Profit Sharingにおける確率的状態遷移下の学習に関する一考察 ～ビギナーズラックの解明～

*植村 渉

龍谷大学 理工学部 電子情報学科

〒 520-2151 滋賀県大津市瀬田大江町横谷 9-5

wataru@rins.ryukoku.ac.jp

http://friede.elec.ryukoku.ac.jp/~wataru

Abstract: 強化学習にはその強化方法によって，環境同定型と経験強化型の二つに分けることができる．本研究では，その二つの強化方法の間に存在する相違点として，価値の扱い方が異なる点について指摘する．環境同定型の強化学習では，最適解を学習するために，価値を理論的な値へと収束させる．この値は，全ての事象を見渡した視点から導かれる値である．それに対して経験強化型の強化学習では，経験したルールから，価値を推定する．そのため，経験していないルールの価値は考慮しない．これらの相違は，期待値計算において，理論的確率を用いるか経験的確率を用いるかの違いである．経験強化型強化学習における価値の扱い方は，賭け事において初心者が勝つ可能性が高いビギナーズラックと類していることを報告する．まずビギナーズラックについて例を挙げて解説し，なぜ初心者は賭け事に勝ち易いと言われているのかを，経験的確率と理論的確率の二つの確率を挙げて解明する．そして，その二つの確率の相違が，環境同定型と経験強化型の強化学習法の価値の扱い方に大きく違いを与える原因であることを明らかにする．

1. はじめに

強化学習の枠組みでは，環境のモデルとしてマルコフ性を持つ環境を仮定することが多い．この場合，学習者であるエージェントが，現在の状態において行動を選択したとき，次状態への遷移が確率的な場合が存在する．強化学習法は試行錯誤によって，各ルールの価値を決定し，そのルールの価値によって行動選択を行う．確率的な状態遷移に対しては，何度も繰り返し経験をつむことで，ルールの価値を推定する．環境同定型の強化学習では，ルールの価値の理想値を，目標までのステップ数に応じた割り引き期待報酬値としている．そのため，確率的な状態遷移においても，ルールの価値は適切な期待値に収束する．それに対して，経験強化型の強化学習である Profit Sharing[Grefenstette 88] では，過去に獲得した報酬を累積してルールの価値を決定するため，確率的な状態遷移が存在する環境では，価値の扱いが不適切になる場合がある．本研究では，この価値の扱いが不適切になる原因を明らかにする．またこの問題は，賭け事において初心者が勝つ可能性が高いビギナーズラックの存在と類していることを報告する．まず，ビギナーズラックについて解析を行い，その後，強化学習における問題へ議論を進める．

以下，2 章では，ビギナーズラックについて定義を行い，なぜ期待値に反して初心者は勝ち易いと言われるのかを，経験的確率と理論的確率の二つの確率を用いて明らかにする．3 章では，経験的確率と理論的確率から生じる確率の違いを基に，経験強化型強化学習が陥る確率的状態遷移下における学習について検討する．そして，4 章でまとめとする．

2. ビギナーズラック

日本における代表的な賭け事として，宝くじ協議会の行う宝くじや，競馬などの公営ギャンブル，そして，直接現金のやり取りは行わないがパチンコ・スロットが挙げられる[1]．これら賭け事において，期待値

[1]勝者投票券（競輪）・勝舟投票券（競艇）・勝車投票券（オートレース）は学生は購入できない．勝馬投票権（競馬）は 2005 年から年齢制限のみとなった．

に反して初心者が勝つ可能性が高い**ビギナーズラック**と呼ばれる現象がある．広辞苑第五版によると，ビギナーズラックとは，以下のように定義されている．

定義 1 (ビギナーズラック)
賭事などで，初心者が往々にして好結果をおさめること．

一般に，統計学では，

$$\text{期待値} = \text{その事象が起こりうる確率} \times \text{その事象が起こったときに得られる利得} \tag{1}$$

として期待値を扱う．賭け事の場合は，掛け金に対して戻ってくると期待できる金額を意味する．

ここでは，この期待値に対して，二つの疑問を検討する．一つ目は，損をするとわかっている期待値の賭け事を，なぜ行うのかという疑問である．二つ目は，初心者が期待値以上に勝つことができるビギナーズラックがなぜ起こるのかという疑問である．

2.1. 損をする期待値の賭け事

賭け事を商売として続けるためには，当たり前であるが購入者の期待値は購入額より少ない．続ければ続けるほど，損をするはずである．ではなぜ，損をするとわかっている期待値の賭け事を，人は続けるのであろうか．競馬やパチンコでは，プレイヤの実力によって勝敗が決まる部分があるため，たとえ損をする期待値であっても,「自分は得をする」と信じる部分があるかもしれない．しかし，宝くじは純粋に確率で当たり外れが決まる[2]．それでも，人は宝くじを買い続けている．

