数理解析研究所講究録
第741巻 1991年 177-196

Learning Algorithms for 2 × 2 Stochastic Games with Incomplete Information

鳥取大学・工学部　小柳 淳二 (Junji Koyanagi)
京都大学・工学部　大西 匡光 (Masamitsu Ohnishi)
京都大学・工学部　茨木 俊秀 (Toshihide Ibaraki)

1 序論

学習は様々な分野で研究され, いろいろなモデルや学習アルゴリズムが提案されている, 例えば, コンピュータ科学では最近ニューラルネットワークによる学習の研究が行われ, パターン認識などを複雑なプログラムなしで行うことの実用化を目指している.

ゲーム理論の分野においては, 従来学習に関連した研究はあまりなかったようであるが, 最近になって, ESS (Evolutionary Stable Strategy) の学習や2人零和ゲームにおける学習を扱ったものがいくつかなされてきている. ESS とは対称な非零和ゲームにおける Nash 均衡の概念を強めたものであり, その特徴として, ESS をとる個体からなるグループの中では, 他の戦略をとる少数の個体は ESS をとる個体より低い期待利得しか受け取り得ないという性質があげられる. よって ESS を学習することはグループ内の小数の突然変異体に対する一種の抵抗力を身につける上で重要である. Harley [1] では過去の履歴に重みづけをし, それに基づいて混合戦略を更新する学習アルゴリズムを考え, シミュレーションを用いて, その学習アルゴリズムの有効性を検証している. 一方2人零和確率的ゲームへの学習アルゴリズムの適用を考察したものには Lakshmivarahan and Narendra [4], [5] がある. [5] では各プレイヤーが2個のアクションを持つ2人零和確率的ゲームを考え, あるアクションが成功したときにはそのアクションをとる確率を増やし, 失敗したときに

は減らすという Linear Reward-Penalty アルゴリズム (以下では L_{R-P} アルゴリズムと略す) を両者が用いた場合, 両者の混合戦略の極限が 2 人零和ゲームの均衡戦略に任意に近づけられることが示されている. [4] では各プレイヤーが複数のアクションを持ち, 純粋戦略で均衡解を持つ 2 人零和確率的ゲームを考え, あるアクションが成功したときにはそのアクションをとる確率を増やすが, 失敗したときには変化させないという Linear Reward-Inaction アルゴリズムを両者が用いた場合, 両者の混合戦略の極限が均衡戦略になる確率が任意に 1 に近づけられることが示されている. Lakshmivarahan [3] では, これらの学習アルゴリズムが成功確率が未知の機械の中から最大の成功確率を持つものを選ぶことを目的とする Multi-Armed Bandit 問題にも適用可能であることが示され, さらに他の分野への応用がまとめられている. これら学習アルゴリズムの漸近的性質を導く上で Norman [7] にある小さなステップで動く Markov 過程における状態の極限分布の性質が重要な役割を演じている. Norman [6] には学習アルゴリズムの数学的側面が詳しく述べられている.

本稿で扱う繰り返し 2 人非零和確率的ゲームは, それぞれのプレイヤーが各ステージごとに 2 個のアクションのうちの一方を選び, それらの選ばれたアクションに依存したある確率でそれぞれのプレイヤーが利得 1 または 0 を得るゲームである. ただし, 各プレイヤーはその確率法則について何の知識もなく, さらにプレイヤーは互いに相手を観察することができない, すなわち自分のとったアクションと利得しか情報として得られないものとする. このような状況で一方のプレイヤーがある固定した混合戦略を用いるなら他方のプレイヤーは 例えば L_{R-P} アルゴリズムを用いて自分に有利な戦略を学ぶことが可能になり, その結果, 固定した混合戦略を用いている方は不利益をうけることがある. ただし, ある固定した戦略が Nash 解ならば不利益を受けることはない. Nash 解を求めるには利得を受ける確率法則を知ることが必要であるが, 本稿ではその確率法則が未知の場合にも両方のプレイヤーが L_{R-P} アルゴリズムを適用し, かつそのゲームの Nash 解が 1 個の場合には双方のプレイヤーの戦略は漸近的に Nash 解に収束することを示す.

本稿の構成は次のとうりである. 次章でゲームと L_{R-P} アルゴリズムについて述べ, 3 章では L_{R-P} アルゴリズムの漸近的性質に関する解析を行う. 4 章では最適反応戦略により 4 つの場合

にゲームを分類し, そのうち唯一の Nash 解をもつ 3 つの場合においては L_{R-P} アルゴリズムにより両者の混合戦略の組は漸近的に Nash 解に収束することを示す. 最後の章では得られた結果についてのまとめと今後の課題について述べる.

2 モデル

2.1 2 × 2–確率的ゲームと Nash 均衡解

本稿で扱う繰り返し 2 人非零和確率的ゲームは次のようなものである. 2 人のプレイヤー PA, PB がそれぞれ 2 個のアクション 1 と 2 をもち, それぞれのプレイヤーが各ステージごとに 2 個のアクションのうちどちらか一方を選ぶ. PA のアクション i と PB のアクション j (i, j は 1 または 2) に対し, 確率 $R_{ij}(k,\ l)$ で PA は利得 k, PB は利得 l (k, l は 1 または 0) を受け取る. PA, PB がアクション 1 をとる確率をそれぞれ p, q とすると両者の混合戦略はそれぞれ $(p,\ 1-p)$ と $(q,\ 1-q)$ となる. 混合戦略の組を $(p,\ q)$ で表すと, それに対して PA の期待利得 $E_{\mathrm{A}}(p,\ q)$ と PB の期待利得 $E_{\mathrm{B}}(p,\ q)$ は次のようになる.

$$
\begin{aligned}
E_{\mathrm{A}}(p,\ q) &= pq(R_{11}(1,\ 1) + R_{11}(1,\ 0)) \\
&\quad + p(1-q)(R_{12}(1,\ 1) + R_{12}(1,\ 0)) \\
&\quad + (1-p)q(R_{21}(1,\ 1) + R_{21}(1,\ 0)) \\
&\quad + (1-p)(1-q)(R_{22}(1,\ 1) + R_{22}(1,\ 0)), \\
E_{\mathrm{B}}(p,\ q) &= pq(R_{11}(1,\ 1) + R_{11}(0,\ 1)) \\
&\quad + p(1-q)(R_{12}(1,\ 1) + R_{12}(0,\ 1)) \\
&\quad + (1-p)q(R_{21}(1,\ 1) + R_{21}(0,\ 1)) \\
&\quad + (1-p)(1-q)(R_{22}(1,\ 1) + R_{22}(0,\ 1)).
\end{aligned}
$$

$E_{\mathrm{A}}(p,\ q)$, $E_{\mathrm{B}}(p,\ q)$ を用いて 2 人非零和ゲームにおいて重要な概念である Nash 解を定義する.

