強化学習　第4章
「モデルフリー型の強化学習」

- モデルフリー (環境非同定) 型 . . . 環境を陽に推定せずに，方策を学習するアプローチ．動的計画法を基礎に，確率的に観測されるデータから学習する．
  - 価値反復法から派生 . . .Q 学習
  - 方策反復法から派生 . . .SARSA 法，アクター・クリティック法
- モデルベース (環境同定) 型 . . . 環境を明示的に推定するアプローチ．データから状態遷移確率 p_T と報酬関数 g を推定し，推定した環境モデルに対して価値反復法などを適用して方策を学習する．

1　データにもとづく意思決定

単一の意思決定系列とは，時点 T までを実行した時の状態と行動，報酬の記録であり，

$$h_T \coloneqq \{s_0, a_0, r_0, \ldots, s_{T-1}, a_{T-1}, r_{T-1}, s_T\}$$

なる集合である．h_T が複数ある場合が複数の意思決定系列である．意思決定系列の最小構成とは,「ある状態 s において行動 a を選択したところ報酬 r を得て状態 s' に遷移した」という実現値の 4 つ組 $\{s, a, r, s'\}$ のことで，これを「経験」という．経験が N 個というのは，時点 1 までを何度も実行した時の意思決定系列の集合であり，

$$\{h_1^{(1)}, \ldots, h_1^{(N)}\} = \{(s_0^{(1)}, a_0^{(1)}, r_0^{(1)}, s_1^{(1)}), \ldots, (s_0^{(N)}, a_0^{(N)}, r_0^{(N)}, s_1^{(N)})\}$$

となる．今後，V^π, V^* を求めるためにベルマン作用素 B を意思決定系列サンプルで近似することになるが，いずれも標本平均が十分統計量となるため，意思決定系列が単一か複数かは考えなくて良く，エピソードで切って何度もサンプリングを行う，といったこともして良い．ただし，単一の意思決定系列だとマルコフ連鎖がエルゴート性でない場合，吸収状態に陥ってサンプリ

ングが機能しなくなり，一方複数の意思決定系列において，系列の構成を小さくしすぎると,「本当はあまり頻繁に訪れることのない状態」も平等に訪れがちになり，サンプリングが非効率になってしまう．エピソードの分割を適切に設定することが重要である．

2 価値関数の推定

p_T や g が未知とし，方策 π を固定した時の

$$\begin{aligned} V^{\pi}(s) &= \mathbb{E}_{A_t|S_t\sim\pi, S_{t+1}|S_t,A_t\sim p_T}[C_0|S_0 = s] \\ &= \mathbb{E}_{A_t|S_t\sim\pi, S_{t+1}|S_t,A_t\sim p_T}[\lim_{T\to\infty}\sum_{t=0}^{T}\gamma^t R_t|S_0 = s] \end{aligned}$$

を推定することを考える．ここで，状態数や行動数は既知であるとし，$\hat{V}$: $\mathcal{S} \to \mathbb{R}$ を履歴データ $\{s_0, a_0, r_0, \ldots, s_T, a_T, r_T\}$ から各 $s \in \mathcal{S}$ について求める．(ルックアップテーブル関数)

- モンテカルロ推定 . . . 確率的に生成されたサンプル集合の平均を期待値の近似値とする最もナイーブな方法．t 期を始点とした時のリターンの実現値を

  $$c_t \coloneqq \sum_{k=t}^{T}\gamma^{k-t} r_k$$

  と置くと，$\hat{V}$ は

  $$\hat{V}(s) \coloneqq \frac{1}{\sum_{t=0}^{T'}\mathbb{I}(s = s_t)}\sum_{t=0}^{T'}\mathbb{I}(s = s_t)c_t, \quad \forall s \in \{s \in \mathcal{S} : \sum_{t=0}^{T'}\mathbb{I}(s = s_t) > 0\}$$

  リターンを正確に計算するには，ハイパーパラメータ $T' \leq T$ を十分小さくする必要があり，利用可能な標本数は限られるため，一般に推定の効率は良くない．

- ベルマン作用素にもとづくアプローチ
  - ベルマン作用素の標本近似 (バッチ学習, オンライン学習 (TD 学習))

2.1 ベルマン作用素の標本近似

状態関数 v にベルマン作用素 B_π を繰り返し適用することで価値関数 V^π を求めることができる．

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

B_π 内の g や p_T は未知なので，履歴データ

$$h_t^\pi \coloneqq \{s_0, a_0, r_0, \ldots, s_{t-1}, a_{t-1}, r_{t-1}, s_t | M(\pi)\} \in \mathcal{H}_t$$

から B_π の標本近似を行う．(方策 π によって時点 t まで収集されたデータを h_t^π と書く)
さて，ベルマン作用素 B_π は

$$(B_\pi v)(s) = \mathbb{E}_{A_t|S_t\sim\pi, S_{t+1}|S_t,A_t\sim p_T}[R_t + \gamma v(S_{t+1})|S_t = s], \quad \forall s \in \mathcal{S}$$

と期待値で書けるので，近似ベルマン作用素 $\hat{B}$ を次のように定義する．

$$\hat{B}(v; h_T^\pi)(s) \coloneqq \begin{cases} \frac{1}{\sum_{t=0}^{T-1}\mathbb{I}(s_t=s)} \sum_{t=0}^{T-1} \mathbb{I}(s_t = s)\ (r_t + \gamma v(s_{t+1})), & (\text{if } \sum_{t=0}^{T-1}\mathbb{I}(s_t=s)) \\ v(s) & \text{otherwise} \end{cases}$$

$\hat{B}$ がモデルベース型強化学習で計算される近似ベルマン作用素と同値であることを示す．B_π を展開すると，

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

となる．ここで，状態について周辺化した報酬関数と状態遷移関数を

$$\begin{aligned} \bar{g}(s) &= \sum_{a\in\mathcal{A}} \pi(a|s) g(s,a) \\ \bar{p}_T(s'|s) &= \sum_{a\in\mathcal{A}} \pi(a|s) p_T(s'|s,a) \end{aligned}$$

