1E4-1

複利型強化学習による危険回避行動の学習

Learning Risk-Averse Behavior Using Compound Reinforcement Learning

松井 藤五郎*1　落合 宏旭*1
Tohgoroh Matsui　Hiroaki Ochiai

*1中部大学
Chubu University

This paper describes a method for learning risk-averse behavior using compound reinforcement learning. Compound reinforcement learning is a reinformentment learning framework for maximizing expected discounted compound return in return-based Markov decision processes. Compound reinforcement learning dislikes negative return rather than likes posive return. In this paper, we use compound reinforcement learning to learn risk-averse behavior in grid-world tasks.

1. はじめに

火星や東日本大震災の被害を受けた東京電力福島第一原子力発電所内など，人間が直接行くことができないところでロボットが活躍している．このようなところでは，ロボットに危険が迫ったとき，人間からロボットへの危険回避命令が間に合わない場合があるため，ロボットが自発的に判断し行動することが必要となる．実際に，福島第一原子力発電所内の様子を調べに行ったロボットが，階段で転んだり，通信ケーブルが切れたりして戻って来れなくなるケースが多発している [間宮 12]．したがって，ロボットが自発的に危険を回避する行動を身につけている必要がある．

複利型強化学習 [松井 11a, Matsui 12] は，報酬の代わりに利益率を観測する利益率型マルコフ決定過程 (MDP) において，割引複利利益率の期待値を最大化する行動規則を学習する枠組みである．これまでに，複利型強化学習は主に投資のドメインに適用され，国際銘柄選択問題，国債取引問題，N 本腕バンディット問題などでの有用性が示されている [Matsui 12, 松井 11b, 松井 13]．

複利強化学習は，エージェントが得た利益率の対数を取ることによって，利益率が負のときは強化値を増強し，利益率が正のときは強化値を抑制する．これにより，正の利益率を好むというよりも負の利益率を嫌うという特性を持っている．とくに，一度でも利益率が -1 になると全財産が無くなり，最終的な複利利益率も -1 になってしまうことを意味するため，複利型強化学習は -1 に近い利益率を獲得しないような振る舞いを学習する傾向がある．

そこで，本論文では，複利型強化学習が負の利益率を嫌うという性質を利用して，複利型強化学習を用いて危険回避行動を学習させることを提案する．

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形で表すことができるため，複利型強化学習では，行動価値を割引複利リターンの対数の期待値と定義する．すなわち，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．

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

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，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．

$$Q_{t+1}(s_t,a_t) = Q_t(s_t,a_t) + \alpha\Delta_t \quad (3)$$

$$\Delta_t = \log(1+R_{t+1}f) + \gamma\max_{a\in A}Q_t(s_{t+1},a) - Q_t(s_t,a_t) \quad (4)$$

ここで，α は学習率パラメーターである．

3. 危険回避行動の学習

本論文では，危険な経路と安全な経路がある格子世界の問題を対象とする．図 1 に，本論文が対象とするタスクの例を示す．図中の S はスタート地点を表し，G はゴール地点を表す．ゴール地点の左隣にある灰色のマスは危険な状態であり，一定の確率で穴に落ちて動けなくなってしまう．動けなくなるということは，このタスクに失敗したことを意味する．エージェントの行動は東西南北の 4 種類であり，危険な状態を除き決定的な状態遷移が行われるものとする．

複利型強化学習では，大きな負の利益率を回避する行動を学習する性質がある．そこで，本論文では，複利型強化学習において失敗に対して -1 の利益率を与えることで，失敗を回避する行動を学習させることを提案する．