定義 2.1 戦略の組 $(p^*,\ q^*)$ が Nash 解であるとは

$$E_{\mathrm{A}}(p,\ q^*) \leq E_{\mathrm{A}}(p^*,\ q^*) \quad for\ all \quad p \neq p^*,$$

かつ

$$E_{\mathrm{B}}(p^*,\ q) \leq E_{\mathrm{B}}(p^*,\ q^*) \quad for\ all \quad q \neq q^*$$

が成り立つときである. □

$$a_{ij} = R_{ij}(1,\ 1) + R_{ij}(1,\ 0), \quad b_{ij} = R_{ij}(1,\ 1) + R_{ij}(0,\ 1)$$

とおけば, a_{ij} (b_{ij}) は PA, PB がアクション i, jを選んだとき PA (PB) が 1 を得る確率であり, ($[a_{ij}]$, $[b_{ij}]$) を期待利得行列とする双行列ゲームはもとのゲームと同じ Nash 解を持つ. a_{ij}, b_{ij} の値をプレイヤーが知っていれば Nash 解を知ることができるが, ここでは双方のプレイヤーは a_{ij}, b_{ij}について何の知識もなく, さらにお互いに相手を観察することができないものとする.

このゲームは繰り返しプレイされるが, もし一方のプレイヤーがすべてのステージで Nash 解以外の固定した混合戦略をとり, 他方のプレイヤーは以下に述べる L_{R-P} アルゴリズムを用いれば自分に有利な戦略を学ぶことができる.

そこで本稿では双方が L_{R-P} アルゴリズムを用いるとどのようになるかを考察する. 以下では煩雑な場合分けを防ぐため a_{ij} (b_{ij}) は相互に全て異なり, かつすべて正で 1 より小さいものとする.

2.2 L_{R-P} アルゴリズム

双方のプレイヤーは利得を受ける確率 $R_{ij}(k,\ l)$ $(i,j = 1,2,\ k,l = 0,1)$ について何の知識もなく, 相手のアクションも利得も観測することができないので, 過去において自分のとったアクションと利得をもとにしてゲームを学習することが必要となる.

以下ではプレイヤーが両方とも L_{R-P} アルゴリズムを適用して混合戦略を更新していく場合を考える. P_n $(n = 1,\ 2, \cdots)$ を第 n ステージにおいて PA がアクション 1 をとる確率とする. すなわち PA の第 n ステージにおける混合戦略は $(P_n,\ 1 - P_n)$ となる. L_{R-P} のもとでは P_n は PA

がそのステージにおいてとったアクションと得た利得により次のように P_{n+1} に更新される.

$$P_{n+1} = \begin{cases} P_n + \theta\beta_{\mathrm{A}}(1-P_n) & \text{アクション 1 をとり利得 1 を得たとき,} \\ P_n - \theta\alpha_{\mathrm{A}}P_n & \text{アクション 1 をとり利得 0 を得たとき,} \\ P_n - \theta\beta_{\mathrm{A}}P_n & \text{アクション 2 をとり利得 1 を得たとき,} \\ P_n + \theta\alpha_{\mathrm{A}}(1-P_n) & \text{アクション 2 をとり利得 0 を得たとき.} \end{cases} \tag{2.1}$$

PB の混合戦略に対応する Q_n の更新も同様である, ただし式 (2.1) における添え字の A を B にかえたものを用いる. β_{A} (β_{B}) とα_{A} (α_{B}) はそれぞれ reward パラメーター, penalty パラメーターと呼ばれる. また θ は両方のプレイヤーの学習のステップサイズを変化させるパラメーターである. なお P_{n+1} が区間 [0, 1] 内にあることを保証するため θ, α_A, α_B, β_A, β_B はすべて区間 (0, 1) 内にあるものとする.

両者の戦略の組の時間変化を表す確率過程 $\{(P_n,\ Q_n);\ n \geq 1\}$ は単位正方形 [0, 1]×[0, 1] 上の時間斉次な Markov 過程であり, その推移確率は学習に関するパラメーターβ_{A}, α_{A}, β_{B}, α_{B}, θ と利得を得る確率 $R_{ij}(k,\ l)$ $(i, j = 1, 2,\ k, l = 0, 1)$ により決定される.

本稿の目的は Markov 過程 $\{(P_n,\ Q_n);\ n \geq 1\}$ の $n \to \infty$ における漸近的な挙動が学習パラメーターやゲームの構造によりどう変化するかを明らかにすることである.

3 混合戦略の漸近的性質

さて両方のプレイヤーが L_{R-P} アルゴリズムを$\alpha_{\mathrm{A}} < \beta_{\mathrm{A}}, \alpha_{\mathrm{B}} < \beta_{\mathrm{B}}$ の条件のもとで用いた場合の, $\{(P_n,\ Q_n);\ n \geq 1\}$ の $n \to \infty$ における漸近的性質について調べる. 上の条件は, 成功を失敗より高く評価することを意味する.

$$(\Delta P_n,\ \Delta Q_n) = (P_{n+1} - P_n,\ Q_{n+1} - Q_n)$$

とする. これは両者の混合戦略の組 $(P_n,\ Q_n)$ の 1 ステージあたりの増分を表わす確率変数である.

$(P_n,\ Q_n) = (p,\ q)$ という条件の下での $(\Delta P_n,\ \Delta Q_n)$ の条件付き期待値を考え，それを θ で割ったものを $w(p,\ q)$ とする．$w(p,\ q)$ は第 n ステージに両者の混合戦略の組が $(p,\ q)$ であるときの $(P_n,\ Q_n)$ の増分の期待方向を示す：

$$\begin{aligned} w(p,\ q) &= (w_{\mathrm{A}}(p,\ q),\ w_{\mathrm{B}}(p,\ q)) \\ &= \frac{1}{\theta}E[(\Delta P_n,\ \Delta Q_n)|\ (P_n,\ Q_n) = (p,\ q)]. \end{aligned} \tag{3.2}$$

この第一成分は

$$w_{\mathrm{A}}(p,\ q) = \frac{1}{\theta}E[\Delta P_n|\ (P_n,\ Q_n) = (p,\ q)]$$

であり，起こり得るすべての事象の組み合わせを考えれば

$$\begin{aligned} E[\Delta P_n|\ (P_n,\ Q_n) = (p,\ q)] &= \theta\beta_{\mathrm{A}}a_{11}(1-p)pq - \theta\alpha_{\mathrm{A}}(1-a_{11})p^2q \\ &\quad +\theta\beta_{\mathrm{A}}a_{12}(1-p)p(1-q) - \theta\alpha_{\mathrm{A}}(1-a_{12})p^2(1-q) \\ &\quad -\theta\beta_{\mathrm{A}}a_{21}p(1-p)q + \theta\alpha_{\mathrm{A}}(1-a_{21})(1-p)^2q \\ &\quad -\theta\beta_{\mathrm{A}}a_{22}p(1-p)(1-q) + \theta\alpha_{\mathrm{A}}(1-a_{22})(1-p)^2(1-q) \end{aligned}$$

となる．

$w_{\mathrm{B}}(p,\ q)$ は上式右辺において a_{ij} を b_{ji} に，p を q に，そして A を B に置き換えることにより得られる．

