2020・10・15　ゼミ資料

Pythonで学ぶ強化学習

3章 強化学習の解法(2):
経験から計画を立てる(後編)

1116 17 9036

山口真哉

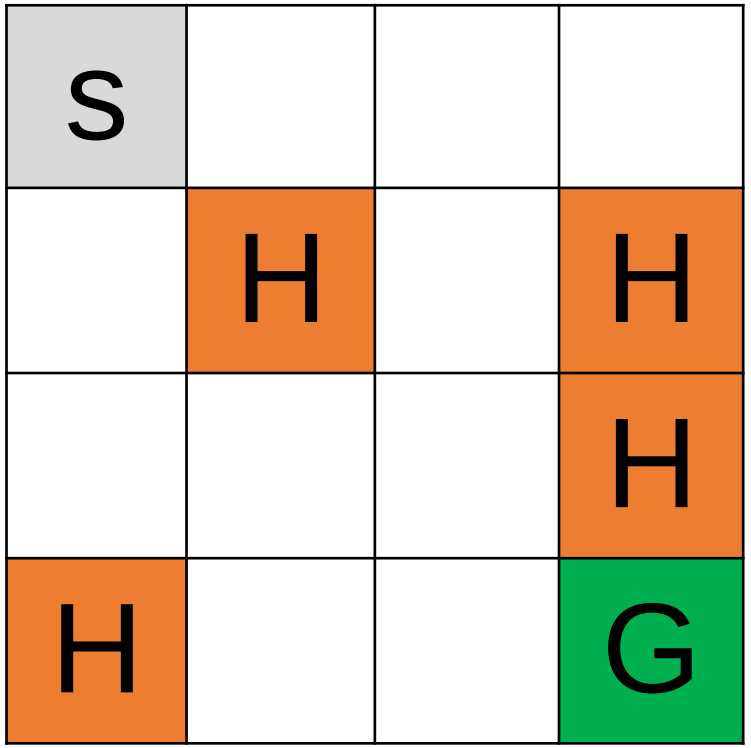
chap 3

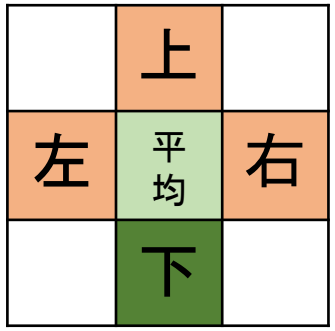
行動した「経験」を活用するにあたって, 検討すべき点が3つある.

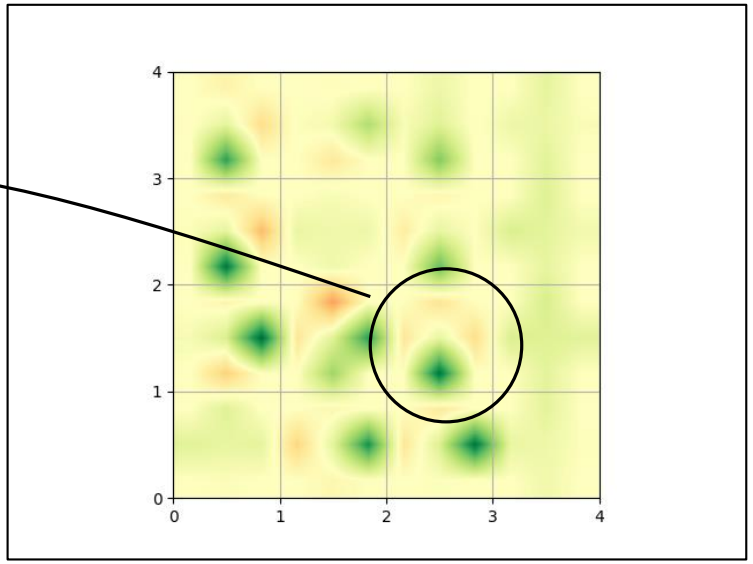
1. 経験の蓄積と活用のバランス (先週 ε – Greedy を使ってコイントスをした)
2. 計画の修正を実績から行うか予測で行うか (理論までやった)
3. 経験を価値評価, 戦略どちらの更新に利用するか

