4E1-OS-15-5

複利型強化学習における投資比率の最適化

Optimizing Betting Fraction in Compound Reinforcement Learning

松井 藤五郎[*1] 後藤 卓[*2] 和泉 潔[*3*4] 陳 ユ[*3]
Tohgoroh Matsui Takashi Goto Kiyoshi Izumi Yu Chen

[*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

This paper describes optimization of the betting fraction parameter in compound reinforcement learning. Compound reinforcement learning maximizes the expected logarithm of compound returns in return-based MDPs. However, a new betting fraction parameter is introduced in order not to diverge values to negative infinity and it causes a problem of choosing the parameter. In this paper, we proposed a method to optimize the betting fraction with on-line gradient ascent in compound reinforcement learning.

1. はじめに

複利型強化学習は，エージェントが獲得するリターンに基づく複利リターン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である．これまでに，複利型強化学習の基本的な枠組みと従来の Q 学習を拡張した複利型 Q 学習のアルゴリズムが提案され，国債銘柄選択問題や国債取引問題での有効性が示されている [Matsui 12, 松井 11a, 松井 11b].

複利型強化学習では，エージェントが自分の資産のうちのどれだけを投資するかを表す投資比率パラメーター f が導入されている．このパラメーターは，複利リターンに大きく影響し，この値によって複利リターンが最大となる行動が異なることがある．

投資比率 f に関しては，リターンの確率分布が既知であるなら，複利リターンを最大化する投資比率を解析的に求められることが明らかとなっている [Kelly, Jr. 56]．この既知のリターン分布の下で複利リターンを最大化する投資比率は，ケリー基準と呼ばれる．しかしながら，一般的には，投資に対するリターンの確率分布は未知であり，真のケリー基準を事前に求めることはできない．

これまでに, optimal f と呼ばれる過去のリターンから良い投資比率を推定する手法が提案されている [Vince 90]．optimal f は，リターンの確率分布が未知の場合でも，良い投資比率を得ることができる．しかしながら，optimal f によって得られる投資比率はケリー基準と同じではなく [Vince 11]，複利リターンを最大化するのにもちることができない．

そこで, 本論文では, オンライン勾配法を用いて投資比率 f を最適化し, 複利リターンを最大化する行動規則を複利型強化学習を用いて学習する方法を提案する．また，Q 学習 [Watkins 92] と Sarsa [Sutton 96] のそれぞれをベースにした提案手法を 3 本腕バンディット問題に適用し，その有効性を確認する．

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

複利型強化学習は，割引複利リターン

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

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．

$$V^{\pi}(s) = \mathrm{E}_{\pi}\left[\log\prod_{k=0}^{\infty}(1+R_{t+k+1}f)^{\gamma^k}\,\middle|\, s_t=s\right] \quad (2)$$

$$Q^{\pi}(s,a) = \mathrm{E}_{\pi}\left[\log\prod_{k=0}^{\infty}(1+R_{t+k+1}f)^{\gamma^k}\,\middle|\, s_t=s, a_t=a\right] \quad (3)$$

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），$\mathrm{R}^a_{ss'}$ 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．

複利型 Q 学習は，従来の Q 学習 [Watkins 92] 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．時刻 t の状態 s_t において行動 a_t を実行し，次の時刻 $t+1$ にリターン R_{t+1} を受け取ると，状態行動対 s_t, a_t に対する Q 値を次のように更新する．

$$Q_{t+1}(s_t,a_t) = Q_t(s_t,a_t) + \alpha\Delta_t \quad (4)$$

$$\Delta_t = \log(1+R_{t+1}f) + \gamma\max_{a\in\mathcal{A}}Q_t(s_{t+1},a) - Q_t(s_t,a_t) \quad (5)$$

ここで，α はステップ・サイズ，γ は割引率，f は投資比率を表す．

同様にして，Sarsa [Sutton 96] も報酬 r_{t+1} を投資比率 f のときのグロス・リターンの対数 $\log(1+R_{t+1}f)$ に置き換えることによって，複利型のアルゴリズムに拡張することができる．すなわち，複利型 Sarsa の更新式は表される．

$$\Delta_t = \log(1+R_{t+1}f) + \gamma Q_t(s_{t+1},a_{t+1}) - Q_t(s_t,a_t) \quad (6)$$

