様　式　Ｃ－１９、Ｆ－１９、Ｚ－１９　（共通）

# 科学研究費助成事業　　研究成果報告書

平成　２７　年　　５　月　１２　日現在

機関番号：　３８００５
研究種目：　基盤研究(C)
研究期間：　2012～2014
課題番号：　２４５００２４９

研究課題名（和文）強化学習のための情報理論に基づく報酬の設計論

研究課題名（英文）Information theoretic optimization of intrinsic rewards for reinforcement learning

研究代表者
内部　英治（Uchibe, Eiji）

沖縄科学技術大学院大学・神経計算ユニット・グループリーダー

研究者番号：２０４２６５７１

交付決定額（研究期間全体）：（直接経費）　　3,800,000円

研究成果の概要（和文）：本研究では密度比推定に基づく新しい逆強化学習法を提案する。アルゴリズムを導出するために、推定される報酬にカルバックライブラー情報量で制約を与える。この結果、最適確率的制御則と基準となる制御則の対数比が報酬と価値関数によって表現される。従来法の大半が制御開始から終了までの状態系列の集合をデータとして用いるのに対し、提案手法は状態遷移の集合をデータとして用いることができるので非常にデータ効率が良い。ロボットのナビゲーション課題に適用し、提案手法は従来法よりも精度よく報酬を推定することができた。また、提案手法はシェーピングの理論と組み合わせることが可能で、順強化学習のスピードも改善できた。

研究成果の概要（英文）：This study investigates novel inverse reinforcement learning methods based on density ratio estimation. To derive the algorithm, we exloit constraints on reward by KL-divergence. We show that the logarithm of the ratio between the optimal policy and the baseline policy is represented by the state-dependent reward and the value function. Our method is data-efficient because those functions can be estimated from a set of state transitions while most of previous methods require a set of trajectories. In addition, we do not need to compute the integral such as evaluation of the partition function. The proposed method is applied into a real-robot navigation task and experimental results show its superiority over conventional methods. In particular, we show that the estimated reward and value functions are useful when forward reinforcement learning is performed with the theory of shaping reward.

研究分野：　知能ロボティクス

キーワード：　強化学習　逆強化学習

## １．研究開始当初の背景

我が国はロボット大国であり、研究者でない一般の人であっても様々なロボットを購入して楽しむことができる。中には自分でプログラミングすることで、ロボットを制御することができるものまである。しかし非専門家にとってロボットを思った通りに動かすのは困難であり、ペットなどのように様々な行動を学習させたいと思うのは自然な欲求であろう。

このような問題に対して強化学習とよばれる行動学習の枠組みがある。近年、制御理論や統計学習との融合による理論的発展だけでなく、ヒトの意思決定の計算モデルとして神経科学分野でも注目を集めている。

強化学習における重要な未解決問題の一つに、目的関数を規定する報酬関数の設計がある。報酬は行動の良し悪しを即時的に評価する値で、問題に応じて設計者が事前に準備することから外的報酬と呼ばれる。実用的な外的報酬を準備するためには、対象となるシステムに関する詳細な知識を必要とすることが多く、ロボットの非専門家である一般のユーザがロボットに学習の目的を与えることは非常に困難であった。

一方でロボット自身が学習の進捗状況に応じて生成する報酬は内的報酬と呼ばれる。これは心理学での好奇心や達成感といった概念で説明される。内的報酬を工学的に実現できればロボット学習はより現実的なものになるが、そのための数理モデルは専門家が試行錯誤的に設計しているだけであった。

## ２．研究の目的

従来の内的報酬は外的報酬に基づく強化学習と独立に設計されていたため、内的・外的報酬の関係が不明確であったことが本質的な問題点である。そこで本研究では、外的報酬は非専門家でも容易に準備できる単純なものに限定し、内的報酬は情報理論に基づき自律的に計算するという新たな報酬関数の設計論を提案する。二種類の報酬を組み合わせることで複雑な報酬を表現できるため、実問題にも適用可能な強化学習法となる。

## ３．研究の方法

これまで提案されてきた内的報酬の多くは情報量基準に基づくものであった。そこから着想を得て、強化学習に用いる報酬のクラスを情報量基準に従って制限し、強化学習で解くべきベルマン最適性方程式がどのように単純化されるかということから調査する。

制約の与え方としては、学習前と学習後の状態遷移の違いを表すカルバックライブラーと呼ばれる距離のようなものを用いる。このとき、強化学習で解かなければならない非線形ベルマン方程式が線形微分方程式に帰着できることが知られている。このことを利用して、非専門家が与えるデータから報酬を設計する新しい枠組みを開発する。

