6. 動的計画法

尾山 大輔

経済学のための数学

2025 年 7 月 10 日，17 日

扱うケース

1. 確定的 (deterministic) なケース
   - 連続状態，連続行動
   - 有界な収益関数 (効用関数)
2. 確率的 (stochastic) なケース
   - 有限状態，有限行動

(有限性はそんなに重要ではない．可測性など確率論的な煩雑性を避けるため．)

確定的なケース

次の動学的最適化問題を考える：

$$(*) \quad \max_{(a_t)_{t=0}^{\infty}} \sum_{t=0}^{\infty} \beta^t r(s_t, a_t)$$

$$\text{s.t.} \quad a_t \in A(s_t), \quad t = 0, 1, \ldots$$

$$s_{t+1} = g(s_t, a_t), \quad t = 0, 1, \ldots$$

$$s_0 \in S : \text{所与}$$

- $S \subset \mathbb{R}^k$ $(\neq \emptyset)$: 状態空間 (state space)
- $A \subset \mathbb{R}^\ell$ $(\neq \emptyset)$: 行動空間 (action space)
- $A(s) \subset A$ $(\neq \emptyset)$ (すべての $s \in S$ に対して)
  (有界閉集合，対応として連続であると仮定する)
- $r : S \times A \to \mathbb{R}$: 収益関数 (return function/reward function)
  (連続であると仮定する)
- $g : S \times A \to S$ (連続であると仮定する)
- $\beta \in (0, 1)$: 割引因子

有界収益の仮定

- 収益関数 r は有界関数であると仮定する.
  - ある M が存在して，すべての $(s, a) \in S \times A$ に対して $|r(s, a)| \leq M$ が成り立つ.
- 多くの典型例で，(修正なしでは) この仮定は満たされない.

  が，理論がきれいになるのでこれを採用する.
- 状態空間 S をうまくコンパクト集合に制限することでこの仮定が満たされるようにできることもある.

例：“Cake Eating”

$S = [0, \bar{k}],\ A = [0, \bar{k}] \qquad (\bar{k} > 0)$

$$\max_{(c_t)_{t=0}^{\infty}} \sum_{t=0}^{\infty} \beta^t u(c_t)$$

$$\text{s.t.} \quad c_t \in [0, k_t]$$

$$k_{t+1} = k_t - c_t, \quad t = 0, 1, \ldots$$

$$k_0 \in S : \text{所与}$$

- $A(k) = [0, k]$
- $r(k, c) = u(c)$
- $g(k, c) = k - c$

達成可能な行動列

- 与えられた $s_0 \in S$ に対して,

  行動の列 $(a_0, a_1, a_2, \ldots)$ が

  $$\begin{aligned} a_0 \in A(s_0) \quad &\Rightarrow s_1 = g(s_0, a_0) \\ a_1 \in A(s_1) \quad &\Rightarrow s_2 = g(s_1, a_1) \\ a_2 \in A(s_2) \quad &\Rightarrow s_3 = g(s_2, a_2) \\ \vdots \end{aligned}$$

  を満たすとき，この列を s_0 に対する達成可能 (feasible) な行動列と呼ぶことにし，それら全体の集合を $\Pi(s_0)$ と書くことにする.

- $\underline{a} = (a_0, a_1, a_2, \ldots) \in \Pi(s_0)$ に対して,

  $$U(\underline{a}) = \sum_{t=0}^{\infty} \beta^t r(s_t, a_t)$$

  (ただし $s_{t+1} = g(s_t, a_t)$) と書く.

再帰性

- $\underline{a} = (a_0, a_1, a_2, \ldots) \in \Pi(s_0)$ に対して，$\underline{a}^t = (a_t, a_{t+1}, a_{t+2}, \ldots) \in \Pi(s_t)$ と書くことにする.
- 任意の T に対して,

$$\sum_{\tau=0}^{T} \beta^\tau r(s_\tau, a_\tau) = \sum_{\tau=0}^{t-1} \beta^\tau r(s_\tau, a_\tau) + \beta^t \sum_{\tau=0}^{T-t} \beta^\tau r(s_{t+\tau}, a_{t+\tau})$$

なので，$T \to \infty$ として

$$U(\underline{a}) = \sum_{\tau=0}^{t-1} \beta^\tau r(s_\tau, a_\tau) + \beta^t U(\underline{a}^t)$$

が成り立つ.

最適価値関数

- 最適価値関数

$$v^*(s) = \sup_{\underline{a} \in \Pi(s)} U(\underline{a})$$

- 有界収益の仮定より，関数 $v^* : S \to \mathbb{R}$ は有界関数
  - $|r(s,a)| \leq M$ (すべての $(s,a) \in S \times A$ に対して) ならば，$|v^*(s)| \leq \frac{M}{1-\beta}$ (すべての $s \in S$ に対して)

政策関数

- 関数 $\sigma\colon S \to A$ を政策関数 (policy function)，あるいは単に政策という．

  $\sigma(s) \in A(s)$ (すべての $s \in S$ に対して) を満たす政策を実行可能 (feasible) な政策という．

  $\cdots$ 各状態 $s \in S$ に対して行動 $\sigma(s) \in A(s)$ をとるというルール．

- 以下，政策といえば実行可能な政策を指すものとする．

- 政策 σ を決めれば，各初期状態 s_0 に対して実行可能な行動列

  $$(\sigma(s_0), \sigma(s_1), \sigma(s_2), \ldots) \in \Pi(s_0)$$

  が一つ定まる (ただし $s_{t+1} = g(s_t, \sigma(s_t))$)．

政策価値関数・最適政策

- 政策 σ に対して関数 $v_\sigma : S \to \mathbb{R}$ を

$$v_\sigma(s) = \sum_{t=0}^{\infty} \beta^t r(s_t, \sigma(s_t))$$

(ただし $s_{t+1} = g(s_t, \sigma(s_t))$, $s_0 = s$) で定義する.

