LF-001

Profit Sharingの学習の合理性に関する理論的考察
A Theoretical Analysis of Rationality in Profit Sharing

北川 哲[†]　田村 直之[‡]
KITAGAWA, Satoshi　TAMURA, Naoyuki

1 はじめに

強化学習は学習者 (エージェント) と学習対象 (環境) との相互作用により得られる情報から適切な行動戦略 (方策) を獲得する学習の枠組みである．エージェントは挙動方策と呼ばれる確率的な方策を用いて環境を探索し，目標達成時に報酬と呼ばれる特別な信号を獲得する．学習は探索と探索により得た知識にもとづいた方策の更新を繰り返すことにより行われる．このとき挙動方策を更新するものを方策オン型学習と呼び，挙動方策と異なる方策を更新するものを方策オフ型学習と呼ぶ．

強化学習は方策の更新に必要な計算量が環境の大きさに依存しないため，動的計画法などの最適化手法では次元の呪いにより現実的な時間に解くことのできない環境でも有効に働くことが期待されている．

強化学習では開始状態もしくは報酬を得た状態から次にエージェントが報酬を得るまでの一連の探索をエピソードと呼ぶ．Profit Sharing [1](PS) はエピソードごとに，開始状態から目標状態までの行動系列 (バックアップ) と目標状態で獲得する報酬信号を用いて方策の更新を行う強化学習法である．

PS は Q 学習などの 1 ステップバックアップによる学習法と比較して高速な学習法となることが知られているが，そのための学習の条件は明らかにされていない．本論文の目的は PS が高速な学習法となる学習の条件を明らかにすることである．

強化学習は主に最適な方策を獲得するのが困難な環境に用いられるため，学習は基準を満たす有効な方策が得られたときに打ち切られる．したがって強化学習における学習の速さは「有効な方策を得るまでの探索時間と方策の更新時間の和の逆数」と定義できる．一般に探索時間と方策の更新時間は比例し，特に PS は探索時間に対する方策の更新時間が最も少ないクラスの学習法である．したがって PS の学習の高速化には探索時間の短縮が重要であり，方策オン型学習において挙動方策を効率よく改善する必要がある．

PS にはこれまでに合理性定理 [2)3)4)5)] と呼ばれる学習の条件が提案されている．しかしそれらは方策オン型学習において挙動方策を効率よく改善するという観点からは導出されておらず．学習の高速化に有効であるかは明らかでない．

本論文では，(1) PS における方策オフ型学習および方策オン型学習の合理性を定義し，それぞれの合理性を満たす学習の条件を合理性定理として提案し，数値実験によりその有効性を示す．(2) 報酬プラン獲得定理 [2)] を方策オフ型学習および方策オン型学習の合理性の観点から考察し，本論文で導出する学習の条件が報酬プラン獲得定理と同等の有効性を持ち，より大規模な環境にも適用できることを示す．

2 章では方策オフ型学習および方策オン型学習の合理性を定義し，それぞれの合理性定理を導出する．3 章では報酬プラン獲得定理 が greedy 方策による方策オフ型学習の合理性とルーレット選択による方策オン型学習の合理性を同時に保証する学習の条件と一致することを示し，報酬プラン獲得定理が PS の学習の高速化に有効となること示す．4 章では数値実験を行い，方策オフ型学習の合理性定理および方策オン型学習の合理性定理の学習性能を確認する．5 章はまとめとする．

2 PS の学習の合理性

PS ではルーレット選択や ϵ-greedy 方策などの確率的な方策により生成した行動系列において行動の重みを更新することで学習を行う．具体的には，エピソードごとに開始状態から目標状態までの離散的な時間ステップ t で訪問した状態 s_t と選択した行動 a_t の組を行動系列として記憶しておき，時間ステップ T で目標状態において報酬を得たときに行動系列に含まれる行動の重みを次式にしたがって更新する．

$$\forall t = 0, \cdots, T.\ W(s_t, a_t) \leftarrow W(s_t, a_t) + f(t, r_T, T) \quad (1)$$

