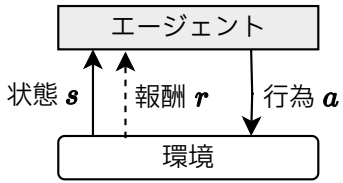
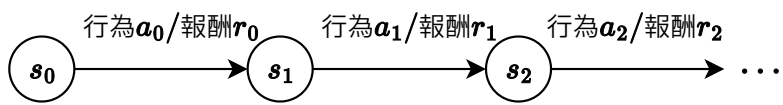
15. 強化学習



(a) 強化学習の枠組み



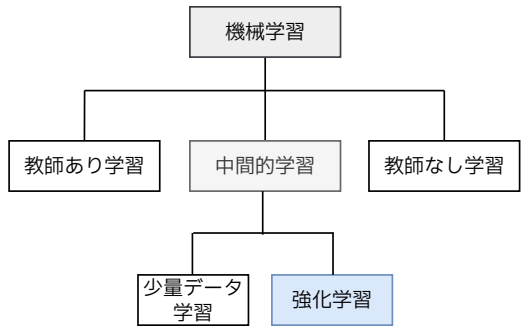
(b) エピソード

- 15.1 強化学習とは
- 15.2 1状態問題 K-armed bandit
- 15.3 複数の状態をもつ問題
- 15.4 深層強化学習
- 15.5 方策勾配法とLLMの学習

- 荒木雅弘：『Pythonではじめる機械学習』（森北出版，2025年）
- スライドとコード

15.1 強化学習とは (1/2)

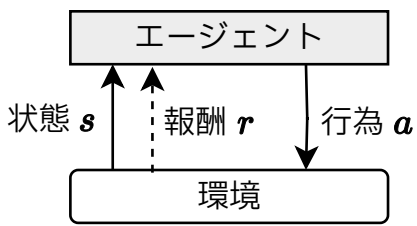
- 強化学習の位置付け：中間的学習
  - 状態変化を伴う環境下で行動するエージェントを想定する
  - 正解（状態に対する正しい行為）は与えられず，時間遅れを伴った報酬（数値）として形を変えて与えられる



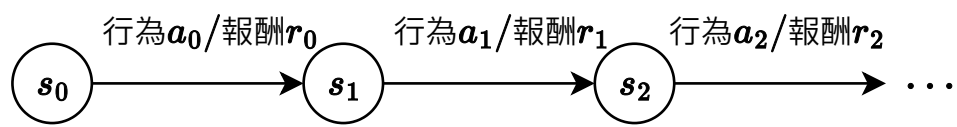
- 報酬仮説
  - 学習目標は累積期待報酬最大化で記述できる

15.1 強化学習とは (2/2)

- 強化学習の設定
  - マルコフ決定過程
    - 離散的に進む時刻 t の各時点において，エージェントは環境に対して行為 a_t を行い，環境から行為の結果変化した状態 s_{t+1} と報酬 r_{t+1} を受け取る
    - 環境にマルコフ性を仮定
      - 遷移先は直前の状態と行為のみに依存し，報酬は直前の状態と遷移先のみに依存する
  - エージェントは報酬の期待値が最大となる政策 π （状態から行為への写像）を学習する



(a) 強化学習の枠組み



(b) エピソード

15.2 1状態問題 -K-armed bandit- (1/3)

- K-armed bandit 問題の定義
  - K 本の腕を持つスロットマシンを考える

  - i 番目の腕を引く行為: a_i，（即時）報酬：$r(a_i)$，行為の価値: $Q(a_i)$ $(i = 1, \ldots, K)$
  - 報酬が確定的な場合
    - すべての a_i を1度ずつ試み，$Q(a_i) = r(a_i)$ が最大となる a_i が最適な行為
  - 報酬が確率的な場合
    - すべての a_i を何度か試み，報酬の平均値 $Q(a_i) = \mathbb{E}(r(a_i))$ が最大となる a_i が最適な行為

15.2 1状態問題 -K-armed bandit- (2/3)

- 時刻 t での報酬の平均値 $Q_t(a_i)$ の計算

$$
\begin{aligned}
Q_t(a_i) &= \frac{1}{t}\sum_{j=1}^{t} r_j(a_i) \\
&= \frac{1}{t}\left(r_t(a_i) + \sum_{j=1}^{t-1} r_j(a_i)\right) \\
&= \frac{1}{t}\left(r_t(a_i) + (t-1)Q_{t-1}(a_i)\right) \\
&= Q_{t-1}(a_i) + \frac{1}{t}(r_t(a_i) - Q_{t-1}(a_i))
\end{aligned}
$$

