迷路探索問題Frozen Lakeにおける最適な行動選択ポリシー及び学習エージェントの提案

1.はじめに

本実験では,迷路探索問題Frozen Lakeにおける最適な行動選択ポリシー及び学習エージェントの決定を行った.性能評価には未学習の状態から1500エピソードの学習を1シミュレーションとし,30回のシミュレーションを行ったときに得られたデータを使用した.また,性能評価の指標として以下の2つを用いた.

-**第一指標**：最後の100エピソードにおける平均獲得報酬.ここで,計算方法として,30回のシミュレーションを行うため,30個の平均獲得報酬が得られるが,さらにその平均を取ることとする.また,最低基準を0.30とする.

-**第二指標**：平均獲得報酬が0.4以上になるまでのエピソード数.

今回の実験では,特にこの第一指標に焦点を当て,最後の100エピソードにおける平均獲得報酬がより大きくなるような学習方法を模索した.第一指標の値の大きい学習では,必然的に第二指標の値も小さくなると考えたため,第二指標については最後のまとめで確認することとする.また,今回想定しうる学習エージェントとして,以下の学習則をもつようなものを考える。

s:状態 a:行動 r:報酬 t:時刻 Q:行動価値関数 V:状態価値関数 α:学習係数 γ:割引率

Qlearner	$Q(s_t,a_t) \leftarrow Q(s_t,a_t) + \alpha\{r_{t+1} + \gamma \max_{a'}(s_{t+1},a'_t) - Q(s_t,a_t)\}$	(1)
SARSAlearner	$Q(s_t,a_t) \leftarrow Q(s_t,a_t) + \alpha\{r_{t+1} + \gamma Q(s_{t+1},a_{t+1}) - Q(s_t,a_t)\}$	(2)
ActorCritic	TD error：$\delta_t = r_{t+1} + \gamma V(s_{t+1}) - V(s_t)$	(3)
	Critic：$V(s_t) \leftarrow V(s_t) + \alpha\delta_t$	(4)
	Actor：$Q(s_t,a_t) \leftarrow Q(s_t,a_t) + \alpha\delta_t$	(5)

さらに,今回想定しうる行動選択ポリシーとして,以下のようなものを考える。

①**greedy法**...常に行動価値関数Qが最大となるような行動を選択する.
②**ε-greedy法**...εの確率でランダムな行動をとり,それ以外は行動価値関数Qが最大となるような行動を選択する.
③**softmax法**...行動価値関数をQ(s,a)としたとき,状態sとしたときの行動確率$P(a_i|s)$を以下のように定める手法.

$$P(a_i|s) = \frac{e^{\frac{Q(s,a_i)}{T}}}{\sum_{j=1}^{n} e^{\frac{Q(s,a_j)}{T}}} \quad (n:\text{状態数}, T:\text{正の定数}) \qquad (6)$$

以上の条件のもと,実験1で行動選択ポリシー及び学習エージェントの決定,実験2,3で最適なパラメータの決定を行った。

2.実験1：行動選択ポリシー及び学習エージェントの決定

【実験の目的】　行動選択ポリシー及び学習エージェントの決定
【実験の手法】　行動選択ポリシー①greedy法②ε-greedy法③softmax法に対して,それぞれ学習エージェントQ,SARSA,ActorCriticにおける第一指標の平均獲得報酬の値を調べた.学習エージェントのパラメータとして学習係数α,割引率γ,行動価値関数の初期値p_0,状態価値関数の初期値v_0があるが,私は特にα,γが重要なのではないかと考えたため,ひとまず$p_0 = 0, v_0 = 0$で固定し,$\alpha(0 < \alpha < 1)$,$\gamma(0 < \gamma < 1)$対して0.1区切りで数値を与えて,横軸α縦軸γとして平均獲得報酬の値をヒートマップを用いて図示を行う.なお,②ε-greedy法では行動選択ポリシーのパラメータ$\varepsilon(0 < \varepsilon \leq 1)$があるので,0.2区切りで数値を与え,一つの学習エージェントにつき5個のヒートマップを作製した.同様に,③softmax法についても行動選択ポリシーのパラメータT(正の定数)があるので,T=0.25,0.50,1.0,2.0,4.0で数値を与え,一つの学習エージェントにつき5個のヒートマップを作成した.ただし,②③に関しては紙面の都合上,すべてのヒートマップを示すことはできないので,各ヒートマップの最大値を表にまとめ,一番大きい値をとるパラメータのヒートマップのみを図として示すこととする.

【実験結果】

①greedy法...Q学習,SARSA学習,ActorCriticの結果を,それぞれ図1,2,3に示す.greedy法ではどの学習エージェントにおいても第一指標の最低基準(0.30)を下回っており,良い結果が得られなかった.また,パラメータの値による法則性も全く見受けられなかった.

