MASによるシングルロボットの行動学習
- 学習空間の異なるエージェント群による意思決定 -

情報電子工学系専攻　12054037　高泉昇太郎

背景

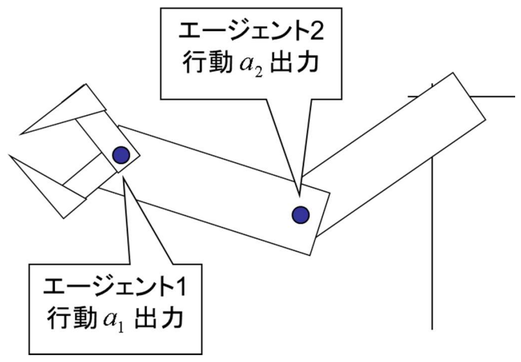
マルチエージェントシステム（MAS）のエージェントの行動学習手法に強化学習が使用される.

ロボットにMASによる強化学習を適用する手法の1つに,
1ロボット内に複数のエージェントを構成する方法が存在する.

1体のロボットがセンサから認識する状態，アクチュエータの動作の組み合わせからエージェントを設定.

↓

各エージェントは自身に割り当てられた動作を出力する.



1ロボット内に複数のエージェントを構成した例

**1ロボット内に複数のエージェントを構成することで
各エージェントが学習する状態と行動の組み合わせ（状態行動対）を分割.
1ロボット1エージェントで学習する場合より効率よく学習することが可能.**

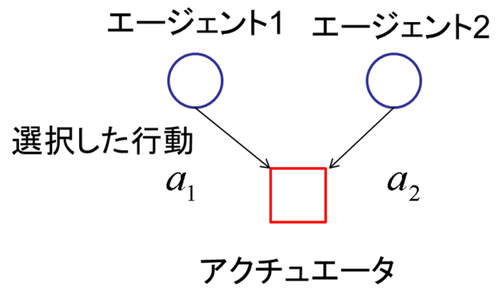
従来研究

- 共通する一つの価値関数を複数のエージェントが非同期並列更新する手法
- MAS概念を用いた運動学計算システムを用いた多関節ロボットの行動生成　　等

複雑なロボットの行動獲得，強化学習の学習時間の短縮にMASが用いられる.

問題点

- センサ，アクチュエータの組み合わせからエージェントを構成.
- 各エージェントは自身に割り当てられたアクチュエータの動作を出力.



アクチュエータの行動のバッティング

↓

1つのアクチュエータの動作を出力するエージェントが2体以上構成される場合が発生する.

- **行動決定の際，各エージェントの選択行動のバッティングが発生.**
- **各アクチュエータの動作を一意的に決定することができない.**
- **従来手法ではバッティングが起こらないように人間があらかじめ各エージェントの動作を調整していた.**

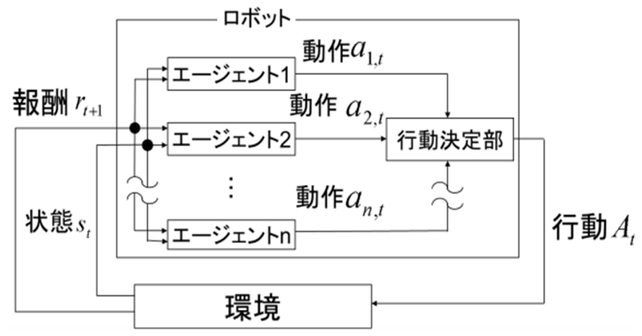
**各アクチュエータに対して各エージェントが選択した行動から
最適とされる行動を一意的に決定したい.**

研究目的

**MASによるシングルロボットの行動学習に対して，各エージェントが選択した動作を調停する方法を提案する.
提案手法によって各アクチュエータの行動を一意に決定.**

提案手法

- **行動選択時，全エージェントが自身に割り振られたアクチュエータの動作を選択.**
- **全エージェントの動作出力後，各エージェントが選択した動作の中から最適行動を選択.**
- **行動後，環境から報酬を獲得．報酬，実際に選択した行動，遷移状態を元に全エージェントが学習.**



提案手法の概要図

ロボットが認識する状態値の集合　$S_{ALL} = \{s_1, s_2, \cdots, s_m\}$

ロボットに搭載されたアクチュエータの動作の集合　$A_{ALL} = \{a_1, a_2, \cdots, a_n\}$

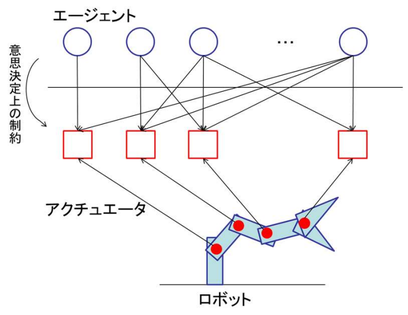
↓

エージェントlが認識する状態　$S_l \subset S_{ALL}$

エージェントlが動作を出力するアクチュエータの集合　$A_l \subset A_{ALL}$

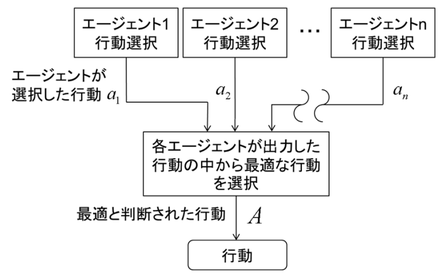
※ただし　A_l は　$A_l \neq \{\phi\}$

- 各エージェントは割り当てられたアクチュエータの動作を出力.