- Q値のインクリメンタルな更新式（更新後の値 = 現在の値 + 学習率 * 誤差）
  - 学習率 η は t の増加に伴って減少させるべきだが，t が大きいときは定数として扱える

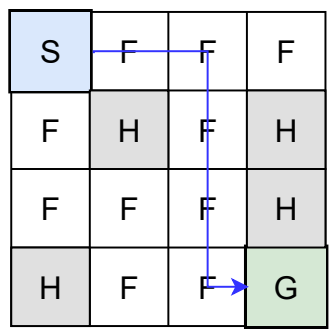
$$
Q_{t+1}(a_i) = Q_t(a_i) + \eta(r_{t+1}(a_i) - Q_t(a_i))
$$

15.2 1状態問題 -K-armed bandit- (3/3)

- どのようにして行為 a_i を選ぶか
  - 常に $Q_t(a_i)$ が最大のものを選ぶ方法
    - もっと良い行為があるのに見逃してしまうかもしれない
  - いろいろな a_i を何度も試みる方法
    - 無駄な行為を何度も行ってしまうかもしれない
- ϵ-greedy法
  - 確率 ϵ でランダムに行為を選ぶ
  - 確率 $1-\epsilon$ でその時点においてもっともQ値が高い行為を選ぶ

15.3 複数の状態をもつ問題 (1/9)

- 例題: frozen lake 問題
  - エージェントはタイル上で初期状態 S から終了状態 G を目指して移動する
    - 初期状態から終了状態に至る期間をエピソードとよぶ
    - 1エピソードで得られる報酬の期待値を最大とする政策の獲得を目標とする
  - タイルの種類
    - F (Frozen) : 歩行可能。ただし，滑る設定では意図した方向に移動できないことがある
    - H (Hole) : 穴に落ちてエピソードは終了する
  - 報酬の例 : G は 1, H は -1



15.3 複数の状態をもつ問題 (2/9)

- マルコフ決定過程：状態遷移を伴う問題の定式化
  - 時刻 t における状態 $s_t \in S$，初期状態分布 $P(s_0)$
  - 時刻 t における行為 $a_t \in A(s_t)$
  - 報酬 $r_t \in \mathbb{R}$，確率分布 $p(r_t \mid s_t, a_t)$
  - 次状態 $s_{t+1} \in S$，確率分布 $P(s_{t+1} \mid s_t, a_t)$
  - 価値関数
    - $V^{\pi}(s_t)$: 状態 s_t から政策 π に従って行動したときに得られる価値
    - $Q(s_t, a_t)$: 状態 s_t における行為 a_t の価値

15.3 複数の状態をもつ問題 (3/9)

- 学習目標
  - 最適政策 π^* の獲得
  - 状態価値関数
    - 時刻 t の状態が s_t で，以降，政策 π に従ったときに得られる累積報酬の期待値
      - γ : 割引率　　$0 \leq \gamma < 1$

$$V^{\pi}(s_t) = \mathbb{E}(r_t + \gamma r_{t+1} + \gamma^2 r_{t+2} + \dots) = \mathbb{E}(\sum_{i=0}^{\infty} \gamma^i r_{t+i})$$

- 累積報酬の期待値（＝将来の平均）が最大となる政策が最適政策

$$\pi^* \equiv \arg\max_{\pi} V^{\pi}(s_t), \ \forall s_t$$

15.3 複数の状態をもつ問題 (4/9)

- 価値反復法
  - 状態価値関数を再帰方程式（ベルマン方程式）で表し，繰り返し計算で収束させる

$$V^*(s_t) = \max_{a_t} Q^*(s_t, a_t) = \max_{a_t} \mathbb{E}\left(\sum_{i=0}^{\infty} \gamma^i r_{t+i}\right) = \max_{a_t} \mathbb{E}\left(r_t + \gamma \sum_{i=0}^{\infty} \gamma^i r_{t+i}\right)$$
$$= \max_{a_t} \mathbb{E}(r_t + \gamma V^*(s_{t+1}))$$

- 状態遷移確率を明示したベルマン方程式

$$V^*(s_t) = \max_{a_t} \{\mathbb{E}(r_{t+1}) + \gamma \sum_{s_{t+1}} P(s_{t+1}|s_t, a_t) V^*(s_{t+1})\}$$

