ファイナンスのための強化学習

Reinforcement Learning for Finance

松井藤五郎*
Tohgoroh Matsui

とうごろう機械学習研究所
Tohgoroh Machine Learning Research Institute

Abstract: This paper describes a framework of reinforcement learning for finance. I propose a new reinforcement learning algorithm based on Q-learning. I show the experimental results using *N*-arms bandit and gridworld problems.

1 はじめに

強化学習 [3] 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 [3] 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ビッグ・シックス・ホイール・マシンがあり，それぞれのホイールを回すためのアーム（腕）がついている．このとき，どのようにホイールを選択するのが最も良いかを学習する．

ここで，1 ドル当たりの払い戻し金が図 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．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，ホイール 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，ホイール A の払い戻し金の期待値は 1.5 ドルであり，ホイール B の払い戻し金の期待値は 1.25 ドルでるからだ．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，ホイール A を選択することを最適とする．

しかしながら，資産を全て賭ける場合にはホイール A は最適ではない．なぜなら，ホイール 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 になって全ての資産を失ってしまうからだ．資産全てを 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 のホイール A

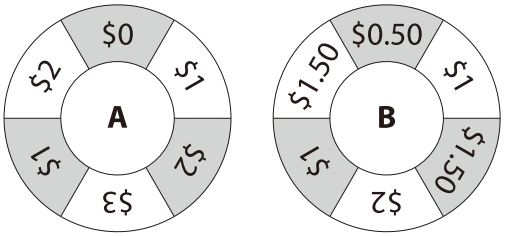

図 1　2 つのホイールを持つバンディット問題．

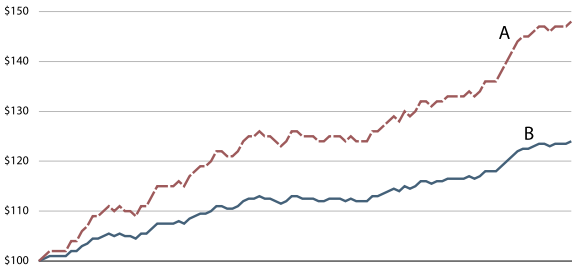

図 2　100 ドルの財産を 1 ドルずついずれか一方に賭け続けたときの資産総額の推移の例．

の財産曲線が途中で止まっているのは，ここで全ての資産を失って賭けが続行できなくなったからである．一方でホイール 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常識的な

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

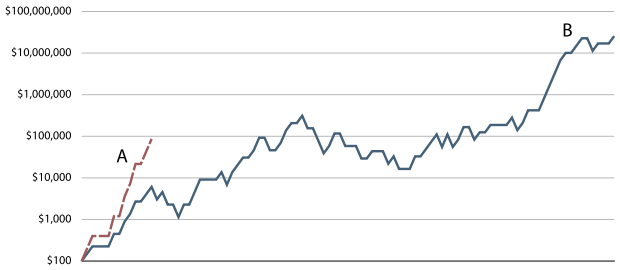

図 3　100 ドルの財産を全額いずれか一方に賭けた続けたときの資産総額の推移の例.

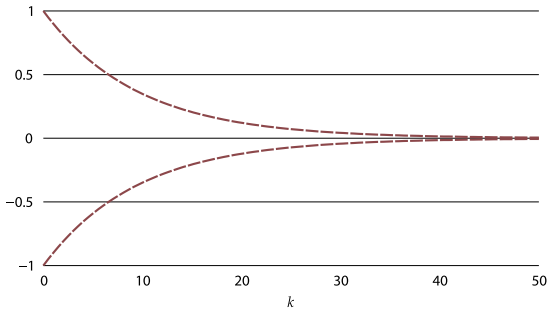

図 4　割引率 $\gamma = 0.9$ のときに ± 1 の報酬がそれぞれ割り引かれる様子.

考え方である [2]. したがって，このような場合には，報酬の替わりに複利式のリターンに基づいて学習するべきである.

そこで，本論文では，複利リターンに基づく強化学習の枠組みと学習アルゴリズムを提案する．また，実験により提案手法の有効性を示す.

2　従来の強化学習

2.1　枠組み

従来の強化学習は，次式で定義される割引収益の期待値を最大にするような行動規則を学習する.

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

ここで，r_{t+1} は報酬，γ は割引率と呼ばれる $0 \leq \gamma \leq 1$ のパラメーターである．報酬 r_{t+1} は資産の増加量に相当する，すなわち，時刻 t における資産の価値を P_t とするとき

$$r_{t+1} = P_{t+1} - P_t$$

と考えることができる.