chap 3

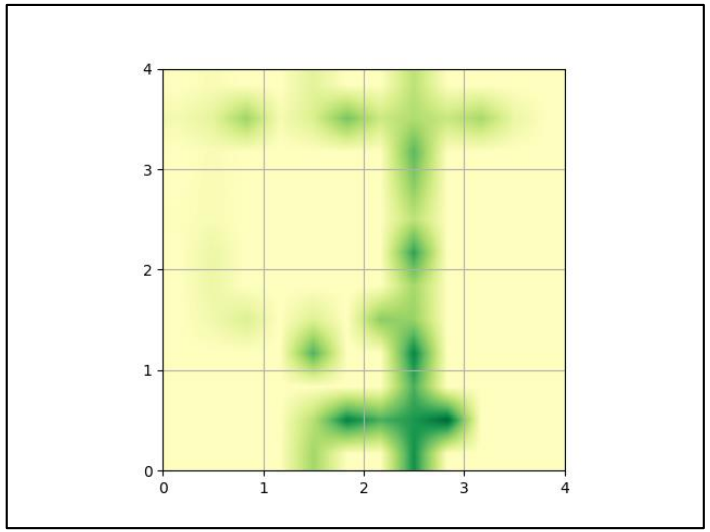
- 今回は実験メインで進めていく.
- OpenALGymのFrozenLakeを使う.
- Gのマスが報酬が1, それ以外は報酬を0を得る.
- H or G のマスに到達するとそのエピソードは終了する.
- 特に断らない限り $\gamma = 0.9$ とする.

