強化学習　第2章
「プランニング」

1　準備

1.1　目的関数

状態 s の期待リターン (価値関数) を目的関数として設定する.

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

目的関数 V^{π} を最大にする方策を最適方策という．最適方策の必要十分条件 (後述) より,「マルコフ決定過程において環境 p_T が観測可能である時」，最適方策は初期条件 s に依存せず定常な決定的方策として一意に求めることができる (環境が観測可能でなければこの定常な最適方策は存在していても求めることができない)．また，価値関数 V^{π} には次の再帰式 (ベルマン方程式) が成り立つ.

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

(π は定常マルコフ方策．非定常なマルコフ方策系列 $\boldsymbol{\pi}$ では，左辺と右辺の価値関数 $V^{\boldsymbol{\pi}}$ が異なってしまい方程式が成り立たない．命題 2.4b でも「定常方策」と言及している．)

1.2　最適価値関数

最適価値関数を次のように設定する.

$$V^{*}(s) \coloneqq \max_{\pi\in\Pi} V^{\pi}(s)$$

マルコフ方策の十分性により，Π はマルコフ方策集合を仮定して良い (実は定常マルコフ方策集合で良い．後述．)．最適価値関数 V^{*} には V^{π} と同様に次の再帰式 (ベルマン方程式) が成り立つ.

$$V^{*}(s) = \max_{a\in\mathcal{A}}\{g(s,a) + \gamma \sum_{s'\in\mathcal{S}} p_T(s'|s,a)V^{*}(s')\}$$

2　ベルマン作用素

ベルマン作用素 B_π, B_* を定義する．すなわち，方策 π が与えられた時，状態関数 $v : \mathcal{S} \to \mathbb{R}$ を別の状態関数に変換する作用素 $B_\pi : \mathbb{R}^{\mathcal{S}} \to \mathbb{R}^{\mathcal{S}}$ を

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

とし，同様に $B_* : \mathbb{R}^{\mathcal{S}} \to \mathbb{R}^{\mathcal{S}}$ を

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

と定義する．
ベルマン作用素を方策系列 $\boldsymbol{\pi}$ に従い k 回適用する場合，$B_\pi^k v$ と書く．例えば，2 時点までの有限時間のマルコフ決定過程で方策系列 $\boldsymbol{\pi} = \{\pi_1, \pi_2\}$ に基づき行動選択することを考える．この時，ベルマン期待作用素 B_π^2 の関数 v への適用は，

$$\begin{aligned} B_\pi^2 v(s) &= (B_{\pi_0}(B_{\pi_1} v))(s) \\ &= \sum_{a\in\mathcal{A}} \pi_0(a|s)\{g(s,a) + \gamma \sum_{s'\in\mathcal{S}} p_T(s'|s,a) \\ &\quad \times \sum_{a'\in\mathcal{A}} \pi_1(a'|s')\{g(s',a') + \gamma \sum_{s''\in\mathcal{S}} p_T(s''|s',a')v(s'')\}\} \end{aligned}$$

ベルマン作用素と V^π, V^* には次のような関係がある．

$$\begin{aligned} V^\pi(s) &= (B_\pi V^\pi)(s), \quad \forall s \in \mathcal{S} \\ V^*(s) &= (B_* V^*)(s), \quad \forall s \in \mathcal{S} \end{aligned}$$

つまり，V^π はベルマン作用素における不動点であり，この方程式はベルマン方程式に一致する．

＜お気持ち＞
ベルマン作用素を上記のように定義することにより，任意の状態方程式にベルマン作用素を繰り返し適用することで最適価値関数

$$V^*(s) = \max_{a\in\mathcal{A}} \{g(s,a) + \gamma \sum_{s'\in\mathcal{S}} p_T(s'|s,a)V^*(s')\}$$

