$$A_t = \begin{cases} \underset{a}{\mathrm{argmax}} Q_t(a), & 1-\varepsilon \\ a \sim \mathrm{Uniform}(\mathcal{A}), & \varepsilon \end{cases}$$

$$V_{k+1}(s) = \max_a \sum_{s'} P(s'|s,a)\{r + \gamma V_k(s')\}$$

$$V(s_t) \leftarrow V(s_t) + \alpha\{G_t - V(s_t)\}$$

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

強化学習のエッセンス

2026/5/7 スタートアップゼミ#7 B4 長田晴人

$$(\mathcal{S}, \mathcal{A}, P, r, \gamma)$$

$$V(s_t) \leftarrow V(s_t) + \alpha\{r_t + \gamma V(s_{t+1}) - V(s_t)\}$$

$$Q(s_t,a_t) \leftarrow Q(s_t,a_t) + \alpha\left\{r_t + \gamma \max_a Q(s_{t+1},a) - Q(s_t,a_t)\right\}$$

強化学習 ざっくり

- 教師あり学習 ➢ 知識を持った外部の講師がいて，正解(と誤差)を示す
  ↪誤差が最小になるような設定を見つけ出す
- 教師なし学習
- 強化学習 ➢ 状態と行動に応じて報酬を得る 報酬最大化を目指す

強化学習エージェント

移動できたら，ごほうびをあげまちゅよ

試行錯誤の間，たまに移動できる動きをする

報酬

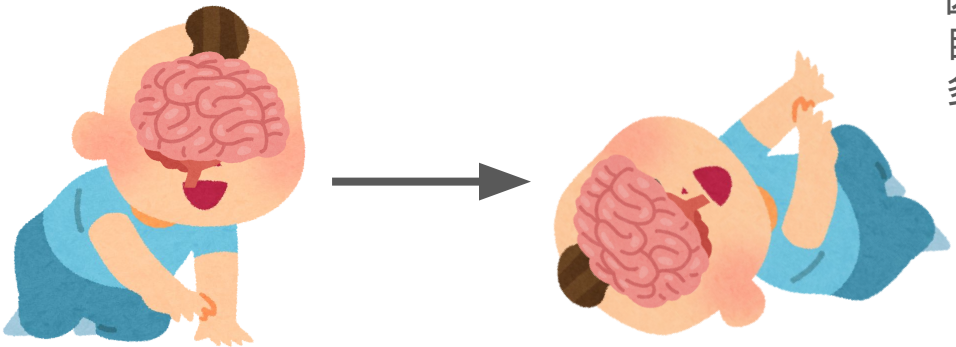
移動できる動きを選択しやすくなる

繰り返し

私たち

強化学習 特徴

- 相互作用 ➢ 選んだ行動に応じて環境の状態も変わる
- 探索と活用 ➢ 新たな行動を試すか，知識から導き出した最良行動か
- 遅延報酬 ➢ 良い行動をしてもすぐに報酬がもらえるとは限らない

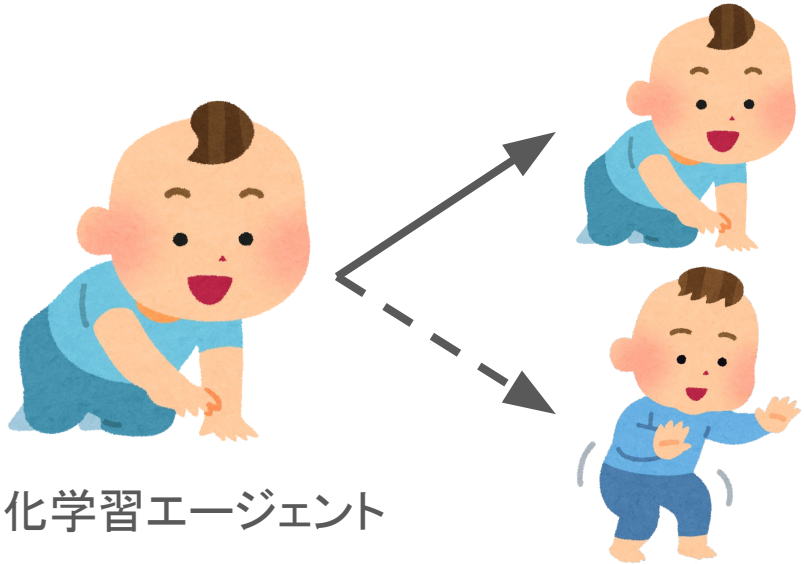


強化学習エージェント

因果・動作の結果・
目標のために何をすべきか など
多くの**情報を自分で作り出す** ことができる

強化学習 特徴

- 相互作用 ➢ 選んだ行動に応じて環境の状態も変わる
- **探索と活用** ➢ **新たな行動を試すか，知識から導き出した最良行動か**
- 遅延報酬 ➢ 良い行動をしてもすぐに報酬がもらえるとは限らない



強化学習エージェント

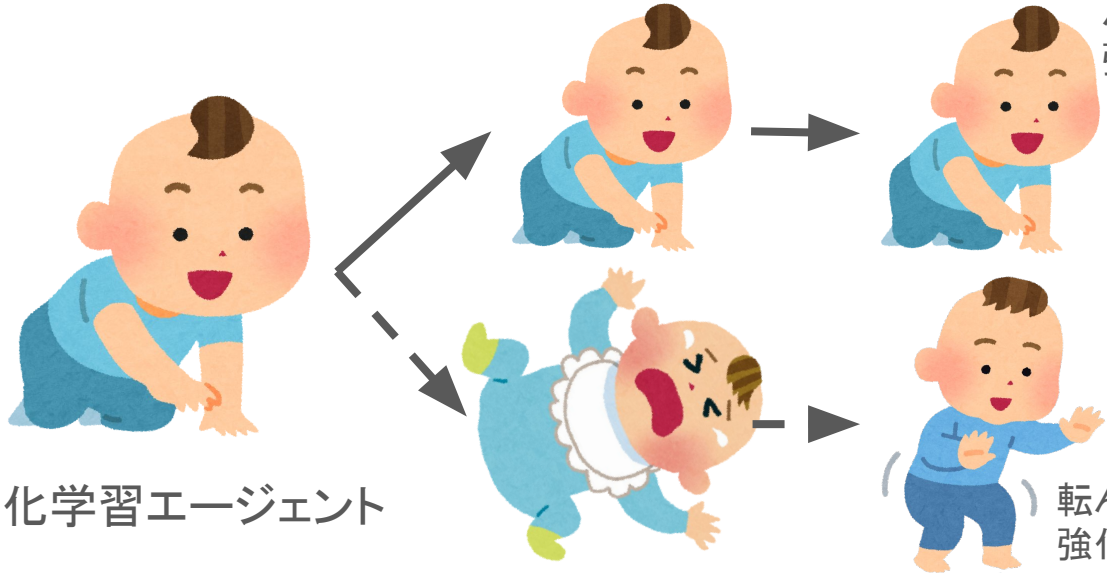
知識利用：過去に試みた行動の中で，
報酬を得るために最も効果的なものを選ぶ

どれくらいのバランスにするか
というジレンマ

探索：過去に試みたことのない行動を選択し，
将来的な行動選択を改善

強化学習 特徴

- 相互作用 ➢ 選んだ行動に応じて環境の状態も変わる
- 探索と活用 ➢ 新たな行動を試すか，知識から導き出した最良行動か
- 遅延報酬 ➢ 良い行動をしてもすぐに報酬がもらえるとは限らない



強化学習エージェント

ハイハイをし続け，速く移動できない
強化学習エージェント

報酬設計を誤ると，望まない学習結果が
もたらされる

転ぶという負の即時報酬の先に，
立って歩くという大きな**遅延報酬**がある

転んだ後立って速く歩くことに成功した
強化学習エージェント

強化学習 成立過程

1950s

H. Robbins, 1952

多腕バンディット問題
統計学・動物心理学
試行錯誤による学習

未知の環境において,
探索と利用のジレンマに
どう立ち向かうか?