とおくと，その最尤推定量は

$$\hat{g}(s; h_T^\pi) = \begin{cases} \frac{1}{\sum_{t=0}^{T-1}\mathbb{I}(s_t=s)} \sum_{t=0}^{T-1} \mathbb{I}(s_t = s) r_t, & (\text{if } \sum_{t=0}^{T-1}\mathbb{I}(s_t=s) > 0) \\ 0 & \text{otherwise} \end{cases}$$

$$\hat{p}_T(s'|s; h_T^\pi) = \begin{cases} \frac{1}{\sum_{t=0}^{T-1}\mathbb{I}(s_t=s)} \sum_{t=0}^{T-1} \mathbb{I}(s_t = s)\mathbb{I}(s_{t+1} = s'), & (\text{if } \sum_{t=0}^{T-1}\mathbb{I}(s_t=s) > 0) \\ \mathbb{I}(s' = s) & \text{otherwise} \end{cases}$$

である．ゆえ，モデルベースによる近似ベルマン作用素は，

$$\begin{aligned}
\hat{B}(v;h_T^\pi)(s) &= \hat{g}(s;h_T^\pi)+\gamma\sum_{s'\in\mathcal{S}}\hat{p}_T(s'|s;h_T^\pi)v(s'),\quad \forall s\in\mathcal{S}\\
&= \frac{1}{\sum_{t=0}^{T-1}\mathbb{I}(s_t=s)}\sum_{t=0}^{T-1}\mathbb{I}(s_t=s)r_t\\
&\quad +\gamma\sum_{s'\in\mathcal{S}}\frac{1}{\sum_{t=0}^{T-1}\mathbb{I}(s_t=s)}\sum_{t=0}^{T-1}\mathbb{I}(s_t=s)\mathbb{I}(s_{t+1}=s')v(s')\\
&= \frac{1}{\sum_{t=0}^{T-1}\mathbb{I}(s_t=s)}\sum_{t=0}^{T-1}\{\mathbb{I}(s_t=s)r_t+\gamma\sum_{s'\in\mathcal{S}}\mathbb{I}(s_{t+1}=s')v(s')\}\\
&= \frac{1}{\sum_{t=0}^{T-1}\mathbb{I}(s_t=s)}\sum_{t=0}^{T-1}\mathbb{I}(s_t=s)(r_t+\gamma v(s_{t+1}))
\end{aligned}$$

と，モデルフリーによる近似ベルマン作用素と一致することがわかった．

近似ベルマン作用素の収束性

マルコフ決定過程がエルゴート性を満たすと仮定する．この時，$T\to\infty$ において，近似ベルマン作用素 $\hat{B}$ はベルマン作用素 B に収束する．

$$\lim_{T\to\infty}\hat{B}(v;h_T^\pi)(s)=(B_\pi v)(s)$$

(証明)
マルコフ過程がエルゴート性を満たす時，初期状態に依存せず各状態が非ゼロな定常分布が存在する．

$$\begin{aligned}
p_\infty^\pi(s) &= \lim_{T\to\infty}Pr[S_t=s|Mc(\pi)],\ \forall s\in\mathcal{S}\\
&= \lim_{T\to\infty}\frac{1}{T}\sum_{t=0}^{T-1}Pr[S_t=s|MC(\pi)],\ \forall s\in\mathcal{S}\\
&= \lim_{T\to\infty}\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{I}(s_t=s)>0,\qquad \forall s_0,s\in\mathcal{S}
\end{aligned}$$

ゆえ，$\hat{B}$ の極限を取ると，

$$\begin{aligned}
\lim_{T\to\infty} \hat{B}(v; h_T^\pi)(s) &= \lim_{T\to\infty} \frac{1}{\sum_{t=0}^{T-1} \mathbb{I}(s_t = s)} \sum_{t=0}^{T-1} \{\mathbb{I}(s_t = s) r_t + \gamma v(s_{t+1})\} \\
&= \lim_{T\to\infty} \frac{\frac{1}{T}\sum_{t=0}^{T-1} \mathbb{I}(s_t = s)(r_t + \gamma v(s_{t+1}))}{\frac{1}{T}\sum_{t=0}^{T-1} \mathbb{I}(s_t = s)} \\
&= \frac{p_\infty^\pi(s) \mathbb{E}_{A_t|S_t\sim\pi, S_{t+1}|S_t, A_t\sim p_T}[R_t + \gamma v(S_{t+1})|S_t = s]}{p_\infty^\pi(s)}
\end{aligned}$$

2 行目→ 3 行目が謎.

2.2 バッチ学習の場合

履歴データ h_T^π が得られているならば，2.1 の近似ベルマン作用素を状態関数に繰り返し適用することで価値関数を求める．また，方策反復法のように連立方程式を解くことによっても価値関数は求まる．

2.3 オンライン学習の場合

2.3.1 TD(0) 法

現時間ステップ t の観測 $\{s_t, r_t, s_{t+1}\}$ のみを用いて $\hat{V}(s_t)$ を微小に更新することを考える．バッチ学習の場合は訪れたことのある任意の s について，

$$\hat{B}(\hat{V}; h_t^\pi)(s) = \frac{1}{\sum_{t=0}^{T-1} \mathbb{I}(s_t = s)} \sum_{t=0}^{T-1} \mathbb{I}(s_t = s)(r_t + \gamma \hat{V}(s_{t+1}))$$

と報酬の標本平均を取ることで価値関数を更新していた．オンライン学習の場合は，t 期に観測した実現値 s_t の価値関数のみを更新する．すなわち，α_t を学習率とすると，

$$\hat{V}(s_t) \quad := \quad (1 - \alpha_t)\hat{V}(s_t) + \alpha_t \hat{B}(\hat{V}; \{s_t, r_t, s_{t+1}\})(s_t) \tag{1}$$

サンプル数 1 の標本平均を α_t 分だけ加えている．(1) について，近似ベルマン作用素を展開して，

$$\begin{aligned}
(1) &= (1 - \alpha_t)\hat{V}(s_t) + \alpha_t\{r_t + \gamma \hat{V}(s_{t+1})\} \\
&= \hat{V}(s_t) + \alpha_t\{r_t + \gamma \hat{V}(s_{t+1}) - \hat{V}(s_t)\} \\
&= \hat{V}(s_t) + \alpha_t \delta_t
\end{aligned}$$

