1A3-2

複利型強化学習

Compound Reinforcement Learning

松井 藤五郎
Tohgoroh Matsui

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

This paper describes a reinforcement learning framework based on compound returns, which is called compound reinforcement learning. Compound reinforcement learning maximizes the compound return in returns-based MDPs. We also describes compound Q-learning algorithm and its convergence. We present experimental results using an ilustrative example.

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 [Sutton 98] 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 個の円盤を持つスロット・マシンがあり，それぞれの円盤にはその円盤を回すためのアーム（腕）が一つずつ付いている．このとき，エージェントはどのアーム，つまり円盤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 のような円盤 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，円盤 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，円盤 A の払い戻し金の期待値は 1.5 ドルであり，円盤 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，円盤 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円盤 A は最適ではない．なぜなら，円盤 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 A 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 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一般的な考え方である [Poundstone 05]．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．

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

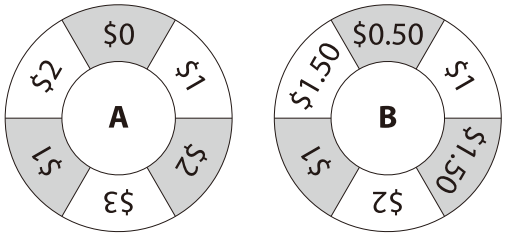

図 1: 2 つのホイールを持つバンディット問題．

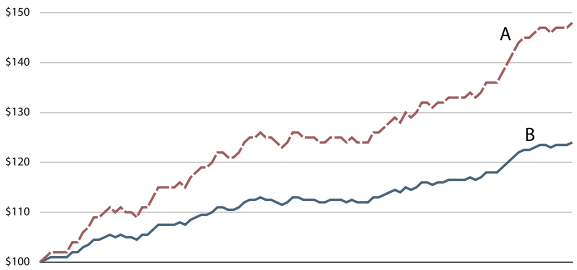

図 2: 100 ドルの財産を 1 ドルずついずれか一方に賭け続けたときの資産総額の推移の例．

そこで，本論文では，複利リターンを最大化するための複利型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. 複利型強化学習

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．

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．

$$R_t = \frac{P_{t+1} - P_t}{P_t} = \frac{P_{t+1}}{P_t} - 1 \qquad (1)$$

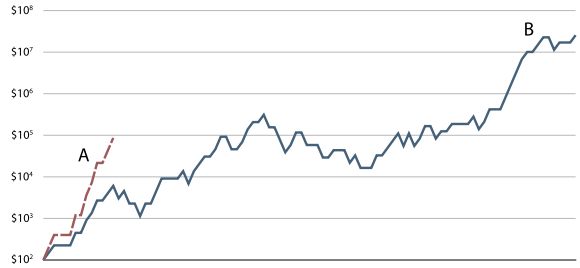

図 3: 100 ドルの財産を全額いずれか一方に賭けた続けたときの資産総額の推移の例.

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 [Campbell 97].

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

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．

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

本論文では，MDP におけるリターン R_{t+k+1} がマフ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 指数関数的割引

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．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

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

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．

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

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

というように，従来の強化学習における多項式的な割引収益と同様に，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．また，指数関数的にグロス・リターンを割り引くことは，遠い将来のリターンほど 0 に近づくと見積もっていると解釈できる．

2.3 投資比率

グロス・リターンの対数 $\log(1+R_t)$ は，リターン R_t が -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，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．

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

この式 (7) は，従来の強化学習の割引収益を表す式の報酬 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．

2.4 価値関数と最適価値関数

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する．

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

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．

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

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，$\mathsf{R}^a_{ss'}$ は獲得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

$$\begin{aligned}Q^\pi(s,a) =& \mathrm{E}_\pi\left[\log\rho_t \middle| s_t = s, a_t = a\right] \\ =& \sum_{s'\in\mathcal{S}}\mathcal{P}^a_{ss'}\left(\log(1+\mathsf{R}^a_{ss'}f) + \gamma V^\pi(s')\right)\end{aligned} \tag{9}$$

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
    $\delta \leftarrow \log(1+Rf) + \gamma \max_{a'} Q(s',a') - Q(s,a)$
    $Q(s,a) \leftarrow Q(s,a) + \alpha\delta$
    $s \leftarrow s'$
  until s が終端状態ならば繰り返しを終了
end loop

と表され，最適価値関数は

$$\begin{aligned} Q^*(s,a) &= \max_{\pi\in\Pi} Q^\pi(s,a) \\ &= \mathrm{E}\left[\log(1+R_{t+1}f) + \gamma \max_{a'} Q^*(s_{t+1},a') \middle| s_t = s, a_t = a\right] \\ &= \sum_{s'} \mathcal{P}^a_{ss'} \left(\log(1+\mathsf{R}^a_{ss'}f) + \gamma \max_{a'} Q^*(s',a')\right) \quad (10) \end{aligned}$$

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 $\log(1+\mathsf{R}^a_{ss'}f)$ に置き換えたものに等しい．

3. 複利型 Q 学習

3.1 アルゴリズム

上に述べたように，式 (10) 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 $\mathcal{R}^a_{ss'}$ 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 (10) 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．

$$Q(s_t,a_t) \leftarrow Q(s_t,a_t) + \alpha\Big(\log(1+R_{t+1}f) + \gamma \max_{a\in\mathcal{A}} Q(s_{t+1},a) - Q(s_t,a_t)\Big) \quad (11)$$

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．

複利型 Q 学習のアルゴリズムを Algorithm 1 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．