エージェントとアクチュエータの関係

- 全エージェントの行動選択後，各エージェントが選択した行動の中から最適な行動を選択.
- 行動の選択方法は，状態に対する各エージェントが選択した行動の獲得報酬の期待値が高いものを選択する.



ロボットの行動決定方法

➢ 期待値算出方法

- 各エージェントの学習時，同時に各状態行動対に対して獲得報酬の期待値を計算.
- 獲得報酬の平均値が高く，分散が低いほど期待値が高くなる.

$$E = \begin{cases} \mu \times \dfrac{1}{\sqrt{2\pi\sigma}} & (\sigma \neq 0) \\ \mu & (\sigma = 0) \end{cases}$$

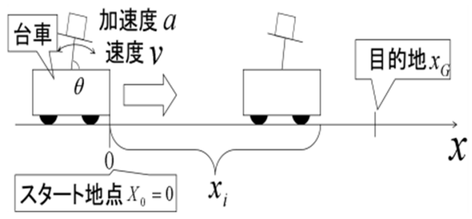
E :期待値
μ :平均値
σ :標準偏差

実験

◆ 実験目的
全エージェントが出力した行動から最適な行動を選択．提案手法がタスクを達成することが可能か検証.

◆ 実験タスク

台車の荷物運搬タスク，台車は荷物を落とさず，目的地に到達することが目的.
台車にはタイヤとテーブルにそれぞれアクチュエータを搭載.



タスクの概要図

◆ 実験概要

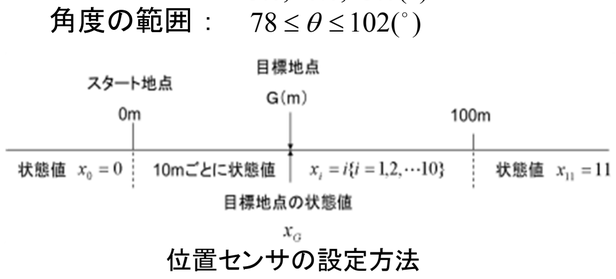
- 従来手法（1ロボット1エージェント）と提案手法（MASによる同時学習）でタスクを実行.
- 提案手法のエージェントの数は，ロボットが持つ全状態行動対に対して設定.
- 各手法の実験結果を比較.

◆ アクチュエータの動作

- タイヤアクチュエータ
  行動　：$-0.5, \pm 0.0, +0.5 (m/s^2)$
  加速度の範囲：$-1.0 \leq a \leq 1.0 (m/s^2)$
  速度の範囲　：$-2.0 \leq x \leq 2.0 (m/s)$
- テーブルアクチュエータ
  行動　：$-3.0, \pm 0.0, +3.0 (^\circ)$
  角度の範囲：$78 \leq \theta \leq 102 (^\circ)$

◆ ロボットの状態設定

ロボットの状態
1. 加速度
2. 角度
3. スタート地点からの位置



位置センサの設定方法

◆ 提案手法の設定エージェント

認識する状態値の組み合わせ：8通り
動作を出力するアクチュエータの組み合わせ：3通り

↓

合計　24エージェント

◆ 報酬設定

$$r = w_1(\theta - R)^2 + w_2(x_G - x)^2$$

台車の現在位置：$x(m)$　テーブルの角度：$\theta(rad)$
台車の目標位置：$x_G(m)$ 合力の角度：$R(rad)$

◆ 実験結果

実験パラメータ

試行回数		5000(回)
1試行の行動回数		1000(回)
行動の間隔		1秒毎
学習を行うタイミング		1行動毎
目的地の位置　状態値x_G		5.0
行動学習手法		Q-learning
行動選択手法		ε-greedy法
初期値	加速度 a	$0.0 (m/s^2)$
	角度 θ	$90 (^\circ)$
	速度 v	$0.0 (m/s)$
	位置センサの値 x	0.0
	各Q値 $Q(s_t, a_t)$	0.0
報酬式の係数		$w_1 = -10.0, w_2 = -0.1$
ε		0.05
ステップ・サイズ・パラメータ α		0.1
割引値 γ		0.9

従来手法

提案手法

各試行の総獲得報酬の推移

考察

- 提案手法は総獲得報酬の値が収束せず，値の大きさも低いものとなった.
- 提案手法ではタスクを達成することのできる行動を獲得することができなかった.

→ **現時点の提案手法では
目的を達成することができない.**

- **期待値**

認識する状態の種類が少ないエージェント→大　認識する状態の種類が多いエージェント　→小
認識する状態数が少ないエージェントほど選択される期待値となっている.

今後の課題

選択したいエージェントの行動が選択されない.

↓

期待値の計算式に問題がある.

**獲得報酬の平均値が高く，かつ分散の小さいエージェントの行動が選択されるようにする.
獲得報酬の期待値の計算方法を検討しなおす.**