ここで,

$$\delta_t = r_t + \gamma \hat{V}(s_{t+1}) - \hat{V}(s_t)$$

とは，$t+1$ 期までの情報を利用した s_t の価値の推定値 $r_t + \gamma \hat{V}(s_{t+1})$ と t 時点での s_t の価値の推定値 $\hat{V}(s_t)$ の差分である．これを時間的差分誤差 (TD 誤差) と呼ぶ．TD 法のアルゴリズムは以下の通り．

1. 推定価値関数 $\hat{V} : \mathcal{S} \to \mathbb{R}$ を任意に初期化し，初期状態 s_0 を観測する．
2. 各 s_t において,
   (a) 方策 $\pi(a|s_t)$ に従い行動 a_t を選択し，a_t を環境に入力する．
   (b) 環境から報酬 r_t と次状態 s_{t+1} を観測する．
   (c) 得られた $\{s_t, r_t, s_{t+1}\}$ から TD 誤差を計算する．

$$\delta \coloneqq r_t + \gamma \hat{V}(s_{t+1}) - \hat{V}(s_t)$$

   (d) TD 誤差をもとに推定価値関数 $\hat{V}(s_t)$ を更新する．

$$\hat{V}(s_t) \coloneqq \hat{V}(s_t) + \alpha_t \delta$$

3. 終了条件 (最大時間ステップ数など) を満たしているならば終了．

2.3.2 TD(λ) 法

TD 誤差 δ_t とは，$t+1$ 時点での s_t の価値の推定値 $r_t + \gamma \hat{V}(s_{t+1})$ と，t 時点での s_t の価値の推定値 $\hat{V}(s_t)$ の差分であった．ここでは，$t+1$ 時点ではなく $t+n$ 時点での s_t の価値の推定値：

$$c_t^{(n)} \coloneqq r_t + \gamma r_{t+1} + \ldots + \gamma^n \hat{V}(s_{t+n})$$

を用いることを考える．この時，TD 誤差 $\delta_t^{(n)}$ とは,

$$\delta_t^{(n)} \coloneqq \{r_t + \gamma r_{t+1} + \ldots + \gamma^n \hat{V}(s_{t+n})\} - \hat{V}(s_t)$$

となる．さらに，$n \in \{1, \ldots\}$ とし，各 $t+n$ 時点での s_t の価値を推定し，その平均を取ることを考える．つまり，重み係数を $\lambda \in [0, 1]$ とすれば.

$$c_{t,\lambda} \coloneqq \begin{cases} (1-\lambda) \sum_{n=1}^{\infty} \lambda^{n-1} c_t^{(n)}, & (\lambda \in [0, 1]) \\ c_t^{(\infty)}, & (\lambda = 1) \end{cases}$$

となり，TDλ 誤差 $\delta_{t,\lambda}$ は

$$\delta_{t,\lambda} = c_{t,\lambda} - \hat{V}(s_t)$$

とすることができる (前方観測的 TD 誤差)．しかし，$\delta_{t,\lambda}$ による $\hat{V}(s_t)$ の更新則では $t+n$ 時点までの情報が必要となってしまうため，各時点で $\delta_{t,\lambda}$ を計算するのはオンライン学習に適さない．そこで，$\delta_{t,\lambda}$ を時間的に分解して，確定している部分のみを用いて価値関数を更新する後方観測的アプローチが用いられる．$\alpha_t = 0$ または α_t が十分に小ならば,

$$
\begin{aligned}
\delta_{t,\lambda} &= c_{t,\lambda} - \hat{V}(s_t) \\
&= (1-\lambda)\sum_{n=1}^{\infty}\lambda^{n-1}c_t^{(n)} - \hat{V}(s_t) \\
&= (1-\lambda)\sum_{n=1}^{\infty}\lambda^{n-1}\{r_t + \gamma r_{t+1} + ... + \gamma^n \hat{V}(s_{t+n})\} - \hat{V}(s_t) \\
&= (1-\lambda)\{\lambda^0(r_t + \gamma\hat{V}(s_{t+1})) \\
&\qquad +\lambda^1(r_t + \gamma r_{t+1} + \gamma^2\hat{V}(s_{t+2})) \\
&\qquad +\lambda^2(r_t + \gamma r_{t+1} + \gamma^2 r_{t+2} + \gamma^3\hat{V}(s_{t+3})) \\
&\qquad +\ldots\} - \hat{V}(s_t) \\
&= (1-\lambda)(\sum_{n=1}^{\infty}\lambda^{n-1}r_t + \gamma\sum_{n=2}^{\infty}\lambda^{n-1}r_{t+1} + \gamma^2\sum_{n=3}^{\infty}\lambda^{n-1}r_{t+2} + \ldots) \\
&\quad +(1-\lambda)(\gamma\hat{V}(s_{t+1}) + \lambda\gamma^2\hat{V}(s_{t+2}) + \lambda^2\gamma^3\hat{V}(s_{t+3}) + \ldots) - \hat{V}(s_t) \\
&= r_t + \lambda\gamma r_{t+1} + \lambda^2\gamma^2 r_{t+2} + \ldots \qquad (\because (1-\lambda)(r_t + \lambda r_t + \lambda^2 r_t + \ldots) = r_t) \\
&\quad +(1-\lambda)(\gamma\hat{V}(s_{t+1}) + \lambda\gamma^2\hat{V}(s_{t+2}) + \lambda^2\gamma^3\hat{V}(s_{t+3}) + \ldots) - \hat{V}(s_t) \\
&= (r_t + \gamma\hat{V}(s_{t+1}) - \hat{V}(s_t)) \\
&\quad +\lambda\gamma(r_{t+1} + \gamma\hat{V}(s_{t+2}) - \hat{V}(s_{t+1})) \\
&\quad +\lambda^2\gamma^2(r_{t+2} + \gamma\hat{V}(s_{t+3}) - \hat{V}(s_{t+2})) \\
&\quad +\ldots \\
&= +\delta_t + \lambda\gamma\delta_{t+1} + \lambda^2\gamma^2\delta_{t+2} + \ldots \\
&= \sum_{n=0}^{\infty}(\lambda\gamma)^n\delta_{t+n}
\end{aligned}
$$

より，$\delta_{t,\lambda}$ は t 以上の任意の時間ステップ T に対して

$$
\delta_{t,\lambda} = \sum_{\tau=t}^{T}(\lambda\gamma)^{\tau-t}\delta_\tau + \sum_{\tau=T+1}^{\infty}(\lambda\gamma)^{\tau-t}\delta_\tau
$$