$\cdots$ 政策 σ の価値関数

- 政策 σ^* が最適政策であるとは,

$$v^*(s) = v_{\sigma^*}(s) \quad (\text{すべての } s \in S \text{ に対して})$$

が成り立つことをいう.

- (以下, 最適政策が存在する, すなわち, 政策関数から生成される行動列の中に最適行動列が存在する, ということが示される.)

動的計画法の原理

1. 関数 $v\colon S \to \mathbb{R}$ に関する方程式 (Bellman 方程式)

   $$(\star)\quad v(s) = \sup_{a \in A(s)} r(s,a) + \beta v(g(s,a)) \qquad (s \in S)$$

   は一意の解を持ち，それは連続関数で，かつ最適価値関数 v^* に等しい.

2. [最適性原理 (Principle of Optimality)]
   政策 σ^* が最適政策であるための必要十分条件は，それが

   $$(\heartsuit)\quad r(s,\sigma^*(s)) + \beta v^*(g(s,\sigma^*(s)))$$
   $$= \sup_{a \in A(s)} r(s,a) + \beta v^*(g(s,a)) \quad (\text{すべての } s \in S \text{ に対して})$$

   を満たすことである.

3. [1 回逸脱の原理 (One-Shot Deviation Principle)]
   政策 σ^* が最適政策であるための必要十分条件は，それが

   $(\diamondsuit)$ $r(s, \sigma^*(s)) + \beta v_{\sigma^*}(g(s, \sigma^*(s)))$

   $$= \sup_{a \in A(s)} r(s,a) + \beta v_{\sigma^*}(g(s,a)) \quad (\text{すべての } s \in S \text{ に対して})$$

   を満たすことである.

4. 最適政策が存在する.

Bellman Operator

- 有界関数 $v\colon S \to \mathbb{R}$ に対して，関数 $Tv\colon S \to \mathbb{R}$ を

$$(Tv)(s) = \sup_{a \in A(s)} r(s,a) + \beta v(g(s,a)) \qquad (s \in S)$$

  で定義する.

- 有界収益の仮定より，Tv は有界関数である.

- つまり，T は有界関数を有界関数に写す関数.

  T を Bellman 演算子 (Bellman operator) という.

- v が Bellman 方程式の解 $\iff$ v が T の不動点 $(Tv = v)$

- v が連続関数ならば，“sup” は “max” に置きかえられる (極値定理より).

  さらに，Tv は連続関数になる (“最大値定理” より).

  つまり，T は連続関数を連続関数に写す.

- $\mathcal{B}(S)$: S 上の実数値有界関数全体からなる集合

  $\mathcal{C}_b(S)$: S 上の実数値有界連続関数全体からなる集合

  と書くことにする.

- T は $\mathcal{B}(S)$ から $\mathcal{B}(S)$ への関数で，さらに $\mathcal{C}_b(S)$ を $\mathcal{C}_b(S)$ に写す，ということ.

One-Shot Return Operator

- 政策 σ，有界関数 $v\colon S \to \mathbb{R}$ に対して，関数 $T_\sigma v\colon S \to \mathbb{R}$ を

$$(T_\sigma v)(s) = r(s, \sigma(s)) + \beta v(g(s, \sigma(s))) \qquad (s \in S)$$

  で定義する.

- 有界収益の仮定より，$T_\sigma v$ は有界関数である.

  つまり，T_σ は $\mathcal{B}(S)$ から $\mathcal{B}(S)$ への関数.

- 定義より，$T_\sigma v \leq Tv$

  ($v \leq w \iff v(s) \leq w(s)$ (すべての $s \in S$ に対して))

単調性

補題 6.1

$v \leq w$ ならば，$Tv \leq Tw$，$T_\sigma v \leq T_\sigma w$.

証明

- $v \leq w$ とする．$s \in S$ を任意にとる.
- すべての $a \in A(s)$ に対して

$$r(s,a) + \beta v(g(s,a)) \leq r(s,a) + \beta w(g(s,a))$$

- したがって，すべての $a \in A(s)$ に対して

$$r(s,a) + \beta v(g(s,a)) \leq (Tw)(s)$$

- したがって，$(Tv)(s) \leq (Tw)(s)$

縮小写像

- $v, w \in \mathcal{B}(S)$ に対して $d(v,w) = \sup_{s \in S} |v(s) - w(s)|$ とする.
  - $d(v,w) \geq 0$
  - $d(v,w) = 0 \iff v = w$
  - $d(v,w) = d(w,v)$
  - $d(v,w) \leq d(v,u) + d(u,w)$　(三角不等式)

補題 6.2

- $d(Tv, Tw) \leq \beta d(v,w)$
- $d(T_\sigma v, T_\sigma w) \leq \beta d(v,w)$

証明

- $s \in S$ を任意に固定する.
- 任意の $a \in A(s)$ に対して,

$$
\begin{aligned}
& r(s,a) + \beta v(g(s,a)) \\
&= r(s,a) + \beta w(g(s,a)) + \beta(v(g(s,a)) - w(g(s,a))) \\
&\leq (Tw)(s) + \beta d(v,w)
\end{aligned}
$$

したがって, $(Tv)(s) \leq (Tw)(s) + \beta d(v,w)$, つまり
$(Tv)(s) - (Tw)(s) \leq \beta d(v,w)$ が成り立つ.

