複利リターンに基づく強化学習による取引戦略の学習

松井藤五郎 [1*] 後藤卓 [2] 和泉潔 [3]

[1] とうごろう機械学習研究所 [2] 三菱東京 UFJ 銀行 [3] 産業技術総合研究所

Abstract: 本論文では，複利リターンに基づく強化学習を用いて日本国債の取引戦略を獲得した結果を報告する．残存期間 10 年の日本国債を対象として，複利リターンに基づく Q 学習と複利リターンに基づく OnPS を用いて 5 年間の金利データから取引戦略を獲得し，その後の 1 年間のデータを用いて評価した．

1 はじめに

強化学習 [6] は，エージェントが獲得する報酬を将来にわたって最大化する行動規則を試行錯誤と通じて学習する枠組みである．

これまでに，強化学習を用いて金融市場における取引戦略を獲得する試みがいくつか行われてきた．Sherstov と Stone は，PXS (Penn Exchange Simulator) を用いた人工市場の中での取引戦略を学習する研究を行った [5]．O らは，強化学習を用いて銘柄と投資比率を決定する戦略を学習する研究を行っている [4]．また，Lee らは，マルチエージェント強化学習を用いてポートフォリオ・マネジメントを行う研究を行っている [1]．筆者らも，強化学習を用いて国債市場における取引戦略を獲得する研究を行ってきた．強化学習を取引戦略を獲得するためのシステム [10] を開発し，獲得された取引戦略の分析を行った [2, 8, 9]．これらの研究は，すべて従来の強化学習の枠組みに基づいて行われたものである．

従来の強化学習では，割引収益の期待値を最大化する行動規則を学習することを目的としている．割引収益とは，将来に得られる報酬を遠い将来のものほど割り引いて合計したものである．しかしながら，ファイナンスの分野では，報酬（すなわち利益）よりもリターンの方が重要視される．たとえば，銘柄 A の株を 1,000 円で購入して 1,100 円で売却するのと銘柄 B の株を 100 円で購入して 200 円で売却するのを比べた場合，利益はどちらも 100 円だが，リターンは前者が 0.1 で後者が 1.0 と大きく異なり，他の条件がすべて同じとすると前者よりも後者が好まれる．また，リターンは，平均リターンではなく複利リターンのほうが重要である．たとえば，3 期のリターンが $-0.5, 0.7, 0.1$ という銘柄 C と同じく $0.1, 0.1, 0.1$ という銘柄 D を比較すると，（算術）平均リターンはともに 0.1 であるが，複利リターンは銘柄 C が 0.935 であるのに対し銘柄 D は 1.331 であり，複利リターンの観点からは銘柄 D の方が好ましい．そこで，筆者は，ファイナンスの分野における取引戦略を獲得するための強化学習の枠組みとして，複利リターンに基づく強化学習を提案している [11]．

本論文では，この複利リターンに基づく強化学習を日本国債の取引戦略の獲得に応用する．まず，複利リターンに基づく強化学習の枠組みを説明し，次に複利リターンに基づく Q 学習と複利リターンに基づく OnPS について述べる．続いて，取引戦略を獲得する方法について述べ，実験結果を示す．

2 複利リターンに基づく強化学習

従来の強化学習 [6] では，割引収益

$$r_{t+1} + \gamma r_{t+2} + \gamma^2 r_{t+3} + \cdots = \sum_{k=0}^{\infty} \gamma^k r_{t+k+1} \qquad (1)$$

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 に獲得した報酬，γ は割引率パラメーターを表す．

これに対し，複利リターンに基づく強化学習 [11] では，割引複利リターン

$$(1 + R_{t+1} f)(1 + R_{t+2} f)^{\gamma} (1 + R_{t+3} f)^{\gamma^2} \cdots = \prod_{k=0}^{\infty} (1 + R_{t+k+1} f)^{\gamma^k}$$

