2H1-OS18-4

複利型強化学習を用いた国債銘柄選択

Global Government Bonds Selection Using Compound Reinforcement Learning

松井藤五郎[*1][*2] 後藤卓[*3] 和泉潔[*4][*5] 陳ユ[*4]
Tohgoroh Matsui Takashi Goto Kiyoshi Izumi Yu Chen

[*1]中部大学生命健康科学部
College of Life and Health Sciences, Chubu University

[*2]中部大学工学部
College of Engineering, Chubu University

[*3]三菱東京 UFJ 銀行
Bank of Tokyo-Mitsubishi UFJ, Ltd.

[*4]東京大学大学院工学系研究科
School of Engineering, The University of Tokyo

[*5]JST さきがけ
PRESTO, JST

This paper describes an application of compound reinforcement learning to global government bonds selection. We formulated the global government bonds selection problem as a n-armed bandit problem based on the yields and the default probabilities. We then confirmed that compoud Q-learning works well in the global government bonds selection problem.

1. はじめに

強化学習 [Sutton 98] 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 [Sherstov 05]．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 [O 06]．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 [Lee 07]．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を開発し，獲得された取引戦略の分析を行った [Matsui 09, 松井 09]．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（利益）よりもリターン（利益の割合）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 -0.065 と負であるのに対し銘柄 D は 0.331 と正であり，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 [松井 11]．

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，国債の利回り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定式化する方法を提案する．また，アメリカ，ドイツ，イギリスの 3 ヵ国の国債を対象として，提案手法を用いて 3 本腕バンディット問題を作成し，このタスクに対して複利型 Q 学習が従来の Q 学習よりも有効であることを示す．

連絡先: 松井藤五郎, TohgorohMatsui@tohgoroh.jp,
http://とうごろう.jp

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 [Sutton 98] では，エージェントは割引収益

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

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．

複利型強化学習 [松井 11] は，従来の MDP における報酬 r_t をリターン R_t に置き換えたリターン型 MDP を対象とする．複利型強化学習では，エージェント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．

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

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

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グロス・リターンの対数の期待値

$$\mathsf{R}^a_{ss'} = \mathrm{E}\left[\log(1+R_{t+1}f)\middle| s_t=s, a_t=a, s_{t+1}=s'\right]$$

を表す．式 (1) および式 (2) は，従来の強化学習の V^π および Q^π の式における

$$\mathcal{R}^a_{ss'} = \mathrm{E}\left[r_{t+1}\middle| s_t=s, a_t=a, s_{t+1}=s'\right]$$

を $\mathsf{R}^a_{ss'}$ に置き換えたものに等しい．

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・リターンの対数の期待値 $\mathsf{R}^a_{ss'}$ に置き換えたものに等しい．複利型 Q 学習 [松井 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．複利型 Q 学習のアルゴリズムを Algorithm 1 に示す．

Q 学習では，報酬 r_t が有界で，ステップ・サイズ・パラメーター α 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 [Watkins 92]．同様にして，複利型 Q 学習でも，$\log(1+R_tf)$ が有界で，ステップ・サイズ・パラメーター α 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 [松井 11]．

3. 国債銘柄選択問題の n 本腕バンディット問題への定式化

3.1 n 本腕バンディット問題

n 本腕バンディット問題は，強化学習の教科書 [Sutton 98] でも取りあげられているシンプルで標準的な連続型タスクである．この問題の環境は，状態数が 1，取りうる行動の数が n であるマルコフ決定過程（MDP）である．エージェントは，n 種類の行動の中から一つを選択することを繰り返す．どの行動を選択しても再び同じ状態に戻るため，状態遷移確率はすべての行動 a に対して $\mathcal{P}^a_{ss}=1$ である．報酬の確率分布は行動ごとに異なる．したがって，n 本腕バンディット問題をタスクとして設定するには，取り得る行動の数 n とそれぞれの行動に対する報酬（リターン）の確率分布を決める必要がある．