ここで $W(s_t, a_t)$ は時間ステップ t で選択した行動の重み，r_T は時間ステップ T で与えられる正値の報酬値，f は行動系列に含まれる行動に報酬を割当てる関数で強化関数と呼ばれる．

学習結果は推定方策と呼ばれる行動の重みにしたがう何らかの方策により取り出される．推定方策には最大の重みを持つ行動を選択する greedy 方策が用いられることが多い．学習の目的は推定方策により単位時間あたりに獲得される報酬を最大化することである．

本章では PS における方策オフ型学習と方策オン型学習の合理性を定義し，それぞれの合理性を保証するための強化関数の条件を導く．

2.1 方策オフ型学習の合理性

PS では学習結果は式 (1) により更新された行動の重みにしたがう推定方策を用いて行動を選択することにより取り出される．推定方策が単位時間あたりの獲得報酬を最大化するには，推定方策により必ず目標状態へ遷移できることが必要条件のひとつとなる．

ここでは学習後の推定方策が必ず目標状態へ遷移できることを方策オフ型学習の合理性として定義する．

定義 1 (方策オフ型学習の合理性) 任意の挙動方策により生成された行動系列における行動の重みの更新を任意の回数繰り返した後に，ある推定方策を用いて目標状態に遷移できるとき PS による方策オフ型学習は合理的であるという　□

[†]神戸大学大学院 自然科学研究科
[‡]神戸大学 学術情報基盤センター

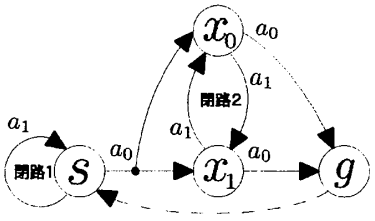

図 1: 閉路を含む単純な環境

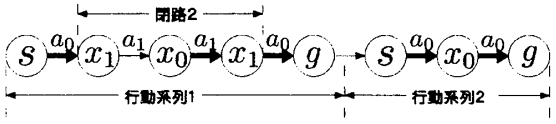

図 2: 閉路を含む単純な環境における行動系列の例

ある推定方策を用いて目標状態に遷移するには，環境に含まれるいかなる閉路においても閉路上のいずれかの状態で推定方策が閉路を出る行動を選択すれば十分である．以降では閉路上の各状態において閉路から出る行動を出力ルール，閉路上の行動を入力ルールと呼ぶ．

ここでは推定方策に greedy 方策を用いたときに方策オフ型学習の合理性を満たす学習の条件を導く．

補題 1 (greedy 方策の Off Policy 合理性) greedy 方策により目標状態へ遷移するための十分条件は

$$\forall s \in \mathcal{S},\ \arg\max_{a\in\mathcal{A}(s)} W(s,a) \neq a_{\mathrm{C}} \qquad (2)$$

かつ，環境に含まれるすべての閉路について

$$\exists s \in \mathcal{S}_{\mathrm{C}}\ \text{s.t.}\ \max_{a_i\in\mathcal{A}_i(s)} W(s,a_i) < \max_{a_o\in\mathcal{A}_o(s)} W(s,a_o) \qquad (3)$$

ここで $\mathcal{S}$ は環境に含まれる状態の集合，$\mathcal{S}_{\mathrm{C}}$ は閉路上の状態の集合，$\mathcal{A}(s)$ は状態 s で選択できる行動の集合，a_{C} は自己閉路，$\mathcal{A}_i(s)$ は状態 s における入力ルールの集合，$\mathcal{A}_o(s)$ は出力ルールの集合である． □

式 (2) は環境のそれぞれの状態に独立した条件なので局所的合理性条件，式 (3) は複数の状態に関係した条件なので大局的合理性条件と呼ばれる．ただし局所的合理性条件は $\mathcal{S}_{\mathrm{C}}$ の要素が 1 のときの大局的合理性条件に含まれるため以降では式 (3) を満たす強化関数の条件を求める．