割引収益は，将来にわたって受け取る収益を時刻が 1 経過するごとに γ 倍して合計したものである．$t = \infty$ までの合計を求めているが，r_t が有限で $\gamma < 1$ ならば割引収益の値は有限となる．例として，割引率が $\gamma = 0.9$ のときに ± 1 の報酬がそれぞれ割り引かれる様子を図 4 に示す．このように，将来に得られる報酬ほど 0 に近づくように割り引く.

行動規則 π を状態 $s \in \mathcal{S}$ において行動 $a \in \mathcal{A}$ を選択する確率 $\pi(s,a) = \Pr(a_t = a | s_t = s)$ と表す．このとき，行動規則 π の下での状態 s の価値 $V^\pi(s)$ は，行動規則 π に従って行動したときの期待割引収益として定義され，次のように表される.

$$V^\pi(s) = \mathrm{E}_\pi \left[\sum_{k=0}^{\infty} \gamma^k r_{t+k+1} \middle| s_t = s \right]$$

この価値関数は再帰的な形に書き直すことができる.

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

これは V^π の **Bellman** 方程式と呼ばれる．ここで，$\mathcal{P}^a_{ss'}$ は状態 s で行動 a を行ったときに状態 s' に遷移する確率，$\mathcal{R}^a_{ss'}$ は状態 s で行動 a を行って状態 s' に遷移したときに得られる報酬の期待値を表す.

最適な行動規則 π^* は価値関数を最大化する行動規則として定義され，最適価値関数は次のように定義される.

$$V^*(s) = \max_{\pi \in \Pi} V^\pi(s)$$

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

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

と表され，最適価値関数は

$$Q^*(s,a) = \max_{\pi \in \Pi} Q^\pi(s,a)$$

と定義される．この式は，次のように書くことができる.

$$\begin{aligned}
Q^*(s,a) &= \mathrm{E} \left[r_{t+1} + \gamma \max_{a' \in \mathcal{A}} Q^*(s_{t+1}, a') \middle| s_t = s, a_t = a \right] \\
&= \sum_{s' \in \mathcal{S}} \mathcal{P}^a_{ss'} \left(\mathcal{R}^a_{ss'} + \gamma \max_{a' \in \mathcal{A}} Q^*(s', a') \right) \qquad (2)
\end{aligned}$$

これが最適行動価値関数 Q^* の Bellman 方程式である.

2.2 Q 学習

Q 学習 [6] は，現在の行動規則にとは独立に Q^* を直接近似するタイプの **TD** 学習法であり，Q の値を次のように更新する．

$$Q(s_t,a_t) \leftarrow Q(s_t,a_t) + \alpha\Big(r_{t+1} + \gamma \max_{a\in\mathcal{A}} Q(s_{t+1},a) - Q(s_t,a_t)\Big)$$

ここで，α はステップ・サイズと呼ばれる $0 \le \alpha \le 1$ のパラメーターである．

Q 学習の最大の利点は，「Q 学習は MDP（マルコフ決定過程）環境において十分な試行錯誤の後に Q の値が必ず Q^* に近づく」ことが Watkins と Dayan [6] によって証明されている点である．

3 複利リターンに対する強化学習

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，本論文では，割引収益の期待値を最大化する替わりに，複利リターンの期待値を最大化することを考える．

複利リターンは次式のように表される．

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

ここで，R_t は時刻 t におけるリターンを表し，資産の価格 P_t が増加した割合を表す．すなわち，

$$R_t = \frac{P_{t+1}-P_t}{P_t} = \frac{r_{t+1}}{P_t}$$

である．また，$1+R_t$ をグロス・リターンという．

この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．ただし，従来の強化学習と同様にして将来のリターンを割り引くと，