3.2 最適解への収束性

複利型 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，複利型強化学習におけ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 $\log(1+\mathsf{R}^a_{ss'})$ に置き換えたものである．したがって，複利型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 Q は最適行動価値 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，複利型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型 Q 学習が最適解に収束するための条件である．これらの条件を満たすとき，リターン型 MDP において複利型 Q 学習を用いて学習すると価値の推定値が真の価値へと近づくことは，Watkins と Dayan の証明 [Watkins 92] において $r_t = \log(1+R_t f)$ とおくことによって示すことができる．

4. 実験結果

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 Q 学習（Simple Q-learning）と本論文で提案した複利型 Q 学習（Compound Q-learning）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，複利型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-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 $\alpha = 0.01$ 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 4, 5 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．複利型 Q 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 Q 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. 考察と関連研究

複利型強化学習ではリターン R が投資比率 f $(0 \le f < 1)$ のときのグロス・リターンの対数 $\log(1+Rf)$ に変換され，強化信号として用いられる．この関係を図 6 に示す．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を抑制し，負の強化信号は増強される．結果として，正の強化信号に対してはリスク追求型となり，負の強化信号に対してはリスク回避型となる．しかしながら，f が小さいときは

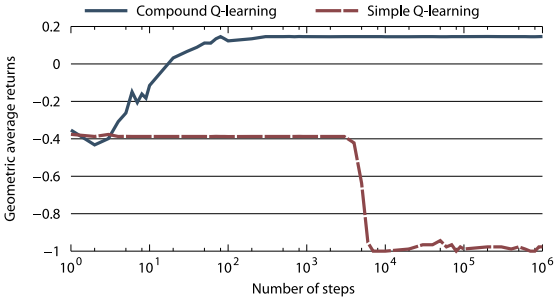

図 4: 幾何平均リターン（1 ドル当たり）

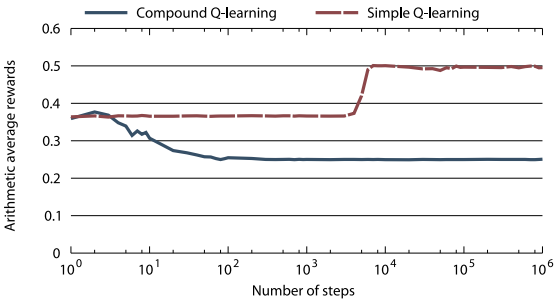

図 5: 算術平均報酬（1 ドル当たり）

負の強化信号も抑制され，常にリスク追求型となる．したがって，投資比率パラメーター f の設定には対数に関する知識が必要となる．

投資比率という概念は，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たものである．確率分布に基づいて計算される最適な投資比率は「ケリー基準」と呼ばれる [Kelly, Jr. 56]．また，「オプティマル f」と呼ばれるケリー基準を推定する方法が提案されている [Vince 90]．

一方，リスク回避型の強化学習としては，期待価値から分散を引く (expected-value-minus-variance) アプローチ [Heger 94, Sato 01] 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 [Geibel 05] などが提案されているが，これらのアプローチでは複利リターンは最大化されない．投資比率パラメーター f によって負のリターンに対するリスク選好を制御でき，かつ，複利リターンを最大化することができることが，複利型強化学習の利点である．

6. まとめ

本論文では，複利リターンに基づいて学習を行う複利型強化学習の枠組みを提案した．複利型強化学習では，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対数の期待値を最大化する．

本論文では，複利型強化学習の枠組みを，従来の報酬型 MDP に対する強化学習における「報酬 r_t」をリターン型 MDP に対する「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 最適方程式を導いた．また，複利型 Q 学習を提案し，本手法が報酬型 MDP における従来の Q 学習と同様にリターン型 MDP において最適な行動規則を獲得できることを示した．複利型 Q 学習と同様に，従来の強化学習アルゴリズムの多くは簡単に複利型に拡張することができる．

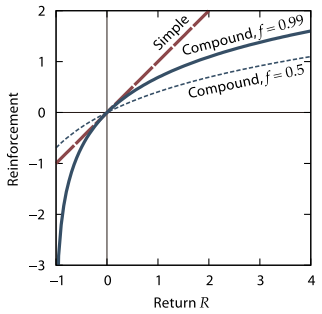

図 6: リターンと強化信号の関係

バンディットの問題を用いた実験の結果より，複利型強化学習の枠組みおよび複利型 Q 学習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．

参考文献

[Campbell 97] 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, 本多 俊毅, 和田 賢治 訳『ファイナンスのための計量分析』共立出版 (2003)

[Geibel 05] Geibel, P. and Wysotzki, F.: Risk-Sensitive Reinforcement Learning Applied to Control under Constraints, *JAIR*, Vol. 24, pp. 81–108 (2005)

[Heger 94] Heger, M.: Consideration of Risk in Reinforcement Learning, in *ICML 1994*, pp. 105–111 (1994)

[Kelly, Jr. 56] Kelly, Jr., J. L.: A new interpretaion of information rate, *Bell System Tech J*, Vol. 35, pp. 917–26 (1956)

[Poundstone 05] 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)

[Sato 01] Sato, M. and Kobayashi, S.: Average-Reward Reinforcement Learning for Variance Penalized Markov Decision Problems, in *ICML 2001*, pp. 473–480 (2001)

[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)

[Vince 90] Vince, R.: Find your optimal f, *Technical Analysis of Stock & Commodities*, Vol. 8, No. 12, pp. 476–477 (1990)

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