強化学習入門

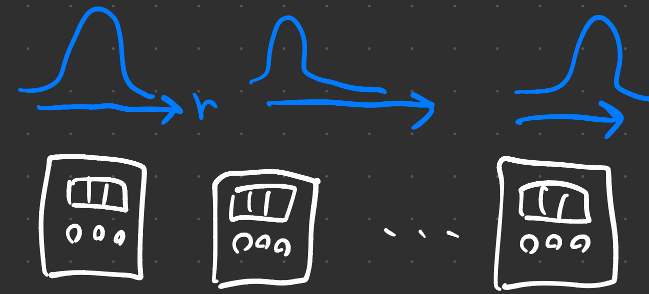
田中 章詞

1. 状態がない場合
2. 状態がある場合

▨1. 状態のない場合

□定義

kコのスロットマシン



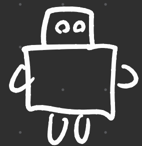
$a = 1 \quad 2 \quad \cdots \quad k$

1. a をえらぶ ← ①
2. ある $r \in \mathbb{R}$ が返される
3. $r > 0 \Rightarrow r$ 円もらえる
   $r < 0 \Rightarrow |r|$ 円しはらう

②.

どの a をえらぶべきか? ← ③

① 行動と方策

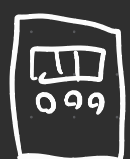


$\longrightarrow a$ ：何らか確率 $\underline{\pi(a)}$ に従う

a ＝ 行動 (action)

$\pi(a)$ ＝ 方策 (policy)

② 報酬と環境



a

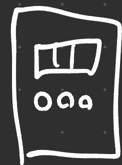
$\longrightarrow r$ ：何らかの確率 $\underline{P(r|a)}$ に従う

条件つき

r ＝ 報酬 (reward)

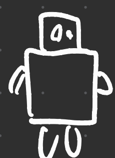
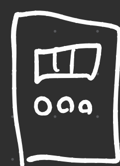
$P(r|a)$ ＝ 環境 (environment)

③ 価値

 $\xrightarrow{p} r$ の期待値 $\mathbb{E}_{r\sim p(\cdot|a)}[r] \overset{\text{def}}{=} \underline{Q(a)}$

a

行動価値
(action value)

 $\xrightarrow{\pi}$  $\xrightarrow{p} r$ の期待値 $\mathbb{E}_{a\sim\pi}[Q(a)] \overset{\text{def}}{=} \underline{J(\pi)}$

a

あえて言うと 方策の価値

$\boxed{\text{目的} \approx \max_{\pi} J(\pi)}$

How?

□学習の方針

$J(\pi)$ の値はいろいろな理由で

アクセスするのが難しい

学習

$$\pi_0 \rightarrow \pi_1 \rightarrow \pi_2 \rightarrow \cdots \rightarrow \pi_*$$

$$\downarrow \quad \downarrow \quad \downarrow \quad \downarrow \quad \downarrow$$

$$J(\pi_0) \le J(\pi_1) \le J(\pi_2) \le \cdots \le J(\pi_*)$$

こうなってほしい

方策改善 (policy improvement)

$\pi \longrightarrow \pi'$

$$J(\pi') - J(\pi) = \mathop{E}_{a \sim \pi'}\Big[\underbrace{Q(a) - J(\pi)}_{\text{"アドバンテージ関数"}\ A_\pi(a)}\Big]$$

$$\boxed{\text{方策改善} \iff \mathop{E}_{a \sim \pi'}\left[A_\pi(a)\right] \geq 0}$$

□ 価値ベース手法

ε-貪欲方策 (ε-greedy policy)

$$\varepsilon \in [0, 1]$$

$$\pi^{\varepsilon}(a) = (1-\varepsilon)\, 1_{a = \underset{\tilde{a}}{\mathrm{argmax}}\, Q(\tilde{a})} + \varepsilon \frac{1}{k}$$

$$\cdots \to \pi^{\varepsilon} \longrightarrow \pi^{(\varepsilon')} \longrightarrow \cdots$$

直感：$\varepsilon \geq \varepsilon'$ ⇒ 証明

$$A_{\pi^{\varepsilon}}(a) = Q(a) - J(\pi)$$

$$= Q(a) - \sum_{\hat{a}} \pi^{\varepsilon}(\hat{a})\, Q(\hat{a})$$

$$\mathbb{E}_{a \sim \pi^{\varepsilon'}} \ \downarrow$$

こうなるには?