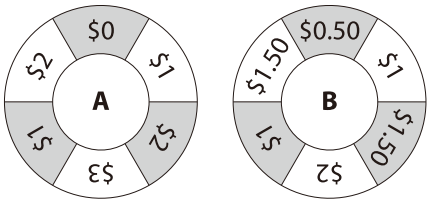

図 1: マネー・ホイールを用いた 2 本腕バンディット問題．

例として，図 1 マネー・ホイールと呼ばれるカジノ・マシンを用いた問題を示す．二つのホイールがあり，エージェントはどちらかを選択する．どちらのホイールを選択しても，同じ状態に戻って再びどちらかを選択する．図に示されている金額は，1 ドルの賭け金に対する払い戻し金の額である．報酬は払い戻し金から賭け金を引いた値であり，リターンは払い戻し金を賭け金で割った値から 1 を引いた値であるから，1 ドルを賭けたときの報酬とリターンは等しい．

この問題においては，1 ドルずつ賭け続ける場合には期待報酬が大きいホイール A の方が良いが，全額を賭け続ける場合，つまり複利リターンを考慮した場合にはホイール B の方が良い．従来の強化学習は算術期待報酬が大きいホイール A を選択する行動規則を学習するが，複利型強化学習は複利リターン（幾何期待リターン）が大きいホイール B を選択する行動規則を学習できる [松井 11]．

3.2 国債の利回りとデフォルト確率

国債は，無リスク資産と考えられることもあるが，実際には，債務不履行（デフォルト）となる可能性がある．つまり，金融危機や経済危機がその国を直撃すると，政府は利払いや償還ができなくなりデフォルトが発生する恐れがある．しがたって，我々が国債を選択する際には，利回りだけでなくデフォルト・リスクを考慮しなければならない．

国債（固定利付債）には，1 年分の利子（表面利率）がパーセント表示されており，半年を 1 期として期ごとに表面利率の半分の利子が支払われる．満期を迎えると償還され，額面と同じ金額で買い戻される．発行された国債を直接購入する場合は表面利率を運用利率とみなすことができるが，国債市場で取引される国債には償還期間が短くなっている（発行されてから 1 期以上経過した）債券が含まれるため，表面利率をそのまま運用利率とすることができない．そこで，残存期間を発行時と同じとして利回り（最終利回り）と価格を換算し，取引が行われる．

国債市場で取引された国債の利回りは，Wall Street Journal の Web サイト，WSJ.com[*1] などで公開されている．また，デフォルトが発生する確率については，Credit Market Analysis（CMA）社が 4 半期ごとに発行している Global Sovereign Credit Risk Report の中で，各国において 5 年以内にデフォルトが発生する確率が公表されている．

*1 http://online.wsj.com

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 年国債の利回りと 5 年以内にデフォルトが発生する確率

国名	利回り	デフォルト確率
アメリカ	1.929%	3.6%
ドイツ	2.222%	5.2%
イギリス	2.413%	6.4%

3.3 n 本腕バンディット問題として定式化する方法

本論文では，主要通貨の流通国であるアメリカ，ドイツ，イギリスを対象として，国債の利回りとデフォルト確率を用いて 3 本腕バンディット問題として定式化した．ここでは，この方法を説明する．

CMA 社の Global Sovereign Credit Risk Report には，5 年以内にデフォルトが発生する確率が掲載されている．今回参照した 2010 年第 4 四半期版のレポート [CMA 11] は，2010 年 12 月 31 日の終値に基づいて計算されたデフォルト確率が掲載されている．そこで，対象を残存期間 5 年の国債とし，2010 年 12 月 31 日の利回りの終値を WSJ.com で調べた．各国の 5 年国債の利回りと 5 年以内にデフォルトが発生する確率を表 1 に示す．