「探索と活用」の原点

N. Metropolis, 1949

モンテカルロ法
シミュレーション
状態遷移の導入

1エピソードを通し,
その結果を全体にFB,
方策を改善する

結果からの**「価値推定」**

2分野の合流

1988

R. S. Sutton, 1988

TD学習
強化学習の成立
価値の更新手法

一歩進んだその直後に
次の一歩の予測を使い,
エピソード終了を待たず
今の自分の評価を直す.
経験と逆算の融合手法.

「予測」で進む方策評価

1990s

Rummery, 1994

SARSA
on-policy TD制御

実際に選んだ次の一手
その現実の価値を使い,
今いる場所を評価する.
実践ベースの更新手法.

「方策」に従うQ制御

マルコフ決定過程
制御工学・数理最適化
フレームワーク

環境がわかっている時の
複雑な問題の定式化

「マルコフ性」と
「ベルマン方程式」定義

R.Bellman, 1952

動的計画法
制御工学・離散最適化
反復計算での求解

環境がわかっている時の
複雑な問題をどう解くか

「方策反復」と
「価値反復」

R.Bellman, 1952

Q学習
off-policy TD制御

次に選ぶ手にかかわらず
一番良いはずの未来で,
今いる場所を評価する.
理想ベースの更新手法.

「貪欲」に近づくQ制御

Watkins, 1989

強化学習 成立過程

1950s

多腕バンディット問題
統計学・動物心理学
試行錯誤による学習

$$A_t = \begin{cases} \underset{a}{\mathrm{argmax}} Q_t(a), & 1-\varepsilon \\ a \sim \mathrm{Uniform}(\mathcal{A}), & \varepsilon \end{cases}$$

「探索と活用」の原点

マルコフ決定過程
制御工学・数理最適化
フレームワーク

$$(\mathcal{S}, \mathcal{A}, P, r, \gamma)$$

「マルコフ性」と
「ベルマン方程式」定義

モンテカルロ法
シミュレーション
状態遷移の導入

$$V(s_t) \leftarrow V(s_t) + \alpha \{G_t - V(s_t)\}$$

結果からの**「価値推定」**

動的計画法
制御工学・離散最適化
反復計算での求解

$$V_{k+1}(s) = \max_a \sum_{s'} P(s'|s,a)\{r + \gamma V_k(s')\}$$

「方策反復」と
「価値反復」

2分野の合流

1988

TD学習
強化学習の成立
価値の更新手法

$$V(s_t) \leftarrow V(s_t) + \alpha\{r_t + \gamma V(s_{t+1}) - V(s_t)\}$$

「予測」で進む方策評価

1990s

SARSA
on-policy TD制御

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

「方策」に従うQ制御

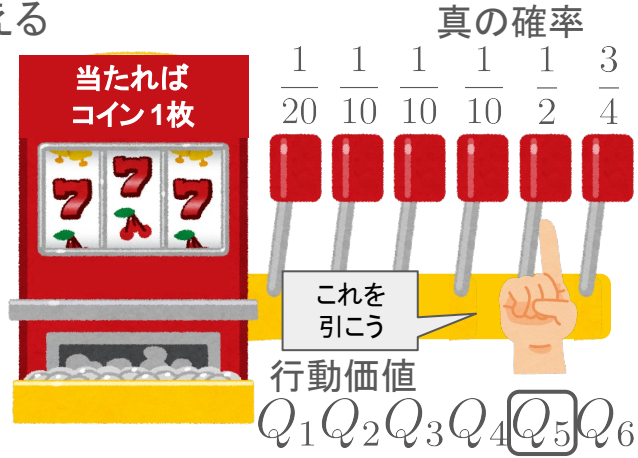
Q学習
off-policy TD制御

$$Q(s_t,a_t) \leftarrow Q(s_t,a_t) + \alpha\left\{r_t + \gamma \max_a Q(s_{t+1},a) - Q(s_t,a_t)\right\}$$

「貪欲」に近づくQ制御

多腕バンディット問題 - 強化学習の基礎概念の理解のための

真の確率(エージェントは知らない)が右図のようになっている6本腕バンディット問題
それぞれの腕で行動価値を考え，何らかの原理でどの腕を選ぶか考える
探索と知識利用のジレンマが腕の選び方にあらわれる



行動価値の決め方に関する手法

- 標本平均手法 → 経験平均をそのまま
- 固定学習率 → 現在行動価値と報酬の差にαをかけたものを足す

腕の選び方に関する手法

- ε-greedy → $1-\varepsilon$で行動価値最大の腕，εで一様ランダムな腕を引く
- Softmax → 行動価値に応じて選択確率を滑らかに割り振る
- UpperConfidenceBound → 行動価値+試行回数の少なさ が最大である腕を引く
- Optimistic初期値 → 初期の行動価値を過大にしておくことで失望を生み，探査を促進
- Thompson sampling → 経験から分布を推定，サンプリングした値が最大の腕を引く

多腕バンディット問題 - それぞれの手法 数式の姿

行動価値の決め方に関する手法

- 標本平均手法 → $Q_t(a) = \frac{r_1 + r_2 + \cdots + r_{t-1}}{N_t(a)}$
- 固定学習率 → $Q_{t+1}(a_t) = Q_t(a_t) + \alpha\left(r_t - Q_t(a_t)\right)$

普通の相加平均
非定常では❌

今後もよく見る形
再帰を解くと加重平均
新しい報酬ほど大きな重みを持つ

$$Q_{n+1} = (1-\alpha)^n Q_1 + \sum_{i=1}^{n} \alpha(1-\alpha)^{n-i} r_i$$

腕の選び方に関する手法

- ε-greedy → $A_t = \begin{cases} \underset{a}{\operatorname{argmax}} Q_t(a), & 1-\varepsilon \\ a \sim \mathrm{Uniform}(\mathcal{A}), & \varepsilon \end{cases}$

εは変化させることも可
最初は探索，後は貪欲に

- Softmax → $P(a_t = a) = \frac{\exp(Q_t(a)/\tau)}{\sum_{b \in \mathcal{A}} \exp(Q_t(b)/\tau)}$

見たことあるような形
τは温度で探索貪欲指標

- UpperConfidenceBound → $a_t = \underset{a \in \mathcal{A}}{\operatorname{argmax}} \left[Q_t(a) + c\sqrt{\frac{\log t}{N_t(a)}} \right]$