$$(1-\varepsilon') \max_{a} Q(a) - \sum_{\hat{a}} \left(\pi^{\varepsilon}(\hat{a}) - \varepsilon' \frac{1}{k} \right) Q(\hat{a}) \geq 0$$

$$\|$$

$$\sum_{\hat{a}} \underbrace{\left(\pi^{\varepsilon}(\hat{a}) - \varepsilon' \frac{1}{k} \right)}_{(1-\varepsilon) 1_{\arg\ldots} + \frac{\varepsilon}{k} - \frac{\varepsilon'}{k} \geq 0} \Bigg(\underbrace{\max_{a} Q(a) - Q(\hat{a})}_{\geq 0} \Bigg)$$

□ 方策ベースの手法

パラメトリックな方策

$\theta \in \mathbb{R}^d$ でパラメータ付される π_θ

<u>例</u>

$(\theta_1, \theta_2, \cdots, \theta_k) \in \mathbb{R}^k$　　$\pi_\theta(a) = \dfrac{e^{\theta_a}}{\sum_{\hat{a}} e^{\theta_{\hat{a}}}}$

(softmax 方策)

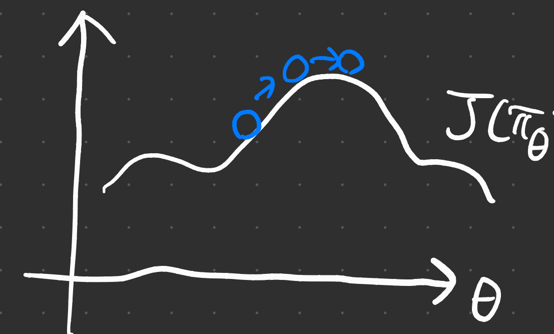
$\cdots \to \pi_\theta \to \pi_{\theta'} \to \cdots$　　方策改善?

まずは $J(\pi_\theta)$ の微分をとってみる

$$\nabla_\theta J(\pi_\theta) = \nabla_\theta \underbrace{\mathbb{E}_{a\sim\pi_\theta}\left[Q(a) \right]}_{\sum_a Q(a)\pi_\theta(a)}$$

$$= \sum_a Q(a) \underbrace{\nabla_\theta \pi_\theta(a)}_{\nabla_\theta \log \pi_\theta(a) \cdot \pi_\theta(a)}$$

$$= \mathbb{E}_{a\sim\pi_\theta}\left[Q(a) \nabla_\theta \log \pi_\theta(a) \right]$$



$$\theta \leftarrow \theta + \eta \nabla_\theta J$$

アドバンテージから再考

こうなってほしい

$$0 \leq \mathop{E}_{\pi_{\theta+\Delta\theta}}\left[A_{\pi_\theta}(a)\right] = J(\pi_{\theta+\Delta\theta}) - J(\pi_\theta)$$

$\Delta\theta$でテイラー展開　$O(\Delta\theta^2)$ムシ

$$\simeq \sum_a \left(\pi_\theta(a) + \Delta\theta \cdot \nabla_\theta \pi_\theta(a)\right) A_{\pi_\theta}(a)$$

$$\mathop{E}_{\pi_\theta}\left[Q(a) - \mathop{E}_{\pi_\theta}[Q(a)]\right] = 0$$

$$= \sum_a \Delta\theta \cdot \underbrace{\nabla_\theta \pi_\theta(a)}_{\nabla_\theta \log \pi_\theta(a) \cdot \pi_\theta(a)} A_{\pi_\theta}(a)$$

$$= \Delta\theta \mathop{E}_{a\sim\pi_\theta}\left[A_{\pi_\theta}(a) \nabla_\theta \log \pi_\theta(a)\right]$$

α

$$\Delta\theta = \eta \cdot E[\cdots]$$

実は

$$\nabla_\theta J(\pi_\theta) = \mathop{E}_{a\sim\pi_\theta}\left[Q(a)\nabla_\theta \log \pi_\theta(a)\right]$$

$$= \mathop{E}_{a\sim\pi_\theta}\left[\underbrace{A_{\pi_\theta}(a)}_{Q(a)-J(\pi_\theta)}\nabla_\theta \log \pi_\theta(a)\right]$$

より一般に

$$= \mathop{E}_{a\sim\pi_\theta}\left[\left(Q(a)-B\right)\nabla_\theta \log \pi_\theta(a)\right]$$