$w_{\mathrm{A}}(p,\ q)$，$w_{\mathrm{B}}(p,\ q)$ は次のように整理される．

$$\begin{aligned} w_{\mathrm{A}}(p,\ q) &= \beta_{\mathrm{A}}p(1-p)\{(1-q)(a_{12}-a_{22}) + (a_{11}-a_{21})q\} \\ &\quad +\alpha_{\mathrm{A}}[(1-p)^2\{1-a_{22}+(a_{22}-a_{21})q\} \\ &\quad -p^2\{1-a_{12}+(a_{12}-a_{11})q\}], \end{aligned} \tag{3.3}$$

$$\begin{aligned} w_{\mathrm{B}}(p,\ q) &= \beta_{\mathrm{B}}q(1-q)\{(1-p)(b_{21}-b_{22}) + (b_{11}-b_{12})p\} \\ &\quad +\alpha_{\mathrm{B}}[(1-q)^2\{1-b_{22}+(b_{22}-b_{12})p\} \\ &\quad -q^2\{1-b_{21}+(b_{21}-b_{11})p\}]. \end{aligned} \tag{3.4}$$

$w(p,\ q)$ の $(p,\ q)$ に関するヤコビアンを

$$J(p,\ q)=\begin{pmatrix} \dfrac{\partial w_{\mathrm{A}}(p,\ q)}{\partial p} & \dfrac{\partial w_{\mathrm{A}}(p,\ q)}{\partial q} \\ \dfrac{\partial w_{\mathrm{B}}(p,\ q)}{\partial p} & \dfrac{\partial w_{\mathrm{B}}(p,\ q)}{\partial q} \end{pmatrix}$$

で表すと，簡単な計算により各成分は以下のようになる．

$$\begin{aligned}
\frac{\partial w_{\mathrm{A}}(p,\ q)}{\partial p} &= \beta_{\mathrm{A}}(1-2p)\{(1-q)(a_{12}-a_{22})+(a_{11}-a_{21})q\} \\
&\quad +\alpha_{\mathrm{A}}[-2(1-p)\{1-a_{22}+(a_{22}-a_{21})q\} \\
&\quad -2p\{1-a_{12}+(a_{12}-a_{11})q\}], \\
\frac{\partial w_{\mathrm{A}}(p,\ q)}{\partial q} &= \beta_{\mathrm{A}}p(1-p)(a_{11}-a_{21}+a_{22}-a_{12}) \\
&\quad +\alpha_{\mathrm{A}}\{(1-p)^2(a_{22}-a_{21})-p^2(a_{12}-a_{11})\}, \\
\frac{\partial w_{\mathrm{B}}(p,\ q)}{\partial p} &= \beta_{\mathrm{B}}q(1-q)(b_{11}-b_{21}+b_{22}-b_{12}) \\
&\quad +\alpha_{\mathrm{B}}\{(1-q)^2(b_{22}-b_{12})-q^2(b_{21}-b_{11})\}, \\
\frac{\partial w_{\mathrm{B}}(p,\ q)}{\partial q} &= \beta_{\mathrm{B}}(1-2q)\{(1-p)(b_{21}-b_{22})+(b_{11}-b_{12})p\} \\
&\quad +\alpha_{\mathrm{B}}[-2(1-q)\{1-b_{22}+(b_{22}-b_{12})p\} \\
&\quad -2q\{1-b_{21}+(b_{21}-b_{11})p\}].
\end{aligned}$$

さらに $(P_n,\ Q_n)=(p,\ q)$ が与えられたという条件のもとでの $(\Delta P_n,\ \Delta Q_n)$ の条件付き共分散行列を

$$C(p,\ q)=E[[(\Delta P_n,\ \Delta Q_n)-\theta w(p,\ q)]^T[(\Delta P_n,\ \Delta Q_n)-\theta w(p,\ q)]|\ (P_n,\ Q_n)=(p,\ q)]$$

とする．ここで ^{T} は転置を表す．

以上で定義された $w(p,\ q)$, $J(p,\ q)$, $C(p,\ q)$ はもちろん L_{R-P} アルゴリズムのパラメーター β_{A}, α_{A}, β_{B}, α_{B} に依存するが，それらを陽には表記しないことにする．

さて小さなステップで動く Markov 過程の極限分布の漸近的性質に関する Norman の定理 ([7] 参照) により次の定理を得る.

定理 3.1 もし $w(p,\ q) = (0,\ 0)$ の解 $(p_{sol},\ q_{sol})$ が唯一で, そこでのヤコビアン $J(p_{sol},\ q_{sol})$ が負定値ならば,

$$\frac{(P_n - p_{sol},\ Q_n - q_{sol})}{\sqrt{\theta}}$$

の分布は $\theta \downarrow 0$ かつ $n\theta \to \infty$ のとき, 平均 $(0,\ 0)$ 共分散行列 $\Sigma(\infty)$ の正規分布に弱収束する. ただし $\Sigma(\infty)$ は次の行列方程式の解として得られる :

$$J(p_{sol},\ q_{sol})\Sigma(\infty) + \Sigma(\infty)J(p_{sol},\ q_{sol})^T + C(p_{sol},\ q_{sol}) = 0. \ \square \tag{3.5}$$

式 (3.5) は Ljapunov 方程式と呼ばれ, 唯一の正定値解をもつことが知られている.

$w(p,\ q) = (0,\ 0)$ の解を吟味するには $w_{\mathrm{A}}(p,\ q) = 0$ となる $(p,\ q)$ のグラフと $w_{\mathrm{B}}(p,\ q) = 0$ となる $(p,\ q)$ グラフを同じ単位正方形 $[0,\ 1] \times [0,\ 1]$ 上に描きそれらの交点をもつか調べればよい.

以下では $w_{\mathrm{A}}(p,\ q) = 0$ となる $(p,\ q)$ のグラフのみについて調べる. いま

$$c_{\mathrm{A}}(q) = 1 - a_{22} + (a_{22} - a_{21})q,$$

$$d_{\mathrm{A}}(q) = 1 - a_{12} + (a_{12} - a_{11})q.$$

と定義する.

$$(1-q)(a_{12} - a_{22}) + q(a_{11} - a_{21}) = c_{\mathrm{A}}(q) - d_{\mathrm{A}}(q)$$

であるので, 式 (3.3) は次のように書き換えることができる :

$$w_{\mathrm{A}}(p,\ q) = \beta_{\mathrm{A}} p(1-p)(c_{\mathrm{A}}(q) - d_{\mathrm{A}}(q)) + \alpha_{\mathrm{A}}\{c_{\mathrm{A}}(q)(1-p)^2 - d_{\mathrm{A}}(q)p^2\}. \tag{3.6}$$

以下簡単のため

$$\gamma_{\mathrm{A}} = \frac{\alpha_{\mathrm{A}}}{\beta_{\mathrm{A}}},$$

$$e_A(q) = \frac{d_A(q)}{c_A(q)}$$

とする.

補題 3.1 任意の $q \in [0, 1]$ に対し $w_A(p, q) = 0$ は区間 $[0, 1]$ 内に次のような唯一解をもち, その解は q に関して連続である :

$$p_A(q) = \begin{cases} \dfrac{2\gamma_A - (1 - e_A(q)) - \sqrt{(1 - e_A(q))^2 + 4\gamma_A^2 e_A(q)}}{2(1 - e_A(q))(\gamma_A - 1)} & c_A(q) \neq d_A(q) \quad \text{のとき}, \\ 1/2 & c_A(q) = d_A(q) \quad \text{のとき}. \end{cases} \tag{3.7}$$

証明 式 (3.6) において $c_A(q) = d_A(q)$ のとき式 (3.6) は

$$\alpha_A c_A(q)\{(1-p)^2 - p^2\} = 0$$

となり, $p = 1/2$ が解であることは容易にわかる.

