○目的:

問題環境(Problem1)において高い報酬を獲得することができるQ学習のプログラムを作成し，そのプログラムの性能を評価する.

○方法:

適切なメカニズムとパラメータを設定したQ学習のプログラムを試行錯誤により作成した後，（1）最終エピソードにおける行動実行回数の平均および獲得報酬の平均と（2）報酬獲得率の推移を調整する. なお，平均は乱数を変えて100回志向した結果を算出する.

○結果:

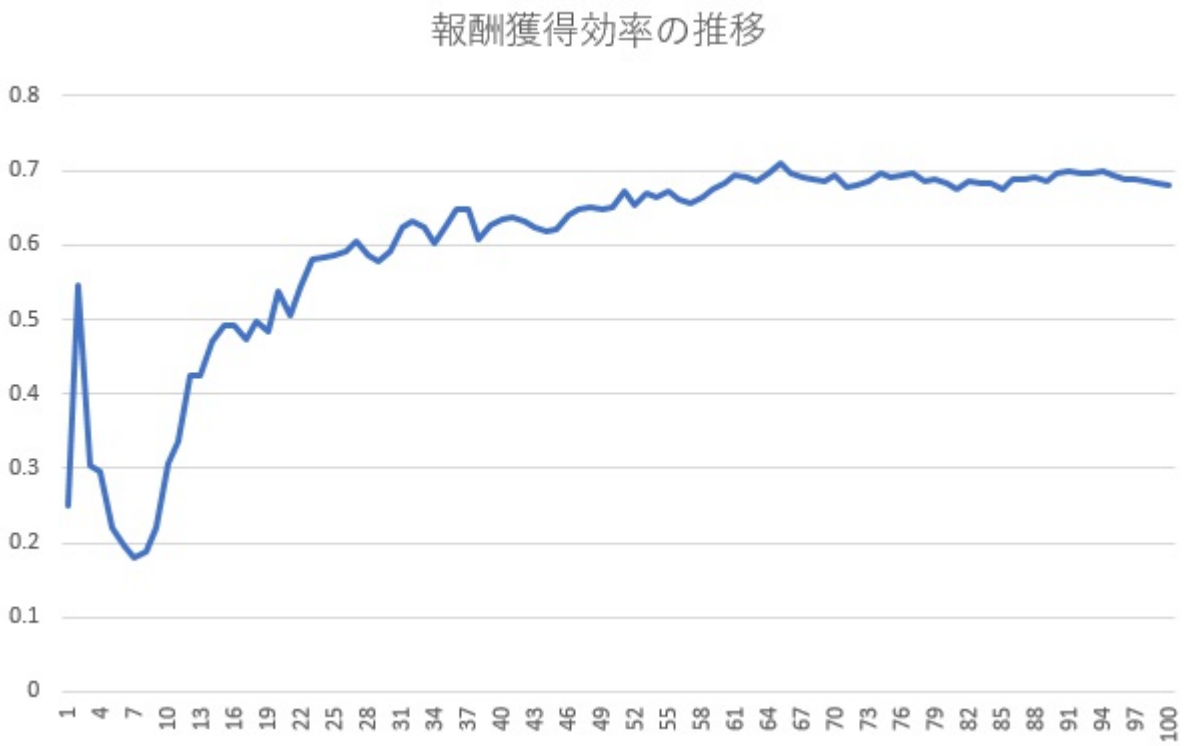
(1) 作成したQ学習のプログラムのメカニズムとパラメータ
本実験で作成したメカニズムは以下の手順である.

1. 状態行動価値Q(s,a)を初期化する.
2. 状態を初期化する. (S <- S0)
3. 条件を満たすまで繰り返し

    1. 行動aを選択
    2. 報酬aを実行
    3. 報酬rと事情体s'を観測
    4. Q(s,a)を更新
    5. s <- s'

(2) 各試行の最終エピソードにおける行動実行回数の平均および獲得報酬の平均
本実験における行動実行回数の平均は4回，獲得報酬の平均は6回であった.

(3) 報酬獲得効率のエピソード推移（1エピソード目からあるエピソードまでの累計獲得報酬を累計行動実行回数で割った値）縦軸は報酬獲得効率，横軸はエピソード数
報酬獲得効率のエピソード推移を以下に示す.



○考察：

- 班員の結果と大きくことなり，報酬獲得効率が1を超えることがなかった．プログラム上の問題だと思ったが原因は不明である．
- Q学習において，学習率などを自分で定義する必要がある．学習率などを自分で定義するため，欲しい結果に近づけるための工作が可能ではないのか．
- 学習していく様子などを毎回図で表示していくとわかりやすいのではないか．

○まとめ

　プログラムがとても怪しいが自分ではどうも改善する場所が全くわからない．また，発展課題では平均獲得報酬や報酬獲得効率を上昇させることを目的としたい．