- 同様に, $(Tw)(s) - (Tv)(s) \leq \beta d(w,v)$ が成り立つ.
- よって $|(Tv)(s) - (Tw)(s)| \leq \beta d(v,w)$
- したがって $d(Tv, Tw) \leq \beta d(v,w)$
- T_σ についても同様.

補題 6.3

T, T_σ は高々一つの不動点を持つ.

証明

- $Tv = v$, $Tw = w$ ならば

$$d(v, w) = d(Tv, Tw) \leq \beta d(v, w)$$

より $(1 - \beta)d(v, w) \leq 0$.

- $\beta < 1$ なので, $d(v, w) \leq 0$.
- したがって, $v = w$ でないといけない.

T_σ の不動点

補題 6.4

任意の政策 σ に対して，v_σ は T_σ の不動点 (したがって唯一の不動点) である.

証明

- $s_0 \in S$ に対して σ が生成する行動列を
  $\underline{a} = (\sigma(s^0), \sigma(s^1), \sigma(s^2), \ldots) \in \Pi(s^0)$　(ただし $s_{t+1} = g(s_t, \sigma(s_t))$)
  とすると,

$$
\begin{aligned}
v_\sigma(s_0) &= U(\underline{a}) \\
&= r(s_0, \sigma(s_0)) + \beta U(\underline{a}^1) \\
&= r(s_0, \sigma(s_0)) + \beta v_\sigma(s_1) = (T_\sigma v_\sigma)(s_0)
\end{aligned}
$$

T の性質

補題 6.5

$v \in \mathcal{B}(S)$ に対して次が成り立つ：

1. $v \geq Tv$ ならば $v \geq v^*$
2. $v \leq Tv$ ならば $v \leq v^*$

証明

1.

- $v \in \mathcal{B}(S)$ に対して $v \geq Tv$ とする.
- 任意に $s_0 \in S$ と $\underline{a} = (a_0, a_1, \ldots) \in \Pi(s_0)$ をとる.

  $((s_1, s_2, \ldots)$ も $s_{t+1} = g(s_t, a_t)$ から定義される$)$
- 仮定より，すべての t に対して

$$v(s_t) \geq \sup_{a \in A(s_t)} r(s_t, a) + \beta v(g(s_t, a)) \geq r(s_t, a_t) + \beta v(s_{t+1})$$

- これを繰り返し使って,

$$\begin{aligned} v(s_0) &\geq r(s_0, a_0) + \beta v(s_1) \\ &\geq r(s_0, a_0) + \beta r(s_1, a_1) + \beta^2 v(s_2) \\ &\vdots \\ &\geq \sum_{t=0}^{T-1} \beta^t r(s_t, a_t) + \beta^T v(s_T) \end{aligned}$$

- ここで $T \to \infty$ とする.

  v は有界で $\beta \in (0,1)$ なので，$\beta^T v(a_T) \to 0$ となり，
  最右辺は $U(\underline{a})$ に収束する.

- したがって，$v(s_0) \geq U(\underline{a})$ である.

- $\underline{a}$ は任意だったので，$v(s_0) \geq v^*(s_0)$ が従う.

2.

- $v \in \mathcal{B}(X)$ に対して $v \leq Tv$ とする.
- 任意に $\varepsilon > 0$ と $s_0 \in S$ をとる.
- $\underline{a} \in \Pi(s_0)$ を以下のように定義する

  $((s_1, s_2, \ldots)$ も $s_{t+1} = g(s_t, a_t)$ から定義される$)$:

  各 t に対して $a_t \in A(s_t)$ を

  $$(Tv)(s_t) \leq r(s_t, a_t) + \beta v(g(s_t, a_t)) + (1 - \beta)\varepsilon$$

  となるものとする.
- 仮定より,

  $$v(s_t) \leq r(s_t, a_t) + \beta v(g(s_t, a_t)) + (1 - \beta)\varepsilon$$

  が成り立つ.

- これを繰り返し使って

$$
\begin{aligned}
v(s_0) &\le r(s_0, a_0) + \beta v(s_1) + (1-\beta)\varepsilon \\
&\le r(s_0, a_0) + \beta r(s_1, a_1) + \beta^2 v(s_2) + (1+\beta)(1-\beta)\varepsilon \\
&\vdots \\
&\le \sum_{t=0}^{T-1} \beta^t r(s_t, a_t) + \beta^T v(s_T) + \sum_{t=0}^{T-1} \beta^t (1-\beta)\varepsilon
\end{aligned}
$$

- ここで $T \to \infty$ とする.

  v は有界で $\beta \in (0,1)$ なので, $\beta^T v(a_T) \to 0$ となり,
  最右辺は $U(\underline{a}) + \varepsilon$ に収束する.

- したがって, $v(s_0) \le U(\underline{a}) + \varepsilon$ である.

- よって, $v(s_0) \le v^*(s_0) + \varepsilon$ (すべての $s_0 \in S$ に対して).

- $\varepsilon > 0$ は任意だったので, $v(s_0) \le v^*(s_0)$ が従う.

T の不動点

命題 6.6

1. T は不動点 (したがって唯一の不動点) を持ち，それは連続関数である.
2. T の不動点は v^* に等しい (したがって，v^* は連続関数である).
3. 任意の $w^0 \in \mathcal{B}(S)$ に対して $\mathcal{B}(S)$ 内の列 $(w^0, w^1, w^2, \ldots)$ を $w^m = Tw^{m-1}$ で定義すると，$\lim_{m\to\infty} d(w^m, v^*) \to 0$.

- 2 は，1 と補題 6.5 から従う.

証明

1, 3.