本論文では，デフォルトがポアソン過程に従って発生すると仮定し，5 年以内にデフォルトが発生する確率から半年以内にデフォルトが発生する確率を次のようにして求める．ポアソン過程の式より，t 期までにデフォルトが k 件発生する確率は

$$\Pr(N_t = k) = \frac{e^{-\lambda t}(\lambda t)^k}{k!}$$

と表される．この式より，10 期（5 年）までにデフォルトが発生しない（0 件発生する）確率は，$t = 10, k = 0$ を代入して $\Pr(N_{10} = 0) = e^{-10\lambda}$ となる．5 年以内にデフォルトが発生する確率を p とすると，

$$p = \Pr(N_{10} > 0) = 1 - \Pr(N_{10} = 0) = 1 - e^{-10\lambda}$$

が成り立つ．ここから，$\lambda = -\frac{1}{10}\log(1-p)$ が求まる．1 期までにデフォルトが発生しない確率は，$\Pr(N_1 = 0) = e^{-\lambda}$ と表され，ここに $\lambda = -\frac{1}{10}\log(1-p)$ を代入すると $\Pr(N_1 = 0) = \sqrt[10]{(1-p)}$ となる．したがって，1 期までにデフォルトが発生する確率は

$$\Pr(N_1 > 0) = 1 - \Pr(N_1 = 0) = 1 - \sqrt[10]{(1-p)}$$

となる．

デフォルトが発生した場合は，利払いが停止され，元本の一部が償還されないものとする．したがって，利回りが y，元本削減率が r のとき，t 期にデフォルトが発生したときのリターンは $(t-1)y/2 - r$ として計算できる．デフォルトが発生しなかったときのリターンは $5y$ である．

このようにして，残存期間 5 年の国債銘柄選択問題を，それぞれの銘柄ごとに異なる 11 通りのリターンが確率的に生じる n 本腕バンディット問題として定式化できる．

例として，ドイツ国債を購入し，2 期（半年後から 1 年後までの間）にデフォルトが発生する場合について考える．元本削減率は 75% とする．1 期の期末（半年後）に年間利回りの半分に相当する 1.111% のリターンが得られるが，その後にデフォルトが発生したことにより，それ以後の利払いが行われず，元本の 75% が失われる．したがって，トータルのリターンは $0.01111 - 0.75 = -0.73889$ となる．ドイツ国債に対して 1 期までにデフォルトが発生する確率を計算すると $\Pr(N_1 > 0) \approx 0.005326$ であるから，この事象が発生する確率は $\Pr(N_1 = 0) \times \Pr(N_1 > 0) \approx 0.005609$ となる．

ここでは，簡単化のため，利回り，デフォルト確率，為替レートの変動を考慮しないものとする．また，このタスクにおける投資比率は $f = 0.99$，割引率は $\gamma = 0.9$，元本削減率は 75% とした．

4. 実験結果

国債銘柄選択問題を用いて，複利型 Q 学習（Compound Q-learning）と従来の Q 学習（Simple Q-learning）の比較を行った．エージェントは，学習中は $\epsilon = 0.2$ の ϵ-グリーディー選択を用いて行動を選択し，評価時はグリーディー選択を用いて行動を選択した．従来の強化学習に対する報酬は，複利型強化学習と同じく投資比率 $f = 0.99$ のときのリターンとした．ステップ・サイズ・パラメーターは $\alpha = 0.001$ とした．これらのパラメーターは，経験的に選択したものである．学習とは独立に 100 万回の試行を行い，その平均パフォーマンスを求めることによって，評価を行った．また，乱数のシードを変えて 100 回の実験を行い，その平均を求めた．

