Journal of the Operations Research
Society of Japan
Vol. 41, No. 4, December 1998

反応曲線が既知なロブ-パス問題の最適解

平岡和幸　　吉澤修治
東京大学

(受理 1996 年 10 月 22 日；再受理 1998 年 5 月 18 日)

和文概要　心理学において,「慣れ」や「飽き」のように，同じ選択を続けると効果が悪くなる現象を記述する，ロブ-パス問題と呼ばれるモデルがある．Abe and Takeuchi は，この問題をオンライン学習問題として定式化し，それが multi-armed bandit 問題の拡張とみなせる事を指摘した．古典的な bandit 問題との違いは，プレイヤーの選択が環境自体に影響を与え，環境を変化させてしまうという点にある．

学習問題としてのロブ-パス問題に対してこれまでに提案された戦略は，すべて基本的に,「未知環境からの反応をもとに，その環境に対する最適 "定常" 戦略を推定し，その戦略に従って選択肢を選ぶ」ということを繰り返すものである．また，戦略の評価には，環境が既知だった場合の最適 "定常" 戦略と比較して，実際には環境が未知な事によるロスが，どの程度におさまるかを基準としている．

このような方針が妥当かどうかを判断するためには，環境が既知だった場合の (定常とは限らない) 最適戦略を知っておく必要がある．本論文はこれを導出する．その系として，従来研究で仮定されていた「マッチング条件」が，最適戦略が打ち切り時刻によらないための必要十分条件となっている事を指摘する．これにより，目標として "定常" 戦略のみを考えることの正当性が保証されることになる．マッチング条件自体の意味や妥当性に関する議論も行う．さらに，漸近最適性を定義し，忘却ありの相手なら定常戦略が漸近最適となるが，忘却なしなら漸近最適戦略は存在しない事を示す．

1. はじめに

実世界においては, 同じ選択を行っても, 周囲の状況 (環境) によって, 得られる結果は異なってくるのが普通である．しかも, 自分の選択自体が環境に影響を与え, その状態を変化させてしまうという場合が多い．このような場合には, 各瞬間瞬間ごとの利益を最大にする選択肢を選ぶのではなく, 環境を自分に有利な方向に誘導するなど, 未来の事まで考えた戦略を取らないと, 全体を通しての合計利益を最大化することはできない．ロブ - パス問題は, その典型的な状況をモデル化したものである．その原型は, 心理学において「慣れ」や「飽き」のように, 同じ選択を続けると効果が悪くなる現象を記述するために用いられていたモデルである [3]. Abe and Takeuchi[1] は，これをオンライン学習問題として定式化し, それが multi-armed bandit 問題の拡張とみなせる事を指摘した．環境が未知であり, 自分の選択に応じて得られる結果を通じて環境を探りながら, しかも全体としての合計利益を最大化するような選択を行っていかねばならない, というのが彼らの問題設定である．古典的な bandit 問題との違いは, プレイヤーの選択が環境に影響を与え, 環境を変化させてしまうという点にある.

オリジナルのロブ - パス問題は, 次の通りである [3]. (本論文での設定は,2章で述べる.)「プレイヤー」と「相手」の 2 名による, 仮想的なテニスの試合を考える. 毎時刻 ($t = 1, 2, 3, \ldots$) ごとにプレイヤーは「ロブ」か「パス」かを選択する．これに対し相手は, プレイヤーがこれまでにロブを選んだ割合を調べていて，それに応じた準備をして待っている．相手がロブ率

を s と推定している時, プレイヤーがロブを出すと勝つ確率は $f_L(s)$, パスを出すと勝つ確率は $f_P(s)$ であるとする.(1回のショットで勝負がつくとし, ラリーが続いたりはしない.) この f_L,f_P を反応曲線と呼ぶ. プレイヤーは勝てば利益1をもらえ, 負ければ利益は0である. このようなゲームを $t = 1, 2, 3, \ldots$ と続けていく時, 合計利益の期待値をできるだけ大きくするには, 各時刻毎にプレイヤーはどのようにロブかパスかを選択すれば良いか.

Abe and Takeuch[1] では「反応曲線 f_L,f_P は未知であり, ゲームを通じて学習しなくてはならない」という点に注目して, Lob-Pass 問題がオンライン学習問題として定式化され, それが Multi-armed Bandit 問題を発展させたものとみなせることが指摘されている. この文献では, 反応曲線 f_L,f_P が線形な場合に対していくつかの戦略が提案され, 様々な設定における性能が調べられている. Kilian et al. [6] は, 同じく線形な反応曲線に対し, ノイズトレラントなバイナリーサーチを応用した戦略を構成し, [1] より弱い前提のもとでも有効に機能することを証明した. ただし [1][4] とは戦略の評価基準が少し違っているうえ, s のダイナミクスが考慮されていない. また,Kilian らは, 単純な greedy algorithm も検討している. これに関しては, ゲームを十分長く続けたときの漸近的な性能が, オーダーとしては理論的限界値を達成しているようだというシミュレーション結果が報告されている. Hiraoka and Amari[4] では, 反応曲線が非線形でノンパラメトリックな場合に対して, 確率近似法を応用した戦略が解析されている.

以上の戦略はすべて, 基本的には, 「未知環境からの反応をもとに, その環境に対する最適 "定常" 戦略を推定し, その戦略に従って選択肢を選ぶ」ということを繰り返すものである. また, これらの論文では, 与えられた戦略 π の評価方法としては,(反応曲線 f_L,f_P を知っているとして)"定常" 戦略の中で最適な戦略の性能に比べて, (反応曲線を知らない) π の性能がどの程度劣っているか, というロスを基準としている. しかし, このような方針が妥当かどうかを判断するためには, 環境が既知だった場合の (定常とは限らない) 最適戦略を知っておく必要がある.

さらに, これらの論文では, 提案されたアルゴリズムがうまく機能するために, 反応曲線 f_L,f_P に関してある条件 ([1][6] ではマッチングショルダー条件, [4] では, その非線形版であるマッチング条件) が仮定されている. しかし, この条件が満たされない場合にどのような現象が生じるかはあまり議論されていない. この条件が, 最適戦略が打ちきり時刻に依存しないための必要条件である事が, [4] で指摘されている程度である.