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

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

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．

$$
\begin{aligned}
V^\pi(s) =& \mathrm{E}_\pi\left[\log\prod_{k=0}^{\infty}(1+R_{t+k+1}f)^{\gamma^k}\middle| s_t=s\right]\\
=& \sum_{a\in\mathcal{A}}\pi(s,a)\\
& \sum_{s'\in\mathcal{S}}\mathcal{P}^a_{ss'}\left(\log(1+\mathsf{R}^a_{ss'}f)+\gamma V^\pi(s')\right)\\
Q^\pi(s,a) =& \mathrm{E}_\pi\left[\log\prod_{k=0}^{\infty}(1+R_{t+k+1}f)^{\gamma^k}\middle| s_t=s, a_t=a\right]\\
=& \sum_{s'\in\mathcal{S}}\mathcal{P}^a_{ss'}\left(\log(1+\mathsf{R}^a_{ss'}f)+\gamma V^\pi(s')\right)
\end{aligned}
$$

と表すことができる．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' になる確率（状態遷移確率），$\mathsf{R}^a_{ss'}$ 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．複利リターンに基づく強化学習では，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．

3　複利リターンに基づく Q 学習

複利リターンに基づく強化学習における価値 $V^\pi(s)$, $Q^\pi(s,a)$ は，従来の強化学習において価値を表す式の中の報酬の期待値 $\mathcal{R}^a_{ss'}$ を投資比率 f のときのグロス・リターンの期待値の対数 $\log(1+\mathsf{R}^a_{ss'}f)$ に置き換えたものに等しい．複利リターンに基づく Q 学習 [11] は，この性質を利用して，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．複利リターンに基づく Q 学習のアルゴリズムを Algorithm 1 に示す．

Q 学習では，報酬が有界で，ステップ・サイズ・パラメーターが適切に設定されているとき，MDP において最適な行動規則を学習できることが証明されている [7]. 同様にして，複利リターンに基づく Q 学習でも，リターンが有界で[*1]，ステップ・サイズ・パラメーターが適切に設定されているとき，MDP において最適な行動規則を学習できることが証明できる [11].

4　複利リターンに基づく OnPS

複利リターンに基づく Q 学習と同様にして，本論文では，OnPS [3] を複利リターンに基づくものに拡張する．複利リターンに基づく profit sharing は，報酬 r_{t+1} の替わりに投資比率 f のときのグロス・リターンの対数 $\log(1+R_{t+1}f)$ を用いた profit sharing を用いる．このアルゴリズムを，Algorithm 2 に示す．従来の OnPS [3] と異なるのは，9 行目で報酬 r の替わりにリターン R を観測している点と，11 行目で報酬 r の替わりに投資比率 f のときのグロス・リターンの対数 $\log(1+Rf)$ を用いて行動優先度 P を更新している点の 2 点である．

profit sharing は，報酬を受け取ると，時間をさかのぼって強化信号を伝えていく．1 ステップさかのぼるごとに強化信号は γ 倍される．これは，複利リターンに基づく profit sharing でも同じである．複利リターンに基づく profit sharing では，R_{t+1} のリターンに対して $\log(1+R_{t+1}f)$ の強化信号が与えられ，これを 1 ステップさかのぼるごとに γ 倍ずつしながら伝えていく．リターン R_{t+1} の範囲を $[-1,\sup(R)]$ とすると，グロス・リターンの対数 $\log(1+R_{t+1})$ の範囲は $[-\infty,\log(1+\sup(R))]$ となる（$\sup(R)$ はリターン R の上界を表す）．グロス・リターンの対数をそのまま強化信号として用いると，リターンが $R_{t+1}=-1$ のときに強化信号が $\log(1+R_{t+1})=-\infty$ となってしまい，強化信号を割り引くことができずにエピソードに含まれるすべての状態行動対の優先度が $-\infty$ になってしまう．また，その後，正の強化信号を獲得しても優先度は $-\infty$ のままとなってしまう．これを防いでいるのが投資比率パラメーター f（$0<f<1$）である．$f<1$ の投資比率を用いることによって，リターンが $R_{t+1}=-1$ のときでも強化信号 $\log(1+R_{t+1}f)$ が $-\infty$ とならない．

5　強化学習を用いた取引戦略の獲得

ここでは，アルゴリズム以外は [2] と同じ方法を用いた．ただし，[2] では評価損益の増分を報酬としていたが，本論文ではリターンをそのままエージェントに渡す．以下に，本手法の概要を簡単に示す．本論文も [2] と同じ週次の国債取引を対象としているため，これに合わせて説明する．