を求めることができる．さらに，任意の s について V^* を達成するような決定論的な定常方策 π^* がある条件を満たすならば存在し，

$$\pi^*(s) = \arg\max_{a\in\mathcal{A}} \{g(s,a) + \gamma \sum_{s'\in\mathcal{S}} p_T(s'|s,a)V^*(s')\}$$

であることを示すことができる．(価値反復法の導出)

2.1 ベルマン作用素の特徴

単調性

二つの状態関数 v, v' が

$$v(s) \leq v'(s), \quad \forall s \in \mathcal{S}$$

を満たす時，以下が成り立つ．

a. ベルマン最適作用素 B_* について，

$$(B_*^k v)(s) \quad \leq \quad (B_*^k v')(s), \quad \forall s \in \mathcal{S},\ k \in \mathbb{N}_0$$

b. 任意のマルコフ方策系列 $\boldsymbol{\pi} \coloneqq \{\pi_0, ...\} \in \Pi_k^M$ のベルマン期待作用素の積 $B_{\pi_0} B_{\pi_1}, \ldots B_{\pi_{k-1}} \coloneqq B_{\boldsymbol{\pi}}^k$ について，

$$(B_{\boldsymbol{\pi}}^k v)(s) \leq (B_{\boldsymbol{\pi}}^k v')(s), \quad \forall s \in \mathcal{S},\ k \in \mathbb{N}_0$$

(ベルマン作用素は状態関数の大小関係を保存する)

(証明)
帰納法によって証明する．

(i)$k = 0$ の時，$(B_*^0 v)(s) = v(s),\ s \in \mathcal{S}$ より補題は自明に成立．

(ii)$k = n$ の時，任意の $v \leq v'$ なる v, v' に対し，

$$(B_*^n v)(s) \leq (B_*^n v')(s), \quad \forall s \in \mathcal{S}$$

が成り立つと仮定する．$p_T \geq 0$ より，正の線形変換は大小関係を保存するので，

$$\sum_{s \in \mathcal{S}} p_T(s'|s,a)(B_*^n v)(s') \leq \sum_{s \in \mathcal{S}} p_T(s'|s,a)(B_*^n v')(s'), \quad \forall (s,a) \in \mathcal{S} \times \mathcal{A}$$

である．従って，任意の $s \in \mathcal{S}$ について，

$$\begin{aligned}(B_*^{n+1} v)(s) &= \max_{a \in \mathcal{A}} \{g(s,a) + \gamma \sum_{s' \in \mathcal{S}} p_T(s'|s,a)(B_*^n v)(s')\} \\ &\leq \max_{a \in \mathcal{A}} \{g(s,a) + \gamma \sum_{s' \in \mathcal{S}} p_T(s'|s,a)(B_*^n v')(s')\} = (B_*^{n+1} v')(s)\end{aligned}$$

が成り立つ．

(i),(ii) より，帰納的に補題は証明された．

線形分離可能

状態関数の和を次のように定義する．$v' : \mathcal{S} \to \mathbb{R}$ を状態関数，b を実数定数と置くと，

$$\begin{aligned}(v+v')(s) &\coloneqq v(s)+v'(s), \quad \forall s \in \mathcal{S} \\ (v+b)(s) &\coloneqq v(s)+b, \quad \forall s \in \mathcal{S}\end{aligned}$$

任意の状態関数 v と定数 b に対して

$$\begin{aligned}(B_*^k(v+b))(s) &= (B_*^k v)(s)+\gamma^k b, \ \forall s \in \mathcal{S}, \ \forall k \in \mathbb{N}_0 \\ (B_{\boldsymbol{\pi}}^k(v+b))(s) &= (B_{\boldsymbol{\pi}}^k v)(s)+\gamma^k b, \quad \forall \boldsymbol{\pi} \in \Pi_k^M, \ \forall s \in \mathcal{S}, \ \forall k \in \mathbb{N}_0\end{aligned}$$