右辺第2項が試行回数の
少なさを表す項

- Optimistic初期値 → $Q_1(a) = Q_0 \ (Q_0 > Q_{\max})$

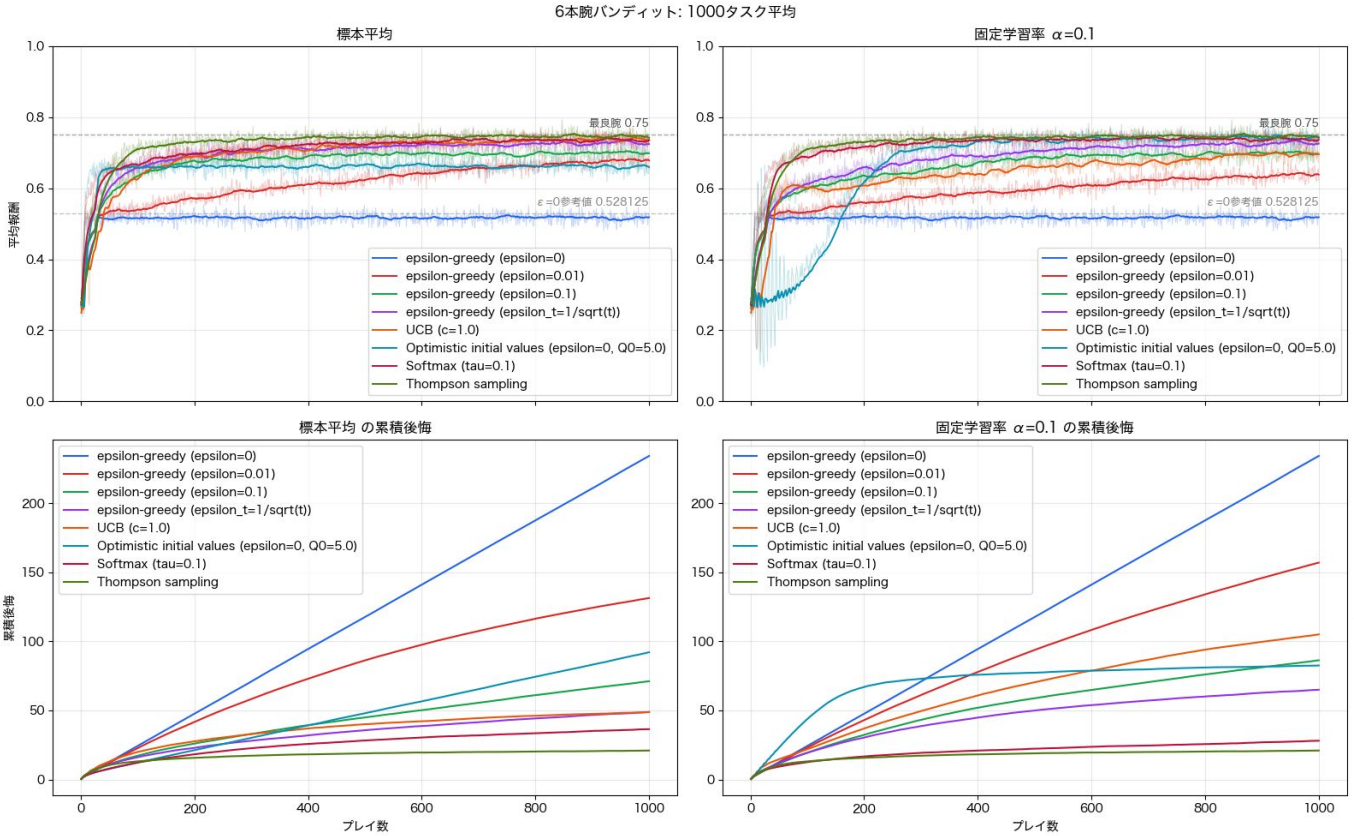
固定学習率の再帰を解くと
初期値が残り続けると分かる

- Thompson sampling → $\theta_a \sim p(\theta_a | \mathcal{D}_t),\ a_t = \underset{a \in \mathcal{A}}{\operatorname{argmax}} \theta_a$

Q値ではなく，Q値が従う
だろう分布から値をとる

多腕バンディット問題 - 手法比較

1000タスク平均報酬 (実線は25プレイ移動平均)



・ε-greedy(ε=0)では，初めて当てた腕を盲信的に信じるため，0.53程度に落ち着く．

・εが大きいほど結果が良くなっているように見えて，だんだんどんな時でもランダムに探索することが足を引っ張っており，限界がある．

・Optimistic初期値は，初期値バイアスが残るα学習率手法で特徴が見られる．はじめは全体を探索し，最適解を確実に見つけ出している．

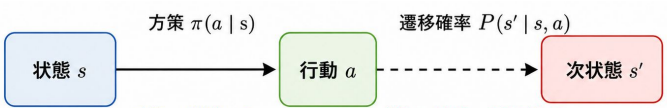
・Thompson samplingは，そもそも行動価値を更新するのではなく推定分布そのものを更新している．その上，ハイパーパラメータが存在しない点からも強い．

前提の枠組み - マルコフ決定過程 (MarkovDecisionProcess)

- 状態空間 $\mathcal{S}$
  - 現在状態 s
  - 次状態　s'
- 行動空間 $\mathcal{A}$
  - 現在行動 a
- 遷移確率 $P(s'|s,a)$
- 報酬関数 $r(s,a,s')$
- 割引率　γ

時刻 t 以降の累積割引報酬リターン　$G_t = r_t + \gamma G_{t+1}$

方策 $\pi(a|s)$ ←で最適化したい



遷移確率 $P(s'|s,a)$ が既知(つまり環境モデルが既知)を想定

状態価値関数: ある状態の時のリターンの期待値

$$V^\pi(s) = \mathbb{E}_\pi[G_t|_s] = \sum_{a\in\mathcal{A}} \pi(a|s) \sum_{s'\in\mathcal{S}} P(s'|s,a)(r(s,a,s') + \gamma V^\pi(s'))$$

行動価値関数: ある状態である行動をとった時のリターンの期待値

$$Q^\pi(s,a) = \mathbb{E}_\pi[G_t|_{s,a}] = \sum_{s'\in\mathcal{S}} P(s'|s,a) \left(r(s,a,s') + \gamma \sum_{a'\in\mathcal{A}} \pi(a'|s') Q^\pi(s',a') \right)$$

πについて $V^\pi(s), Q^\pi$ を最大化

$$V^*(s) = \max_\pi V^\pi(s), Q^*(s,a) = \max_\pi Q^\pi(s,a)$$

最適状態価値関数：**状態の価値** は取れる行動のうち**価値が最も高い** もの

$$V^*(s) = \max_a Q^*(s,a)$$

最適行動価値関数：**状態と行動を選んだ際** の期待値は**次の状態価値** で決まる

$$Q^*(s,a) = \sum_{s'} P(s'|s,a)\,(r + \gamma V^*(s'))$$

前提の枠組み - 動的計画法 (Dynamic Programming)