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

となるが，本論文では，グロス・リターンを指数関数的に割り引いた複利リターン

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

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．

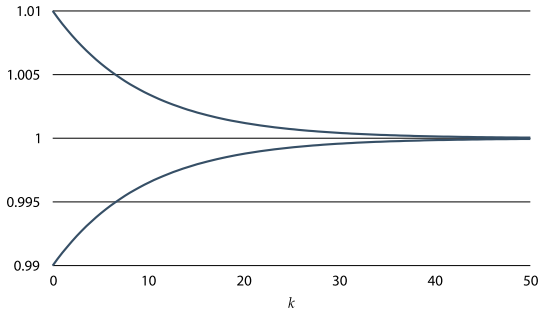

図 5　割引率 $\gamma = 0.9$ のときに ± 0.01 のリターンがそれぞれ指数関数的に割り引かれる様子．

指数関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

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

というように，従来の強化学習における割引収益と同様に，再帰的に表すことができる．本論文では，これを対数割引複利リターンと呼ぶ．割引複利リターンの期待値を最大化することは，この対数割引複利リターンの期待値を最大化することに等しい．

指数関数的にグロス・リターンを割り引くことは，遠い将来のグロス・リターンほど 1 に近づくようにしていることを意味する．例として，図 5 に ± 0.01 のリターン——すなわち，1.01 と 0.99 のグロス・リターンが割引率 $\gamma = 0.9$ で指数関数的に割り引かれる様子を示す．

ところが，グロス・リターンの対数 $\log(1+R_t)$ は，リターンが $R_t = -1$ 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，本論文では，投資比率の概念を導入する．投資比率は，保有資産のうち賭け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良く用いられている．投資比率が f のときのリターンは $R_t f$ であり，グロス・リターンは $1+R_t f$ となる．投資比率を 1 未満，すなわち $0 \le f < 1$ とすることにより，グロス・リターンの対数が $-\infty$ となっ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．

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

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．

$$\log \prod_{k=0}^{\infty} (1+R_{t+k+1}f)^{\gamma^k} = \sum_{k=0}^{\infty} \log (1+R_{t+k+1}f)^{\gamma^k} = \sum_{k=0}^{\infty} \gamma^k \log (1+R_{t+k+1}f) \quad (4)$$

この式 (4) は，従来の強化学習の割引収益を表す式 (1) の報酬 r_t を投資比率 f のときのグロス・リターンの対数 $\log(1+R_t f)$ に置き換えたものに等しい．

そこで，本論文では，行動規則 π の下での状態 s の価値 $V^\pi(s)$ を対数割引複利リターンの期待値として次のように定義し直す．

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

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．

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

ここで，f は投資比率，γ は割引率，$\mathsf{R}^a_{ss'}$ は状態 s で行動 a を行って状態 s' に遷移したときに得られるリターンの期待値である．

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

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

と表され，最適価値関数は

$$\begin{aligned} Q^*(s,a) =& \max_{\pi\in\Pi} Q^\pi(s,a) \\ =& \mathrm{E} \Bigg[\log(1+R_{t+1}f) \\ & \quad + \gamma \max_{a'\in\mathcal{A}} Q^*(s_{t+1},a') \Bigg| s_t = s, a_t = a \Bigg] \\ =& \sum_{s'\in\mathcal{S}} \mathcal{P}^a_{ss'} \left(\log(1+\mathsf{R}^a_{ss'}f) + \gamma \max_{a'\in\mathcal{A}} Q^*(s',a') \right) \end{aligned} \quad (5)$$

と表される．これが複利リターンに対する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，式 (2) に示された従来の強化学習における Q^* の Bellman 方程式の $\mathcal{R}^a_{ss'}$ を $\log(1+\mathsf{R}^a_{ss'}f)$ に置き換えたものに等しい．

4 複利リターンに対する Q 学習

4.1 アルゴリズム

上に述べたように，式 (5) に示した複利リターンのための強化学習における最適行動価値 Q^* に関する Bellman 方程式は，式 (2) の従来の強化学習における Bellman 方程式の $\mathcal{R}^a_{ss'}$ を $\log(1+\mathsf{R}^a_{ss'}f)$ に置き換えたものに等しい．

したがって，式 (5) の Q^* を推定するには従来の Q 学習の報酬 r_{t+1} を対数リターン $\log(1+R_{t+1}f)$ 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．

$$\begin{aligned} Q(s_t,a_t) \leftarrow Q(s_t,a_t) + \alpha \Big(& \log(1+R_{t+1}f) \\ & + \gamma \max_{a\in\mathcal{A}} Q(s_{t+1},a) - Q(s_t,a_t) \Big) \end{aligned}$$

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ーである．

複利リターンのための Q 学習のアルゴリズムを図 6 に示す．従来の Q 学習と異なるのは，(i) 報酬 r の替わりにリターン R を観測し，(ii) 更新式の報酬 r を投資比率 f のときのグロス・リターンの対数 $\log(1+Rf)$ に置き換えている点である．