(証明)
和の定義に従ってベルマン作用素を計算する．

$$\begin{aligned}(B_*(v+b))(s) &= \max_{a\in\mathcal{A}}\{g(s,a)+\gamma\sum_{s'\in\mathcal{S}} p_T(s'|s,a)(v(s')+b)\} \\ &= \max_{a\in\mathcal{A}}\{g(s,a)+\gamma\sum_{s'\in\mathcal{S}} p_T(s'|s,a)v(s')+\gamma b\sum_{s\in\mathcal{S}} p_T(s'|s,a)\} \\ &= \max_{a\in\mathcal{A}}\{g(s,a)+\gamma\sum_{s'\in\mathcal{S}} p_T(s'|s,a)v(s')\}+\gamma b\end{aligned}$$

より，

$$(B_*(v+b))(s) = (B_*v)(s)+\gamma b$$

である．この計算を繰り返すことで補題が得られる．

動的計画法の収束性

a. 任意の有界の状態関数 $v : \mathcal{S} \to \mathbb{R}$ に対して，ベルマン作用素 B_* を繰り返し適用した関数 $(B_*^k v)$ は最適価値関数 V^* に漸近的に等しくなる．

$$V^*(s) = \lim_{k\to\infty} (B_*^k v)(s), \quad \forall s \in \mathcal{S}$$

b. 任意の有界の状態関数 $v : \mathcal{S} \to \mathbb{R}$ に対し，マルコフ方策系列 $\tilde{\boldsymbol{\pi}} \coloneqq \{\pi_0, \ldots, \pi_{k-1}\} \in \Pi_k^M$ のベルマン期待作用素 $(B_{\tilde{\boldsymbol{\pi}}}^k v)$ は $\boldsymbol{\pi} = \{\tilde{\boldsymbol{\pi}}, \pi_k, \pi_{k+1}, \ldots\} \in \Pi^M$ の価値関数 $V^{\boldsymbol{\pi}}$ に漸近的に等しくなる．

$$V^{\boldsymbol{\pi}}(s) = \lim_{k\to\infty} (B_{\boldsymbol{\pi}}^k v)(s), \quad \forall s \in \mathcal{S}$$

(b に関しては $k, k+1, ...$ とかはそこまで考えなくてよくて，a と同じように「ベルマン期待作用素を繰り返し適用することで価値関数に近似できる」と考える)

(証明 a.)
有界関数 v には $|v(s)| < m,\ \forall s \in \mathcal{S}$ なる実数定数 m が存在する．従って

$$V^*(s) - b \le v(s) \le V^*(s) + b$$

を満たすような実数 $b \in \mathbb{R}$ が存在する．この不等式に B_*^k を適用すれば，ベルマン作用素の単調性より，

$$(B_*^k(V^* - b))(s) \le (B_*^k v)(s) \le (B_*^k(V^* + b))(s), \quad \forall s \in \mathcal{S}$$

さらに，定数 b を線形分離して

$$V^*(s) - \gamma^k b \le (B_*^k v)(s) \le V^*(s) + \gamma^k b, \quad \forall s \in \mathcal{S}$$

$0 < \gamma < 1$ なので，$k \to \infty$ とすれば，はさみうちの定理により，命題 a が成り立つ．

(証明 b 準備)

マルコフ期待作用素 $B^k_{\boldsymbol{\pi}}$ は次のように期待値で表現できる.

