3O1-6in

逆強化学習による報酬関数推定における目的関数の影響の考察

Effect of Objective Function on Estimated Rewards by Inverse Reinforcement Learning

*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

Inverse Reinforcement Learning (IRL) is a promising framework for estimating a reward function under the given optimal policy. An original idea of IRL is proposed by Russell et.al. where the objective function is calculated by summation of the difference of maximum Q-value and other Q-value of a state over the whole states, and then maximized this value. While, in this paper, the objective function is defined by each state. That is, there are multiple objective functions. Each objective function is calculated by the difference of maximum Q-value and other Q-value of a state.

Then, we solve an IRL problem as the multiobjective optimization problem. This paper is shown the effectiveness of our defined objective function for estimating the reward function via some experiments.

1. はじめに

強化学習 [Sutton 98] 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規則を獲得するための枠組みである．強化学習においてこれまで報酬は所与とされてきたが，複雑な問題では報酬を手動で設定することが難しい．

この問題を解決するための手法として逆強化学習がある．逆強化学習は，Russell [Russell 98] 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 [Ng 00] 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 [Abbeel 04] 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．

今回注目する Ng の逆強化学習では，最適行動を獲得するために報酬が満たすべき条件を制約条件として表し，制約条件を満たす報酬の中から妥当なひとつを選ぶためにヒューリスティックな目的関数を導入した．この目的関数は，各状態の最適行動と，それ以外の行動の Q 値の差を全状態に対して算出し，この合計を最大化するというものである．

本研究では，各状態ごとに Q 値の差を最大化する多目的最適化問題とした 3 通りの解法を示し，それぞれの目的関数を評価する．多目的最適化問題を解くアルゴリズムには，メタヒューリスティクスのひとつである NSGA-II[Deb 02] を用いる．

2. 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 を調整することで，獲得する報酬の大きさをコントロールする．$\boldsymbol{R}_{max}(>0)$ は報酬の上下限を設定するパラメータである．

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

$$\begin{aligned}\text{subject to}: (\boldsymbol{P}_{a_1}-\boldsymbol{P}_a)(\boldsymbol{I}-\gamma\boldsymbol{P}_{a_1})^{-1}\boldsymbol{R} \geq 0 \\ |\boldsymbol{R}_i| \leq R_{max}, \quad i=1,\ldots,N\end{aligned}$$

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

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

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，その他の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．

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

本稿では，式 (4) を用いる．

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，arai@tu.chiba-u.ac.jp

3. NSGA-II

NSGA-II は制約付き多目的最適化問題を解くために GA を拡張した手法である．比較的実装が容易で高性能であることから，EMO（Evolutionary Multiobjective Optimization : 進化型多目的最適化）の分野で最もよく用いられている．NSGA-II では，ランキング割り当て，多様性維持メカニズム，エリート保存戦略，制約条件の取り扱いの 4 つの特徴がある．次にそれぞれについて詳しく説明する．

3.1 優越関係に基づくランキング

多くの EMO 手法では，個体の評価は優越関係に基づき行われる．しかし，優越関係を用いて直接的に 2 個体の比較を行うと，多くの場合で比較不可能となる．そのため，優越関係を用いて個々の個体にランクを割り当てるランキングという操作を行う．NSGA-II のランキングでは，はじめに，現在の個体群の中でほかの個体に優越されない非劣個体に対してランク 1 が与えられる．次に，ランク 1 が与えられた非劣個体を除いた個体群のなかで，ほかの個体に優越されない非劣個体に対してランク 2 が与えられる．すべての個体に対してランクが割り当てられるまで，この操作を繰り返す.