図1

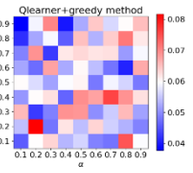


図2

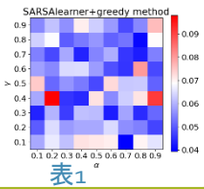


図3

②ε-greedy法...パラメータεに対する,Q学習,SARSA学習,ActorCriticの各ヒートマップにおける最大平均獲得報酬を表1に示す.表より,ε=0.2の時,最大平均獲得報酬が一番大きいものとなったので,ε=0.2の時のQ学習,SARSA学習,ActorCriticのヒートマップをそれぞれ図4,5,6に示す.結果として全ての場合において,第一指標の最低基準を下回っていたが,共通してγを大きくしたり,εを小さくすることで平均獲得報酬が大きくなることや,Q学習及びSARSA学習ではαを小さくしたり,ActorCriticではαを中間あたりの値にすると平均獲得報酬が大きくなるといったパラメータの値による法則性が見られた.

表1

平均獲得報酬(最大)	ε=0.2	ε=0.4	ε=0.6	ε=0.8	ε=1.0
Qlearner	0.18(α=0.1,γ=0.9)	0.08(α=0.1,γ=0.9)	0.04(α=0.1,γ=0.9)	0.03(α=0.3,γ=0.9)	0.02(α=0.1,γ=0.8)
SARSAlearner	0.14(α=0.1,γ=0.9)	0.07(α=0.1,γ=0.9)	0.04(α=0.2,γ=0.9)	0.02(α=0.2,γ=0.8)	0.02(α=0.5,γ=0.9)
ActorCritic	0.18(α=0.4,γ=0.9)	0.09(α=0.7,γ=0.9)	0.05(α=0.3,γ=0.8)	0.03(α=0.4,γ=0.8)	0.02(α=0.8,γ=0.8)

図4

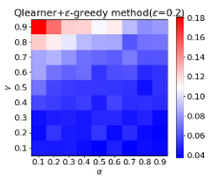


図5

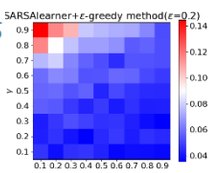
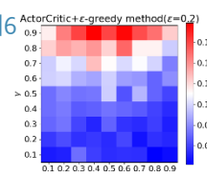


図6



③softmax法...パラメータTに対する,Q学習,SARSA学習,ActorCritcの各ヒートマップにおける最大平均獲得報酬を表2に示す.表より,T=0.25の時,最大平均獲得が一番大きいものとなったので,T=0.25の時のQ学習,SARSA学習,ActorCriticのヒートマップをそれぞれ図7,8,9に示す.結果としてActorCritic(T=0.25, 0.5,1.0)のとき最低基準を上回るものが現れた.

また,全てに共通してγを大きくしたり,Tを小さくすることで平均獲得報酬が大きくなることや,ActorCriticではαを中間あたりの値にすると平均獲得報酬が大きくなるといったパラメータの値による法則性が見られた.

⇒今回は,第一指標の最低基準を上回ることがあるActorCritic+softmax法を採用する.

表2

平均獲得報酬(最大)	T=0.25	T=0.5	T=1.0	T=2.0	T=4.0
Qlearner	0.03(α=0.9,γ=0.9)	0.02(α=0.6,γ=0.9)	0.02(α=0.8,γ=0.9)	0.02(α=0.2,γ=0.7)	0.02(α=0.1,γ=0.7)
SARSAlearner	0.02(α=0.2,γ=0.9)	0.02(α=0.6,γ=0.7)	0.02(α=0.1,γ=0.5)	0.02(α=0.5,γ=0.7)	0.02(α=0.7,γ=0.9)
ActorCritic	0.62(α=0.5,γ=0.9)	0.59(α=0.6,γ=0.9)	0.38(α=0.8,γ=0.9)	0.11(α=0.9,γ=0.9)	0.04(α=0.9,γ=0.9)

図7

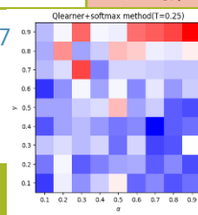


図8

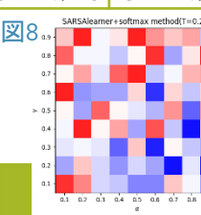
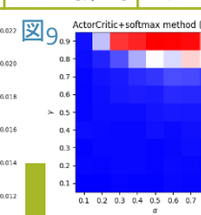