PS は目標状態に遷移したときにのみ行動の重みの更新を行う．このため行動の重み更新時の行動系列上の各状態において最後に選択された行動は，その行動系列において目標状態に遷移することが可能な行動であり，有効ルールと呼ばれる．

図 2 は 図 1 の閉路を含む単純な環境での探索により生成される行動系列の例である．図 2 の太線矢印は各行動系列における有効ルールである．ここで行動系列上の各状態において有効ルールの重みが最大となるように報酬を割当てるとき，行動系列 1 では greedy 方策が閉路 1 上の状態 s で閉路 1 を出る行動 a_0 を，閉路 2 上の状態 x_1 で閉路 2 を出る行動 a_0 を選択するように更新される．同様に行動系列 2 では greedy 方策が状態 s で閉路 1 を出る行動 a_0 を，状態 x_0 で閉路 2 を出る行動 a_0 を選択するように更新される．これらから次の補題が導出される．

補題 2 (エピソード学習の Off Policy 合理性) 目標状態に遷移したとき，行動系列上の各状態において有効ルールの重みが最大となるように報酬を割当てると，各状態がどのような閉路上にあっても行動系列上のいずれかの状態で greedy 方策が出力ルールを選択するように行動の重みが更新される． □

補題 2 からただちに次の定理が導出される．

定理 1 (greedy 方策の Off Policy 合理性定理) 任意の方策により生成した無限個の行動系列において行動の重みを更新した後に greedy 方策により目標状態に遷移するための強化関数の条件は

$$f(t,r_T,T) = \max_{a\in\mathcal{A}(s)} W(s,a) - W(s,a_t) + \Delta \quad if\ a_t = a_e$$

$$\Delta = \begin{cases} 0 & if\ \ \arg\max_{a\in\mathcal{A}(s)} W(s,a) = a_e \\ \theta & else. \end{cases}$$

ここで a_e は行動系列上の状態 s における有効ルール．θ は任意の正値である． □

greedy 方策が有効ルールを選択することを保証するには，greedy 方策によって遷移する可能性のあるすべての状態において有効ルールが特定されていなければならない．マルコフ決定過程のように確率的な遷移をする環境において greedy 方策が未知の状態へ遷移しないことを保証するため，ここでは無限個の行動系列を仮定した．

2.2 方策オン型学習の合理性

方策オフ型学習の合理性は推定方策が報酬を獲得することを保証するため理論的に重要だが，学習の高速化には方策オン型学習において挙動方策の改善を保証する学習の条件が必要となる．ここでは方策オン型学習における行動の重みの更新ごとの学習の合理性を定義し，その合理性を満たす強化関数の条件を導出する．

探索時間を短縮し，学習を高速化するには挙動方策が閉路となる行動を選択すべきではない．したがって行動の重み更新により，閉路上の状態のいずれかで挙動方策が出力ルールを選択する確率が増加することが学習の高速化の必要条件のひとつとなる．

定義 2 (方策オン型学習の合理性) 行動系列上の各状態がどのような閉路上にあっても，行動の重み更新により行動系列上のいずれかの状態で挙動方策による出力ルールの選択確率が増加するとき PS による方策オン型学習は合理的であるという． □

ルーレット選択は状態 s で行動 a を選択する確率が $P(s,a) = \frac{W(s,a)}{\sum_{a\in\mathcal{A}(s)} W(s,a)}$ で定義される方策で，PS で

は挙動方策としてよく用いられる．ここでは探索にルーレット選択を用いたときに方策オン型学習の合理性を満たす学習の条件を導く．

補題 3 (エピソード学習の On Policy 合理性) 行動系列上の各状態においてルーレット選択による有効ルールの選択確率が増加するように報酬を割当てると，各状態がどのような閉路上にあっても行動系列上のいずれかの状態でルーレット選択による出力ルールの選択確率が増加する． □

