1E4-2in

PSO 導入による学習効率を考慮した報酬関数の推定

Search for Reward Function that Makes Faster Convergence to the Optimum Policy by PSO

*1北里勇樹　*2荒井幸代
Kitazato Yuki　Arai sachiyo

*1千葉大学大学院工学研究科建築・都市科学専攻
Graduate School of Engineering, Chiba University, Architecture and Urban Science

*2千葉大学大学院工学研究科都市環境システムコース
Graduate School of Engineering, Chiba University, Department of Urban Environmental Systems course

Though reinforcement learning has been popular for its applicability to control systems using minimum knowledge about their dynamics, it is impractical in real world applications because of the long time it takes them to learn. In the context of reinforcement learning context, the minimum knowledge means the sparse rewards during the learning process. If extra origins of reward would be introduced within the process of learning, the optimal policy could be found in In the effective way. In this study, we proposed a method to find the additional origin of rewards via inverse reinforcement learning algorithm that combines the *Particle Swarm Optimization*. Here, PSO is applied to search the most appropriate reward function to reach the optimal policy within the less learning process. The effectiveness of the proposed method is shown by some empirical experiments of maze problems.

1. はじめに

強化学習 [1] は，報酬と呼ばれるスカラー量を手掛かりに，ゴールに至る適切な行動を獲得する枠組みである．ゴールにおける報酬だけを定義すればよいことから，その応用が期待されてきたが，ゴールまでに多くの遷移を必要とする大規模な問題では，報酬の遅れが大きく，学習時間を要するため，実問題への応用のボトルネックとなっている．

この問題に対してサブゴールを設定し，学習の高速化を実現する方法がある．この方法は，ゴール状態以外にも報酬を設定することによって，報酬の遅れを緩和し，学習を高速化する．しかし，途中の状態に対して適切な報酬値を設定することは難しいとされている．

サブゴールを用いて学習を高速化する手法には，事前知識を用いず学習中の試行錯誤によるものと，事前知識を用いるものがある．前者には，ゴールに至るまでに遷移した状態に対して等間隔にサブゴールを設定し，サブゴールの評価・再設定の繰返しによってサブゴールを発見する手法 [2]，学習中に遷移した状態の出現頻度からサブゴールを発見する手法 [3][4] がある．後者には,「望ましい行動系列」を所与とし，報酬関数を推定する逆強化学習 [5][6] がある．逆強化学習では，学習により最適行動が得られるような報酬関数を推定するが，この報酬関数は学習の速度を考慮していない．そこで，本研究では既存の逆強化学習に PSO を導入し，学習速度を考慮したアルゴリズムへ拡張することで，学習高速化のためのサブゴールを発見する手法を提案する．既存の逆強化学習と提案手法を迷路問題に適用し，提案手法の有用性を t 検定と報酬関数の解析結果から考察する．

連絡先: 北里勇樹，千葉大学大学院工学研究科建築・都市科学専攻，〒 263-8522 千葉市稲毛区弥生町 1-33，043-290-3316，aaca0294@chiba-u.jp

2. 逆強化学習

逆強化学習は，Russell [7] によって最適な行動系列や環境モデルを所与として報酬関数を求める問題として定義され，様々な手法が提案されている．Ng ら [5] は有限状態空間を持つ環境に対しては線形計画法，無限の状態空間を持つ環境に対してはモンテカルロ法を用いて報酬関数を推定する手法を示し，Abbeel ら [6] は報酬関数を推定する過程で最適な方策を獲得する “Apprenticeship Learning”(見習い学習) の手法を示した．

2.1 Ng の逆強化学習

各状態 s における最適な行動 a_1 を所与とし，式 (1) の線形計画問題を解くことによって報酬関数 $\boldsymbol{R}$ を推定する．式 (1) において，報酬関数ベクトル $\boldsymbol{R}$ は状態 s における報酬 r_s で与えられる．状態遷移行列 $\boldsymbol{P}_a$ は行動 a の状態遷移確率で与えられる $M \times M$ 行列であり，状態 s から行動 a をとり s' に遷移する確率を $P^a_{ss'}$ とすると，$\boldsymbol{P}_a$ は式 (2) で表される．$\boldsymbol{P}_a(i)$ は，$\boldsymbol{P}_a$ の第 i 行ベクトルで式 (3) のように表される．λ はペナルティ係数であり，λ を小さくすることで多くのサブゴールが得られる．$\boldsymbol{R}_{max}(>0)$ は報酬の制約として設定する値である．