最適状態価値関数：状態の価値 は取れる行動のうち**価値が最も高い**もの

$$V^*(s) = \max_a \sum_{s'} P(s'|s,a)\,(r + \gamma V^*(s'))$$

環境モデル既知という状況で
この**ベルマン最適方程式**を解き，
最適化するような方策　を知る

方策反復手法

現在の方策のもと
価値を伝播させる

更新価値に応じて
方策を改善する

方策が変化しなくなるまで反復

数式では…

$\pi_0 \to V^{\pi_0} \to \pi_1 \to V^{\pi_1} \to \pi_2 \to \cdots$ と反復

$$V^{\pi_k}(s) = \sum_a \pi_k(a|s) \sum_{s'} P(s'|s,a)\,(r + \gamma V^{\pi_k}(s'))$$

$$\pi_{k+1}(s) = \arg\max_a \sum_{s'} P(s'|s,a)\,(r + \gamma V^{\pi_k}(s'))$$

$\pi_{k+1} = \pi_k$ となったとき，　$V^* = V^{\pi_k}, \pi^* = \pi_k$

価値反復手法

最善選択をする
方策のもと
価値を伝播させる

最適価値から
最適方策を得る

価値が
ほとんど変化しなくなるまで反復

数式では…

$V_0 \to V_1 \to V_2 \to \cdots$　と反復

$$V_{k+1}(s) = \max_a \sum_{s'} P(s'|s,a)\,(r + \gamma V_k(s'))$$

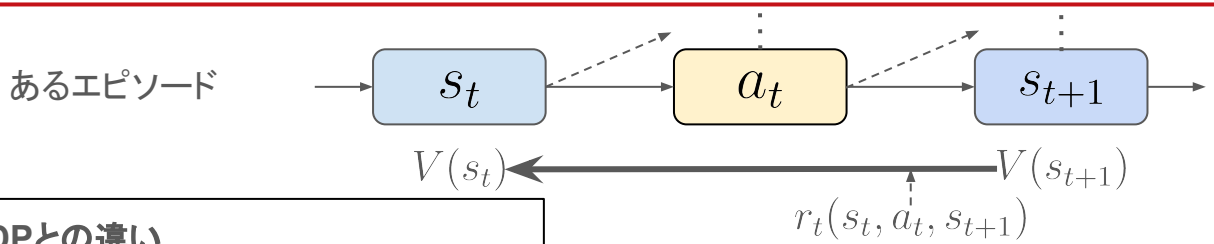
例えば　$\Delta = \max_s |V_{k+1}(s) - V_k(s)| < \theta$　となったとき，　$V^*(s) = V_k(s)$

$$\pi^*(s) = \arg\max_a \sum_{s'} P(s'|s,a)\left(r + \gamma V^*(s')\right)$$

Temporal Difference 学習

モンテカルロ法の価値更新式(TD学習ではない!)　$V(s_t) \leftarrow V(s_t) + \alpha(\quad G_t \quad - V(s_t))$

TD(0)学習　の価値更新式　$V(s_t) \leftarrow V(s_t) + \alpha(r_t + \gamma V(s_{t+1}) - V(s_t))$



実際に得るリターンの代わりに
得ると予測している期待値で更新

DPとの違い
DPでは環境モデルが既知のため,
全状態の全行動をスイープできる
TD学習は**通った状態した行動**のみ
価値が更新できる.

モンテカルロ法との違い
エピソード終了後に報酬をロール
バックするのではなく, **ステップ**
ごとに価値を更新する

価値の伝播と価値関数への学習の融合

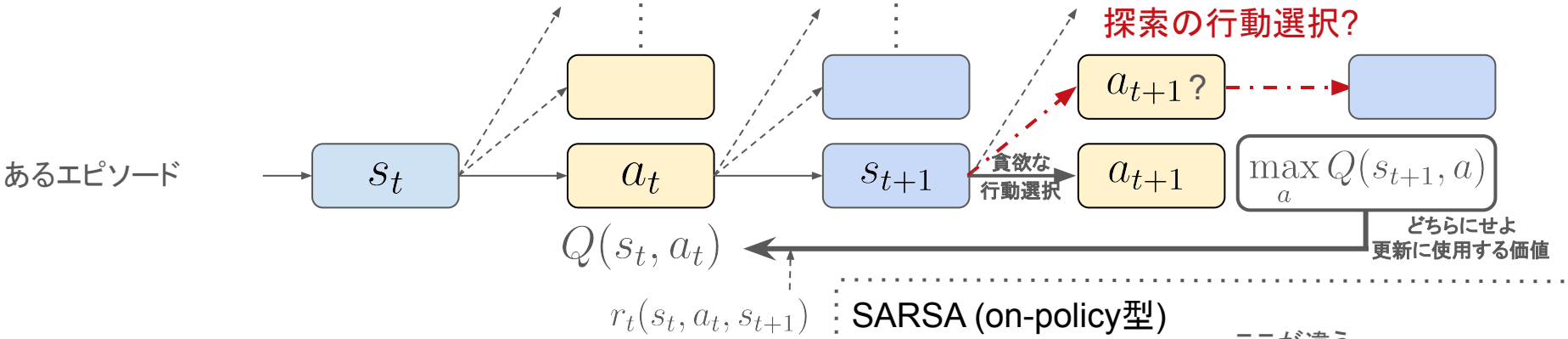
これで全て解決…かと思いきや

方策改善がTD学習では**できない**

(環境モデル(遷移関数)未知のため,
状態価値関数がわかっていても
どの行動を選べばいいかはわからないので)

Q学習 $Q(s_t,a_t) \leftarrow Q(s_t,a_t) + \alpha \left\{ r_t + \gamma \max_a Q(s_{t+1}, a) - Q(s_t, a_t) \right\}$

状態価値ではなく行動価値を参照することで解決!



環境モデル(遷移確率P)を知らなくても,
行動価値 が**最も大きな行動** を選べば**貪欲方策** になる.

モデル既知では可能だった,**「最適方策を知る」**
というゴールに**モデル未知** で辿りついた!

SARSA (on-policy型)

ここが違う

$$Q(s_t,a_t) \leftarrow Q(s_t,a_t) + \alpha \left\{ r_t + \gamma \boxed{Q(s_{t+1}, a_{t+1})} - Q(s_t, a_t) \right\}$$

Q学習 (off-policy型)では, 探索で最大価値でない行動を選んだ時も, Q値の更新は最大価値で行っていた.
対して, SARSAでは, 価値に関わらず実際に進んだ行動の価値で更新をする. つまり探索の際にそれによる低報酬で過去の行動価値を更新してしまう.

強化学習 成立過程

1950s

多腕バンディット問題
統計学・動物心理学
試行錯誤による学習

未知の環境において,
探索と利用のジレンマに
どう立ち向かうか?

「探索と活用」の原点

マルコフ決定過程
制御工学・数理最適化
フレームワーク

環境がわかっている時の
複雑な問題の定式化

「マルコフ性」と
「ベルマン方程式」定義

モンテカルロ法
シミュレーション
状態遷移の導入

1エピソードを通し,
その結果を全体にFB,
方策を改善する

結果からの**「価値推定」**

動的計画法
制御工学・離散最適化
反復計算での求解

環境がわかっている時の
複雑な問題をどう解くか

「方策反復」と
「価値反復」

2分野の合流

1988

TD学習
強化学習の成立
価値の更新手法

一歩進んだその直後に
次の一歩の予測を使い,
エピソード終了を待たず
今の自分の評価を直す.
経験と逆算の融合手法.

