情報処理学会論文誌　Vol. 52　No. 12　3300–3308 (Dec. 2011)

複利型強化学習の枠組みと応用

松井　藤五郎[†1]　後藤　卓[†2]
和泉　潔[†3,†4]　陳　ユ[†3]

本論文では，強化学習において複利リターンを最大化する複利型強化学習の枠組みを示し，ファイナンス分野のタスクへの応用例を示す．複利型強化学習は，報酬の代わりにリターンがマルコフ性を満たすリターン型 MDP を対象とする．複利型強化学習では，二重指数的割引と投資比率の概念を導入し，対数をとることによって従来の強化学習と同様の方法で割引複利リターンを最大化する．続いて，従来の強化学習のアルゴリズムである Q 学習と OnPS を複利型に拡張した複利型 Q 学習と複利型 OnPS のアルゴリズムを示す．また，3 本腕バンディット問題に対する実験結果と日本国債取引問題への応用例を示し，複利型強化学習の有効性を確認する．

Compound Reinforcement Learning: Framework and Application

Tohgoroh Matsui,[†1] Takashi Goto,[†2]
Kiyoshi Izumi[†3,†4] and Yu Chen[†3]

This paper describes an extended framework of reinforcement learning, called compound reinforcement learning, which maximizes the compound return and shows its application to financial tasks. Compound reinforcement learning is designed for return-based MDP in which an agent observes the return instead of the rewards. We introduce double exponential discounting and betting fraction into the framework and then we can write the logarithm of double-exponentially discounted compound return as the sum of a polynomially discounted logarithm of simple gross return. In this paper, we show two algorithms of compound reinforcement learning: compound Q-learning and compound OnPS. We also show the experimental results using 3-armed bandit and an application to a financial task: Japanese government bond trading.

1. はじめに

強化学習[16] 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を通じて学習する枠組みとして定式化されている．

N 本腕バンディット問題は，強化学習の教科書[16] で強化学習の枠組みを説明するために用いられているシンプルな例題である．それぞれ払い戻し金とその確率が異なる N 個のホイールを持つマネー・ホイールがあり，それぞれのホイールを回すためのアーム（腕）が 1 つずつ付いている．エージェントは，どのアーム——つまり，どのホイールを選択するのが最も良いかを学習する．

ここで，**図 1** のようなホイール A，B を持つ 2 本腕バンディット問題を考えよう．ホイール上の金額は，賭け金 1 ドルあたりの払い戻し金である．最初に 100 ドル持っていて，このゲームに 1 ドルずつ 100 回連続して賭ける場合には，A を選択する方が払い戻し金が多くなると期待できる．なぜなら，A の払い戻し金の期待値は 1.5 ドルであり，B の払い戻し金の期待値は 1.25 ドルだからである．実際にこの賭けを行ったときの資産総額の推移の例を**図 2** に示す．従来の強化学習は，これと同じように考えて学習を行い，A を選択することを最適とする．

しかしながら，資産をすべて賭ける場合には，A を選択することは最適ではない．なぜな

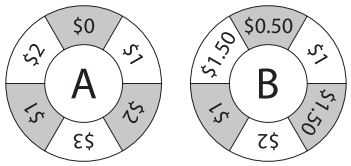

図 1　2 本腕バンディット問題
Fig. 1　2-armed bandit problem.

†1 中部大学
Chubu University
†2 三菱東京 UFJ 銀行
Bank of Tokyo-Mitsubishi UFJ, Ltd.
†3 東京大学
The University of Tokyo
†4 JST さきがけ
JST PRESTO

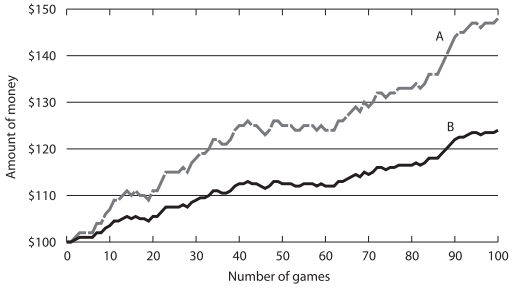

図 2　100 ドルの財産を 1 ドルずついずれか一方に賭け続けたときの資産総額の推移の例
Fig. 2　Examples of betting $1 each.

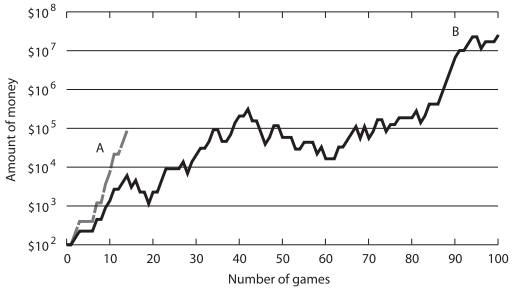

図 3　100 ドルの財産を全額いずれか一方に賭け続けたときの資産総額の推移の例.
Fig. 3　Examples of betting all money.

