強化学習における情報量を用いた探索と利用の自律的制御

認知ロボティクス研究室　11054032　澁谷 和

研究背景

探索と利用

◆ 利用(exploitation)
- 過去に試みた行動の中で，多くの報酬を得るような行動を取ること

トレードオフ

◆ 探索(exploration)
- 未知状態を経験するために行動すること
- 現在，所有している知識が最適とは限らない
- 多くの報酬を得るためには未知状態を探索することが不可欠である

問題点

- 探索と利用のバランスが崩れると…
  → 学習効率の低下
- 学習環境によって適したパラメータは変化する
  → 環境に応じて，適したパラメータの値を人の手で設定するのは手間がかかる

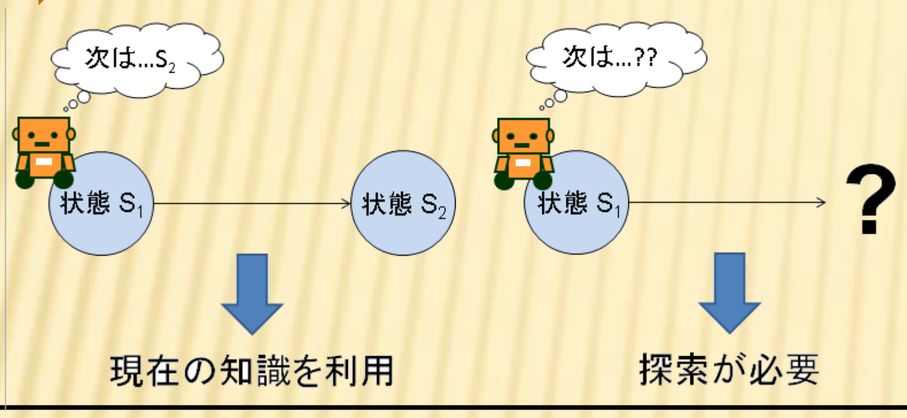
研究目的

ロボットが探索と利用のバランスを自律的に制御するシステムの構築

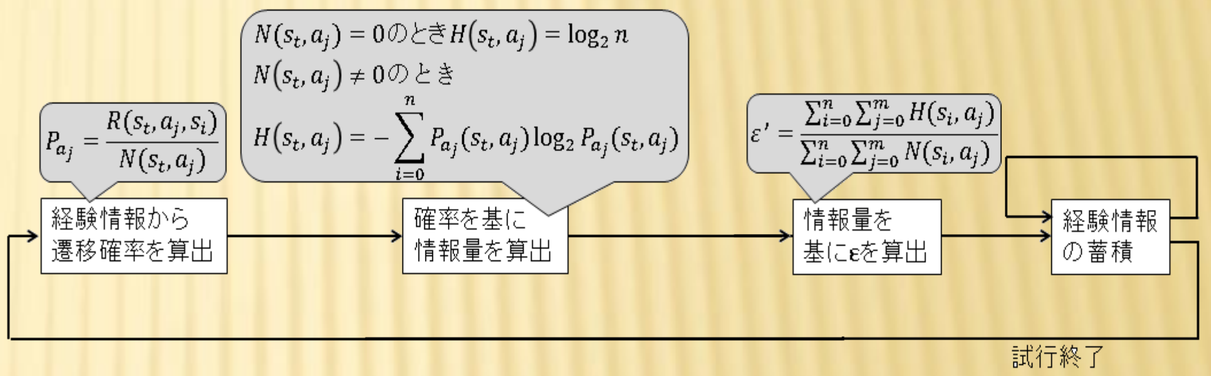
アプローチ

強化学習の目的：獲得報酬量を最大に
→ 獲得報酬量の予測が必要
- 予測が可能→利用
- 予測が不可能→探索

→ 予測可能，不可能の判断を情報量で行う



提案手法



$N(s_t,a_j)$ ：s_tにおいてa_jを取った回数
$R(s_t,a_j,s_i)$ ：s_tにおいてa_jを取り，s_iに遷移した回数
$P_{a_j}(s_t,s_i)$ ：s_tにおいてa_jを取り，s_iに遷移する確率
$H(s_t,a_j)$ ：(s_t,a_j)における平均情報量

迷路問題への適応

実験目的

提案手法によってパラメータ調整ができているかを検証する

実験概要

- 迷路問題に適応
- ゴールまでの最短経路を探索
- 以下の2種類のエージェントで比較
  - 全試行通して，ε=0.05
  - 全試行通して，ε変動(提案手法)

