

POMDPs 環境のための決定的政策を学習する Profit Sharing による 四足歩行ロボットの歩行モーションの獲得

森野友哉 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。そのような手法のひとつとして POMDPs (Partially Observable Markov Decision Processes) 環境のための決定的政策を学習する Profit Sharing[1] 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 を用いて、四足歩行ロボットの歩行モーションの獲得を実現する。

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

2.1 概要

POMDPs 環境のための決定的政策を学習する Profit Sharing では、エージェントが環境から観測を受け取り、観測が過去に不完全知覚状態であると判断されていれば過去の観測の履歴を考慮した行動の選択を行う。不完全知覚状態であると判断されていない場合は、受け取った観測のみを考慮して行動の選択を行う。報酬を得るまでの観測と行動の系列をエピソードと呼び、エピソード終了後に、獲得した報酬をエピソード中の各時刻におけるルール の価値に分配していくことで、学習を進めていく。決定的な行動選択を行うために十分な観測の履歴の情報を参照できないと判断された場合には、さらに過去の観測の履歴を考慮することになる。また、不完全知覚状態かどうかの判定は観測ごとの行動の決定度と学習の進行度に基づいて行う。

POMDPs 環境のための決定的政策を学習する Profit Sharing の流れを図 1 に示す。

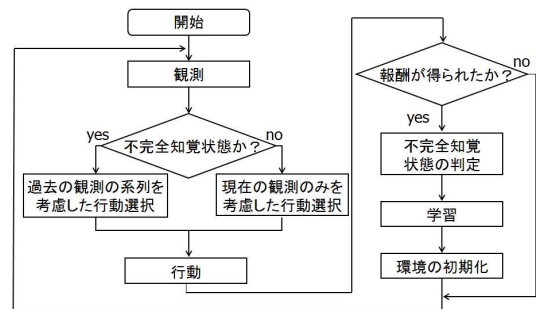


図 1: POMDPs 環境のための決定的政策を学習する Profit Sharing の流れ

2.2 行動選択

この手法では、エージェントの観測が不完全知覚状態と判定されていない場合、現在の観測におけるルール の価値の比に基づいて行動を選択する。また、観測が不完全知覚状態であると判断されている観測においては、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決定する。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の選択が可能となる。

2.3 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 四足歩行ロボットの歩行モーションの獲得

3.1 概要

本研究では、四足歩行ロボットにおいて少ないモーションで遠くまで移動できるような歩行モーション

Learning of Waking Motion in Four-legged Robot by Profit Sharing that can Learn Deterministic Policy for POMDPs Environments
Yuya Morino and Yuko Osana (Tokyo University of Technology, osana@stf.teu.ac.jp)

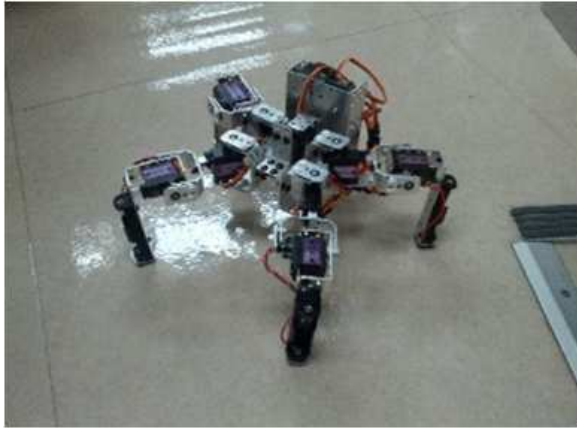


図 2: 使用するロボット

を獲得することを目的として学習を行う。また、ロボットのモーションを手動で再設定する必要があるが、POMDPs 環境のための決定的政策を学習する Profit Sharing[1] を用いて学習を行うことにより、環境の変化に応じたモーションの獲得を実現できる。

3.2 ロボットの仕様

図2に実験で用いる四足歩行ロボットを示す。本研究で使用するロボットキットは共立電子産業から発売されているプチロボ XL を使用している。アクチュエータには同メーカーのサーボモータ (WR-MG90S) を使用しており、関節の数は9つとなっている。また、移動距離を検出する為にロボットの前面に PSD (Position Sensitive Detector) 距離センサを搭載しており、このセンサは約 10~80cm 間の距離を測定することが可能である。

3.3 四足歩行ロボットの効率的な歩行モーションの獲得

本研究では、図3に示すような環境で実験を行う。四足歩行ロボットを■で示したゴール (障害物) に衝突する寸前まで移動することを目的として学習を行う。

本研究ではロボットの歩行モーションをあらかじめ大まかに設定した状態から実験を開始する。モーションとしては

- 左前脚、右後ろ脚の移動
- 元のポジションに戻す
- 右前脚、左後ろ脚の移動
- 元のポジションに戻す

の4ポーズの流れを1区切りとして考え、それを1モーションと定義する。1モーションごとに PSD 距離

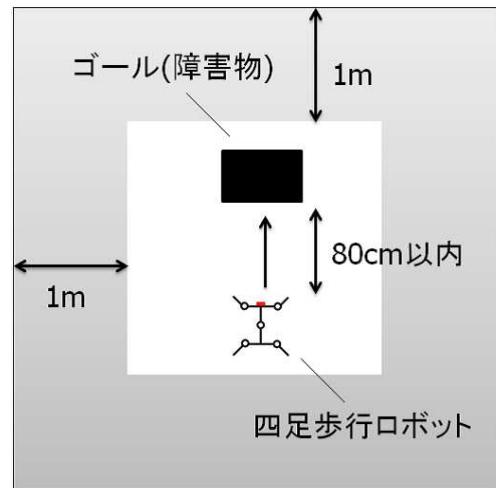


図 3: 実験環境

センサの値を求め、そこから1モーションごとの移動距離を取得し、移動距離が長いほど多くの報酬を獲得できるように設定する。モーション開始前とモーション後の距離の差が負の場合には、ロボットがゴールから遠ざかっていると考えられるため、報酬は0としている。また、モーション後に PSD 距離センサが測定可能範囲内の値を計測できなかった場合には、ロボットがゴールに向かって進むことができず、コースから外れてしまっていると考えられるため、その時点でその試行を終了する。

4 実験

POMDPs 環境のための決定的政策を学習する Profit Sharing を用いて四足歩行ロボットの歩行モーションの獲得実験を行った。材質の異なる2種類の床において実験を行い、それぞれ歩行モーションを獲得できることを確認した。また、一方の環境において学習を行った後、別の環境で再度学習を行った場合に、何も学習していない状態から開始する場合に比べ、短時間で効率的な歩行モーションが獲得できることも確認した。

参考文献

- [1] Y. Takamori and Y. Osana: "Profit sharing that can learn deterministic policy for POMDPs environments," Proceedings of IEEE International Conference on System, Man and Cybernetics, Anchorage, 2011.