5T-4

POMDPs環境のための決定的政策を学習するProfit Sharingによるアーム型ロボットの行動学習

吉田匠汰　長名優子

東京工科大学 コンピュータサイエンス学部

1　はじめに

教師信号を用いずに環境との相互作用により適切な行動系列を獲得するための学習手法として，強化学習に関する様々な研究が行われている．強化学習で扱う問題環境のうちエージェントの知覚能力に限界がある問題環境を部分観測マルコフ決定過程 (POMDPs: Partially Observable Markov Decision Processes) 環境といい，POMDPs 環境で起こる問題を不完全知覚問題という．

POMDPs 環境でも比較的頑健な学習を行うことができる手法として，POMDPs 環境のための報酬獲得効率に基づく強化学習法 [1] が提案されている．この手法では報酬を獲得するために同じ観測に対して複数の行動をとる必要があるような環境では，報酬獲得のために必要な同一観測での行動が同じ確率で選択されるように学習が行われる．また，POMDPs 環境のための報酬獲得効率に基づく強化学習法を拡張し，過去の観測の系列を用いることで，同じ観測に対して複数の行動をとる必要がある場合にもより適切な確率で行動の選択が行える手法として POMDPs 環境のための決定的政策を学習する Profit Sharing[2] が提案されている．

本研究では，POMDPs 環境のための決定的政策を学習する Profit Sharing[2] を用いて，アーム型ロボットの行動学習を実現する．

2　POMDPs環境のための決定的政策を学習するProfit Sharing

POMDPs 環境のための決定的政策を学習する Profit Sharing では不完全知覚の生じている観測上において，報酬獲得に必要な行動の選択確率が等しくなるように学習が行われる．学習がある程度進行した段階で，行動の選択が決定的に行われているかどうかを行動の決定度を用いて判断し，行動の選択が確率的に行われているようであれば，それ以降のエピソードにおいてより過去の観測の情報を考慮した行動の選択を行えるようにする．

2.1　行動選択

POMDPs 環境のための決定的政策を学習する Profit Sharing では，現在の観測が不完全知覚状態であると判断されていない場合は，現在の観測におけるルールの価値の比に基づいてボルツマン選択により行動を選択する．また，現在の観測が不完全知覚状態であると判断されている場合は，現在の観測だけでなく観測の履歴を考慮したルールの価値の比に基づいて行動を選択する．ボルツマン選択において，温度パラメータは学習の初期には大きな値に設定され，学習が進むにつれて，小さくなる．このようにすることにより，学習の初期には確率的な行動選択が行われやすく，学習が進行するにつれて，より決定的な行動選択が行われるようにしている．この手法では，最終的には行動決定に必要な観測の履歴を考慮できるように学習が進むため，不完全知覚状態においても決定的な行動選択が可能となる．

2.2　学習

学習はエージェントが報酬を獲得した時のみ行われる．観測が不完全知覚状態と判断されていない場合は時刻 x での観測に関するルールの価値のみを更新する．また，観測が不完全知覚状態と判断されている場合は時刻 x での観測に関するルールの価値だけではなく，時刻 x での行動決定に関わる観測の系列に関するルールの価値も更新する．

3　ロボットの行動学習

本研究では，POMDPs 環境のための決定的政策を学習する Profit Sharing を用いてアームロボットの行動学習を実現する．実験は，アームロボットが物体を保持している状態から開始し，保持している物体を目標地点まで正確に運ぶことを目的として学習を行う．

Learning in Arm Robot by Profit Sharing that can Learn Deterministic Policy for POMDPs Environments
Shota Yoshida and Yuko Osana (Tokyo University of Technology, osana@stf.teu.ac.jp)

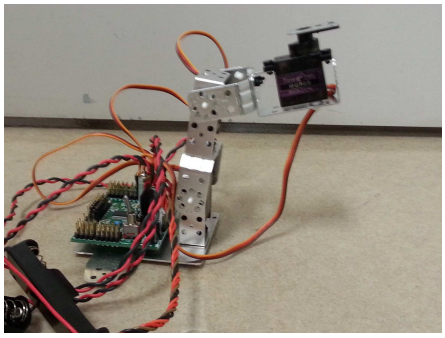

図 1: アーム型ロボット

3.1 ロボットの仕様

図 1 に実験で用いる自由度 3 のアーム型ロボットを示す．本研究ではロボットキットとしてプチロボ XL (共立電子産業) を使用する．アクチュエータとしてはサーボモータ (WR-MG90S, 共立電子産業) を用いる．

3.2 位置検出

アーム型ロボットによって移動させた物体の位置は，ZMP (Zero Moment Point) を用いて検出する．ロボットが物体を置く目標地点は四隅に圧力センサを取り付けた板の上とする．ZMP とは，床全体に分布してかかっている床反力をある一点にかかっているとして置き換えたときの作用点のことである．各点にかかる床反力によるモーメントの和とすべての力が ZMP にかかっているとしたときのモーメントの間には以下のような関係が成り立つ．

$$\sum_{i=1}^{N} x_i f_i = ZMP_x \sum_{i=1}^{N} f_i \tag{1}$$

$$\sum_{i=1}^{N} y_i f_i = ZMP_y \sum_{i=1}^{N} f_i \tag{2}$$

ここで，N はセンサの数，f_i はセンサ i の値，x_i はセンサ i の x 座標，y_i はセンサ i の y 座標，ZMP_x と ZMP_y はそれぞれ ZMP の x 座標と y 座標である．床反力を受けているすべての点の座標と床反力の大きさが分かれば，ZMP は次のように求めることができる．

$$ZMP_x = \frac{\sum_{i=1}^{N} x_i f_i}{\sum_{i=1}^{N} f_i} \tag{3}$$

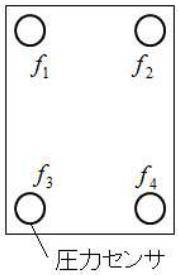

図 2: 圧力センサを取り付けた床

$$ZMP_y = \frac{\sum_{i=1}^{N} y_i f_i}{\sum_{i=1}^{N} f_i} \tag{4}$$

圧力センサを設置したところだけで床反力を受けるようにすれば，ZMP はこれらの式を用いて求めることができる．

3.3 行動の学習

本研究では，初期位置で物体を持った状態から板の上に物体を置くまでを一つのエピソードとして扱う．また，サーボモータの角度を観測，サーボモータの角度変化量を行動とする．目標地点に物体が置かれているときの ZMP とロボットが物体を床の上に置いたときの ZMP との差 (距離) に基づいて報酬を設定し，距離が短いほど多くの報酬を獲得できるものとする．

4 実験

POMDPs 環境のための決定的政策を学習する Profit Sharing[2] を用いてアームロボットの学習を行い，目標地点に物体を置く行動を学習できることを確認した．

参考文献

[1] 河合宏和, 上野敦志, 辰巳昭治： “POMDPs 環境のための報酬獲得効率に基づく強化学習法,” 人工知能学会論文誌, Vol.23, No.1, pp.1–12, 2008.

[2] Y. Takamori and Y. Osana : “Profit sharing that can learn deterministic policy for POMDPs environments,” Proceedings of SMC, Anchorage, 2011.