3.2 密集する個体の適応度を相対的に下げる多様性維持メカニズム

個体群のなかで同じランクの個体だけを考え，個々の目的関数ごとに，対象とする個体の左右に位置する 2 個体間の距離を計算し，すべての目的関数に関する距離の総和を密集度の測度として個体に割り当てる．もし，対象とする個体が同じランクの個体のなかで少なくとも一つの目的関数に関して最小値または最大値となる場合は，その個体の密集度の測度を無限大とする．

3.3 非劣個体の集合を保存するエリート保存戦略

親個体群と子個体群をあわせて個体群サイズが 2 倍の合併個体群を生成し，合併個体群に対してランキングと密集度の計算が行われる．ランクの高い順 (ランクの値が小さい順) に合併個体群から個体を順番に選択することで次世代の個体群が構成される．

3.4 制約条件の取り扱い方法

NSGA-II では，ペナルティ係数を使わずによりよい解を探索する．具体的には，以下の三つのルールを用いる．以下の条件を満たすとき，解 i が解 j よりよい解である

- 解 i が実行可能解 (feasible) であり，解 j が実行不可能解 (infeasible) であるとき
- 解 i, j がともに実行不可能解であり，解 i が制約条件を破る個数が少ないとき
- 解 i, j がともに実行可能解であり，解 i が解 j を優越するとき

このルールを用いることで，実行可能解はどの実行不可能解よりもよいランクを持つ．また，制約条件を破る個数が少ない解がよいランクを持つ.

3.5 NSGA-II のアルゴリズム

NSGA-II のアルゴリズムを疑似コードで以下に示す．なお，P は親個体群 (Population)，P' は子個体群を表わす.

1. P:= Initialize(P)
2. while stop_criterion not satisfied do
3. P':=Genetic Operations(P)
4. P:= Replace($P \cup P'$)
5. end while
6. return(P)

NSGA-II では，親個体選択においてランキングと多様性維持メカニズムを用い，世代更新でエリート保存戦略を実現している.

4. 逆強化学習の多目的化と好ましい解の選定法

Ng が提案した目的関数は，各状態の最適行動と，それ以外の行動の Q 値の差を全状態に対して算出し，この合計を最大化するというものである．本稿では，さらに最適行動を獲得しやすい，精度の高い報酬を獲得するために，状態ごとに Q 値の差を最大化する目的関数を提案する.

4.1 逆強化学習の目的関数の多目的化

Ng の逆強化学習に対して，3 通りの多目的の目的関数を提案する.

$$\text{maximize}: \ (\boldsymbol{P}_{a_1} - \boldsymbol{P}_{a_2})(\boldsymbol{I} - \gamma \boldsymbol{P}_{a_1})^{-1}\boldsymbol{R} - \lambda ||\boldsymbol{R}||_1 \boldsymbol{I} \tag{5}$$

$$\text{maximize}: \ \sum_{j=1}^{K} \{(\boldsymbol{P}_{a_1} - \boldsymbol{P}_{a_j})(\boldsymbol{I} - \gamma \boldsymbol{P}_{a_1})^{-1}\boldsymbol{R}\} - \lambda ||\boldsymbol{R}||_1 \boldsymbol{I} \tag{6}$$

$$\text{maximize}: \ (\boldsymbol{P}_{a_1}(i) - \boldsymbol{P}_{a_2}(i))(\boldsymbol{I} - \gamma \boldsymbol{P}_{a_1})^{-1}\boldsymbol{R}) - \lambda ||\boldsymbol{R}||_1 \boldsymbol{I} \quad i = 1, 2, \ldots, N \tag{7}$$

式 (5) は，状態ごとに最適行動と 2 番目によい行動の Q 値の差の最大化を行う．これは，状態ごとの Q 値を最大化することにより，式 (4) のすべての状態の Q 値の合計を最大化するよりも精度の高い報酬の推定が期待できる．式 (6) は，状態ごとに最適な行動とその他のすべての行動との Q 値の差の合計の最大化を行う．これは 2 番目によい行動だけでなく，すべての行動を考慮することで最適行動を獲得しやすい報酬の推定が期待できる．式 (7) は，状態ごとに最適な行動とその他のすべての行動との Q 値の差の最大化を行う．これは，式 (5) と式 (6) を組み合わせたものであり，決定変数が行動数だけ増加するが，さらに精度の高い報酬の推定が期待できる．以後，式 (5)，式 (6)，式 (7) を用いた解法をそれぞれ NSGA-1, NSGA-2，NSGA-3 と呼ぶ.