ら，A の払い戻し金の幾何期待値（幾何平均）は 0 ドルであり，長期的に観るといつかは払い戻し金が 0 になってすべての資産を失ってしまうからである．資産すべてを 100 回連続して賭けたときの資産総額の推移の例を**図 3** に示す．図 3 の A の財産曲線が途中で止まっているのは，ここですべての資産を失って賭けが続行できなくなったことを示している．一方で B の払い戻し金の幾何期待値は約 1.14 ドルであり，この賭けの最終的な資産の期待値は約 7,400 万ドルにもなる．

このように，払い戻し金を賭け金に上乗せする，すなわち複利式のリターンを考える場合には，従来の強化学習のような期待割引収益の最大化は意味をなさない．「(無分配型の）投資信託を選択する際にリターンの算術平均ではなくリターンの幾何平均が高い商品を選ぶべきである」というのは，ファイナンスの分野では一般的な考え方である[13]．したがって，このような場合には，報酬の代わりに複利リターンに基づいて学習するべきである．

本論文では，まず，複利リターンを最大化するための強化学習の枠組みである複利型強化学習[8] について述べる．複利型強化学習では，二重指数的割引と投資比率の概念を導入し，対数をとることによって従来の強化学習と同様の方法で割引複利リターンを最大化する．続いて，従来の強化学習アルゴリズムである Q 学習と OnPS を複利型に拡張した複利型 Q 学習と複利型 OnPS のアルゴリズムを示す．また，3 本腕バンディット問題に対する実験結果と日本国債取引問題への応用例を示し，複利型強化学習の有効性を確認する．

2. 複利型強化学習の枠組み

ファイナンスの分野では，リターンの算術平均よりもリターンの幾何平均——すなわち，複利リターンが重視される．そこで，複利型強化学習では，割引収益の期待値を最大化する代わりに，複利リターンの期待値を最大化する．

2.1 複利リターン

時刻 t における資産の価格を P_t とすると，この資産を時刻 t から時刻 $t+1$ まで保持したときの**リターン** R_{t+1} は次のように計算される．

$$R_{t+1} = \frac{P_{t+1} - P_t}{P_t} = \frac{P_{t+1}}{P_t} - 1 \tag{1}$$

また，$1 + R_{t+1}$ を**グロス・リターン**という．このとき，**複利リターン**は次式のように定義される[3]．

$$\rho_t = (1 + R_{t+1})(1 + R_{t+2}) \cdots (1 + R_T) \tag{2}$$

ここで，T は資産を保有していた最終時刻を表す．強化学習の連続型タスクのため，本論文では次のように T を無限大とする．

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

複利型強化学習は，MDP におけるリターン R_{t+k+1} がマルコフ性を持つ確率変数である**リターン型 MDP** を対象とする．

2.2 二重指数的割引

複利リターンに対し，従来の強化学習と同様に，割引の概念を導入する．従来の強化学習では，$\gamma^k r_{t+k+1}$ というように，k ステップ後の報酬 r_{t+k+1} に対して指数的に割り引いた重み γ^k を掛けることによって将来の報酬を割り引く．この割引は，行動経済学の分野では指

数的割引と呼ばれている．複利型強化学習では，従来の強化学習が用いている指数的割引の代わりに，$(1+R_{t+k+1})^{\gamma^k}$ というように，k ステップ後のグロス・リターン $1+R_{t+k+1}$ に対して指数的に割り引いた重み γ^k をべき乗することによって将来のリターンを割り引く．この割引は，二重指数関数 $f(x)=a^{b^x}$ の形になっていることから，本論文ではこれを二重指数的割引と呼ぶ．

複利型強化学習では，グロス・リターンを二重指数的に割り引いた複利リターン

$$\begin{aligned}\rho_t &= (1+R_{t+1})(1+R_{t+2})^{\gamma}(1+R_{t+3})^{\gamma^2}\cdots \\ &= \prod_{k=0}^{\infty}(1+R_{t+k+1})^{\gamma^k} \end{aligned} \tag{4}$$

を**割引複利リターン**と呼び，これを最大化することを考える．二重指数的に割り引かれるグロス・リターンは，遠い将来のリターンほど割引リターンを 0 に近づけることに相当し，ファイナンスにおけるリスク調整後リターンの一種と考えることができる．

二重指数的に割り引くことによって，割引複利リターンの対数を

$$\begin{aligned}\log\rho_t &= \log\prod_{k=0}^{\infty}(1+R_{t+k+1})^{\gamma^k} \\ &= \sum_{k=0}^{\infty}\log(1+R_{t+k+1})^{\gamma^k} \\ &= \sum_{k=0}^{\infty}\gamma^k\log(1+R_{t+k+1}) \\ &= \log(1+R_{t+1})+\gamma\sum_{k=0}^{\infty}\gamma^k\log(1+R_{t+k+2}) \\ &= \log(1+R_{t+1})+\gamma\log\rho_{t+1} \end{aligned} \tag{5}$$

というように，従来の強化学習における指数的な割引収益と同様に，再帰的に表すことができる．複利型強化学習では，これを**対数割引複利リターン**と呼び，この対数割引複利リターンの期待値を最大化する．