「予測」で進む方策評価

1990s

SARSA
on-policy TD制御

実際に選んだ次の一手
その現実の価値を使い,
今いる場所を評価する.
実践ベースの更新手法.

「方策」に従うQ制御

Q学習
off-policy TD制御

次に選ぶ手にかかわらず
一番良いはずの未来で,
今いる場所を評価する.
理想ベースの更新手法.

「貪欲」に近づくQ制御

強化学習 成立過程

1950s

多腕バンディット問題
統計学・動物心理学
試行錯誤による学習

$$A_t = \begin{cases} \mathrm{argmax}_a Q_t(a), & 1-\varepsilon \\ a \sim \mathrm{Uniform}(\mathcal{A}), & \varepsilon \end{cases}$$

「探索と活用」の原点

マルコフ決定過程
制御工学・数理最適化
フレームワーク

$$(\mathcal{S}, \mathcal{A}, P, r, \gamma)$$

「マルコフ性」と
「ベルマン方程式」定義

モンテカルロ法
シミュレーション
状態遷移の導入

$$V(s_t) \leftarrow V(s_t) + \alpha \{G_t - V(s_t)\}$$

結果からの**「価値推定」**

動的計画法
制御工学・離散最適化
反復計算での求解

$$V_{k+1}(s) = \max_a \sum_{s'} P(s'|s,a)\{r + \gamma V_k(s')\}$$

「方策反復」と
「価値反復」

2分野の合流

1988

TD学習
強化学習の成立
価値の更新手法

$$V(s_t) \leftarrow V(s_t) + \alpha\{r_t + \gamma V(s_{t+1}) - V(s_t)\}$$

「予測」で進む方策評価

1990s

SARSA
on-policy TD制御

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

「方策」に従うQ制御

Q学習
off-policy TD制御

$$Q(s_t,a_t) \leftarrow Q(s_t,a_t) + \alpha\left\{r_t + \gamma \max_a Q(s_{t+1},a) - Q(s_t,a_t)\right\}$$

「貪欲」に近づくQ制御

さらなる発展

- DQN　　Q学習は，状態の次元が多いと破綻するため，
Qをニューラルネットワークで求められるよう近似する．

$Q(s,a) \approx Q(s,a;\theta)$ （θはネットワークのパラメータ）は，　$y_t = r_t + \gamma \max_{a'} Q(s_{t+1},a';\theta^-)$ を

目標に更新される．具体的には，　$L(\theta) = \left(y_t - Q(s_t,a_t;\theta)\right)^2$　が小さくなるよう　を変化させる．

- Actor-Critic　　連続行動空間では，最適方策が簡単には表せない．
そこで，価値推定を利用して方策そのものを関数近似する．

方策と状態価値がそれぞれパラメータを　として　　と表され，

$\theta \leftarrow \theta + \alpha_\theta (r_t + \gamma V_w(s_{t+1}) - V_w(s_t)) \dfrac{\nabla_\theta \pi_\theta}{\pi_\theta}$ ，$w \leftarrow w + \alpha_w (r_t + \gamma V_w(s_{t+1}) - V_w(s_t)) \nabla_w V_w(s_t)$

で更新される．

他にも色々ありますがテキスト以上の説明ができないため割愛

協調型自律運転のための交差順序最適制御への応用

十字交差点があり，信号制御を強化学習で試みる

phase group $\mathcal{G} = \{\mathrm{NS}, \mathrm{EW}\} \times \{\mathrm{Straight}, \mathrm{Left}, \mathrm{Right}\}$

phase group features
$\mathcal{F} = \{\mathrm{QueueLength}, \mathrm{ApproachCount}, \mathrm{HeadDistance}, \mathrm{OldestWait}, \mathrm{RecentArrival}\}$

current signal phase $\mathcal{P} = \mathcal{A} = \mathcal{G} \sqcup \{\mathrm{AllRed}\}$

global features $\mathcal{F}' = \{\mathrm{PhaseDuration}, \mathrm{CrossingCount}, \mathrm{TotalQueue}\}$

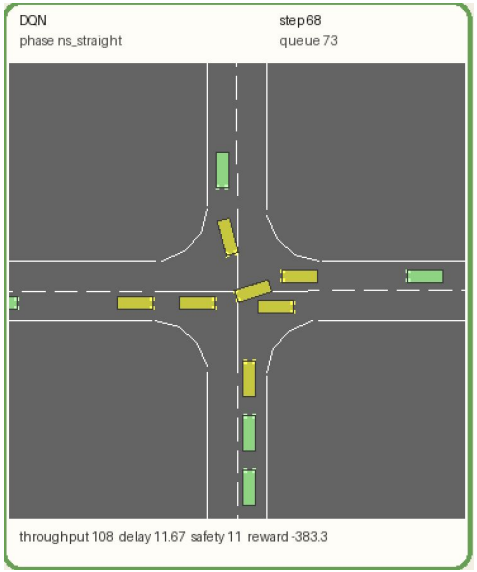
として，状態空間 $\mathcal{S} = \mathcal{G} \times \mathcal{F} \sqcup \mathcal{P} \sqcup \mathcal{F}'$

$\mathcal{S}$ は40次元，$\mathcal{A}$は7つ ←次元数が多く，表形式Q学習では　は
9次元以外の状態情報を捨象して考えている

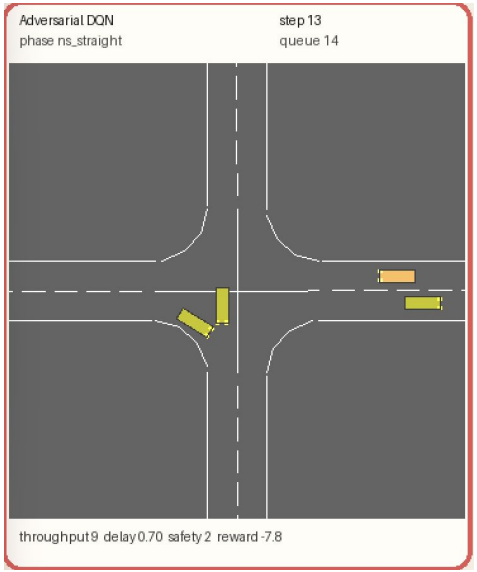
状態に応じて信号指示という行動を変え，
円滑な交通に導くことが目標．

←はじめから説明用に入っていたgif

協調型自律運転のための交差順序最適制御への応用



デッドロック



事故

防がなければならない (大きな負報酬)