エージェントが取り得る行動は，ショートあるいはロングのいずれかである．ショートは日本国債を売ることを表し，日本国債を信用売りしているときをショート・ポジションという．ロングは日本国債を買うことを表し，日本国債を保有しているときをロング・ポジションという．また，取引量は常に一定とする．

エージェントが置かれている状態，すなわち国債市場の状態は，金利水準とボリンジャー・バンドの幅という

[*1] 最小値が -1 であるため，上界だけあればいい．

Algorithm 1 複利リターンに基づく Q 学習

Input: discount rate γ, step size α, betting fraction f
Initialize $Q(s,a)$ arbitrariliy, for all s, a
loop (for each episode)
  Initialize s
  repeat (for each step of episode)
    Choose a from s using policy derived from Q (e.g., ϵ-greedy)
    Take action a, observe return R, next state s'
    $Q(s,a) \leftarrow Q(s,a) + \alpha\left[\log(1+Rf) + \gamma \max_{a'} Q(s',a') - Q(s,a)\right]$
    $s \leftarrow s'$
  until s is terminal
end loop

Algorithm 2 複利リターンに基づく OnPS

Input: discount rate γ, step size α, betting fraction f, initial preference c
Initialize $P(s,a) \leftarrow c$, for all s, a
loop (for each episode)
  Initialize s
  $c(s,a) \leftarrow 0$, for all s, a
  repeat (for each step of episode)
    Choose a from s using policy derived from P (e.g., ϵ-greedy)
    $c(s,a) \leftarrow c(s,a) + 1$
    Take action a, observe return R, next state s'
    for all s, a **do**
      $P(s,a) \leftarrow P(s,a) + \alpha \log(1+Rf)c(s,a)$
      $c(s,a) \leftarrow \gamma c(s,a)$
    end for
    $s \leftarrow s'$
  until s is terminal
end loop

2 つのパラメーターによって表す．ボリンジャー・バンドの幅の広さは移動標準偏差の大きさを表しており，金利の動きが激しいときにバンド幅は広くなり，金利の動きが穏やかなときはバンド幅が狭くなる．いずれのパラメータも，直近 14 週の移動平均と移動標準偏差を用いて，直近 14 週の値とくらべて高いか低いかを表すように変換する．例えば，金利 y_t をその移動平均 μ_t と移動標準偏差 σ_t を用いて次のように変換する．

$$o_1(t) = \frac{y_t - \mu_t}{3\sigma_t} \tag{2}$$

これによって，絶対値の大小に関わらず，直近の値との相対的な大小で表すことができる．

もうひとつの特徴は，エピソードの切り方である．通常は前のエピソードの最後の状態の次の状態を次のエピソードの初期状態として用いるが，本手法では，前のエピソードの最後の状態を次の状態の初期状態として用いる．ひとつのエピソードは同じポジションを取っている間続き，ポジションが変更されるとそれまでのエピソードが終了すると同時にその状態から新しいエピソードとなる．

6 実験

6.1 実験方法

残存期間 10 年の日本国債の週次の金利を対象とした．2001 年から 2008 年までのデータを用いた．この期間の金利の動きを図 1 に示す．

強化学習の訓練期間を 5 年間とし，同じデータに対する評価と直後 1 年間のデータに対する評価を行った．

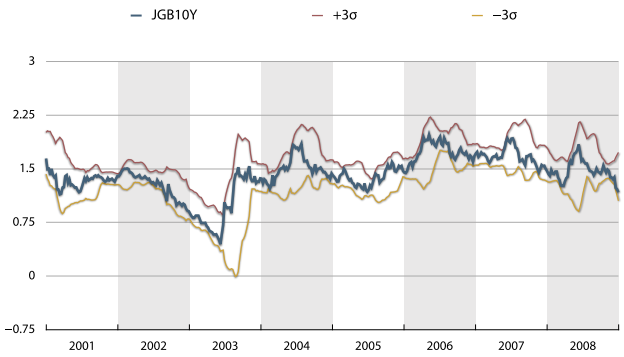

図 1　残存期間 10 年の日本国債の金利と $\pm 3\sigma$ ボリンジャー・バンド（週次）.