2.3 投資比率

グロス・リターンの対数 $\log(1+R_{t+1})$ の値は，$R_{t+1}=-1$ のときに $-\infty$ となってしまうため，対数割引複利リターンは発散してしまう可能性がある．そこで，複利型強化学習では，**投資比率**が割引率と同様にタスクによって決められているものとする．投資比率は，保有資産のうち実際に投資する資産の割合を表すもので，ファイナンスの分野では過剰投資を避けるために用いられている．投資比率が f のときのリターンは $R_{t+1}f$ であり，グロス・リターンは $1+R_{t+1}f$ となる．$R_{t+1}f>-1$ となるように投資比率 f を設定することにより，破産して対数割引複利リターンが発散してしまうことを回避することができる．

投資比率が f のときの割引複利リターンは次のように表される．

$$\rho_t = \prod_{k=0}^{\infty}(1+R_{t+k+1}f)^{\gamma^k} \tag{6}$$

また，投資比率が f のときの対数割引複利リターンは次のように表される．

$$\log\rho_t = \sum_{k=0}^{\infty}\gamma^k\log\left(1+R_{t+k+1}f\right) \tag{7}$$

この式 (7) の右辺は，従来の強化学習の割引収益を表す式の報酬 r_{t+1} を投資比率 f のときのグロス・リターンの対数 $\log(1+R_{t+1}f)$ に置き換えたものに等しい．

2.4 価値関数と最適価値関数

行動規則 π の下での状態 s の価値 $V^\pi(s)$ は，対数割引複利リターンの期待値として次のように定義される．

$$\begin{aligned}V^\pi(s) &= \mathrm{E}_\pi\left[\log\rho_t \middle| s_t=s\right] \\ &= \mathrm{E}_\pi\left[\sum_{k=0}^{\infty}\gamma^k\log(1+R_{t+k+1}f)\middle| s_t=s\right]\end{aligned}$$

この式は，従来の強化学習と同様にして，次のように書くことができる．

$$\begin{aligned}&= \mathrm{E}_\pi\left[\log(1+R_{t+1}f)+\gamma\sum_{k=0}^{\infty}\gamma^k\log(1+R_{t+k+2}f)\middle| s_t=s\right] \\ &= \sum_{a\in\mathcal{A}(s)}\pi(s,a)\sum_{s'\in\mathcal{S}}\mathcal{P}^a_{ss'}\left(\mathsf{R}^a_{ss'}+\gamma\mathrm{E}_\pi\left[\sum_{k=0}^{\infty}\gamma^k\log(1+R_{t+k+2}f)\middle| s_{t+1}=s'\right]\right) \\ &= \sum_{a\in\mathcal{A}(s)}\pi(s,a)\sum_{s'\in\mathcal{S}}\mathcal{P}^a_{ss'}\left(\mathsf{R}^a_{ss'}+\gamma V^\pi(s')\right)\end{aligned} \tag{8}$$

ここで，$\pi(s,a)$ は行動選択確率，$\mathcal{P}^a_{ss'}$ は状態遷移確率，f は投資比率，γ は割引率，$\mathsf{R}^a_{ss'}$ は投資比率 f のときのグロス・リターンの対数の期待値，すなわち

$$\mathsf{R}^a_{ss'} = \mathrm{E}\left[\log(1+R_{t+1}f)\middle|\, s_t=s, a_t=a, s_{t+1}=s'\right] \tag{9}$$

である．

同様に，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

$$\begin{aligned} Q^\pi(s,a) &= \mathrm{E}_\pi\left[\log\rho_t\middle|\, s_t=s, a_t=a\right] \\ &= \sum_{s'\in\mathcal{S}} \mathcal{P}^a_{ss'}\left(\mathsf{R}^a_{ss'} + \gamma V^\pi(s')\right) \end{aligned} \tag{10}$$

と表され，最適価値関数は

$$\begin{aligned} Q^*(s,a) &= \max_{\pi\in\Pi} Q^\pi(s,a) \\ &= \mathrm{E}\left[\log(1+R_{t+1}f) + \gamma\max_{a'} Q^*(s_{t+1},a')\middle| s_t=s, a_t=a\right] \\ &= \sum_{s'} \mathcal{P}^a_{ss'}\left(\mathsf{R}^a_{ss'} + \gamma\max_{a'} Q^*(s',a')\right) \end{aligned} \tag{11}$$

と表される．これが複利型強化学習における最適行動価値関数 Q^* の Bellman 方程式である．この式は，従来の強化学習のための Q^* の Bellman 方程式における獲得報酬の期待値 $\mathcal{R}^a_{ss'}$ を投資比率 f のときの対数グロス・リターンの期待値 $\mathsf{R}^a_{ss'}$ に置き換えたものに等しい．

3. 複利型強化学習のアルゴリズム

3.1 複利型 Q 学習

上に述べたように，式 (11) に示した複利型強化学習における最適行動価値 Q^* に関する Bellman 方程式は，従来の強化学習における Bellman 方程式の期待報酬 $\mathcal{R}^a_{ss'}$ を期待対数グロス・リターン $\mathsf{R}^a_{ss'}$ に置き換えたものに等しい．