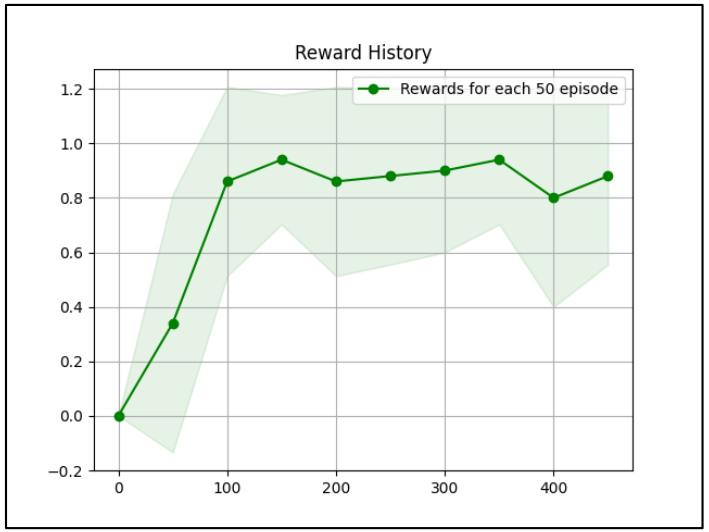






モンテカルロ法

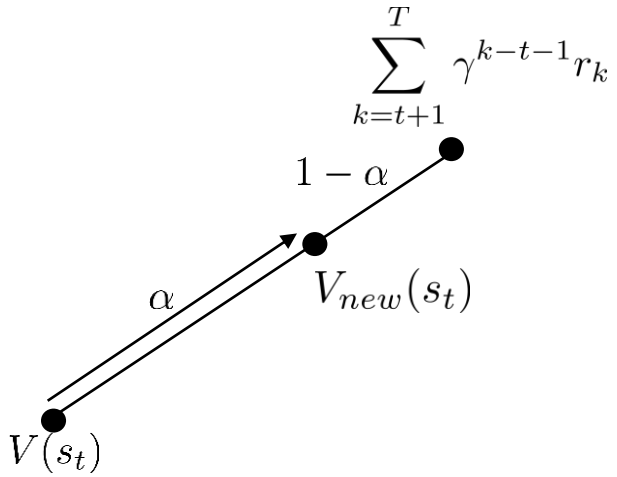




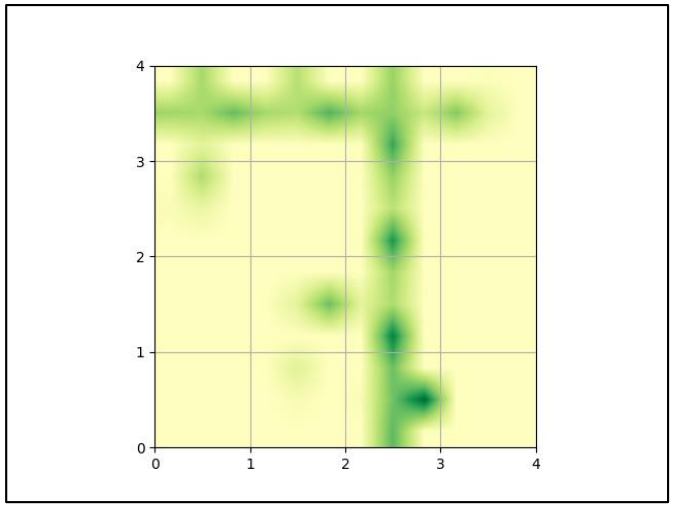
更新アルゴリズム（モンテカルロ法)

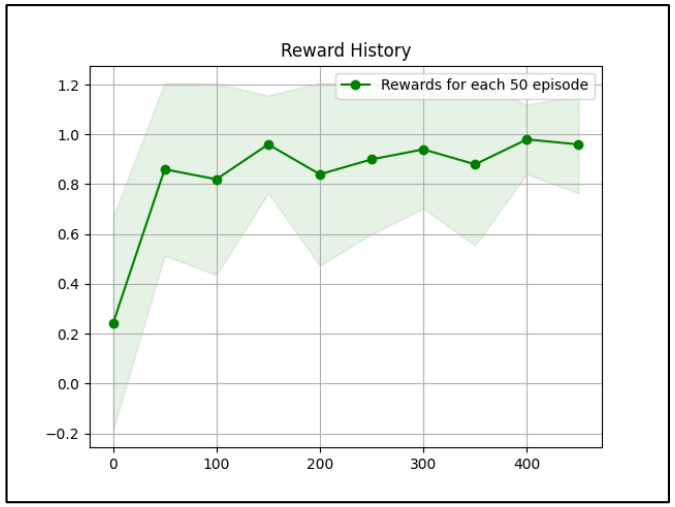
$$V(s_t) \leftarrow V(s_t) + \alpha \left(\left(\sum_{k=t+1}^{T} \gamma^{k-t-1} r_k \right) - V(s_t) \right)$$

α は学習率と呼ばれる.



TD法 (Q – learning)





更新アルゴリズム(Q - learning)

$$Q(s_t, a_t) \leftarrow Q(s_t, a_t) + \alpha(r_{t+1} + \gamma \max_{a \in A}(Q(s_{t+1}, a))) - Q(s_t, a_t))$$

TD法とモンテカルロ法の比較

- この環境だと状態数が少ないので時間差は感じられなかった.
- ややわかりづらいがTD法は1つ後しか見ないので学習が安定せずグラフがガタガタしている.
- モンテカルロ法はゴール付近で緑が濃いことからゴール前をうろうろしていることがわかる.
- それに比べてTD法は比較的ゴールに直行している.
- 両方とも穴マスが緑になっていないため, 穴を通っていないことがわかる.

SARSA

- Q-learningの更新対象は「価値評価」で行動選択の基準はOff-Policyであった.
(価値が最大になるような行動を取った.)