$$\begin{aligned}(B^k_{\boldsymbol{\pi}}v)(s) &= \sum_{a_0\in\mathcal{A}}\pi_0(a_0|s)\{g(s,a_0)+\gamma\sum_{s_1\in\mathcal{S}}p_T(s_1|s,a_0)\\ &\quad\times\sum_{a_1\in\mathcal{A}}\pi_1(a_1|s_1)\{g(s_1,a_1)+\gamma\sum_{s_2\in\mathcal{S}}p_T(s_2|s_1,a_1)\\ &\quad\cdots\\ &\quad\times\sum_{a_{k-1}\in\mathcal{A}}\pi_{k-1}(a_{k-1}|s_{k-1})\{g(s_{k-1},a_{k-1})+\gamma\sum_{s_k\in\mathcal{S}}p_T(s_k|s_{k-1},a_{k-1})v(s_k)\}\ldots\}\\ &= \mathbb{E}_{A_t|S_t\sim\boldsymbol{\pi},S_{t+1}|S_t,A_t\sim p_T}[\sum_{t=0}^{k-1}\gamma^t g(S_t,A_t)+\gamma^k v(S_k)|S_0=s]\end{aligned}$$

(証明 b)
方策 $\boldsymbol{\pi}=\{\tilde{\boldsymbol{\pi}},\pi_k,...\}$ についての価値関数 $V^{\boldsymbol{\pi}}$ を次のように分解する.

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

右辺第二項について，g は有界より，$|g(S_t,A_t)|\le R_{\max}$ なので,

$$|\mathbb{E}_{A_t|S_t\sim\boldsymbol{\pi},S_{t+1}|S_t,A_t\sim p_T}[\sum_{t=k}^{\infty}\gamma^t g(S_t,A_t)|S_0=s]|\le\sum_{t=k}^{\infty}\gamma^t R_{\max}=\frac{\gamma^k R_{\max}}{1-\gamma}$$

これは

$$|V^{\boldsymbol{\pi}}-\mathbb{E}_{A_t|S_t\sim\tilde{\boldsymbol{\pi}},S_{t+1}|S_t,A_t\sim p_T}[\sum_{t=0}^{k-1}\gamma^t g(S_t,A_t)|S_0=s]|\le\frac{\gamma^k R_{\max}}{1-\gamma}$$

と書き換えられるので,

$$-\frac{\gamma^k R_{\max}}{1-\gamma}\le V^{\boldsymbol{\pi}}-\mathbb{E}_{A_t|S_t\sim\tilde{\boldsymbol{\pi}},S_{t+1}|S_t,A_t\sim p_T}[\sum_{t=0}^{k-1}\gamma^t g(S_t,A_t)|S_0=s]\le\frac{\gamma^k R_{\max}}{1-\gamma}\quad(1)$$

なる不等式が成り立つことがわかる．
次に，ベルマン期待作用素 $B^k_{\tilde{\boldsymbol{\pi}}}$ を関数 v に適用することで，

$$\begin{aligned}(B^k_{\tilde{\boldsymbol{\pi}}}v)(s) &= \mathbb{E}_{A_t|S_t\sim\tilde{\boldsymbol{\pi}},S_{t+1}|A_t,S_t\sim p_T}[\sum_{t=0}^{k-1}\gamma^t g(S_t,A_t)+\gamma^k v(S_k)|S_0=s] \\ &= \mathbb{E}_{A_t|S_t\sim\tilde{\boldsymbol{\pi}},S_{t+1}|A_t,S_t\sim p_T}[\sum_{t=0}^{k-1}\gamma^t g(S_t,A_t)|S_0=s] \\ &\quad +\gamma^k\mathbb{E}_{A_t|S_t\sim\tilde{\boldsymbol{\pi}},S_{t+1}|A_t,S_t\sim p_T}[v(S_k)|S_0=s]\end{aligned}$$

と書き下せるので，任意の $s\in\mathcal{S}$ について

$$\gamma^k\min_{s\in\mathcal{S}} \le (B^k_{\tilde{\boldsymbol{\pi}}}v)(s)-\mathbb{E}_{A_t|S_t\sim\tilde{\boldsymbol{\pi}},S_{t+1}|A_t,S_t\sim p_T}[\sum_{t=0}^{k-1}\gamma^t g(S_t,A_t)|S_0=s] \le \gamma^k\max_{s\in\mathcal{S}} v(s) \quad (2)$$