たとえば，2003 年から 2007 年の 5 年間のデータを用いて取引戦略を学習し，この取引戦略を同じ期間のデータと 2008 年 1 年間のデータに対してそれぞれ適用して評価を行った．これを 1 年ずつ前にずらして 3 回ずつ行った.

強化学習エージェントは, 観測値を半径 $2/14 \approx 0.143$ の RBF カーネルを 1 次元当たり 15 個ずつ，$15 \times 15 = 225$ 個並べて特徴ベクトル化し，線形関数近似を行った．エージェントの行動選択には Gibbs ソフトマックス選択を用い，温度パラメーターを学習中は $\tau = 0.2$，評価中は $\tau = 0.2$ とした．割引率パラメーターは $\gamma = 0.9$，ステップ・サイズ・パラメータは $\alpha = 1.0$，投資比率パラメーターは $f = 0.99$ とした．これらをランダム・シードを変えてそれぞれ 30 回ずつ実験を行い，その平均を求めた.

複利リターンに基づく Q 学習と従来の Q 学習，複利リターンに基づく OnPS と従来の OnPS をそれぞれ比較した.

6.2　結果

学習曲線を図 2–7 に示す．図 2, 3 は 2003 年から 2007 年までのデータから学習したもの，図 4, 5 は 2002 年から 2006 年までのデータから学習したもの，図 6, 7 は 2001 年から 2005 年までのデータから学習したものである．図の横軸は学習したステップ数，縦軸は年間平均利益を表している．エラー・バーは標準偏差 $\pm 1\sigma$ を表す.

複利リターンに基づく Q 学習の結果は，2002 年から 2006 年のデータから学習して 2007 年のデータに適用した場合を除いて，従来の Q 学習の結果とほぼ同じか僅かに良い結果となった．複利リターンに基づく OnPS の結果も，同じように，従来の OnPS とほぼ同じか僅かに良い結果となった.

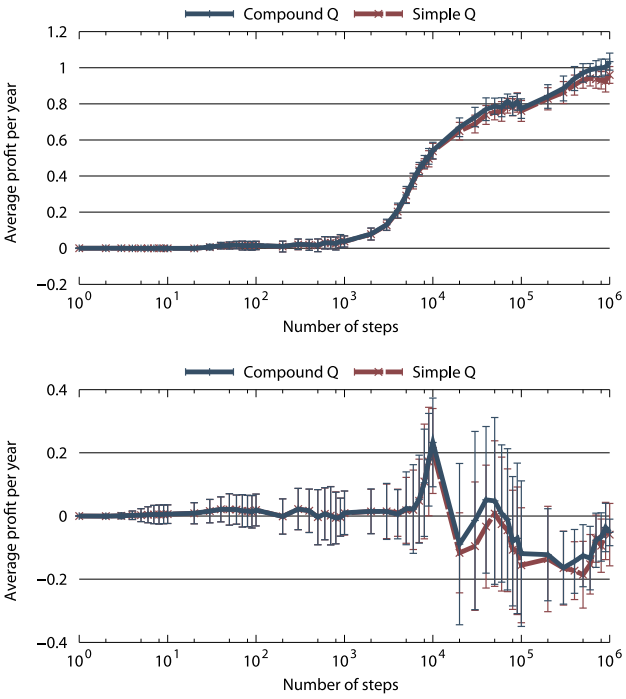

図 2　複利リターンに基づく Q 学習を用いて 2003 年から 2007 年のデータで学習し，同じデータで評価したとき（上）と 2008 年のデータで評価したとき（下）の学習曲線.

