルを用意し，図 2のアルゴリズムをそのまま適用

【デモプログラム】 コマンドライン上で

```
appletviewer AccesCtrlQLDemo.html
```

を実行すると，本例題を Q-learning で学習する．状態遷移や報酬の履歴はコンソールへ出力される．
Q-learning アルゴリズムのパラメータ (学習率や割引率等) は `qagent0.cnf` の中に記述する．

6.2 確率的傾斜法を用いた学習

ここでは単層ニューラルネットと Bernoulli (ボロノイ) logistic unit を用いた確率的傾斜法を示す．

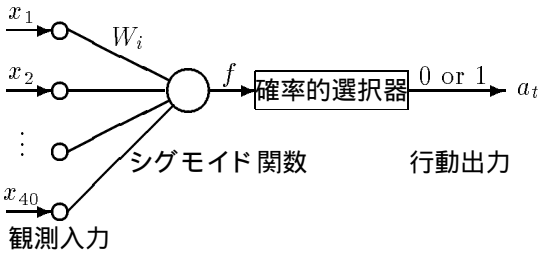

図 8: サーバアクセスコントロール問題で使用した確率的傾斜法の政策表現

(4×10) の状態があるので，図 8に示すように 40 コの入力ユニット $(x_1, x_2, \cdots, x_{40})$ を用意し，対応する状態の入力ユニットのみ 1，それ以外の入力ユニットは全てゼロとなるようにする．任意の i 番めの入力ユニットとシグモイド関数を結ぶ内部変数を w_i と表す．エージェントは入力ユニット x_i と内部変数 w_i の加重和をとり，f を以下のように計算する．

$$f = \frac{1}{1 + \exp\left(-\sum_{i=1}^{40} x_i\, w_i\right)} \ . \tag{5}$$

この f は明らかに $0 < f < 1$ なので，f の値を行動 a_t が 1 をとる確率とする．エージェントが時間 t で行動として任意の a_t を出力したとき，適正度は以下のように計算できる．

$$e_i(t) = x_i(a_t - f) \ . \tag{6}$$

これを 4の一般形アルゴリズムにあてはめると，図 9のようになる．

1. 環境の観測 X_t を受けとる．
2. 式 5により f を計算し，f の確率で行動 $a_t = 1$ を選択，それ以外の場合 $a_t = 0$ を選択
3. 環境から報酬 r_t を受け取る．
4. 内部変数 W の全ての要素 w_i について以下の $e_i(t)$ と $D_i(t)$ を求める．ただしγは割引率 $(0 \le \gamma < 1)$ である．

$$\begin{aligned} e_i(t) &= x_i(a_t - f) \ , \\ D_i(t) &= e_i(t) + \gamma D_i(t-1) \ , \end{aligned}$$

5. 次式を用いて$\Delta w_i(t)$ を求める：　$\Delta w_i(t) \ = \ (r_t - b)\, D_i(t)$
6. 政策の改善: 次式で全ての w_i を更新　$w_i \ \leftarrow \ w_i + \alpha\, \Delta w_i(t)$
7. 時間ステップ t を $t+1$ へ進めて 1 へ戻る．

図 9: サーバアクセスコントロールのエージェントに実装した確率的傾斜法

【デモプログラム】 コマンドライン上で

```
appletviewer AccesCtrlSGADemo.html
```

を実行すると，本例題を確率的傾斜法で学習する．状態遷移や報酬の履歴はコンソールへ出力される．
確率的傾斜法のパラメータ (学習率や割引率等) は `sgaagent0.cnf` の中に記述する．