(2) − (1) より，

$$\gamma^k(\min_{s\in\mathcal{S}} v(s)-\frac{R_{\max}}{1-\gamma}) \le (B^k_{\boldsymbol{\pi}}v)(s)-V^{\boldsymbol{\pi}}(s) \le \gamma^k(\max_{s\in\mathcal{S}} v(s)+\frac{R_{\max}}{1-\gamma}),\quad \forall s\in\mathcal{S},\ \forall k\in\mathbb{N}_0$$

となるので，$k\to\infty$ とすれば，はさみうちの定理により，命題 b が成り立つ．

ベルマン方程式の解の一意性

a. ベルマン方程式の解になる関数 $v:\mathcal{S}\to\mathbb{R}$ は

$$(B_*v)(s)=v(s),\quad \forall s\in\mathcal{S}$$

を満たすが，それは最適価値関数 V^* ただ一つである．

b. 定常方策 $\pi\in\Pi$ のベルマン期待方程式の解になる関数 $r:\mathcal{S}\to\mathbb{R}$ は

$$(B_\pi v)(s)=v(s),\quad \forall s\in\mathcal{S}$$

を満たすが，それは π の価値関数 V^π ただ一つである．

(証明)
背理法によって示す．方程式：

$$(B_*v')(s)=v'(s),\quad \forall s\in\mathcal{S}$$

を満たすような $V' = V^*$ が存在すると仮定する．上の方程式が成り立つということは，B_* を何度 v' に適用しても v' のままなので

$$v'(s) = \lim_{k\to\infty} (B_*^k v')(s) \tag{1}$$

また，動的計画法の収束性により

$$V^*(s) = \lim_{k\to\infty} (B_*^k v')(s) \tag{2}$$

(1),(2) により，

$$v'(s) = V^*(s), \qquad \forall s \in \mathcal{S}$$

が成り立つので，$v' \neq V^*$ という仮定に矛盾する．

ベルマン作用素の縮小性

任意の有界関数 $v : \mathcal{S} \to \mathbb{R},\ v' : \mathcal{S} \to \mathbb{R}$ と $k \in \mathbb{N}_0$ に対し，

a. ベルマン作用素 B_*^k について，

$$\begin{aligned}
\gamma^k \min_{s\in\mathcal{S}}\{v(s) - v'(s)\} &\le \min_{s\in\mathcal{S}}\{(B_*^k v)(s) - (B_*^k v')(s)\} \\
\max_{s\in\mathcal{S}}\{(B_*^k v)(s) - (B_*^k v')(s)\} &\le \gamma^k \max_{s\in\mathcal{S}}\{v(s) - v'(s)\} \\
\max_{s\in\mathcal{S}} |(B_*^k v)(s) - (B_*^k v')(s)| &\le \gamma^k \max_{s\in\mathcal{S}} |v(s) - v'(s)|
\end{aligned}$$

b. 任意の $\boldsymbol{\pi} \in \Pi$ のベルマン期待作用素 $B_{\boldsymbol{\pi}}^k$ について

$$\begin{aligned}
\gamma^k \min_{s\in\mathcal{S}}\{v(S) - v'(s)\} &\le \min_{s\in\mathcal{S}}\{(B_{\boldsymbol{\pi}}^k v)(s) - (B_{\boldsymbol{\pi}}^k v')(s)\} \\
\max_{s\in\mathcal{S}}\{(B_{\boldsymbol{\pi}}^k v)(s) - (B_{\boldsymbol{\pi}}^k v')(s)\} &\le \gamma^k \max_{s\in\mathcal{S}}\{v(s) - v'(s)\} \\
\max_{s\in\mathcal{S}} |(B_{\boldsymbol{\pi}}^k v)(s) - (B_{\boldsymbol{\pi}}^k v')(s)| &\le \gamma^k \max_{s\in\mathcal{S}} |v(s) - v'(s)|
\end{aligned}$$

