複利型強化学習における投資比率最適化手法の検討

松井藤五郎 [1*] 後藤卓 [2] 和泉潔 [3,4]

[1] 中部大学 [2] 三菱東京 UFJ 銀行 [3] 東京大学 [4] JST さきがけ

Abstract: 本論文では，複利型強化学習において投資比率を最適化する方法について検討する．これまでに提案した方法ではオンライン勾配法を用いて投資比率を最適化するが，強化学習によって有効な行動規則を学習できない場合には投資比率が 0 に収束してしまう場合がある．そこで，本論文では，複利型共学習における投資比率の最適化方法について検討し，新たな方法を提案する．

1 はじめに

複利型強化学習 [6] は，通常の強化学習 [3] と同様に，試行錯誤を通じて行動規則を学習する枠組みである．通常の強化学習が将来にわたって得られる報酬の和を最大化するのに対し，複利型強化学習では将来にわたって得られる利益率の複利効果を最大化する．これまでに，複利型強化学習の基本的な枠組みと従来の Q 学習を拡張した複利型 Q 学習のアルゴリズムが提案され，国債銘柄選択問題や国債取引問題での有効性が示されている [2, 7].

複利型強化学習では，エージェントが自分の資産のうちのどれだけを投資するかを表す投資比率パラメーター f が導入されている．このパラメーターは複利効果に大きく影響し，この値によって複利効果が最大となる行動が異なる場合がある．

投資比率 f に関しては，リターンの確率分布が既知であるなら，複利リターンを最大化する投資比率を解析的に求められることが明らかとなっている [1]．この既知のリターン分布の下で複利リターンを最大化する投資比率は，ケリー基準と呼ばれる．しかしながら，一般的には，投資に対するリターンの確率分布は未知であり，真のケリー基準を事前に求めることはできない．

投資比率に関しては，Vince が，optimal f と呼ばれる過去のリターンから良い投資比率を推定する手法が提案している [4]．optimal f は，リターンの確率分布が未知の場合でも良い投資比率を得ることができるが，optimal f によって得られる投資比率はケリー基準と同じではなく [5]，複利リターンを最大化することができない．

我々は，これまでに，オンライン勾配法を用いて投資比率 f を最適化し，複利型強化学習における複利効果を最大化する方法を提案した [8, 9]．ところが，オンライン勾配法を用いて投資比率を最適化すると，複利型強化学習によって適切な行動規則を学習できないときに投資比率が 0 に収束してしまうことがある．

そこで，本論文では，複利型共学習における投資比率の最適化方法について検討し，新たな方法を提案する．

2 複利型強化学習と投資比率の最適化

複利型強化学習は，割引複利リターン

$$\begin{aligned}&(1+R_{t+1}f)(1+R_{t+2}f)^{\gamma}(1+R_{t+3}f)^{\gamma^2}\dots\\ =&\prod_{k=0}^{\infty}(1+R_{t+k+1}f)^{\gamma^k}\end{aligned}$$

の期待値を最大化するような行動規則を学習する．ここで，R_t は時刻 t に観測されたリターン，γ は割引率パラメーター，f は投資比率パラメーターを表す．割引複利リターンは，対数を取ることで，従来の強化学習と同じように再帰的な形で表すことができる．すなわち，行動規則 π の下での状態 s の価値 $V^{\pi}(s)$ と行動規則 π の下での状態 s における行動 a の価値 $Q^{\pi}(s,a)$ は次のように表される．

$$V^{\pi}(s) = \mathrm{E}_{\pi}\left[\log\prod_{k=0}^{\infty}(1+R_{t+k+1}f)^{\gamma^k}\,\middle|\, s_t=s\right] \quad (1)$$

$$Q^{\pi}(s,a) = \mathrm{E}_{\pi}\left[\log\prod_{k=0}^{\infty}(1+R_{t+k+1}f)^{\gamma^k}\,\middle|\, s_t=s, a_t=a\right] \quad (2)$$

ここで，$\pi(s,a)$ は行動規則 π の下で状態 s において行動 a が選択される確率（行動選択確率），$\mathcal{P}^{a}_{ss'}$ は状態 s において行動 a を行ったときに次の状態が s' になる確