したがって，式 (11) の Q^* を推定するには従来の Q 学習の報酬 r_{t+1} を対数グロス・リターン $\log(1+R_{t+1}f)$ に置き換えればいい．すなわち，状態 s_t において行動 a_t を実行した後に状態 s_{t+1} に遷移してリターン R_{t+1} を受け取ったとき，Q の値を次のように更新する．

$$Q(s_t,a_t) \leftarrow Q(s_t,a_t) + \alpha\left(\log(1+R_{t+1}f) + \gamma\max_{a\in\mathcal{A}} Q(s_{t+1},a) - Q(s_t,a_t)\right) \tag{12}$$

ここで，α はステップ・サイズ，γ は割引率，f は投資比率をそれぞれ表すパラメータである．

複利型 Q 学習のアルゴリズムを**図 4** に示す．従来の Q 学習と異なるのは，(i) 報酬 r の代わりにリターン R を観測し，(ii) 更新式の報酬 r を対数グロス・リターン $\log(1+Rf)$ に置き換えている点である．

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返し）
　s を初期化
　repeat（エピソードの各ステップに対して繰返し）
　　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返しを終了
end loop

図 4　複利型 Q 学習アルゴリズム
Fig. 4　Compound Q-learning algorithm.

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．リターン型 MDP においては，リターン R_{t+1} がマルコフ性を満たすので，$\log(1+R_{t+1}f)$ もマルコフ性を満たす．したがって，複利型強化学習における $\log(1+R_{t+1}f)$ を従来の強化学習における報酬 r_{t+1} と考えれば，リターン型 MDP における複利型 Q 学習の行動価値 Q は最適行動価値 Q^* に近づく．

ただし，Watkins と Dayan による報酬型 MDP における Q 学習の収束性についての証明[19] では，報酬が有界である——という条件が付いている．複利型強化学習においては，リターン型 MDP において $\log(1+R_{t+1}f)$ が有界，すなわち，$R_{t+1}f$ が -1 より大きく，かつ，上界を持つことが条件となる．以上のことから，次の定理が導ける．

定理 1. $-1 < \mathrm{R}_{\min} \le R_t f \le \mathrm{R}_{\max}$ を満たすリターン R_t，投資比率 f，$0 \le \alpha_t < 1$，かつ，

$$\sum_{i=1}^{\infty} \alpha_{t^i} = \infty, \quad \sum_{i=1}^{\infty} [\alpha_{t^i}]^2 < \infty, \tag{13}$$

を満たすステップ・サイズが与えられたとき，複利型 Q 学習において，確率 1 で $\forall s$, $a[Q_t(s,a) \to Q^*(s,a)]$ が成り立つ．ここで，$\mathrm{R}_{\min}$ は $R_t f$ の下界，$\mathrm{R}_{\max}$ は $R_t f$ の上界である．

Proof. $r_{t+1} = \log(1 + R_{t+1})$ とおくと，式 (12) に示された複利型 Q 学習の更新式は，従来の Q 学習の更新式に等しい．$R_t f$ が有界であるから，$\log(1 + R_t f)$ は有界である．また，リターン型 MDP において $\log(1 + R_t f)$ はマルコフ性を満たす．したがって，Watkins と Dayan の証明において r_{t+1} を $\log(1 + R_{t+1} f)$ に置き換えることによって，定理 1 を証明できる．□

3.2 複利型 OnPS

Profit sharing [11] は，それぞれの行動に優先度を割り当てるタイプの強化学習法であり，それぞれの状態ごとに行動の優先度を学習する．状態 s における行動 a の優先度を $P(s, a)$ とし，エピソードに含まれるすべての状態行動対 s_t，a_t に対する優先度をエピソード終了後に一括して次のように更新する．

$$P(s_t, a_t) \leftarrow P(s_t, a_t) + g(t, r_T, T) \tag{14}$$

ここで，T はエピソードの最終ステップ，g は信用割当関数である．信用割当関数には次の等比減少関数がよく用いられている．

$$g(t, r_T, T) = \gamma^{T-t-1} r_T \tag{15}$$

エピソードごとに一括更新を行う従来の profit sharing をステップごとの更新ができるように改良したものが OnPS [10] である．

OnPS は，信用トレースと呼ばれる技法を用いて，ステップごとに訪れた状態行動対の信用トレース c の値を 1 増やし，行動優先度と信用トレースを次のように更新する．

$$c(s_t, a_t) \leftarrow c(s_t, a_t) + 1 \tag{16}$$

$$P(s, a) \leftarrow P(s, a) + \alpha r_{t+1} c(s, a) \quad \text{for all } s, a \tag{17}$$

$$c(s, a) \leftarrow \gamma c(s, a) \quad \text{for all } s, a \tag{18}$$

本論文では，複利型 Q 学習と同様にして，OnPS を複利型に拡張する．複利型 OnPS は，報酬 r_{t+1} の代わりに対数グロス・リターン $\log(1 + R_{t+1} f)$ を用いてすべての状態行動対の優先度 P を更新する．すなわち，OnPS の式 (17) を次の式に置き換える．

$$P(s, a) \leftarrow P(s, a) + \alpha \log(1 + R_{t+1} f) c(s, a) \tag{19}$$

このアルゴリズムを，**図 5** に示す．従来の OnPS アルゴリズムと異なるのは，(i) 報酬 r を観測する代わりにリターン R を観測している点と，(ii) 報酬 r を用いて行動優先度 P を更新する代わりに対数グロス・リターン $\log(1 + Rf)$ を用いて行動優先度 P を更新している点の 2 点である．この違いは，複利型 Q 学習と従来の Q 学習の間の違いと同じである．

