強化学習　第1章
「準備」

逐次的意思決定問題は，次の三点によって規定される．

- 環境を規定するマルコフ決定過程
- エージェントの行動選択ルールを規定する方策モデル
- 方策について最適化される目的関数

1　マルコフ決定過程

状態，行動に関する確率分布を次のように定義する．

- $p_{S_0}(s) \coloneqq Pr[S_0 = s]$
  (初期状態 S_0 が s である確率)
- $p_T(s' \mid s, a) \coloneqq Pr[S_{t+1} = s' \mid S_t = s, A_t = a] \quad \forall t \in \mathbb{N}_0$
  (t 期の状態と行動が s, a で与えられている時，$t+1$ 期に状態 s' に遷移する確率)
- $\pi(a \mid s) \coloneqq Pr[A = a \mid S = s]$
  (状態が s で与えられた時，行動 a を選択する確率)

上のように分布を定義すると，マルコフ決定過程 $M(\pi) = \{\mathcal{S}, \mathcal{A}, p_{S_0}, p_T, g, \pi\}$ の時間発展 (確率変数を観測する順番) は次のようになる．

1. 時間ステップ t を $t = 0$ と初期化して，初期状態確率 p_{S_0} に従い初期状態 $S_0 \sim p_{S_0}$ を観測する．$(S_0 = s_0)$
2. $t = 0, ..., T-1$ に対し，
   (a) 状態 $S_t = s_t$ が与えられたとし，行動 $A_t \sim \pi(\cdot \mid s_t)$ を選択する．$(A_t = a_t)$
   (b) 行動 a_t を実行し，報酬 $r_t = g(s_t, a_t)$ と次期の状態 $s_{t+1} \sim p_T(\cdot \mid s_t, a_t)$ を観測する．$(S_{t+1} = s_{t+1})$

2　方策モデル

方策は，次のように大別できる．

- マルコフ方策 . . . 過去の経験とは独立に状態 s の情報から行動を決める．
  - 定常 . . . s の条件付き分布．確率論的 or 決定論的
  - 非定常 . . . s_t の条件付き分布．
- 非マルコフ方策 . . . 過去の経験をもとに行動を決める．

2.1　マルコフ方策

過去の経験とは独立に行動を選択することをマルコフ方策と呼ぶ．

- 定常 (時不変) なマルコフ方策集合 Π の定義

$$\Pi \coloneqq \{\pi : \sum_{a \in \mathcal{A}} \pi(a|s) = 1,\ \forall s \in \mathcal{S}\}$$

(確率的方策 $\pi(s|a) = Pr[A = a|S = s]$ の集合．分布 π は t に依存しない．)

  - 決定論的なマルコフ方策集合 Π^d の定義

$$\Pi^d \coloneqq \{\pi^d : S \to \mathcal{A}\}$$

(決定的方策 $\pi(a|s) = \mathbb{I}(\pi^d(s) = a)$ の集合．Π の部分集合．)

- 時間ステップ t の進展に伴い π が変化する一般の (非定常) マルコフ方策系列集合 $\mathbf{\Pi}^M$ の定義

$$\mathbf{\Pi}^M \coloneqq \{\boldsymbol{\pi}^m \coloneqq \begin{pmatrix}\pi_0 \in \Pi & \pi_1 \in \Pi & \ldots\end{pmatrix}\}$$

2.2　非マルコフ方策

時点 t までの全ての確率変数の実現値の履歴を

$$h_t \coloneqq \{s_0, a_0, r_0, \ldots, s_{t-1}, a_{t-1}, r_{t-1}, s_t\} \in \mathcal{H}_t$$

と置くと，履歴依存の方策 $\pi_t^h : \mathcal{A} \times \mathcal{H}_t \to [0, 1]$ とは

$$\pi_t^h(a_t \mid h_t) \coloneqq Pr[A_t = a_t \mid H_t = h_t]$$

と定義できる．この時，非マルコフ方策の集合 Π_t^H とは，

$$\Pi_t^h \coloneqq \{\pi_t^h : \sum_{a \in \mathcal{A}} \pi_t^h(a_t \mid h_t) = 1\}$$

であり．方策 π_t^h の系列 $\boldsymbol{\pi}^h$ とは

$$\boldsymbol{\pi}^h \coloneqq \{\pi_0^h, \pi_1^h, \ldots\} \in \boldsymbol{\Pi}^H \coloneqq (\Pi_t^h)_{t \in \mathbb{N}_0}$$

※方策が非マルコフ方策であるとき，マルコフ決定過程の中に方策は含まれず，$M = \{\mathcal{S}, \mathcal{A}, p_{S_0}, p_T, g\}$ となる．