補題 3 からただちに次の定理が導出される．

定理 2 (ルーレット選択の On Policy 合理性定理) ルーレット選択により生成した行動系列において行動の重みを更新した後に，閉路上のいずれかの状態でルーレット方策による出力ルールの選択確率が増加するための強化関数の条件は

$$\forall t = 1, \cdots, T.\ \ k_t \sum_{i=0}^{t-1} f(i, r_T, T) < f(t, r_T, T)$$
$$k_t = \frac{W(s_t, a_e)}{\sum_{a' \in A(s_t), a' \neq a_e} W(s_t, a')} \tag{4}$$

ここで a_e は時刻 t で訪問された状態 s_t における有効ルールである． □

導出は 付録 A に示す．ただし，この条件は k_t の値が状態によって変化するため実用性に欠ける．以下では方策オン型の学習の合理性を $k_t \leq 1$ の時にのみ保証するようにして実用性の高い強化関数の条件を提案する．

系 1 (ルーレット選択の合理性条件) ルーレット選択による方策オン型学習が高速化する強化関数の条件は

$$\forall t = 1, \cdots, T.\ \ \sum_{i=0}^{t-1} f(i, r_T, T) < f(t, r_T, T) \tag{5}$$

□

式 (5) は $k \leq 1$ のとき，すなわち入力ルールの選択確率が高い状態では定理 2 を満たし，方策が改善することが保証される．また $k > 1$ のとき，すなわち出力ルールの選択確率が高い状態においても出力ルールの選択確率が増加傾向にあることが保証される．さらに，無限回の行動の重み更新の後にはすべての状態で $k > 1$ となり式 (3) が成り立つため方策オフ型学習の合理性も保証できる実用的な条件である．

式 (5) を満たす強化関数には公比 0.5 の等比減少関数がある．これはどのような環境に対しても使用できる汎用性の高い強化関数である．

3 関連研究

PS にはすでにいくつかの合理性定理が提案されている．報酬プラン獲得定理 [2] は系 1 よりも強化関数の条件が厳しい定理であるが学習速度に関する解析はされておらず，学習の高速化に有効であるかは明らかでない．

定理 3 (報酬プラン獲得定理) 学習後に greedy 方策が目標状態へ遷移することを保証する強化関数の条件は

$$\forall i = 1, 2, \cdots, W.\ \ L \sum_{j=i}^{W} f_j < f_{i-1} \tag{6}$$

ここで W はエピソードの最大長，L はある状態に存在する有効ルールの最大個数である． □

系 1 に示すルーレット選択の合理性条件は定理 3 において $L = 1$ とした条件と一致する．式 (6) は $L = 1$ の時に最も多く報酬を割当てるので系 1 は定理 3 を完全に含む．

このことから報酬プラン獲得定理による学習において，挙動方策にルーレット選択を用いる場合には，行動の重みの改善が必要な状態での方策オン型の合理性が保証されるため学習の高速化に有効となることが示される．

系 1 は報酬プラン獲得定理と同様に学習後に greedy 方策により目標状態へ遷移することを保証するので，挙動方策にルーレット選択を用いるのであれば定理 3 の合理性を満たすには系 1 の条件を用いれば十分である．

また期待値合理性 [3]，拡張合理性 [4]，統計的合理性 [5] は報酬プラン獲得定理における強化関数の条件を緩和したものだが，いずれも等比減少関数にもとづいた強化関数を用いている．定理 1 は合理性を保証するのに必要なだけの報酬を割当てることができるため，既存の定理にもとづく報酬割当てよりも長い行動系列に対して効率よく報酬を割当てることができる．

