複利型強化学習を用いた
国債の金利とデフォルト確率に基づく銘柄選択

松井藤五郎 [1*] 後藤卓 [2] 和泉潔 [3,4] 陳昱 [3]

[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，国債の銘柄選択問題を金利とデフォルト確率に基づいて N 本腕バンディット問題としてとらえ，これを複利型強化学習を用いて投資戦略を学習する方法を提案する．提案手法を用いて 2010 年第 2 四半期の日米欧各国の国債を対象にした強化学習タスクを作成し，複利型 Q 学習を用いて学習を行った．また，学習した行動価値に基づいてポートフォリオを構成し，モンテ・カルロ・シミュレーションによってパフォーマンスを評価した結果を示す．

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 [2]．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 [12] を開発し，獲得された取引戦略の分析を行った [3, 10, 11]．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，複利型強化学習を提案している [8, 9]．

本論文では，この複利型強化学習を国債の銘柄選択問題に適用する．具体的には，国債の金利とデフォルト確率に基づいて銘柄選択問題を N 本腕バンディット問題としてタスク設計を行い，複利型強化学習を用いて各銘柄に対する行動価値を学習する．その後，学習した行動価値に基づいてポートフォリオを作成し，モンテ・カルロ・シミュレーションによってパフォーマンスを評価する．

2 複利型強化学習

2.1 複利型強化学習の枠組み

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，複利型強化学習 [8, 9] では，割引複利リ

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

ターン

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

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．

$$
\begin{aligned}
V^{\pi}(s) =& \mathrm{E}_{\pi}\left[\log\prod_{k=0}^{\infty}(1+R_{t+k+1}f)^{\gamma^k}\middle| s_t=s\right] \\
=& \sum_{a\in\mathcal{A}}\pi(s,a) \\
& \sum_{s'\in\mathcal{S}}\mathcal{P}^{a}_{ss'}\left(\log(1+\mathsf{R}^{a}_{ss'}f)+\gamma V^{\pi}(s')\right) \\
Q^{\pi}(s,a) =& \mathrm{E}_{\pi}\left[\log\prod_{k=0}^{\infty}(1+R_{t+k+1}f)^{\gamma^k}\middle| s_t=s, a_t=a\right] \\
=& \sum_{s'\in\mathcal{S}}\mathcal{P}^{a}_{ss'}\left(\log(1+\mathsf{R}^{a}_{ss'}f)+\gamma V^{\pi}(s')\right) \qquad (2)
\end{aligned}
$$

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．

2.2 複利型 Q 学習アルゴリズム

複利型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．複利型 Q 学習 [8, 9] 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．複利型 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，報酬型 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，複利型 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，リターン型 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 [8, 9]．ここで，報酬型 MDP とは従来の強化学習が対象としている MDP のことを表し，リターン型 MDP とは従来の MDP における報酬をリターンに置き換えたものである．

3 国債の金利とデフォルト確率に基づく銘柄選択

本論文では，国債の金利とデフォルト確率に基づいて，国債銘柄選択問題を N 本腕バンディット問題としてタスク化する方法を提案する．

デフォルト確率は Credit Market Analysis (CMA) 社が発行している Global Sovereign Credit Risk Report の 2010 年第 2 四半期版 [1] から取得した．このレポートは，各国の国債に対し 5 年以内にデフォルトが発生する確率に基づいて各国をレーティングしたものである．本レポートにおけるデフォルト確率の期間が 5 年であるため，本論文では 5 年国債を対象とした．また，本レポートは 2010 年 6 月 29 日の終値に基づいてデフォルト確率を算定しているため，実質金利は同日の終値を用いた．

銘柄選択の対象国は，主要通貨の流通国である日本，アメリカ，イギリス，ドイツに加え，ユーロ圏内主要国でデフォルト確率が特徴的なイタリア，スペイン，ギリシャ，そして最もデフォルト確率が低いノルウェーとした[*2]．日本，アメリカ，イギリス，ドイツ，イタリア，スペインの各国債の実質金利は，WSJ.com[*3]を参照した．ギリシャ国債の実質金利は Bloomberg[*4]，ノルウェー国債の実質金利はノルウェー中央銀行 Norges Bank[*5]のサイトを参照した．対象国の 5 年国債の実質金利とデフォルト確率を表 1 と図 1 に示す．一般に，デフォルト確率と実質金利には正の相関がある．図 1 から，日本国債はデフォルト確率に対して実質金利が非常に低く，ノルウェー国債はデフォルト確率に対して実質金利が高いことがわかる．