2.3 方策の特徴

2.2 で定義した，履歴依存の非マルコフ方策系列の集合 $\boldsymbol{\Pi}^H$ から最良な方策系列 $\boldsymbol{\pi}^{h*}$ を求めることができれば，それより良い方策は存在しない．しかし，非マルコフ方策系列では方策サイズが時間ステップ数に対して組合せ爆発を起こしてしまうため，一般に $\boldsymbol{\Pi}^H$ に対する方策の最適化は困難である．実は，マルコフ方策のみを扱っても多くの場合十分な最適解を得ることができる．

命題

任意のマルコフ決定過程 $M = \{\mathcal{S}, \mathcal{A}, p_{S_0}, o_T, g\}$ と履歴依存の方策系列 $\boldsymbol{\pi}^h = \{\pi_0^h, \pi_1^h, \ldots\} \in \boldsymbol{\Pi}^H$ に対し，次を満たすようなマルコフ方策の系列が存在する．

$$Pr[S_t = s, A_t = a \mid M(\boldsymbol{\pi}^h)] = Pr[S_t = s, A_t = a \mid M(\boldsymbol{\pi}^m)], \quad \forall (t, s, a) \in \mathbb{N}_0 \times \mathcal{S} \times \mathcal{A}$$

(任意の非マルコフ方策による状態，行動の時系列同時分布に対し，同様の分布を与えるマルコフ方策が必ず存在する)

3 逐次的意思決定問題

方策の最適化問題のことを逐次的意思決定問題という．逐次的意思決定問題は，その問題設定に応じて次の 2 種類の方法で解かれる．

- プランニング...環境モデルであるマルコフ決定過程 $M = \{\mathcal{S}, \mathcal{A}, p_{S_0}, p_T.g\}$ が既知である，という問題設定．動的計画法や線形計画法によって環境モデルから方策を最適化する．

- 強化学習問題 . . . 環境 M が未知の場合の方策の学習問題．バッチ学習 (オフライン学習) とオンライン学習に大別される．
  - バッチ学習 (オフライン学習). . . 環境との相互作用から得たデータから方策を学習
  - オンライン学習 . . . 逐次的に環境と相互作用してデータを収集.「探索と活用のトレードオフ」といって，新たなデータを発見する目的で行動を選択するのか，データから最良と思われる行動を選択するのか，という二つの意思決定戦略のバランスを取ることが問題となる．

3.1 リターン

リターン (割引累積報酬) なる確率変数を導入する．すなわち，

$$\begin{aligned} C_t &= \lim_{K\to\infty}\sum_{k=0}^{K}\gamma^k R_{t+k} \\ &= R_t + \lim_{K\to\infty}\sum_{k=1}^{K}\gamma^k R_{t+k} \\ &= R_t + \gamma C_{t+1} \end{aligned} \tag{1}$$

ただし，$t+k$ 期の報酬を確率変数 $R_{t+k} = g(s_{t+k}, a_{t+k})$ とおいた．これは $A_{t+k}|S_{t+k} \sim \pi$，$S_{t+k}|A_{t+k-1}, S_{t+k-1} \sim p_T$ に依存する．また (1) より，リターン C_t は再帰構造をもつ．
一般に，逐次的意思決定問題はリターン C_t に関する何かしらの統計量 $f : \Pi \to \mathbb{R}$ を方策 π について最適化する，という構造を取る．つまり，最適方策は

$$\pi^* \coloneqq \arg\max_{\pi\in\Pi}\{f(\pi)\}$$

と書くことができる．

3.2 目的関数 1

時間ステップ $t = 0$ からのリターン C_0 の期待値を目的関数 f_0 と定義する．

$$f_0(\pi) \coloneqq \mathbb{E}_{S_0\sim p_{S_0}, A_t|S_t\sim\pi, S_{t+1}|S_t, A_t\sim p_T}[C_0]$$

f_0 は目的関数の十分性の条件を満たすので，方策集合を定常マルコフ方策のみを想定して良い．次に，価値関数 $V^\pi(s)$ を

$$V^\pi(s) \coloneqq \mathbb{E}_{A_t|S_t\sim\pi, S_{t+1}|S_t, A_t\sim p_T}[C_0 \mid S_0 = s]$$

と定義し，初期状態 S_0 に関して期待値を取ることで，f_0 を次のように p_{S_0} を取り出した形で書きなおすことができる．

$$\begin{aligned} f_0(\pi) &= \mathbb{E}_{S_0 \sim p_{S_0}}[V^\pi(S_0)] \\ &= \sum_{s \in \mathcal{S}} p_{S_0}(s) V^\pi(s) \end{aligned}$$