このように，複利型強化学習は，従来の強化学習アルゴリズムを自然な形で拡張して用いることができる．

3. オンライン勾配法による投資比率の最適化

リターン R_{t+1} を受け取ったとき，時刻 $t+1$ までの複利リターンは次のように計算される．

$$G_{t+1} = \prod_{k=1}^{t+1}(1+R_k f) \quad (7)$$

これは，f をパラメーターとする関数と見ることができる．そこで，複利リターン G_{t+1} を最大化するような投資比率 f を求める．

まず，式 (7) の両辺の対数を取ると次のようになる．

$$\log G_{t+1} = \sum_{k=1}^{t+1}\log(1+R_k f) \quad (8)$$

この両辺を f で偏微分する．

$$\frac{\partial}{\partial f}\log G_{t+1} = \sum_{k=1}^{t+1}\frac{\partial}{\partial f}\log(1+R_k f) \quad (9)$$

$$= \sum_{k=1}^{t+1}\frac{R_k}{1+R_k f} \quad (10)$$

目的関数 $\log G_{t+1}$ は上に凸な関数であるから，この値が 0 となる f が G_{t+1} を最大化する f である．

このとき，最急降下法を用いることによって，G_{t+1} を最大化する f を求めることができる．

$$f_{n+1} = f_n + \eta\sum_{k=1}^{t+1}\frac{R_k}{1+R_k f_n} \quad (11)$$

ここで，η は学習率と呼ばれるパラメーターである．ただし，強化学習の各ステップごとにこれを求めるには，過去のリターン $R_1,\ldots,R_{t+1}$ を全て記憶しておかなければならない．

そこで，本論文では，最急降下法の替わりにオンライン勾配法を用いることを提案する．すなわち，強化学習の各ステップ t において f を次のように更新する．

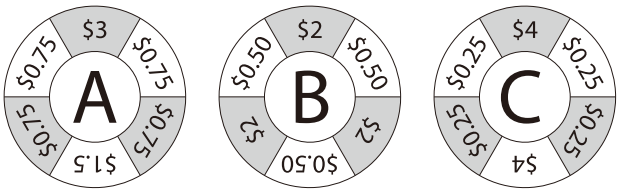
$$f_{t+1} = f_t + \eta\frac{R_{t+1}}{1+R_{t+1}f_t} \quad (12)$$

オンライン勾配法では，f の更新に R_{t+1} しか使用しないので，過去のリターンを記憶しておく必要がない．

実際には，G_{t+1} を最大化する f の計算は状態行動対ごとに行うため，次のように更新する．

$$f_{t+1}(s_t,a_t) = f_t(s_t,a_t) + \eta\frac{R_{t+1}}{1+R_{t+1}f_t(s_t,a_t)} \quad (13)$$

オンライン勾配法による投資比率最適化付きの複利型 Q 学習と複利型 Sarsa のアルゴリズムを，それぞれ，Algorithm 1, Algorithm 2 に示す．



Ari. Avg.: $1.25	Ari. Avg.: $1.25	Ari. Avg.: $1.50
Std. Dev.: $0.91	Std. Dev.: $0.82	Std. Dev.: $1.94
Geo. Avg.: $1.06	Geo. Avg.: $1.00	Geo. Avg.: $0.63

図 1: バンディット問題

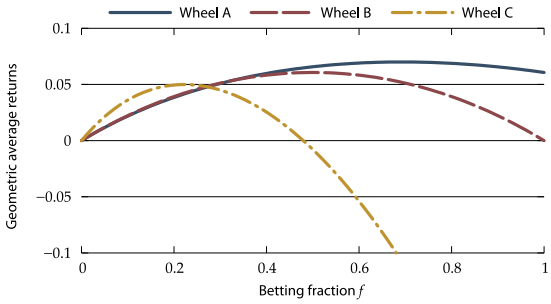


図 2: バンディット問題の各ホイールおける投資比率に対する幾何平均リターン

4. 実験

4.1 実験方法

図 1 に示す 3 本腕バンディット問題を用いて実験を行った．このマシーンには 3 つのホイールがあり，それぞれのホイールを回すための腕が 3 本ついている．ホイールに表示されているのは，賭け金 $1 あたりの払い戻し金の額である．ホイールの下には，賭け金 $1 あたりの払い戻し金の算術平均，標準偏差，幾何平均が示されている．

ホイール A は幾何平均リターンが最も大きく，算術平均リターンがホイール B と等しい．ホイール B はリターンの分散が最も小さく，算術平均リターンがホイール A と等しい．ホイール C は算術平均リターンが最も大きいが，幾何平均リターンは最も小さく，リターンの分散が最も大きい．

各ホイールにおける投資比率に対する幾何平均リターンを図 2 に示す．幾何平均リターンが最大となるのは，ホイール A は $f \approx 0.69$ のとき，ホイール B は $f = 0.5$ のとき，ホイール C は $f \approx 0.22$ のときである．

この問題を用いて，オンライン勾配法による投資比率最適化付き複利型 Q 学習，固定投資比率の複利型 Q 学習 [Matsui 12, 松井 11a, 松井 11b]，固定投資比率の Q 学習 [Watkins 92] を比較した．同様に，オンライン勾配法による投資比率最適化付き複利型 Sarsa，固定投資投資比率の複利型 Sarsa，固定投資比率の Sarsa [Sutton 96] を比較した．

