練1.1

- 相手の方策も変化するので，同じ方策が学習されるとは限らないし，最適方策が得られるとも限らない。
- $\alpha \to 0$ としない場合，ループに落ち入るかもしれない？
- もし両者が最適方策を学習した場合，結果は必ず引き分けになるので，それ以降方策は変化しない。

練1.2

- 対称性を用いて同じ盤面とみなせる複数の盤面を1つの状態に対応させる。状態数が減らせるので，最適方策の推定がより上手くできる（対称的な盤面には必ず同じ価値が割り当てられるから）。
- 相手が対称性を考慮しない方策の場合，対称的盤面であっても異なる状態であるとみなすべき。相手は環境の一部なので，環境が異なることになるから。

練1.3

- greedyなAgentは、探索を行わないので長期的視点の価値関数が得られず、non-greedyより弱い可能性が高い。

練1.4

- 探索的な打ち手を継続する場合、得られる状態価値関数は、「探索を含む方策」におけるもの。一方、探索的な手から学習しなければ、最適方策が得られる。

練1.5

- 三目並べは状態数が少ないので完全解析できる。
- 引き分けの盤面の状態価値を、負けの盤面の状態価値より大きくする。(勝率だと両方0になってしまう)。
- 複数iteration行ってから状態価値を更新すると学習が安定する？