ベースライン

ふつうは $J(\pi_\theta)$ をとる

□実際

- ε-greedy π^{ε} $\leftarrow$ $Q(a) = E[r]$ の値が必要

- parametric π_θ $\leftarrow$ $\nabla_\theta J(\pi_\theta) = E[\cdots]$ の値が必要

$\Downarrow$

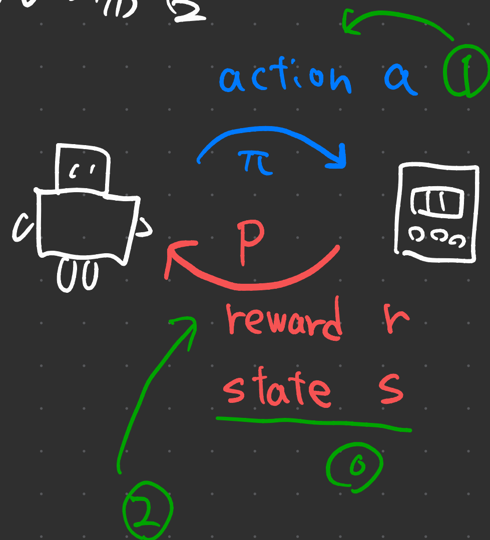
✓ $E[\cdots]$ 知ってる ⇒ OK.

✓ $E[\cdots]$ 知らない ⇒ モンテカルロ

(次回のsection1よてい)

図 2. 状態のある場合

"Agent"



"Environment"

「ほどく」と

$$* \xrightarrow{P_{int}} (s_0 \xrightarrow{\pi} a_0) \xrightarrow{P} (s_1 \xrightarrow{\pi} a_1) \xrightarrow{P} (s_2 \xrightarrow{\pi} a_2) \xrightarrow{P} \cdots$$

$$r_1 \qquad r_2 \qquad \cdots$$

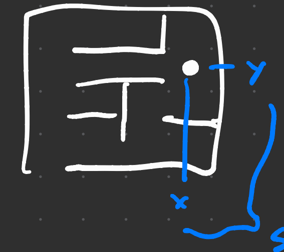
$\{(s_t, a_t, r_{t+1})\}_t$

マルコフ性.

→ マルコフ決定プロセス (MDP)

⓪ 状態 (state)

Agentが置かれている状態を表す変数



① 行動と方策

$s \xrightarrow{\pi} a$: 条件つき確率 $\pi(a|s)$

② 報酬. 状態更新

$(s,a) \to s'$

$\quad\;\;\searrow r'$

: 条件つき確率 $P(s',r'|s,a)$

③ 価値

$$* \xrightarrow{P_{init}} (s \color{blue}{\rightarrow} a) \rightarrow (s_1 \color{blue}{\rightarrow} a_1) \rightarrow (s_2 \color{blue}{\rightarrow} a_2) \rightarrow (s_3 \color{blue}{\rightarrow} a_3) \rightarrow \cdots$$

$$r_1 \qquad r_2 \qquad r_3 \qquad \cdots$$

$$E\left[\; E\left[\; E\left[\, r_1 + \gamma r_2 + \gamma^2 r_3 + \cdots \right] \right] \right]$$

割引率 (discount rate) $\in [0,1]$

$= Q_\pi(s, a)$

$= V_\pi(s)$ 状態価値 (state value)

$= J(\pi)$

$$\boxed{\text{目的：} \max_\pi J(\pi)}$$

□ 学習の方針

$\max_{\pi} J(\pi)$ は難しいので…

学習

$$
\begin{array}{ccccccccc}
\pi_0 & \to & \pi_1 & \to & \pi_2 & \to & \cdots & \to & \pi_* \\
\downarrow & & \downarrow & & \downarrow & & \downarrow & & \downarrow \\
J(\pi_0) & \le & J(\pi_1) & \le & J(\pi_2) & \le & \cdots & \le & J(\pi_*)
\end{array}
$$

方策改善.