$c_A(q) \neq d_A(q)$ の時には $p = 0$ の時と $p = 1$ の時の $w_A(p, q)$ の値を調べると

$$w_A(0, q) = \alpha_A c_A(q),$$

$$w_A(1, q) = -\alpha_A d_A(q)$$

であり, すべての $q \in [0, 1]$ にたいし $c_A(q)$, $d_A(q) > 0$ であることから 区間 $[0, 1]$ 内の $w_A(p, q) = 0$ の解は唯一であることがわかる.

$w_A(p, q)$ は p に関して 2 次式であり, $c_A(q) > d_A(q)$ の時上に凸な関数で, $c_A(q) < d_A(q)$ の時上に凹な関数である. よって 2 次方程式 $w_A(p, q) = 0$ の 2 個の解

$$p_A(q) = \frac{2\gamma_A - (1 - e_A(q)) \pm \sqrt{(1 - e_A(q))^2 + 4\gamma_A^2 e_A(q)}}{2(1 - e_A(q))(\gamma_A - 1)} \tag{3.8}$$

のうち区間 $[0, 1]$ 内にある解は $c_A(q) > d_A(q)$ の時大きい方で, $c_A(q) < d_A(q)$ の時小さい方である.

$c_A(q) > d_A(q)$ のとき $e_A(q) < 1$ であり, $\gamma_A < 1$ であるから式 (3.8) の右辺の分母は負になる. 分子は正号 + をとると正になり, 負号 − をとると負になる. よって大きい方の解は負号 − をとることにより得られる.

一方 $c_A(q) < d_A(q)$ のときは $e_A(q) > 1$, また $\gamma_A < 1$ であるから式 (3.8) の右辺の分母は正になる. 分子は常に正であり, 小さい方の解は負号 $-$ をとることにより得られる.

以上のことから任意の $q \in [0,\ 1]$ に対し区間 $[0,\ 1]$ 内にある $w_A(p,\ q) = 0$ の解は式 (3.7) で与えられる.

$p_A(q)$ の q に関する連続性は $c_A(q) = d_A(q)$ のときに問題となる. よって $e_A(q) \to 1$ のとき式 (3.7) の右辺の上式が $1/2$ に収束することを以下で示す.

$$\begin{aligned}
&\lim_{e_A(q)\to 1} \frac{2\gamma_A - (1-e_A(q)) - \sqrt{(1-e_A(q))^2 + 4\gamma_A^2 e_A(q)}}{2(1-e_A(q))(\gamma_A - 1)} \\
=\ &\lim_{x\to 1} \frac{2\gamma_A - (1-x) - \sqrt{(1-x)^2 + 4\gamma_A^2 x}}{2(1-x)(\gamma_A - 1)} \\
=\ &\lim_{x\to 1} \frac{1 - \frac{-2(1-x)+4\gamma_A^2}{2\sqrt{(1-x)^2+4\gamma_A^2 x}}}{-2(\gamma_A - 1)} = 1/2,
\end{aligned}$$

ただし第 1 の等号は $e_A(q)$ を x と置くことにより, 第 2 の等号は l'Hospital の定理を用いることにより得られる. よって $p_A(q)$ は q に関して連続である. □

補題 3.2 (1) $p_A(q)$ は q に関し単調増加あるいは単調減少である.

(2) $p_A(q)$ は γ_A に関し $c_A(q) < d_A(q)$ のとき単調増加であり, $c_A(q) > d_A(q)$ のとき単調減少である.

証明 (1) $p_A(q)$ を q に関して微分すると,

$$\begin{aligned}
p'_A(q) &= \frac{\left(e'_A(q) - \frac{-2e'_A(q)(1-e_A(q))+4\gamma_A^2 e'_A(q)}{2f_A(q)}\right)(1-e_A(q)) + e'_A(q)(2\gamma_A - 1 + e_A(q) - f_A(q))}{2(\gamma_A - 1)(1-e_A(q))^2} \\
&= g(q)\left[\left(1 + \frac{(1-e_A(q)) - 2\gamma_A^2}{f_A(q)}\right)(1-e_A(q)) + 2\gamma_A - 1 + e_A(q) - f_A(q))\right] \\
&= g(q)\left[\frac{(1-e_A(q))^2 - 2\gamma_A^2(1-e_A(q))}{\sqrt{(1-e_A(q))^2 + 4\gamma_A e_A(q)}} + 2\gamma_A - \sqrt{(1-e_A(q))^2 + 4\gamma_A e_A(q)}\right] \\
&= g(q)\left[2\gamma_A + \frac{(1-e_A(q))^2 - 2\gamma_A^2(1-e_A(q)) - (1-e_A(q))^2 - 4\gamma_A^2 e_A(q)}{\sqrt{(1-e_A(q))^2 + 4\gamma_A e_A(q)}}\right] \\
&= g(q)\left[2\gamma_A + \frac{-2\gamma_A^2 - 2\gamma_A^2 e_A(q)}{\sqrt{(1-e_A(q))^2 + 4\gamma_A e_A(q)}}\right]
\end{aligned}$$

$$= \frac{\gamma_A e'_A(q)}{(1-e_A(q))^2(\gamma_A-1)}\left[1-\frac{\gamma_A(1+e_A(q))}{\sqrt{(1-e_A(q))^2+4\gamma_A^2 e_A(q)}}\right], \tag{3.9}$$

ここで $'$ は q に関する微分を表し,

$$f_A(q)=\sqrt{(1-e_A(q))^2+4\gamma_A e_A(q)},$$

$$g(q)=\frac{e'_A(q)}{2(\gamma_A-1)(1-e_A(q))^2}$$

とおいた.

式 (3.9) の右辺の $[\cdot]$ 内の式は常に正であることが以下で示される.

$$\begin{aligned}
& 1-\frac{\gamma_A(1+e_A(q))}{\sqrt{(1-e_A(q))^2+4\gamma_A^2 e_A(q)}} \\
> \; & 1-\frac{\gamma_A(1+e_A(q))}{\sqrt{\gamma_A^2(1-e_A(q))^2+4\gamma_A^2 e_A(q)}} \\
= \; & 1-\frac{\gamma_A(1+e_A(q))}{\sqrt{\gamma_A^2(1+e_A(q))^2}} \\
= \; & 0.
\end{aligned}$$

以上のことから $p'_A(q)$ と $e'_A(q)$ とは逆の符号をもつことがわかる. また $e_A(q)$ は q に関する線形分数関数なので, $e'_A(q)$ はすべての $q\in[0,\,1]$ にたいし一定の符号をもつ. よって $p_A(q)$ は q に関して単調である.