7　具体例 3：倒立振子の強化学習– 状態・行動空間が連続な場合

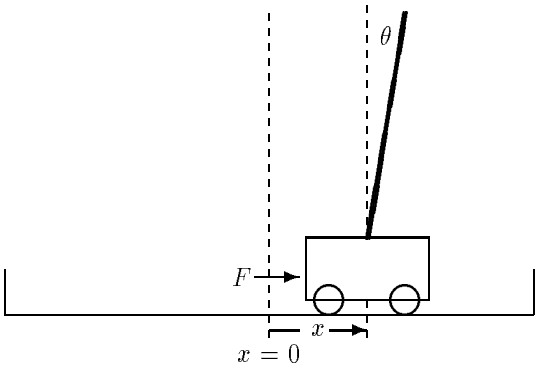

図 10: 倒立振子の強化学習問題

図 10に示す倒立振子を倒れないように学習させることを考える．エージェントは$\Delta t = 0.02$sec の離散時間間隔おきにシステムの状態 $(x, \dot{x}, \theta, \dot{\theta})$を観測し，行動として台車に加える力 Fを出力する．エージェントは行動出力として任意の連続値をとることができるが，$F = \pm 20$(N) の範囲を超える場合には，環境はこの制限を超える分を無視して行動を実行する．環境は $(x, \dot{x}, \theta, \dot{\theta}) = (0, 0, 0, 0)$ の初期状態から始まる．ポールの角度θが± 12 度を超えるか，または台車の中心 x が± 2.4 の範囲からはみ出すと，環境からエージェントへ-1 の報酬が与えられ，環境は初期状態へ戻る．実際の倒立振子のダイナミクスは微分方程式で記述されるが，計算機シミュレ– タ上では$\Delta t = 0.02$sec の離散時間システムとして近似する．シミュレ– タは [Barto et al. 83] の実験設定を参考にしている．

7.1　Q-learning を用いた学習

Q-learning を適用するためには，連続な状態・行動空間を離散化しなければならない．最も一般的なのは，連続な空間を等間隔な多次元のグリッドに分割し，各ボックスを一つの状態として量子化する方法である．このとき，グリッドが粗すぎると，離散マルコフ決定過程として見たときに状態遷移のマルコフ性が満たされなくなる (隠れ状態・不完全知覚) という問題が生じる．そのため十分に細かいグリッドを用いなければならない．

7.2　確率的傾斜法を用いた学習：連続値の状態入力 / 行動出力への拡張

図 4に示した確率的傾斜法の一般形アルゴリズムは，政策πを確率密度関数とすることにより連続値の行動へ拡張可能である．正規分布は連続値の確率変数のためのマルチパラメータの分布として単純なためよく知られている．正規分布は平均値μと標準偏差σの 2 つのパラメータを持つ．政策πを式 (7) で表される正規分布で与えるとき平均値μと標準偏差σをパラメータと考えた場合の適正度は以下のように計算できる．

$$\pi(a, \mu, \sigma) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(\frac{-(a-\mu)^2}{2\sigma^2}\right) \tag{7}$$

$$e_\mu = \frac{\partial}{\partial \mu} \ln\Big(\pi(a, \mu, \sigma)\Big) = \frac{a_t - \mu}{\sigma^2} \tag{8}$$

$$e_\sigma = \frac{\partial}{\partial \sigma} \ln\Big(\pi(a, \mu, \sigma)\Big) = \frac{(a_t - \mu)^2 - \sigma^2}{\sigma^3} \tag{9}$$

μとσが内部変数 w_iの関数として決定される場合，適正度は合成関数の微分公式より以下の式で与えられる．

$$e_i = \frac{\partial}{\partial w_i} e_\mu + \frac{\partial}{\partial w_i} e_\sigma \tag{10}$$

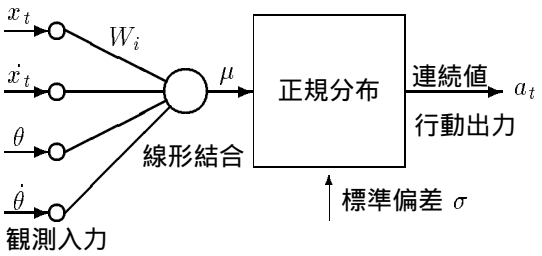

図 11: 倒立振子問題で使用した確率的傾斜法の政策表現

ここで示す例では，エージェントは 5 つの内部変数 $w_1 \cdots w_5$を持ち，これを用いてμとσを以下のように計算する．

$$\begin{aligned} \mu &= w_1\, x_t + w_2\, \dot{x_t} + w_3\, \theta_t + w_4\, \dot{\theta_t} \quad , \\ \sigma &= 0.1 + \frac{1}{1+\exp(-w_5)}. \end{aligned} \tag{11}$$

適正度 $e_1, \cdots, e_5$は以下に与えられる．

$$\begin{aligned} e_1 &= \frac{a_t-\mu}{\sigma^2}\, x_t \ , \quad e_2 = \frac{a_t-\mu}{\sigma^2}\, \dot{x_t} \ , \quad e_3 = \frac{a_t-\mu}{\sigma^2}\, \theta_t \ , \quad e_4 = \frac{a_t-\mu}{\sigma^2}\, \dot{\theta_t} \\ e_5 &= \frac{(a_t-\mu)^2-\sigma^2}{\sigma^3}\sigma(1+0.1-\sigma) \ . \end{aligned}$$