と時間分解できる．時間ステップ T 時点で右辺第一項は計算できるが，第二項は計算できない．
次に，時間ステップ T までにある状態 s に訪問した時間ステップの集合を $\{t_1, \ldots, t_n\}$ と置くと，時間ステップ T までの TD(λ) 誤差の和とは，

$$\begin{aligned}
\Delta_T(s) &\coloneqq \delta_{t_1,\lambda} + \ldots + \delta_{t_n,\lambda} \\
&= \sum_{\tau=t_1}^{T} (\lambda\gamma)^{\tau-t_1}\delta_\tau + \sum_{\tau=T+1}^{\infty} (\lambda\gamma)^{\tau-t_1}\delta_\tau \\
&\quad + \ldots \\
&\quad + \sum_{\tau=t_n}^{T} (\lambda\gamma)^{\tau-t_n}\delta_\tau + \sum_{\tau=T+1}^{\infty} (\lambda\gamma)^{\tau-t_n}\delta_\tau \\
&= \sum_{\tau=t_1}^{T} (\lambda\gamma)^{\tau-t_1}\delta_\tau + \ldots + \sum_{\tau=t_n}^{T} (\lambda\gamma)^{\tau-t_n}\delta_\tau \\
&\quad + \sum_{\tau=T+1}^{\infty} (\lambda\gamma)^{\tau-t_1}\delta_\tau + \ldots + \sum_{\tau=T+1}^{\infty} (\lambda\gamma)^{\tau-t_n}\delta_\tau \\
&= \Delta_T^{\text{past}}(s) + \Delta_T^{\text{future}}(s)
\end{aligned}$$

で書ける．ただし，

$$\begin{aligned}
\Delta_T^{\text{past}}(s) &= \sum_{\tau=t_1}^{T} (\lambda\gamma)^{\tau-t_1}\delta_\tau + \ldots + \sum_{\tau=t_n}^{T} (\lambda\gamma)^{\tau-t_n}\delta_\tau \\
\Delta_T^{\text{future}}(s) &= \sum_{\tau=T+1}^{\infty} (\lambda\gamma)^{\tau-t_1}\delta_\tau + \ldots + \sum_{\tau=T}^{\infty} (\lambda\gamma)^{\tau-t_n}\delta_\tau
\end{aligned}$$

とおいた．さて，$\Delta_T^{\text{past}}(s)$ について，

$$\sum_{\tau=t_1}^{T} (\lambda\gamma)^{\tau-t_1}\delta_\tau + \ldots + \sum_{\tau=t_n}^{T} (\lambda\gamma)^{\tau-t_n}\delta_\tau = \sum_{t=0}^{T}\delta_t \sum_{\tau=0}^{t} \mathbb{I}_{t-\tau\in\{t_1\ldots t_n\}}(\lambda\gamma)^\tau$$

であることを示す．これは，t_1 のみについて考えれば，

$$\sum_{\tau=t_1}^{T} (\lambda\gamma)^{\tau-t_1}\delta_\tau = \sum_{t=0}^{T}\delta_t \sum_{\tau=0}^{t} \mathbb{I}_{t-\tau=t_1}(\lambda\gamma)^\tau$$

が成り立てば良い．$t-\tau = t_1 \Leftrightarrow \tau = t - t_1$ より，右辺は

$$\sum_{t=0}^{T}\delta_t \sum_{\tau=0}^{t} \mathbb{I}_{t-\tau=t_1}(\lambda\gamma)^\tau = \sum_{t=t_1}^{T}\delta_t(\lambda\gamma)^{t-t_1}$$

となるため，上式が成り立つことが示された．従って，

$$\begin{aligned}\Delta_T^{\text{past}}(s) &= \sum_{t=0}^{T}\delta_t\sum_{\tau=0}^{t}\mathbb{I}_{t-\tau=t_1}(\lambda\gamma)^{\tau}\\ &= \sum_{t=0}^{T}\delta_t\sum_{\tau=0}^{t}\mathbb{I}_{s_{t-\tau}=s}(\lambda\gamma)^{\tau}\\ &= \sum_{t=0}^{T}\delta_t\sum_{\tau=0}^{t}\mathbb{I}_{s_{\tau}=s}(\lambda\gamma)^{t-\tau} \quad \forall s\in\mathcal{S}\end{aligned}$$

最終行は $\tau = t-\tau$ なる平行移動を行なった．さて，

$$\delta_{t,\lambda}^{\text{back}}(s) = \delta_t\sum_{\tau=0}^{t}\mathbb{I}_{s_\tau=s}(\lambda\gamma)^{t-\tau} = \delta_t z_{t,\lambda}(s)$$

と置く．$z_{t,\lambda}$ をエリジビリティ・トレースという．これを用いて，Δ_T^{past} は

$$\Delta_T^{\text{past}}(s) = \sum_{t=0}^{T}\delta_{t,\lambda}^{\text{back}}(s)$$

である．$\delta_{t,\lambda}^{\text{back}}(s)$ を後方観測的な TDλ 誤差という．以上より，前方観測的な TDλ アプローチの近似として，後方観測的な TDλ アプローチが導出される．

$$\hat{V}(s) \coloneqq \hat{V}(s) + \alpha_t\delta_{t,\lambda}^{\text{back}}(s),\ \ \forall s\in\mathcal{S}$$

true online TDλ アルゴリズムは以下の通り．

1. 推定価値関数 $\hat{V}:\mathrm{S}\to\mathbb{R}$ とエリジビリティ・トレース $z:\mathcal{S}\to\mathbb{R}$ を任意に初期化し，初期状態 s_0 を観測する．
2. s_t において，
   (a) 方策 $\pi(a|s_t)$ に従い行動 a_t を選択し，a_t を環境に入力する．
   (b) 環境から報酬 r_t と次状態 s_{t+1} を観測する．
   (c) 全ての $s\in\mathcal{S}$ に対し，エリジビリティ・トレースと TD 誤差を求め，$\hat{V}(s)$ を更新する．

$$\begin{aligned}z(s) &\coloneqq \mathbb{I}_{s=s_t}+\gamma\lambda z(s), \quad \forall s\in\mathcal{S}\\ \delta &\coloneqq r_t+\gamma\hat{V}(s_{t+1})-\hat{V}(s_t)\\ \hat{V}(s) &\coloneqq \hat{V}(s)+\alpha_t\delta z(s), \quad \forall s\in\mathcal{S}\end{aligned}$$

3. 終了要件 (最大時間ステップなど) を満たすならば終了.