(2) $p_A(q)$ をパラメーターγ_A に関して偏微分すれば,

$$\begin{aligned}
\frac{\partial p_A(q)}{\partial \gamma_A} &= \frac{\left(2-\frac{8\gamma_A e_A(q)}{2f_A(q)}\right)(\gamma_A-1)-(2\gamma_A-1+e_A(q)-f_A(q))}{2(1-e_A(q))(\gamma_A-1)^2} \\
&= h(q)\left[2(\gamma_A-1)-\frac{4\gamma_A e_A(q)(\gamma_A-1)}{f_A(q)}-2\gamma_A+1-e_A(q)+f_A(q)\right] \\
&= h(q)\left[-1-e_A(q)+\frac{(1-e_A(q))^2+4\gamma_A^2 e_A(q)-4\gamma_A e_A(q)(\gamma_A-1)}{\sqrt{(1-e_A(q))^2+4\gamma_A^2 e_A(q)}}\right] \\
&= h(q)\left[-1-e_A(q)+\frac{4\gamma_A e_A(q)+(1-e_A(q))^2}{\sqrt{(1-e_A(q))^2+4\gamma_A^2 e_A(q)}}\right]
\end{aligned} \tag{3.10}$$

ただし

$$h(q)=\frac{1}{2(e_A(q)-1)(\gamma_A-1)^2}$$

とおいた.

式 (3.10) の右辺の $[\cdot]$ 内の式は負であることを以下の不等式を証明することにより示す.

$$1+e_A(q) > \frac{(1-e_A(q))^2+4\gamma_A e_A(q)}{\sqrt{(1-e_A(q))^2+4\gamma_A^2 e_A(q)}}. \tag{3.11}$$

式 (3.11) の両辺を 2 乗した後整理すれば

$$(1+e_A(q))^2\{(1-e_A(q))^2+4\gamma_A^2 e_A(q)\} > \{(1-e_A(q))^2+4\gamma_A e_A(q)\}^2 \tag{3.12}$$

を得る. 式 (3.12) の左辺から右辺を引けば

$$\begin{aligned}
& (1+e_A(q))^2(1-e_A(q))^2+(1+e_A(q))^2 4\gamma_A^2 e_A(q)-(1-e_A(q))^4 \\
& -8\gamma_A e_A(q)(1-e_A(q))^2-16\gamma_A^2 e_A(q)^2 \\
= \; & (1-e_A(q))^2(1+e_A(q))^2-(1-e_A(q))^4-8\gamma_A e_A(q)(1-e_A(q))^2 \\
& +(1+e_A(q))^2 4\gamma_A^2 e_A(q)-16\gamma_A^2 e_A(q)^2 \\
= \; & (1-e_A(q))^2\{(1+e_A(q))^2-(1-e_A(q))^2-8\gamma_A e_A(q)\} \\
& +4e_A(q)\gamma_A^2\{(1+e_A(q))^2-4e_A(q)\} \\
= \; & (1-e_A(q))^2\{4e_A(q))-8\gamma_A e_A(q)\}+4\gamma_A^2 e_A(q)(1-e_A(q))^2 \\
= \; & 4e_A(q)(1-e_A(q))^2(\gamma_A^2-2\gamma_A+1) \\
= \; & 4(1-e_A(q))^2 e_A(q)(\gamma_A-1)^2 \\
> \; & 0.
\end{aligned}$$

よって式 (3.10) は $c_A(q) < d_A(q)$ $(e_A(q) > 1)$ のとき単調増加であり, $c_A(q) > d_A(q)$ $(e_A(q) < 1)$ のとき単調減少である.

補題 3.3 (1) $c_A(q) > d_A(q)$ のとき, $\gamma_A \downarrow 0$ とすれば $p_A(q)$ は下から単調に 1 に収束する.

(2) $c_A(q) < d_A(q)$ のとき, $\gamma_A \downarrow 0$ とすれば $p_A(q)$ は上から単調に 0 に収束する.

(3) $c_A(q) = d_A(q)$ のとき, $p_A(q) = 1/2$.

証明 $\gamma_A \downarrow 0$ とすると式 (3.7) より $e_A(q) < 1$ か > 1 によって $p_A(q)$ が 1 または 0 に収束することがわかる. 単調性は補題 3.2 (2) からあきらかである. □

q が 0 から 1 まで動くとき $c_A(q) - d_A(q)$ の符号は

$$q^* = \frac{a_{22} - a_{12}}{a_{22} - a_{21} - a_{12} + a_{11}} \tag{3.13}$$

が区間 [0, 1] 内にあればそこで変わることになるが, これは $a_{11} - a_{21}$ の符号と $a_{12} - a_{22}$ の符号が異なるときである.

$w_B(p, q)$ についても添え字 A を B, p を q, a_{ij} を b_{ji} に置き換えれば同様の補題が得られる.

4 4つの場合

通常双行列ゲーム ($[a_{ij}]$, $[b_{ij}]$) は a_{ij} と b_{ij} の大小の順によって分類されるが, ここで重要なのは $a_{12} - a_{22}$, $a_{11} - a_{21}$ ($b_{21} - b_{22}$, $b_{11} - b_{12}$) の符号である. 符号の組合せには 4 通りあり, それらはそのゲームの PA の最適反応戦略と以下に示すような関係がある

A1: $a_{12} - a_{22} < 0$, $a_{11} - a_{21} < 0$ のとき, PA の最適反応戦略は PB のアクションにかかわらずアクション 2 をとることである.

A2: $a_{12} - a_{22} < 0$, $a_{11} - a_{21} > 0$ のとき, PA の最適反応戦略は PB のアクションと同じアクションをとることである.

A3: $a_{12} - a_{22} > 0, a_{11} - a_{21} < 0$ のとき, PA の最適反応戦略は PB のアクションと逆のアクションをとることである.

A4: $a_{12} - a_{22} > 0, a_{11} - a_{21} > 0$ のとき, PA の最適反応戦略は PB のアクションにかかわらずアクション 1 をとることである.

PB に対しても A を B にかえ, a_{ij} を b_{ji} にかえれば同様の場合わけができる. 全部で $4 \times 4 = 16$ の場合があるように思われるが, アクションやプレイヤーの名前のつけかえにより 4 つの場合を調

べれば十分であることがわかる. 以下ではそれら 4 つの場合について調べる.

Case 1. $a_{12}-a_{22}<0,\ a_{11}-a_{21}<0,\ b_{21}-b_{22}<0,\ b_{11}-b_{12}<0.$

この場合 Nash 解は (0, 0) すなわち双方のプレイヤーがアクション 2 をとることである. また $c_A(q)<d_A(q)$, $c_B(p)<d_B(p)$ が任意の $(p,\ q)$ について成り立つので補題 3.3 (2) を用いれば $(p_{sol},\ q_{sol})$ は$\gamma_A \downarrow 0,\ \gamma_B \downarrow 0$ のとき (0, 0) に収束することがわかる (図 1 参照).

次にヤコビアン $J(p_{sol},\ q_{sol})$ は

$$J(0,\ 0)=\begin{pmatrix} \beta_A(a_{12}-a_{22})-2\alpha_A(1-a_{22}) & \alpha_A(a_{22}-a_{21}) \\ \alpha_B(b_{22}-b_{12}) & \beta_B(b_{21}-b_{22})-2\alpha_B(1-b_{12}) \end{pmatrix}$$

で近似される. これは負定値であるから $J(p_{sol},\ q_{sol})$ もγ_A, γ_B を十分小さくとることにより負定値にすることができる.