- $w^0 \in \mathcal{B}(S)$ を任意にとる.
- $\mathcal{B}(S)$ 内の列 $(w^0, w^1, w^2, \ldots)$ を $w^m = Tw^{m-1}$ で定義する.
- 各 m に対して,

$$\begin{aligned} d(w^m, w^{m+1}) &= d(Tw^{m-1}, Tw^m) \\ &\leq \beta d(w^{m-1}, w^m) \\ &\leq \cdots \leq \beta^m d(w^0, Tw^0) \end{aligned}$$

が成り立つことに注意する.

- 任意の $\varepsilon > 0$ に対して，$M \in \mathbb{N}$ を $\frac{\beta^M}{1-\beta} d(w^0, Tw^0) < \varepsilon$ なるものとすると，

  $M \leq m \leq n$ ならば

$$\begin{aligned} d(w^m, w^n) &\leq d(w^m, w^{m+1}) + d(w^{m+1}, w^{m+2}) + \cdots + d(w^{n-1}, w^n) \\ &\leq \beta^m d(w^0, Tw^0) + \beta^{m+1} d(w^0, Tw^0) + \cdots + \beta^{n-1} d(w^0, Tw^0) \\ &= \frac{\beta^m - \beta^n}{1-\beta} d(w^0, Tw^0) \leq \frac{\beta^M}{1-\beta} d(w^0, Tw^0) < \varepsilon \end{aligned}$$

  が成り立つ.

- $d(w^m, w^n) = \sup_{s \in S} |w^m(s) - w^n(s)|$ だったので，これより，各 $s \in S$ に対して実数列 $\{w^m(s)\}_{m=0}^{\infty}$ は Cauchy 列であることがわかる.

- $\mathbb{R}$ の完備性より，実数列 $\{w^m(s)\}_{m=0}^{\infty}$ は何らかの実数に収束する.

  その収束先を $w^*(s)$ とおく.

- $w^*: S \to \mathbb{R}$ は有界関数 ($\mathcal{B}(S)$ の要素) である．(証明略—容易)
- $w^0 \in \mathcal{C}_b(S)$ として上の手続きを行うと，w^* は連続関数 ($\mathcal{C}_b(S)$ の要素) である．(証明略—少々ややこしい)
- $\lim_{m\to\infty} d(w^m, w^*) = 0$
  - 任意に $\varepsilon > 0$ をとり，M を，すべての $m, n \geq M$ に対して $d(w^m, w^n) < \frac{\varepsilon}{2}$ が成り立つものとする．
  - 任意に $s \in S$ をとる．
  - N を，$N \geq M$ かつ $|w^N(s) - w^*(s)| < \frac{\varepsilon}{2}$ が成り立つものとする．
  - すると，$m \geq M$ ならば，
$$\begin{aligned} |w^m(s) - w^*(s)| &\leq |w^m(s) - w^N(s)| + |w^N(s) - w^*(s)| \\ &< \frac{\varepsilon}{2} + \frac{\varepsilon}{2} = \varepsilon \end{aligned}$$
  - $s \in S$ は任意だったので，$d(w^m, w^*) \leq \varepsilon$ が成り立つ．

- 最後に，w^* が T の不動点であることを示す.
- 任意に $\varepsilon > 0$ をとる.
- M を，任意の $m \geq M$ に対して $d(w^m, w^*) < \frac{\varepsilon}{1+\beta}$ となるものとする.
- すると,

$$\begin{aligned} d(Tw^*, w^*) &\leq d(Tw^*, Tw^M) + d(Tw^M, w^*) \\ &\leq \beta d(w^*, w^M) + d(w^{M+1}, w^*) \\ &< \beta \times \frac{\varepsilon}{1+\beta} + \frac{\varepsilon}{1+\beta} = \varepsilon \end{aligned}$$

- $\varepsilon > 0$ は任意だったので，$d(Tw^*, w^*) \leq 0$ である.
- したがって，$Tw^* = w^*$ でないといけない.

動的計画法の原理 (再掲)

命題 6.7

1. Bellman 演算子 T は一意の不動点を持ち，それは連続関数で，かつ最適価値関数 v^* に等しい.

2. [最適性原理]
   政策 σ^* が最適政策である $\iff$ $Tv^* = T_{\sigma^*}v^*$

3. [1 回逸脱の原理]
   政策 σ^* が最適政策である $\iff$ $Tv_{\sigma^*} = T_{\sigma^*}v_{\sigma^*}$

4. 最適政策が存在する.

証明

1. 命題 6.6.
2. $Tv^* = T_{\sigma^*} v^* \iff v^* = T_{\sigma^*} v^* \quad (\because\ v^* \text{ は } T \text{ の不動点})$

   $\iff v^* = v_{\sigma^*} \quad (\because\ v_{\sigma^*} \text{ は } T_{\sigma^*} \text{ の唯一の不動点})$
3. $Tv_{\sigma^*} = T_{\sigma^*} v_{\sigma^*} \iff Tv_{\sigma^*} = v_{\sigma^*} \quad (\because\ v_{\sigma^*} \text{ は } T_{\sigma^*} \text{ の不動点})$

   $\iff v_{\sigma^*} = v^* \quad (\because\ v^* \text{ は } T \text{ の唯一の不動点})$

4. 仮定より，各 $s \in S$ に対して

   - $r(s,a)$ と $g(s,a)$ は a について連続
   - $A(s)$ は有界閉集合

   また，1 より

   - v^* は連続

   よって，

   $$r(s,a) + \beta v^*(g(s,a))$$

   は a について連続.

   それを有界閉集合 $A(s)$ 上で最大化する行動 $a \in A(s)$ が存在する.
   そのようなものをひとつ選んで $\sigma^*(s)$ とする.

   すると，2 から政策 σ^* は最適政策である.

