帯大研報, II-5 (1979) : 77～85.

多段階ゲームの構成とダイナミックプログラミングの手法によるゲーム行動の分析について

中　原　淳　一
(帯広畜産大学心理学研究室)

1978年8月31日受理

Construction of Multi Stage Game and Analysis of Game Behavior by Dynamic Programming Method

By

Jun-ich NAKAHARA

I　マトリックスゲームの表現形式について

社会関係，あるいは社会的相互交渉過程を実験的にとらえようとする実験ゲーム研究 (Experimental game study) においては，ゲーム事態は利得行列 (payoff matrix) によってしめされることが多い。この場合，ゲームは標準形 (normal form) によって表現されており，ゲームに参画する各々のプレイヤーの戦略 (strategy) s_i ($s_i \in S_i$, S_i はプレイヤー i の戦略の集合とし，s_i はその要素とする。K 人ゲームであれば $i=1, 2,..., K$ である) がすべて定まれば，ひとつのプレイ α が完結し，この α について各プレイヤーの効用関数 (utility function, あるいは return function) $M_i(\alpha)=u_i(\alpha)$ が定義されている。したがってあるプレイ α を，α 生成させた戦略のある組合せと同形 (isomorphic) であるとし，さらに α を α における各プレイヤーへのリターン u_i を要素とするベクトル $U(\alpha)$ に対応させれば，

$$\alpha(s_1, s_2,..., s_K)=(u_1(\alpha), u_2(\alpha),..., u_K(\alpha))$$

のように表現できる。したがってひとつのゲーム事態 Γ とは，各々の S_i から s_i を選んでできるすべての可能な s_i の組に，それぞれ利得ベクトル U を対応させる関数であると考えることができる。すなわち，

$$\Gamma(\alpha)=U(\alpha)$$

ただし α は $(s_1, s_2,..., s_K)$ に 1−1 に対応する。

ところでこの関数 Γ は一般的な対応関係でしかないが，ゲーム事態を静態的 (static) に問題にする限り，実証的研究のためにも，理論的にとらえる場合にも，それ以上の構造性を与える必要はあまりなさそうである。ただわれわれの関心は実験的研究に利便を与えるような構

造性をもった動態的ゲーム (dynamic game) を構成することにあるから，上述したような von Neumann[1] によってあたえられた形式をすこし変更してみて，多段階決定過程 (Multi stage decision process) としてゲーム事態を構成するうえで，都合がよくなるように改変してみることにする。以下に報告するのは上述の目的のために筆者がこころみている方法のひとつである。

うえにのべたように関数 Γ は可能な戦略の組合せを，それぞれ利得ベクトルに対応させるが，ここで利得ベクトルはもちろん数値ベクトルと考えられるものであり，一方，戦略の組合せは当然のことながら数値の組合せではなく，数値ベクトルではない。しかしながら，このように関数 Γ によって対応し合う戦略の組合せ，つまり名目の組合せと数値ベクトルとの間に，もうひとつの数値ベクトルを媒介的に挿入させることができると考えてみて，この媒介的にとり入れられた数値ベクトルと利得ベクトルを対応させることにすれば，両者はともに数値ベクトルであるから，この間の対応関係についてはかなりの構造性を持ちこむことができよう。すなわち，関数 Γ は戦略のある組合せ α をある数値ベクトル $x(\alpha)$ に対応させるが，さらにこの $x(\alpha)$ を利得ベクトル $U(\alpha)$ に変換する変換 T があるとしてみる。この場合ゲーム事態は以下のようにあらわされる。

$$\Gamma(\alpha)=x(\alpha)$$
$$U(\alpha)=T\{x(\alpha)\}$$

ここで Γ および T をどのように具体化するかによって種々のゲーム事態が得られるであろうが，Γ を 1−1 の対応関係とし，戦略の組合せ α がそれぞれある定められた数値ベクトル $x(\alpha)$ に対応させられるとし，さらに $x(\alpha)$ が行列 R によって利得ベクトル $U(\alpha)$ に変換されるとしてみよう。この場合ゲーム事態は次のようになる。