本論文では, 反応曲線 f_L,f_P が既知の場合の, 厳密な最適戦略を求める. その系として, マッチング条件は, 最適戦略が打ちきり時刻に依存しないための必要十分条件であることが導かれる. この結果, 上のように "定常" 戦略に限定した中で最適な戦略を考える事に根拠が与えられる. また, マッチング条件自体の意味や妥当性についても議論する. マッチング条件が成立していない場合には, 打ち切り時刻に応じて最適戦略は異なったものになる. そのため, 打ち切り時刻が特に定まっていない時の戦略を議論する際には, 「最適」性の定義に注意が必要である. 本論文では, 漸近最適性を定義し, 忘却ありなら定常戦略が漸近最適となるが, 忘却なしの時は漸近最適戦略は存在しない事を示す.

なお, 離散性ゆえに生じる繁雑さをさけ, 本質的な現象に注目するために, 本論文では時間が連続な場合を扱う. このような, 反応曲線既知, 連続時間のロブ - パス問題は, 入力の大きさに制限がある時の, 一次遅れ系の制御問題へと書き換えられる.(ただし, 通常の制御問題とは目的関数が異なっている.) 従って, 変分法により最適解を求めることができる.

2章で, 本論文に用いる記号を定義し, 扱う問題を述べる. そして, 一般の戦略を扱う前に, 準備として定常戦略について調べる. 3章が本論文の主要な結果である. この章では, 相手がロブ

率を推定する際に, 忘却がありの場合, なしの場合のそれぞれに対して, 厳密な最適戦略を示す. 4章では, ゲームの打ち切り時刻が特に定められていない場合に関して議論する. まず,[1] で導入されたマッチング条件が, 最適戦略が打ち切り時刻によらないための必要十分条件となっている事を 4.1節で指摘する. 次に,4.2節で漸近最適性を定義し, 定常戦略の漸近最適性 / 非最適性を調べる. 5章では, 本論文で仮定した条件の妥当性や, オリジナルのロブ - パス問題との相違について議論する. 本文中で示す最適戦略の導出は, 付録 A,Bで行われる.

2. 問題設定

この章では, 本論文で用いる記号の定義を与え, ロブ - パス問題を定式化する. その後, 一般の戦略を議論する準備として, 定常戦略について述べる.

2.1. ロブ - パス問題

本論文で扱うのは, 連続時間, 既知反応関数を持つロブ - パス問題である. 前述のように, 「プレイヤー」と「相手」の 2 名による, 仮想的なテニスの試合を考える. 時刻 t における, プレイヤーの選択を $r(t)$, 相手が推定しているロブ率を $s(t)$ で表す. 各時刻 $t \in \mathrm{R}$ ごとに, プレイヤーはその瞬間のロブ率 $r(t)$ を選択する. 時間を連続化しているので, プレイヤーは $0 \leq r(t) \leq 1$ の任意の値を選択することができるとする. これは, ロブ 70%, パス 30% のような混合戦略をとることができるということである (5章参照). 関数 $r(\cdot)$ は区分的に連続であるとする.

一方, 相手の方はある決まったアルゴリズムで, 過去の結果に基づいて, プレイヤーのロブ率 $r(t)$ を推定している. そのアルゴリズムとしては, 本論文では次の 2 通りの場合を考える.

$$s(t) = \begin{cases} s_0 e^{-(t-t_0)} + \displaystyle\int_{t_0}^{t} r(\tau) e^{-(t-\tau)} d\tau & (\text{忘却あり}) \\ \dfrac{1}{t}\left(t_0 s_0 + \displaystyle\int_{t_0}^{t} r(\tau) d\tau \right) & (\text{忘却なし}) \end{cases}$$

ここに, $s(t_0) = s_0$ はゲーム開始時刻 t_0 における初期状態である. 微分形で書くと

$$\frac{d}{dt} s(t) = \begin{cases} r(t) - s(t) & (\text{忘却あり}) \\ \dfrac{1}{t}(r(t) - s(t)) & (\text{忘却なし}) \end{cases} \tag{2.1}$$

となる. つまり, 忘却なしでは, 過去のロブ率の平均値が $s(t)$ であり, 忘却ありでは, 遠い過去の影響が小さくなるよう重みをつけて平均をとったものが $s(t)$ である. $r(\cdot)$ が区分的に連続であるとしたので, $s(\cdot)$ は連続であり, さらに区分的に C^1 級となる. この $s(t)$ は, 時刻 t における相手の「状態」とみなすことができる. なお, 忘却ありの場合, 一般には忘却因子 $\beta > 0$ を明示して

$$\frac{d}{dt} s(t) = \beta(r(t) - s(t))$$

とすべきだが, 時間軸のスケールを取り直し, βt を改めて t とおけば, 忘却因子 $\beta = 1$ の場合に帰着される.

ある瞬間に, プレイヤーの選択が r, 相手の状態が s であったら, プレイヤーは

$$w(r, s) \equiv r f_L(s) + (1 - r) f_P(s)$$

という利益を得る. 反応曲線 $f_L(s)$,$f_P(s)$ はそれぞれ, 相手の状態が s の時, プレイヤーがロブ $(r = 1)$, またはパス $(r = 0)$ を選んだ場合の利益である (図 1). 本論文では, 反応曲線 $f_L(s)$,$f_P(s)$