TDλ 法まとめ

1. 前方観測的な TDλ 誤差 $\delta_{t,\lambda}$ をそのまま使おうとすると効率が悪い.
2. (準備) $\delta_{t,\lambda}$ は時間ステップ T において計算できる部分とできない部分に分離できる.
3. 各状態 s について，時間ステップ T まで前方観測的な TDλ 誤差 $\delta_{t,\lambda}$ で $\hat{V}(s)$ を更新してきたとする．任意の状態 s について，TDλ 誤差の和 $\Delta_T(s)$ は 2.(準備) より

$$\Delta_T(s) = \Delta_T^{\text{past}}(s) + \Delta_T^{\text{future}}(s)$$

と，計算できる部分とできない部分に分離できる.

4. 3. の Δ_T^{past} は

$$\Delta_T^{\text{past}}(s) = \sum_{t=0}^{T} \delta_{t,\lambda}^{\text{back}}(s)$$

という，各時間 t で計算できる $\delta_{t,\lambda}^{\text{back}}$ の和として書ける.

5. 各時間ステップ t において，$\delta_{t,\lambda}^{\text{back}}$ を更新していけば，時間ステップ T までで更新量の和は $\Delta_T^{\text{past}}(s)$ である．これは，$T \to \infty$ で,

$$\Delta_T(s) \approx \Delta_T^{\text{past}}$$

である．このように，後方観測的 TDλ 誤差 $\delta_{t,\lambda}^{\text{back}}$ を用いることで，前方観測的な TDλ 誤差 $\delta_{t,\lambda}$ を近似できる．(計算できない未来の情報を使わないことで少し更新量は落ちるが T が大ならばその影響が些少される)

3 方策と行動価値関数の学習

3.1 ベルマン行動作用素と最適行動価値関数

＜概観＞

1. V^π と同様に V^* を推定したい．$\Rightarrow B_*$ を推定できれば良い.

2. B_* は B_π と異なり期待値の外側に max 演算子があるため直接標本近似することができない.

3. $V^*(s) = \max_{a\in\mathcal{A}} Q^*(s,a)$ と置くことで，Q^* を推定する問題に置き換える．$\Rightarrow \Upsilon_*$ を推定できれば良い.

4. Υ_* は期待値なので標本近似が可能！

V^* の推定を考える．価値反復法の第 n 繰り返し目の推定価値関数を $\hat{V}_n$ と書けば，ベルマン最適作用素 B_* は

$$\begin{aligned}\hat{V}_{n+1}(s) = (B_*\hat{V}_n)(s) &= \max_{a\in\mathcal{A}}\{g(s,a) + \gamma\sum_{s'\in\mathcal{S}} p_T(s'|s,a)\hat{V}_n(s')\}, \quad \forall s\in\mathcal{S} \\ &= \max_{a\in\mathcal{A}} \mathbb{E}_{S_{t+1}|S_t,A_t\sim p_T}[g(S_t,A_t) + \gamma\hat{V}_n(S_{t+1})|S_t = s, A_t = a]\end{aligned}$$

と期待値の形で表せる．しかし，B_π と異なり，B_* は期待値の外側に max 演算子を持つため標本近似ができない．そこで，上式の max の内側を s,a を引数に取る推定行動価値関数として

$$\hat{Q}_n(s,a) \coloneqq \mathbb{E}_{S_{t+1}|S_t,A_t\sim p_T}[g(S_t,A_t) + \gamma\hat{V}_n(S_{t+1})|S_t = s, A_t = a], \quad \forall (s,a)\in\mathcal{S}\times\mathcal{A}$$

と定義すると，$\hat{V}_{n+1}$ は

$$\hat{V}_{n+1}(s) = \max_{a\in\mathcal{A}}\hat{Q}_n(s,a), \quad \forall s\in\mathcal{S}$$

なので，これを $\hat{Q}_{n+1}(s,a)$ の式に代入することで,

$$\hat{Q}_{n+1}(s,a) \coloneqq \mathbb{E}_{S_{t+1}|S_t,A_t\sim p_T}[g(S_t,A_t) + \gamma\max_{a'\in\mathcal{A}}\hat{Q}_n(S_{t+1},a')|S_t = s, A_t = a], \quad \forall (s,a)\in\mathcal{S}\times\mathcal{A}$$

を得る．以上より，価値反復法によって $\hat{V}_n$ は n が大ならば V^* に収束する，という命題を用いて

$$\begin{aligned}V^*(s) &\approx \hat{V}_{n+1}(s), \qquad \forall s\in\mathcal{S}, \quad \text{as } n\to\infty \\ &= \max_{a\in\mathcal{A}}\hat{Q}_n(s,a), \quad \forall s\in\mathcal{S}\end{aligned}$$

であることがわかった．言い換えれば価値反復法の要領で

$$Q^*(s,a) \approx \hat{Q}_n(s,a), \quad \forall s\in\mathcal{S}, \text{ as } n\to\infty$$

なる Q^* を求めることで V^* が推測できる，ということである．すなわち，ベルマン行動作用素 Υ_* を任意の $q \in \mathcal{R}^{\mathcal{S}\times\mathcal{A}}$ に対して

$$\Upsilon_* q(s,a) \coloneqq \mathbb{E}_{S_{t+1}|S_t,A_t\sim p_T}[g(S_t,A_t) + \gamma \max_{a'\in\mathcal{A}} q(S_{t+1},a')|S_t = s, A_t = a]$$

と定義し，行動価値関数：

$$Q^\pi(s,a) = \mathbb{E}_{S_{t+1}|S_t,A_t\sim p_T}[g(S_t,A_t) + \gamma V^\pi(S_{t+1})|S_t = s, A_t = a]$$

に対し最適行動価値関数を

$$Q^*(s,a) \coloneqq \max_{\pi\in\Pi} Q^\pi(s,a), \quad \forall (s,a) \in \mathcal{S}\times\mathcal{A}$$