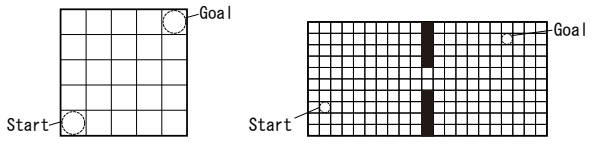
4.2 好ましい解の選定法

目的関数を多目的化したことにより，多数のパレート解が存在する最適化問題となる．多目的最適化問題を解いた後，得られたパレート解集合 R_{pareto} の中から適切なひとつの解（好ましい解）を選ぶために，次の 3 通りの解の選定法を提案する.

$$\arg\max_{R \in R_{\text{pareto}}} \ \sum_{i}^{N} (\boldsymbol{P}_{a_1}(i) - \boldsymbol{P}_{a_2}(i))(\boldsymbol{I} - \gamma \boldsymbol{P}_{a_1})^{-1}\boldsymbol{R} \tag{8}$$

$$\arg\max_{R \in R_{\text{pareto}}} \ \sum_{i}^{N} b_i \tag{9}$$

$$b_i = \begin{cases} 1 & ((\boldsymbol{P}_{a_1}(i) - \boldsymbol{P}_{a_2}(i))(\boldsymbol{I} - \gamma \boldsymbol{P}_{a_1})^{-1}\boldsymbol{R} > 0) \\ 0 & (\text{otherwise}) \end{cases}$$



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

図 1: Experimental Environment

$$\arg\max_{R \in R_{\text{pareto}}} \sum_i^N \sum_j^K c_{ij} \tag{10}$$

$$c_{ij} = \begin{cases} 1 & ((\boldsymbol{P}_{a_1}(i) - \boldsymbol{P}_{a_j}(i))(\boldsymbol{I} - \gamma \boldsymbol{P}_{a_1})^{-1}\boldsymbol{R} > 0) \\ 0 & (\text{otherwise}) \end{cases}$$

式 (8) は，Ng の目的関数を踏襲したものであり，最適行動の Q 値と 2 番目によい行動の Q 値の差の全状態の合計が大きい報酬を選択する．式 (9) は，最適行動の Q 値と 2 番目によい行動の Q 値に差がある状態の数が大きくなる報酬を選択する．これは，多くの状態で最適行動を獲得しやすい報酬の選択が期待できる．式 (10) は，最適行動の Q 値とその他の行動の Q 値に差がある状態の数が大きくなる報酬を選択する．これは，式 (9) に 2 番目以外の行動も考慮に入れたものである．

5. 計算機実験

図 1(a),(b) に示す環境を用いて，提案した目的関数の評価を行う．図 1(a) は，スタートを座標 (0,0)，ゴールを座標 (4,4) に持つ迷路問題である．図 1(b) は，スタートを座標 (1,2)，ゴールを座標 (17,8) に持ち，スタートとゴールの間に壁がある複雑な迷路問題である．

各目的関数の評価は最適性と収束性に関して行う．最適性は，各目的関数で獲得した報酬を用いて学習し得られた最適行動と，逆強化学習で所与とした最適行動との一致率で評価する．収束性は，エピソードとステップをプロットしたグラフから評価する．強化学習のアルゴリズムは Q 学習を用いる．また，好ましい解の選択法は，予備実験の結果から式 (9) を用いる．

5.1 5x5-GridWorld の実験結果

Q 学習のパラメータは，学習率 0.1，割引率 0.9，ϵ=0.1，上限ステップ数なし，エピソード数 2000 (1900 エピソード以降 ϵ=0)，とする．また，行動選択には ϵ-greedy 選択を用いる．また，NSGA-II のパラメータは，世代数 500，個体数 50，交叉率 0.9，突然変異率 0.01 を用いる．