例：“Cake Eating”

- $A(k) = [0, k]$
- $r(k, c) = u(c)$
- $g(k, c) = k - c$
- Bellman 方程式

$$v(k) = \sup_{c \in [0,k]} u(c) + \beta v(k - c)$$

- $u(c) = \dfrac{c^{1-\gamma}}{1-\gamma}$ $(\gamma > 0,\ \gamma \neq 1)$ とする.
- ヒントなしには解けない.
- ヒント：$v^*(k) = A\dfrac{k^{1-\gamma}}{1-\gamma}$ の形をしている (定数 A を定める).

確率的なケース

最初から「最適な政策を選ぶ」という形で定式化する.

- S: 状態空間; $|S| = n$ (≥ 1) とする
- A: 行動空間; $|A| = m$ (≥ 1) とする
- $A(s) \subset A$ $(\neq \emptyset)$ (すべての $s \in S$ に対して)
  状態 s において実行可能な行動の集合
- $r\colon S \times A \to \mathbb{R}$: 収益関数
- $q(s'|s,a)$:
  今期の状態が s で行動 a をとったときに，来期の状態が s' である確率
- $\beta \in (0,1)$: 割引因子
- $\sigma\colon S \to A$: 政策
  (ただし，すべての $s \in S$ に対して $\sigma(s) \in A(s)$)

- 政策 σ をとったとすると：
  - 第 0 期の収益は $r(s_0, \sigma(s_0))$
  - 第 1 期の期待収益は $\sum_{s'} q(s'|s_0, \sigma(s_0)) r(s', \sigma(s'))$
  - 第 2 期の期待収益は $\sum_{s', s''} q(s'|s_0, \sigma(s_0)) q(s''|s', \sigma(s')) r(s'', \sigma(s''))$
- 政策 σ に対して：
  - $n \times n$ 行列 Q_σ を
    $$Q_\sigma(s, s') = q(s'|s, \sigma(s))$$
    で定義する (第 s-s' 要素).
  - n 次元列ベクトル r_σ を
    $$r_\sigma(s) = r(s, \sigma(s))$$
    で定義する (第 s 要素).
- すると：
  - 第 1 期の期待収益は $(Q_\sigma r_\sigma)(s_0)$　(第 s_0 要素)
  - 第 2 期の期待収益は $(Q_\sigma^2 r_\sigma)(s_0)$　(第 s_0 要素)

けっきょく，最適化問題は次のように書ける：

$$(**) \quad \max_{\sigma} \sum_{t=0}^{\infty} \beta^t (Q_\sigma^t r_\sigma)(s_0)$$

$$\text{s.t.} \quad s_0 \in S : \text{所与}$$

- 政策 σ の価値関数 v_σ (n 次元ベクトルと見てもよい):

$$v_\sigma(s) = \sum_{t=0}^{\infty} \beta^t (Q_\sigma^t r_\sigma)(s)$$

- 最適価値関数 v^* (n 次元ベクトルと見てもよい):

$$v^*(s) = \max_{\sigma} v_\sigma(s)$$

- 最適政策 σ^*:

$$v^*(s) = v_{\sigma^*}(s) \quad (\text{すべての } s \in S \text{ に対して})$$

- Bellman operator

$$(Tv)(s) = \max_{a \in A(s)} r(s,a) + \beta \sum_{s' \in S} q(s'|s,a)v(s')$$

- One-shot return operator

$$(T_\sigma v)(s) = r(s,\sigma(s)) + \beta \sum_{s' \in S} q(s'|s,\sigma(s))v(s')$$

動的計画法の原理

1. [Bellman 方程式]
   最適価値関数 v^* は Bellman 方程式

$$(\star) \quad v(s) = \max_{a \in A(s)} r(s,a) + \beta \sum_{s' \in S} q(s'|s,a) v(s') \qquad (s \in S)$$

   の唯一の解である.

2. [最適性原理]
   政策 σ^* が最適政策であるための必要十分条件は，それが

$$(\heartsuit) \quad r(s,\sigma^*(s)) + \beta \sum_{s' \in S} q(s'|s,\sigma^*(s)) v^*(s')$$
$$= \max_{a \in A(s)} r(s,a) + \beta \sum_{s' \in S} q(s'|s,a) v^*(s') \quad (\text{すべての } s \in S \text{ に対して})$$

   を満たすこと.

3. 最適政策が存在する.

例 (Putterman 2005, Example 6.2.1)

- $S = \{s_1, s_2\}$
- $A(s_1) = \{a_1, a_2\}$, $A(s_2) = \{a_3\}$
- $r(s_1, a_1) = 5$, $r(s_1, a_2) = 10$, $r(s_2, a_3) = -1$
- $q(s_1|s_1, a_1) = 1/2, \quad q(s_2|s_1, a_1) = 1/2,$
  $q(s_1|s_1, a_2) = 0, \quad q(s_2|s_1, a_2) = 1,$
  $q(s_1|s_2, a_3) = 0, \quad q(s_2|s_2, a_3) = 1$
- $\beta \in (0, 1)$