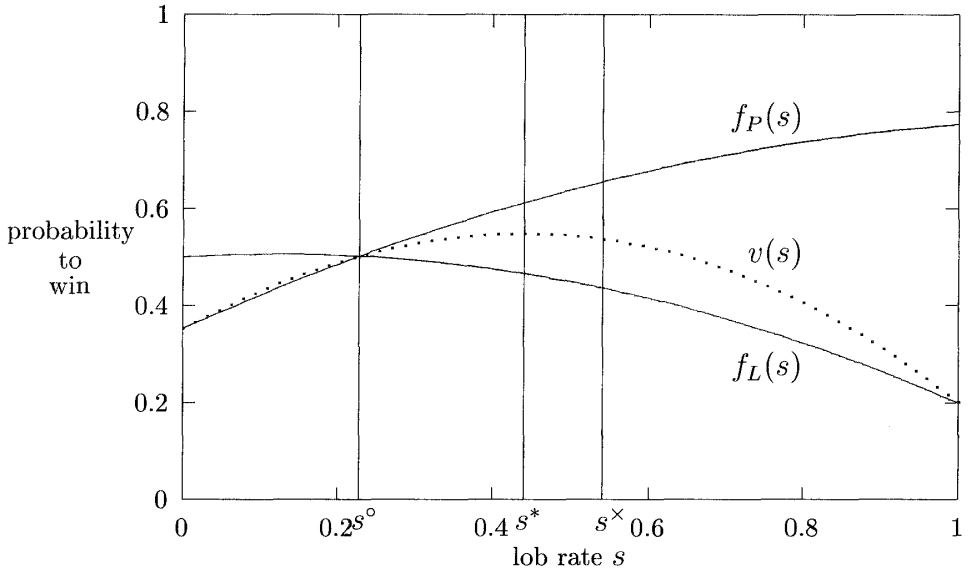

図 1: 反応曲線 $f_L(s)$, $f_P(s)$

が既知の場合を扱う. プレイヤーの目標は, 合計利益

$$G_T[r] \equiv \int_{t_0}^{T} w(r(t), s(t))dt$$

を最大にするよう $r(t)$ を選ぶことである. ゲームの開始時刻 t_0, 終了時刻 T, および相手の初期状態 $s(t_0) = s_0$ は前もって知らされているとする.

反応曲線 f_L,f_P に関しては,C^2 級であること, 及び以下の条件を仮定する.

仮定 1 (単調性)

$$f_L'(s) < 0, \quad f_P'(s) > 0 \quad (\text{for } 0 < s < 1) \tag{2.2}$$

仮定 2 (非自明性)

$$f_L(0) > f_P(0), \quad f_L(1) < f_P(1) \tag{2.3}$$

これらの条件は, 大雑把に言うと次のようなことを主張している: 単調性については, 相手が推定している「プレイヤーがロブを出す確率」が s であるから, それが高いほど, ロブを出した時に得られる利益が少なくなると言うことである. そもそもそういった現象 (同じ選択を続けると効果が下がる, 慣れ・飽きなど) をモデル化したのがロブ - パス問題だったのだから, 単調性は当然の仮定である. また非自明性は, $s = 0, 1$ の両極端の状態では, それぞれ, パス, ロブの方が勝ちやすいということである. もし (2.3) が成り立っておらず, 例えば $f_L(0) \leq f_P(0)$ だったら, 相手が「ロブは来ない」と確信している時にロブを出してもパスを出すより少ない利益しか得られないのであるから, プレイヤーがロブを出すべき場合が全くなくなってしまう. 即ち, 単調性からすべての s で $f_L(s) \leq f_P(s)$ となり, 常に $r(t) = 0$ が最適戦略であるという自明な結果となってしまう.

さらに, 技術的な理由から, 以下を仮定する.

仮定 3 (単峰性) s の関数 $w(s, s)$ は単峰である. すなわち, $dw(s, s)/ds$ は, ある点 s^* を境に

$$\frac{d}{ds}w(s, s) > 0 \ (s < s^*), \quad = 0 \ (s = s^*), \quad < 0 \ (s > s^*) \tag{2.4}$$

となっている.

仮定 4 (符号の単調性 (忘却なしの場合)) s の関数 $(\partial/\partial s)w(r,s)|_{r=s}$ は, ある点 $s^\times$ を境に

$$\left.\frac{\partial w(r,s)}{\partial s}\right|_{r=s} \geq 0\ (s<s^\times),\quad =0\ (s=s^\times),\quad \leq 0\ (s>s^\times) \tag{2.5}$$

となっている. この条件は忘却なしの場合のみ仮定する.

なお, [1][6] のように反応曲線が線形のときは, 「単調性」さえ満たされていれば, 自動的に「単峰性」「符号の単調性」も成立する. また, 「単峰性」「符号の単調性」の妥当性に関する別の解釈は, 5.1節で議論される.

まとめると, 本論文で解く問題は,

> $0 \leq r(t) \leq 1$ なる, 区分的に連続な関数 $r(\cdot)$ をうまく決めて, $G_T[r] = \int_{t_0}^{T} w(r(t), s(t))dt$ を最大化せよ. ここに
>
> $$\begin{aligned} s(t_0) &= s_0, \quad \frac{d}{dt}s(t) = \begin{cases} r(t)-s(t) & \text{(忘却ありの場合)} \\ \dfrac{1}{t}(r(t)-s(t)) & \text{(忘却なしの場合)} \end{cases} \\ w(r,s) &= rf_L(s)+(1-r)f_P(s) \end{aligned}$$
>
> f_L,f_P は既知で, 条件 (2.2)(2.3)(2.4) および (2.5)(忘却なしの場合) を満たしている.

となる. 関数 $r(\cdot)$ をどう決めるかが, プレイヤーの戦略である. これは, 一次遅れ系に対する一般の最適制御問題の形であり, 変分法により解くことができる. ただし, 通常の制御問題とは, 目的関数 w が異なっている.

なお, オリジナルのロブ - パス問題との相違点に関しては, 5.2節で議論する.

2.2. 最適定常戦略と瞬間利益最大化戦略

一般の戦略を議論する前に, 準備としてまず定常戦略について検討しておく. 定常戦略とは, $r(t) = r_0$ という, ある一定の選択肢を選び続ける戦略のことである. このとき相手の状態 s は r_0 に漸近するので, 十分時間がたてば瞬間利益は,

$$v(r_0) \equiv w(r_0, r_0) = r_0 f_L(r_0) + (1-r_0)f_P(r_0)$$

に漸近する. このとき, 単峰性条件 (2.4) から, s^* が, 定常戦略の中では (長い時間ゲームを行うなら) 最適なロブ率となる (図 1). そこで, s^* を「最適定常ロブ率」と呼ぶことにする.