これらを用いて図 4の手順で政策を学習する．

式 8,9においてパラメータσが分母となっていることより，σが 0 へ近付くと適正度が発散することに注意しなければならない．この問題に対処するための一つの方法として，σの値に応じて内部変数の更新のステップ幅を制御することが考えられる．例えば，更新のステップ幅をσ^2に比例させるなど．

8　具体例 4：マルコフ性を満たさない環境下での強化学習– 隠れ状態問題

マルコフ決定過程の環境では，エー ジェントの状態観測 x_tは完全，つまり環境の状態を s_tとすると $x_t = s_t$であることを仮定している．しかし，実問題では状態観測が不完全であることが多い．

部分観測マルコフ決定過程 (POMDP) は，マルコフ決定過程を拡張し，エー ジェントの状態観測に不完全性や不確実性を付加した数理モデルである．POMDP 環境下のエージェントは，環境の MDP の状態 s を直接知ることはできないが，状態 s についての情報を含んだ “観測” を受けとる．観測の集合を $\mathcal{X} = \{X_1, X_2, \cdots, X_m\}$ と表す．MDP における状態が s のとき，エージェントは $P(X|s)$ の確率で観測 Xを受けとる．$P(X|s)$ は現在の状態 s だけに依存する．ある時間ステップ t においてエージェントは環境中の状態 s_tにあるとする．このときエージェントは $P(X_t|s_t)$ の確率で環境から観測 X_tを受けとり，行動決定を行う．強化学習エー ジェントは，状態遷移規則や報酬の与えられ方に加えて観測の与えられ方すなわち $P(X_t|s_t)$ に関する知識をあらかじめ持っていないものとする．

エー ジェントの状態観測に不完全性や不確実性が存在することは「 隠れ状態 」と呼ばれ，困難な非マルコフ性の問題として知られる．

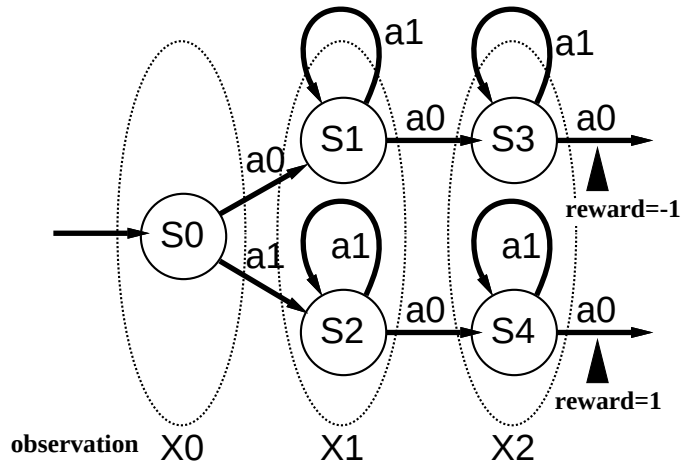

図 12: 隠れ状態問題の例．$S0, \cdots, S4$ は状態，$X0, X1, X2$ はエー ジェントの状態観測，$a0, a1$ および矢印は行動とそのときの状態遷移をあらわす．

図 12に示すような状態遷移をする環境を考える．環境の状態が $S0$ のとき，エー ジェントの状態観測は $X0$ である．環境の状態が $S1$ または $S2$ のとき，エー ジェントの状態観測は $X1$ である．このように状態観測に不完全性があるため，エー ジェントは環境の状態 $S1$ と $S2$ を区別できない同様に環境の状態 $S3$ と $S4$ を区別できず，$X2$ を観測する．より多くの報酬を得るためには，観測 $X0$–行動 $a1$，観測 $X1$–行動 $a1$，観測 $X2$–行動 $a1$ という政策を得る必要がある．

【デモプログラム】
ブラウザ上で `HiddenStateDemo.html` を開くと，自分がエージェントとして行動選択 / 状態遷移するデモプログラムが走る．または，コマンドライン上で

```
appletviewer HiddenStateDemo.html
```

を実行しても同様である．この場合はコンソールへ状態遷移や報酬の履歴が出力される．

8.1 Q-learning を用いた学習

状態観測も行動も離散なので，(3 状態) × (2 行動) の Q 値のテーブルを用意し，図 2のアルゴリズムをそのまま適用

【デモプログラム】 コマンドライン上で

```
appletviewer HiddenStateQLDemo.html
```

を実行すると，本例題を Q-learning で学習する．状態遷移や報酬の履歴はコンソールへ出力される．
Q-learning アルゴリズムのパラメータ (学習率や割引率等) は `qagent3.cnf` の中に記述する．