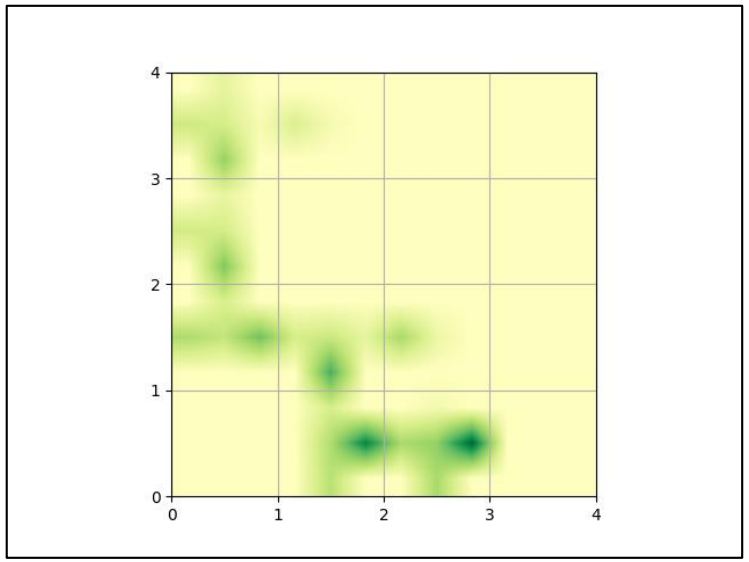
- 更新対象が「戦略」で基準が「On-Policy」であるSARSAを紹介する.

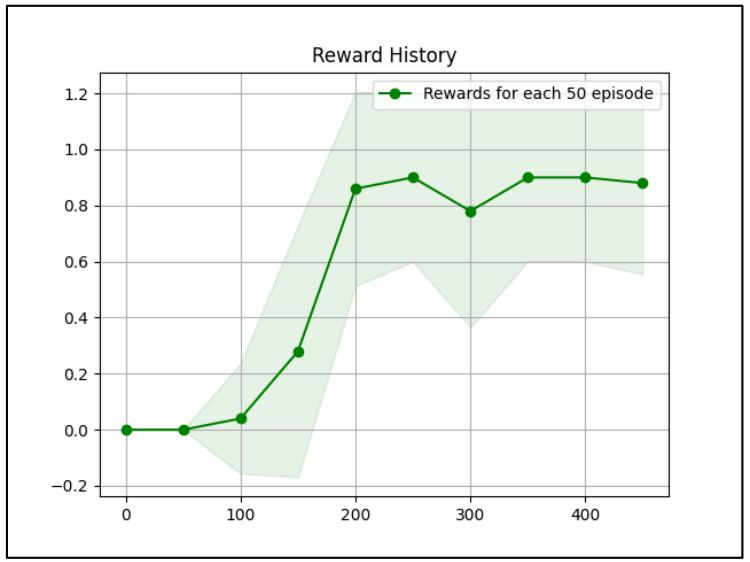
更新アルゴリズム（SARSA)

$$Q(s_t, a_t) \leftarrow Q(s_t, a_t) + \alpha(r_{t+1} + \gamma Q(s_{t+1}, a_{t+1})) - Q(s_t, a_t))$$

- Q-learningは状態行動価値が最大になるように動いたため少し異なる.