- この例ではとりうる政策は 2 つしかない：
  - $\sigma^1(s_1) = a_1,\ \sigma^1(s_2) = a_3$
  - $\sigma^2(s_1) = a_2,\ \sigma^2(s_2) = a_3$
- σ^1 の価値関数 v_{σ^1}：

$$v_{\sigma^1}(s_1) = 5 + \beta\left(\frac{1}{2}v_{\sigma^1}(s_1) + \frac{1}{2}v_{\sigma^1}(s_2)\right)$$

$$v_{\sigma^1}(s_2) = (-1) + \beta v_{\sigma^1}(s_2)$$

$\cdots\ v_{\sigma^1}(s_1), v_{\sigma^1}(s_2)$ を変数とする 2 元連立線形方程式

- σ^2 の価値関数 v_{σ^2}：

$$v_{\sigma^2}(s_1) = 10 + \beta v_{\sigma^2}(s_2)$$

$$v_{\sigma^2}(s_2) = (-1) + \beta v_{\sigma^2}(s_2)$$

- $v_{\sigma^1}(s_2) = v_{\sigma^2}(s_2)$
  - $v_{\sigma^1}(s_1) \geq v_{\sigma^2}(s_1) \implies \sigma^1$ が最適
  - $v_{\sigma^1}(s_1) \leq v_{\sigma^2}(s_1) \implies \sigma^2$ が最適

- Bellman 方程式：

(1) $v(s_1) = \max\left\{5 + \beta\left(\frac{1}{2}v(s_1) + \frac{1}{2}v(s_2)\right), 10 + \beta v(s_2)\right\}$

(2) $v(s_2) = (-1) + \beta v(s_2)$

$\cdots$ $v(s_1), v(s_2)$ を変数とする 2 元連立 (非線形) 方程式

求解アルゴリズム

- 価値反復法 (Value iteration)
- 政策反復法 (Policy iteration)
- 修正政策反復法 (Modified policy iteration)
- 線形計画法 (Linear programming)

- 確率的なケース (有限状態，有限行動) で説明する.

用語

- ベクトル (関数) v に対して，政策 σ が v-greedy であるとは,

$$Tv = T_\sigma v$$

が成り立つこと，すなわち,

$$\sigma(s) \in \arg\max_{a \in A(s)} r(s,a) + \beta \sum_{s' \in S} q(s'|s,a) v(s') \quad \text{(すべての } s \in S \text{ に対して)}$$

が成り立つことをいう.

- この用語を使うと：

$$\sigma^* \text{ が最適政策} \iff \sigma^* \text{ が } v^*\text{-greedy}$$

価値反復法

- 命題 6.6 より，任意のベクトル (関数) v^0 に対して，

$$\lim_{n\to\infty} d(T^n v^0, v^*) = 0.$$

- したがって，任意にとった v^0 に対して，$T^n v^0$ を計算していき，
  $T^n v^0$ と $T^{n+1} v^0$ の差が十分小さくなったところで止めると，
  最後の $T^{n+1} v^0$ $(= \hat{v}$ とおく$)$ は真の最適価値関数 v^* に十分近く，
  $\hat{v}$-greedy な政策 $\hat{\sigma}$ は真の最適政策 σ^* に十分近い.

- 「差が十分小さい」をちゃんと書くと $\cdots$

価値反復法

$\varepsilon > 0$ を定める.

1. $n = 0$ する.

   v^0 を任意にとる.

2. $v^{n+1} = Tv^n$ とする.

3. もし

$$d(v^{n+1}, v^n) < \frac{1-\beta}{2\beta}\varepsilon$$

   ならば，ここで停止し，$\hat{v} = v^{n+1}$ と $\hat{v}$-greedy な政策 $\hat{\sigma}$ を返す.

   さもなくば，$n = n + 1$ として 2 に戻る.

価値反復法

命題 6.8

与えられた $\varepsilon > 0$ に対し，価値反復法は有限回で終了し，

- $\hat{v}$ は v^* の $\frac{\varepsilon}{2}$-近似であり，
- $\hat{\sigma}$ は ε-最適政策である．

ただし：

- v が v^* の δ-近似であるとは $d(v, v^*) < \delta$ が成り立つこと．
- $\hat{\sigma}$ が ε-最適政策であるとは $v_{\hat{\sigma}}$ が v^* の ε-近似であること．

証明

- 命題 6.6 の証明中にやったとおり，任意の m に対して $d(Tv, T^m v) \leq \frac{\beta}{1-\beta} d(v, Tv)$ なので,

$$\begin{aligned} d(Tv, v^*) &\leq d(Tv, T^m v) + d(T^m, v^*) \\ &\leq \frac{\beta}{1-\beta} d(v, Tv) + d(T^m, v^*) \end{aligned}$$

- 第 2 項はいくらでも小さくなるので $d(Tv, v^*) \leq \frac{\beta}{1-\beta} d(v, Tv)$.
- よって，$d(v^n, v^{n+1}) < \frac{1-\beta}{2\beta}\varepsilon$ ならば $d(v^{n+1}, v^*) < \frac{\varepsilon}{2}$.

- 次に，$u = Tv$ と書くとする.
- σ を u-greedy な政策であるとする.

  $v_\sigma = T_\sigma v_\sigma$, $T_\sigma u = Tu$ であることに注意する.
- すると,

$$\begin{aligned} d(v_\sigma, u) &= d(T_\sigma v_\sigma, u) \\ &\leq d(T_\sigma v_\sigma, Tu) + d(Tu, u) \\ &= d(T_\sigma v_\sigma, T_\sigma u) + d(Tu, Tv) \leq \beta d(v_\sigma, u) + \beta d(u, v) \end{aligned}$$

  なので，整理して $d(v_\sigma, u) \leq \frac{\beta}{1-\beta}\beta d(u, v)$.
- よって，$d(v^{n+1}, v^n) < \frac{1-\beta}{2\beta}\varepsilon$ のとき，σ を v^{n+1}-greedy な政策とすると,

$$\begin{aligned} d(v_\sigma, v^*) &\leq d(v_\sigma, v^{n+1}) + d(v^{n+1}, v^*) \\ &\leq \frac{\beta}{1-\beta} d(n^{n+1}, v^n) + \frac{\beta}{1-\beta} d(v^n, v^{n+1}) \\ &= \frac{2\beta}{1-\beta} d(n^{n+1}, v^n) < \varepsilon \end{aligned}$$

政策反復法

- 価値反復法は，Bellman operator T が縮小写像であるということだけを使っていて，いま考えている問題固有の構造はまったく使っていない.
- 構造を使ってもう少し賢く解く.