$$\Gamma(\alpha)=x(\alpha)$$
$$U(\alpha)=x(\alpha)\cdot R$$

数値例として，簡単に2人2戦略 (2×2) の場合をとりあげてみよう。2人のゲーム参画者 A と B の戦略をそれぞれ $\langle a_1, a_2\rangle$, $\langle b_1, b_2\rangle$ とし，これらから得られる4個の戦略の組合せに対して，Γ が以下のようにそれぞれの数値ベクトルを対応させているとする。

$$\langle a_1, b_1\rangle \quad (1, \ 1)$$
$$\langle a_1, b_2\rangle \quad (1, -1)$$
$$\langle a_2, b_1\rangle \quad (-1, \ 1)$$
$$\langle a_2, b_2\rangle \quad (-1, -1)$$

ここで変換のための行列 R として以下のようなものを選んでいるとすれば，それぞれの

$$R=\begin{pmatrix} -2 & 8 \\ 8 & -2 \end{pmatrix}$$

戦略の組合せに対して，利得ベクトルは次のように定まる。ここでは先の要素がプレイヤーAへのリターン，後の要素がプレイヤーBへのリターンである。

$$\left\{\begin{array}{l}\langle a_1, b_1\rangle\\ \langle a_1, b_2\rangle\\ \langle a_2, b_1\rangle\\ \langle a_2, b_2\rangle\end{array}\right\} \rightarrow \Gamma \rightarrow \left\{\begin{array}{r}(1, \ \ 1)\\ (1, -1)\\ (-1, \ \ 1)\\ (-1, -1)\end{array}\right\} \rightarrow R \rightarrow \left\{\begin{array}{r}(6, \ \ 6)\\ (-10, \ \ 10)\\ (10, -10)\\ (-6, \ -6)\end{array}\right\}$$

これを通常の利得行列表のように書き改めれば以下のようになるが，これは2人非ゼロ和ゲームのうちの，いわゆる囚人のジレンマゲーム (prisoner's dilemma game) を構成している。

$$\begin{array}{c} \\ a_1 \\ a_2 \end{array}\begin{array}{c} \begin{array}{cc} b_1 & \qquad b_2 \end{array} \\ \begin{pmatrix} 6, \ \ 6 & -10, \ \ 10 \\ 10, -10 & -6, \ -6 \end{pmatrix} \end{array}$$

このように数値ベクトル x と変換行列 R を適当にえらんでいけば，われわれはさまざまな個別ゲームを構成していくことができるが，もちろんわれわれの関心はそのような個別ゲームの構成にあるわけではない。もしそうであれば対応関係 Γ だけで十分であり，x や R を媒介させる必然性はなにもない。しかし先にものべたように実験ゲームを人間の諸関係の動的な相互交渉過程を実験的にとらえていくうえでの道具として使用しようとする時には事情がちがってくるであろう。相互交渉の動的なプロセスをとらえようとする時には，その動的なプロセスを生みだしていく人間の行動をとらえるための構造性をゲーム事態がそなえていなければならないのはいうまでもないが，同時にその構造性は実験に参画するプレイヤー達，つまり人間の認知の構造に無理なく入りきるもの，あるいはその構造性を被験者達が利用しうるようなものであることが，少なくとも研究の初期の段階では要請されねばならないであろう。もしもそのような考慮をはらわないとするならば，動態的構造をもったゲーム事態を構成することなどは，元来は単純な作業である。たとえば代表的な個別ゲームをいくつか準備して，それぞれを状態であると想定して，各状態間の推移確率行列を与えれば，マルコフ過程として記述しうるプロセスが得られるであろう。しかしこのような構造性を用いて実験を行ってみても，あまりうるところはなさそうである。なぜならば，構造をうごかしてプロセスを生成させていくのは確率行列であって人間ではなく，被験者達の選択行動は，ゲームの動態的構造とは直接的にはすこしもかかわらない。しかしだからといって被験者達のゲーム選択行動は，個々のゲームを全く個別にプレイしているだけだともいいきれない。ゲームが明らかな動態的構造性をもって呈示されてくるならば，プレイヤー達も当然それに対応した選択行動を行うことを予想しなければならないが，それを直接リフアーする契機をこのようなかたちでの実験は持ち得ないはずである。ゲーム事態の確率構造と，選択行動の確率的構造（それが得られたとして）を比較し対照させるぐらいでしかないだろう。