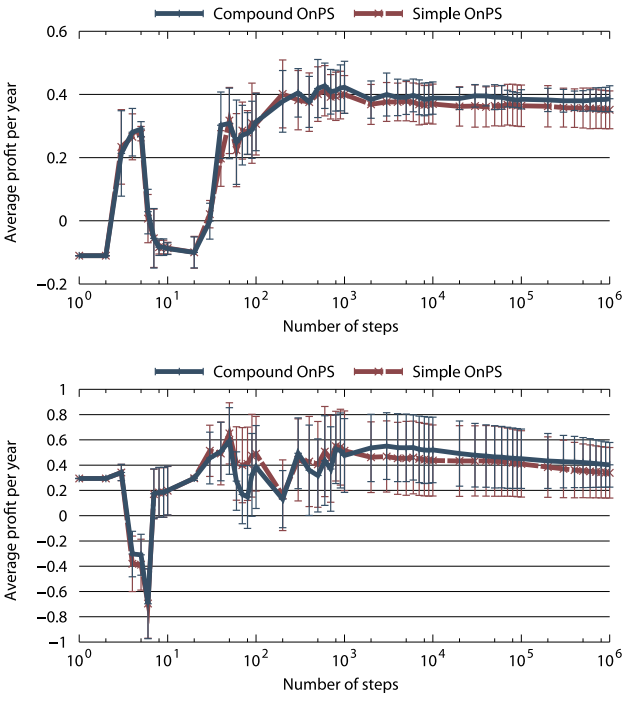

図 3　複利リターンに基づく OnPS を用いて 2003 年から 2007 年のデータで学習し，同じデータで評価したとき（上）と 2008 年のデータで評価したとき（下）の学習曲線.

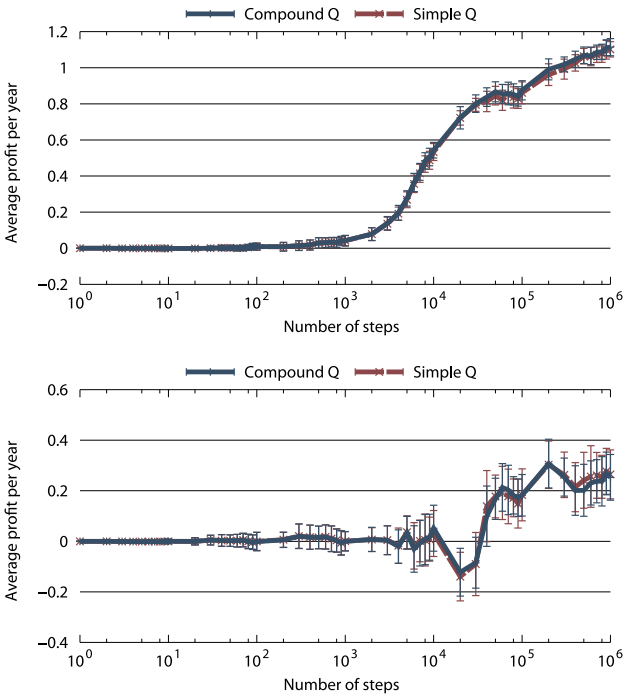

図 4　複利リターンに基づく Q 学習を用いて 2002 年から 2006 年のデータで学習し，同じデータで評価したとき（上）と 2007 年のデータで評価したとき（下）の学習曲線.

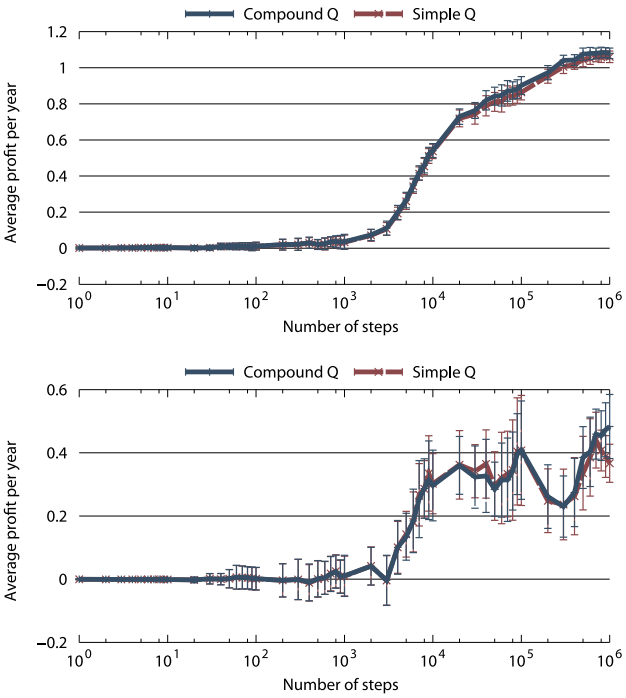

図 6　複利リターンに基づく Q 学習を用いて 2001 年から 2005 年のデータで学習し，同じデータで評価したとき（上）と 2006 年のデータで評価したとき（下）の学習曲線.