割引率は $\gamma = 0.9$ とした．オンライン勾配法による投資比率最適化における投資比率の初期値は $f_0 = 1.0$ とし，固定したときの投資比率は $f = 1.0$ とした．この問題では，破産 ($R_t = -1$) が生じないため，$f = 1.0$ でもグロス・リターンの対数は発散しない．ステップ・サイズ α とオンライン勾配法における投資比率学習率 η は共に 0.001 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．

Algorithm 1 オンライン勾配法による投資比率最適化付き複利型 Q 学習アルゴリズム

入力: 割引率 γ, ステップ・サイズ α, 投資比率学習率 η
$Q(s,a)$ を任意に初期化
$f(s,a)$ を $0 \leq f(s,a) \leq 1$ の範囲で任意に初期化
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
    $Q(s,a) \leftarrow Q(s,a) + \alpha\left(\log(1 + Rf(s,a)) + \gamma \max_{a'} Q(s',a') - Q(s,a)\right)$
    $f(s,a) \leftarrow f(s,a) + \eta \frac{R}{1+Rf(s,a)}$
    $s \leftarrow s'$
  until s が終端状態ならば繰り返しを終了
end loop

Algorithm 2 オンライン勾配法による投資比率最適化付き複利型 Sarsa アルゴリズム

入力: 割引率 γ, ステップ・サイズ α, 投資比率学習率 η
$Q(s,a)$ を任意に初期化
$f(s,a)$ を $0 \leq f(s,a) \leq 1$ の範囲で任意に初期化
loop (各エピソードに対して繰り返し)
  s を初期化
  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
  repeat (エピソードの各ステップに対して繰り返し)
    行動 a を実行し，リターン R と次の状態 s' を観測
    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
    $Q(s,a) \leftarrow Q(s,a) + \alpha\left(\log(1 + Rf(s,a)) + \gamma Q(s',a') - Q(s,a)\right)$
    $f(s,a) \leftarrow f(s,a) + \eta \frac{R}{1+Rf(s,a)}$
    $s \leftarrow s', a \leftarrow a'$
  until s が終端状態ならば繰り返しを終了
end loop

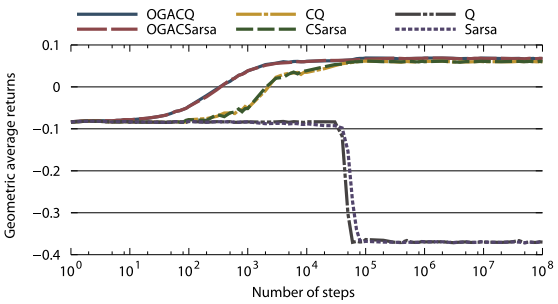


図 3: 幾何平均リターンの推移

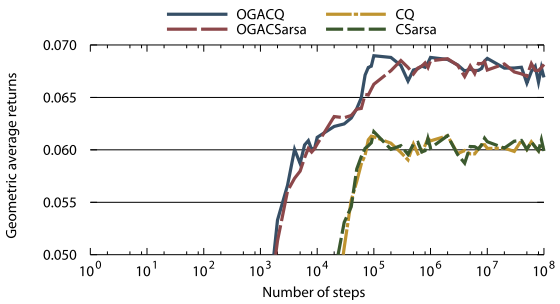


図 4: 幾何平均リターンの推移（一部拡大）

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，最も価値が高いと学習した行動を選択し続けた場合の幾何平均リターンを求めた．

4.2 結果

結果を図 3 に示す．横軸は学習ステップ数，縦軸は幾何平均リターン

$$\bar{G} = \left(\prod_{t=1}^{n}(1 + R_t f)\right)^{\frac{1}{n}} - 1 \qquad (14)$$

を表す．OGACQ と OGACSarsa は，それぞれ，オンライン勾配法による投資比率最適化付きの複利型 Q 学習と複利型 Sarsa，CQ と CSarsa は，それぞれ，投資比率固定の複利型 Q 学習と複利型 Sarsa，Q と Sarsa は，それぞれ，投資比率固定の Q 学習と Sarsa を表している．

図 3 の一部を拡大したものを図 4 に示す．最終的な幾何平均リターンが最も高かったのは，提案手法であるオンライン勾配法による投資比率最適化付きの複利型 Q 学習と複利型 Sarsa だった．投資比率を固定した複利型 Q 学習と複利型 Sarsa も同じ行動，すなわち，ホイール A を選択する行動規則を学習したが，投資比率が $f = 1.0$ で固定されているため，オンライン勾配法による投資比率最適化付きの提案手法よりも幾何平均リターンが小さかった．従来の Q 学習と Sarsa は，算術平