4. 実験

4.1 実験 1: 危険度

図 1 に示したタスクを用いて，提案手法の有効性を確認するための実験をおこなった．灰色のマスにおいて東の行動を選

連絡先: 松井藤五郎, TohgorohMatsui@tohgoroh.jp

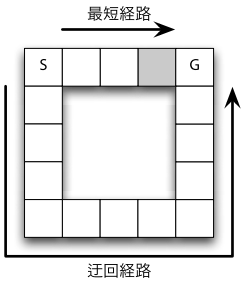

図 1: 危険な経路が存在する格子世界の問題

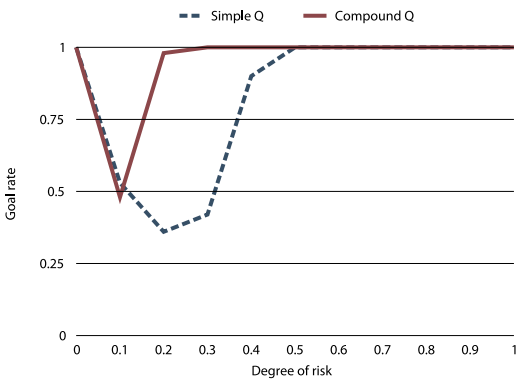

図 2: 実験 1 の結果

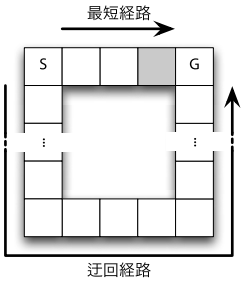

図 3: より長い迂回経路がある問題

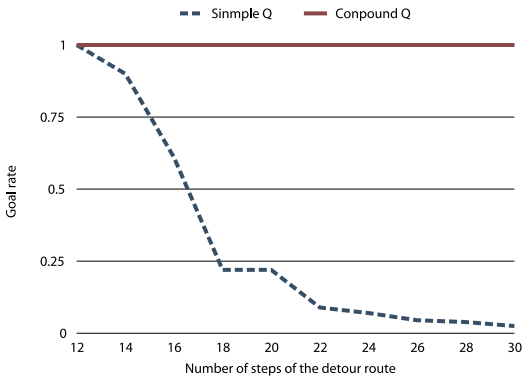

図 4: 実験 2 の結果

択したときに失敗する確率（危険度）を 0 から 1 まで 0.1 ごとに変えて 100 回の学習を行い，学習した行動価値が最大の行動を選択し続けたとき（すなわち，グリーディー選択を用いたとき）のゴール地点への平均到達率を調べた．

複利型強化学習のアルゴリズムには複利型 Q 学習 [Matsui 12] を使用し，学習率を $\alpha = 0.1$，割引率を $gamma = 0.9$，投資比率を $f = 0.9$ とした．比較のため，同じパラメーターを用いた従来の Q 学習と比較した．

結果を図 2 に示す．Simple Q は従来の Q 学習，Compound Q は複利型 Q 学習を表す．

従来の Q 学習は，危険度が 0.5 以上になると迂回経路を確実に学習したが，危険度が 0.5 より小さいときには危険な最短経路を学習することがあった．これに対し，複利型 Q 学習は，危険度が 0.1 のときは最短経路を学習する傾向があったものの，危険度 0.2 のときはほぼ，危険度 0.3 以上では確実に迂回経路を学習して危険を回避することができた．

4.2 実験 2: 迂回経路の長さ

続いて，危険度を従来の Q 学習，複利型強化学習ともに確実に迂回経路を学習できた 0.5 に固定し，迂回経路の長さを 3 に示すように（実験 1 と同じ）12 から 30 まで 2 ずつ伸ばして実験 1 と同様の実験を行った．この結果を図 4 に示す．従来の Q 学習は迂回経路が長くなるに連れて最短経路を学習する傾向が強くなったが，複利型強化学習は迂回経路が長くなっても確実に迂回経路を学習して危険を回避することができた．

5. まとめ

本論文では，複利型強化学習を用いて危険回避行動を学習することを提案した．複利型強化学習が -1 の利益率を嫌う性質を利用して，失敗行動に対して -1 の利益率を与えることによって失敗行動につながる危険を回避させることができた．従来の Q 学習では安全な迂回経路の長さが長くなると危険がある最短経路を選ぶ傾向があるが，複利型 Q 学習は迂回経路が長くなっても危険がある最短経路ではなく安全な迂回経路を選択することができた．

複利型強化学習は，危険回避に有効ではあるが，危険がない経路がなく，最も危険が小さい経路を選ばなければならないような問題でも，危険を回避して立ち往生しまうという問題がある．このような状況にどう対応するかは今後の課題である．

参考文献

[Matsui 12] Matsui, T., Goto, T., Izumi, K., and Chen, Y.: Compound Reinforcement Learning: Theory and An Application to Finance, *EWRL 2011*, pp. 321–332 (2012)

[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)

[松井 13] 松井 藤五郎, 後藤 卓, 和泉 潔, 陳 ユ：複利型強化学習における投資比率の最適化, 人工知能学会論文誌, Vol. 28, No. 3, pp. 267–272 (2013)

[間宮 12] 間宮 利夫：戻ってこないロボットたち, しんぶん赤旗, 2012 年 12 月 31 日付 (2012)