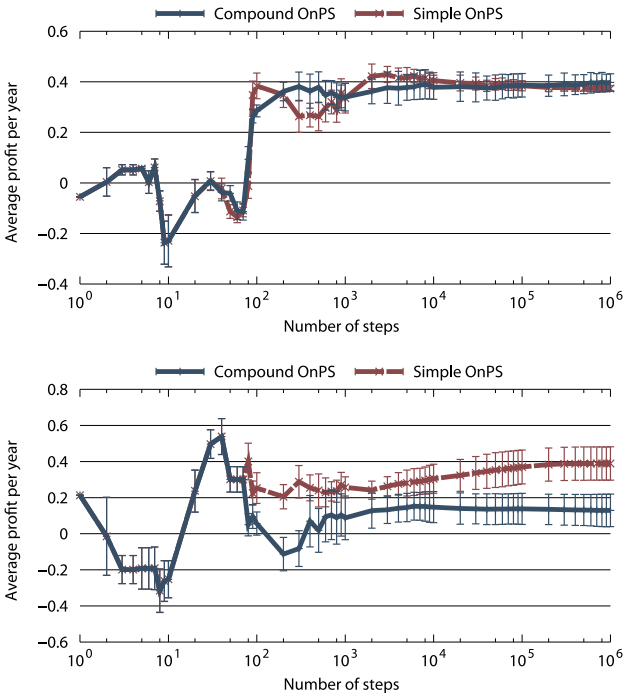

図 5　複利リターンに基づく OnPS を用いて 2002 年から 2006 年のデータで学習し，同じデータで評価したとき（上）と 2007 年のデータで評価したとき（下）の学習曲線.

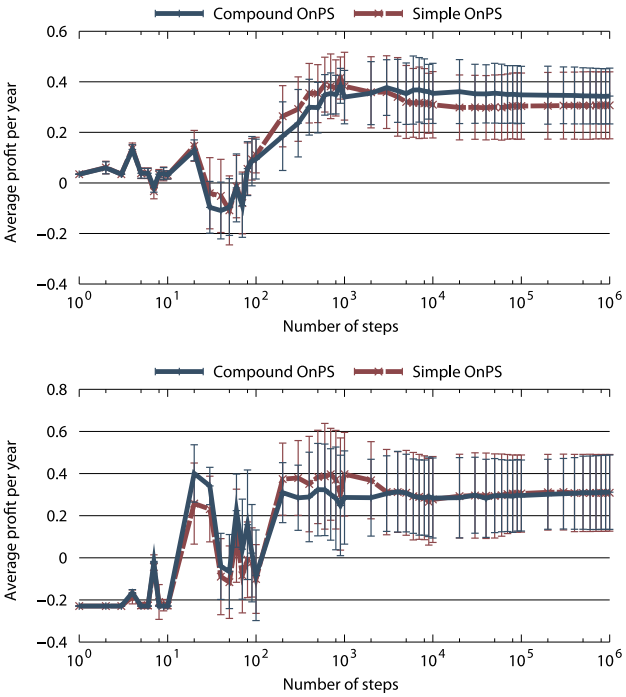

図 7　複利リターンに基づく OnPS を用いて 2001 年から 2005 年のデータで学習し，同じデータで評価したとき（上）と 2006 年のデータで評価したとき（下）の学習曲線.

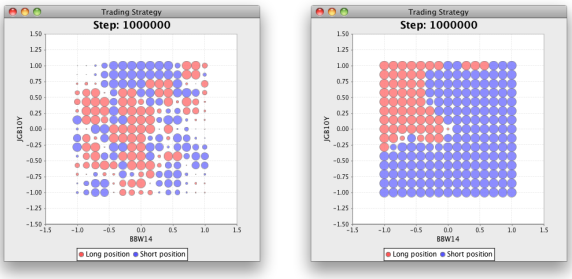

図 8　複利リターンに基づく Q 学習（左）と複利リターンに基づく OnPS（右）を用いて 2003 年から 2007 年のデータで学習したときに獲得された最も良い取引戦略.