5 年国債のため，デフォルトが発生しない場合は満期までに実質金利（年利）の 5 倍のリターンが得られる．デフォルトは 2 年半後にデフォルト確率に応じて発生するものと仮定し，デフォルトが発生した場合は一定割

[*1] 最小値が -1 であるため，上界だけあればいい．
[*2] 最もデフォルト確率が高いのはベネズエラ（58.4%）で，ギリシャは 2 番目に高い．
[*3] http://online.wsj.com
[*4] http://www.bloomberg.com
[*5] http://www.norges-bank.no

Algorithm 1 複利型 Q 学習アルゴリズム

入力: 割引率 γ, ステップ・サイズ α, 投資比率 f
$Q(s,a)$ を任意に初期化
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
    $Q(s,a) \leftarrow Q(s,a) + \alpha\left(\log(1+Rf) + \gamma \max_{a'} Q(s',a') - Q(s,a)\right)$
    $s \leftarrow s'$
  until s が終端状態ならば繰り返しを終了
end loop

表 1　対象国の 5 年国債の実質金利とデフォルト確率

国名	略号	実質金利	デフォルト確率
日本	JP	0.354%	8.3%
アメリカ	US	1.788%	3.4%
イギリス	UK	2.125%	6.6%
ドイツ	DE	1.429%	3.9%
イタリア	IT	2.973%	15.5%
スペイン	ES	3.908%	20.7%
ギリシャ	GR	10.948%	53.0%
ノルウェー	NO	2.550%	2.3%

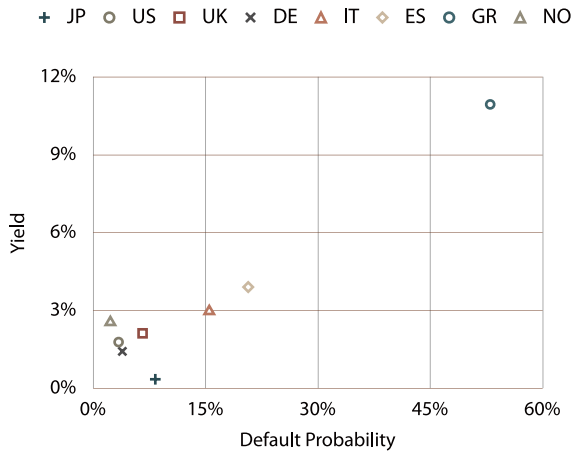

図 1　デフォルト確率と実質金利の関係.

合の元本が削減されるものとした．すなわち，デフォルト確率 p，実質金利 y の国債を購入したときのリターンは，確率 p で $5y$，確率 $1-p$ で $2.5y - ppr$ となる．ここで，ppr はデフォルト発生時の元本削減率を表す．

国債を購入しない場合は，現金として保有するものとした．現金を選択した場合のデフォルト発生確率は 0 で，かつ，リターンも 0 である．

また，ここでは為替レートの変動を考慮していない．

4　実験

4.1　実験方法

上で述べたデータを基にして国債銘柄選択問題を作成し，実験を行った．選択できる銘柄は，上記の 8 ヵ国に「現金」を加えた 9 銘柄とした．本論文では，デフォルトが発生したときの元本削減率を 75%，50%，25% とした三通りのシナリオを想定して実験を行った．それぞれのシナリオにおいて，各銘柄を選択したときに得られるリターンとその生起確率を表 2 に示す．

割引率パラメーター γ は，単利型，複利型ともに 0.9 とした．複利型 Q 学習における投資比率 f は 0.99 とした．ステップ・サイズ・パラメーター α は，単利型，複利型ともに 0.05 とした．学習時の行動選択には $\epsilon = 0.2$ の ϵ-グリーディ選択を用いた．