「夢を買うから」と答える人がいるかもしれない．確かに，期待値以上に，夢が大きければ購入する価値があるだろう．しかし，くじを買ったときの夢とは，期待値そのものである．事象が起こりうる確率が小さすぎるため，適切な値が設定できず，本来の期待値以上の価値を見出していると考えられる．例えば，一等が一億円の宝くじが当たったと想像するとき，このくじの期待値は 500 円だからといって 500 円の使い道を検討する人はいない．当たる確率が 0 に近いと言うことは無視して，一億円の使い道を検討する．この場合，期待値の計算は，事象が起こる前提で計算していると考えられる．

損をする期待値への賭け事は，事象が起こった場合のみを想定して計算するため，本来の期待値よりも大きな値を期待するため生じると考えられる．

2.2. ビギナーズラックと期待値

なぜ初心者は好成績を収めることができるのだろうか．無欲だからだろうか．深く考えないからであろうか．ここでは，期待値における確率事象の扱い方に着目して解明する．

例えば，1/100 の確率で 8,000 円が当たる一枚 100 円の宝くじを考える．このくじの期待値は 80 円であるため，明らかに損をする．この宝くじを購入し，外れ続けた場合，80 回目で当たりが出れば収支は 0 である．それまでの 79 回のどこかで当たりを引けば収支は正になる．80 回目まで報酬を得られない確率は約 $45\%(= (1 - 1/100)^{80})$ である．つまり，100 回に 1 回当たる確率ではあるが，1/2 以上の確率で収支を正にすることができる．ただし 80 回以上はずれを引いた場合，コストは膨らみ続ける．80 回目以降も外れ続けた場合を考慮すると，全ての事象の期待値になり損をするが，収支が正になる可能性のみを考えると，かなりの確率で収支を正にすることができる．このことより，初心者の段階では期待値に反して収支を正にすることができ，それに味を占めて買い続けると，本来の期待値に近づくことがわかる．これが，ビギナーズラックの正体であろう．

余談ではあるが，常に負けないギャンブルの方法として，前回の負けたときの掛け金の倍額をかけ続ければ，勝った場合は必ず報酬が正になる手法がある．これも同様に，期待値のしわ寄せを負けた場合に詰め込むため，負けた場合の負担額が相当なものになる．

[2] パチプロという言葉は存在するが，プロの宝くじ師は聞いたことがない

2.3. 大数の法則

確率論・統計論の極限定理の一つとして大数の法則がある．これは，理論的確率と経験的確率が一致するという法則である．理論的確率とは，全ての事象に対するある事象の割合のことで，いわゆる「**確率**」である．経験的確率とは，何千回，何万回と試行を繰り返すことで，標本的に確率を求めることであるが，根本にあるのは「**平均**」である．大数の法則の下では，理論的な確率と経験的な平均値を混同しても問題は生じない．しかし，先の例では，一回の当たりを引くまでの統計でしかないため，理論的な確率と経験的な平均値が一致しない．そのため，統計的に意味を持つ回数を経験していないため，初心者は期待値以上に当たりを引きやすく，繰り返して経験者になるにつれ，期待値に近づくため当たりを引きにくくなる．

3. 確率的状態遷移下の強化学習

強化学習の確率的状態遷移下における価値の扱いについて検討する．環境同定型の強化学習と，経験強化型強化学習における価値の扱いの違いを，理論的確率と経験的確率を用いて明らかにする．

3.1. 環境同定型の強化学習

環境同定型の強化学習では，ルールの価値の理想値を，目標までのステップ数に応じた割り引き期待報酬値としている．例えば，Q-Learning[Watkins 92] では，時刻 t で経験したルール (s, a) の価値 $Q(s, a)$ を以下の式に従い更新する．

$$Q(s,a) \leftarrow (1-\alpha)Q(s,a) + \alpha\left[r + \gamma \max_{a'} Q(s',a')\right] \tag{2}$$

ここで，α は学習率を表し，r は報酬，γ は割引率，そして，$\max_{a'} Q(s', a')$ は次状態 s' における最大の Q 値を持つルール (s', a') の価値である．確率的な状態遷移の場合を考える．10 回に 1 回は 100 の報酬を得られ，残り 9 回は 10 の報酬を得られるルールの場合，学習率 α を適切に 0 に近づけることで，Q 値は 19 に近づく．この値は，理論的な確率による期待値 19（$= 100 \times 0.1 + 10 \times 0.9$）に等しい．そのため，Q-Learning による学習は，最適解を獲得できる．

