複数目的下での行動決定手法の提案

認知ロボティクス研究室　情報電子工学系専攻 12054055　三浦丈典

研究背景

実世界でロボットを制御するためには，通常複数の目的を満たすことが求められる．

例） 移動ロボット　1．エネルギー残量を一定値以上に保つ
2．障害物を避ける
3．目的地に到達する

ロボットの行動学習手法の1つである強化学習によって，
複数目的を学習する研究が行われている

従来研究

- 多目的最適化によってパレート最適解などの最適行動候補を発見する研究
- 階層型強化学習によって複数目的を学習する研究
- 複数の報酬に重みを付加し，一つの報酬として学習を行う研究　　等

問題点

- 最適行動候補を見つける研究では，候補を絞り込むだけで最終的な行動決定は人間が行う必要がある．
- 従来の複数目的の学習手法では，環境の変化などによって各目的の重要性が変わった場合には再学習を行う必要がある．

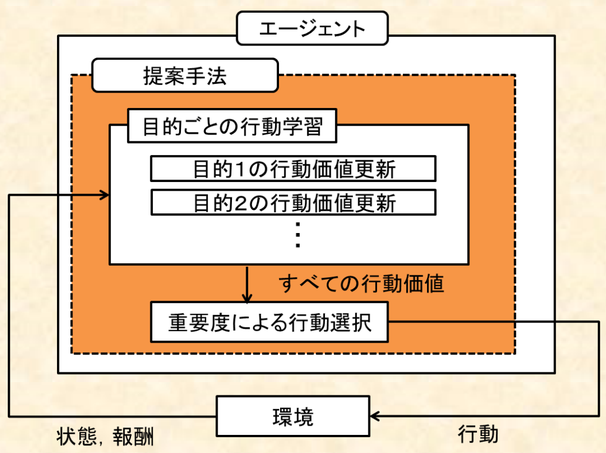
研究目的

目的ごとに独立した学習を行い，状況に応じた行動を選択することで
目的間の重要性の変化に対応した行動決定手法を提案する

提案手法

- システム概念図

- 目的ごとに強化学習によって行動価値を更新
- 複数の最良行動候補から行動を決定

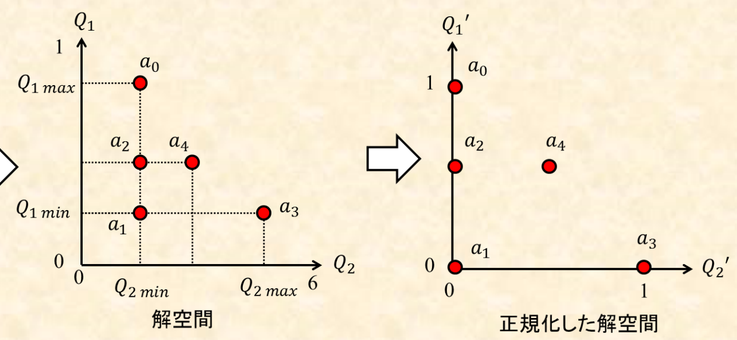


複数の最良行動候補から行動決定を行うためのアルゴリズム

1.　各行動を取った場合に得られる報酬の期待値（Q値）から解空間を生成する．
生成した解空間を(1)式により0～1の空間へ正規化する．

行動に対する行動価値（Q値）

行動	Q_1の値	Q_2の値
a_0	0.9	3
a_1	0.3	3
a_2	0.5	3
a_3	0.3	7
a_4	0.5	5



ある行動a_iでの各Q値を$Q_n(a_i)$とする

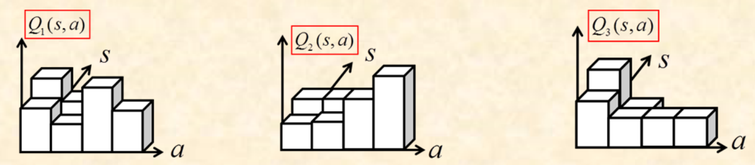
$$Q_n'(a_i) = \frac{Q_n(a_i) - Q_{n\,min}}{Q_{n\,max} - Q_{n\,min}} \quad (1) \qquad (n\text{は目的数})$$

アプローチ

- 再学習が必要となるのは，　一つの学習空間で学習を行なっているため．

- 目的ごとに別々の学習空間で学習を行い，目的ごとに最良となる行動を決定する．

目的1に対する学習　目的2に対する学習　目的3に対する学習



目的1のための行動　目的2のための行動　目的3のための行動

各学習ごとに最良となる行動候補が存在する

- 複数の最良行動候補の中から状況に合わせた行動を選択する必要がある．
- 各目的の重要度を定義し，重要度を基に行動決定を行う．

重要度は，現状態からその目的を達成した時に得られる報酬の期待値を基に決定する

2.　各目的の重要度を算出する．

各目的に与えられる報酬に関連性が無ければ重要度は算出できない．（検証実験を踏まえて）
エネルギーなどの最大・最小値のある報酬を指標に各目的の報酬設定を行うことを前提とする．