* 連絡先: TohgorohMatsui@tohgoroh.jp, http://とうごろう.jp

率（状態遷移確率），$R^a_{ss'}$ は状態 s において行動 a を行って次の状態が s' になったときに得られるリターンの期待値を表す．複利型強化学習では，すべての s, a に対してこの $Q^\pi(s,a)$ を最大化するような行動規則 π を学習する．

複利型 Q 学習は，従来の Q 学習の報酬 r_{t+1} を投資比率 f のときのグロス・リターンの対数 $\log(1+R_{t+1}f)$ に置き換えたものである．ただし，投資比率は，状態行動対ごとに投資比率を使い分ける場合と，すべての状態行動対に対して共通の投資比率を使う場合がある．状態行動対ごとに投資比率を使い分ける場合，時刻 t の状態 s_t において行動 a_t を実行し，次の時刻 $t+1$ にリターン R_{t+1} を受け取ると，状態行動対 s_t, a_t に対する Q 値を次のように更新する．

$$Q_{t+1}(s_t,a_t) = Q_t(s_t,a_t) + \alpha\Delta_t \tag{3}$$

$$\Delta_t = \log(1+R_{t+1}f(s_t,a_t)) + \gamma \max_{a\in\mathcal{A}} Q_t(s_{t+1},a) - Q_t(s_t,a_t) \tag{4}$$

ここで，α はステップ・サイズ，γ は割引率，$f(s,a)$ は状態行動対 s,a に対する投資比率を表す．

状態行動対ごとに異なる投資比率を使い分け，オンライン勾配法を用いて投資比率を最適化する場合は，強化学習の各ステップ t において投資比率 f を次のように更新する．

$$f_{t+1}(s_t,a_t) = f_t(s_t,a_t) + \eta\frac{R_{t+1}}{1+R_{t+1}f_t(s_t,a_t)} \tag{5}$$

ここで，η はオンライン勾配法の学習率を表す．

3 投資比率ゼロ収束問題とその対策

オンライン勾配法では，エージェントの行動履歴に基づいて，複利効果を最大化する投資比率 f を求める．しかしながら，エージェントの行動が不適切な場合には，正の複利効果を得ることができず，学習の途中で投資比率が 0 となることがある．

投資比率が 0 になると，どのようなリターンが発生してもグロス・リターンの対数が常に $\log(1+R_{t+1}f) = \log 1$ になる．このようなとき，どのような行動をとってもフィードバックとして得られるグロス・リターンの対数が同じであるため，複利型強化学習で適切な行動規則を学習することができなくなってしまう．その後，強化学習の探査行動によって偶然高いリターンが連続して得られ，その結果として，オンライン勾配法によって投資比率が増加する可能性もあるが，適切な行動規則を学習できていないため，ほとんどの場合は再び投資比率は 0 に収束してしまう．この問題を本論文では投資比率最適化におけるゼロ収束問題と呼ぶ．

そこで本論文では，投資比率最適化におけるゼロ収束問題を解決するために，投資比率がゼロになった場合に学習をリセットする仕組みを導入することを提案する．すなわち，状態行動対 s, a に対する投資比率 $f(s,a)$ を更新し，$f(s,a) \le 0$ になった場合には，状態 s における全ての状態行動対 s, a' について投資比率 $f(s,a')$ と行動価値 $Q(s,a')$ を初期化する．このアルゴリズムを Algorithm 1 に示す．

投資比率 f の更新式はオンライン勾配法と同じであるが，$f(s,a) = 0$ となったとき，s における投資比率 f と行動価値 Q を初期化する．

4 実験

4.1 ブラックジャック問題

カジノ・ゲームの一つであるブラックジャック問題を用いて実験を行った．ブラックジャックは，プレイヤーとディーラーがトランプのカードを用いた 1 対 1 の勝負を行う．21 点を越えない範囲で相手よりも大きい得点を獲得すると勝ちとなる．絵札（J, Q, K）は 10 点と数える．A は 11 点と数えるが，11 点と数えたときに得点が 21 点を超える場合には 1 点とする．

プレイヤーには 2 枚のカードが表向きに配られ，ディーラーには 2 枚のカードが 1 枚が表向き，もう 1 枚が裏向きに配られる．プレイヤーは，もう一枚のカードを引く（ヒット）か，それ以上のカードを引かずに勝負をする（スタンド）かを選択する．プレイヤーは 21 点を超えない範囲では何度でもヒットを選択できるが，21 点を超えた時点でプレイヤーの負けとなる．ディーラーは 17 点未満では自動的にヒットを選択し，17 点以上では自動的にスタンドを選択する．