一方, 各瞬間ごとにその瞬間の利益を最大化する瞬間利益最大化戦略 (その日暮らし戦略) は, 「$f_L(s(t)) > f_P(s(t))$ なら $r(t)=1$, $f_L(s(t)) < f_P(s(t))$ なら $r(t)=0$」である. 今, $f_L(s) = f_P(s)$ となる点 (非自明性 (2.3) と単調性 (2.2) から唯一存在する) を $s = s^\circ$ とおくと, 瞬間利益最大化戦略は, 「$s(t) < s^\circ$ なら $r(t)=1$, $s(t) > s^\circ$ なら $r(t)=0$」とも書ける. この s° を「マッチング点」と呼ぶことにしよう. $s(t)$ は $r(t)$ に追従するから, 瞬間利益最大化戦略を続けると $s(t) \to s^\circ$ となり, 各瞬間の利益は $v(s^\circ) = f_L(s^\circ) = f_P(s^\circ)$ になる.

一般には $s^* \neq s^\circ$ であり, 瞬間利益最大化戦略は, 累積の合計利益に関しては最適定常戦略に劣る [1][3][4]. 以上の議論は, 忘却ありでもなしでも成立する.

3. 厳密な最適戦略

3.1. 忘却ありの場合

この節では, 相手が忘却ありの場合の厳密な最適戦略を述べる.

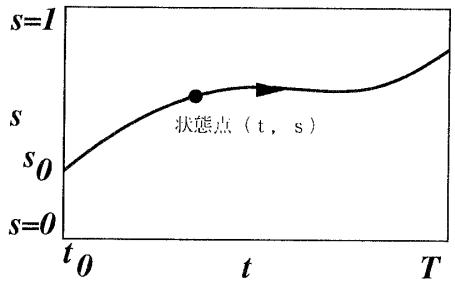

図 2: ゲームの状態点 (t, s)

ゲームの状態は, 現在の時刻 t とその時の相手の状態 $s(t)$ の組で記述されるから, t-s 平面上の一点で表現される (図 2). この「状態点」は時刻 t を進めるにつれて図の右方向へ推移してゆく. その際, $r(t) > s(t)$ であれば状態点は図の右上方向に進み, $r(t) < s(t)$ であれば右下方向に進む. $r(t) = s(t)$ であれば, 状態点は右へ水平に進む. $r(t) = s(t) + ds(t)/dt$ という関係があるから, 制約 $0 \leq r(t) \leq 1$ は, 状態点の軌跡の傾きがある範囲の値 (s に依存する) しか取れないということを意味している.

次の定理の証明は, 付録 A で述べる.

定理 1 *(2.1)* において, 忘却ありとし, 条件 *(2.2)(2.3)(2.4)* を仮定する. このとき, 最適戦略は *bang-bang* コントロールの形である. 即ち, t-s 平面が 2 つの領域 L,P に分割され, 領域 L では $r = 1$*(ロブを出す)*, 領域 P では $r = 0$*(パスを出す)* とするのが最適となる. ここに

- $s^* \geq s^\circ$ の場合*:*

$$L = \{(t, s) | s < s^* \text{ かつ } D_P(t, s) \geq 0)\} \tag{3.1}$$

- $s^* < s^\circ$ の場合*:*

$$P = \{(t, s) | s > s^* \text{ かつ } D_L(t, s) \geq 0\} \tag{3.2}$$

であり, 他方はその補集合として決定される. D_L,D_P は, それぞれ

$$\begin{aligned} D_L(t, s) &= v(s) - f_L(1 - (1 - s)e^{-(T-t)}) \\ D_P(t, s) &= v(s) - f_P(se^{-(T-t)}) \end{aligned}$$

と定義する. ただし, L と P の境界線上で $s = s^*$ の部分では, $r = s^*$ とする. *(図 3)* □

例えば,

$$f_L(s) = -a_L s + b_L, \quad f_P(s) = a_P s + b_P \tag{3.3}$$

という線形の場合には

$$s^* = \frac{a_P + b_L - b_P}{2(a_L + a_P)}, \quad s^\circ = \frac{b_L - b_P}{a_L + a_P}$$

であり, $s^* \geq s^\circ$ (すなわち, $b_L - b_P \leq a_P$) なら, 曲線 $D_P = 0$ は,

$$s = \frac{(1 - e^{-(T-t)})a_P + b_L - b_P}{a_L + a_P}$$

という指数関数の形になる. 具体例として,

$$a_L = 0.2, \quad a_P = 0.8, \quad b_L = 0.4, \quad b_P = 0.1, \quad T = 10 \tag{3.4}$$

という設定でこれをプロットすると図 4になる. 点線が $s = s^*$, 実線が曲線 $D_P = 0$ である.