現在の状態からその目的を行うことで獲得できる総報酬量を，目的を達成した場合に得られる期待報酬と考える．現状態からある目的iを達成した時の期待報酬R_iは(2)式で求める．

$$R_i = \sum_{t=t_s}^{m} r_t \qquad (2)$$

tはステップ
mは目的達成までに要したステップ数
t_sは現在のステップ数

報酬の期待値は各報酬によって正負が異なる可能性があるので，すべての期待値を正の値に変換，その比をとることで0～1の範囲の重要度を決定する．

目的1の期待報酬　10
目的2の期待報酬　100
⇨
目的1の重要度　0.1
目的2の重要度　1

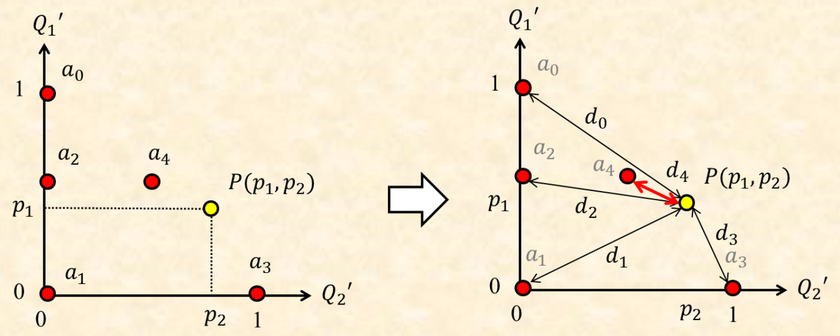
重要度　p_n　$(0 \le p_n \le 1)$　（nは目的数）

3.　正規化した空間に各目的の重要度をプロットし，各行動とのユークリッド距離を算出する．

解空間上での各重要度の交点$P(p_1, p_2)$　　(3)式により距離を算出

p_1：目的1に対する重要度
p_2：目的2に対する重要度

$$d_i = \sqrt{\sum_{k=1}^{n}\{p_k - Q_k'(a_i)\}^2} \qquad (3)$$



4.　算出した距離が最も小さい行動a_iをロボットの取る行動とする．

上記の例ではa_4が選択される

※但し，局所解への迷い込みを防ぐためεの確率でランダム行動を取る

検証実験

- ロボットに2つの目的（報酬関数）を与えた場合のシミュレーション実験
  重要度については事前に設計者が設定し，重要度による行動決定が可能かどうか検証する

目的1：エネルギーの獲得

- 報酬　：　$R_1 = \Delta E$（ΔEは一回の行動に対するエネルギー変位）
  ※エネルギー残量が0%となった場合のみR_1 = -100
- ロボットの状態による重要度p_1の変化　：　$p_1 = \frac{1}{1+e^{(0.2E-10)}}$　（Eはエネルギー残量）

目的2：ゴールへの到達

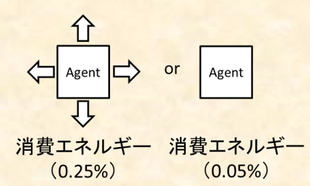
- 報酬　：　$R_2 = 50$　（ゴール到達時のみ）
- ロボットの状態による重要度p_2の変化　：　$p_2 = 0.5$ (固定)

ロボットが認識する状態

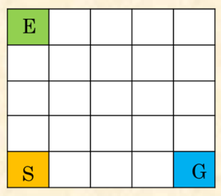
- マスの位置S
- エネルギー残量E

ゴールへ到達すると
スタート位置へ戻る

エージェントの行動

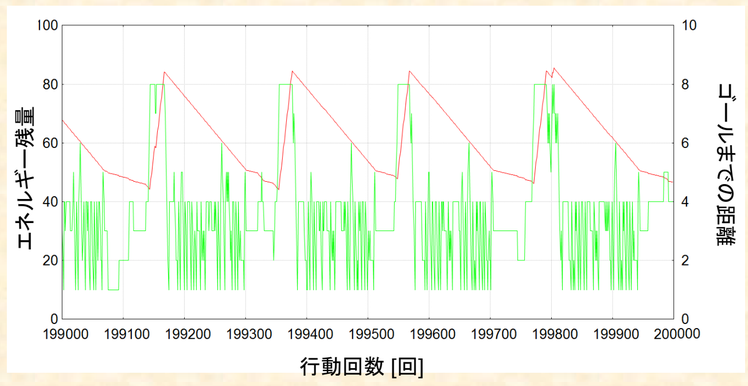


実験環境



S：スタート位置
G：タスクのゴール位置
E：エネルギー充電ポイント

- 実験結果



行動回数に対するエネルギー残量の変位とゴールまでの距離
（行動回数199000～200000）

実験パラメータ

学習回数	200000 [回]
行動学習手法	Q学習
初期行動価値	0.0
ステップサイズパラメータα	0.1
割引率γ	0.9
ランダム行動確率ε	0.05
充電ポイントでのエネルギー変位	+2.0 [%]
前後左右移動時のエネルギー変位	-0.25 [%]
停止時のエネルギー変位	-0.05 [%]

エネルギー残量に応じた行動選択を実現できている．

今後の予定

- 重要度を自律的に獲得する部分を実装し検証実験を行う
- 目的達成までにかかる行動数を考慮することで，より状況に合わせた行動決定を行う
- 他の手法との比較実験を行う