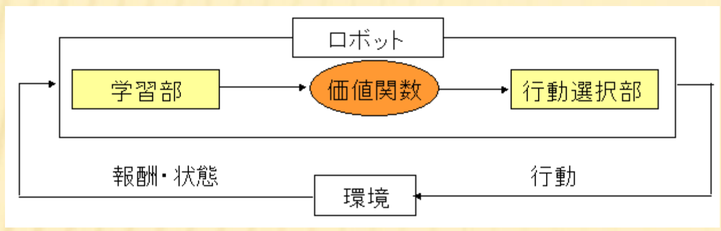
強化学習における情報量を用いた explorationとexploitationの自律的制御

認知ロボティクス研究室　11054032　澁谷 和

背景：強化学習とは



- 試行錯誤を通じて環境に適応する学習法
- 報酬を最大化するためにはexplorationとexploitationのバランスが重要

問題点

- 同じ状態に対して，explorationすれば良いのか，exploitationすれば良いのか一意に定まらない
- 強化学習においてexploration-exploitationを決定するパラメータは一意に決定される

⇒ 既知状態に対して，explorationする

⇒ 学習効率の低下

- ロボットが学習する環境によって，適したパラメータの値は変化する

⇒ 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

explorationとexploitation

◆ exploitation(利用)
- 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

◆ exploration(探索)
- 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

研究目的

ロボットがexplorationとexploitation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

✓ 問題点の把握

✓ 提案システムの構築・実装

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→exploitation(利用)
予測が不可能→exploration(探索)

⇒ 予測可能，不可能の判断を情報量で行う

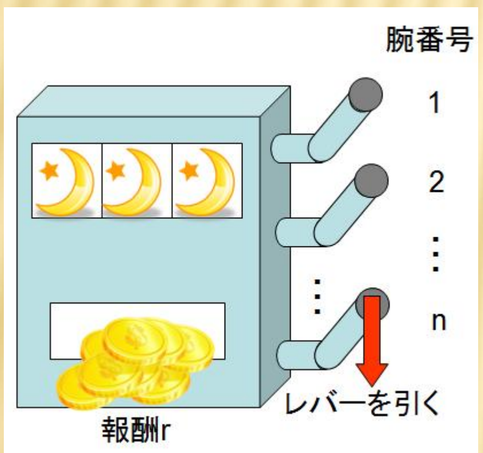
✓ 静的・動的環境下への適応

実験：問題点の検証

実験目的

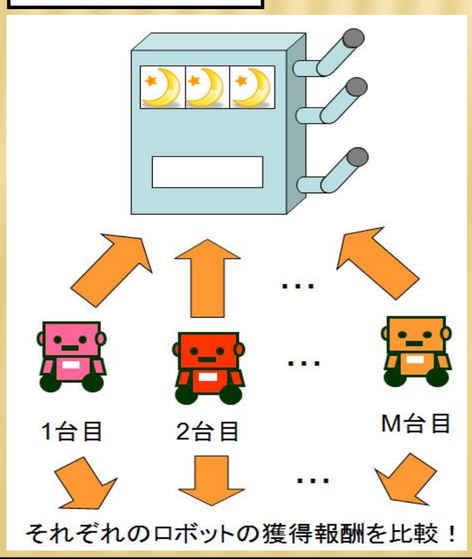
exploitation-explorationのバランスを一意に決定することにおける問題点を示す．

実験タスク：N本腕バンディット問題



- 腕を引くたびに，「当たり」か「はずれ」が決定
- 「当たり」であれば，その腕に設定された報酬を獲得できる
- 獲得報酬を最大化することが目的

実験概要



- M台のロボットにバンディット問題を適用
- exploitation-explorationのバランスを決定するパラメータを異なった値に設定
- 各ロボットの獲得報酬を比較

実験設定

■ エージェントの設定

➢ 学習手法：標本平均化手法

$$Q_t(a) = \frac{r_1 + r_2 + \cdots + r_{k_a}}{k_a}$$

➢ 行動選択法：ε-greedy法

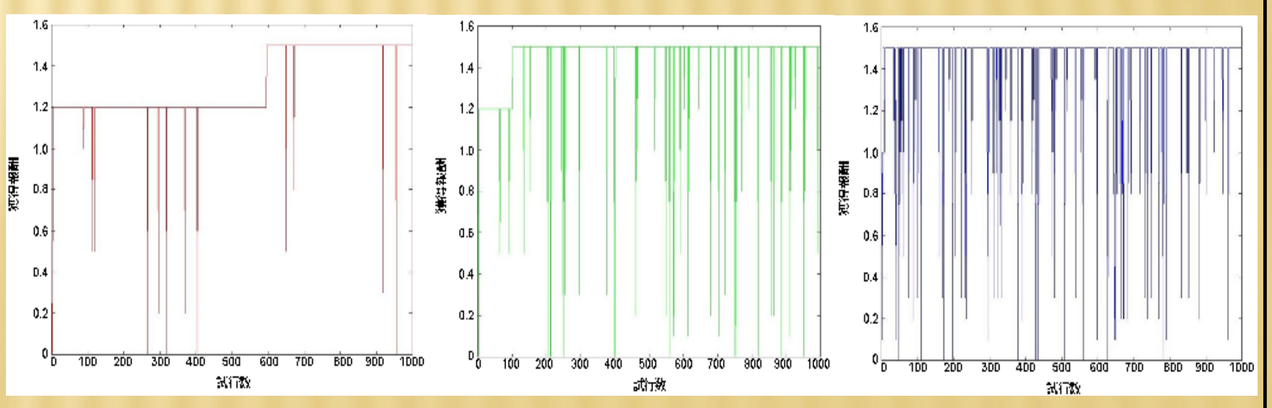
■ パラメータ

試行回数	1000試行
ε	0.01，0.05，0.10
ロボットの台数	3台
Q値の初期値	1.00

■ タスクの設定

腕の番号	1	2	3	4	5	6	7	8	9	10
報酬	0.0	0.1	0.0	0.2	0.3	1.2	1.0	0.8	0.5	1.5

実験結果・考察



ε =0.01　　ε =0.05　　ε =0.10

総獲得報酬:1305.2　　総獲得報酬:1415.4　　総獲得報酬:1407.3

⇒ 利用重視：最適値を発見しにくい．発見後は安定

探索重視：最適値を発見しやすい．発見後は安定せず