$$\begin{aligned}\text{maximize}: \sum_{i=1}^{N} \min_{a \in a_2,\ldots,a_k} \{(\boldsymbol{P}_{a_1}(i) - \boldsymbol{P}_a(i)) \\ (\boldsymbol{I} - \gamma \boldsymbol{P}_{a_1})^{-1}\boldsymbol{R}\} - \lambda ||\boldsymbol{R}||_1 \qquad (1)\\ \text{subject to}: (\boldsymbol{P}_{a_1} - \boldsymbol{P}_a)(\boldsymbol{I} - \gamma \boldsymbol{P}_{a_1})^{-1}\boldsymbol{R} \geq 0\end{aligned}$$

$$\boldsymbol{P}_a = \begin{pmatrix} P^a_{11} & P^a_{12} & \cdots & P^a_{1j} & \cdots & P^a_{1M} \\ P^a_{21} & P^a_{22} & \cdots & P^a_{2j} & \cdots & P^a_{2M} \\ \vdots & \vdots & \ddots & \vdots & \ddots & \vdots \\ P^a_{i1} & P^a_{i2} & \cdots & P^a_{ij} & \cdots & P^a_{iM} \\ \vdots & \vdots & \ddots & \vdots & \ddots & \vdots \\ P^a_{M1} & P^a_{M2} & \cdots & P^a_{Mj} & \cdots & P^a_{MM} \end{pmatrix} \qquad (2)$$

$$\boldsymbol{P}_a(i) = (P^a_{i1}, P^a_{i2}, \cdots, P^a_{iM}) \qquad (3)$$

また，式 (1) は最適な行動と二番目に良い行動の期待報酬の差を最大化する報酬関数 $\boldsymbol{R}$ を求めるものであるが，二番目に良い行動だけでなく，すべての行動における期待報酬の差を最大化する報酬関数 $\boldsymbol{R}$ を求める目的関数についても Ng は述べており，この目的関数を式 (4) に示す.

$$\begin{aligned}\text{maximize}: \sum_{i=1}^{N}\sum_{j=2}^{K}\{(\boldsymbol{P}_{a_1}(i)-\boldsymbol{P}_{a_j}(i))\\ (\boldsymbol{I}-\gamma\boldsymbol{P}_{a_1})^{-1}\boldsymbol{R}\}-\lambda||\boldsymbol{R}||_1 \quad (4)\\ \text{subject to}: (\boldsymbol{P}_{a_1}-\boldsymbol{P}_a)(\boldsymbol{I}-\gamma\boldsymbol{P}_{a_1})^{-1}\boldsymbol{R}\geq 0\end{aligned}$$

本稿では式 (1) を用いるものを Ng-1，式 (4) を用いるものを Ng-2 と呼ぶ.

2.2 Abbeel の逆強化学習

各状態で最適な行動をとるエージェントをエキスパートと定義する．Abbeel の逆強化学習ではエキスパートの行動軌跡を所与とし，エキスパートに近い行動軌跡が得られる報酬関数 $\boldsymbol{R}$ を推定する.

具体的には，特徴量 $\boldsymbol{\phi}$ と特徴期待値 $\boldsymbol{\mu}$ を定義し，エキスパートの特徴期待値との差が ϵ 以下となる特徴期待値が得られる報酬関数を推定する．ここで，状態は $\mathcal{S}\rightarrow[0,1]^k$ で定義される特徴量ベクトルで表す．また，方策 $\boldsymbol{\pi}$ に従ったときの期待割引累積特徴量 $\boldsymbol{\mu}(\boldsymbol{\pi})=\boldsymbol{E}[\sum_{t=0}^{\infty}\gamma^t\boldsymbol{\phi}(s_t)|\boldsymbol{\pi}]\in\mathbb{R}^k$ を用いて行動軌跡を定量化する．なお，文献 [6] では max-margin 法と projection 法が提案されている．それぞれ QPsolver，正射影ベクトルの計算を用いて報酬関数を推定する．文献 [6] の実験では projection 法が収束率の点でわずかに良い性能を示していたため，本研究では projection 法を用いる．projection 法のアルゴリズムを図 1 に示す.