3.3 目的関数 2

時不変の定常マルコフ方策 $\pi \in \Pi$ について，$f_\infty(\pi)$ を次のように定義する．

$$\begin{aligned} f_\infty(\pi) &\coloneqq \lim_{T\to\infty} \mathbb{E}_{S_0 \sim p_{S_0}, A_t|S_t \sim \pi, S_{t+1}|S_t, A_t \sim p_T}[\frac{1}{T}\sum_{t=0}^{T-1} C_t] \\ &= \lim_{T\to\infty} \mathbb{E}_{S_0 \sim p_{S_0}, S_t|S_{t-1}, A_{t-1} \sim p_T}[\frac{1}{T}\sum_{t=0}^{T-1} \\ &\qquad \mathbb{E}_{A_t|S_t \sim \pi, S_{t+1}|S_t, A_t \sim p_T}[C_t \mid S_t = s_t]] \\ &= \lim_{T\to\infty} \mathbb{E}_{S_0 \sim p_{S_0}, S_t|S_{t-1}, A_{t-1} \sim p_T}[\frac{1}{T}\sum_{t=0}^{T-1} V^\pi(S_t)] \qquad (2) \end{aligned}$$

全ての時点から期待リターンをとって，その平均を取る．
f_∞ は (2) のように，状態に関する期待値として書ける．そこで，マルコフ決定過程 $M(\pi)$ において状態にフォーカスした確率過程をマルコフ連鎖 $MC(\pi) \coloneqq \{\mathcal{S}, p_{S_0}, p^\pi_{MC}\}$ と呼ぶ．ただし，$p^\pi_{MC} : S \times S \to [0,1]$ とは状態 s から s' へと遷移する確率であって，

$$p^\pi_{MC}(s'|s) \coloneqq \sum_{a \in \mathcal{A}} \pi(a|s) p_T(s'|s,a)$$

命題

マルコフ連鎖 $MC(\pi)$ がエルゴート性を満たすならば，期待リターンの時間平均 f_∞ は報酬の時間平均：

$$\lim_{T\to\infty} \mathbb{E}_{S_0 \sim p_{S_0}, A_t|S_t \sim \pi, S_{t+1}|S_t, A_t \sim p_T}[\frac{1}{T}\sum_{t=0}^{T-1} R_t]$$

の定数倍として表せる．
($\Leftrightarrow$ 目的関数 f_∞ を最大化する逐次的意思決定問題は，割引率を導入しない時間平均報酬の最大化と同値)

(証明の準備)
エルゴート性とはマルコフ連鎖の特徴のことで，以下を満たすことをいう．

- 既約的 (全ての状態が行き来可能)

$$Pr[S_t = j \mid S_0 = i, MC(\pi)] > 0,\ \forall i, j \in S, \exists t \in \mathbb{N}$$

- 非周期的 (状態 s から s へと戻ってくる現象が k の倍数回のみで見られる，というときに，k の最大公約数が 1 である．)

$$\mathcal{T}(s) \coloneqq \{t \geq 1 : Pr[S_t = s \mid S_0 = s] > 0\}$$

  の最大公約数が 1 である．$\gcd\mathcal{T}(s) = 1,\ \forall s \in \mathcal{S}$

マルコフ連鎖がエルゴート的である時，状態空間には定常分布 $p^{\pi}_{\infty} : S \to [0, 1]$ が存在する．

$$\begin{aligned} p^{\pi}_{\infty}(s) &= \lim_{T\to\infty} Pr[S_T = s \mid MC(\pi)], \quad \forall s \in \mathcal{S} \\ &= \lim_{T\to\infty} \frac{1}{T} \sum_{t=0}^{T-1} Pr[S_t = s \mid MC(\pi)], \quad \forall s \in \mathcal{S} \end{aligned} \tag{3}$$

式 (3) は定常分布が初期状態確率 p_{S_0} に依存しないことを意味する，

(証明)
状態の関数 $v : S \to \mathbb{R}$ について $v(S_t)$ の期待値とは

$$\mathbb{E}_{S_t \sim p^{\pi}_{\infty}}[v(S_t)] = \sum_{s\in\mathcal{S}} Pr[S_t = s \mid MC(\pi)]v(s) \tag{4}$$