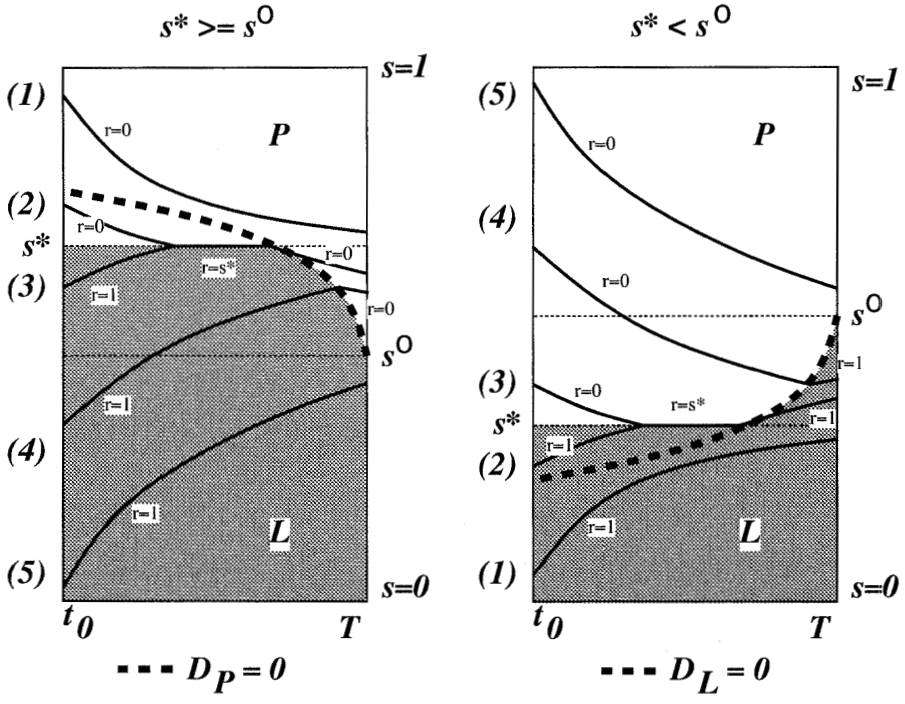

図 3: 忘却ありの場合の最適戦略

3.2. 忘却なしの場合

忘却ありの時と同様にして, 次の定理が得られる. なお, ロブとパスは対称なので, $s^\times \geq s^\circ$ と仮定して一般性を失わない. 実際, $s^\times < s^\circ$ のときは,

$$R = 1 - r, \quad S = 1 - s, \quad F_L(S) = f_P(s), \quad F_P(S) = f_L(s)$$

とおいて, (r, s, f_L, f_P) のかわりに (R, S, F_L, F_P) を考えれば, $s^\times \geq s^\circ$ の場合に帰着される.

定理 2 *(2.1)* において, 忘却なしとし, 条件 *(2.2)(2.3) (2.4)(2.5)* を仮定し, $s^\times \geq s^\circ$ とする. このとき, 最適戦略は *bang-bang* コントロールの形である. 即ち, t-s 平面が 2 つの領域 L,P に分割され, 領域 L では $r = 1$*(*ロブを出す*)*, 領域 P では $r = 0$*(*パスを出す*)* とするのが最適となる. 切替え曲線 *(*L と P の境界線*)*$s_c(t)$ は,

- $t_0 \leq t \leq t^\times$ では

$$s_c(t) = s^\times \tag{3.5}$$

- $t^\times < t \leq T$ では

$$\begin{cases} \dfrac{ds_c}{dt} = \dfrac{f_P'(s_c(t)t/T)}{\dfrac{\partial w(r,s)}{\partial s}\Big|_{r=s=s_c(t)} - f_P'(s_c(t)t/T)} \dfrac{s_c(t)}{t} \\ s_c(T) = s^\circ \end{cases} \tag{3.6}$$

により決定される. ここに $t^\times$ は, 微分方程式 *(3.6)* の解が $s_c = s^\times$ と交わる時刻であり,

$$\int_{t^\times/T}^{1} f_P'(s^\times \zeta) \frac{d\zeta}{\zeta} = f_P(s^\times) - f_L(s^\times) \tag{3.7}$$

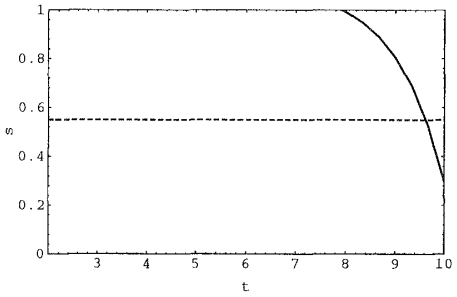

図 4: 忘却ありの場合の切り替え曲線 (数値計算例). 点線: $s = s^*$, 実線: $D_P = 0$

という関係からも求められる. この曲線より s が大きい側が P, 小さい側が L である. ただし, L と P の境界線上で $s = s^\times$ の部分, 即ち線分 $\{(t, s^\times) \mid t_0 \le t \le t^\times\}$ 上では, $r = s^\times$ とする. □

この結果を図示すると, 図5左のようになる. (3.3)(3.4) の設定で計算した具体例は図5右である.

4. 打ち切り時刻の影響

ここまでで述べたように, 最適戦略は, 打ちきり時刻までの残り時間に依存する非定常なものとなる. 従って, 打ちきり時刻 T が未知の場合には, 厳密な意味での (どんな T に対しても一様に) 最適な戦略は存在しない. この点を克服するには, 前提条件を強める, 「最適」の意味を弱める, 戦略の範囲を限定する, T に事前分布を仮定する, といった対策が考えられる. この事情は, 統計的推定論において, 「最適」な推定量を定義する際の話と類似している. 本論文では

1. 一様な最適が存在するよう, 反応曲線 f_L,f_P に条件を追加する.
2. T が 1 より十分大きい時の「漸近最適性」を考える.

という 2 通りのアプローチを考察する. 4.1節では, 最適戦略が打ちきり時刻に依存しなくなるような, 「マッチング条件」と呼ぶ条件について述べる. 4.2節では漸近最適性を定義し, 忘却ありの時は定常戦略が漸近最適なこと, 忘却なしの時は漸近最適戦略が存在しない事を示す.

4.1. マッチング条件

マッチング条件を述べる前に, これまで定義した 3 点 $s^\circ, s^\times, s^*$ についてまとめておく. 最適定常ロブ率 s^* は, $v(s) = w(s,s)$ が最大となる点であった. マッチング点 s° は $f_L(s) = f_P(s)$ となる点であり, 瞬間利益最大化戦略と関係していた. 3 点はそれぞれ,

$$s^\circ \quad : \quad \left.\frac{\partial}{\partial r} w(r,s)\right|_{r=s} \tag{4.1}$$

$$s^\times \quad : \quad \left.\frac{\partial}{\partial s} w(r,s)\right|_{r=s} \tag{4.2}$$

$$s^* \quad : \quad \frac{d}{ds} w(s,s) \tag{4.3}$$

の零点として定義された. 次の事実は容易に示される.