Case 2. $a_{12}-a_{22}<0,\ a_{11}-a_{21}<0,\ b_{21}-b_{22}<0,\ b_{11}-b_{12}>0.$

この場合も Nash 解は (0, 0) すなわち双方のプレイヤーがアクション 2 をとることである. しかし Case 1 では双方のプレイヤーは相手のアクションにかかわらずアクション 2 をとるのが最適反応戦略であったが, この場合は PB は PA がどちらのアクションをとるかにより自分のアクションを変える必要がある. $c_A(q)<d_A(q)$ であるから, 補題 3.3 (2) より $p_A(q)$ は$\gamma_A \downarrow 0$ とすれば 0 に収束する. 一方 $p<p^*$ に対しては $c_B(p)<d_B(p)$, $p>p^*$ 対しては $c_B(p)>d_B(p)$ であるから $q_B(p)$ は $\gamma_B \downarrow 0$ とすれば 0 ($p<p^*$ のとき) または 1 ($p>p^*$ のとき) に収束することが補題 3.3 (1) および (2) からわかる. $p=p^*$ のときには 補題 3.3 (3) より $q_B(p)=1/2$ である. 以上より $(p_{sol},\ q_{sol})$ は$\gamma_A \downarrow 0,\ \gamma_B \downarrow 0$ のとき (0, 0) に収束することがわかる (図 2 参照).

ヤコビアンは Case 1 と同様であるので省略する.

Case 3. $a_{12}-a_{22}<0,\ a_{11}-a_{21}>0,\ b_{21}-b_{22}>0,\ b_{11}-b_{12}<0.$

Cases 1 と 2 では Nash 解は純粋戦略の組であったがこの場合は混合戦略の組

$$(p^*,\ q^*)=\left(\frac{b_{22}-b_{21}}{b_{22}-b_{21}-b_{12}+b_{11}},\ \frac{a_{22}-a_{12}}{a_{22}-a_{21}-a_{12}+a_{11}}\right). \tag{4.14}$$

が Nash 解である．Case 2 と同様の議論を $p_A(q)$，$q_A(p)$ に適用すれば $\gamma_A \downarrow 0$ かつ$\gamma_B \downarrow 0$ のとき $(p_{sol}, q_{sol}) \to (p^*, q^*)$ がわかる (図 3 参照).

ヤコビアン $J(p_{sol}, q_{sol})$ は

$$J(p^*, q^*) = \begin{pmatrix} & \beta_A p^*(1-p^*) \\ -2\alpha_A(1-a_{22} & (a_{22}-a_{12}-a_{21}+a_{11}) \\ +(a_{22}-a_{21})q^*) & +\alpha_A[(1-p^*)^2(a_{22}-a_{21}) \\ & -p^{*2}(a_{12}-a_{11})] \\ \\ \beta_B q^*(1-q^*) & \\ (b_{22}-b_{21}-b_{12}+b_{11}) & -2\alpha_B(1-b_{22} \\ +\alpha_B[(1-q^*)^2(b_{22}-b_{12}) & +(b_{22}-b_{12})p^*) \\ -q^{*2}(b_{21}-b_{11})] & \end{pmatrix}.$$

で近似される．$\mathrm{tr}(J(p^*, q^*))$ と $[\mathrm{tr}(J(p^*, q^*))]^2 - 4[\det(J(p^*, q^*))]$ は共に負であるので，この行列は負定値であることがわかる．

Case 4. $a_{12} - a_{22} < 0$, $a_{11} - a_{21} > 0$, $b_{21} - b_{22} < 0$, $b_{11} - b_{12} > 0$.

この場合 Nash 解は 3 つ (0, 0), (1, 1), (p^*, q^*) あり，それに応じて $p_A(q)$, $q_B(p)$ の交点も 3 つある．各交点は$\gamma_A \downarrow 0$ かつ$\gamma_B \downarrow 0$ のときそれぞれ 3 つの Nash 解に収束し（図4 参照），(0, 0), (1, 1) の近傍にある交点ではヤコビアンは負定値であることが容易にわかる．直感的には混合戦略の組はこのどちらかの Nash 解に収束しそうだがはっきりしたことは現在のところ不明である．

5 結論

本稿では不完全情報を持つ繰り返し 2×2-確率的ゲームを扱い，もしゲームが唯一の Nash 解を持ち，両方のプレイヤーが L_{R-P} アルゴリズムを適切なパラメーターのもとで用いれば，両者の混合

戦略の組は漸近的に Nash 解に収束することを示した. .

しかし 2 × 2–確率的ゲームの中で Nash 解を 3 つ持つものについては, L_{R-P} アルゴリズムの漸近的性質について十分な考察を加えることができなかった. 今後の課題としたい.

References

[1] Harley, C. B., Learning the evolutionarily stable strategy, *J. Theoretical Biology*, 89 (1981), pp. 611-633.

[2] Hines W. G. S. and Bishop D. T., On learning and the evolutionarily stable strategy, *J. Applied Probability*, 20 (1983), pp. 689-695.

[3] Lakshmivarahan, S., *Learning Algorithms Theory and Applications*, Springer-Verlag, New York, 1981.

[4] Lakshmivarahan, S. and Narendra, K. S., Learning algorithms for two-person zero-some stochastic games with incomplete information, *Mathematics of Operations Research*, 6 (1981), pp. 379-386.

[5] Lakshmivarahan, S. and Narendra, K. S. , Learning algorithms for two-person zero-some stochastic games with incomplete information: a unified approach, *SIAM J. Control and Optimization*, 20 (1982), pp. 541-552.

[6] Norman, M. F., *Markov Processes and Learning Models*, Academic Press, New York, 1972.

[7] Norman, M. F., A central limit theorem for Markov processes that move by small steps, *Annals of Probability*, 6 (1974), pp. 1065-1074.

[8] Norman, M. F., Markovian learning processes, *SIAM Review*, 16 (1974), pp. 143-162.

Appendix

増分の規模を示すパラメーター $\theta \in (0,\ 1)$ に対し，$\{X_n^\theta;\ n \geq 0\}$ を $\mathcal{R}^M$ の部分集合 I を状態空間に持つ時間斉次なマルコフ過程とする．このとき $\theta \downarrow 0$ かつ $n\theta \to \infty$ の時の漸近的な性質について次のことが解っている．

$\Delta X_n^\theta = X_{n+1}^\theta - X_n^\theta$ として以下を仮定する．

A1: $E[\Delta X_n^\theta \mid X_n^\theta = x] = \theta w(x) + O(\theta^2)$,

A2: $E[(\Delta X_n^\theta - \theta w(x))^T(\Delta X_n^\theta - \theta w(x)) \mid X_n^\theta = x] = \theta^2 C(x) + o(\theta^2)$,

A3: $E[|\Delta X_n^\theta|^3 \mid X_n^\theta = x] = O(\theta^3)$, ここで $|\cdot|$ は I で定義されたノルムであり，

A1, A2, A3 に現れるオーダーは $x \in I$ で一様とする．

A4: I はコンパクト，

A5: $w(x)$ は bounded Lipschitz derivative をもち，

A6: $C(x)$ は I で Lipschitz，

A7: $w(x) = 0$ となる解 $x_{sol} \in I$ が唯一，

A8: $w(x)$ のヤコビアン $J(x)$ が x_{sol} で負定値．

ここで

$$\begin{aligned} \mu_n^\theta(x) &= E[X_n^\theta \mid X_0^\theta = x], \\ \omega_n^\theta(x) &= E[(X_n^\theta - \mu_n^\theta(x))^T(X_n^\theta - \mu_n^\theta(x)) \mid X_0^\theta = x]. \end{aligned}$$

と置く．

仮定 A1 から A8 が満たされていると，以下のことが成り立つ．

定理 5.1 (Norman の定理)

(a) $x \in I$ と $n \geq 0$ に関し一様に $\omega_n^\theta(x) = O(\theta)$.