＜お気持ち＞
B_*^k が収縮写像である，ということは v, v' にそれぞれ B_*^k を適用した時に，$(B_*^k v)$ と $(B_*^k v')$ の距離が v, v' の距離よりも小さくなる (縮む)，ということ．この補題が成り立つと，次のような嬉しいことがある．a. の第三式：

$$\max_{s\in\mathcal{S}} |(B_*^k v)(s) - (B_*^k v')(s)| \le \gamma^k \max_{s\in\mathcal{S}} |v(s) - v'(s)|$$

の v' に V^* を代入する．$V^* = B_*^k V^*$ より，

$$\max_{s\in\mathcal{S}} |(B_*^k v)(s) - V^*(s)| \leq \gamma^k \max_{s\in\mathcal{S}} |v(s) - V^*(s)|$$

と，ベルマン作用素を適用するたびに v が V^* に近付いていくことがわかる．

(証明)
簡便化のため，

$$\underline{\epsilon} \coloneqq \min_{s\in\mathcal{S}}\{v(s) - v'(s)\}, \quad \overline{\epsilon} \coloneqq \max_{s\in\mathcal{S}}\{v(s) - v'(s)\}$$

と定義すると，

$$v(s) - v'(s) \geq \underline{\epsilon},\ v(s) - v'(s) \leq \overline{\epsilon},\ \forall s \in \mathcal{S}$$

が成立するので，

$$v'(s) + \underline{\epsilon} \leq v(s) \leq v(s) + \overline{\epsilon}$$

と書くことができる．この不等式に B_*^k を適用すると，

$$\begin{aligned} & (B_*^k v')(s) + \gamma^k \underline{\epsilon} \leq (B_*^k v)(s) \leq (B_*^k v')(s) + \gamma^k \overline{\epsilon}, \quad \forall s \in \mathcal{S} \\ \Leftrightarrow \quad & \gamma^k \underline{\epsilon} \leq (B_*^k v)(s) - (B_*^k v')(s) \leq \gamma^k \overline{\epsilon}, \quad \forall s \in \mathcal{S} \end{aligned}$$

$$\begin{aligned} \therefore \quad & \gamma^k \min_{s\in\mathcal{S}}\{v(s) - v'(s)\} \leq \min_{s\in\mathcal{S}}\{(B_*^k v)(s) - (B_*^k v')(s)\} \\ & \max_{s\in\mathcal{S}}\{(B_*^k v)(s) - (B_*^k v')(s)\} \leq \gamma^k \max_{s\in\mathcal{S}}\{v(s) - v'(s)\} \end{aligned}$$

3 最適方策

任意の初期状態 $s \in \mathcal{S}$ からの期待リターンを最大化する方策 π^* を最適方策と定義する．

$$V^*(s) = V^{\pi^*}(s),\ \forall s \in \mathcal{S}$$

最適方策の存在性と必要十分条件

- 最適方策 π^* は存在する．
- ある定常な決定的方策 $\pi^* \in \Pi^d$ が最適方策である

$$\Leftrightarrow V^*(s) = (B_{\pi^*} V^*)(s), \quad \forall s \in \mathcal{S}$$

(証明)

($\Leftarrow$) ある π^* が $V^*(s) = (B_{\pi^*}V^*)(s),\ \forall s \in \mathcal{S}$ を満たすとする．この時，ベルマン方程式の解の一意性により，$V^* = V^{\pi^*}$ が成り立つ．

($\Rightarrow$) π^* が $V^*(s) = V^{\pi^*}$ を満たすとする．この時，ベルマン方程式の解の一意性により，$V^*(s) = (B_{\pi^*}V^*)(s),\ \forall s \in \mathcal{S}$ が成り立つ．

最後に，

$$\pi^{d*} := \arg\max_{a\in\mathcal{A}}\{g(s,a) + \gamma \sum_{s'\in\mathcal{S}} p_T(s'|s,a)V^*(s')\}$$