4 数値実験

提案する合理性の有効性を評価するために図 3 に示す Sutton の Dyna 迷路 [6] において実験を行う． これは開始状態 (S) から目標状態 (G) までの経路を学習する問題で，状態数は 47， 最短路は 14 ステップである．エージェントは各マスを別々の状態として知覚でき，各マスにおいて上下左右からひとつ行動を選択し隣接するマスに決定論的に移動する．ただし移動先が障害物か迷路の縁の場合はその場にとどまる．エージェントは目標状態 (G) へ遷移すると報酬 1 を獲得し，再び開始状態 (S) から探索を開始する．

この問題において報酬プラン獲得定理の条件を満たす公比 0.25 の等比減少関数，提案するルーレット選択の合理性条件を満たす公比 0.5 の等比減少関数，統計的合理性定理，期待値合理性定理の条件を満たす公比 0.95 の等比減少関数を強化関数に用いた場合の Profit Sharing の

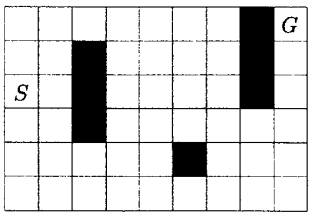

図 3: Sutton の Dyna 迷路

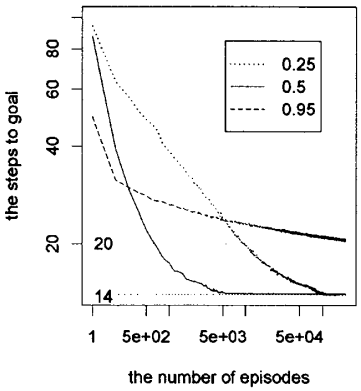

図 4: Dyna 迷路タスクにおける学習曲線

学習速度を比較する．挙動方策はいずれもルーレット選択を用い，行動の重みの初期値はそれぞれ 0.0001，0.01，0.1 とした．これらの初期値は予備実験においてよい結果を示すものを選んだ．

各強化関数の学習性能はルーレット選択による開始状態 (S) から目標状態 (G) までのステップ数により評価する．20 万エピソードについて 100 エピソードごとのステップ数の平均値を記録する実験を乱数の種を変えて 30 回繰り返し，その平均を実験値とする．

実験結果を図 4 に示す．縦軸は各エピソードにおけるルーレット選択によるステップ数，横軸はエピソード数である．縦軸，横軸ともに対数で表示している．学習に必要な時間は有効な方策を得るまでの各エピソードにおけるステップ数の和に比例するため，図 4 ではグラフの面積が学習速度を表す．

図 4 から，学習初期では公比 0.95 の等比減少関数の学習速度が速いが，すぐに公比 0.5 の関数が逆転することがわかる．また公比 0.95 の関数は 20 万エピソードの学習後にも 20 ステップの方策しか獲得していない．公比 0.25 と提案する合理性条件にもとづく公比 0.5 の等比減少関数はどちらも 14 ステップの最適方策を獲得するが，公比 0.25 の関数は最適方策を獲得するまでに公比 0.5 の関数の約 10 倍のエピソードを必要とする．

これらより本論文で提案するルーレット選択の合理性条件は Profit Sharing において高速で安定した学習を行うための有効な条件であるといえる．

なお，定理 1 に示した方策オフ型の合理性条件に関しては図 3 の迷路を縦横ともに 40 倍した迷路において最短ステップ数 521 の 2.5 倍程度の方策を獲得できることを確認している．

5 おわりに

本論文では Profit Sharing において (1) 方策オン型学習の合理性を提案し，定理にもとづく強化関数を用いれば学習が安定して高速化することを示した．(2) 報酬プラン獲得定理が方策オン型と方策オフ型の両方の合理性を満たすことを示し，報酬プラン獲得定理により学習が加速する理由を明らかにした．(3) 方策オフ型学習の合理性を満たす減少関数以外の強化関数の条件を導出し目標までのステップ数の多い問題への PS の適用を可能とした．