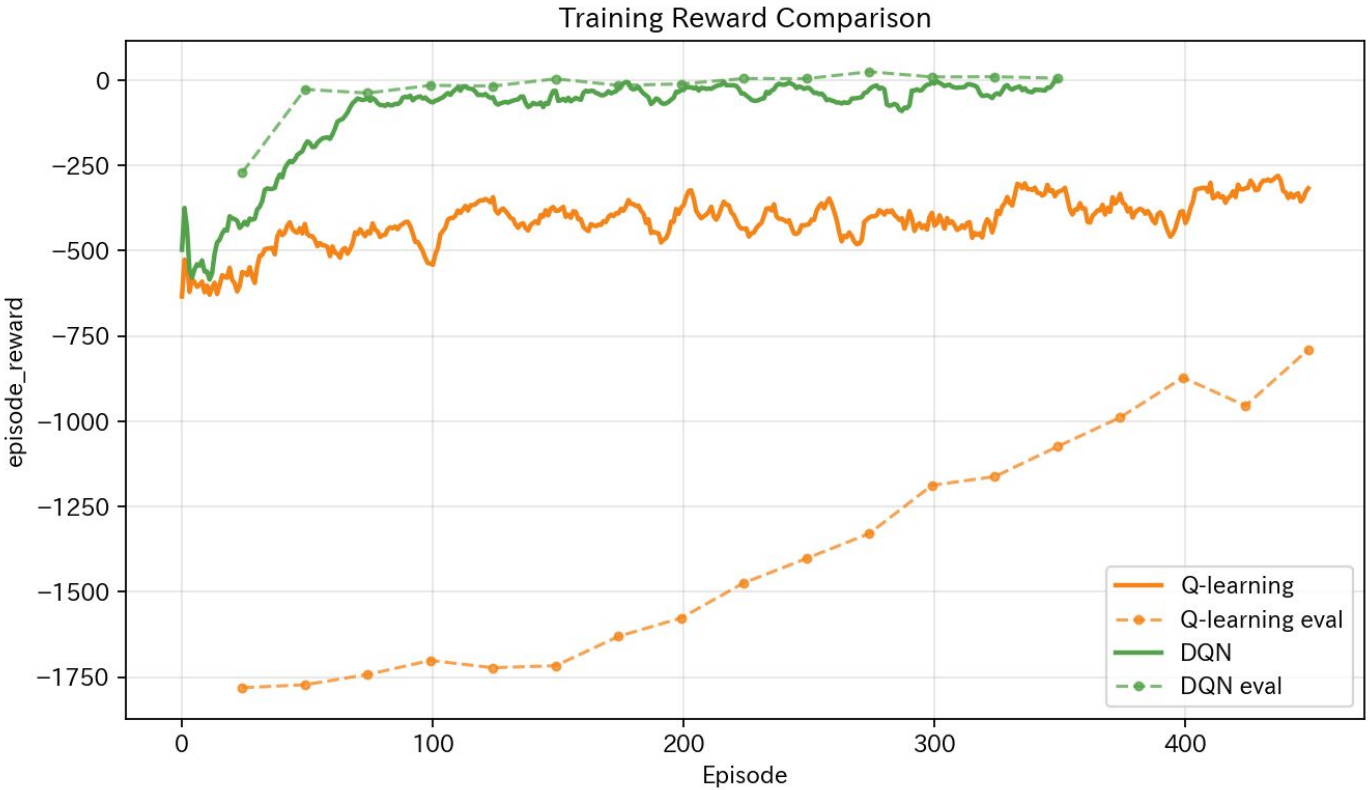
十字交差点があり，信号制御を強化学習で試みる

$$r_t = 1.5N_t + 0.2(W_{t-1} - W_t) - 0.08Q_t - 5.0V_t - 0.15I_t - 0.05A_t$$

N_t :number of vehicles that start crossing at time t
W_t :total waiting time of all queued vehicles at time t
Q_t :total queue length at time t
V_t :1 if a safety violation occurs, and 0 otherwise
I_t :1 if the signal phase is switched, and 0 otherwise
A_t :1 if the all−red phase is selected, and 0 otherwise

- 車両通過，待ち時間減少に正報酬
- 待ち時間増加，待ち行列，事故，
  頻繁な指示変更，長時間の全赤指示に負報酬

協調型自律運転のための交差順序最適制御への応用

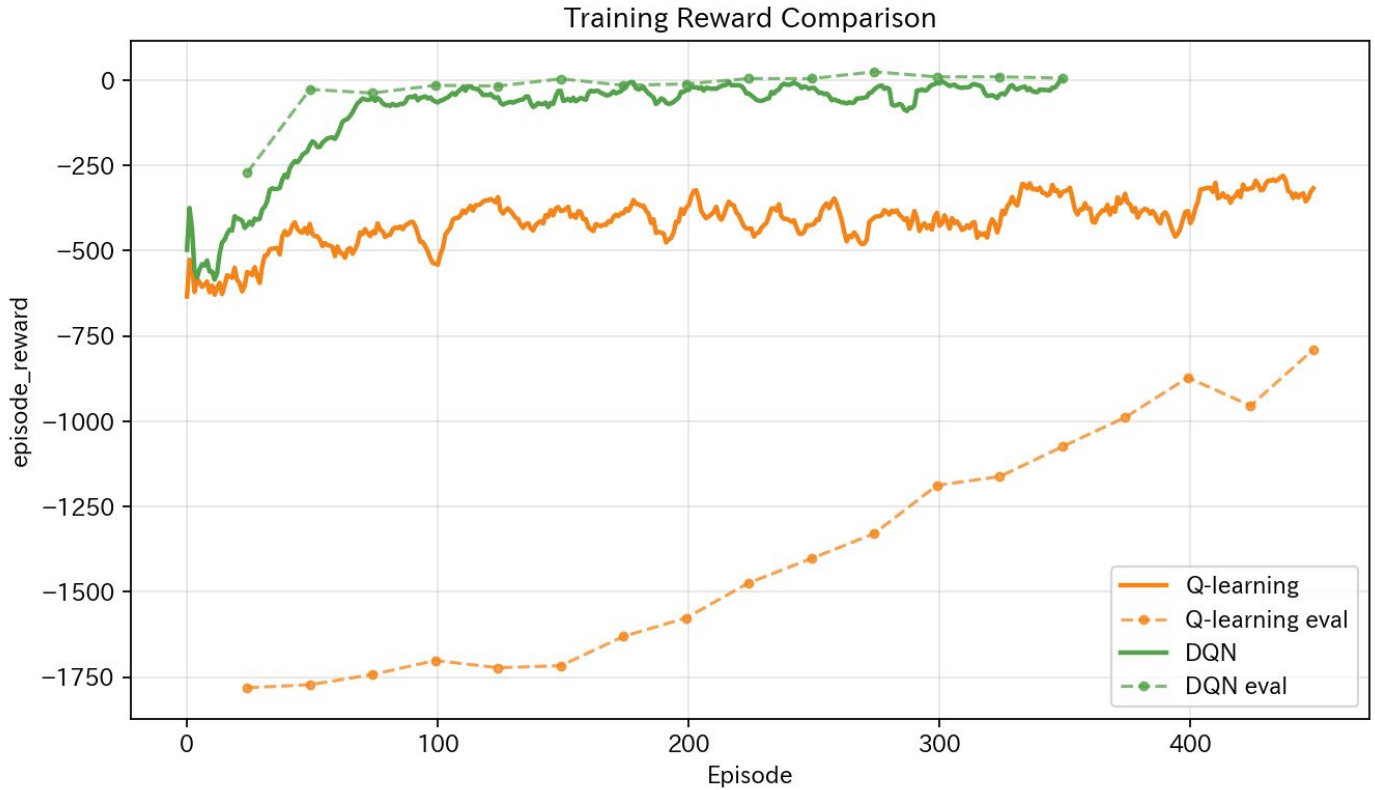


学習エピソード数：Q　　450
　　DQN　350
エピソード長　：120 steps
ε-greedy exploration
- Q：エピソード単位減衰
- DQN：ステップ単位減衰