われわれがとらえようとするのは，人間の相互交渉過程の動態的構造性であって，ゲーム事態それ自身の動態構造ではない。ゲーム事態は被験者の選択行動に動態的様相を与えうる構造性をもち，かつまたそのレファラントとなりうるものであることが求められるのである。前述のような方法で個別ゲームを構成するのはそのための準備である。

われわれの場合，個別ゲーム Γ は個別のある変換 T によって構成される。その限りでは個別ゲームは個別的でしかないが，しかし変換 T をある構造性をもった変域 $\{T\}$ に組み入れることもできる。前述のように T として行列を用いるとすれば，たとえば他の元素は一定であるが，主対角元素が同時に同方向へ同じ距離だけ変動するという変域 $\{R\}$ を考えることもできる。そうすれば，個別ゲームは変域 $\{R\}$ の値域 $\{\Gamma\}$ の値として位置づけることができる。すなわち，個別ゲームのそれぞれは $\{R\}$ を媒介項にして相互に関連し合うわけである。この場合さらにプレイヤー達の反応に依存して $\{R\}$ の要素が，したがってまた $\{\Gamma\}$ の要素である個別ゲームが構成されるようにゲーム事態を構成することができるとすれば，プレイヤーの選択行動によって生成されていく R のあるいは Γ の系列を分析することによって，われわれは相互交渉の動態的構造に近づくことができるはずである。

II 多段階ゲームの構成とダイナミックプログラミングモデルによる分析法

この節では，前節で定義したゲームを基礎にして，動態的ゲームを多段階決定過程として構成し，さらにそれを用いて実験的研究を行った場合の分析法について考察してみる。すでにのべたように，われわれの場合ゲーム事態は，

$$U(\alpha)=x(\alpha)\cdot R$$
$$x(\alpha)=\Gamma(\alpha)$$

として形式化されているが，これへ動態的な構造を持ち込むことはさまざまな方向から可能であろう。ここではそれを R を動態化することによるゲームの構成についてのべてみたい。

いま全体で N 段階の離散的なゲーム過程を考え，このなかの第 n 段階目の変換行列を R_n とあらわし，さらにこの R_n は第 $n-1$ 段階目の変換行列 R_{n-1} と，その時の戦略の組合せから得られた数値ベクトル $x_{n-1}=\Gamma(\alpha)$ に依存し，かつ新しい変換 H によって得られるものとしよう。すなわち，

$$R_n=H(R_{n-1},\ x_{n-1})$$

変換行列 R を R_n として上のように定めれば，前節で定義したゲームは，N 段階の離散的なゲームとして以下のようにあらわされる。

$$U_n(\alpha^n)=x_n(\alpha^n)\cdot R_n$$
$$R_n=H(R_{n-1},\ x_{n-1})$$
$$x_n(\alpha^n)=\Gamma(\alpha^n)\qquad \text{ただし } n=1,\ 2,\ldots,\ N$$

ここでサフィックスは離散的な段階をあらわし，H は R に対する変換をあらわしているが，この H については，ここでは具体的には問題にしないことにしておく。もちろん実験的研究を行う場合には H は厳密に決定されねばならず，またその決定にあたっては実際上の種々の制約があるが，以下の議論には直接関係しない。

さて，上にさだめたようなゲームでは，初期値としてある R_1 が与えられれば，プレイヤー達の戦略の選択に伴ってゲームが進行していくが，そのさい R_n は R_{n-1} に依存するとともに，プレイヤー達の選択行動によってえらばれる x_{n-1} にも依存する。どんな変換行列 R_n が得られるのか，したがってまたどんな利得ベクトル U_n が得られるかが，プレイヤーの戦略の選択に依存しているわけであり，その意味でプレイヤー達は，全体のプロセスを制御 (control) する立場にあるとみることもできる。