と書けるので，目的関数 $f_{\infty}(\pi)$ は次のように表せる．

$$f_{\infty}(\pi) = \lim_{T\to\infty} \mathbb{E}_{S_0 \sim p_{S_0}, S_t \mid S_{t-1}, A_{t-1} \sim p_T}[\frac{1}{T} \sum_{t=0}^{T-1} V^{\pi}(S_t)] \tag{5}$$

$$\begin{aligned} &= \lim_{T\to\infty} \frac{1}{T} \sum_{t=0}^{T-1} \mathbb{E}_{S_t \sim p^{\pi}_{\infty}}[V^{\pi}(S_t)] \\ &= \lim_{T\to\infty} \frac{1}{T} \sum_{t=0}^{T-1} \sum_{s\in\mathcal{S}} Pr[S_t = s \mid MC(\pi)]V^{\pi}(s) \quad \because (4) \\ &= \sum_{s\in\mathcal{S}} V^{\pi}(s) \lim_{T\to\infty} \{\frac{1}{T} \sum_{t=0}^{T-1} Pr[S_t = s \mid MC(\pi)]\} \\ &= \sum_{s\in\mathcal{S}} p^{\pi}_{\infty}(s)V^{\pi}(s), \quad \forall \pi \in \Pi \end{aligned} \tag{6}$$

(5) = (6) より，価値関数の時間平均 = 空間平均となることがわかる．次に，価値関数の定義式より，

$$\begin{aligned} V^{\pi}(s) &= \mathbb{E}_{A_t|S_t\sim\pi,S_{t+1}|S_t,A_t\sim p_T}[C_0 \mid S_0 = s] \\ &= \mathbb{E}_{A_t|S_t\sim\pi,S_{t+1}|S_t,A_t\sim p_T}[R_0 + \gamma C_1 \mid S_0 = s] \quad \because (1) \\ &= \sum_{a\in\mathcal{A}} \pi(a|s)g(s,a) + \gamma \mathbb{E}_{A_t|S_t\sim\pi,S_{t+1}|S_t,A_t\sim p_T}[C_1|S_0 = s] \quad (7) \end{aligned}$$

右辺第二項について，1 時点目まで期待値を展開して，

$$\begin{aligned} & \mathbb{E}_{A_t|S_t\sim\pi,S_{t+1}|S_t,A_t\sim p_T}[C_1|S_0 = s] \\ = & \sum_{a\in\mathcal{A}} \pi(a|s) \sum_{s'\in\mathcal{S}} p_T(s'|s,a) \mathbb{E}_{A_t|S_t\sim\pi,S_{t+1}|S_t,A_t\sim p_T}[C_1|S_1 = s'] \quad (8) \end{aligned}$$

(7),(8) と価値関数の定義式により，ベルマン方程式を得る．

$$V^{\pi}(s) = \sum_{a\in\mathcal{A}} \pi(a|s)\{g(s,a) + \gamma \sum_{s'\in\mathcal{S}} p_T(s'|s,a)V^{\pi}(s')\}, \quad \forall s \in \mathcal{S} \qquad (9)$$

(6) に (9) を代入して，

$$\begin{aligned} f_{\infty}(\pi) &= \sum_{s\in\mathcal{S}}\sum_{a\in\mathcal{A}} p^{\pi}_{\infty}(s)\pi(a|s)g(s,a) + \gamma \sum_{s\in\mathcal{S}}\sum_{s'\in\mathcal{S}} p^{\pi}_{\infty}(s)p^{\pi}_{MC}(s'|s)V^{\pi}(s') \\ &= \lim_{T\to\infty} \mathbb{E}_{S_0\sim p_{S_0},S_t|S_{t-1},A_{t-1}\sim p_T}[\frac{1}{T}\sum_{t=0}^{T-1} R_t] + \gamma \sum_{s'\in\mathcal{S}} p^{\pi}_{\infty}(s')V^{\pi}(s') \\ &= \lim_{T\to\infty} \mathbb{E}_{S_0\sim p_{S_0},S_t|S_{t-1},A_{t-1}\sim p_T}[\frac{1}{T}\sum_{t=0}^{T-1} R_t] + \gamma f_{\infty}(\pi), \quad \forall \pi \in \Pi \end{aligned}$$

以上より，

$$f_{\infty}(\pi) = \frac{1}{1-\gamma} \lim_{T\to\infty} \mathbb{E}_{S_0\sim p_{S_0},S_t|S_{t-1},A_{t-1}\sim p_T}[\frac{1}{T}\sum_{t=0}^{T-1} R_t]$$

が証明された．よって，エルゴート性のもと，目的関数を f_{∞} とする逐次的意思決定問題は平均報酬の最大化問題と一致し，最適方策 $\pi^* = \arg\max_{\pi}\{f_{\infty}(\pi)\}$ は割引率 γ の設定に依存せず平均報酬を最大化する．

※ベルマン方程式を導出するところで定常分布を使っていないことに注意．マルコフ連鎖がエルゴート的でなくともベルマン方程式は用いることができる．