DQNでは評価(探索なし)で学習時より良い結果だが，Q学習では評価時に学習時より大幅に悪い結果となっている．

そもそも状態空間が異なる，εの減衰が異なるなどあるが，大幅にDQNが早く高い報酬を得ている．

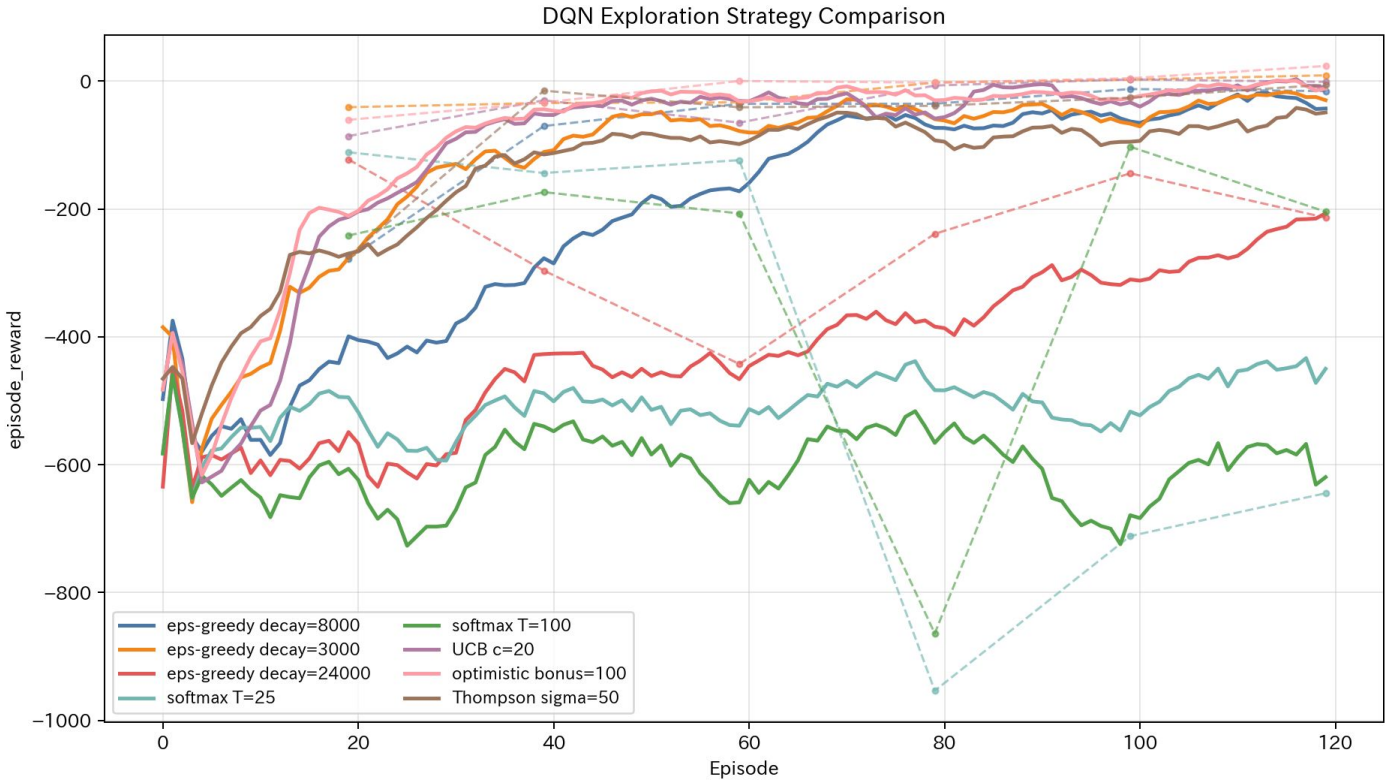
協調型自律運転のための交差順序最適制御への応用



DQNでも報酬が0を超えていない．通過の報酬を罰が凌駕している．

いったんの目標
DQNの探索方法の変更で報酬にどんな影響が出るか比較する．

協調型自律運転のための交差順序最適制御への応用



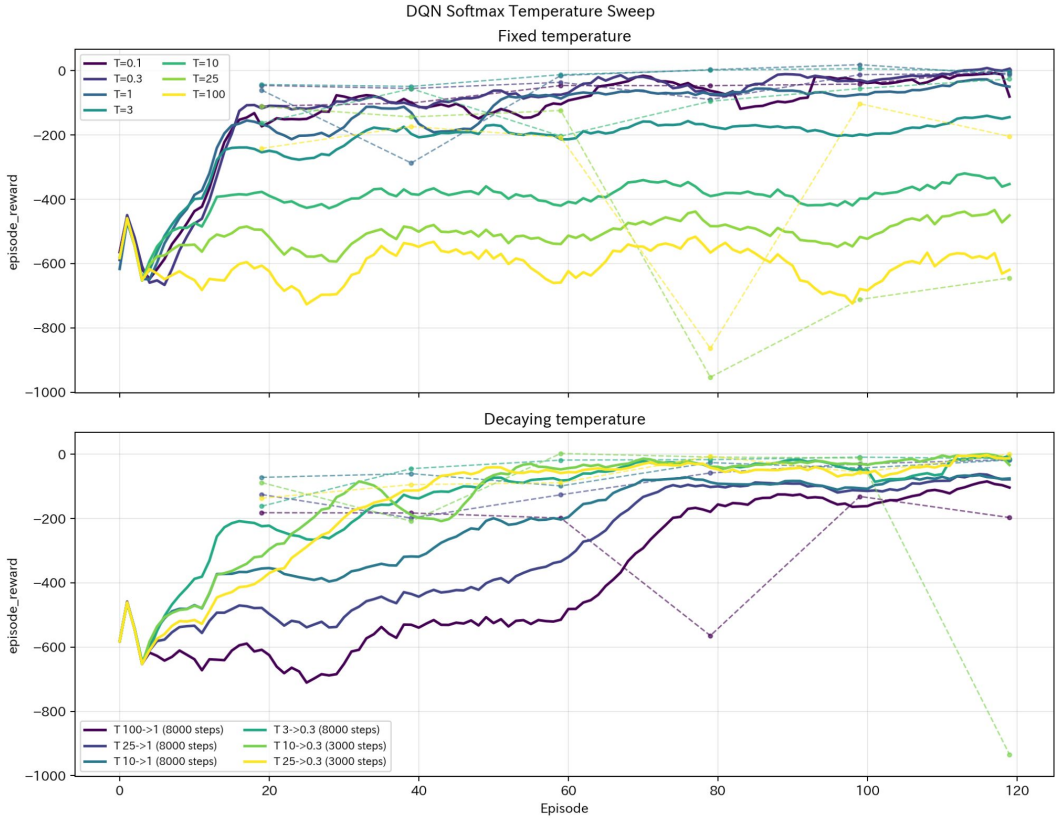
バンディットの時に調べた
色々な手法とDQN

バンディットの時にかなり良かった
たsoftmaxが全然よくない
εと異なり，温度は価値の次元を
持つ

$$P(a_t=a)=\frac{\exp(Q_t(a)/\tau)}{\sum_{b\in A}\exp(Q_t(b)/\tau)}$$

温度のスケールを合わせたら
もっと良い結果になるか?