Compute $\boldsymbol{\mu}^{(0)}=\boldsymbol{\pi}^{(0)}$
Set $\boldsymbol{w}^{(1)}=\boldsymbol{\mu}_E-\boldsymbol{\mu}^{(0)}, i=1$
Repeat (until $t^{(i)}\leq\epsilon$)
  Compute $\boldsymbol{\pi}^{(i)}$ using the RL algorithm and rewards $\boldsymbol{R}=(\boldsymbol{w}^{(i)})^T\boldsymbol{\phi}$
  Compute $\boldsymbol{\mu}^{(i)}=\boldsymbol{\mu}(\boldsymbol{\pi}^{(i)})$
  Set $i=i+1$
  Set $\bar{\boldsymbol{\mu}}^{(i-1)}=\bar{\boldsymbol{\mu}}^{(i-2)}+\frac{(\boldsymbol{\mu}^{(i-1)}-\bar{\boldsymbol{\mu}}^{(i-2)})^T(\boldsymbol{\mu}_E-\bar{\boldsymbol{\mu}}^{(i-2)})}{(\boldsymbol{\mu}^{(i-1)}-\bar{\boldsymbol{\mu}}^{(i-2)})^T(\boldsymbol{\mu}^{(i-1)}-\bar{\boldsymbol{\mu}}^{(i-2)})}(\boldsymbol{\mu}^{(i-1)}-\bar{\boldsymbol{\mu}}^{(i-2)})$
  Set $\boldsymbol{w}^{(i)}=\boldsymbol{\mu}_E-\bar{\boldsymbol{\mu}}^{(i-1)}$
  Set $t^{(i)}=||\boldsymbol{\mu}_E-\bar{\boldsymbol{\mu}}^{(i-1)}||_2$

図 1: Algorithm of projection method

なお，エキスパートの特徴期待値 $\boldsymbol{\mu}_E$ は，エキスパートの m 試行の行動軌跡 $\{s_0^{(i)},s_1^{(i)},\ldots\}_{i=1}^m$ から式 (5) によって推定する.

$$\hat{\boldsymbol{\mu}}_E=\frac{1}{m}\sum_{i=1}^{m}\sum_{t=0}^{\infty}\gamma^t\boldsymbol{\phi}(s_t^{(i)}) \quad (5)$$

3. 提案手法

3.1 既存研究の限界

事前知識を用いない手法では，一度ゴールにたどり着く必要があることから学習初期に多くの時間を要する，試行錯誤によりサブゴールを発見するため不適切なサブゴールを設定する可能性があるなどの問題がある.

事前知識を用いる手法では，得られた報酬関数で最適な行動を学習できるが，学習に要する速度については考慮していないという問題がある．すなわち，既存の逆強化学習 [5][6] によって複数の状態に報酬値が与えられるが，それらをサブゴールとみなした場合，これによって学習速度を最小化できるとは限らない．そこで，学習速度を最小化するサブゴールを発見する仕組みを導入した逆強化学習を考える.

表 1: Comparison of problem definition of the proposed method and the previous method

	Ng の逆強化学習	提案手法
目的関数	式 (1)	Q 学習に要するステップ数の最小化
制約条件	式 (2)	式 (2)
要素技術	LPsolver	PSO

3.2 提案手法の特徴

本研究では Ng の逆強化学習に着目し，目的関数を学習速度を考慮したものに拡張する．提案手法の既存手法からの変更点を表 1，アルゴリズムを図 2 に示す．提案手法の目的関数は，Q 学習を終了するまでに要するステップ数の最小化とする．Q 学習の終了には次の二つの指標が考えられる.

1. 最短経路を発見
2. 全状態で最適行動を発見

本手法では，あるエピソードを終えるまでに要したステップ数を用いて評価している．なお，提案手法では制約条件をペナルティ法を用いて表現し，要素技術にメタヒューリスティクスの一つである Particle Swarm Optimization （PSO）を用いる.