15.3 複数の状態をもつ問題 (5/9)

- 価値反復法のアルゴリズム
  - 入力 : マルコフ決定過程
  - 出力 : 状態価値関数 $V(s)$

    1. 初期化: $V(s) = 0$
    2. **repeat**:

       for $s \in S$:

       $V(s) \leftarrow \max_{a \in A(s)} \left\{ r(s,a) + \gamma \sum_{s' \in S} P(s' \mid s,a) V(s') \right\}$

       until $V(s)$ の変化量が一定値以下になる

15.3 複数の状態をもつ問題 (6/9)

- 方策反復法
  - 直接的に方策を繰り返し改善することで，最適な行動価値関数 $Q(s, a)$ を推定する
  - 一般に，安定的に収束する

$$Q^{\pi}(s_t, a_t) = r(s_t, a_t) + \gamma \sum_{s_{t+1} \in S} P(s_{t+1} \mid s_t, a_t) V^{\pi}(s_{t+1})$$

15.3 複数の状態をもつ問題 (7/9)

- 方策反復法のアルゴリズム
  - 入力 : マルコフ決定過程
  - 出力 : 最適方策 π^*

1. **repeat**:

   方策評価を行い，$V^{\pi}(s)$ を計算する

   for $a \in A$:

   for $s \in S$:

   $Q(s,a) \leftarrow r(s,a) + \gamma \sum_{s' \in S} P(s' \mid s,a) V(s')$

   $\pi(s) \leftarrow \arg\max_{a} Q(s,a)$

   until 方策 π が変化しなくなる

2. **return** 最適方策 π^*

15.3 複数の状態をもつ問題 (8/9)

- Q 学習
  - 状態遷移確率や報酬分布が未知の場合に，探索によってQ値を更新していく方法
  - 探索戦略
    - ϵ-greedy法
    - ソフトマックス法（学習が進むにつれて温度 Tを小さくする）

$$P(a \mid s) = \frac{\exp(Q(s,a)/T)}{\sum_{b \in A} \exp(Q(s,b)/T)}$$

- Q値の更新（更新分を時間とともに減らしていく）

$$Q(s,a) \leftarrow Q(s,a) + \eta(r + \gamma \max_{a'} Q(s',a') - Q(s,a))$$

15.3 複数の状態をもつ問題 (9/9)

- Q学習のアルゴリズム
  - 入力 : 行為と報酬の系列
  - 出力 : 最適方策 π^*

1. 初期化: $Q(s,a)$ を0に設定する
2. **for each** エピソード:

   $s \leftarrow s_0$

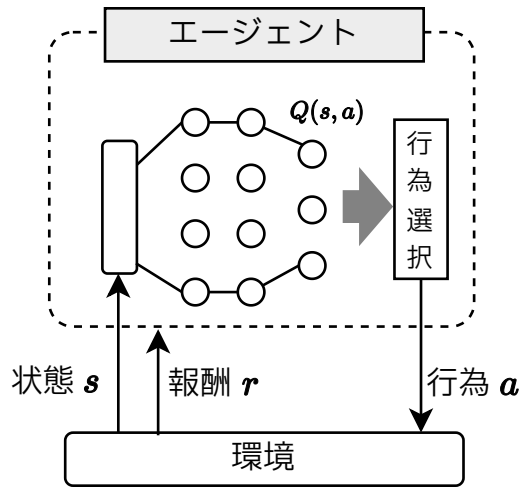
   while s が終了状態でない:

   探索基準に基づき行為 a を選択・実行し，報酬 r と次状態 s' を観測

   $Q(s,a) \leftarrow Q(s,a) + \eta(r + \gamma \max_{a'} Q(s',a') - Q(s,a)), \quad s \leftarrow s'$
3. $\pi^*(s) \leftarrow \arg\max_a Q(s,a)$
4. **return** 最適方策 π^*

15.4 深層強化学習 (1/2)

- 深層学習の強化学習への導入
  - DNN を用いて，強化学習で用いる関数を近似する
  - 近似対象 : 状態価値関数 $V(s)$，Q関数 $Q(s,a)$，方策 $\pi(a|s)$
  - 画像入力の場合など，状態数が非常に多い場合に有効



15.4 深層強化学習 (2/2)