とおくと，以下で示すとおり任意の $\hat{Q}_0 : \mathcal{S}\times\mathcal{A} \to \mathbb{R}$ に対しベルマン行動作用素 υ_* を繰り返し適用することで $\hat{Q}_0$ は Q^* に収束し，$V^*(s) = \max_{a\in\mathcal{A}} Q^*(s,a)$ が成り立つ．

Υ_* による動的計画法の収束性

任意の $\hat{Q}_0 : \mathcal{S}\times\mathcal{A} \to \mathbb{R}$ に対しベルマン行動作用素 Υ_* を繰り返し適用することで $\hat{Q}_0$ は Q^* に収束する．

$$\lim_{n\to\infty} (\Upsilon_*^n \hat{Q}_0)(s,a) = Q^*(s,a), \quad \forall (s,a) \in \mathcal{S}\times\mathcal{A}$$

(証明の方針)
$Q^*, \hat{Q}_n$ をそれぞれ $V^*, \hat{V}_n$ で書くと，動的計画法の収束性により $V^* = \lim_{n\to\infty} \hat{V}_n$ なので命題が成り立つ．確率変数列 $\hat{Q}_n, \hat{V}_n$ はなんか非負っぽいので期待値と極限は任意に交換して良い．

(証明)
$Q^*(s,a)$ の式を変形して，

$$\begin{aligned} Q^*(s,a) &= \max_{\pi\in\Pi} Q^\pi(s,a) \\ &= \max_{\pi\in\Pi} \mathbb{E}_{S_{t+1}|S_t,A_t\sim p_T}[g(S_t,A_t) + \gamma V^\pi(S_{t+1})|S_t = s, A_t = a] \\ &= \mathbb{E}_{S_{t+1}|S_t,A_t\sim p_T}[g(S_t,A_t) + \gamma V^*(S_{t+1})|S_t = s, A_t = a], \quad \forall (s,a) \in \mathcal{S}\times\mathcal{A} \end{aligned}$$

となる．また，

$$\begin{aligned}
\lim_{n\to\infty}(\gamma_*^n\hat{Q}_0)(s,a) &= \lim_{n\to\infty}\hat{Q}_n(s,a) \\
&= \lim_{n\to\infty}\mathbb{E}_{S_{t+1}|S_t,A_t\sim p_T}[g(S_t,A_t)+\gamma\hat{V}_n(S_{t+1})|S_t=s,A_t=a] \\
&= \mathbb{E}_{S_{t+1}|S_t,A_t\sim p_T}[g(S_t,A_t)+\gamma\lim_{n\to\infty}\hat{V}_n(S_{t+1})|S_t=s,A_t=a] \\
&= \mathbb{E}_{S_{t+1}|S_t,A_t\sim p_T}[g(S_t,A_t)+\gamma V^*(S_{t+1})|S_t=s,A_t=a], \quad \forall(s,a)\in\mathcal{S}\times\mathcal{A}
\end{aligned}$$

以上より，命題の等式は成立.

3.2 ベルマン行動作用素の標本近似

履歴データ $h_T=\{s_0,a_0,r_0,\ldots,s_{T-1},a_{T-1},r_{T-1},s_T\}$ から Υ_*：

$$\Upsilon_* q(s,a)=\mathbb{E}_{S_{t+1}|S_t,A_t\sim p_T}[g(S_t,A_t)+\gamma\max_{a'\in\mathcal{A}}q(S_{t+1},a')|S_t=s,A_t=a]$$

を標本近似した近似ベルマン行動最適作用素 $\hat{\Upsilon}_*$ を

$$\hat{\Upsilon}_*(q;h_T)(s,a):=\begin{cases}\frac{1}{\sum_{t=0}^{T-1}\mathbb{I}(s_t=s)\mathbb{I}(a_t=a)}\sum_{t=0}^{T-1}\mathbb{I}(s_t=s)\mathbb{I}(a_t=a)\{r_t+\gamma\max_{a'\in\mathcal{A}}q(s_{t+1},a_{t+1})\}, \\ \qquad\qquad \text{if } \sum_{t=0}^{T-1}\mathbb{I}(s_t=s)\mathbb{I}(a_t=a)>0 \\ q(s,a), \qquad\qquad \text{otherwise}\end{cases}$$

と定義する．今，履歴データ h_T を収集する際に用いる方策を行動方策，$\hat{Q}$ から最終的に計算される方策を目的方策と呼ぶことにすると，行動方策 $\boldsymbol{\pi}$ が

$$\lim_{T\to\infty}\frac{1}{T}\sum_{t=1}^{T}Pr[S_t=s,A_t=a|M(\boldsymbol{\pi})]>0, \quad \forall(s,a)\in\mathcal{S}\times\mathcal{A}$$

を満たすならば，近似ベルマン行動最適作用素 $\hat{\Upsilon}_*$ は $T\to\infty$ で真のベルマン行動最適作用素 Υ_* に収束し，$\hat{\Upsilon}_*$ による動的計画法から求まる目的方策は行動方策に依存しない.

- 方策オフ型の学習 ... 行動方策に非依存の $\hat{\Upsilon}_*$ を用いて目的方策を求める学習手法.
- 方策オン型の学習 ... 行動方策に依存する $\hat{B}_\pi$ や $\hat{\Upsilon}$ を用いて目的方策を求める学習手法.