Compute constraints of inverse Reinforcement Learning
Repeat (for each iterations)
  Update the reward function by PSO
  If satisfying the all constraints
    Compute total number of steps using Q learning
    Set the total number of steps in the evaluation value
  If breaking the constraints
    Set the penalty of number of breaking the constraints

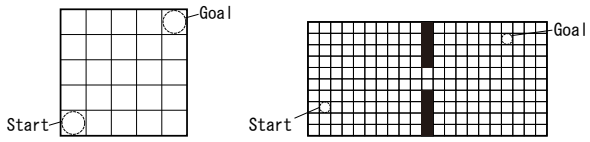
図 2: Algorithm of proposed method

4. 実験

提案手法の性能を評価するために実験を行う．実験には図 3(a),(b) に示す 5×5-GridWorld と Two-Rooms と呼ぶスタートからゴールまでの最短経路を求める問題を用いる．Two-Rooms は，スタートとゴールの間に存在する壁によって，二つの部屋に分けられた環境で，有効なサブゴールの設定によって学習の高速化が期待できる問題である.

4.1 5 × 5-GridWorld の実験結果

図 4 に既存手法の学習速度との比較を示す．縦軸はゴールに至るまでに要したステップ数，横軸はエピソード数である.



(a) 5x5-GridWorld (b) Two-Rooms

図 3: Experimental Environment

なお，Normal はゴールだけに報酬を与えたものを表す．また，Ng の逆強化学習の評価は，予備実験として Ng-1 と Ng-2 の比較を行った結果，最も良い性能を示した Ng-2 の $|\boldsymbol{R}| < 1$, λ=0 を用いる．実験は学習率 0.03，割引率 0.9，ϵ=0.3 を用いて，ゴールにたどり着くまでを 1 エピソードとして 2000 エピソードまでを 1 試行として 100 試行の平均値を出力した．

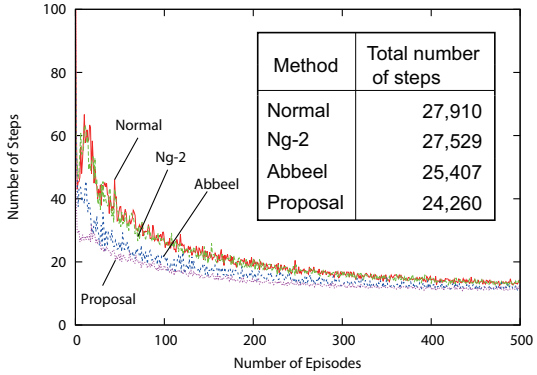


図 4: Comparison of convergence curves (5×5-GridWorld)

図 4 から，提案手法，Abbeel，Ng，Normal の順に，学習速度と収束が速いことがわかる．

4.2 Two-Rooms の実験結果

図 5 に既存手法の学習速度との比較を示す．図 5 は比較を容易にするため，10 エピソードの平均値をプロットした．縦軸はゴールに至るまでに要したステップ数，横軸はエピソード数である．なお，Normal はゴールだけに報酬を与えたものを表す．また，Ng の逆強化学習の評価は，予備実験として Ng-1 と Ng-2 の比較を行った結果，最も良い性能を示した Ng-2 の $|\boldsymbol{R}| < 1$,λ=1 を用いる．実験は学習率 0.03，割引率 0.9，ϵ=0.3 を用いて，ゴールにたどり着くまでを 1 エピソードとして 5000 エピソードまでを 1 試行として 10 試行の平均値を出力した．

提案手法は Two-Rooms では実行可能な解を得ることができなかった．図 5 から，Ng，Normal，Abbeel の順に学習速度と収束が速いことがわかる．

5. 考察

5.1 5 × 5-GridWorld

学習速度の差が誤差によるものか確かめるため，Ng と Normal，Abbeel と提案手法に対して両側 5%の t 検定を行ったところそれぞれ有意差があるとなった．各手法で得られた報酬関数から学習高速化の要因を考察する．各手法で得られた報酬関数を図 6 に示す．各状態で 1 つ先の状態と 2 つ先の状態に遷

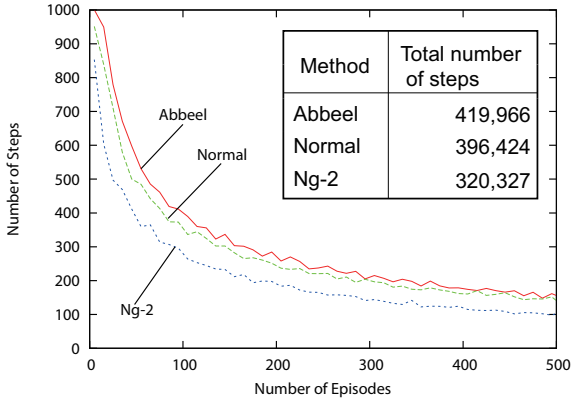


図 5: Comparison of convergence curves (Two-Rooms)