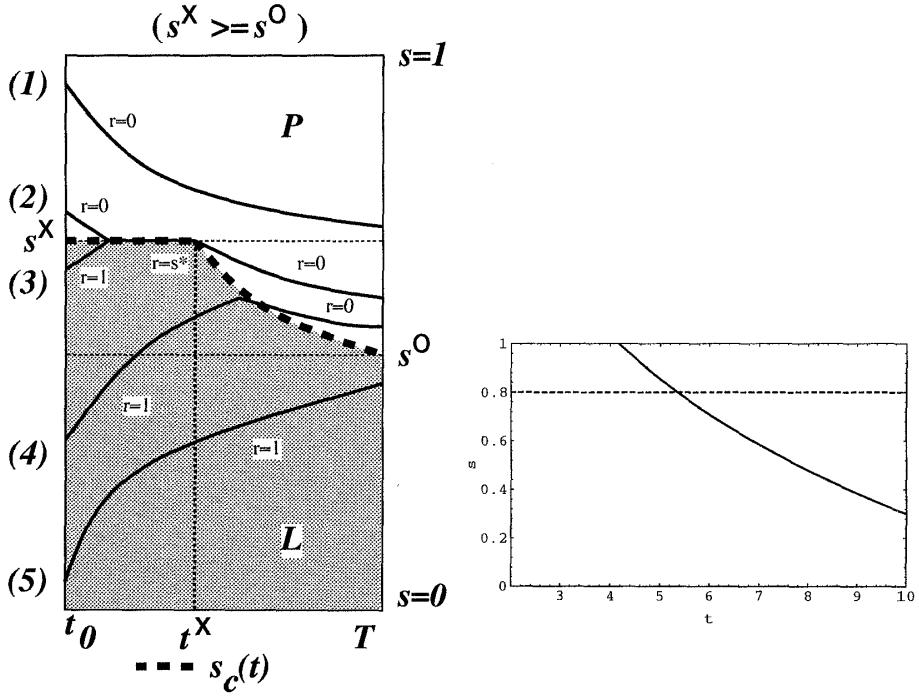

図 5: 忘却ありの場合の最適戦略 (左) と数値計算例 (右)(点線: $s = s^*$, 実線: $D_P = 0$)

補題 1 条件 *(2.2)(2.3) (2.4)(2.5)* を仮定すると, s^* は必ず s° と $s^\times$ の間にある. 特に, 反応曲線 f_L, f_P が線形の場合, s^* は s° と $s^\times$ の中点である. □

さて, マッチング条件とは, これらが一致するという主張である.

条件 1 (マッチング条件)

$$\left.\frac{\partial}{\partial r}w(r,s)\right|_{r=s=s^*} = \left.\frac{\partial}{\partial s}w(r,s)\right|_{r=s=s^*} = 0 \tag{4.4}$$

ここで実は,

$$\frac{d}{ds}w(s,s) = \left.\frac{\partial}{\partial r}w(r,s)\right|_{r=s} + \left.\frac{\partial}{\partial s}w(r,s)\right|_{r=s}$$

だから,

$$\left.\frac{\partial}{\partial r}w(r,s)\right|_{r=s=s^*} = 0 \quad \text{か} \quad \left.\frac{\partial}{\partial s}w(r,s)\right|_{r=s=s^*} = 0$$

の片方が成り立てば, 他方も成り立つ, すなわちマッチング条件がなりたつことがすぐにわかる.

文献 [1] では, 反応曲線が線形の場合に対して, ある推定がうまくいくための条件としてマッチングショルダーと呼ばれる条件が導入された. 上で述べたマッチング条件は, [4] において, 最適戦略が定常となる必要条件として導入されたものである. マッチング条件はマッチングショルダー条件を非線形な反応曲線の場合に拡張したもので, 反応曲線が線形なら両者は一致する.

3.1,3.2節で示した最適戦略の形から, ただちに次のことがわかる.

定理 3 最適戦略がゲームの打ち切り時刻 T に依存しないための必要十分条件は，マッチング条件 *(4.4)* である．その時の最適戦略は，忘却ありの場合も，なしの場合も，「$s(t) < s^*$ ならロブを出し $(r(t) = 1)$, $s(t) > s^*$ ならパスを出す $(r(t) = 0)$. $s(t) = s^*$ なら $r(t) = s^*$.」となる．

□

学習問題としてのロブ - パス問題 (反応曲線 f_L,f_P が未知) に関するこれまでの論文 [1][4][6] ではすべて，マッチング条件が仮定されている．そして，提案されている戦略はすべて基本的に，最適 "定常" ロブ率 s^* を推定してロブ率を s^* に近づけていく，という形をしている．さらに，戦略の評価には，最適 "定常" 戦略 $r(t) = s^*$ と比較しての性能のロスが基準として用いられている．このような方針を取ることの根拠は，定理 3によって与えられる．なぜなら，$s(t)$ のダイナミクス (2.1) から，定理 3のような戦略を続ければ有限時間で $s(t) = s^*$ となり，以降は $r(t) = s(t) = s^*$ となる．すなわち，ある時刻以降は最適定常戦略と同じになるのである．(論文 [1][4][6] は，打ち切り時刻 $T \to \infty$ での漸近的な性質を議論している.)

4.2. 漸近最適性

前節では，最適戦略が時間に依存しないよう条件を付加することを考えた．この節では，それとは別のアプローチとして，打ち切り時刻 T が 1 より十分大きい場合の漸近的な振舞いによる評価を検討する．

打ち切り時刻 T が特に定められていない場合，得られた利益を，ゲームを行った時間でノーマライズして，

$$\bar{G}_T[r] \equiv \frac{1}{T - t_0} G_T[r] = \frac{1}{T - t_0} \int_{t_0}^{T} w(r(t), s(t)) dt$$

という量を導入するのは自然である．上に述べたように，どんな T に対しても $G_T[r]$ が最大となるような，T に依存しない同一の戦略 $r(\cdot)$ は，一般には存在しない．しかし，T が十分大きい場合を考えると，$\bar{G}_T[r]$ が (各 T ごとの) 最大値に漸近するような戦略 $r(\cdot)$ なら存在する場合がある．そこでまず，$\bar{G}_T^{\mathrm{opt}} \equiv \max_{r(\cdot)} \bar{G}_T[r]$ と置く．$\bar{G}_T^{\mathrm{opt}}$ は，本論文で求めた最適戦略による，単位時間あたりの平均利益である．そして，$\bar{G}_T^{\mathrm{opt}} - \bar{G}_T[r_1] \to 0$ as$T \to \infty$ を満たす戦略 $r_1(\cdot)$ を「漸近最適である」ということにする．