2003 年から 2007 年までのデータから学習したときに獲得された最も良い取引戦略を図 8 に示す．縦軸が金利，横軸がボリンジャー・バンドの幅（標準偏差）を表す．赤い円がロング（購入），青い円がショート（売却）を表し，円の半径はポジションの大きさを表している．複利リターンに基づく Q 学習がかなり細かい戦略を獲得していることがわかる．

7　考察

本論文では，複利リターンに基づく強化学習を用いて日本国債の取引戦略を獲得する実験の結果を報告した．実験の結果から，複利リターンに基づく強化学習が金融市場における取引戦略の獲得に従来の強化学習と同じくらい有効であることが確認できた．

通常の強化学習を用いた実験 [2] でも見られたように，複利リターンに基づく Q 学習は学習データに対して過学習してしまうため，学習期間でない期間に対する評価が著しく低下する傾向が見られた．また，複利リターンに基づく Q 学習は，図 2（下）のように，年間平均利益がマイナスとなる場合があった．これに対し，複利リターンに基づく OnPS は，安定的にプラスの利益を獲得することができた．したがって，安定的に運用できる取引戦略を獲得するには，複利リターンに基づく Q 学習よりも複利リターンに基づく OnPS が適している．

複利リターンに基づく強化学習と従来の強化学習の結果があまり変わらないのは，リターンの絶対値が小さいからだと考えられる．リターンが -1 となる行動が存在するときには，複利リターンに基づく強化学習はこれを回避できることが分かっている [11]．しかしながら，今回対象とした日本国債の市場では極端なリターンが生じないため，複利リターンに基づく強化学習の効果が大きく現れなかったと考えられる．

参考文献

[1] J. W. Lee, J. Park, J. O, J. Lee, and E. Hong. A multiagent approach to Q-learning for daily stock trading. *IEEE Trans. on Systems, Man and Cybernetics, Part A*, Vol. 37, No. 6, pp. 864–877, 2007.

[2] T. Matsui, T. Goto, and K. Izumi. Acquiring a government bond trading strategy using reinforcement learning. *Journal of Advanced Computational Intelligence and Intelligent Informatics*, Vol. 13, No. 6, pp. 691–696, 2009.

[3] T. Matsui, N. Inuzuka, and H. Seki. On-line profit sharing works efficiently. In *Proc. of the 7th Int'l Conf. on Knowledge-Based Intelligent Information & Engineering Systems*, pp. 317–324, 2003.

[4] J. O, J. Lee, J. W. Lee, and B.-T. Zhang. Adaptive stock trading with dynamic asset allocation using reinforcement learning. *Information Science*, Vol. 176, pp. 2121–2147, 2006.

[5] A. A. Sherstov and P. Stone. Three automated stock-trading agents: A comparative study. In *Proceedings of the AAMAS 2004 Workshop on Agent-Mediated Electronic Commerce*, pp. 173–187, 2005.

[6] R. S. Sutton and A. G. Barto. *Reinforcement Learning: An Introduction*. The MIT Press, 1998. 三上, 皆川 共訳. 強化学習. 森北出版, 2000.

[7] C.J.C.H. Watkins and P. Dayan. Technical note: Q-learning. *Machine Learning*, Vol. 8, No. 3/4, pp. 279–292, 1992.

[8] 松井, 後藤, 和泉, 大和田. 強化学習を用いた債券取引戦略の獲得. 2008 年度人工知能学会（第 22 回）全国大会講演論文集, pp. 2C3–1, 2008.

[9] 松井, 後藤. 強化学習を用いた金融市場取引戦略の獲得と分析. 人工知能学会誌, Vol. 24, No. 3, pp. 400–407, 2009.

[10] 松井, 後藤, 和泉, 大和田. 強化学習を用いた金融市場取引戦略分析システムの試作. 人工知能学会 ファイナンスにおける人工知能応用研究会 (第 1 回) 研究会資料, pp. 12–17, 2008.

[11] 松井. ファイナンスのための強化学習. 第 3 回ファイナンスにおける人工知能応用研究会（SIG-FIN）, pp. 81–88, 2009.