SARSA

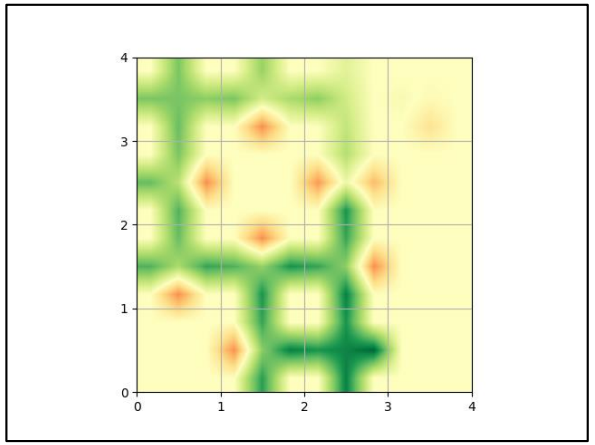




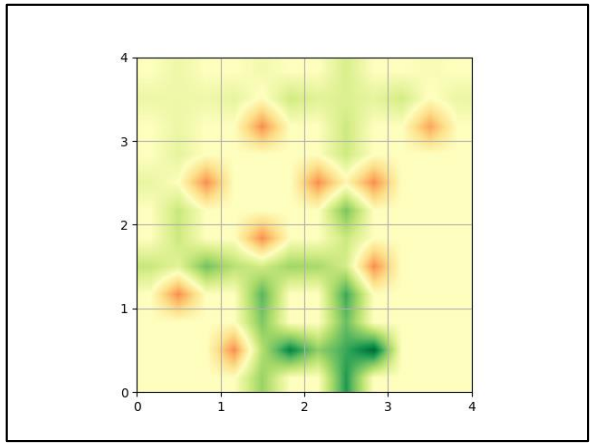
- TD法ベースなので学習速度は遅くて分散は大きいが,
  ちゃんと前進していることがわかる.

Q-learning と SARSA の比較

- わかりやすくするために, $\varepsilon = 0.33$ とし, 落ちた時のペナルティ=0.5を与えた.
- Q-learningでは最善の行動が前提となるため状態行動価値が大きくなる.
- SARSAは戦略による行動, つまり穴に落ちてしまうような行動
  (ゴールと逆方向に行く)を抑制されるようになる.
- イメージはQ-learningは楽観的, SARSAは現実的なエージェントである.