協調型自律運転のための交差順序最適制御への応用



温度が低いほど高い報酬を得ている
初めのτ=25,100はスケールが大きすぎた

そもそも温度スケールはどう決まるのか

$$P(a_t=a)=\frac{\exp(Q_t(a)/\tau)}{\sum_{b\in A}\exp(Q_t(b)/\tau)}$$

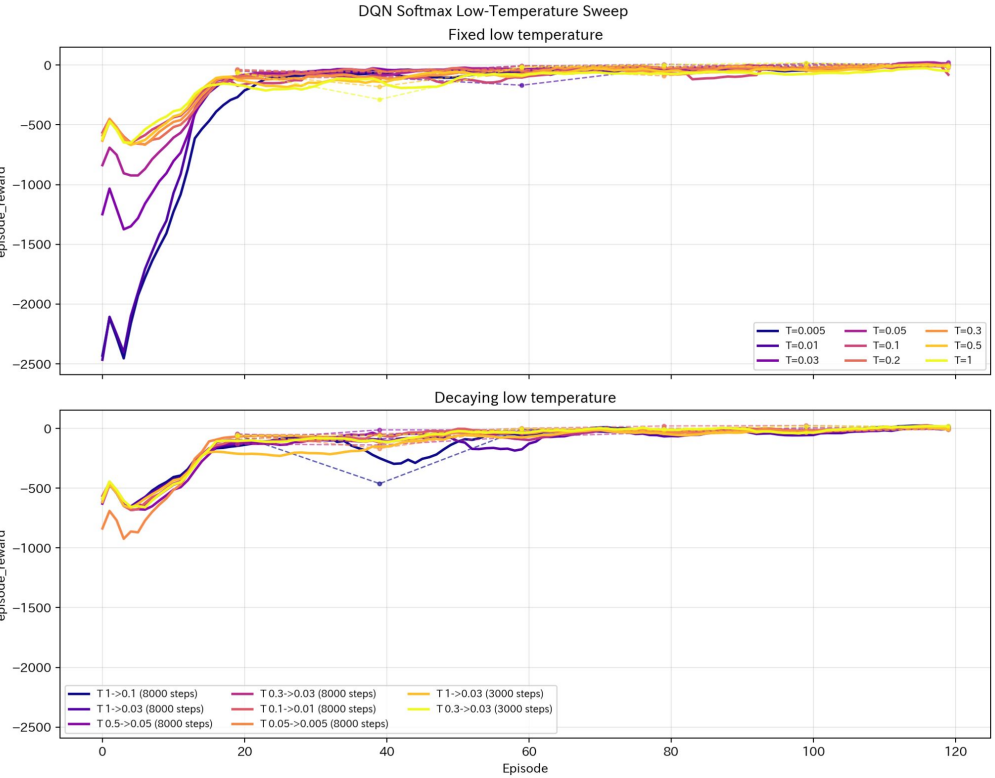
指数のため，Qの絶対的な大きさは意味を持たない．Qの範囲のスケールと関係がありそう

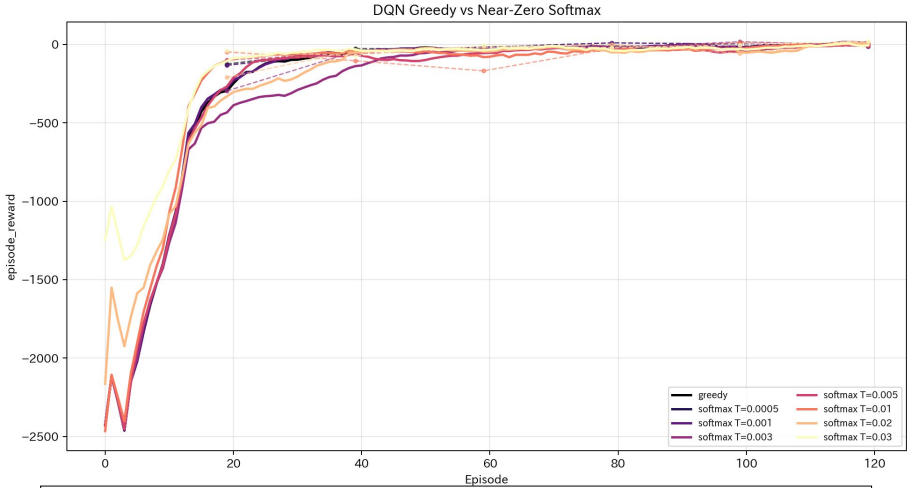
$\frac{Q(s,a_i)-Q(s,a_j)}{\tau}$ や，特に

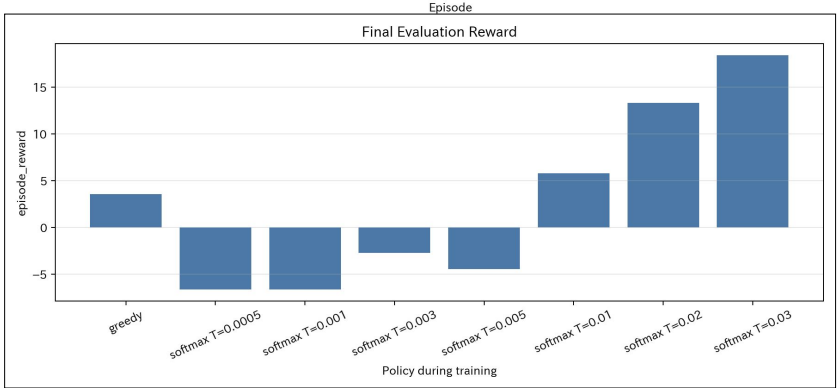
$$\Delta(s)=Q(s,a^*)-Q(s,a^{(2)})$$

に注目してみる．
今回の学習内での $\Delta(s)$ の中央値は0.023と，
最も良い結果の出たτ=0.03と同じオーダー
探索有意と活用優位を切り替えられる値．

協調型自律運転のための交差順序最適制御への応用







所感

- Q学習という名前で手法がまとめられたのが1989年のクリストファー・ワトキンズの博士論文とのことで，学生でもこんなものを考えられるのかと感動し，自分が博士でどんなことができるだろうとわくわくした．
- 本を読み出したら面白く，テキスト要約に時間を割いてしまい気づいたら課題にかけるための時間がほとんど残っていなかったことがかなり心残り，特に古橋先輩が用意してくださったたくさんのテーマに手をつけないまま基礎的な部分の変化だけで終わってしまったのが残念な上，申し訳ない．
- Recursive Logit Modelが逆強化学習，特に最大エントロピー型の逆強化学習に近いモデルであるらしく，その点からもこの分野について深く知ることが研究の助けになるかもしれないと感じた．

https://github.com/OSADA-Haruto/bandit
バンディットの学習のコードです

交差点の方のコードはプライベートリポジトリにあります
誰に見えていて誰に見えていないのかわからないです