すなわち，ゲーム事態をひとつのコントロールシステムとみなせば，R_n はシステムの n 段階目での状態をあらわし，x_n は n 段階のプレイヤー達による入力であり，U_n はシステムの状態と，その時の入力に依存する出力である。一方，次のシステムの状態 R_{n+1} は R_n と入力 x_n に依存している。このように，われわれのゲームでは，プレイヤーの選択行動によって決まる変量 x_n は利得 U_n にかかわり，かつまた利得ベクトルにかかわる変換 R_n にかかわる。つまり，x_n はひとつの制御変量であるとみなすことができるわけである。このような考察は，プレイヤー達が N 段階のゲーム事態の全体にわたって，ある計画された x_n の系列を生成し，それによって R_n および U_n の系列を制御し，プロセスの全体を通じてある目的が果されるように行動するであろうという，ゲーム選択行動に関するひとつのモデルをみちびく。すなわち，N 段階のゲーム過程の全体についての，次のような評価関数を与え，F の値を基準に

$$F(R_1, R_2,..., R_N;\ x_1, x_2,..., x_N)$$

して，変量 x_n および R_n の系列を評価し，最適な x_n および R_n の系列を得て，実際のプレイヤー達の選択行動を分析する際のレファラントにとろうとするわけである。もちろん現実のゲーム実験において，被験者がプロセスの全体を厳密に最適化する行動をとることはまれであろうが，われわれはいまひとつのベースになるモデルについてのべているのであり，より細かく実際のゲーム行動に立入る方法については，後にゆずる。

ところで，$R_n=H(R_{n-1}, x_{n-1})$ であり，かつまた R_1 が初期値として与えられていれば，

$$F(R_1, R_2,..., R_N;\ x_1, x_2,..., x_N)=G(x_1, x_2,..., x_N)$$

に書き改めることができる。したがってそのプロセスの最適化の問題は，関数 G を最適化する x_n を求めることであり，これは N 個の同時方程式

$$\frac{\partial G}{\partial x_n}=0 \qquad n=1, 2,..., N$$

を解くことに帰着する。ただし周知のとおり，このことは一般には容易でない。しかしながら，

この種の多段階の各段階ごとの計画が求められるような問題に対しては，特にその数値的な解法として Bellman, R.[2-4] によって与えられた Dynamic Programming (DP) の手法が有力な武器となる場合がある。DP では，解法の一般化も，あるいは問題の数式による表現も，そう容易ではないが，われわれがこれまでに定義してきたゲーム事態はその守備範囲にうまく入り切るようである。

DP では，ある評価関数に基づいて多段階にわたるシステムの制御を問題にする時，その評価関数が，いわゆるマルコフ特性 (Markov properties) を持っていることが要請されている。これは簡単にいえば，履歴効果がないようにすることであるが，たとえば，全体で N 段階のプロセスで，いま第 k 番目にいるとすれば，のこりの $N-k$ 段階の制御過程が全体に及ぼす効果は k 番目での状態と k 番目，およびそれ以後の制御にのみ依存し，それ以前の制御とは無関係であるという性質である。したがってこの要請に合うように評価関数を設定することが，まず問題になるわけである。

われわれの問題に戻ろう。先にものべたように，第1次近似のモデルとして，われわれはプレイヤーが，プロセスの全体にわたってある評価関数を最適化することを想定する。そのような評価のための基準としてまず手はじめに，ゲームに参画している全てのプレイヤーの利得の和 (total joint gain) を最大化することをとりあげてみよう。

いま初期状態 R_1 で出発する N 段階のゲームでの，すべてのプレイヤーの利得の和を $W_N(R_1)$ であらわせば

$$W_N(R_1)=\sum_{n=1}^{N}\sum_{i=1}^{K}u_{in}$$

である。ただし u_{in} は i 番目のプレイヤーの n 段階目の利得をあらわしている。ところで，u_{in} は利得ベクトル U_n の i 番目の要素であり，$U_n=x_n\cdot R_n$ であるから $\sum_i u_{in}$ を簡単に $w(R_n\cdot x_n)$ とあらわしておこう。そうすれば，