移するときに得られる累積報酬を最大化する行動を表 2，その行動が最適行動である割合を表 3 に示す．

表 2: Comparison of acquired actions by each methods

	0	1	2	3	4
4	→	←	→	→	-
3	↑	↓	→	↑	↑
2	→	→	↓	↑	↑
1	→	→	↑	←	↑
0	→	↑	→	↑	↑

	0	1	2	3	4
4	→	→	→	→	-
3	→	↑	→	→	↑
2	→	→	→	↑	↑
1	→	→	↑	↑	↑
0	→	↑	↑	↑	↑

(a) By Abbeel's : Evaluate $\boldsymbol{R}$ by each step (left), every 2 steps (right)

	0	1	2	3	4
4	-	←	→	→	-
3	↑	-	→	↑	↑
2	-	-	-	↑	-
1	-	-	-	-	↑
0	-	-	-	↑	-

	0	1	2	3	4
4	-	→	→	→	-
3	↑	→	→	↑	↑
2	↑	-	→ or ↑	↑	↑
1	-	-	-	↑	↑
0	-	-	→	→	↑

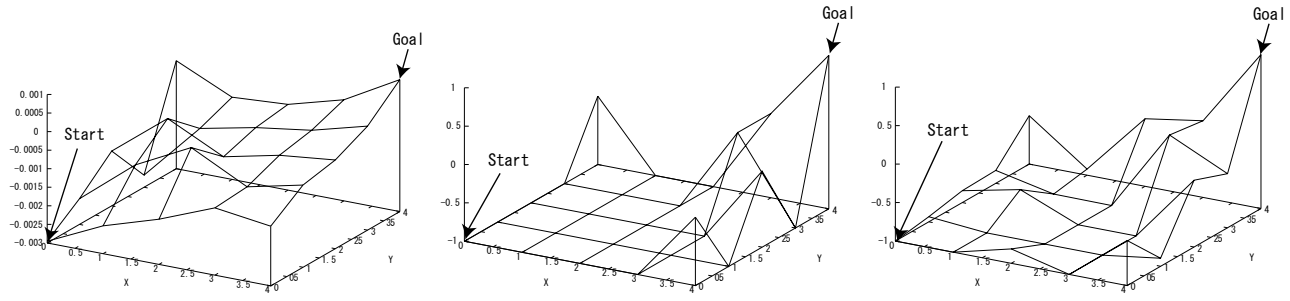
(b) By Ng's : Evaluate $\boldsymbol{R}$ by each step (left), every 2 steps (right)

	0	1	2	3	4
4	↓	→	→	→	-
3	↑	→	→	↑	↑
2	→	←	←	↑	↑
1	↑	↑	↓	→	↑
0	↑	→	-	→	↑

	0	1	2	3	4
4	→	→	→	→	-
3	↑	→	→	↑	↑
2	↑	←	→ or ↑	↑	↑
1	↑	↑	↑	→	↑
0	↑	→	→	→	↑

(c) By Proposal's : Evaluate $\boldsymbol{R}$ by each step (left), every 2 steps (right)

表 3 より，二つ先の報酬まで見ると，提案手法と Abbeel の報酬関数はそれぞれ 96 %，100 %の割合で最適な行動を示しており，ほとんどの状態でゴールにたどり着かなくても最適な行動を学習する報酬関数が得られていることがわかる．一方 Ng では 71 %である．このことから提案手法と Abbeel は Ng より学習が速くなった．また，Abbeel より提案手法の報酬関数で学習が速くなった理由は，Abbeel では報酬関数の値にほとんど差がないことから，学習中の探索により，状態価値がすぐに逆転してしまうため，最適行動の発見に時間がかかるのではないかと考えている．以上の理由から，提案手法では多くの状態で適切な報酬が得られることから報酬の遅れが小さくなり，学習が速くなったと考えられる．



(a) Abbeel (b) Ng (c) Proposal

図 6: Reward function of 5 × 5-GridWorld

表 3: The mostly rewarded action of each state

	最適行動の割合 [%]	
手法	5 × 5-GirdWorld	Two-Rooms
Abbeel (each step)	83	56
Abbeel (2 step)	100	52
Ng (each step)	33	63
Ng (2 step)	71	73
Proposal (each step)	79	-
Proposal (2 step)	96	-

5.2 Two-Rooms

5 × 5-GridWorld と同様に，各状態で 1 つ先の状態と 2 つ先の状態に遷移するときに得られる累積報酬を最大化する行動が最適行動である割合を表 3 に示す.