(b) 任意の $x \in I$ に対し，次の（ベクトル）微分方程式

$$\frac{d}{dt}f(t) = w[f(t)], \quad f(0) = x$$

の解は唯一であり，$\mu_n^\theta(x) = f(n\theta) + O(\theta)$ が $x \in I$ と $n \geq 0$ に関して一様に成り立つ．

(c) 次の（行列）微分方程式

$$\frac{d}{dt}\Sigma(t) = J[f(t)]\Sigma(t) + \Sigma(t)J[f(t)]^T + C[f(t)], \quad \Sigma(0) = 0$$

は唯一の解 $\Sigma(t)$ を持つ．

(d) $(X_n^\theta - f(n\theta))/\sqrt{\theta}$ の分布は $\theta \downarrow 0$，$n\theta \to t \leq \infty$, の時，平均0，共分散行列 $\Sigma(t)$ の正規分布に弱収束する．特に $n \to \infty$ のときには $f(n\theta) \to x_{sol}$ で$\Sigma(\infty)$ は次の行列方程式を解くことにより得られる．

$$J(x_{sol})\Sigma(\infty) + \Sigma(\infty)J(x_{sol})^T + C(x_{sol}) = 0. \ \square$$

図 1

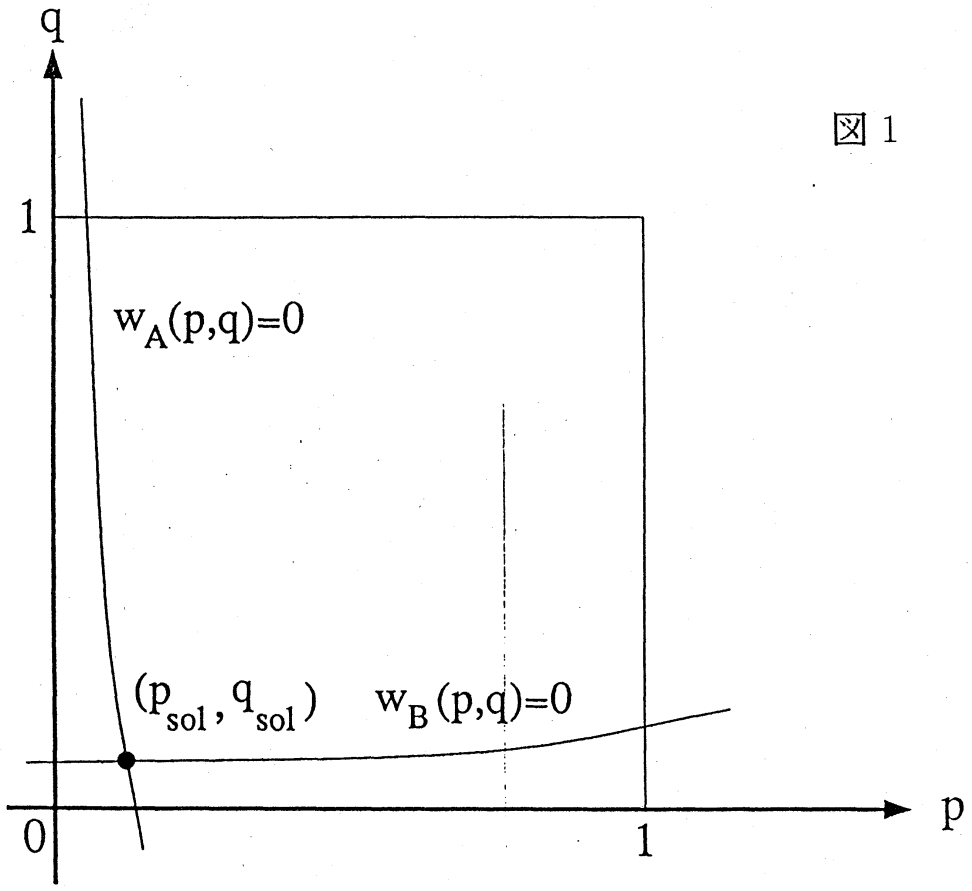
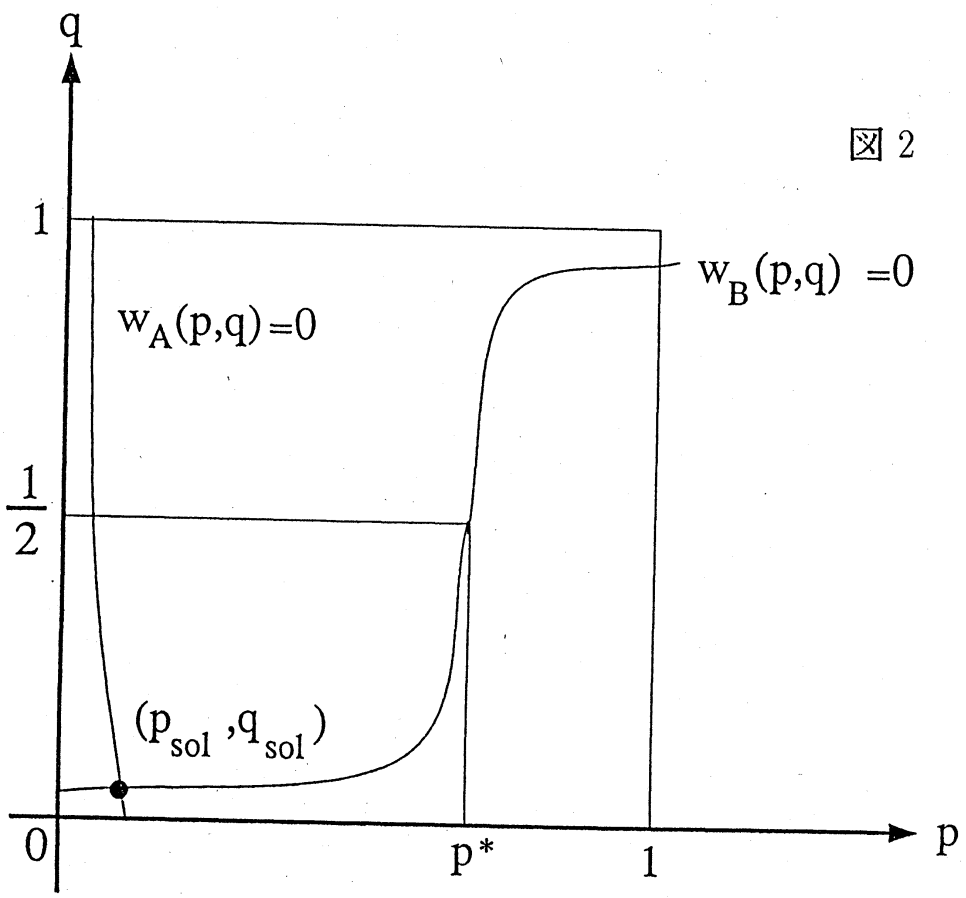