$$W_N(R_1)=\sum_{n=1}^{N}w(R_n\cdot x_n)$$

である。このように評価関数の値が，個々の段階で得られる値を加え合わせて得られる時，評価関数は加法的であるというが，この場合評価関数 W は先にのべたマルコフ特性をみたしている。したがってわれわれは DP によって W_N を最適化する x_n の系列を求めればよいわけである。ここで $W_N(R_1)$ の最大値を $f_N(R_1)$ であらわせば

$$\begin{aligned}f_N(R_1)&=\mathrm{Max}\ W_N(R_1)\\&=\mathop{\mathrm{Max}}_{x_1}\mathop{\mathrm{Max}}_{x_2}\ldots\mathop{\mathrm{Max}}_{x_N}\{w(R_1,x_1)+w(R_2,x_2)+\ldots+w(R_N,x_N)\}\end{aligned}$$

評価関数は加法的であるから

$$f_N(R_1)=\mathop{\mathrm{Max}}_{x_1}\{w(R_1,x_1)\}+\mathop{\mathrm{Max}}_{x_2}\mathop{\mathrm{Max}}_{x_3}\ldots\mathop{\mathrm{Max}}_{x_N}\{w(R_2,x_2)+\ldots+w(R_N,x_n)\}$$

ところで $\underset{x_2}{\text{Max}}\ \underset{x_3}{\text{Max}} \ldots \underset{x_N}{\text{Max}}\{w(R_2, x_2)+w(R_3, x_3)+\ldots+w(R_N, x_N)\}$ は状態 R_2 で始まる $N-1$ 段階のゲーム過程であるから $f_{N-1}(R_2)$ とあらわすことができる。また $R_2=H(R_1, x_1)$ である。つまり最初の決定 x_1 がなされれば，R_1 は変換 H によって $H(R_1, x_1)$ にうつり，全体で N 段階の過程は $N-1$ 段階の過程へ移行する。DP ではここで最適性の原理 (principle of optimality) を導入して重要な漸化式 (recurrence relation) をみちびくが，それは Bellman[3] によれば次のようである。"最適な計画，あるいは政策 (policy) とは，最初の状態，およびそこでの最初の決定がなんであれ，その最初の決定によって結果する状態に関して，残っている決定は最適なポリシイを構成しているという特性を持っていなければならない"。これを適用し，また上に定めた関係を利用すれば，

$$f_N(R_1)=\underset{x_1}{\text{Max}}[w(R_1, x_1)+f_{N-1}\{H(R_1, x_1)\}]$$

である。一方 $N=1$ の場合には $f_1(R_1)$ を求めることは通常容易であり

$$f_1(R_1)=\underset{x_1}{\text{Max}}\ w(R_1, x_1)$$

である。これと $f_N(R_1)$ の上の漸化式を用いて関数方程式の系列 $\{f_N(R_1)\}$ を解いていくことができる。数値解を実際に求める場合の具体的な方法についてはここではのべないが，それが得られれば，われわれは N 段階のゲームの全体にわたって評価関数 W_N を最大にする最適決定の系列としての x_n の系列をうることができる。

さてこのようにして得られた x_n の系列は，プロセス全体を通じてのひとつのモデルとなるが，先にものべたように，実際のプレイで被験者が生成していく x_n の系列（以後 x_n^* とあらわす）は，DP によって生成される x_n とはしばしばくい違うであろう。それでは x_n をどのようにつかって x_n^* を分析していったらよいだろうか。この場合によりどころとなるのが，先にあげた DP の基本的条件である評価関数のマルコフ特性と最適性の原理である。評価関数が履歴効果を持たないということと，最適性はそれまでの経験がなんであれ，ある段階のある状態以後の決定について求められているということは，DP 的解法をどの段階からでも開始しうるということを保証している。また DP 的解法は x_n の系列と同時に $\{f_n(R_1)\}_{n=1,2,\ldots,N}$ の系列を与えるということは，全体の段階数を適宜に伸縮させながら，そのそれぞれについて x_n の系列をうることができるということである。これらのことは x_n^* に x_n をモデルとしてあてはめてみる場合に，実際に被験者達の生成した x_n^* の系列のどの部分が評価関数 W_N を用いた最適化のモデルによる x_n に適合するかを，DP 的解法の開始段階およびその長さを適宜に変更しながら探索していくことを可能にしてくれている。たとえば第 k 番目以後のプロセスが，長さ $N-k$ 段階にわたって x_n がよく x_n^* に適合しているとすれば，われわれは k 番目の段階以降，被験者達は結合利得 (joint gain) を最大にしようとしてプレイしたが，それ以前については別のモデルを適合させる必要があると推論できる。あるいはまた第 k 番目の段階