と定義すれば，

$$(B_{\pi^{d*}}V^*)(s) = \max_{a\in\mathcal{A}}\{g(s,a) + \gamma \sum_{s'\in\mathcal{S}} p_T(s'|s,a)V^*(s')\} = V^*(s)$$

より，最適方策は存在することが証明された．

※以上の議論により，マルコフ決定過程において最適方策 π^* は定常な決定的方策として常に存在し，

$$\pi^{d*} := \arg\max_{a\in\mathcal{A}}\{g(s,a) + \gamma \sum_{s'\in\mathcal{S}} p_T(s'|s,a)V^*(s')\}$$

である．従って，p_T が既知であるならば後述の動的計画法によってこの最適方策を求めることができる．

4 動的計画法

動的計画法とは，最適化問題を部分問題に分割して部分問題を再帰的に繰り返し解くことで最適解を求めるアプローチである．次の最適性の原理により，逐次的意思決定問題は時間ステップごとの最適化問題を解くことで時間整合的な最適方策を求められるので，動的計画法を適用することができる．

最適性の原理

時間ステップ T までの逐次的意思決定問題の最適方策系列を

$$\boldsymbol{\pi}^* = \{\pi_0^*, \pi_1^*, \ldots, \pi_T^*\}$$

とし，$\boldsymbol{\pi}^*$ に従い意思決定をし，時間ステップ t で到達可能な状態集合を $\tilde{\mathcal{S}}_t$ とする．この時，任意の $t = \{0, \ldots, T\}$ と $s \in \tilde{\mathcal{S}}_t$ について，時間ステップ t で状態 s から再開し，時間ステップ T までの元の問題の部分問題を考えた場合，$\boldsymbol{\pi}^*$ の部分系列 $\{\pi_t^*, \ldots, \pi_T^*\}$ が最適解であることを最適性の原理という．

4.1 価値反復法

1. 価値関数 $v : \mathcal{S} \to \mathbb{R}$ を任意に初期化する．特に事前知識がない場合，$v(s) \coloneqq 0,\ \forall s \in \mathcal{S}$ とする．
2. v にベルマン作用素を適用することで v' に更新する．
$$v'(s) \coloneqq \max_{a \in \mathcal{A}} \{g(s,a) + \gamma \sum_{s' \in \mathcal{S}} p_T(s'|s,a) v(s')\},\ \forall s \in \mathcal{S}$$
3. もし $\max_{s \in \mathcal{S}} |v(s) - v'(s)| < \epsilon$ ならば，決定的方策：
$$\pi_{v'}^d(s) \coloneqq \arg\max_{a \in \mathcal{A}} \{g(s,a) + \gamma \sum_{s' \in \mathcal{S}} p_T(s'|s,a) v'(s')\},\quad \forall s \in \mathcal{S}$$
を求め終了する．そうでない場合は 2. に戻る．

4.2 方策反復法

1. 決定的方策 $\pi^d : \mathcal{S} \to \mathcal{A}$ を任意に初期化する．
2. 方策 π^d のベルマン方程式を連立一次方程式で解いて，π^d の価値関数 V^{π^d} を求める．
$$V^{\pi^d}(s) = g(s,a) + \gamma \sum_{s \in \mathcal{S}} p_T(s'|s,a) V^{\pi}(s'),\ \forall s \in \mathcal{S}$$
3. 改善方策 $\pi^{d'} : \mathcal{S} \to \mathcal{A}$ を求める．
$$\pi^{d'}(s) \coloneqq \arg\max_{a \in \mathcal{A}} \{g(s,a) + \gamma \sum_{s' \in \mathcal{S}} p_T(s'|s,a) V^{\pi^d}(s')\},\ \forall s \in \mathcal{S}$$
4. もし $\pi^d(s) = \pi^{d'}(s),\ \forall s \in \mathcal{S}$ ならば終了する．