参考文献

1) John J. Grefenstette: Credit Assignment in Rule Discovery Systems Based on Genertic Algorithms, Machine Learning 3, 225/245, (1988)
2) 宮崎 和光, 山村 雅幸, 小林 重信: 強化学習における報酬割当ての理論的考察, 人工知能学会誌, vol.9, no.4, 580/587, (1994)
3) 松井 藤五郎, 自律型エージェントの行動学習に関する研究, 名古屋工業大学学位審査論文, (2003)
4) 植村 渉, 辰巳 昭治: Profit Sharing における強化関数に関する一考察, 人工知能学会論文誌, vol.19, 197/203, (2004)
5) 河合 宏和, 上野 敦志, 辰巳 昭治: ルーレット選択を用いた Profit Sharing 強化学習における合理性についての一考察, Proc. of The 19th Annual Conference of the Japanese Society for Artifical Intelligence, 1D3-03, (2005)
6) Richard S. Sutton, Andrew G. Barto, Reinforcement Learning: An Introduction, The MIT Press, (1998)

付録 A　定理 2 の導出

行動の重み更新を $W(s,a) \leftarrow W(s,a)+\Delta W(s,a)$ と表記するとき，ルーレット選択による行動の選択確率の増分 $\Delta P(s_t,a)$ は

$$\Delta P(s_t,a)$$
$$= \frac{W(s_t,a)+\Delta W(s_t,a)}{\sum_{a'\in A(s_t)}(W(s_t,a')+\Delta W(s_t,a'))} - \frac{W(s_t,a)}{\sum_{a'\in A(s_t)} W(s_t,a')}$$
$$= \frac{\Delta W(s_t,a)\sum_{a'} W(s_t,a') - W(s_t,a)\sum_{a'}\Delta W(s_t,a')}{\sum_{a'}(W(s_t,a')+\Delta W(s_t,a'))\sum_{a'} W(s_t,a')}$$

となる．分母は行動 a に関して定数となるので以降では分子のみを記述する．行動の重み更新により有効ルール a_e の選択確率が増加する条件は

$$\sum_{a'\in A(s_t),a'\neq a_e} \Delta P(s_t,a') - \Delta P(s_t,a_e) < 0 \qquad (7)$$

式 (7) の左辺を展開する．ここで状態 s_t で選択可能な行動の重みの和を $\sum_{a'} W$，増分の和を $\sum_{a'}\Delta W$，有効ルール a_e の重みを W_e，増分を ΔW_e，有効ルール以外の行動の重みの和を $\sum_{a'\neq a_e} W$，増分の和を $\sum_{a'\neq a_e}\Delta W$ と略記する．

$$\sum_{a'\neq a_e}(\Delta W\sum_{a'} W - W\sum_{a'}\Delta W) - (\Delta W_e\sum_{a'} W - W_e\sum_{a'}\Delta W)$$
$$= (\sum_{a'\neq a_e}\Delta W - \Delta W_e)\sum_{a'} W - (\sum_{a\neq a_e} W - W_e)\sum_{a'}\Delta W$$

$\sum_{a'} W = W_e + \sum_{a'\neq a_e} W$, $\sum_{a'}\Delta W = \Delta W_e + \sum_{a'\neq a_e}\Delta W$ を代入．

$$= 2(W_e\sum_{a'\neq a_e}\Delta W - \Delta W_e\sum_{a'\neq a_e} W)$$

よって

$$k_t\sum_{a'\neq a_e}\Delta W < \Delta W_e,\quad k_t = \frac{W_e}{\sum_{a'\neq a_e} W} \qquad (8)$$

のとき式 (7) を満たす．有効ルール a_e は状態 s_t において最後に選択された行動より

$$\Delta W_e = f(t,r_T,T),\quad \sum_{a'\neq a_e}\Delta W = \sum_{i=0}^{t-1} f(i,r_T,T)$$

を式 (8) へ代入する．　(導出終り) □