5.1.1 最適性

各目的関数で獲得した報酬を用いて学習し，最終エピソードで最も Q 値が大きい行動と，逆強化学習で所与とした行動の一致率を表 1 に示す．Ng の逆強化学習と比較すると提案した目的関数のほうが最適性が高い．

5.1.2 収束性

横軸をエピソード数，縦軸をステップ数でプロットした結果を図 2 に示す．

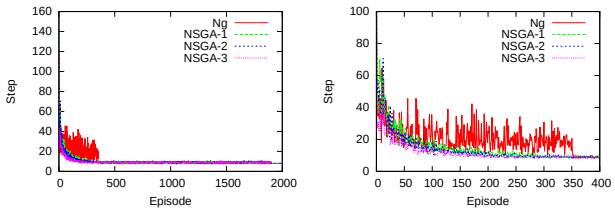
NSGA-3，NSGA-2，NSGA-1，Ng の順に収束性が高い．

5.2 Two-Rooms の実験結果

Q 学習のパラメータは，学習率 0.1，割引率 0.9，ϵ=0.6，上限ステップ数なし，エピソード数 5000 (4900 エピソード以降 ϵ=0)，とする．また，行動選択には ϵ-greedy 選択を用いる．また，NSGA-II のパラメータは，世代数 500，個体数 50，交叉率 0.9，突然変異率 0.01 を用いる．

表 1: Comparison of concordance rate by each methods

Method	Concordance rate[%]
Ng	75
NSGA-1	92
NSGA-2	96
NSGA-3	96



(a) Full scale (b) Enlarged

図 2: Comparison of convergence curves (5x5-GridWorld)

5.2.1 最適性

Two-Rooms では，試行ごとに得られた最適行動に大きなばらつきがあったため，所与とした最適行動との一致率にもばらつきがあり，比較することができなかった．

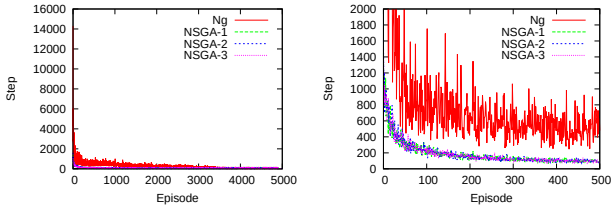
5.2.2 収束性

横軸をエピソード数，縦軸をステップ数でプロットした結果を図 3 に示す．Ng の収束性が低く，NSGA-1，NSGA-2，NSGA-3 の収束性が高い．NSGA-1，NSGA-2，NSGA-3 の間には大きな差はない．

6. 考察

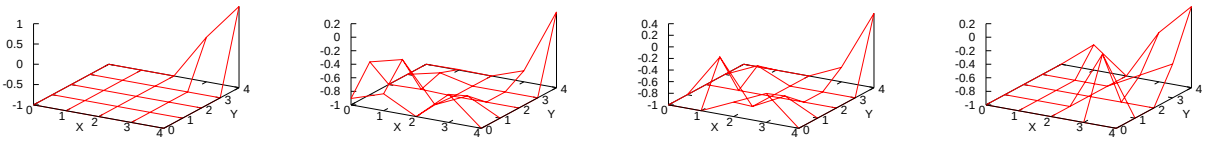
図 4 と図 5 にそれぞれ 5x5-GridWorld と Two-Rooms に各目的関数を適用し，得られた報酬関数をまとめる．

はじめに 5x5-GridWorld において，提案した目的関数で最適性と収束性が向上した理由を考察する．最適性は，Ng の目的関数では，Q 値の差を全状態で合計したものを最大化しているため，図 4(a) に見られるように，Q 値の差を大きくしやすい特定の状態に集中して報酬を与えている．提案した目的関