q
1
$w_A(p,q)=0$
(p_{sol}, q_{sol})
$w_B(p,q)=0$
0
1
p

図 2


q
1
$w_B(p,q)=0$
$w_A(p,q)=0$
$\frac{1}{2}$
(p_{sol}, q_{sol})
0
p*
1
p

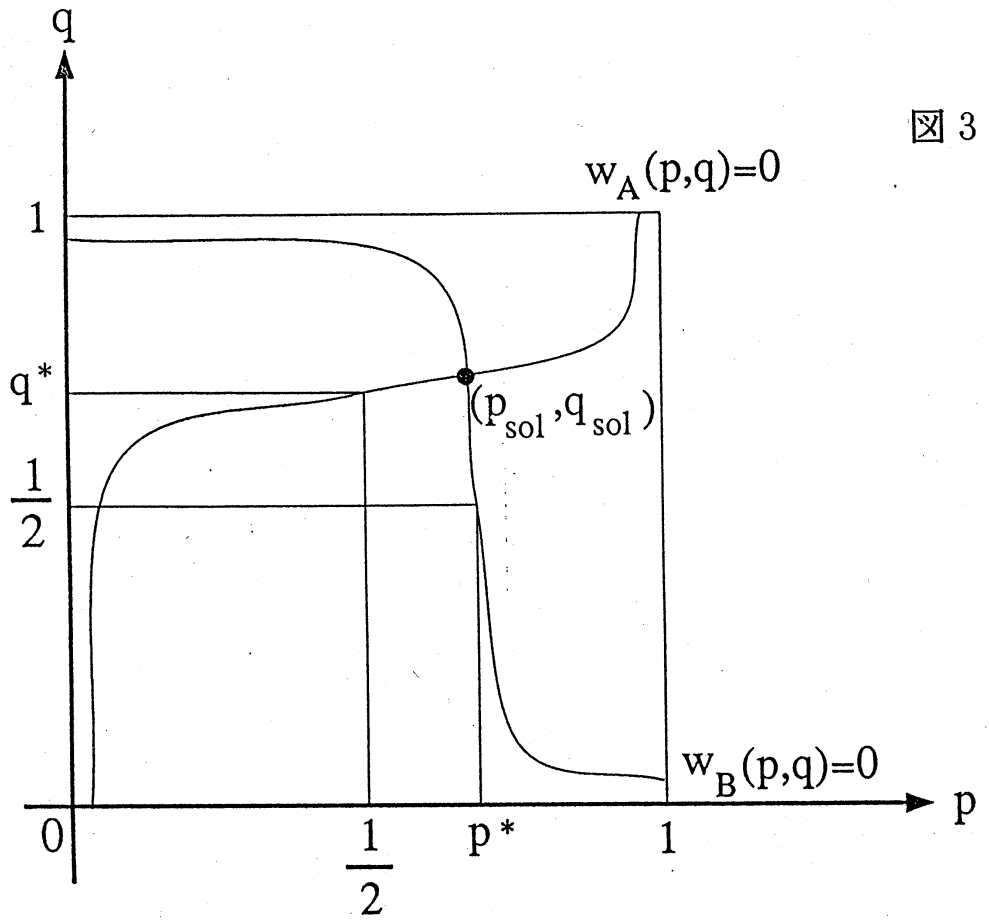

q
1
q*
1/2
0
1/2
p*
1
p
$w_A(p,q)=0$
$w_B(p,q)=0$
(p_{sol},q_{sol})

図 3

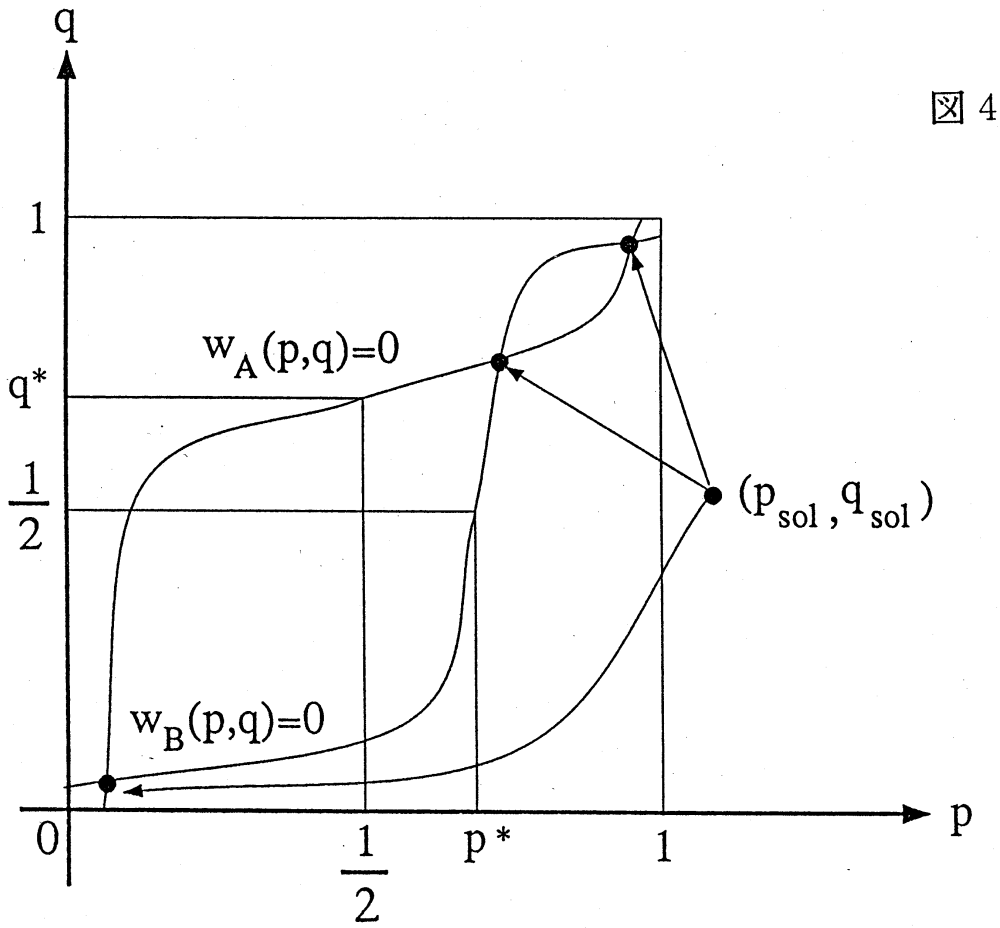

q
1
q*
1/2
0
1/2
p*
1
p
$w_A(p,q)=0$
$w_B(p,q)=0$
(p_{sol},q_{sol})

図 4