21 点の状態をブラックジャックと呼ぶ．最初に配られた 2 枚のカードで 21 点の状態をナチュラル 21 と呼ぶ．プレイヤーが勝つとベットした金額の 2 倍の払い戻しを受ける．すなわち，このときのリターンは 1 である．ただし，プレイヤーがナチュラル 21 で勝つとベットした金額の 2.5 倍の払い戻しを受け，このときのリターンは 1.5 である．プレイヤーが負けるとベットした金額はすべて没収され，このときのリターンは -1 である．

プレイヤーは，カードが配られる前に，次のゲームにベットする金額を決めなければならない．また，ゲーム中にベット金額を変更することはできない．したがって，この問題では，全ての状態行動対において同じ投資比率が用いられる．

Algorithm 1 提案アルゴリズム

入力: 割引率 γ, ステップ・サイズ α, 投資比率 f, 投資比率学習率 η
$Q(s,a)$ を任意に初期化
$f(s,a)$ を $0 \leq f(s,a) < 1$ の範囲で任意に初期化
loop (各エピソードに対して繰り返し)
  s を初期化
  repeat (エピソードの各ステップに対して繰り返し)
    Q から導かれる行動規則（行動選択確率）に従って s での行動 a を選択
    行動 a を実行し，リターン R と次の状態 s' を観測
    $f(s,a) \leftarrow f(s,a) + \eta \frac{R}{1+Rf(s,a)}$
    $Q(s,a) \leftarrow Q(s,a) + \alpha \left(\log(1+Rf(s,a)) + \gamma \max_{a'} Q(s',a') - Q(s,a)\right)$
    if $f(s,a) \leq 0$ **then**
      for all a' **do**
        $Q(s,a')$ を任意に初期化
        $f(s,a')$ を $0 \leq f(s,a') < 1$ の範囲で任意に初期化
      end for
    end if
    $s \leftarrow s'$
  until s が終端状態ならば繰り返しを終了
end loop

通常のカジノでは，6 デッキ[*1]以上のカードを使ってゲームが行われるが，簡単化のため，ここではデッキ数を無限大とした．すなわち，カードのマークは無視してそれぞれのカードが 1/13 の確率で出現する．

図 1 に，ブラックジャック問題におけるゲーム進行の例を示す．最初にプレイヤーに配られた 2 枚のカードによる得点は 9，ディーラーには 9 ともう一枚のカードが配られている．プレイヤーがヒットを選択したところ，10 のカードが配られた．プレイヤーの得点は 19 点となり，ここでプレイヤーはスタンドを選択した．その後，ディーラーが裏向きのカードを開いたところ，このカードは 6 だった．このときのディーラーの得点は 15 点で 17 点未満であるため，ディーラーは自動的にヒットし，5 のカードが配られた．これによってディーラーの得点が 20 点となり，17 点以上になったため，ディーラーは自動的にスタンドした．最終的に，19 点対 20 点でプレイヤーの負けとなり，このゲームのリターンは -1 だった．

4.2 実験方法

この問題を用いて，従来のオンライン勾配法を用いて投資比率を最適化する複利型 Q 学習と提案手法である再初期化付きオンライン勾配法を用いて投資比率を最適化する複利型 Q 学習を比較した．割引率 γ は 0.9 とし，投資比率の初期値は $f = 0.5$ とした．提案手法において投資比率を再び初期化する場合も，$f = 0.5$ とした．ステップ・サイズは $\alpha = 0.01$，オンライン勾配法における学習率は $\eta = 0.001$ とした．学習時の行動選択には $\epsilon = 0.1$ の ϵ-グリーディ選択を用いた．

強化学習は，10^8 ステップの学習をランダム・シードを変えて 100 回行い，最も価値が高いと学習した行動を選択し続けた場合の幾何平均リターンを求めた．幾何平均リターンは複利効果を評価するための指標である．n 期のリターンが $R_1, R_2, \ldots, R_n$ のときの幾何平均リターンは

$$G = \left(\prod_{i=1}^{n}(1+R_i)\right)^{\frac{1}{n}} - 1 \tag{6}$$

と表される．たとえば，幾何平均リターンが 0.05 のとき，毎期 0.05 のリターンによって，資産が 5% ずつ増えることが期待できる．

4.3 結果

まずはじめに，オンライン勾配法を用いたとき（従来手法）と提案手法を用いたときの投資比率 f の推移を図 2, 3 に示す． オンライン勾配法を用いた従来手法に

[*1] 1 デッキは 1 組 52 枚.