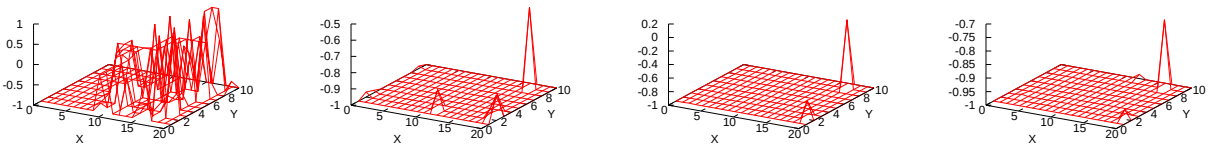
(a) Full scale (b) Enlarged

図 3: Comparison of convergence curves (Two-Rooms)



(a) Ng (b) NSGA-1 (c) NSGA-2 (d) NSGA-3

図 4: Reward function (5x5-GridWorld)



(a) Ng (b) NSGA-1 (c) NSGA-2 (d) NSGA-3

図 5: Reward function (Two-Rooms)

数では，各状態で別々に Q 値の差を最大化しているため，図 4(b)(c)(d) のように，多くの状態に報酬を与えている．このことから，多くの状態で最適行動を獲得しやすい，精度の高い報酬が獲得できたと考えられる．収束性は，多くの状態に報酬が設定されていること，精度の高い報酬が設定されたことにより向上した．

次に Two-Rooms において，獲得した最適行動にばらつきが生じた理由を考察する．Two-Rooms では，各状態においてゴールに向かう行動が最適行動となるが，最適行動が複数存在する状態が多数存在する．例えば，スタート地点（座標 (1,7)）では，上と右の両方の行動が最適行動となる．ここで獲得した報酬を確認すると，図 5(a) では左半分，(b)(c)(d) では全体的に報酬が設定されていない状態が多数存在する．このため，初めに設定した最適行動以外の最適行動にも収束する可能性があることから，ばらつきが生じたと考える．各状態で多目的化することにより，多くの状態に報酬が設定されることを期待していたが，多目的化したことで問題が複雑になったこと，決定変数が増加したことが原因となり，局所解に収束したと考えている．収束性については，Ng の目的関数で得られた報酬（図 5(a)）では多くの状態に報酬が設定されているが，上述のように Q 値の差を大きくしやすい状態に集中して報酬を与えているため，最短経路以外の状態で最適行動の Q 値を大きしやすい状態が存在した場合，最短経路の学習に悪影響を与えた可能性がある．このことから収束性が下がったと考える．また提案した目的関数の間では，図 5(b)(c)(d) のように，ほとんど同じ形状の報酬関数が得られているため，収束性に差が生じなかったと考えられる．

7. まとめ

本稿では，報酬設定が困難な問題に対して期待されている逆強化学習において，Ng によって提案された逆強化学習の目的関数に着目し，各状態の Q 値の差を最大化する多目的最適化問題として定式化し，各目的関数によって得られた報酬関数について考察した．

今後の課題は，状態数が多い環境では，問題が複雑になり決定変数が増加することによる，解の精度の低下があげられる．大規模な問題でも所与とした最適行動を獲得できる目的関数を考える必要がある．また，ほかの問題にも提案した目的関数を適用し，有用性を確認する必要がある．

参考文献

[Sutton 98] 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)

[Russell 98] Stuart Russell : Learning agents for uncertain environments (extended abstract), In Proceedings of the 16th International Conference on Machine Learning, pp.278-287, (1998)

[Ng 00] Andrew Y. Ng, Stuart Russell : Algorithms for Inverse Reinforcement Learning, In Proceedings of the Seventeenth International Conference on Machine Learning, pp.663-670, (2000)

[Abbeel 04] 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)

[Deb 02] Kalyanmoy Deb, Amrit Pratap, Sameer Agarwal, and T. Meyarivan : A Fast and Elitist Multi-Objective Genetic Algorithm: NSGA-II, IEEE Transactions on Evolutionary Computation, pp.1-20, (2002)