# An Introduction to Dynamic Programming

一人の意思決定者 (DM) の長期最大化問題

$t = 1$: *initial state* $s_1 \in S$ (exogenous), choose $a_1 \in A$

get *reward* $r(a_1, s_1) \in \Re$

$t = 2$: state is generated by $s_2(a_1, s_1)$. Choose $a_2 \in A$

get reward $r(a_2, s_2)$

$\ldots$

$a_t$ は歴史に依存してかまわない

$A$ が $s_t$ に依存したり、$r(\cdot, \cdot)$ が $t$ に依存したりもある

$s_t$ は確率的に生成されることもある

## 2回繰り返しゲームにあてはめると

$G = (\{1, 2\}, A_1, A_2, u_1, u_2)$

Focus on Player $i$ as the DM

Player $j$'s strategy $(s_{j1}, s_{j2})$ determines states

$s_{j1} \in A_j$: initial state

next period state $s_{j2} : A_i \times A_j \to A_j$

set of DM's feasible actions $A = A_i$

reward function $u_i : A_i \times A_j \to \Re$

$A_i$ が有限ならば、２期間最大化問題とは

$$\max_{a_{i1},\ a_{i2}\in A_i}\left\{u_i(a_{i1},s_{j1})+u_i\Big(a_{i2},s_{j2}(a_{i1},s_{j1})\Big)\right\}=:f(s_{j1}).$$

($f(s_{j1})$ は Optimal value function と呼ばれる。)

明らかに

$$\max_{a_{i1},\ a_{i2}\in A_i}\left\{u_i(a_{i1},s_{j1})+u_i\Big(a_{i2},s_{j2}(a_{i1},s_{j1})\Big)\right\}$$

$$=\max_{a_{i1}\in A_i}\left\{u_i(a_{i1},s_{j1})+\max_{a_{i2}\in A_i}u_i\Big(a_{i2},s_{j2}(a_{i1},s_{j1})\Big)\right\}.$$

# Backward Induction

Optimal value function に書き換える

$$f(s_{j1}) = \max_{a_{i1} \in A_i} \left\{ u_i(a_{i1}, s_{j1}) + f_2\Big(s_{j2}(a_{i1}, s_{j1})\Big) \right\},$$

ここで

$$f_2\Big(s_{j2}(a_{i1}, s_{j1})\Big) = \max_{a_{i2} \in A_i} u_i\Big(a_{i2}, s_{j2}(a_{i1}, s_{j1})\Big).$$

これを一般化して、任意の有限期間最大化の解は後ろ向き帰納法で求められる！

# Infinite Horizon Dynamic Programming

割引和を最大化する定常的問題を扱う（一般にはそうでもない）

$t = 1$: initial state $s_1 \in S$. Choose $a_1 \in A$. Get reward $u(a_1, s_1)$.

$t \geqq 2$: state transition function $s_t : \{A \times S\}^{t-1} \to S$

Choose $a_t \in A$ based on the history $h_{t-1} \in \{A \times S\}^{t-1}$.

$\to$ choose a *strategy* $\mathbf{a} = (a_1, a_2, \ldots, )$

$a_1 \in A$, $a_t : \{A \times S\}^{t-1} \to A$, $t = 2, 3, \ldots$.

a sequence of transition functions $\mathbf{s} = (s_1, s_2, \ldots)$ も似たような形

(例：other players' pure strategy combination)

目的関数は

$$V(\mathbf{a}; s_1) := \sum_{t=1}^{\infty} \delta^{t-1} \cdot u(a_t(h_t), s_t(h_t)),$$

where $h_1 = (a_1, s_1)$ and $h_t = (a_t(h_{t-1}), s_t(h_{t-1}))$.

Optimal value function

$$f(s_1) := \sup_{\mathbf{a}} V(\mathbf{a}; s_1).$$

## Unimprovable Strategy in One Step

**Def**. Fix an initial state $s_1$ and a strategy $\mathbf{a}$.

For each $t = 1, 2, \ldots$ and each history until $t$-th period,

consider a *one-step deviation strategy* : chooses a different action in period $t$ from $\mathbf{a}$ but follows $\mathbf{a}$ from $t+1$-th period on.

If no one-step deviation strategy gives a greater total discounted payoff than that of $\mathbf{a}$,

then $\mathbf{a}$ is called *unimprovable in one step*.

**Lemma**. Fix an initial state $s_1$. If a strategy $\mathbf{a}$ is unimprovable in one step, then it is unimprovable in any finite steps.

**Proof**. Unimprovable in one step より

$$V(\mathbf{a}; s_1) = \max_{x_1}\Big[u(x_1, s_1) + \delta V(\mathbf{a}(x_1, s_1); s_2(x_1, s_1))\Big],$$

ここで、$\mathbf{a}(x_1, s_1)$ は continuation strategy と呼ばれ、 $s_2(x_1, s_1)$ を initial state とみなし、$a_2(x_1, s_1) \in A$ を初期行動としてあとは $\mathbf{a}$ と同じことをするもの

$h_1 = (x_1, s_1)$ の後も unimprovable in one step より

$$V(\mathbf{a}(h_1); s_2(h_1)) = \max_{x_2}\Big[\, u(x_2, s_2(h_1)) + \delta V\big(\mathbf{a}(x_2, s_2(h_1)); s_3(x_2, s_2(h_1))\big)\Big],$$

合わせると

$$
\begin{aligned}
V(\mathbf{a}; s_1) = \max_{x_1} \Big[ & u(x_1, s_1) + \delta \cdot \max_{x_2} \big\{ u(x_2, s_2(x_1, s_1)) \\
& + \delta V\big(\mathbf{a}(x_2, s_2(x_1, s_1)); s_3(x_2, s_2(x_1, s_1))\big) \big\} \Big] \\
= \max_{x_1,\ x_2} \Big[ & u(x_1, s_1) + \delta u(x_2, s_2(x_1, s_1)) \\
& + \delta V\big(\mathbf{a}(x_2, s_2(x_1, s_1)); s_3(x_2, s_2(x_1, s_1))\big) \Big].
\end{aligned}
$$

つまり、$\mathbf{a}$ は unimprovable in two steps。これを繰り返せばよい。

□

**Proposition**. Fix an initial state $s_1$. If a strategy $\mathbf{a}$ is unimprovable in one step, then $\mathbf{a}$ is an optimal strategy that attains $f(s_1)$.

**Proof**. 背理法の仮定として、$\mathbf{a}$ は unimprovable in one step なのに optimal でないとする。すると他の戦略 $\mathbf{x}$ と $\epsilon > 0$ が存在して

$$V(\mathbf{a}; s_1) + 2\epsilon \leqq V(\mathbf{x}; s_1).$$

割引和なので、$\epsilon$ を固定すると、十分大きい $T$ が存在して、$\mathbf{x}$ と最初の $T$ 期間は同じ行動計画をする任意の戦略 $\mathbf{y}$ について、二つの割引総和はほとんど同じにできる。

$$V(\mathbf{x}, s_1) - \epsilon \leqq V(\mathbf{y}, s_1).$$

特に $\mathbf{y}$ として $T$ 期以降は $\mathbf{a}$ と同じ行動計画を持つ戦略にしてもよい。

合わせて

$$V(\mathbf{a}; s_1) + \epsilon \leqq V(\mathbf{y}, s_1).$$

しかし、$\mathbf{y}$ は $T$ 期間しか $\mathbf{a}$ と違う行動をしないから、これは Lemma より $\mathbf{a}$ が unimprovable in finite steps であることに矛盾する。　□