$\pi \to \pi'$

$$J(\pi') - J(\pi) = \sum_{t=0}^{\infty} \gamma^t \mathop{E}_{\substack{a_0 \cdots a_t \sim \pi' \\ s_0 \cdots s_t \sim P}} \Big[\underbrace{Q_\pi(s_t, a_t) - V_\pi(s_t)}_{\text{アドバンテージ関数}} \Big]$$

"証明"

$$A_\pi(s_t, a_t)$$

$$\boxed{\text{方策改善} \Longleftrightarrow \sum_{t=0}^{\infty} \gamma^t \mathop{E}_{\substack{a_{0:t} \sim \pi' \\ s_{0:t} \sim P}} \left[A_\pi(s_t, a_t) \right] \geq 0}$$

$J(\pi') - J(\pi) =$

$$(s_0 \xrightarrow{\pi'} a_0) \to (s_1 \xrightarrow{\pi} a_1) \to (s_2 \xrightarrow{\pi} a_2) \to \cdots$$
$$\searrow r_1 \qquad \searrow r_2 \qquad \searrow \cdots$$
$$\boxed{\mathop{E}_{a_0\sim\pi'}\Big[\ \underbrace{\mathop{E}_{\pi}\big[r_1 + \gamma r_2 + \cdots\big]}_{Q_\pi(s_0,a_0)}\ \Big]}$$

$-$

$$(s_0 \xrightarrow{\pi} a_0) \to (s_1 \xrightarrow{\pi} a_1) \to (s_2 \xrightarrow{\pi} a_2) \to \cdots$$
$$\searrow r_1 \qquad \searrow r_2 \qquad \searrow \cdots$$
$$\underbrace{E\big[\ r_1 + \gamma r_2 + \cdots\ \big]}_{\mathop{E}_{a_0\sim\pi'}[V_\pi(s_0)]}$$

$$(s_0 \xrightarrow{\pi'} a_0) \to (s_1 \xrightarrow{\pi'} a_1) \to (s_2 \xrightarrow{\pi} a_2) \to \cdots$$
$$\searrow r_1 \qquad \searrow r_2 \qquad \searrow \cdots$$
$$E\big[\ \cancel{r_1} + \gamma(r_2 + \cdots)\ \big]$$

$$(s_0 \xrightarrow{\pi'} a_0) \to (s_1 \xrightarrow{\pi} a_1) \to (s_2 \xrightarrow{\pi} a_2) \to \cdots$$
$$\searrow r_1 \qquad \searrow r_2 \qquad \searrow \cdots$$
$$E\big[\ \cancel{r_1} + \gamma(r_2 + \cdots)\ \big]$$

$$(s_0 \xrightarrow{\pi'} a_0) \to (s_1 \xrightarrow{\pi'} a_1) \to (s_2 \xrightarrow{\pi'} a_2) \to \cdots$$
$$\searrow r_1 \qquad \searrow r_2 \qquad \searrow \cdots$$
$$E\big[\ r_1 + \gamma r_2 + \cdots\ \big]$$

$$(s_0 \xrightarrow{\pi'} a_0) \to (s_1 \xrightarrow{\pi'} a_1) \to (s_2 \xrightarrow{\pi'} a_2) \to \cdots$$
$$\searrow r_1 \qquad \searrow r_2 \qquad \searrow \cdots$$
$$E\big[\ r_1 + \gamma r_2 + \cdots\ \big]$$

$$\mathop{E}_{a_0\sim\pi'}\big[\,Q_\pi(s_0,a_0) - V_\pi(s_0)\,\big] \qquad \mathop{E}_{a_0 a_1}\big[\,\gamma\,(Q_\pi(s_1,a_1) - V_\pi(s_1))\,\big] \qquad 0$$

□ 価値ベースの手法

ε-貪欲方策

$$\underline{\pi}^{\varepsilon}(a|s) = (1-\varepsilon)\,\mathbb{1}_{a=\arg\max Q_{\underline{\underline{\pi}}}(s,a)} + \varepsilon\frac{1}{k}$$

π　　π

$$\pi^{\varepsilon=1} \to \cdots\cdots \to \pi^{\varepsilon} \to \pi^{\varepsilon'} \to \cdots$$

直感：$\varepsilon : 1 \searrow 0$ がいいのでは？

$(\varepsilon \geq \varepsilon')$

$^\forall s, a, \ ^\exists \varepsilon' > 0 \ \text{s.t.} \ \mathop{E}_{a \sim \pi^{\varepsilon'}}[A_{\pi^\varepsilon}(s, a)] \geq 0$

状態なしのときと同じ

方策改善

$$0 \leq \sum_{t=0}^{\infty} \gamma^t \underbrace{\mathop{E}_{\substack{a_0 \cdots a_t \\ s_0 \cdots s_t}} \left[A_{\pi^\varepsilon}(s_t, a_t) \right]}_{\mathop{E}_{\substack{a_0 \cdots a_{t-1} \\ s_0 \cdots s_{t-1}\, s_t}} \mathop{E}_{a_t}[\cdots]}$$