強化学習は，10^9 ステップの学習をランダム・シードを変えて 100 回行い，学習した行動価値（Q-values）の平均を求めた．この平均行動価値から銘柄を組み合わせてポートフォリオを作成し，そのパフォーマンスを評価した．ポートフォリオは，(1) 平均行動価値に対して比例配分したものと，(2) 現金（Cash）の平均行動価値を超過した平均行動価値に対して比例配分したものを用意した．これらのポートフォリオに対して，1 期のモンテ・カルロ・シミュレーションを 10,000 回行い，その幾何平均リターンを求めた．これは，利息を再投資しない場合の幾何平均リターンに相当する．

表 2　各銘柄を選択したときに得られるリターン．上段はデフォルトが発生しないとき，下段はデフォルトが発生するときを表す．

行動	確率	削減率 75%	削減率 50%	削減率 25%
JP	0.917	1.770%	1.770%	1.770%
	0.083	−74.115%	−49.115%	−24.115%
US	0.966	8.94%	8.94%	8.94%
	0.034	−70.53%	−45.53%	−20.53%
UK	0.934	10.6250%	10.6250%	10.6250%
	0.066	−69.6875%	−44.6875%	−19.6875%
DE	0.961	7.1450%	7.1450%	7.1450%
	0.039	−71.4275%	−46.4275%	−21.4275%
IT	0.840	14.8650%	14.8650%	14.8650%
	0.160	−65.5675%	−42.5675%	−17.5675%
ES	0.790	19.54%	19.54%	19.54%
	0.210	−65.23%	−40.23%	−15.23%
GR	0.470	54.74%	54.74%	54.74%
	0.530	−47.63%	−22.63%	2.37%
NO	0.977	12.75%	12.75%	12.75%
	0.023	−68.625%	−43.625%	−18.625%
Cash	1.000	0	0	0

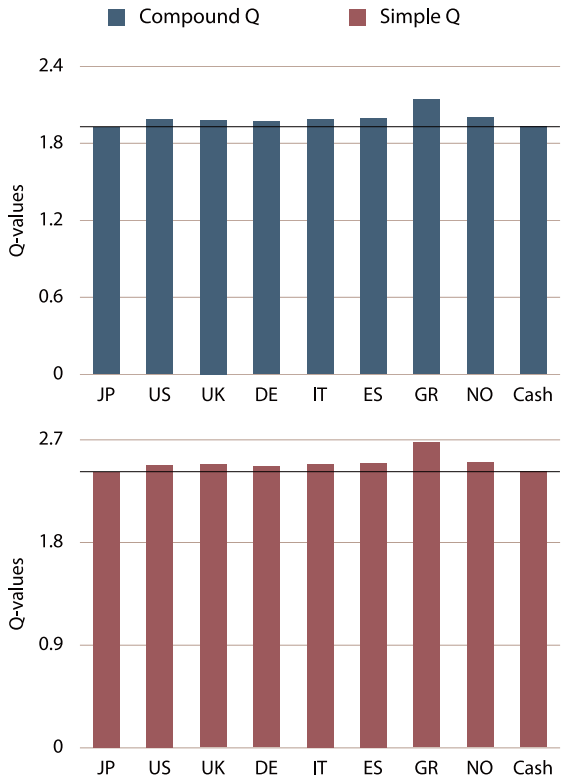

図 2　デフォルト発生時の元本削減率を 25% と想定したときに複利型 Q 学習と単利型 Q 学習が学習した行動の価値（Q 値）．

4.2　結果

まず，デフォルト発生時の元本削減率を 25% と想定したシナリオでの結果を図 2 に示す．上のグラフは複利型 Q 学習が学習した行動価値（Q-values）の平均値，下のグラフは単利型 Q 学習が学習した行動価値の平均値である．行動価値の定義が異なるため，複利型 Q 学習と単利型 Q 学習で行動価値の絶対的な大きさが違うことにはあまり意味はなく，それぞれが学習した他の行動（銘柄）との相対的な関係が重要である．このシナリオでは，複利型 Q 学習，単利型 Q 学習ともにギリシャ国債を選択する行動の価値が最も高いと学習した．

同様にして，デフォルト発生時の元本削減率を 50% と想定したシナリオの結果を図 3 に，75% と想定したシナリオの結果を図 4 に示す． この二つのシナリオでは，複利型，単利型ともにノルウェー国債を選択する行動の価値が最も高いと学習した．