図9



3.実験2：最適なパラメータの決定(α,γ,T)

【実験の目的】　ActorCritc+softmax法におけるパラメータ(α,γ,T)の決定
【実験の手法】　実験2では,先ほど採用したActorCritic+softmax法における最適なパラメータを一部決定する.先ほどと同様にひとまず$p_0 = 0, v_0 = 0$とし,それ以外のパラメータα,γ,Tを決定する.まず,γに関しては,実験1の図9のヒートマップの傾向からγを大きくすると平均獲得報酬が大きくなることがわかったので,γ=1.0で確定させた.残り,αとTについて考えればよいので,αとTの値を変えて第一指標の平均獲得報酬を調べれば良いということとなる.ということで,α(0<α<1)は0.1区切りで数値を与え,Tは実験1の表2の結果からActorCriticでは,1以下で平均獲得報酬が最低基準を上回ることがわかったので,Tの範囲を0<T<1に絞り,0.1区切りで数値を与え,第一指標の平均獲得報酬を調べた.その結果を横軸α縦軸平均獲得報酬scoreとして,折れ線グラフにて示す.折れ線グラフでは,平均獲得報酬の値と共に標準偏差も表示し,さらに図10ではT=0.1～0.3,図11ではT=0.4～0.6,図12ではT=0.7～0.9のグラフを同時に表示することとする.
【実験結果】　前述の図10～12を以下に示す.αについては中間あたりの値をとると平均獲得報酬が大きくなることがわかった.また,Tについてはαが中間あたりの値を取る場合,平均獲得報酬は多少の違いはあれど,ほとんど違いが無いことがわかった.
⇒今回は,γ=1.0,Tはαの値によって平均獲得報酬の値の変動が比較的小さいT=0.4,αはT=0.4の中で標準偏差の小さいα=0.4を採用する.

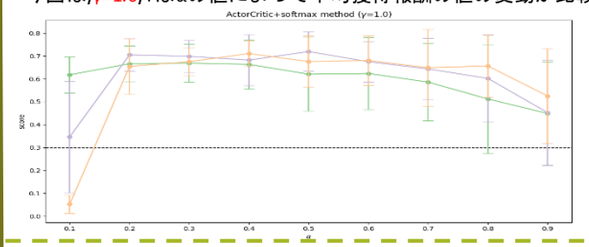


図10

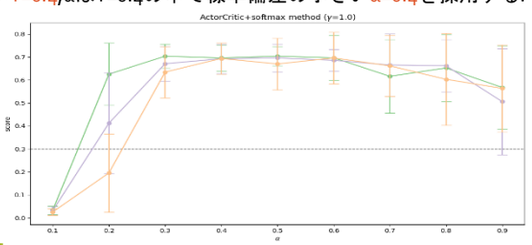


図11

4.実験3：最適なパラメータの決定(q_0, v_0)

【実験の目的】　ActorCritic+softmax法におけるパラメータ(q_0, v_0)の決定
【実験の手法】　実験3では,ActorCritic+softmax法における最適なパラメータの最終決定を行う.先ほど最適なパラメータとしてγ=1.0,T=0.4,α=0.4を決定したので,最後にこのパラメータの下で最適なパラメータq_0, v_0の決定を行う.q_0, v_0にそれぞれ-2.0,-1.0,0.0,+1.0,+2.0の数値を与えて,第一指標の平均獲得報酬の値を調べた.その結果を横軸v_0縦軸平均獲得報酬scoreとして,折れ線グラフにて示す.折れ線グラフでは,平均獲得報酬の値と共に標準偏差も表示し,さらにq_0=-2.0,-1.0,0.0,+1.0,+2.0の時のグラフを同時に表示することとする.
【実験結果】　実験結果を図13に示す.v_0が負だと極端に報酬が小さくなり,非負ならばほとんど報酬の違いが生じなかった.また,q_0の値による違いは多少標準偏差の違いはあれどほとんど見受けられなかった.
⇒今回は,v_0は一番平均獲得報酬の高くなったv_0=1.0,q_0は一番標準偏差の小さくなったq_0=2.0を採用する.

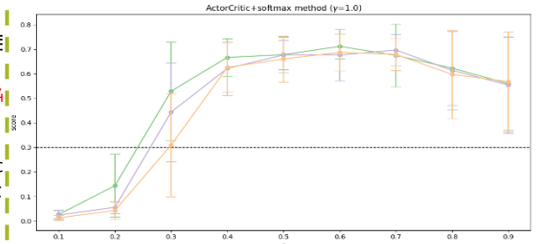


図12

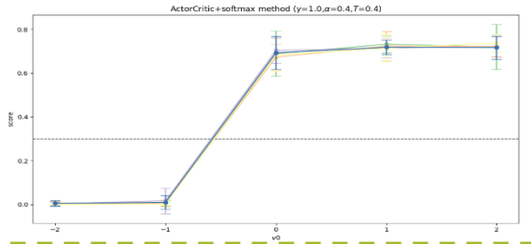


図13

5.考察

これまでの実験結果をもとに,考察を行っていく.大きく①行動選択ポリシー及び学習エージェント②パラメータに分けて考察を行う.