8.2 確率的傾斜法を用いた学習

ここでは状態観測が離散で行動も離散な 2 値なので，サ– バの例題で用いたのと同様の政策表現を用いる．

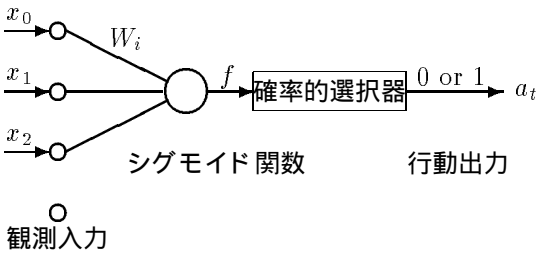

図 13: 隠れ状態の例題で使用した確率的傾斜法の政策表現

3 つの状態観測があるので，図 13に示すように 3 コの入力ユニット (x_0, x_1, x_2) を用意し，対応する状態観測の入力ユニットのみ 1，それ以外の入力ユニットは全てゼロとなるようにする．後は図 9に示した手順で学習する．

【デモプログラム】 コマンドライン上で

```
appletviewer HiddenStateSGADemo.html
```

を実行すると，本例題を確率的傾斜法で学習する．状態遷移や報酬の履歴はコンソールへ出力される．
確率的傾斜法のパラメータ (学習率や割引率等) は `sgaagent3.cnf` の中に記述する．

課題

1. タクシー の例題において，(A) 最も短期的な利益の最大化を目指す場合，および(B) 長期的な利益の最大化を目指す場合，各都市においてどの行動を選択すべきか? Q-learning および確率的傾斜法の割引率等のパラメー タを適切に設定して学習させ，その学習結果より考察せよ．
   学習結果より，それぞれの手法の特徴について考察せよ．

2. サーバアクセスコントロールの例題において，長期的な利益の最大化を目指す場合，どのような行動を選択すべきか? Q-learning および確率的傾斜法の割引率等のパラメー タを適切に設定して学習させ，その結果より考察せよ．学習結果より，それぞれの手法の特徴について考察せよ．

3. 隠れ状態の例題において，Q-learning および確率的傾斜法の学習を比較し，違いとその原因について考察せよ．

4. マルコフ決定過程の環境において 4.2.1節の定理 1が成り立つことを示せ．

5. サー バアクセスコントロー ルの例題で示した単層ニューラルネットと Bernoulli logistic unit を用いた確率的傾斜法において，適正度が式 6になることを示せ．

6. 強化学習の応用できそうな問題を挙げよ．このとき，エー ジェントの状態観測，行動，報酬がそれぞれ何に対応するのかを示し，産業応用上の利点を考察せよ．

レポー ト提出期限：8 月 28 日 (金) 午後 5 時
レポー ト提出先　：G5 棟 8F 801 号室

担当：木村 元
gen@fe.dis.titech.ac.jp
http://www.fe.dis.titech.ac.jp/ gen/indexj.html

参考文献

[Barto et al. 83] Barto, A. G., Sutton, R. S. and Anderson, C. W.: Neuronlike Adaptive Elements That Can Solve Difficult Learning Control Problems, *IEEE Transactions on Systems, Man, and Cybernetics*, vol. SMC-13, no.5, September/October 1983, pp. 834-846.

[Barto et al. 95] Barto, A. G., Bradtke, S. J. and Singh, S. P.: Learning to act using real-time dynamic programming, *Artificial Intelligence 72 (1995)*, 81-138.

[Chrisman 92] Chrisman, L.: Reinforcement learning with perceptual aliasing: The Perceptual Distinctions Approach, *Proceedings of the 10th National Conference on Artificial Intelligence*, pp. 183-188 (1992).

[Goldberg 89] Goldberg, D. E.: Genetic Algorithms in Search, Optimization and Machine Learning, Addison-Wesley (1989).

[Jaakkola 94] Jaakkola, T., Singh, S. P., & Jordan, M. I.: Reinforcement Learning Algorithm for Partially Observable Markov Decision Problems, *Advances in Neural Information Processing Systems 7*, pp.345-352 (1994).

[Kaelbling et al.96] Kaelbling, L. P., & Littman, M. L., & Moore, A. W.: Reinforcement Learning: A Survey, *Journal of Artificial Intelligence Research*, Vol. 4, pp. 237–277 (1996).

[Kimura 95] Kimura, H., Yamamura, M., & Kobayashi, S.: Reinforcement Learning by Stochastic Hill Climbing on Discounted Reward, *Proceedings of the 12th International Conference on Machine Learning*, pp.295-303 (1995).