3.2.1 バッチ学習の場合

関数 $\hat{Q}: \mathcal{S} \times \mathcal{A} \to \mathbb{R}$ を適当に初期化して，$\hat{Q}$ に対しベルマン行動最適作用素を繰り返し作用させ，

$$\hat{Q}(s,a) \coloneqq \hat{\Upsilon}_*(\hat{Q}; h_T)(s,a), \quad \forall (s,a) \in \mathcal{S} \times \mathcal{A}$$

として $\hat{Q}$ を繰り返し更新する事で，最適行動価値関数 Q^* を推定できる．この時，最適方策は定常な決定的方策として

$$\hat{\pi}^{d*}(s) \coloneqq \arg\max_{a \in \mathcal{A}} \hat{Q}(s,a)$$

として得ることができる．

3.3 オンライン学習の場合

現時間ステップ t の観測 $\{s_t, a_t, r_t, s_{t+1}\}$ のみを用いて $\hat{Q}(s_t, a_t)$ を微小に更新することを考える．バッチ学習の際は経験したことのある任意の s, a の組について，

$$\hat{\Upsilon}(\hat{Q}; h_T)(s,a) = \frac{1}{\sum_{t=0}^{T-1} \mathbb{I}(s_t = s)\mathbb{I}(a_t = a)} \sum_{t=0}^{T-1} \mathbb{I}(s_t = s)\mathbb{I}(a_t = a)\{r_t + \gamma \max_{a' \in \mathcal{A}} \hat{Q}(s_{t+1}, a')\}$$

と報酬の標本平均を取ることで行動価値関数を更新していた．オンライン学習の場合は，t 期に経験した実現値 s_t, a_t の行動価値関数のみを更新する．すなわち，α_t を学習率とすると，

$$\hat{Q}(s_t, a_t) \coloneqq (1 - \alpha_t)\hat{Q}(s_t, a_t) + \alpha_t \hat{\Upsilon}_*(\hat{Q}; \{s_t, a_t, r_t, s_{t+1}\})(s_t, a_t)$$

これについて近似行動最適作用素を展開して

$$\begin{aligned}
\hat{Q}(s_t, a_t) &= (1 - \alpha_t)\hat{Q}(s_t, a_t) + \alpha_t\{r_t + \gamma \max_{a' \in \mathcal{A}} \hat{Q}(s_{t+1}, a')\} \\
&= \hat{Q}(s_t, a_t) + \alpha_t\{r_t + \gamma \max_{a' \in \mathcal{A}} \hat{Q}(s_{t+1}, a') - \hat{Q}(s_t, a_t)\} \\
&= \hat{Q}(s_t, a_t) + \alpha_t \delta_t^{(q)}
\end{aligned}$$

ただし，$\delta_t^{(q)}$ を Q 学習法の TD 誤差として

$$\delta_t^{(q)} \coloneqq r_t + \gamma \max_{a' \in \mathcal{A}} \hat{Q}(s_{T+1}, a') - \hat{Q}(s_t, a_t)$$

とおいた．Q 学習法のアルゴリズムは以下の通り．

1. 推定値 $\hat{Q}: \mathcal{S} \times \mathcal{A} \to \mathbb{R}$ を任意に初期化し，初期状態 s_0 から行動方策 $\boldsymbol{\pi}$ に従い観測を開始する．

2. 任意の t において，$\{s_t, a_t, r_t, s_{t+1}\}$ を得た時，TD 誤差 $\delta_t^{(q)}$ を

$$\delta_t^{(q)} \coloneqq r_t + \gamma \max_{a' \in \mathcal{A}} \hat{Q}(s_{t+1}, a') - \hat{Q}(s_t, a_t)$$

と計算し，$\hat{Q}$ を

$$\hat{Q}(s_t, a_t) \coloneqq \hat{Q}(s_t, a_t) + \alpha_t \delta_t^{(q)}$$

と更新する．

3. 終了条件を満たすならば最適方策：

$$\hat{\pi}^{d*}(s) \coloneqq \arg\max_{a \in \mathcal{A}} \hat{Q}(s, a), \quad \forall s \in \mathcal{S}$$

を求め終了．

3.4 アクタークリティック法

行動価値関数：

$$\begin{aligned} Q^\pi(s,a) &= \mathbb{E}_{S_{t+1}|S_t,A_t \sim p_T}[g(S_t, A_t) + \gamma V^\pi(S_{t+1}) | S_t = s, A_t = a] \\ &= g(s,a) + \gamma \sum_{s' \in \mathcal{S}} p_T(s'|s,a) V^\pi(s') \\ &= g(s,a) + \gamma \sum_{s' \in \mathcal{S}} \sum_{a' \in \mathcal{A}} p_T(s'|s,a) \pi(a'|s') Q^\pi(s', a') \end{aligned}$$

を考える．任意の関数 $q: \mathcal{S} \times \mathcal{A} \to \mathbb{R}$ に対し，ベルマン期待作用素 Υ_π を

$$\begin{aligned} \Upsilon_\pi q(s,a) &\coloneqq g(s,a) + \gamma \sum_{s' \in \mathcal{S}} \sum_{a' \in \mathcal{A}} p_T(s'|s,a) \pi(a'|s) q(s', a'), \quad \forall (s,a) \in \mathcal{S} \times \mathcal{A} \\ &= \mathbb{E}_{A_t|S_t \sim \pi, S_{t+1}|S_t,A_t \sim p_T}[g(S_t, A_t) + \gamma q(S_{t+1}, A_{t+1}) | S_t = s, A_t = a] \end{aligned}$$

と定義すれば，次のベルマン行動期待作方程式が成り立つ．

$$Q^\pi(s,a) = \Upsilon_\pi Q^\pi(s,a), \quad \forall (s,a) \in \mathcal{S} \times \mathcal{A}$$

また Υ_π についても動的計画法は収束し，任意の $\hat{Q}^\pi$ に対して Υ_π を繰り返し作用させることで，Q^π に近似できる．
また，Υ_π は期待値なので，次のように標本近似できる．

$$\hat{\Upsilon}(q; h_T)(s,a) = \begin{cases} \frac{1}{\sum_{t=0}^{T-1} \mathbb{I}(s_t = s)\mathbb{I}(a_t = a)} \sum_{t=0}^{T-1} \mathbb{I}(s_t = s)\mathbb{I}(a_t = a)(r_t + \gamma q(s_{t+1}, a_{t+1})) \\ \qquad\qquad\qquad\qquad (\text{if} \quad \sum_{t=0}^{T-1} \mathbb{I}(s_t = s)\mathbb{I}(a_t = a) > 0) \\ q(s,a), \qquad\qquad (\text{otherwise}) \end{cases}$$

$\hat{\Upsilon}$ は行動方策 $\boldsymbol{\pi}$ が

$$\lim_{T\to\infty}\frac{1}{T}\sum_{t=1}^{T} Pr[S_t = s, A_t = a|M(\boldsymbol{\pi})] > 0, \quad \forall(s,a)\in\mathcal{S}\times\mathcal{A}$$