エピソード型タスクにおいて OnPS は従来の profit sharing に等価[10] なので，OnPS はエピソード型タスクにおいて宮崎らによる profit sharing の合理性定理[11] を満たすことができる．しかしながら，宮崎らの合理性定理では報酬を $r_T > 0$ と仮定しているのに対し，対数グロス・リターンは $\log(1 + R_T f) > 0$ を満たすとは限らないため，複利型 OnPS は宮崎らの合理性定理を満たさないことに注意が必要である．

入力： 割引率 γ, ステップ・サイズ α, 投資比率 f, 初期優先度 c
for all s, a **do**
  $P(s, a) \leftarrow c$
end for
loop （各エピソードに対して繰返し）
  s を初期化
  for all s, a **do**
    $c(s, a) \leftarrow 0$
  end for
  repeat （エピソードの各ステップに対して繰返し）
    P 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
    $c(s, a) \leftarrow c(s, a) + 1$
    行動 a を実行し，リターン R と次の状態 s' を観測
    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

図 5　複利型 OnPS アルゴリズム
Fig. 5　Compound OnPS algorithm.

4. 実験結果

図 6 に示された 3 つのホイールを持つバンディット問題を用いて，複利型 Q 学習（Compound）と従来の Q 学習（Simple），リスク回避型強化学習の 1 つである文献 15) に基づく分散ペナルティ型 Q 学習（Variance-Penalized）の比較を行った．ホイール A は，払い戻し金の算術期待値が最も大きいが，幾何期待値は最も小さい．払い戻し金の幾何期待値が

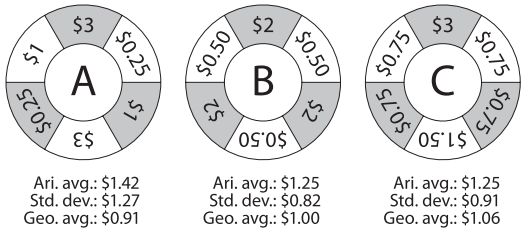

図 6　3 本腕バンディット問題
Fig. 6　3-armed bandit problem.

最も大きいのは，ホイール C である．ホイール B は，払い戻し金の算術期待値がホイール C と同じであるが，その標準偏差はホイール C よりも小さい．

従来の強化学習では払い戻し金から出資金を引いた値が報酬となり，複利型強化学習では払い戻し金を出資金で割った値から 1 を引いた値がリターンとなる．したがって，この問題ではエージェントが受け取る報酬とリターンは等しい．割引率と投資比率はいずれも $\gamma = 0.9$，$f = 1$ とした．

実験では，ランダム・シードを変えて 100 回の学習を行い，その平均を求めた．このとき，それぞれの評価値は学習とは独立に 100 回の試行を行うことによって求めた．学習中は $\epsilon = 0.2$ の ϵ-グリーディー選択を用い，評価時は最も価値が高い行動を選択するグリーディー選択を用いた．また，ステップ・サイズを $\alpha = 0.001$，分散ペナルティ型 Q 学習のリスク・パラメータを $\kappa = 1$ とした．これらのパラメータと行動選択法は，予備実験を行って経験的に定めた．

結果を**図 7** に示す．横軸は学習ステップ数，縦軸は，それぞれ，獲得リターン（報酬）の幾何平均と算術平均を表している．従来の Q 学習は，算術平均リターンは大きいが幾何平均リターンが小さい行動規則——すわなち A を選択する行動規則を学習した．分散ペナルティ型 Q 学習は，従来の Q 学習が学習した幾何平均が小さい行動規則を学習してしまうことは回避できたが，幾何平均リターンが最大の行動規則——すなわち C を選択する行動規則を学習することはできなかった．これに対し，複利型 Q 学習は，幾何平均リターンが最大の行動規則を学習することができた．

5. 日本国債取引への応用

続いて，複利型強化学習をより実際的な問題である日本国債取引問題[9] に適用した．国債市場は，株式市場のように個別企業のファンダメンタルズのようなミクロの影響を受けにくく，テクニカル分析に基づいた取引戦略を策定しやすい．Matsui ら[9] は，残存期間 10 年の日本国債の週次取引を対象として，OnPS を用いて取引戦略を獲得する手法を提案している．本論文では，このタスクに複利型 OnPS（Compound）を適用し，従来の OnPS（Simple）と比較する．