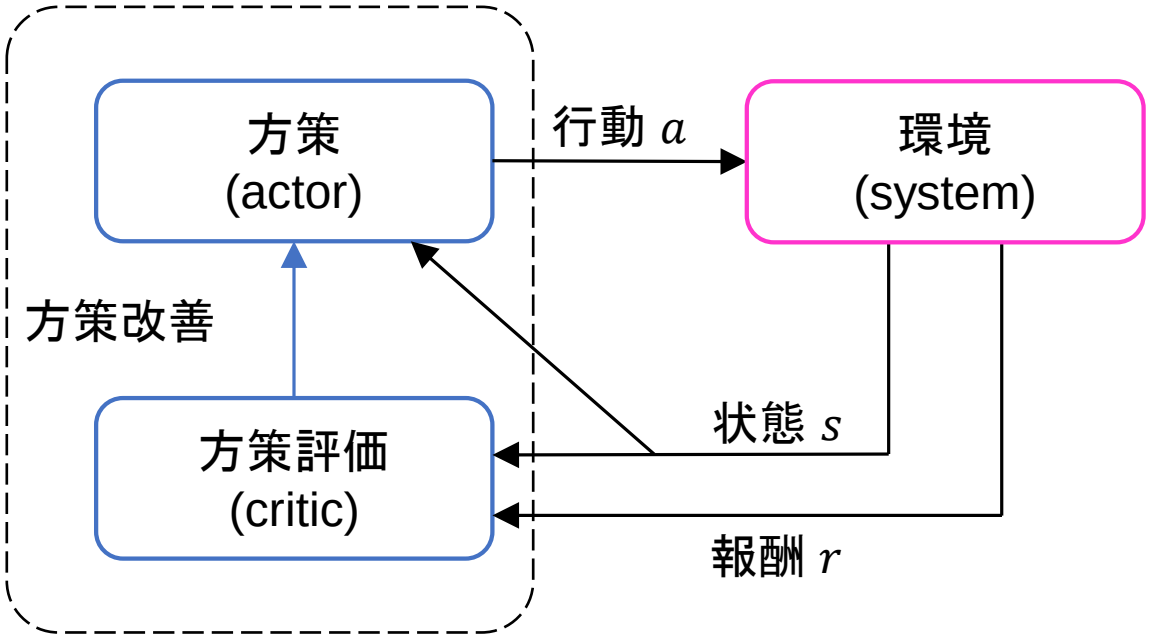
Q-learning



SARSA

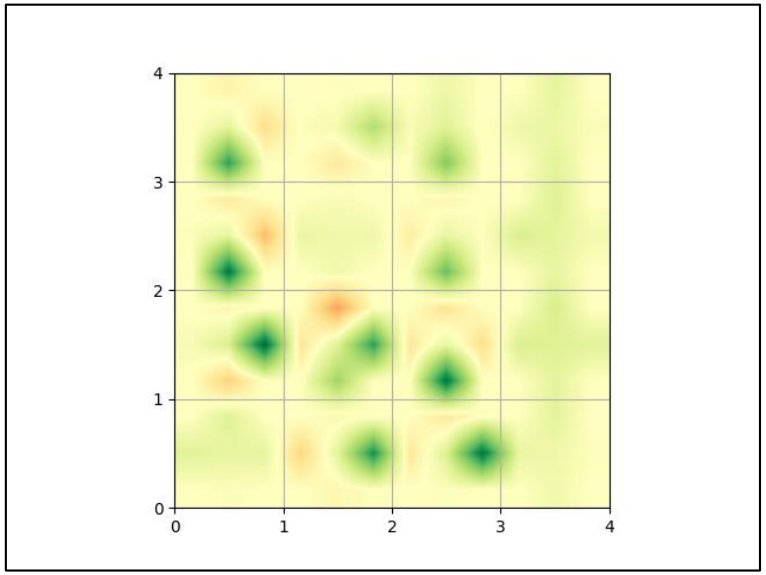
Actor Critic

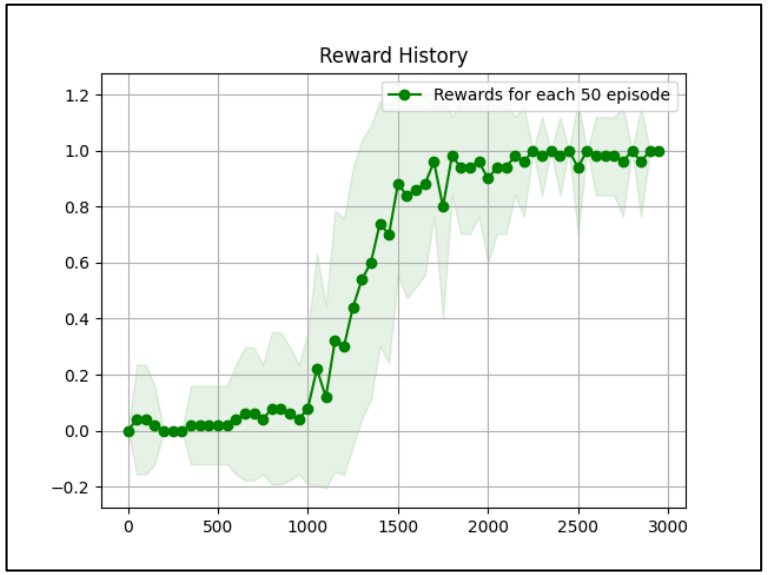
- ValueベースとPolicyベースを組み合わせた手法.
- 詳細は省くがcriticによってactorとcriticを更新する. (Agentが2人)



Actor Critic

- 学習に時間はかかるが, めっちゃいい性能が出ている.
- 特にゴール方向がちゃんと評価されているのが印象的.





epiosode数が他より多くしているのに注意

まとめ

- OpenAI Gymを使って強化学習の手法を見てきた.
- 実装の詳細は紹介しなかったがQテーブルを使って実装した.
  (Q[s] [a] := 状態 s で行動 a した時の状態行動価値)
- しかしQテーブルというデータの持ち方は連続値だと破綻する.
  - メモリと実行時間が有限ではないため
- 次章ではパラメータを持った関数でQ値の算出を行う.
- Q値の近似方法が鍵になる.

何か質問はありますか？