定理 4

1. 忘却ありの場合は，最適定常戦略 $r(t) = s^*$ は漸近最適である．
2. 忘却なしの場合は，漸近最適戦略が存在するためには，マッチング条件 *(4.4)* が必要十分である．その時，最適定常戦略 $r(t) = s^*$ は漸近最適である．

□

証明の前に，忘却ありとなしでこのような定性的な差が生じる理由を説明しておく．打ちきり時刻 T が十分大きい場合，忘却なしの最適戦略は $r(t) = s(t) = s^{\times}$ という定常な部分を含む．この部分での瞬間利益は $v(s^{\times})$ であり，最適定常戦略 $r(t) = s^*$ による瞬間利益 $v(s^*)$ より劣っている．厳密な最適戦略が最適定常戦略よりまさっているのは，その後の $r(t) = 0$ $(s^{\times} > s^{\circ}$ の場合) の部分である．打ちきり時刻 T を大きくしていくと，この $r(t) = 0$ の部分の長さ (時間) は T に比例して大きくなる[1]．したがって，この部分の $\bar{G}_T$ への影響は $T \to \infty$ でも残る．一方，

[1] $T = T^{(1)}$ のときの (3.6) の解 $s_c(t) = s_c^{(1)}(t)$ と，$T = T^{(2)}$ のときの (3.6) の解 $s_c(t) = s_c^{(2)}(t)$ との間に，

$$s_c^{(2)}(t) = s_c^{(1)}\left(\frac{T^{(1)}}{T^{(2)}} t\right)$$

の関係が成立することは容易に確かめられる．

忘却ありの場合には, 厳密な最適戦略で定常な部分は $r(t) = s(t) = s^*$ である. つまり, この部分では最適定常戦略と一致する. しかも, その後の $r(t) = 0$ ($s^* > s^\circ$ の場合) の部分の長さは T を大きくしても一定のままである. (f_L,f_P を固定した時, 領域 L,P が $T-t$ にのみ依存する形であることに注意.) したがって, この部分の $\bar{G}_T$ への影響が, $T \to \infty$ では 0 になるのである.

定理 4の証明:

忘却ありの場合には, 定理 1および上の説明より $\bar{G}_T^{\mathrm{opt}} \to v(s^*)$ as $T \to \infty$ だから, 最適定常戦略 $r(t) = s^*$ が漸近最適なことは容易にわかる.

忘却なしの場合も, マッチング条件 (4.4) が成立していれば, 定理 3より $\bar{G}_T^{\mathrm{opt}} \to v(s^*)$ ($T \to \infty$) であり, 最適定常戦略 $r(t) = s^*$ は漸近最適である.

一方, マッチング条件不成立の場合には, B.2節と同様に $u = \log t, u_0 = \log t_0, U = \log T$ と変数変換し, $R(u) = r(e^u)$, $S(u) = s(e^u)$ とおく. すると,

$$\bar{G}_{T=e^u}[r] = \frac{1}{e^U - e^{u_0}} \int_{u_0}^{U} w(R(u), S(u)e^u) du = \frac{1}{1 - e^{-(U-u_0)}} \int_0^{U-u_0} w(R(U-\eta), S(U-\eta)) e^{-\eta} d\eta$$

となる. したがって,$u = U$ 付近での $S(u)$ が, 最適なもの (B.2節参照) と異なっていた場合, それによるロス $\bar{G}_{T=e^U}^{\mathrm{opt}} - \bar{G}_{T=e^U}$ は $U \to \infty$ でも 0 にならず残ってしまう. そして, マッチング条件不成立としたので最適戦略は U に依存している. そのため, 任意の $U \gg 1$ で一様に $u = U$ 付近での $S(u)$ を最適なものに漸近させるということはできない. ∎

5. 本論文の設定に関して

5.1. 反応曲線に関する仮定の妥当性について

本論文では, 反応曲線に関して単峰性などの条件を仮定した. また, マッチング条件というものも導いた. これらの条件の妥当性に関して, 一つの解釈を本節で述べる.

ロブ - パス問題を一旦忘れて, 次のような状況を考える. 「プレイヤー」と「相手」が, ある二人零和ゲームを行なっているとしよう. 相手が「プレイヤーの利益を最小化しよう」という意志を持って「手」を選ぶ点が, これまでの設定と異なっている. プレイヤーは「ロブ」「パス」の 2 通りから手を選び, 相手は $[0,1]$ の連続値から手 q を選ぶとする. そのときのプレイヤーの利益を, プレイヤーの手が「ロブ」のとき $F_L(q)$, プレイヤーの手が「パス」のとき $F_P(q)$ とする. プレイヤーが, 確率 r で「ロブ」, 確率 $(1-r)$ で「パス」, という混合戦略をとったときには, プレイヤーの利益は $W(r,q) \equiv rF_L(q) + (1-r)F_P(q)$ となる. ここで, 関数 F_L,F_P について, C^2 級であること, および

- 単調性: $F_L'(q) < 0, \quad F_P'(q) > 0$
- 非自明性: $F_L(0) > F_P(0), \quad F_L(1) < F_P(1)$
- 狭義下凸性: $F_L''(q) > 0, F_P''(q) > 0$

を仮定する. さて, 相手は, 何らかの方法で, プレイヤーの今回のロブ率 r を s と推定しているとしよう. 相手はこの推定を信じているため, ミニマックス利益ではなく, $r = s$ のときの期待利益を基準に手 q を決めるとする. その決め方を $Q(\cdot)$ という関数で表すことにし, 相手は $q = Q(s)$ という手を選ぶとする[2]. このとき, $F_L(Q(\cdot))$,$F_P(Q(\cdot))$ をそれぞれ $f_L(\cdot)$,$f_P(\cdot)$ とおき直せば, 推定値 s からロブ, パスの利益 $f_L(s)$,$f_P(s)$ が決まるというロブ - パス問題の設定と同じになる.

[2]プレイヤーのロブ率 r を固定した時, 狭義下凸性の仮定から, (相手が) 混合戦略をとるのは (相手にとって) 損なので, 純粋戦略のみを考える.