4.2 最適解への収束性

複利リターンに対する Q 学習は，従来の Q 学習の報酬 r_{t+1} を対数リターン $\log(1+R_{t+1}f)$ に置き換えたものである．一方，複利リターンに対する最適行動価値関数 Q^* の Bellman 方程式も，従来の最適行動価値関数

1. $Q(s,a)$ を任意に初期化する
2. 各エピソードに対して繰り返し：
3. 　s を初期化する
4. 　エピソードの各ステップに対して繰り返し：
5. 　　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 を選択する
6. 　　行動 a を実行し，リターン R と次の状態 s' を観測する
7. 　　$Q(s,a) \leftarrow Q(s,a) + \alpha\left[\log(1+Rf) + \gamma \max_{a'} Q(s',a') - Q(s,a)\right]$
8. 　　$s \leftarrow s'$
9. 　s が終端状態ならば繰り返しを終了

図 6　複利リターンに対する Q 学習アルゴリズム

の Bellman 方程式における報酬の期待値 $\mathcal{R}^a_{ss'}$ を対数リターンの期待値 $\log(1+\mathrm{R}^a_{ss'})$ に置き換えたものである．したがって，複利リターンに対する強化学習における対数リターンを従来の強化学習における報酬と考えれば，Q 学習の行動価値関数 Q は最適行動価値関数 Q^* に近づく．

ただし，Watkins と Dayan の証明における強化学習の報酬には「報酬が有界である」という条件が付いている．したがって，複利リターンに対する強化学習においては，対数リターンが有界でなければならない．すなわち，Rf が -1 より大きく，かつ，上界を持つことが条件となる．

リターン R の最小値は -1 であるから，

1. 投資比率 f が 1 よりも小さいことと
2. リターン R が上界を持つこと

が複利リターンに対する Q 学習が最適解に収束するための条件である．

以上のことより，複利リターンに対する Q 学習に対して次の定理が成り立つ．

定理 1. $0 \leq f < 1$ の投資比率 f，$-1 \leq R_t \leq \mathrm{R}$ のリターン R，

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

のステップ・サイズ α_t が与えられたとき，複利リターンに対する Q 学習において $t \to \infty$ のとき確率 1 で次式が成り立つ．

$$\forall s,a[Q_t(s,a) \to Q^*(s,a)]$$

ここで，R は R_t の上界である．

Proof. $r_t = \log(1+R_t f)$ とおくと，複利リターンに対する Q 学習は報酬に対する従来の Q 学習 [6] と一致する．また，$0 \leq f < 1$ かつ $-1 \leq R_t \leq \mathrm{R}$ であることから r_t は有界である．

したがって，Watkins と Dayan の証明 [6] において $r_t = \log(1+R_t f)$ とおくことによって，定理 1 が証明される．□

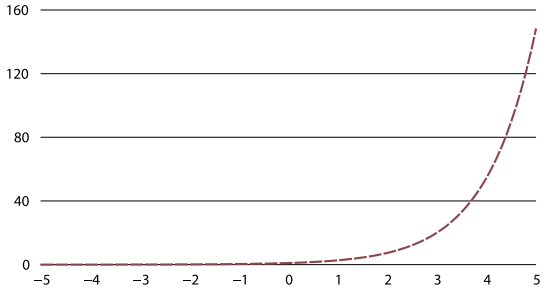

図 7　Gibbs 分布（$\tau = 1$）に用いられる指数関数

4.3　シグモイド・ソフトマックス行動選択法

複利リターンに対する Q 学習では，Q は有界ではあるものの極めて小さい値から極めて大きい値まで取りうる．Gibbs 分布は図 7 に示すように指数的に増加するため，複利リターンに対する Q 学習では，Gibbs 分布を用いたソフトマックス行動選択法

$$\Pr(a_t = a | s_t = s) = \frac{e^{Q(s,a)/\tau}}{\sum_{a' \in \mathcal{A}} e^{Q(s,a')/\tau}}$$

において温度パラメーター τ をどのように与えても適切な行動選択確率を導くことができない．なぜなら，従来の Gibbs 分布を用いたソフトマックス行動選択法は暗黙のうちに Q の絶対値の上限を 1 と仮定しているからである．