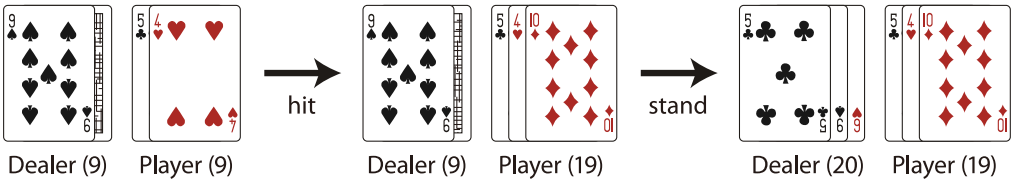

図 1　ブラックジャック問題におけるゲームの例

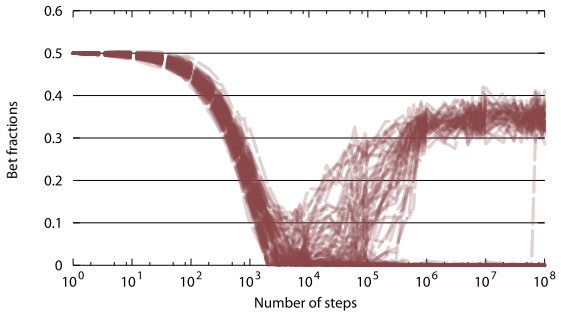

図 2　オンライン勾配法（従来手法）による投資比率最適化を行ったときの投資比率 f の推移

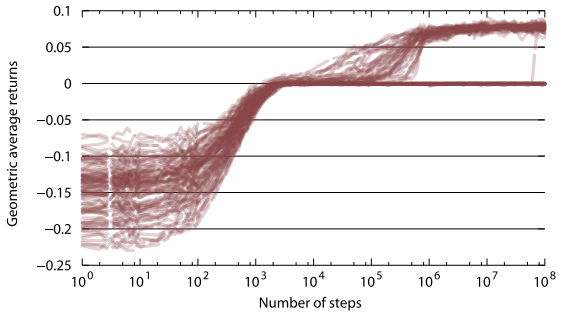

図 4　オンライン勾配法による投資比率最適化を行ったときの幾何平均リターンの推移

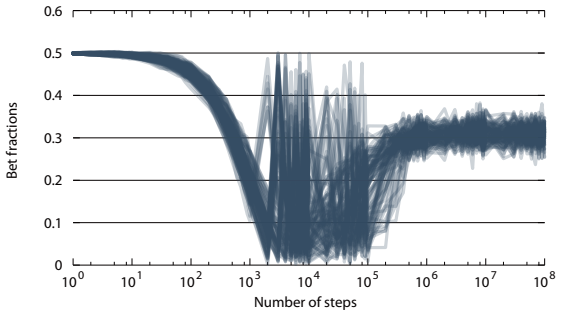

図 3　提案手法による投資比率最適化を行ったときの投資比率 f の推移

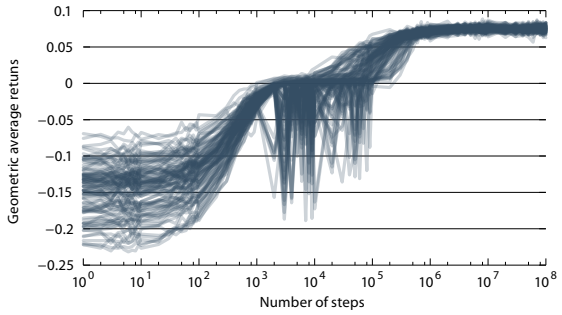

図 5　提案手法による投資比率最適化を行ったときの幾何平均リターンの推移

よって投資比率を最適化したときは，約半数のランダム・シードにおいて投資比率が 0 に収束してしまった．これに対し，提案手法によって投資比率を最適化したときは，投資比率が 0 に収束することはなかった．グラフから，投資比率がリセットされているケースが 10^3 ステップから 10^5 ステップの間に存在することがわかる．