政策反復法

1. $n = 0$ とする.

   σ^0 を任意にとる.
   (または，v^{-1} を任意にとり，σ^0 を v^{-1}-greedy な政策とする．)

2. [政策評価]
   線形方程式 $v = T_{\sigma^n} v$ を解くことで σ^n の価値 v_{σ^n} を計算し，
   それを v^n とする.

3. [政策改善]
   v^n-greedy な政策を計算し，それを σ^{n+1} とする (つまり $T_{\sigma^{n+1}} v^n = T v^n$).
   (可能ならば $\sigma^{n+1} = \sigma^n$ とする．)

4. もし $\sigma^{n+1} = \sigma^n$ ならば，ここで停止し，$\hat{\sigma} = \sigma^n$ と $\hat{v} = v^n$ を返す.

   さもなくば，$n = n + 1$ として 2 に戻る.

政策反復法

命題 6.9

1. $v_{\sigma^n} = T_{\sigma^n} v_{\sigma^n} \leq T v_{\sigma^n} = T_{\sigma^{n+1}} v_{\sigma^n} \leq T^2_{\sigma^{n+1}} v_{\sigma^n} \leq \cdots \leq v_{\sigma^{n+1}}$.
2. $T^n v_{\sigma^0} \leq v_{\sigma^n}$ $(\leq v^*)$.
3. $\lim_{n\to\infty} v_{\sigma^n} = v^*$.
4. $v_{\sigma^n} = v_{\sigma^{n+1}}$ ならば $v_{\sigma^n} = T v_{\sigma^n}$ で，したがって σ^n は最適政策である.
   $v_{\sigma^n} \neq v_{\sigma^{n+1}}$ ならば $v_{\sigma^n} \neq T v_{\sigma^n}$ で，したがって σ^n は最適政策ではない.

証明

1. ▶ まず,

$$\begin{aligned} v_{\sigma^n} &= T_{\sigma^n} v_{\sigma^n} && (v_{\sigma^n} \text{ は } T_{\sigma^n} \text{ の不動点}) \\ &\leq T v_{\sigma^n} && (\text{任意の } v, \sigma \text{ に対して } T_\sigma v \leq T v) \\ &= T_{\sigma^{n+1}} v_{\sigma^n} && (\sigma^{n+1} \text{ の定義}) \end{aligned}$$

▶ よって，$T_{\sigma^{n+1}}$ の単調性より

$$T_{\sigma^{n+1}} v_{\sigma^n} \leq T^2_{\sigma^{n+1}} v_{\sigma^n} \leq T^3_{\sigma^{n+1}} v_{\sigma^n} \leq \cdots \leq T^k_{\sigma^{n+1}} v_{\sigma^n}$$

▶ $k \to \infty$ とする.

任意の v に対して $T^k_{\sigma^{n+1}} v \to v_{\sigma^{n+1}}$ なので,
とくに $T^k_{\sigma^{n+1}} v_{\sigma^n} \to v_{\sigma^{n+1}}$.

2. 帰納法による：
   - $n = 0$ については自明.
   - $T^n v_{\sigma^0} \leq v_{\sigma^n}$ とする.
   - T の単調性より
     $$T^{n+1} v_{\sigma^0} \leq T v_{\sigma^n}$$
   - 1 より，$T v_{\sigma^n} \leq v_{\sigma^{n+1}}$.
3. $T^n v_{\sigma^0} \to v^*$ なので，2 より $v_{\sigma^n} \to v^*$.
4. 1 より.

政策反復法

命題 6.10

政策反復法は有限回で終了し,

- $\hat{\sigma}$ は最適政策であり,
- $\hat{v}$ は最適価値関数である.

修正政策反復法

- 政策 σ^n の価値 v_{σ^n} は線形方程式 $v = T_{\sigma^n} v$ を解くことで得られる.
- 状態の数が大きいと解くのに時間がかかる.
- $\lim_{k\to\infty} T_{\sigma^n}^k v_{\sigma^{n-1}} = v_{\sigma^n}$ なので，十分大きい k に対して $T_{\sigma^n}^k v_{\sigma^{n-1}} \approx v_{\sigma^n}$.
- v_{σ^n} を $T_{\sigma^n}^k v_{\sigma^{n-1}}$ で置き換えた算法は「修正政策反復法」と呼ばれる.
- $k = 1$ ならば $T_{\sigma^n}^k v_{\sigma^{n-1}} = T v_{\sigma^{n-1}}$ $\cdots$ 価値反復法

  $k \to \infty$ ならば $T_{\sigma^n}^k v_{\sigma^{n-1}} \to v_{\sigma^n}$ $\cdots$ 政策反復法

線形計画法

- 補題 6.5 と命題 6.6 より,
  - $v \geq Tv$ ならば $v \geq v^*$
  - $v^* = Tv^*$
- つまり，v^* は $v \geq Tv$ を満たす関数の中で最小のもの.