## ４．研究成果

### (1) 密度比推定に基づく逆強化学習アルゴリズム

学習前と後で状態遷移確率がどのように変化したかを報酬に与える制約として用いると、状態遷移確率の密度比が状態依存の報酬と価値関数の差分で表現されることがわかった。学習前と学習後の状態遷移に関するデータが十分に与えられた場合、密度比は直接推定できる。その後、報酬と価値関数は通常の最小二乗法による回帰問題として簡単に解ける。これは逆強化学習と呼ばれるアルゴリズムに分類され、一般にデータから報酬を唯一に定めることはできないが、カルバックライブラーによる制約がこの問題を緩和している。

開発した手法を人の倒立振子の学習やラットのレバー押し課題に適用した。最初の実験では複数の被験者に長さの異なる振子の振り上げ・安定化課題を学習してもらい、学習前と後の振子の状態の変化をもとに被験者がどのような報酬を用いていたかを推定した（図 1）。

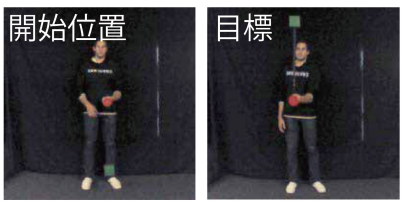


図 1 振子の振り上げ安定化課題

結果を図 2 に示す。横軸は振子の角度、縦軸は振子の角速度、色は報酬の強度を表し、青色の方が望ましい状態と解釈できる。振子の長さにかかわらずうまく学習できた被験者 4,7 のデータから推定された報酬は同じような形をしている。長い振子はうまく学習できたが、短い振子はそれほどうまくなかった被験者 5 の場合は、両者に対応する報酬の形状が大きく異なっていることも確認できた。

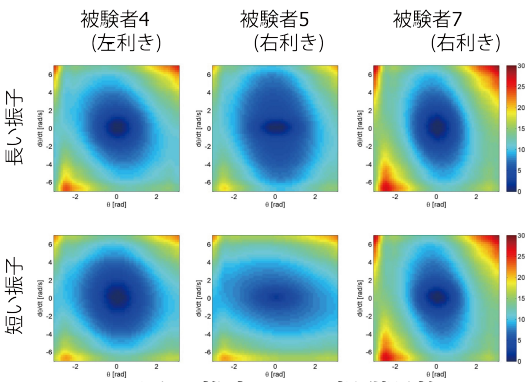


図 2 推定された報酬関数

次に異なる制約の与え方として、学習前と後で行動則がどのように変化したかを用いたとき、別の逆強化学習法を開発できた。この手法を検証するために、図 3a に示す環境においてロボットのナビゲーション実験を実施した。図 3b はロボットに搭載されたカメラから得られる画像で、簡単な色抽出処理によって環境の情報を取得している。ロボットの目的は図 3 中のスタート地点 A-E から緑色のターゲットまで移動することである。被験者はロボットを遠隔操作し、ロボットを緑色のターゲットまで誘導させ、そのときのデータをもとに報酬を推定する。


図 3 ロボットのナビゲーション課題

ここでは推定した報酬をもとに順強化学習を用いて行動を再現する際に、提案手法が同時に推定している価値関数を用いることで学習スピードが大幅に改善されることを示す。図 4 で赤線が提案手法で推定された報酬と価値関数を使って学習した場合、青線が提案手法で報酬だけを使って学習した場合、黒線が従来法で推定された報酬を使って学習した場合である。従来法は価値関数を推定しないため学習に時間がかかり、それは本件研究も同様であるが、価値関数を使うと学習スピードが大幅に改善されていることがわかる。

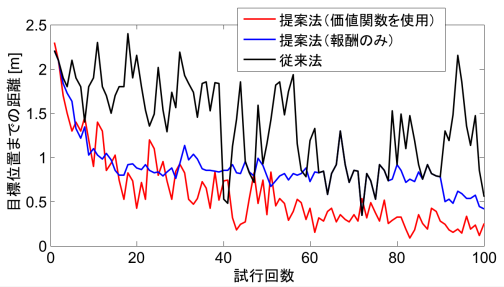


図 4 推定された報酬関数を使って順強化学習したときの学習速度の比較

アルゴリズムは ICDL-EpiRob 2014 の口頭発表として選ばれ、PCT 出願（国際特許出願）した。現在、目的関数を修正することで計算コストの低減と推定精度の改善を実現した本年度提案したアルゴリズムの改良版を国際英語論文誌に投稿中である。