次に，オンライン勾配法を用いたとき（従来手法）と提案手法を用いたときの幾何平均リターンの推移を図 4, 5 に示す．従来手法を用いた場合は，約半数のケースにおいて投資比率が 0 に収束したことによって幾何平均リターンも 0 に収束した．提案手法を用いた場合は，投資比率が 0 になって Q 値がリセットされたときに一旦幾何平均リターンが悪くなるが，その後，正の幾何平均リターンに収束した．

100 回ずつランダム・シードを変えて行った実験の幾何平均リターンの平均を図 6 に示す．従来手法は約半数のランダム・シードにおいて幾何平均リターンが 0 に収束してしまっているため，従来手法における幾何平均リターンの平均は提案手法に比べて約半分となった．

5　まとめ

本論文では，複利型強化学習において，投資比率をオンライン勾配法を用いて最適化すると投資比率が 0 に収束してしまうことがある問題（投資比率ゼロ収束問題）に対し，学習を初期化することによってこれを解決する方法を提案した．実際に，提案手法をブラックジャック問題に適用し，提案手法が有効に働くことを確認した．

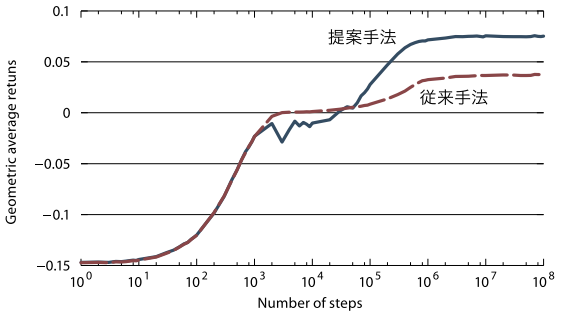

図 6　幾何平均リターンの平均

提案手法を用いることによって，適切な行動が学習できないために投資比率が 0 に収束してしまう問題を回避することはできる．しかしながら，適切な行動が存在せず投資比率を 0 とすることが最適なタスクにおいても，投資比率が 0 になると学習が初期化されてしまい，学習が収束しない．したがって，今後，投資比率を 0 とすることが最適な場合には投資比率が 0 に収束する手法を考案する必要がある．

謝辞

本研究は科研費（23700182）の助成を受けたものである．

留意事項

本論文は三菱東京 UFJ 銀行の公式見解を表すものではありません．

参考文献

[1] John Larry Kelly, Jr. A new interpretation of information rate. *Bell System Technical Journal*, Vol. 35, pp. 917–926, 1956.

[2] Tohgoroh Matsui, Takashi Goto, Kiyoshi Izumi, and Yu Chen. Compound reinforcement learning: Theory and an application to finance. In Scott Sanner and Marcus Hutter, editors, *Recent Advanced in Reinforcement Learning: Revised and Selected Papers of the European Workshop on Reinforcement Learning 9 (EWRL 2011)*, Vol. 7188 of *Lecture Notes in Computer Science*, pp. 321–332, 2012.

[3] Richard S. Sutton and Andrew G. Barto. *Reinforcement Learning: An Introduction*. The MIT Press, 1998. 三上貞芳, 皆川雅章 共訳. 強化学習. 森北出版, 2000.

[4] Ralph Vince. Find your optimal f. *Technical Analysis of Stock & Commodities*, Vol. 8, No. 12, pp. 476–477, 1990.

[5] Ralph Vince. Optimal f and the Kelly criterion. *IFTA Journal*, pp. 21–28, 2011.

[6] 松井藤五郎. 複利型強化学習. 人工知能学会論文誌, Vol. 26, No. 2, pp. 330–334, 2011.

[7] 松井藤五郎, 後藤卓, 和泉潔, 陳ユ. 複利型強化学習の枠組みと応用. 情報処理学会論文誌, Vol. 52, No. 12, pp. 3300–3308, 2011.

[8] 松井藤五郎, 後藤卓, 和泉潔, 陳ユ. オンライン勾配法による投資比率最適化付き複利型強化学習. 第 8 回ファイナンスにおける人工知能応用研究会 (SIG-FIN), pp. 42–45, 2012.

[9] 松井藤五郎, 後藤卓, 和泉潔, 陳ユ. 複利型強化学習における投資比率の最適化. 第 26 回人工知能学会全国大会 (JSAI 2012), pp. 4E1–OS–15–5, 2012.