- Q 関数の近似の例
  - Q関数の表現 : DNN f のパラメータ θ, 入力状態 $\boldsymbol{x}$

$$Q(\boldsymbol{x}, a; \theta) = f_\theta(a \mid \boldsymbol{x})$$

  - 教師信号として用いられるターゲット出力（以下，定式化は状態を s とする）

$$y = r + \gamma \max_{a'} Q(s', a'; \theta^-)$$

  - 損失関数：学習対象のネットワークとターゲットネットワークの二乗誤差

$$L(\theta) = \mathbb{E}_{s,a,r,s' \sim \mathcal{D}} \left[\left(y - Q(s, a; \theta) \right)^2 \right] = \mathbb{E}_{(s,a,r,s') \sim \mathcal{D}} \left[\left(r + \gamma \max_{a'} Q(s', a'; \theta^-) - Q(s, a; \theta) \right)^2 \right]$$

15.5 方策勾配法と LLM の学習 (1/2)

- 方策勾配法
  - 方策 $\pi(a \mid s)$ をパラメータ化し，累積報酬の期待値を最大化するようにパラメータ θ を更新する
  - 期待累積報酬

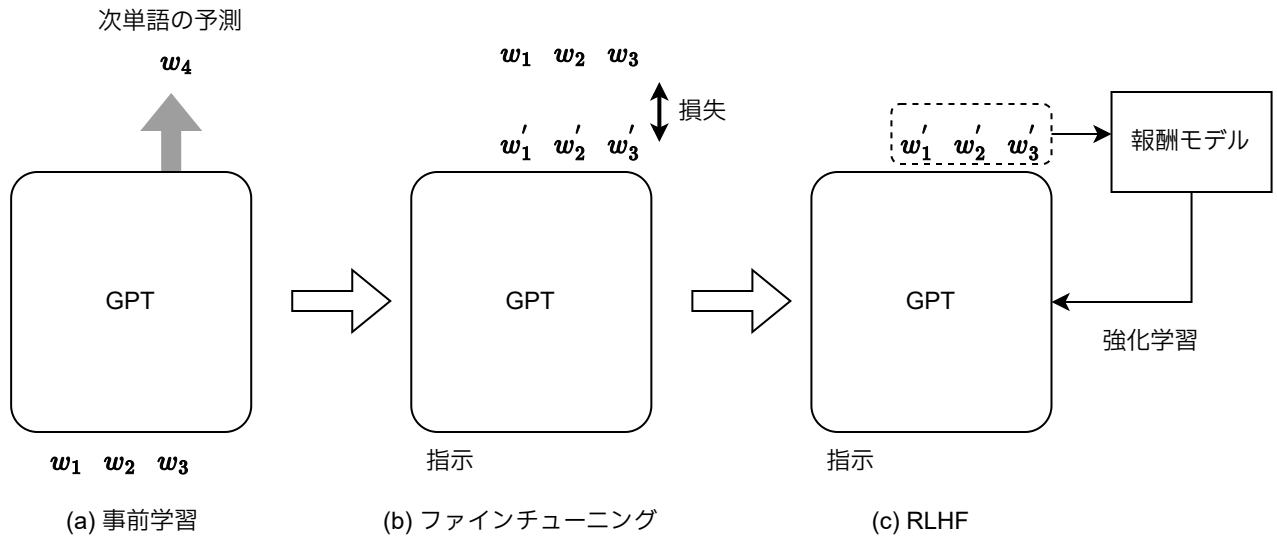
$$J(\theta) = \mathbb{E}^{\pi_\theta}\left[\sum_{t=0}^{T-1} \gamma^t r_t\right]$$

  - 方策の勾配を用いたパラメータ更新

$$\theta \leftarrow \theta + \alpha \nabla_\theta J(\theta)$$

15.5 方策勾配法と LLM の学習 (2/2)

- 方策勾配法を用いた LLM の学習
  - LLM の出力を行為，文脈を状態として，方策勾配法で学習する
  - ヒューマンフィードバックから報酬モデルを学習して報酬とする



まとめ

- 強化学習の問題設定
  - 学習目標は累積期待報酬最大化で記述できるという報酬仮説に基づく
  - 目標は状態から行為を決める関数の獲得だが，正解情報は示されず，遅延した報酬が確率的に得られる
- 最適政策の求め方
  - 価値反復法，方策反復法，Q学習など
- 状態数が多い場合など，深層学習との統合が有効
- 方策勾配法を用いて，LLMの学習が可能