を満たすならば Υ_π に収束する．

3.4.1 SARSA 法

推定価値 $\hat{Q}$ に依存する行動方策 $\pi_t(a|s;\hat{Q})$ を用いて，$\hat{Q}$ を次のように更新する．

$$\begin{aligned}\hat{Q}(s_t,a_t) &\coloneqq (1-\alpha_t)\hat{Q}(s_t,a_t) + \alpha_t\hat{\Upsilon}(\hat{Q};\{s_t,a_t,r_t,s_{t+1},a_{t+1}\})(s_t,a_t)\\ &= \hat{Q}(s_t,a_t) + \alpha_t\delta_t^{(\mathrm{sarsa})}\end{aligned}$$

行動方策が推定価値に依存していることで，SARSA 法は毎回「方策評価」と「方策改善」を繰り返していることになる．

3.4.2 アクタークリティック法

- クリティック ... V^π の推定値 $\hat{V}$ を用いて行動方策 π を評価する．現時点ステップ t の観測 $\{s_t, r_t, s_{t+1}\}$ を用いて，推定価値 $\hat{V}:\mathcal{S}\to\mathbb{R}$ を

$$\hat{V}(s_t) \coloneqq \hat{V}(s_t) + \alpha_t^{(\mathrm{critic})}\delta_t, \qquad (\delta_t = r_t + \hat{V}(s_{t+1}) - \hat{V}(s_t))$$

- アクター ... 効用関数 q に依存する方策 $\pi(a|s;q)$ を用いて方策改善する．クリティックから TD 誤差 δ_t を信号として受けた時，$q(s_t,a_t)$ は

$$q(s_t,a_t) \coloneqq q(s_t,a_t) + \alpha_t^{(\mathrm{actor})}\delta_t$$

アドバンテージ関数 $A^\pi:\mathcal{S}\times\mathcal{A}\to\mathbb{R}$ を

$$A^\pi(s,a) \coloneqq Q^\pi(s,a) - V^\pi(s), \quad \forall(s,a)\in\mathcal{S}\times\mathcal{A}$$

と定義する．クリティックが $\hat{V}(s)$ を計算したとすると，行動価値関数は

$$\hat{Q}(s,a) \coloneqq \mathbb{E}_{S_{t+1}|S_t,A_t\sim p_T}[R_t + \gamma\hat{V}(S_{t+1})|S_t = s, A_t = a], \quad \forall(s,a)\in\mathcal{S}\times\mathcal{A}$$

と書ける (値は求まらないけど)．さらにアドバンテージ関数を

$$\hat{A}(s,a) \coloneqq \hat{Q}(s,a) - \hat{V}(s)$$

と書ける (値はもとまらないけど)．すると，δ_t とは，

$$\begin{aligned}\delta_t &= r_t + \hat{V}(s_{t+1}) - \hat{V}(s) \\ &\approx (\hat{Q}(s_t, a_t) - \hat{V}(s)) + (r_t + \gamma\hat{V}(s_{t+1}) - \hat{Q}(s_t, a_t)) = \hat{A}(s_t, a_t) + X_t\end{aligned}$$

ただし，ノイズ項 X_t を

$$X_t \coloneqq R_t + \gamma\hat{V}(S_{t+1}) - \hat{Q}(s_t, a_t)$$

とおいた．$\hat{Q}$ の定義より明らかに $\mathbb{E}_{S_{t+1},S_t,A_t\sim p_T}[R_t + \gamma\hat{V}(S_{t+1})|S_t = s, A_t = a] = 0$ なので，アクターの方策改善はアドバンテージ関数が 0 より大きい行動を取る確率を上方修正するようなものとなっている．

4 収束性

確率的近似とは，確率的勾配法の確率過程版であり，次のように定義される．

確率的近似

状態関数 $v : \mathcal{S} \to \mathbb{R}$ の更新則を次のように定める．

$$v_{t+1}(s) \coloneqq (1 - \alpha_t(s))v_t(s) + \alpha_t(s)\{B_t v_t(s) + X_t(s) + Y_t(s)\}$$

ただし，$\alpha_t \in \mathbb{R}^{\mathcal{S}}_{\geq 0}$ を学習率，$B_t : \mathbb{R}^{\mathcal{S}} \to \mathbb{R}^{\mathcal{S}}$ を状態関数の作用素，$X_t, Y_t \in \mathbb{R}^{\mathcal{S}}$ をノイズ (確率変数) と置いた．

ここで．確率変数が全て以前の確率変数に依存せず iid であれば，確率的近似は確率的勾配法に対応する．次に，確率的近似の収束性を示す．

確率的近似の収束性

確率的近似が次を満たすとする．

1. 学習率 $\alpha_t(s)$ は全ての $s \in \mathcal{S}$ でロビンス・モンローの条件を満たす．

$$\sum_{t=0}^{\infty} \alpha_t(s) = \infty, \ \ \sum_{t=0}^{\infty} \alpha_t^2(s) < \infty, \ \forall s \in \mathcal{S}$$

   $\alpha_t = \frac{1}{1+t}$ など．(t 軸で積分してみよう！)

2. 作用素 B_t は任意の $t \in \mathbb{N}_0$ で同一の不動点 v^* をもつ縮小写像である．つまり，次を満たす $\tau \in [0, 1)$ が存在する．

$$||B_t v_t - v^*||_\infty \leq \tau ||v_t - v^*||_\infty, \quad \forall t \in \mathbb{N}_0$$

   (作用させると v^* との距離が必ず縮む．)

3. ノイズ X_t の期待値はゼロである．

$$\mathbb{E}[X_t(s)] = 0, \quad \forall t \in \mathbb{N}_0, \ s \in \mathcal{S}$$

   また与えられた任意のノルムに対し

$$\mathbb{E}[X_t^2(s)] \leq c + d||v_t||^2, \ \ \forall t \in \mathbb{N}_0, \ s \in \mathcal{S}$$

   を満たす $c, d \in \mathbb{R}_{\geq 0}$ が存在する．(有限分散)

4. ノイズ Y_t に対し，次式を満たすような 0 に収束する系列 $\{\beta_t \in \mathbb{R}_{\geq 0}\}$ が存在する．

$$|Y_t(s)| \leq \beta_t(||v_t||_\infty + 1), \quad \forall t \in \mathbb{N}_0, \ s \in \mathcal{S}$$

この時，v_t は v^* に収束する．