□方策ベースの手法

パラメトリック方策

$\theta \in \mathbb{R}^{\cdot}$ 毎に定まる $\pi_\theta(a|s)$

例

$$\{(\theta_{s1}, \theta_{s2}, \cdots, \theta_{sk})\}_s \rightarrow \pi_\theta(a|s) = \frac{e^{\theta_{sa}}}{\sum_{\tilde{a}} e^{\theta_{s\hat{a}}}}$$

$$\cdots \rightarrow \pi_\theta \rightarrow \pi_{\theta+\Delta\theta} \rightarrow \cdots$$

$$\sum_{t=0}^{\infty} \gamma^t \underset{\substack{a_0 \cdots a_t \sim \pi_{\theta+\Delta\theta} \\ s_0 \cdots s_t \sim p}}{E} \left[A_{\pi_\theta}(s_t, a_t) \right] = J(\pi_{\theta+\Delta\theta}) - J(\pi_\theta)$$

↓ テイラー展開 ← $\underset{a_t \sim \pi_\theta}{E}\left[A_{\pi_\theta}(s_t, a_t) \right] = 0$

$$= \underbrace{\Delta\theta \sum_{t=0}^{\infty} \gamma^t \underset{\substack{a_0 \cdots a_t \sim \pi_\theta \\ s_0 \cdots s_t \sim p}}{E} \left[A_{\pi_\theta}(s_t, a_t) \nabla_\theta \log \pi_\theta (a_t | s_t) \right]}_{\nabla_\theta J(\pi_\theta)} + O(\Delta\theta^2)$$

$$A_{\pi_\theta}(s_t, a_t) = Q_\pi(s_t, a_t) - \underbrace{V_\pi(s_t)}_{B(s_t)}$$

ベースライン

$$\nabla_\theta J(\pi_\theta)$$

$$\parallel$$

$$\sum \cdots \quad A \to Q$$

□ 実際.

- ε-greedy $\pi^{\varepsilon} \longrightarrow Q_{\pi}(s,a) = E[\cdots]$

  が必要

- param $\pi_{\theta} \longrightarrow \nabla_{\theta} J(\pi_{\theta}) = E[\cdots]$

  が必要

$\Downarrow$

$$E[\cdots] \;\to\; \frac{1}{N}\sum_{i=1}^{N}\cdots \quad \begin{cases} \text{モンテカルロ} \\ \text{Temporal Difference 法.} \end{cases}$$

□ Temporal Difference に向けて.

ベルマン方程式

ポイント： $r_1 + \underbrace{\gamma r_2 + \gamma^2 r_3 + \cdots}_{\gamma(\underbrace{r_2 + \gamma r_3 + \cdots}_{g_2})} = g_1$

$\Downarrow$

$$Q_\pi(s_0, a_0) = \mathop{E}_{\substack{a_1 \sim \pi \\ s_1 r_1 \sim p}}\left[r_1 + \gamma Q_\pi(s_1, a_1)\right]$$

$$V_\pi(s_0) = \mathop{E}_{p}\left[r_1 + \gamma V_\pi(s_1)\right]$$

ベルマン最適方程式

アイデア： $\cdots \to \pi^{\varepsilon} \to \pi^{\varepsilon'} \to \cdots \underbrace{\pi^{\varepsilon=0}}_{\pi^*(a|s) = 1_{a=\underset{\tilde{a}}{\mathrm{argmax}}\, Q_{\pi^*}(s,\tilde{a})}} = 1_{\mathrm{argmax}\, Q}$

ベルマン eq

$$Q_{\pi_*}(s_0, a_0) = \underset{\substack{a_1 \sim \pi_* \\ s_1 r_1 \sim p}}{E}\left[r_1 + \gamma Q_{\pi_*}(s_1, a_1)\right]$$

$$= \underset{s_1 r_1 \sim p}{E}\left[r_1 + \gamma \max_{\tilde{a}} Q_{\pi_*}(s_1, \tilde{a})\right]$$

Made with Goodnotes

$$F(s_0, t_0) = E[\bullet + F(s_1, a_1)]$$

時間が
1に進んだところ

$$\Downarrow$$

$$\frac{1}{N}\sum_{i=1}^{N} \rightarrow \mathrm{TD}.$$

参考文献

- R. Suttun, A. Barto "Reinforcement Learning : An Introduction"
  - └ 訳 : ver1 : 森北出版 (小ぶりなサイズ)
  - : ver2 : (A4くらい?)
- 牧野貴樹さん ほか "これからの強化学習"
- 森村哲郎さん "強化学習"