3.2. 経験強化型の強化学習

経験強化型の強化学習である Profit Sharing では，報酬を獲得した時に，それまでに経験した行動系列を，ルール価値を以下の式に従い更新する．

$$Q(s_x, a_x) \leftarrow Q(s_x, a_x) + r \times f(x) \tag{3}$$

ここで x は，報酬を獲得したルールを基点とした時系列と逆方向のインデックスである．$f(x)$ は，報酬を分配する関数で，一般に等比減少関数 [宮崎 94][植村 04] を用いる．

Profit Sharing は，問題環境がマルコフ性を持たない一部の場合にも学習を進められることが示されている [宮崎 99][植村 05]．そのため，実世界における学習手法として今後が期待されている．Profit Sharing の報酬分配には，次の二点の問題がある．1) 報酬獲得時にしか学習を行わないオフライン型の強化方法である．2) 割引率に相当する強化関数の公比が 0 に近いため，ステップ数あたりの報酬の価値を大きく扱う．

前者の問題により，行動選択時に確率的なルールを更新することができない．そのため，まれに大きな報酬を得られるルールを早いうちに経験すると，そのルールの価値を減少させることができず，経験に固執する [植村 06]．これは，経験的確率と同様，学習者にとって適切な期待値が扱えない場合である．

後者の問題により，遠回りで安定した報酬を獲得するルールよりも，一か八かで大きな報酬を得られるルールの方が価値が高くなる．これも，学習者にとって適切な期待値が扱えない場合である

いずれの場合も，0 に近い確率で大きな報酬を得られるルールを早いうちに経験すると，その経験に固執し，適切な価値に修正できない．

Profit Sharing の問題 1 (適切に価値を扱えない場合)
0 に近い確率で大きな報酬を得られるルールを早いうちに経験すると，その経験に固執し，適切な価値に修正できない．

3.3. Profit Sharing の実験における問題点

前述の通り，Profit Sharing では確率的状態遷移を持つ環境では，適切な期待値が扱えない問題がある．しかし，現在，提案されている Profit Sharing に基づいた強化学習は，学習環境として決定的な状態遷移に限定していない．これは，性能評価実験において，試行回数を増やすことにより，経験的確率と理論的確率とが交えた結果となっているため，確率的状態遷移においてもうまく学習しているように見えていると考えられる．例えば実験において 10,000 回のステップによる学習を 100 回行ったとする．このとき，それぞれの確率による統計は，a)1 回の学習における 10,000 回のステップによる統計と，b)100 回学習を行うことで分散を抑える統計である．Profit Sharing では，0 に近い事象で大きな報酬を得られるルールを早いうちに経験すると，その経験に固執するため，a) による統計においては学習が適切に修復されない．経験の固執は，b) のおかげであたかも改善されたように見える．

4. おわりに

経験強化型強化学習である Profit Sharing 法は，環境のモデルにマルコフ性を仮定しない環境でも学習を適切に進めることができるため，現実社会への適応など今後が期待されている．しかし，研究を進めるにつれ，確率的状態遷移下での学習性能が，環境同定型強化学習に対して劣る場合が見つかった．本研究では，Profit Sharing における強化が，理論的な確率に基づくものではなく，経験的な確率に基づくものであり，報酬獲得時のみの強化を行うため期待値が適切でなくなる場合があることを示した．また，このことは，経験から強化を行う人間でも同様であり，ビギナーズラックという言葉が生まれる原因であることを示した．

Profit Sharing の学習性能を，理論的確率の影響と経験的確率の影響を分けて表現する必要があり，そのための実験方法や性能評価方法が必要とされる．今後の課題は，そのような方法を考えることである．

参考文献

[Grefenstette 88] Grefenstette, J.J., “Credit Assignment in Rule Discovery Systems Based on Genetic Algorithms”, Machine Learning, Vol.3, pp.225–245 (1988).

[Watkins 92] Watkins, C.J.C.H. and Dayan, P., “Technical Note:Q-Learning”, Machine Learning, Vol.8, pp.279–292 (1992).

[植村 04] 植村 渉, 辰巳 昭治, “Profit Sharing 法における強化関数に関する一考察”, 人工知能論文誌, Vol.19, No.4, pp.197–203 (2004).

[植村 05] 植村 渉, 上野 敦志, 辰巳 昭治, “POMDPs 環境のためのエピソード強化型強化学習法”, 電子情報通信学会論文誌, Vol. J88-A, pp. 761–774, (2005).

[植村 06] 植村 渉, 上野 敦志, 辰巳 昭治, “経験に固執しない Profit Sharing 法”, 人工知能論文誌, Vol. 21, pp. 81–93, (2006).

[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ての理論的考察”，人工知能誌, Vol.9, No.4, pp.580–587 (1994).

[宮崎 99] 宮崎 和光，荒井 幸代，小林 重信, “POMDPs 環境下での決定的政策の学習”, 人工知能誌, Vol.14, No.1, pp.148–156 (1999).