強化学習の基礎的な知識

- 強化学習の基礎的な知識についてまとめる。
- 環境：エージェントが行動する場所
- エージェント：行動する主体
- 状態 s∈S：環境があるタイムステップでどうなっているか
- 行動 a∈A：エージェントがある状態に対して取る行動
- 報酬 r∈R：エージェントがある状態である行動を取ったときに得る報酬
- 価値
    - 状態価値関数 V(s)：エージェントがある状態sにいるときの価値。行動とは関係ない。
    - 行動価値関数 Q(s, a)：エージェントがある状態sである行動aを取ったときに得られる報酬の期待値(報酬とは違い、今回の行動で報酬はなくても次回の行動で報酬が得られる可能性が高まったら価値はあるといえる)
- 方策 π：エージ 方策勾配定理：方策の勾配は報酬の期待値の勾配と一致ェントがある状態でどの行動aを取るかを決める関数
- 強化学習の目的は、価値を最大化するような方策を見つけることである。
- 強化学習の手法には、価値ベースと方策ベースがある。
    - 価値ベース：行動価値Qを最大化するような行動aを決める
    - 方策ベース：状態価値Vを最大化するような方策を決める
    - モデルベース：環境が完全にわかっているという過程を置く。価値最大化を計画する方法。← よくわかんない
- 方策オンと方策オフ
    - 方策オン：状態と行動を現在の方策からサンプリングする
    - 方策オフ：これまでの方策からサンプリングする 正直よくわかんないけど。。。経験再生（過去の経験をストックしておく手法）の利用ができるか否かの違いがあるらしい。オフのことだと思う。
- 最適化問題の解き方
    - DP:パラメータが既知の場合全探索で完全な解を求める
    - モンテカルロ：サンプリングで解を求める。ロールアウトではシュミレーションで試した結果で方策を決める。
    - TD：temporal difference。現在の推定価値と時期までの推定価値の差分を使って価値を更新し、十分小さくなったとき収束したとして終了する。
- Q学習 行動価値関数Qを最大化して報酬最大化を目指す。TD法を使って価値を更新する。 価値ベースであり、方策オフである。経験再生ができる。ε貪欲法より探索と活用のトレードオフを行うことができる。

- SARSA Q学習と同じ。 異なる点はTD法を今期と次期に加え何期も考慮すること。何期も見るからリスク回避的が魅力 何期も考慮する場合は、方策オンになるらしい。←なんで？これある時点で推定した未来の行動と実際にその次期にになったときに取る行動は必ずしも一致しないから、過去の歴史で動いてないよってことかな？

- 方策勾配法 方策を最適化する際に誤差から勾配によって求める手法。 方策ベースであり、方策オンである。経験再生は一般にできない。

- Actor-Critic法 方策勾配法と同じ。異なる点は、方策を決める部分と価値を決める部分を分けて考える。これにより推定価値のバイアスを減らすことができる。（過大評価しない）のがメリット。方策ベースで方策オン。