そこで，本論文では，シグモイド関数を利用して極めて小さい値から極めて大きい値まで取りうる Q に対し

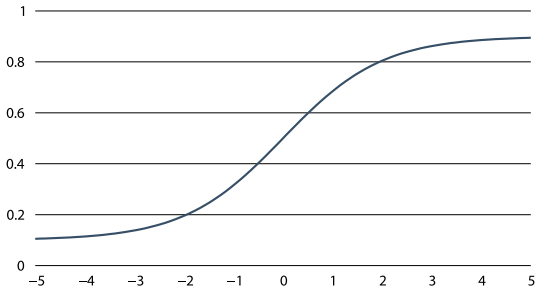

図 8　シグモイド関数（$g = 1, \epsilon = 0.1$）

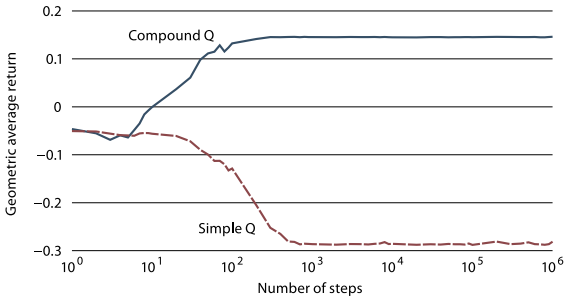

図 9　バンディットにおける幾何平均リターン

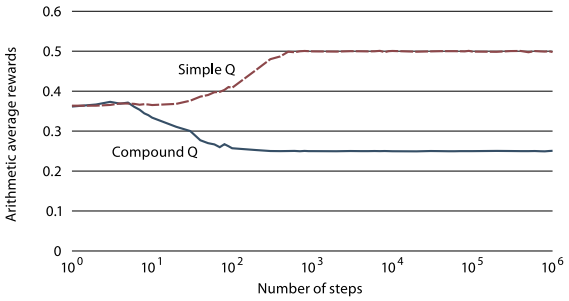

図 10　バンディットにおける算術平均報酬

ても適切な行動選択確率を導出するシグモイド・ソフトマックス行動選択法を提案する．本論文で用いた ϵ ロジスティック・シグモイド・ソフトマックス行動選択法における行動選択確率は次のように表される．

$$\Pr(a_t = a | s_t = s) = \frac{\frac{1-2\epsilon}{1+e^{-gQ(s,a)}} + \epsilon}{\sum_{a' \in \mathcal{A}} \frac{1-2\epsilon}{1+e^{-gQ(s,a')}} + \epsilon}$$

ここで，g はゲインと呼ばれるシグモイド曲線の傾きを決めるパラメーター，ϵ は最小優先度を決めるパラメーターである．例として，図 8 に $g = 1, \epsilon = 0.1$ のシグモイド曲線を示す．

5　実験結果

5.1　2 本腕バンディット問題

図 1 に示された 2 つのホイールを持つバンディット問題を用いて，従来の報酬に対する Q 学習と本論文で提案した複利リターンに対する Q 学習の比較を行った．

エージェントは 100 ドルの資金を持っており，100 回繰り返しホイールを選択して賭ける．従来の強化学習では 1 ドルずつ賭け，複利リターンに対する強化学習では保有資産の 99% を賭けるものとした．すなわち，投資比率 $f = 0.99$ とした．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，複利リターンに対する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．割引率はいずれも $\gamma = 0.9$ とした．

実験では，ランダム・シードを変えて 101 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 251 回の試行を行うことによって求めた．学習中は $g = 1.0, \epsilon = 0.1$ の ϵ ロジスティック・シグモイド・ソフトマックス選択を用い，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 $\alpha = 0.01$ とした．これらのパラメーターと行動選択法は，予備実験を行って経験的に定めた．

結果を図 9, 10 に示す．それぞれ，幾何平均リターンと算術平均報酬を表している．複利リターンに対する強化学習は正の幾何平均リターンを得られる行動規則，すなわち B を選択する行動規則を学習し，従来の強化学習は算術平均報酬がより大きい行動規則，すわなち A を選択する行動規則を学習した．

5.2　格子世界の問題

従来の強化学習で良く用いられる例題に「迷路を解く移動ロボット」がある．そこで，図 11 のような 5×5 の格子世界において，複利リターンに対する強化学習と従来の強化学習の違いを見てみよう．

S のマスが初期状態，G のマスが目標状態である．ロボットが取りうる行動は「東」「西」「南」「北」の 4 種類とする．上側の灰色のマスでは，ロボットが行動したときに 0.1 の確率で動けなくなってしまう．その他の行動遷移確率は全て 1 である．

