状態価値と行動価値を併用したモンテカルロ木探索による強化学習

情報・ネットワーク工学専攻　兼岩研究室　高橋大樹

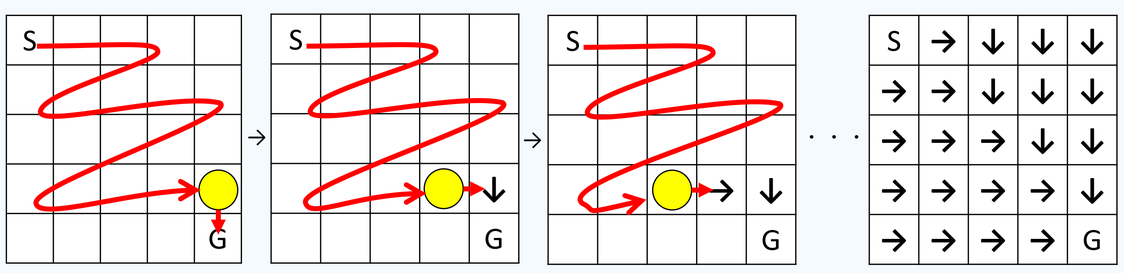
強化学習

目的：

環境を探索し、タスクを達成できる「方策」を学習すること.

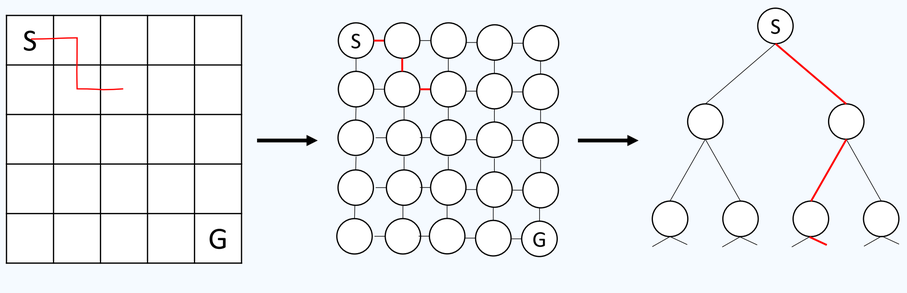
- 行動の結果に対して報酬が与えられる
- 報酬を基に方策を更新する

学習初期はあらゆる行動を試行錯誤し，知識・経験を得る.
学習終盤では知識・経験を活用し，報酬の期待値が高い行動を選択する.
最終的に効率的な行動選択指針（方策）を学習できる.



モンテカルロ木探索(MCTS)

- 各状態をノード・各行動をエッジとみなした木構造を用いる
- 木に知識・経験を記録し，探索する
- 囲碁AI「AlphaGo」等に用いられる



利点：

実際の勝率を基に探索領域を選択するため，評価関数を設計できないような専門的・複雑な問題にも適応可能.

勝率推定に信頼上限を用いて統計的に探索する（UCT）.

提案手法

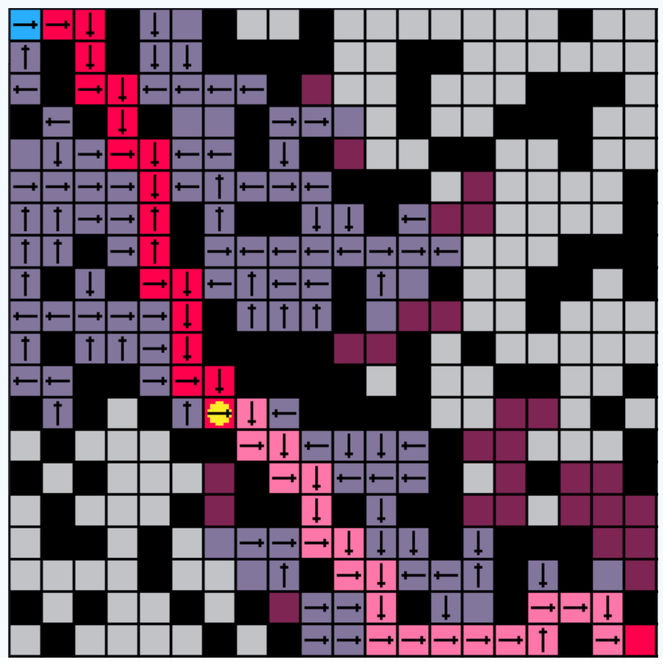
従来手法の問題点：

- スタート付近を重点的に探索
- 同じマスを何度も通過

改善点：

- スタートから遠いほど探索優先
- 同じマスの複数回通過を禁止
- 最短経路の推定で成績の良いQ学習の手法を導入

探索中の様子→



実験

サイズ20×20，穴の割合15％の迷路において，従来手法よりも早く，確実に最適解を導出することができた.

手法	初ゴール /step	学習時間 /step	最終推定経路長
Q学習	21450.27	51659.66	**38.00**
MCTS	89625.33	100000.00	-
MCTS/Q (提案手法)	18702.59	**45593.12**	**38.00**

今後の課題：

より複雑な迷路への対応
迷路以外の実問題への応用