提案手法では Two-Rooms に対して制約条件を全て満たす解を得ることができなかった．これは，状態数と制約条件が多いことや，制約条件を緩和するにしても，PSO による解の探索時には，どの制約条件を破っているかが特定できないためである．一方，Ng と Abbeel の逆強化学習では，表 3 より，二つ先の報酬まで見ると，Abbeel と Ng の報酬関数はそれぞれ 52 %，73 %の割合で最適な行動を示している．このため，Ng の手法によって得られた報酬関数を用いた学習が最も速くなったと考えられる.

6. 結論

本研究は，強化学習における報酬の遅れに起因する学習時間の増大に着目し，適切なサブゴールの設定によって学習を高速化する方法を提案した．サブゴールの設定方法として，逆強化学習を適用するにあたり，既存の逆強化学習を拡張し，学習の速度を考慮に入れたものに変更することで学習を高速化する報酬関数を得ることができた.

今後の課題として次の二つを考えている．一つ目は，提案手法で導入した PSO の制約条件に関する課題である．PSO は制約条件の増加に伴い実行可能解を得ることが困難になる問題があり，この結果 Two-Rooms の解を得ることができなかった．PSO の制約条件の処理法に何等かの工夫が必要である.

二つ目は，学習速度の評価方法に関してである．現状では収束とみなすことができる適当なエピソード数までの累積ステップ数を評価指標に用いているが，この数値は exploration のパラメータである ϵ など「環境の同定に要する時間」を考慮した評価である．しかし，学習の高速化といった場合,「ゴールまでの最短経路さえ発見できればよい」とする立場では，必ずしも全ての状態で最適な行動を獲得しなくてもよいとも考えられる．特に有限でない空間を対象とする場合は後者を評価指標とすべきとも考えられるが，Q-learning のような Bootstrap の方法をベースにした場合には，最短経路上にない状態に対しても正しい行動が獲得されている必要があるため，考察でも述べた通り，学習速度と，各状態での最適行動の獲得割合は非常に密接な関係がある．今後は，高速化に対する考え方を明確にし，PSO の評価関数を定義し直すことによって，提案手法の洗練化をめざす.

参考文献

[1] Richard S. Sutton, Andrew G. Barto: Reinforcement Learning: An Introduction, 三上貞芳，皆川 雅章訳: ”強化学習”, 森北出版， pp.142-170, (2000)

[2] 木村卓哉 : 強化学習における事前知識を用いないサブゴールの設定，北海道大学大学院情報科学研究科コンピュータサイエンス専攻修士論文要旨，pp.1-4,（2012）

[3] 萩原史典，高野浩貴 , 村田純一 : 強化学習における多段サブゴールの逐次的発見による学習高速化，システム・情報部門学術講演会，pp.203-208,（2011）

[4] Amy McGovern，Andrew G. Barto : Automatic Discovery of subgoals in Reinforcement Learning using Diverse Density, To appear in the 2001 International Conference on Machine Learning， pp.1-8,（2001）

[5] Andrew Y. Ng，Stuart Russell : Algorithms for Inverse Reinforcement Learning，In Proceedings of the Seventeenth International Conference on Machine Learning，pp.663-670, (2000)

[6] Pieter Abbeel，Andrew Y. Ng : Apprenticeship Learning via Inverse Reinforcement Learning, In Proceedings of the 21st International Conference on Machine Learning, pp.1–8, (2004)

[7] Stuart Russell : Learning agents for uncertain environments (extended abstract), In Proceedings of the 16th International Conference on Machine Learning, pp.278-287, (1998)

[8] 相吉英太郎，安田恵一郎 : メタヒューリスティクスとその応用，電気学会，pp.69-90, (2007)