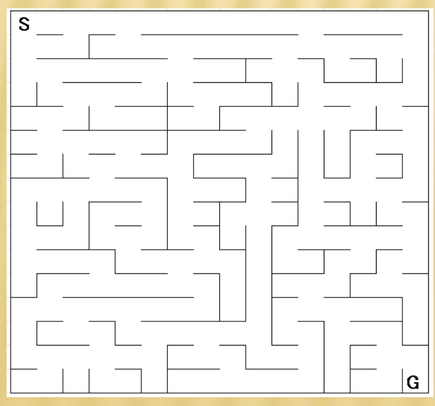
実験設定

■ エージェントの設定
- 学習手法：Q学習

$$Q(s_t,a_t)\leftarrow Q(s_t,a_t)+\alpha\left[r_{t+1}+\gamma\max_a Q(s_t,a)-Q(s_t,a_t)\right]$$

- 行動選択法：ε-greedy法

■ タスクの設定

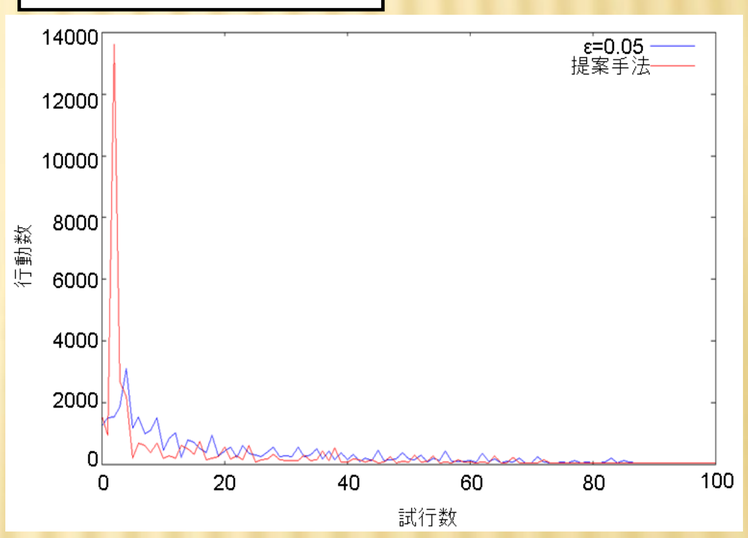


実験で使用した迷路(16×16)

■ パラメータの設定

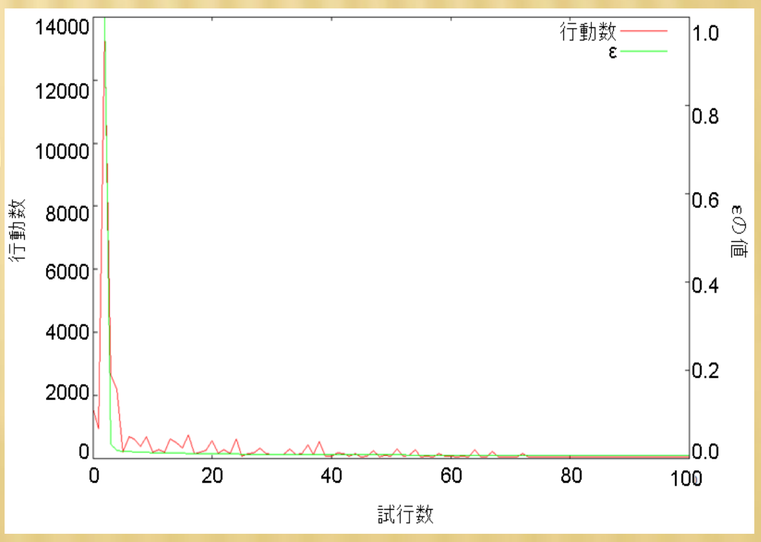
迷路の大きさ	16×16
試行数	100
報酬(ゴールのみ)	100
Q値の初期値	0.001
学習率α	0.5
割引率γ	0.5
探査率ε	0.05

実験結果・考察



各試行における行動数の推移

- 学習初期
  →経験していない状態・行動が多い
  →εが高い値
  →探索行動
- 学習が進むにつれて
  →ほとんどの状態・行動を経験済み
  →εは0に向かって減少
  →利用行動



提案手法における試行毎の行動数とεの値の推移

今後の課題

- εの正規化
- softmax法のτの制御
- 複雑なタスクでの実験
- 他のεや従来手法との比較
- 動的環境への適応