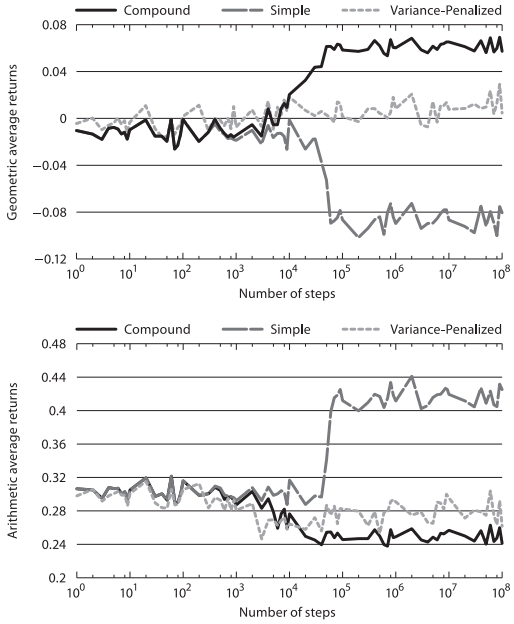

図 7　3 本腕バンディット問題の実験結果．上が幾何平均，下が算術平均
Fig. 7　Experimental results for 3-armed bandit.

このタスクでは，状態は，金利と 14 週移動標準偏差の対 (y_t, σ_t) によって表される．エージェントは，この順序対が一定の範囲に収まるよう 14 週相対的観測値として観測する．エージェントがとりうる行動は，買いまたは売りのいずれかである．割引率と投資比率は，従来の OnPS と複利型 OnPS のいずれも $\gamma = 0.9$，$f = 1$ とした．連続的な観測値を扱うため，半径 1/14 の RBF 特徴を格子状に 15×15 個配置して線形関数近似を行った．行動選択には Gibbs ソフトマックス選択を用い，学習時は $\tau = 0.2$，評価時は $\tau = 0.1$ とした．

実験は，学習期間を 3 年とし，評価期間をその直後の 1 四半期として行った．たとえば，2004 年 1 月 1 日から 2006 年 12 月 31 日のデータを用いて学習を行い，2007 年 1 月 1 日

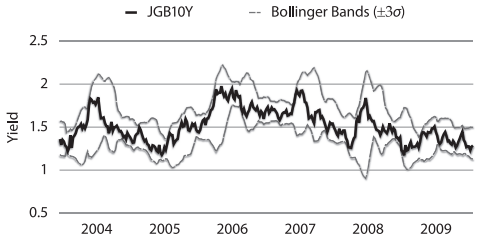

図 8　2004 年から 2009 年までの残存期間 10 年日本国債の金利とそのボリンジャー・バンド（$\pm 3\sigma$）
Fig. 8　Yields of 10-year Japanese government bond (JGB).

から 2007 年 3 月 31 日のデータに対して学習した取引戦略を用いて運用したときの収益を評価した．これを，期間を四半期ずつずらしながら 12 回行い，3 年間の累積収益を求めた．また，乱数のシードを変えて 10 回の実験を行った．

学習期間および評価期間の対象となった 2004 年から 2009 年にかけての残存期間 10 年の日本国債の金利の動きと $\pm 3\sigma$ のボリンジャー・バンド（移動標準偏差）を**図 8** に，実験結果を**図 9** に示す．上のグラフは，乱数のシードを変えて行った 10 回の実験の平均累積収益を表している．下のグラフは，12 期すべてにおける最良のケースと最悪のケースの累積収益を表している．

累積収益のいずれも，複利型 OnPS の収益は従来の OnPS の収益よりも高くなった．中央累積収益において最も大きな差が生じたのが 2008 年第 2 四半期であった．これはサブプライム・ローン問題によってリーマン・ブラザーズが破綻した 2008 年 9 月 15 日の直前の期間である．

6. 考察と関連研究

複利型強化学習では，リターン R が対数グロス・リターン $\log(1+Rf)$ に変換され，強化信号として用いられる．この関係を**図 10** に示す．左のグラフは投資比率が $f=1$ のとき，右のグラフは $f=0.5$ のときの強化信号を比較したものである．同じ投資比率を導入した従来の強化学習（Simple）と比較すると，この対数変換によって正の強化信号は抑制され，負の強化信号は増強される．結果として，正のリターンに対してはリスク追求型となり，負のリターンに対してはリスク回避型となる$^{\star 1}$．

$\star 1$ ここでは，「リスク」を期待効用理論[14),18)] の文脈で使用しており，ファイナンスやリスク回避型強化学習の文脈で用いられる「リスク」とは異なる．

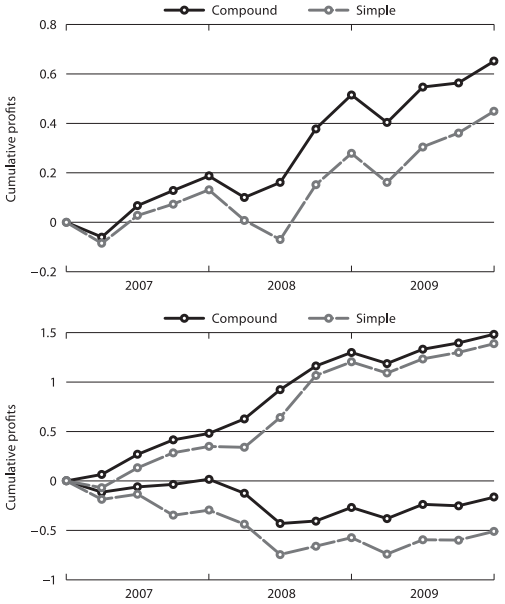

図 9　日本国債取引問題の実験結果．上が平均累積収益，下が累積収益の最良ケースと最悪ケース
Fig. 9　Experimental results for 10-year JGB.