の付近で，プレイヤー達になにかドラスティックなことがあったと予想することもできる。

さらにわれわれは評価関数 W_N の代わりに，別の評価関数 Y_N を用いてプロセスを解くこともできる。たとえばプレイヤー達は joint gain maximum でなく，プレイヤー間になるべく利得格差がないように，なるべく利得が水平化するようにプレイするというモデルを考えることができる。この場合には，たとえば各プレイでの利得の平均からの差の2乗和をとれば，この評価関数を最小化するという最適化のモデルが考えられる。すなわち

$$Y_N=\sum_n\sum_i(u_{in}-\bar{u}_{in})^2$$

のように Y_N を定め，これを最小にする x_n の系列を求めるわけである。このようにして，たとえば x_n^* の系列の前半に Y_N によるモデルがよく適合し，後半に W_N によるモデルがよく適合したとすれば，われわれは x_n^* の動態的構造性を Y_N と W_N の特性に基づいて検出していることになろう。このことを一般的にのべれば，実際のゲームプレイで被験者によって生成された x_n^* の系列は，段階 N の種々の分割と，そのそれぞれに対する各種の評価関数による計画の系列 x_n と対照せしめられ，もっとも適合度の高い分割と評価関数を選んでモデル構成を行っていくということになろう。かくして，このような DP 的手法による分析法は，x_n^* を単一のプロセスとしてとらえる分析法よりも，よりよくその微細構造に入り込めるはずであり，よりよくその動態的構造をはあくできるはずである。

References

1) von Neumann, J. & Morgenstern, O.: Theory of games and economic behavior. Princeton, Princeton University Press, 1953.
2) Bellman, R. E.: Dynamic Programming. Princeton, Princeton University Press, 1957.
3) Bellman, R. E. & Dreyfus, S. E.: Applied Dynamic Programming. Princeton, Princeton University Press, 1962.
4) Bellman, R. E.: Adaptive Control Processes; A guide Tour. Princeton, Princeton University Press, 1961.

Summary

In a matrix game situation, the selection of strategies, one by each player, determines a play α, and to this strategy complex utility function to each player is defined. Therefore, a matrix game could be considered as a function Γ which defines one to one correspondence between strategy complex and utility vector U.

$$U=\Gamma(\alpha)$$

In the report, the author expands the above definition of matrix game by incorporating scalar vector x and transformation matrix R. Now, a matrix game is redefined as

$$U=x\cdot R$$
$$x=\Gamma(\alpha)$$

Then, the N-stage finite multi-stage matrix game is constructed by using another

transformation H.

$$U_n = x_n \cdot R_n$$
$$R_n = H(R_{n-1}, x_{n-1})$$
$$x_n = \Gamma(\alpha) \qquad n = 1, 2, \ldots, N$$

As a model of player's decisional game behavior in this multi-stage game situation, we can preliminarily consider that they will make a sequence of optimum decisions which will maximize some evaluation function of the game. For instance, they may play the game to make the total joint gain maximum. In this case, the evaluation function is able to be defined as

$$W_N(R_1) = \sum_{n=1}^{N} \sum_{i=1}^{K} u_{in}$$
$$= \sum_{n=1}^{N} w(R_n, x_n)$$

Now, because of additive W, we are able to apply the Dynamic Programming method, and are able to have the recurrence relation

$$f_N(R_1) = \underset{x_1}{\mathrm{Max}}[w(R_1, x_1) + f_{N-1}\{H(R_1, x_1)\}]$$

The numerical solution of the above functional equation is the optimum policy which designates an optimum decision for each stage of the game, and maximizes the evaluation function.

Now, we may shift the analytical first stage to an actual later stage, or we may replace the evaluation function by another strategic nature, and in this way, we will be able to analyze an actual dynamic human game behavior more in detail.