F_004

方策こう配法を用いた行動学習
——方策中での状態遷移確率の表現——

Behavior Learning Based on a Policy Gradient Method
—— an Expression of State Transition Probabilities Included in the Policies ——

石原 聖司† 五十嵐 治一‡
Seiji Ishihara Harukazu Igarashi

1．まえがき

強化学習によりエージェントの行動決定方法（方策）を学習する際には，通常，環境の状態遷移確率を学習する必要はない．例えばQ学習では，このような情報は，行動価値関数$Q(s,a)$の値として，問題解決のための行動決定に関する知識と一体的に学習されていく．しかし，環境に依存しないで有効な行動決定をもたらす普遍的な知識は多くの問題で存在する．例えば，追跡問題において，ハンターが獲物に接近する行動を選択した方が捕獲に貢献し高報酬につながるといった知識はハンターの動作特性に無関係である．ところが，ぬかるみに足を取られたり強風に煽られたりするなどの理由で，必ずしも思い通りの方向にハンターが進めず確率的にしか状態を遷移できない場合，学習により得られた方策はこの動作特性（＝状態遷移確率）の影響を受ける．もし，方策に関する知識が状態遷移確率に依存するものと依存しないものとに分離された形で学習できれば，後者の知識を別の環境下でも容易に再利用できるという利点がある．

方策こう配法による強化学習では，Q学習などとは異なり，状態価値関数や行動価値関数を求めることなく，方策中のパラメータを直接学習することができる[1]～[3]．さらに，各時刻における行動決定問題をある目的関数の最小化問題として定式化し，その目的関数の期待値を一定にするという条件下で行動決定に関する不確定さ（エントロピー）を最大にする確率的方策を求めると，それはボルツマン型の確率分布関数となる．これまで我々は，このようなボルツマン型の確率的方策を用いた方策こう配法を追跡問題に適用してきた [4]．ただし，そこでの実験は状態遷移が決定論的な場合にとどまっていた．

本研究では，状態の遷移，すなわちエージェントの動作特性が確率的である場合を対象とする．以下では，エージェントの動作特性を表すパラメータ（動作特性パラメータ）と，状態遷移確率に依存しない行動知識に関するパラメータ（行動知識パラメータ）の2種類のパラメータを含む目的関数の例とその学習則とを示す．さらに，最短経路問題への適用実験によって，双方のパラメータが方策こう配法により適切に学習できることを示す．

2．方策こう配法による行動学習

2.1 目的関数と方策

状態$s\in S$におけるエージェントの行動$a\in A$を決定する方策$\pi(a;s,\{\theta(s)\},\{\omega(s,s';a)\})$を以下のボルツマン型の分布関数で定義する．

$$\pi(a;s,\{\theta(s)\},\{\omega(s,s';a)\})\equiv\frac{e^{-E(a;s,\{\theta(s)\},\{\omega(s,s';a)\})/T}}{\sum_a e^{-E(a;s,\{\theta(s)\},\{\omega(s,s';a)\})/T}} \quad (1)$$

ただし，目的関数Eを次のように定義する．

$$E(a;s,\{\theta(s)\},\{\omega(s,s';a)\})\equiv-\sum_{s'}\theta(s')\omega(s,s';a) \quad (2)$$

ここで，$\theta(s)$は状態sの価値を表す行動知識パラメータであり，$\omega(s,s';a)$は状態sで行動aを選択したときに状態s'へ遷移する度合いを表す動作特性パラメータである．

2.2 学習則

今，エピソードごとのパラメータ更新を考える．各エピソードは実際に選択した行動列$\{a(t)\}$と実現された状態列$\{s(t)\}$で表される．エピソードごとに与えられる報酬rの期待値$E[r]$を最大にするパラメータを求めるために確率的こう配法[4]を用いると，次の学習則を得る．

$$\Delta\theta(s')=\varepsilon\cdot r\sum_{t=0}^{L-1}\frac{1}{T}\left[\omega(s(t),s';a(t))-\sum_a\omega(s(t),s';a)\pi(a;s(t))\right] \quad (3)$$

ここで，tは（離散）時刻，Lはエピソードの長さをそれぞれ表す．式(3)では，エピソード中に出現した状態$s(t)$から実際に選択した行動$a(t)$により遷移可能な状態s'の価値$\theta(s')$が更新される．

同様に，動作特性パラメータ$\omega(s,s';a)$の学習則は，

$$\Delta\omega(s,s';a)=\varepsilon\cdot r\sum_{t=0}^{L-1}\frac{1}{T}\left[\delta_{a,a(t)}-\pi(a;s(t))\right]\theta(s')\delta_{s,s(t)} \quad (4)$$

となる．なお，$\delta_{i,j}$（$i\in S,j\in S$）は，$i=j$ならば1，$i\neq j$ならば0をとる関数である．式(4)では，エピソード中に出現した状態$s(t)$から状態s'へと，行動aの選択によりその状態が遷移する度合い$\omega(s,s';a)$が更新される．この際，実際に選択された行動$a(t)$については，式(4)の右辺の括弧内が非負なので状態s'の価値$\theta(s')$の大きさに比例して$\omega(s,s';a)$の値が増加する．一方，選択されなかった行動$a\neq a(t)$については，状態s'への遷移が抑制されるよう$\omega(s,s';a)$の値が減少する．この強化と抑制の度合いは，そのエピソードに与えられた報酬rに比例している．