結果を図 2 に示す．横軸は学習ステップ数を表す．上のグ

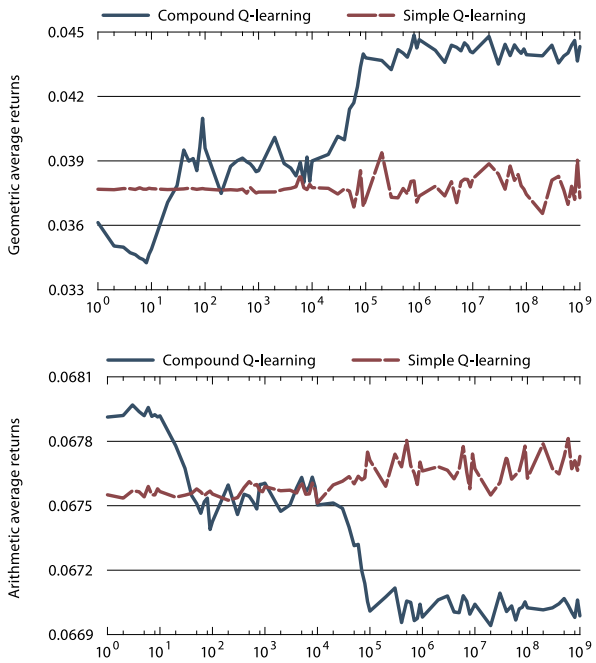

図 2: 実験結果．横軸は学習ステップ数，縦軸は幾何平均リターン（上）と算術平均リターン（下）．

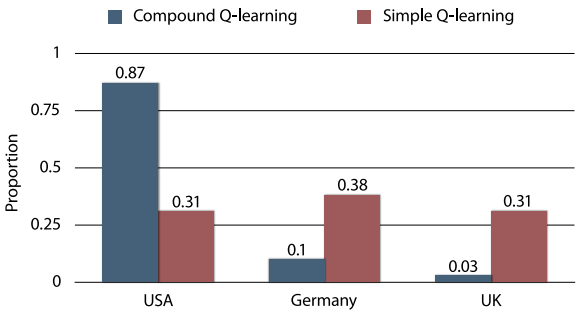

図 3: 各国債が学習された取引戦略によって選択された割合．

ラフの縦軸は幾何平均リターン，下のグラフの縦軸は算術平均リターンを表している．複利型 Q 学習では幾何平均リターンが上昇したが，従来の Q 学習の幾何平均リターンは上昇しなかった．

学習された取引戦略によって選択された国債銘柄の割合を図 3 に示す．従来の Q 学習は特定の銘柄を選択するように収束せず，乱数のシードによる選択される銘柄のバラツキが大きかった．それに対し，複利型 Q 学習は，大半のシードにおいて米国債を選択する取引戦略を学習した．

5. 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，複利型 Q 学習をこのタスクに応用し，複利型強化学習が従来の強化学習に比べてファイナンス分野への応用に適していることを確認した．

従来の強化学習は，割引収益を最大化するよう学習しているため，学習の進行とともに算術平均リターンは増加したものの幾何平均リターンは増加しなかった．これに対し，複利型強化学習では，割引複利リターンを最大化するよう学習しているため，学習の進行とともに幾何平均リターンが増加した．このことから，複利型強化学習が幾何平均リターン，つまり複利リターンを最大化するのに有効であることが確認できた．

複利型強化学習の有効性をさらに明らかにするために，ファイナンス分野のタスクをより一般的な多状態の MDP に定式化することが次の課題である．

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．

謝辞

本研究は科研費（課題番号 23700182）の助成を受けています．

参考文献

[CMA 11] CMA: Global Sovereign Credit Risk Report, 4th quarter 2010, Credit Market Analysis Ltd. (2011)

[Lee 07] Lee, J. W., Park, J., O, J., Lee, J., and Hong, E.: 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)

[Matsui 09] Matsui, T., Goto, T., and Izumi, K.: 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)

[O 06] O, J., Lee, J., Lee, J. W., and Zhang, B.-T.: Adaptive stock trading with dynamic asset allocation using reinforcement learning, *Information Science*, Vol. 176, pp. 2121–2147 (2006)

[Sherstov 05] Sherstov, A. A. and Stone, P.: Three Automated Stock-Trading Agents: A Comparative Study, in *Proceedings of the AAMAS 2004 Workshop on Agent-Mediated Electronic Commerce (AMEC 2004)*, pp. 173–187 (2005)

[Sutton 98] Sutton, R. S. and Barto, A. G.: *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

[Watkins 92] Watkins, C. J. C. H. and Dayan, P.: Q-Learning, *Machine Learning*, Vol. 8, No. 3/4, pp. 279–292 (1992)

[松井 09] 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)

[松井 11] 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)