したがって，最短経路を通ると 4 ステップで目標に到達できるが，途中で動けなくなってしまう可能性がある．目標に到達できる確率は $(1-0.1)^3 = 0.729$ である．一方，迂回経路は目標に到達するのに 12 ステップ

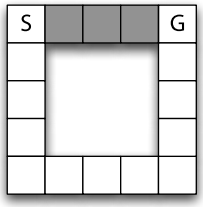

図 11　格子世界の問題．S は初期状態，G は目標状態を表す．上側の灰色のマスでは行動したときに 0.1 の確率でロボットが動けなくなってしまう．

を要するが，確実に目標に到達することができる．

従来の強化学習では，目標に到達したときに報酬 1，動けなくなったときは報酬 -1 がロボットに与えられる．その他の報酬は全て 0 である．複利リターンに対する強化学習においては，目標に到達したときのリターンを 1，動けなくなったときのリターンを -1 とする．その他のリターンはすべて 0 である．つまり，この問題でも報酬とリターンが等しい．割引率は $\gamma = 0.9$，投資比率は $f = 0.99$ とした．

結果を図 12 から図 15 に示す．それぞれ，幾何平均リターン，算術平均報酬，目標到達率，目標到達に要した平均ステップ数を表している．複利リターンに対する Q 学習は 12 ステップを要するが確実に目標に到達できる迂回経路を選択する行動規則を学習し，従来の Q 学習は 4 ステップしか要しないが目標に到達できる割合が 3/4 未満しかない最短経路を選択する行動規則を学習した．バンディット問題と同様に，獲得した行動規則によって得られる幾何平均リターンは，複利リターンに対する Q 学習の場合は正であるのに対し，従来の Q 学習の場合は負である．また，算術平均報酬は従来の Q 学習の方が高かった．

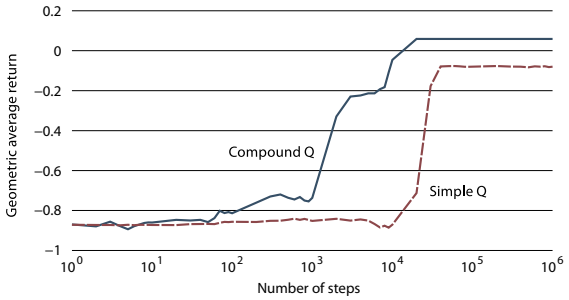

図 12　格子世界における幾何平均リターン

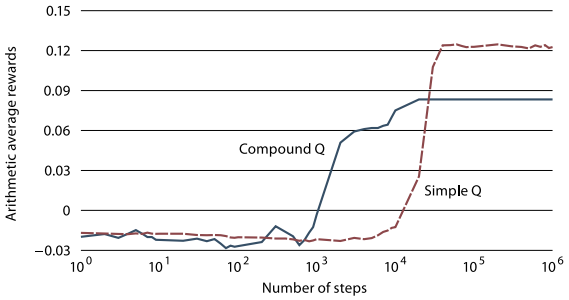

図 13　格子世界における算術平均報酬

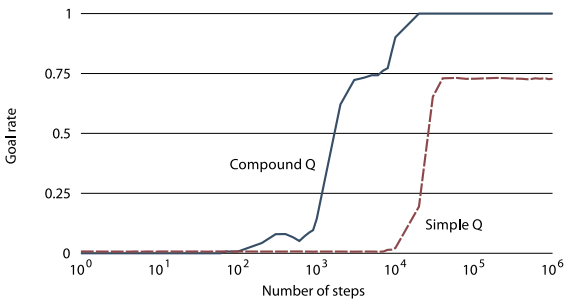

図 14　格子世界における目標到達率

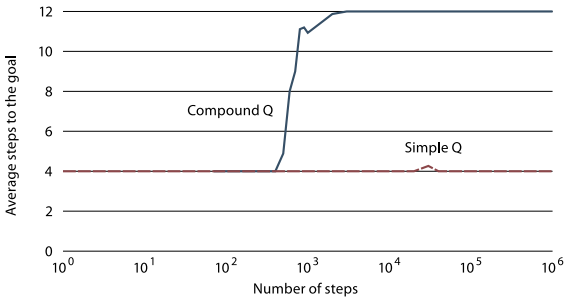

図 15　格子世界における目標到達に要した平均ステップ数

6　考察

6.1　獲得した行動規則の違い

従来の Q 学習が獲得した行動規則は，従来の強化学習の基準において最適なものである．すなわち，多項式的に割り引いた収益の期待値を最大化するものである．実際に，図 10 に示されたバンディットにおける算術平均報酬と図 13 に示された格子世界における算術平均報酬は，理論的最適値に等しく，複利リターンに対する Q 学習のものよりも高かった．