[木村 96] 木村 元，山村 雅幸，小林 重信: 部分観測マルコフ決定過程下での強化学習：確率的傾斜法による接近，人工知能学会誌, Vol.11, No.5, pp.761-768 (1996).

[Kimura et al. 97] Kimura, H., Miyazaki, K. and Kobayashi, S.: Reinforcement Learning in POMDPs with Function Approximation, *Proceedings of the 14th International Conference on Machine Learning*, pp. 152–160 (1997).

[木村 97] 木村 元，Kaelbling, L. P.: 部分観測マルコフ決定過程下での強化学習，人工知能学会誌, Vol.12, No.6, pp.822-830 (1997).

[Kimura et al. 98] Kimura, H. & Kobayashi, S.: Reinforcement Learning for Continuous Action using Stochastic Gradient Ascent, *Submitted to International Conference on Intelligent Autonomous Systems 5*, (1998).

[Lin 92] Lin, L. J. & Mitchell, T. M.: Reinforcement Learning With Hidden States, *Proceedings of the 2nd International Conference on Simulation of Adaptive Behavior (ICSAB)*, pp. 271-280 (1992).

[Littman 94] Littman, M. L.: Memoryless Policies: Theoretical Limitations and Practical Results, *Proceedings of the 3rd International Conference on Simulation of Adaptive Behavior (ICSAB)*, pp. 238-245 (1994).

[Littman 95] Littman, M. L., Cassandra, A. R., & Kaelbling, L. P.: Learning policies for partially observable environments: Scaling up, *Proceedings of the 12th International Conference on Machine Learning*, pp. 362-370 (1995).

[McCallum 93] McCallum, R. A.: Overcoming Incomplete Perception with Utile Distinction Memory, *Proceedings of the 10th International Conference on Machine Learning*, pp. 190-196 (1993).

[McCallum 95] McCallum, R. A.: Instance-Based Utile Distinctions for Reinforcement Learning with Hidden State, *Proceedings of the 12th International Conference on Machine Learning*, pp. 387-395 (1995).

[宮崎 94] 宮崎 和光，山村 雅幸，小林 重信: 強化学習における報酬割当の理論的考察, 人工知能学会誌, Vol.9, No.4, pp.580-587 (1994).

[小笠原正巳，坂本武司 67] 小笠原正巳，坂本武司 著，北川 敏男 編: 情報科学講座 (全 62 巻)A・5・1 マルコフ過程，共立出版 (1967).

[Singh 94] Singh, S. P.: Reinforcement Learning Algorithms for Average-Payoff Markovian Decision Processes, *Proceedings of the 12th National Conference on Artificial Intelligence*, Vol. 1, pp. 700-705 (1994).

[Singh 94] Singh, S. P., Jaakkola, T., & Jordan, M. I.: Learning Without State-Estimation in Partially Observable Markovian Decision Processes, *Proceedings of the 11th International Conference on Machine Learning*, pp. 284-292 (1994).

[Singh 96] Singh, S. P., & Sutton, R.S.: Reinforcement Learning with Replacing Eligibility Traces, *Machine Learning 22*, pp. 123-158 (1996).

[Singh et al. 96] Singh, S., and Bertsekas, D.: Reinforcement Learning for Dynamic Channel Allocation in Cellular Telephone Systems, *Advances in Neural Information Processing Systems 9*, pp. 974-980 (1996).

[Sutton 88] Sutton, R. S.: Learning to Predict by the Methods of Temporal Differences, *Machine Learning 3*, pp. 9-44 (1988).

[Sutton 90] Sutton, R. S.: Reinforcement Learning Architectures for Animats, *Proceedings of the 1st International Conference on Simulation of Adaptive Behavior*, pp. 288-295 (1990).

[Sutton et al. 98] Sutton, R. S. & Barto, A.: Reinforcement Learning: An Introduction, *A Bradford Book*, The MIT Press (1998).

[Watkins et.al 92] Watkins, C. J. C. H., & Dayan, P.: Technical Note: Q-Learning, *Machine Learning 8*, pp. 55-68 (1992).

[Whitehead 95] Whitehead, S. D., & Lin, L. J.: Reinforcement learning of non-Markov decision processes, *Artificial Intelligence 73*, pp.271-306 (1995).

[Williams 92] Williams, R. J.: Simple Statistical Gradient Following Algorithms for Connectionist Reinforcement Learning, *Machine Learning 8*, pp. 229-256 (1992).