これらの平均行動価値に比例配分してポートフォリオを構成し，モンテ・カルロ・シミュレーションによるパフォーマンス評価を行った結果を図 5 に示す．また，平均行動価値が現金（Cash）の平均行動価値を超過した部分（黒い横線の上の部分）に比例半分してポートフォリオを構成して同様に評価した結果を図 6 に示す．

デフォルト発生時の元本削減率を 50% あるいは 75% と想定したときは，超過行動価値に比例配分してポートフォリオを構成したときの複利型 Q 学習のパフォーマンスが単利型 Q 学習のパフォーマンスよりも大きかった．それ以外のケースでは大きな差は見られなかった．

5　考察

学習した平均行動価値に比例配分してポートフォリオを構成したときのパフォーマンスは，従来の Q 学習と複利型 Q 学習で大きな差がない（図 5）．これは，構成されるポートフォリオに大きな違いがないためである．

現金に対する超過行動価値に比例配分してポートフォリオを構成したときのパフォーマンスは，デフォルト発生時の元本削減率を 50% と想定したときと 75% と想定したときに複利型 Q 学習の方が従来の Q 学習に比べて良かった（図 6）．これは，複利型 Q 学習のポートフォリオのほとんどがノルウェー国債で占められているのに対し，従来の Q 学習のポートフォリオにおけるノルウェー国債の割合が約 2/3 であったためである．また，デフォルト発生時の元本削減率を 25% と想定したときは，ポートフォリオに大きな違いはなく，その結果，パフォーマンスに大きな差が生じなかった．

デフォルト発生時の元本削減率を 25% と想定したと

図 3　デフォルト発生時の元本削減率を 50% と想定したときに複利型 Q 学習と単利型 Q 学習が学習した行動の価値（Q 値）.

き，ギリシャ国債はデフォルトが発生したとしても 2.5 年分の利息で元本削減分をカバーでき，結果として正のリターンとなる．したがって，複利型 Q 学習と単利型 Q 学習ともにギリシャ国債を選択する行動の価値が最も高いと学習したことは妥当であると考えられる．

デフォルト発生時の元本削減率を 50% あるいは 75% と想定したとき，複利型 Q 学習はギリシャ国債を選択する行動の価値を比較的低く学習している．これは，複利型 Q 学習が今回のタスクに対して有効に働いていることを示していると考えられる．

6　まとめ

本論文では，国債の金利とデフォルト確率に基づいて国債銘柄選択問題を N 本腕バンディット問題としてタスク化する方法を提案した．また，このタスクに従来の Q 学習と複利型 Q 学習を適用し，その結果を比較した．

学習した行動価値に基づいてポートフォリオを構成し，モンテ・カルロ・シミュレーションによってパフォーマンスを評価した結果，いずれも幾何平均リターンが正であった．今回の実験では複利型 Q 学習のほうが従来

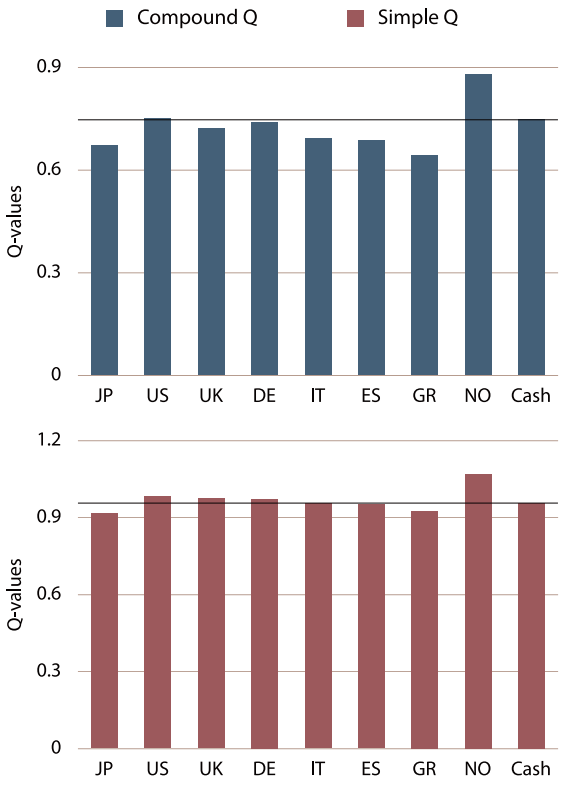

図 4　デフォルト発生時の元本削減率を 75% と想定したときに複利型 Q 学習と単利型 Q 学習が学習した行動の価値（Q 値）.