ただし提案手法によって推定された報酬は内的報酬と外的報酬の和であって、両者を分割することはできなかった。ロボットの制御則を学習させることが最終目的である場合は現状でも問題ないが、人や動物の行動の原因を解析するためには報酬関数の近似器を工夫する必要があることが判明した。

(2) ロボット実験プラットホームの開発

申請時に実験に用いる予定であったロボットのパーツが 2014 年時点で製造中止になったことに伴い、使用するロボットを変更しなければならなかった。結果として申請時の予定には含まれなかった新しいロボットの開発をする必要があり、そのために予定外の時間を要した。

図 5 スマートフォンロボット

図 5 に示す新しいロボットはアンドロイドスマートフォンをベースにした安価な倒立二輪型の移動台車であり、可能な限りスマートフォン自体のセンサを用いる。ただし車輪のトルク不足のため、ロボットが転倒したときに自分自身では立ち上がることはできない。この問題に対処するために、右側のロボットは先端にばねを取り付けたバンパーを持っている。これによって転倒したロボットは振動子をベースにした単純な制御則でバネの運動エネルギーを蓄えたのちに、それを使って起き上がることができる。大半のパーツに市販品を採用しているため、パーツの製造中止にも柔軟に対応できる。またアンドロイドスマートフォンは毎年更新されるが、それにも対応できるようにソフトウェアを設計している。これは次年度以降の研究において非常に重要な成果であった。

## ５．主な発表論文等

（研究代表者、研究分担者及び連携研究者には下線）

〔学会発表〕（計　10　件）


① Wang, J., Uchibe, E., & Doya, K. (2015). Two-wheeled smartphone robot learns to stand up and balance by EM-based policy hyper parameter exploration. In Proc. of the 20th International Symposium on Artificial Life and Robotics.

② Uchibe, E., & Doya, K. (2014). Inverse reinforcement learning using dynamic policy programming. In Proc. of the 4th Joint IEEE International Conference on Development and Learning and on Epigenetic Robotics, pp. 222-228.

③ Uchibe, E., & Doya, K. (2014). Combining

learned controllers to achieve new goals based on linearly solvable MDPs. In Proc. of IEEE International Conference on Robotics and Automation, pp. 5252-5259.
④ Kinjo, K., Uchibe, E., & Doya, K. (2014). Robustness of linearly solvable Markov games with inaccurate dynamics model. In Proc. of the 19th International Symposium on Artificial Life and Robotics.
⑤ Wang, J., Uchibe, E., & Doya, K. (2014). Control of two-wheeled balancing and standing-up behaviors by an Android phone robot. Proc. of the 32nd Annual Conference of Robotics Society of Japan, Kyushu Sangyo University.
⑥ **内部英治, 銅谷賢治** (2014)**. 密度比推定を用いた逆強化学習. 第** 32 **回日本ロボット学会学術講演会予稿集, 九州産業大学.**
⑦ Wang, J., Uchibe, E., & Doya, K. (2013). Standing-up and balancing behaviors of Android phone robot. In Proc. of IEICE-NLP2013-122, 49-54.
⑧ Uchibe, E., Ota, S., & Doya, K. (2013). Inverse reinforcement learning for analysis of human behaviors. The 1st Multidisciplinary conference on reinforcement learning and decision making, Princeton, New Jersey, USA.
⑨ Ota, S., Uchibe, E., & Doya, K. (2013). Analysis of human behaviors by inverse reinforcement learning in a pole balancing task. The 3rd International Symposium on Biology of Decision Making, Paris, France.
⑩ **内部英治, 銅谷賢治** (2013)**. 密度比推定を用いた逆強化学習. 第** 16 **回情報論的学習理論ワークショップ** (IBIS2013)**.**

**〔産業財産権〕**
**○出願状況（計　1　件）**

**名 称 ：** Estimating goals using inverse reinforcement learning based on density ratio estimation
**発明者：**E. Uchibe and K. Doya
**権利者：**E. Uchibe and K. Doya
**種類：特許**
**番号：**US62/034510
**出願年月日：**2014 **年** 7 **月** 31 **日**
**国内外の別：外国**

**〔その他〕**
**ホームページ等**
https://groups.oist.jp/ja/ncu/adaptive-systems-group

**６．研究組織**
**(1)研究代表者**
**内部　英治**　（EIJI UCHIBE）
**沖縄科学技術大学院大学・神経計算ユニット・グループリーダー**
**研究者番号：**20426571