これに対し，複利リターンに対する Q 学習が獲得した行動規則は，複利リターンに対する強化学習の基準において最適なもの——すなわち，指数関数的に割り引い

たグロス・リターンの期待値を最大化するものである．実際に，図 9 に示されたバンディットにおける幾何平均リターンと図 12 に示された格子世界における幾何平均リターンは，理論的最適値に等しく，従来の Q 学習のものよりも高かった．

このように，最大化している対象が異なるため，従来の Q 学習が獲得する行動規則と複利リターンに対する Q 学習が獲得する行動規則は異なるものとなる．

6.2 割引率 γ の意味

従来の強化学習における割引率 γ は，遠い将来の報酬ほど多項式的に割り引いて 0 に近づける効果を持つ．これに対し，複利リターンに対する強化学習における割引率 γ は，遠い将来のグロス・リターンほど指数関数的に割り引いて 1 に近づける効果を持つ．

いずれも，$\gamma < 1$ のとき，遠い将来に得られる利益よりも近い将来に得られる利益を重視する．あるいは，遠い将来に得られる利益の見込みほど不確実性が高いと考えて報酬またはグロス・リターンを割り引く．

6.3 投資比率 f の意味

複利に対する強化学習にとって，$0 \leq f < 1$ の投資比率 f は，状態の価値 V および行動の価値 Q が $-\infty$ になることを回避するために導入されたものである．

他方で，ファイナンスの分野では，投資比率を導入する考え方は賭け過ぎを防止するための方法として一般的なものである．最適投資比率を求め，それに基づいて投資を行う手法は「オプティマル f」と呼ばれている [4, 5, 7]．最適な投資比率を求める方法としてはケリー基準 [1] がよく知られており，ケリー基準の半分を投資比率とするハーフ・ケリーと呼ばれる投資手法がよく用いられる [2].

7 まとめ

本論文では，複利リターンに対する強化学習の枠組みを提案した．複利リターンに対する強化学習では，多項式的割引収益の期待値を最大化する替わりに，投資比率 f $(0 \leq f < 1)$ のときの指数関数的割引複利リターンの期待値を最大化する．本論文では，複利リターンに対する強化学習の枠組みを，従来の強化学習における「報酬 r_t」を「投資比率が f のときのグロス・リターンの対数 $\log(1 + R_t f)$」に置き換えたものとして定式化し，従来の強化学習と同様に Bellman 最適化方程式を導いた．

また，本論文では，複利リターンに対する Q 学習を提案し，本手法が従来の Q 学習と同様に MDP において最適な行動規則を獲得できることを示した．

バンディットと格子世界の問題を用いた実験の結果より，提案手法の有効性を確認した．従来手法が近視眼的な報酬（単利リターン）の算術平均を最大化する行動規則を学習したのに対して，提案手法は複利リターンの幾何平均を最大とする行動規則を獲得することができた．

本論文で提案した複利リターンに対する強化学習は，失敗する可能性が高い最短経路よりも遠回りでも確実な経路を選択する．したがって，複利リターンに対する強化学習は，ファイナンスだけでなく，火星で岩石の採取と運搬を何度も繰り返す探査ロボットや放射能汚染事故現場で被災者を繰り返し運び出す救助ロボットのような分野への応用も期待できる．

参考文献

[1] J. L. Kelly, Jr. A new iterpretaion of information rate. *Bell System Technical Journal*, Vol. 35, pp. 917–26, 1956.

[2] William Poundstone. *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.

[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. *Portfolio management formulas: mathematical trading methods for the futures, options, and stock markets*. Wiley, 1990. 長尾 慎太郎 訳：投資家のためのマネー・マネジメント——資産を最大限に増やすオプティマル f, パン・ローリング, 2005.

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

[7] Larry R. Williams. *Long-term secrets to short-time trading*. Wiley, 1999. 清水 昭男, 長尾 慎太郎, 柳谷雅之 訳：ラリー・ウィリアムズの短期売買法——投資で生き残るための普遍の真理, パン・ローリング, 1999.