3．最短経路問題への適用

3.1 最短経路問題

2次元のグリッド上において，一つのエージェントが目的地までマス目を移動する問題を考える．本研究では，条件を以下のように設定する．

・目標達成時にのみエージェントに報酬を与える．

†近畿大学工学部
‡芝浦工業大学工学部

・エージェントの行動は，上下左右いずれかの方向へ1マス移動するという四つに限る（静止はない）．
・グリッドを7×7の格子状トーラスとする．
・エージェントの初期配置と目的地の配置とは，エピソードごとに毎回ランダムとする．
・エージェントの視界を7×7とする（完全知覚）．
・状態 s で行動 a が選択されたときの遷移先の状態 s' は，予め定められた状態遷移確率 $P(s,s';a)$ により決定される．

初期状態から目標達成時までを1エピソードとし，目標達成に要した時間ステップ数をエピソード長 L とする．報酬の値 r を $1/L^2$ とし，パラメータの更新を各実験において10万エピソード繰り返した．

3.2 実験1：決定論的な状態遷移の場合

エージェントが選択した方向へ必ず移動する場合，すなわち，状態遷移確率の値が0または1に限られる場合について，次のような2種類の行動学習実験を行った．なお，このときの状態遷移確率を $P_1(s,s';a)$ で表す．

[実験1.1] ω を P_1 に固定し，θ を学習．
[実験1.2] ω を P_1 と正反対（選択した方向と逆の方向へ必ず移動する場合）に固定し，θ を学習．

実験1.1および実験1.2から得られた θ の値を等高線で表現した図を図1(a)および図1(b)にそれぞれ示す．これらの図は，2次元格子の中心を目的地としたときの各グリッドの重み，すなわちエージェントの現在地となるグリッドに対応する状態の価値を表している．図1(a)は，中心に近づくほど状態の価値が上昇することから，最短経路問題において高い報酬を得るための行動知識を正しく表していると言える．一方，ω の値が P_1 と正反対である場合の図1(b)は，中心から離れるほど状態の価値が上昇する傾向が見てとれる．つまり，目的地に近づくためには，あえて反対方向への移動を選択しなければならないという行動知識を正しく表現していると言える．

3.3 実験2：確率的な状態遷移の場合

エージェントの移動先が選択行動に対応した方向から確率 p の割合で右にずれる（例えば，上方向を選択しても移動先は確率 p の割合で右方向となる）場合について，次のような2種類の行動学習実験を行った．このときの状態遷移確率を $P_2(s,s';a)$ で表す．なお，以下の実験では，$p=\{0.1,0.2,0.3,0.4,0.5\}$ の5通りについて学習を行った．

[実験2.1] ω を P_2 に固定し，θ を学習．
[実験2.2] θ を実験1.1で求めた値に固定し，P_2 の下で ω を学習．

実験2.1から得られた θ の値は，いずれの p についても実験1.1で求めた θ の値と同じく，中心に近いほど高い値となった．これは，方策に関する知識を分離することで，行動知識の再利用が可能であることを示している．$p=0.1$ および $p=0.5$ のそれぞれの場合について，θ の値を図1と同様の方法で図2(a)および図2(b)にそれぞれ示す．また，実験2.2から得られた動作特性パラメータ ω の値と，対応する状態遷移確率 P_2 の値との二乗誤差の平均値は，いずれの (s,s',a) についても0.003以下とかなり小さくなった．つまり，方策こう配法によりエージェントの動作特性を推定することが出来た．

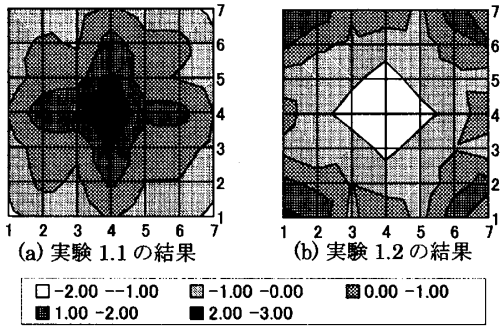

図1. 実験1により得られた $\theta(s)$ の値

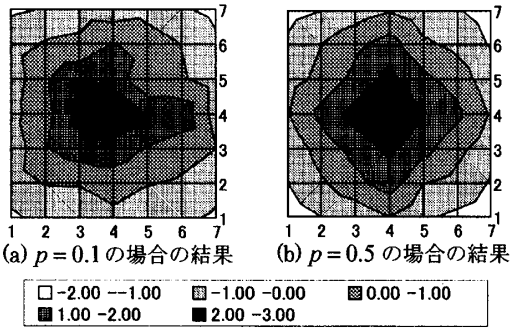

図2. 実験2.1により得られた $\theta(s)$ の値

3.4 実験3：θ と ω の値を同時に学習する場合

実験2で用いた5通りの p に関する P_2 の下，θ と ω をそれぞれ同時に学習する実験を行った．その結果，実験3から得られた θ および ω の値は，実験2の結果に近い値となった．このことから，2種類のパラメータ θ と ω を同時に学習することが可能であることが示された．

4. まとめ

本研究では，状態の遷移が確率的である場合の方策こう配法による強化学習を扱った．最短経路問題への適用実験の結果，方策こう配法では，状態遷移確率とそれによらない問題固有の行動知識とを分離して学習できることを確認した．今後は，非マルコフ決定過程やマルチエージェント系への適用を試みて行く予定である．

参考文献

[1] R.J. Williams, "Simple statistical gradient-following algorithms for connectionist reinforcement learning," Machine Learning, vol. 8, pp. 229-256, 1992.
[2] 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, Sept. 1996.
[3] L.C. Baird and A.W. Moore, "Gradient descent for general reinforcement learning," Advances in Neural Information Processing Systems 11, The MIT Press, pp.968-974, July 1999.
[4] 石原聖司，五十嵐治一，"マルチエージェント系における行動学習への方策勾配法の適用—追跡問題—，" 電子情報通信学会論文誌，Vol.J87-D-I, No.3, pp.390-397, Jul. 2004.