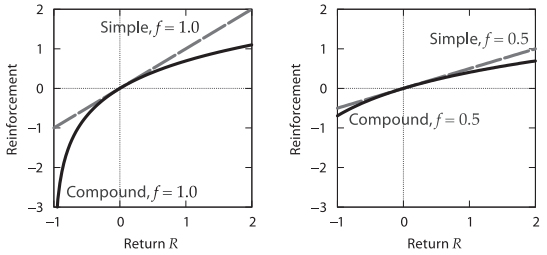

図 10　リターンと強化信号の関係
Fig. 10　Relations between returns and reinforcements.

複利型強化学習では負の強化信号が増強されるため，なるべくリターンが負にならないようにする取引戦略を獲得することが期待できる．この効果は，日本国債取引問題の結果に見ることができる．各期間の年換算収益の平均 μ と標準偏差 σ は，複利型 OnPS では $\mu=0.217$，$\sigma=0.510$ であったのに対し，従来の OnPS では $\mu=0.150$，$\sigma=0.578$ であっ

た．すなわち，複利型 OnPS による収益は従来の OnPS による収益に比べて平均が高く分散が小さかった．また，最良ケースと最悪ケースを比較すると，最良ケースの収益には大きな差は生じなかったが最悪ケースの収益に大きな差が生じた．これらの結果から，複利型 OnPS は大きな負のリターンを回避することで安定的に高い収益をあげる取引戦略を獲得できたと考えられる．

投資比率の概念は，ファイナンスの分野では過剰投資を避けるためのものとしてよく知られている．リターンの分布から求められる利益を最大化する投資比率[6] はケリー基準と呼ばれる．また，オプティマル f と呼ばれる過去のリターンからケリー基準を推定する方法[17] が提案されている．

複利型強化学習において，投資比率は割引率と同様にエージェントに対して外から与えられる．与えられる投資比率の値によって，複利型強化学習における最適な行動が異なる場合がある．投資比率が 1 に近いほど複利型強化学習と従来の強化学習の差異は大きくなり，投資比率が 0 に近いほどその差違は小さくなる．複利型強化学習は，複利リターンを最大化する割引率や投資比率を探すのではなく，与えられた割引率と投資比率の下で割引複利リターンを最大化する行動規則を学習する枠組みである．

一方，リスク回避型強化学習として，期待価値から分散を引くアプローチ[5),15] や望ましくない状態への到達確率をリスクと定義するアプローチ[4]，コスト関数を導入するアプローチ[1),2] などが提案されている．ファイナンスの分野では投資リスクはリターンの分散として定義されるため，期待価値から分散を引くアプローチはファイナンスへの応用に向いているように思われる．しかしながら，期待価値から分散を引くアプローチは，3 本腕バンディット問題においてホイール C よりもホイール B を良しとする．実際に，分散ペナルティ型 Q 学習は複利リターンが最大となるホイール C を選択する行動規則を学習できなかった．すなわち，リスク回避型強化学習では複利リターンを最大化することはできない．

強化学習をファイナンスに応用する論文は，これまでにもいくつか発表されている[1),7),12]．しかしながら，これらの論文では，複利リターンを最大化することは考えていない．複利型強化学習の最大の特徴は，複利リターンを最大化することである．

7. ま と め

本論文では，まず，複利型強化学習の枠組みについて述べた．複利型強化学習では，二重指数的な割引の概念を導入し，リターン型 MDP における割引複利リターンの対数の期待値を最大化する．対数をとることによって二重指数的割引複利リターンを指数的割引対数リターンの和に変換し，従来の強化学習のアルゴリズムを拡張することを可能としている．また，複利型 Q 学習と複利型 OnPS アルゴリズムを示し，実際に従来の強化学習アルゴリズムが容易に複利型に拡張できることを示した．さらに，バンディットの問題を用いた実験の結果と日本国債取引への応用を示すことより，複利型強化学習がファイナンスへの応用に有効であることを確認した．

複利型強化学習は複利リターンに基づいて学習することからファイナンスへの応用に適している．今後は，ファイナンス以外の分野において複利リターンを最大化することが有用な問題を調査し，応用範囲を広げていきたい．