①行動選択ポリシー及び学習エージェント...最適なものとしてActor Critic+softmax法になった理由について考察をしてみる.

まず,学習エージェントActorCriticとなった理由として,学習則のTD誤差((3)式)によるものだと考える.Frozen Lakeではスリップという必ずしも自分の思った行動がとれるとは限らない要素があるため,行動を含まない学習則の方が良いのではないかと考えた.したがってそのような学習則を唯一もつActorCriticで平均獲得報酬が高くなり,行動を含む学習則をもつQ学習及びSARSA学習では平均獲得報酬がそこまで高くならなかったのではないかと考える.

また,行動選択ポリシーがsoftmax法となった理由として,学習エージェントと行動選択ポリシーの相性によるものだと考える.実験1の結果から,同じ学習エージェントをもつ場合でも,行動選択ポリシーによって平均獲得報酬が大きく異なり,Q学習とSARSA学習はε-greedy法,ActorCriticはsoftmax法と組み合わせると報酬が高くなった.したがって,学習エージェントとしてActorCriticが選ばれたことで,相性の良いsoftmax法が選ばれたのではないかと考える.

②パラメータ...最適なパラメータとしてα=0.4,γ=1.0,T=0.4,v_0=1.0,q_0=2.0を採用したが,この値になった理由について考察してみる.

α→学習の速度を表すパラメータであり,α=0.4とαが中間あたりの値を取るとき平均獲得報酬が高くなった.これはActorCriticの適正学習係数が中間あたりの値であるからだと考えられる.図6,9からαを中間あたりに設定すると高い報酬を得られることがわかるので,ActorCriticでは学習速度を速くもなく遅くもなくすることで良い結果が得られると考えられる.ちなみに図3を考慮しなかった理由としては,図1,2よりgreedy法ではそもそも学習が上手くいっていないと判断したからである.

γ→目先の報酬をどれくらい重視するかを表すパラメータであり,小さければ小さいほど目先の報酬を重視してしまう.したがって,γ=1.0と大きな値にし,将来の報酬を最大限に重視させることにより,高い平均獲得報酬が得られたのだと考える.

T→softmax法で用いられるパラメータであり,Tが大きくなるほどrandomに近い行動選択をしてしまう.したがって,T=0.4と比較的小さい値にし,行動確率をより行動価値関数Qの値に依存させたことにより,高い平均獲得報酬が得られたのだと考える.

q_0, v_0→ 実験3ではq_0による報酬の変化はほぼなかったので,v_0についてのみ考察を行う.実験3ではv_0の値が負になると極端に獲得報酬が低くなったが,これはFrozenLakeでは報酬が0～1であるので,初期値を負にすることで,学習がうまく進まなかったことが原因だと考えられる.したがって,$v_0 = 1.0$と非負にすることで学習がうまく進み高い平均獲得報酬が得られたのだと考える.

6.まとめ

以上より,私はActorCritic+softmax法(α=0.4,γ=1.0,T=0.4,$v_0 = 1.0, q_0 = 2.0$)を提案する.

最後にこのときの第一指標と第二指標がどうなったか確認を行う.第一指標は,最後の100エピソードにおける平均獲得報酬(30シミュレーション分の箱ひげ図である図14,第二指標は,各エピソードで30シミュレーション分の獲得報酬の平均をとったものの推移を横軸episode縦軸平均獲得報酬として表した図15にて確認を行う.

まず,図14より,第一指標の平均獲得報酬が0.702となり,一番悪いシミュレーションで最小値0.63,一番良いシミュレーションで最大値0.78をとることがわかった.

次に,図15では,最低基準である0.30を赤線,第二指標の基準である0.40を青線で表示しているが,約200エピソードを超えたあたりで報酬0.40を超えることがわかった.

図14

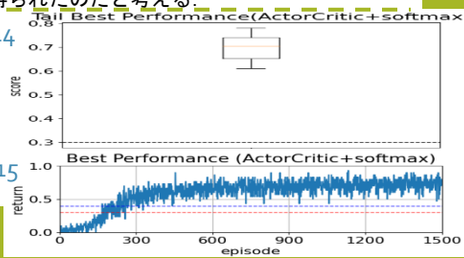


図15