- v^* は次の最適化問題の (唯一の) 解である：

$$\begin{aligned} \min_{v \in \mathbb{R}^{|S|}} \quad & \sum_{s \in S} v(s) \\ \text{s. t.} \quad & v(s) \geq (Tv)(s) \quad (\text{すべての } s \in S \text{ に対して}) \end{aligned}$$

(ただし $(Tv)(s) = \max_{a \in A(s)} r(s,a) + \beta \sum_{s' \in S} q(s'|s,a) v(s')$)

- T の定義中の「max」を書き下すと:

$$\begin{aligned} \min_{v \in \mathbb{R}^{|S|}} \quad & \sum_{s \in S} v(s) \\ \text{s. t.} \quad & v(s) \geq r(s,a) + \beta \sum_{s' \in S} q(s'|s,a) v(s') \\ & (a \in A(s) \text{ なるすべての } (s,a) \in S \times A \text{ に対して}) \end{aligned}$$

$\cdots$ 線形計画問題

包絡線公式・包絡線定理

- 確定的ケースを考える．定式化を少し変える．
- 各 $s \in S$ に対して，$\Gamma(s) = \{g(s,a) \in S \mid a \in A(s)\}$ とする．

  また，$\operatorname{graph}\Gamma = \{(s,s') \in S \times S \mid s' \in \Gamma(s)\}$ とする．
- 関数 $F\colon \operatorname{graph}\Gamma \to \mathbb{R}$ を

$$F(s,s') = \max_{a \in A(s)} r(s,a) \text{ s.t. } g(s,a) = s'$$

  で定義する．
- Bellman 方程式は

$$v(s) = \max_{s' \in \Gamma(s)} F(s,s') + \beta v(s')$$

  と書ける．
- 簡単化のため $S \subset \mathbb{R}$ とする (つまり，状態 s は 1 次元)．

例：“Cake Eating”

- $A(k) = [0, k]$
- $r(k, c) = u(c)$
- $g(k, c) = k - c$
- $\Gamma(k) = [0, k]$
- $F(k, k') = u(k - k')$
- $v(k) = \max_{k' \in [0,k]} u(k - k') + \beta v(k')$

包絡線公式

- v^* を s で微分することを考える.
- $F(s, s')$ は s について微分可能であるとする.
  - Cake-eating の例：u が微分可能
- そもそも v^* が微分可能であると仮定するならば,
  $v^{*\prime}$ が満たすべき式を求めることは単なる練習問題 (すでに何度もやった).
- Bellman 方程式

$$v^*(s) = \max_{s' \in \Gamma(s)} F(s, s') + \beta v^*(s')$$

の右辺の最大化問題の解を s^* とすると，包絡線公式は

$$v^{*\prime}(s) = F_1(s, s^*)$$

となる (右辺を s で微分してから s^* を代入するのでよい).

(ただし，F_1 は $F(s, s')$ の第 1 変数 s による微分．)

最適価値関数の微分可能性の十分条件

- 包絡線定理の実質的な内容は v^* の微分可能性.
- S および $\operatorname{graph}\Gamma$ は凸集合であるとする.
- 最適価値関数の微分可能性の十分条件：
  $F(s, s')$ は (s, s') について凹関数.
  - Cake-eating の例：u が凹関数
- このとき，v^* は凹関数.

  (v^* の定義式から示す—証明容易)
- $G(s) = \arg\max_{s' \in \Gamma(s)} F(s, s') + \beta v^*(s')$ と書くことにする.

包絡線定理

命題 6.11

以下を仮定する：

1. F は凹関数.
2. s_0 は S の内点，$v^*(s_0) < \infty$，$s^* \in G(s_0)$.
3. s_0 に十分近いどんな $s \in S$ に対しても $s^* \in \Gamma(s)$.
4. $F(s, s^*)$ は $s = s_0$ において s について微分可能.

このとき，$v^*(s)$ は $s = s_0$ において微分可能で,

$$v^{*\prime}(s_0) = F_1(s_0, s^*)$$

が成り立つ.

- この定理は「Benveniste-Scheinkman の包絡線定理」と呼ばれることがある.
- 定理の実質的な内容は v^* の微分可能性であり,
  (微分可能性を所与として) 包絡線公式を導出することは単なる練習問題.
- Benveniste-Scheinkman は「劣微分」に関する定理 (ふつうは分離定理から証明される) を使って証明したが, そのような高等な定理を使うまでもなく初等的に証明できる.
- “On the Differentiability of the Value Function” も参照のこと

証明

- 仮定 2, 3 より，十分小さな $\varepsilon > 0$ に対して

$$v^*(s_0) = F(s_0, s^*) + \beta v^*(s^*)$$
$$v^*(s_0 - \varepsilon) \geq F(s_0 - \varepsilon, s^*) + \beta v^*(s^*)$$
$$v^*(s_0 + \varepsilon) \geq F(s_0 + \varepsilon, s^*) + \beta v^*(s^*)$$

- 仮定 1 より，v^* は凹関数となる.

したがって,

$$\frac{v^*(s_0) - v^*(s_0 - \varepsilon)}{\varepsilon} \geq \frac{v^*(s_0 + \varepsilon) - v^*(s_0)}{\varepsilon}$$

- これらを合わせて

$$\begin{aligned}\frac{F(s_0, s^*) - F(s_0 - \varepsilon, s^*)}{\varepsilon} &\geq \frac{v^*(s_0) - v^*(s_0 - \varepsilon)}{\varepsilon} \\ &\geq \frac{v^*(s_0 + \varepsilon) - v^*(s_0)}{\varepsilon} \\ &\geq \frac{F(s_0 + \varepsilon, s^*) - F(s_0, s^*)}{\varepsilon}\end{aligned}$$

- ここで $\varepsilon \to 0$ とすると，仮定 4 より最左辺と最右辺は $F_1(s_0, s^*)$ に収束し，したがって間の 2 項も $F_1(s_0, s^*)$ に収束する.
- つまり，v^* は s_0 において微分可能で，$v^{*\prime}(s_0) = F_1(s_0, s^*)$.