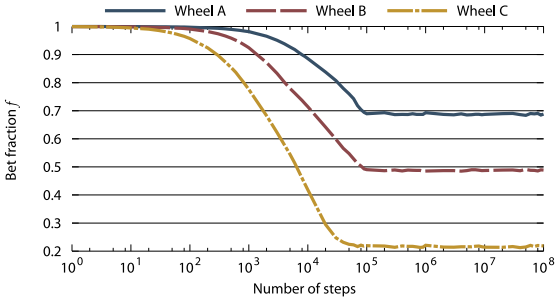


図 5: オンライン勾配法による投資比率最適化付き Q 学習における投資比率の推移

均リターンが最大となるホイール C を選択する行動を学習し，幾何平均リターンは最低だった．

今回の問題では，オンライン勾配法による投資比率最適化付きの複利型強化学習，投資比率固定の複利型強化学習，従来の強化学習のいずれにおいても，Q 学習と Sarsa の性能の違いは大きくなかった．

オンライン勾配法による投資比率最適化付き複利型 Q 学習における投資比率の推移を図 5 に示す．それぞれのホイールにおいて，ほぼ最適な投資比率に収束していることがわかる．

5. 考察

複利型強化学習は，複利リターンを最大化する行動規則を学習することができたが，これまでは投資比率が固定されていた．複利リターンの値は投資比率 f に依存しているため，この投資比率をどのように決めるかが複利型強化学習における一つの問題となっていた．

本論文で提案したオンライン勾配法による投資比率最適化付き複利型強化学習は，オンライン勾配法を用いて幾何平均リターン，つまり複利リターンを最大化する投資比率を学習することによって，複利型強化学習の問題点を解決している．また，最急降下法の替わりにオンライン勾配法を用いることによって，過去のリターンを全て記憶する必要がなくなっている．

提案手法において投資比率を最適化するために用いられているオンライン勾配法は，複利型強化学習には全く依存していない．オンライン勾配法においては，目的関数が強凸である場合は，Regret 解析 [Hazan 07] によって最適解へ収束することが明らかとなっている [岡野原 11]．提案手法の目的関数 $\log G_{t+1}$ は，f で二階偏微分すると

$$\frac{\partial^2}{\partial^2 f}\log G_{t+1} = \frac{\partial}{\partial f}\sum_{k=1}^{t+1}\frac{R_k}{1+R_k f} \tag{15}$$

$$= \sum_{k=1}^{t+1}\frac{{R_k}^2}{(1+R_k f)^2} \geq 0 \tag{16}$$

であり，$\exists R_k[R_k \neq 0]$ のとき強凸である．したがって，提案手法では最適な投資比率 f を学習することができる．

6. まとめ

本論文では，複利型強化学習において，オンライン勾配法を用いて投資比率 f を最適化する手法を提案した．また，提案手法に基づくオンライン勾配法による投資比率最適化付きの Q 学習と Sarsa を 3 本腕バンディット問題に適用し，投資比率固定の複利型強化学習および従来の強化学習と比較した．

実験の結果から，提案手法が最適な投資比率を獲得できることが確認された．また，投資比率を最適化することによって，提案手法は幾何平均リターンを最大化することができた．

今後は，より実際的で複雑なタスクに提案手法を適用し，提案手法が一般的に有効であることを確認したい．

謝辞

本研究は科研費（23700182）の助成を受けたものである．

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．

参考文献

[Hazan 07] Hazan, E., Agarwal, A., and Kale, S.: Logarithmic regret algorithms for online convex optimization, *Machine Learning*, Vol. 69, No. 169–192 (2007)

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

[Matsui 12] Matsui, T., Goto, T., Izumi, K., and Chen, Y.: Compound Reinforcement Learning: Theory and An Application to Finance, in Sanner, S. and Hutter, M. eds., *Recent Advanced in Reinforcement Learning: Revised and Selected Papers of the European Workshop on Reinforcement Learning 9 (EWRL 2011)*, Vol. 7188 of *Lecture Notes in Computer Science*, pp. 321–332 (2012), in press

[Sutton 96] Sutton, R. S.: Generalization in reinforcement learning: Successful examples using sparse coarse coding, in Touretzky, D. S., Mozer, M. C., and Hasselmo, M. E. eds., *Advances in Neural Information Processing Systems*, Vol. 8, pp. 1038–1044, MIT Press (1996)

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

[Vince 11] Vince, R.: Optimal f and the Kelly Criterion, *IFTA Journal*, pp. 21–28 (2011)

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

[岡野原 11] 岡野原 大輔：オンライン凸最適化と線形識別モデルの最前線, 第 14 回情報論的学習理論ワークショップ (IBIS 2011) (2011)

[松井 11a] 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)

[松井 11b] 松井 藤五郎, 後藤 卓, 和泉 潔, 陳 ユ：複利型強化学習の枠組みと応用, 情報処理学会論文誌, Vol. 52, No. 12, pp. 3300–3308 (2011)