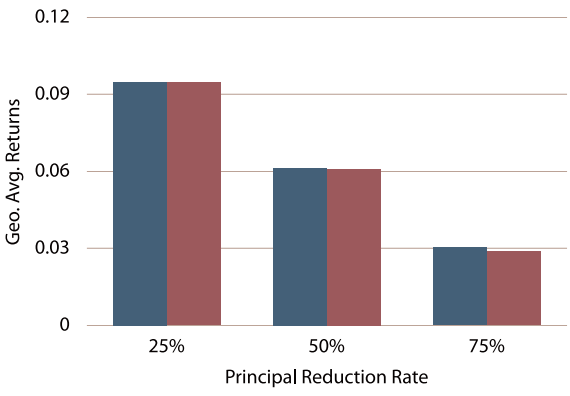

図 5　行動価値 (Q-values) に比例配分してポートフォリオを構成したときの幾何平均リターン.

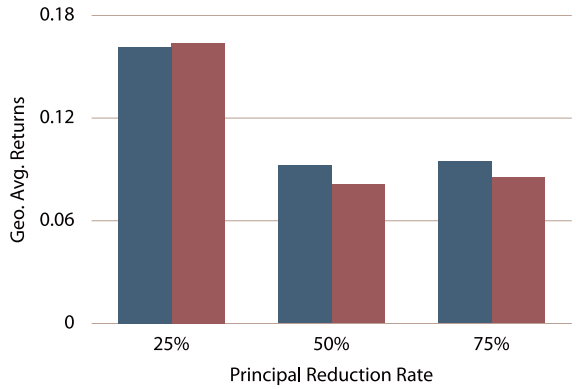

図 6　現金（cash）に対する超過行動価値（Q-values）に比例配分してポートフォリオを構成したときの幾何平均リターン．

の Q 学習よりもパフォーマンスが良いケースがあったが，これは行動価値からポートフォリオを構成する方法に依存する．

今後は，行動価値からポートフォリオを構成する方法について検討する必要がある．また，強化学習アルゴリズムの評価方法についても検討する必要がある．

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．本論文に記載された情報は信頼すべき情報源から入手したものですが，誤りが存在する可能性があります．したがって，当該情報および結果の正確性について一切保証するものではありません．また，意思決定に関してなんらの推奨をするものでもありません．

参考文献

[1] Credit Market Analysis Limited (CMA). Global sovereign credit risk report, 2nd quarter 2010.

[2] Jae Won Lee, Jonghun Park, Jangmin O, Jongwoo Lee, and Euyseok Hong. A multiagent approach to Q-learning for daily stock trading. *IEEE Transactions on Systems, Man and Cybernetics, Part A*, Vol. 37, No. 6, pp. 864–877, 2007.

[3] Togoroh Matsui, Takashi Goto, and Kiyoshi 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.

[4] Jangmin O, Jongwoo Lee, Jae Won Lee, and Byoung-Tak Zhang. Adaptive stock trading with dynamic asset allocation using reinforcement learning. *Information Science*, Vol. 176, pp. 2121–2147, 2006.

[5] Alexander A. Sherstov and Peter Stone. Three automated stock-trading agents: A comparative study. In *Proceedings of the AAMAS 2004 Workshop on Agent-Mediated Electronic Commerce (AMEC 2004)*, Vol. 3435 of *Lecture Notes in Artificial Intelligence*, pp. 173–187, 2005.

[6] 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.

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

[8] 松井藤五郎. 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.

[9] 松井藤五郎. 複利型強化学習. 2010 年度人工知能学会全国大会 (JSAI 2010), 1A3-2, 2010.

[10] 松井藤五郎, 後藤卓, 和泉潔, 大和田勇人. 強化学習を用いた債券取引戦略の獲得. 2008 年度人工知能学会全国大会 (JSAI 2008), 2C3-1, 2008.

[11] 松井藤五郎, 後藤卓. 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.

[12] 松井藤五郎, 後藤卓, 和泉潔, 大和田勇人. 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. 第 1 回ファイナンスにおける人工知能応用研究会 (SIG-FIN), pp. 12–17, 2008.