謝辞　本研究は科研費（21013049，23700182）の助成を受けたものである．

参 考 文 献

1) Basu, A., Bhattacharyya, T. and Borkar, V.S.: A Learning Algorithm for Risk-Sensitive Cost, *Mathematics of Operations Research*, Vol.33, No.4, pp.880–898 (2008).
2) Borkar, V.S.: Q-Learning for Risk-Sensitive Control, *Mathematics of Operations Research*, Vol.27, No.2, pp.294–311 (2002).
3) Campbell, J.Y., Lo, A.W. and MacKinlay, A.G.: *The Econometrics of Financial Markets*, Princeton University Press (1997). 祝迫得夫，大橋和彦，中村信弘ほか（訳）：ファイナンスのための計量分析，共立出版 (2003).
4) Geibel, P. and Wysotzki, F.: Risk-Sensitive Reinforcement Learning Applied to Control under Constraints, *Journal of Artificial Intelligence Research*, Vol.24, pp.81–108 (2005).
5) Heger, M.: Consideration of Risk in Reinforcement Learning, *Proc. 11th International Conference on Machine Learning* (*ICML 1994*), pp.105–111 (1994).
6) Kelly, Jr., J.L.: A new interpretaion of information rate, *Bell System Technical Journal*, Vol.35, pp.917–926 (1956).
7) Li, J. and Chan, L.: Reward Adjustment Reinforcement Learning for Risk-averse Asset Allocation, *Proc. International Joint Conference on Neural Networks* (*IJCNN 2006*), pp.534–541 (2006).
8) 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).
9) 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).
10) Matsui, T., Inuzuka, N. and Seki, H.: On-Line Profit Sharing Works Efficiently, *Proc. 7th International Conference on Knowledge-Based Intelligent Information & Engineering Systems* (*KES 2003*), pp.317–324 (2003).

11) 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，人工知能学会誌，Vol.9, No.4, pp.580–587 (1994).
12) Nevmyvaka, Y., Feng, Y. and Kearns, M.: Reinforcement Learning for Optimized Trade Execution, *Proc. 23rd International Conference on Machine Learning* (*ICML 2006*), pp.673–680 (2006).
13) Poundstone, W.: *Fortune's Formula: The untold story of the scientific betting system that beat the casinos and wall street*, Hill and Wang (2005). 松浦俊輔（訳）：天才数学者はこう賭ける—だれも語らなかった株とギャンブルの話，青土社 (2006).
14) Russell, S. and Norvig, P.: *Artificial Intelligence: A Modern Approach*, 2nd edition, Prentice Hall (2003). 古川康一（監訳）：エージェントアプローチ 人工知能，第 2 版，共立出版 (2008).
15) Sato, M. and Kobayashi, S.: Average-Reward Reinforcement Learning for Variance Penalized Markov Decision Problems, *Proc. 18th International Conference on Machine Learning* (*ICML 2001*), pp.473–480 (2001).
16) 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).
17) Vince, R.: Find your optimal f, *Technical Analysis of Stock & Commodities*, Vol.8, No.12, pp.476–477 (1990).
18) Von Neumann, J. and Morgenstern, O.: *Theory of Games and Economic Behavior*, Princeton University Press (1944).
19) Watkins, C.J.C.H. and Dayan, P.: Technical Note: Q-Learning, *Machine Learning*, Vol.8, No.3/4, pp.279–292 (1992).

(平成 23 年 4 月 11 日受付)
(平成 23 年 9 月 12 日採録)

松井藤五郎（正会員）

中部大学生命健康科学部臨床工学科兼工学部情報工学科講師．1997 年名古屋工業大学工学部知能情報システム学科卒業．2003 年名古屋工業大学大学院工学研究科電気情報工学専攻博士課程修了，博士（工学）．2003 年東京理科大学理工学部経営工学科助手，2007 年同助教．2009 年とうごろう機械学習研究所を設立．2010 年より現職．機械学習およびデータ・マイニングに関する研究に従事．人工知能学会，AAAI，ACM 各会員．

後藤　　卓

三菱東京 UFJ 銀行融資企画部．1997 年名古屋大学工学部情報工学科卒業．同年株式会社東海銀行（現，株式会社三菱東京 UFJ 銀行）入社，2010 年より現職．1998 年より ALM および債券運用業務に従事し，2001 年から 2007 年まで日本および英国にてプロップ・トレーディング業務に従事．帰国後，円貨資金証券部，市場企画部を経て現在に至る．日本証券アナリスト協会検定会員．

和泉　　潔（正会員）

東京大学大学院工学系研究科准教授．1993 年東京大学教養学部基礎科学科第二卒業．1998 年同大学大学院博士課程修了．博士（学術）．同年より 2010 年まで，電子技術総合研究所（現，産業技術総合研究所）勤務．2010 年より現職．マルチエージェントシミュレーション，特に社会シミュレーションに興味がある．人工知能学会，電子情報通信学会，電気学会各会員．

陳　　ユ

東京大学大学院新領域創成科学研究科准教授．1989 年上海交通大学熱工学科卒業．1994 年東京大学大学院工学系研究科システム量子工学博士課程修了，工学博士．同年東京大学工学部助手．その後，同講師，助教授，同大学院情報学環助教授，同工学系研究科准教授を経て，2011 年より現職．複雑系のシミュレーションに関する研究に従事．日本原子力学会，応用物理学会各会員．