さて, 相手が意志を持ってゲームをしているのなら, 相手の戦略としては $Q(s) = \arg\min_q W(s,q)$ がまず考えられるであろう[3]. このとき, 次のことがなりたつ.

補題 2 $Q(s)$ は連続かつ単調非減少で, $Q(0) = 0$, $Q(1) = 1$ である. 特に, $0 < Q(s) < 1$ の領域では, 以下が成立する.

1. $Q'(s) > 0$
2. 与えられた s に対し, $\partial W(s,q)/\partial q = 0$ となる q が唯一存在する. この q が $Q(s)$ である.

□

証明:

$Q(0) = 0$, $Q(1) = 1$ は単調性から明らかである. さらに, 狭義下凸性から $\partial^2 W(s,q)/\partial q^2 > 0$ である. したがって, $Q(s) = 0, 1$ の場合以外は, 各 s に対し, $\partial W(s,q)/\partial q = 0$ なる q が唯一存在し, この q が $Q(s)$ となる:

$$\left.\frac{\partial}{\partial q} W(s,q)\right|_{q=Q(s)} = 0 \tag{5.1}$$

$0 < Q(s) < 1$ の時, $Q'(s) > 0$ であることを示すために,

$$\tilde{W}(s,q) \equiv \frac{\partial}{\partial q} W(s,q) = sF_L'(q) + (1-s)F_P'(q)$$

と定義する. すでに示された関係 $\tilde{W}(s,Q(s)) = 0$ の両辺を微分すると,

$$\frac{d}{ds}\tilde{W}(s,Q(s)) = \Big(F_L'(Q(s)) - F_P'(Q(s))\Big) + \Big(sF_L''(Q(s)) + (1-s)F_P''(Q(s))\Big)Q'(s) = 0$$

となる. よって, 単調性と狭義下凸性から,

$$Q'(s) = \frac{F_P'(Q(s)) - F_L'(Q(s))}{sF_L''(Q(s)) + (1-s)F_P''(Q(s))} > 0$$

が得られる. そのうえ, $W(s,q)$ は s に関して一様連続だから, q に関する狭義下凸性とあわせると, 最小点 $Q(s)$ が「飛び移る」ことはありえない. すなわち, $Q(s)$ は s に関して連続となっている. ∎

以上から, $f_L(s) = F_L(Q(s))$, $f_P(s) = F_P(Q(s))$ のグラフは, 図6のようになる. ここに, $Q(s) = 0$ をみたす最大の s を $s^\flat$ とし, $Q(s) = 1$ をみたす最小の s を $s^\sharp$ とする. $s \le s^\flat$ および $s \ge s^\sharp$ の領域では, $f_L'(s) = f_P'(s) = 0$ となる. $s^\flat < s < s^\sharp$ の領域では, $f_L'(s) < 0$, $f_P'(s) > 0$ である. さらに, (5.1) から, $w(r,s) = rf_L(s) + (1-r)f_P(s)$ に関して,

$$\begin{aligned}\left.\frac{\partial}{\partial s} w(r,s)\right|_{r=s} &= \left.\frac{\partial}{\partial s} W(r,Q(s))\right|_{r=s} \\ &= Q'(s)\left.\frac{\partial}{\partial q} W(s,q)\right|_{q=Q(s)} = 0 \qquad (\text{for all } s,\ 0 \le s \le 1)\end{aligned} \tag{5.2}$$

が成り立っている. ($s \le s^\flat$ および $s \ge s^\sharp$ では, それぞれ $Q(s) = 0$ および $Q(s) = 1$ と定数だから, $Q'(s) = 0$ なことに注意.) つまり, 符号の単調性条件 (2.5) は自明に満たされている. す

[3]相手はプレイヤーの戦略の「ダイナミクス」を知らないと想定している. そのため相手は, 「プレイヤーを誘導して長期的な合計利益を最小化する」といったことは図りようがなく, 瞬間利益を基準として戦略を決めているとする.

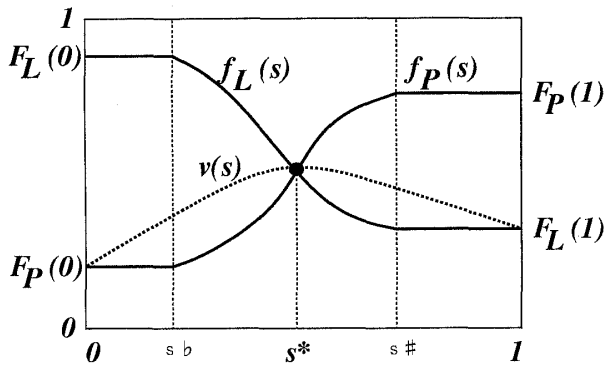

図 6: F_L,F_P から構成された反応曲線 f_L,f_P

ると, $v(s) = w(s,s)$ に関して $v'(s) = f_L(s) - f_P(s)$, $v''(s) = f_L'(s) - f_P'(s) > 0$ となり, 単峰性条件 (2.4) も自動的に満たされる. 特に, $v(s)$ の極大点 s^* は, $s^\flat < s^* < s^\sharp$ を満たす. さらに, $f_L(s^*) = f_P(s^*)$ だから, マッチング条件 (4.4) も自動的に満たされている.

ちなみに, (5.2) は強い制約であり, これが成り立っていると, 忘却ありの時の,B.1節の変分の一次の項が常に 0 となってしまう. このことは, ある戦略をとった時の利益が, $s(t_0)$ と $s(T)$ (T は打ち切り時刻) という両端点の値のみで定まり, その間の軌道によらないことを意味する. この場合の最適戦略は, 「打ち切り時刻の相手の状態 $s(T)$ が $s^*(= s^\circ = s^\times)$ にできるだけ近くなるように, できれば一致するようにする」という簡単なものになる. たとえば, 反応曲線 f_L,f_P は未知, 打ち切り時刻 T は既知, という設定では, ([1][4][6] のようなオーダー評価ではなく) 精密な解析が要求されるが, (5.2) の条件はそれを容易にすると期待される. このような場合については今後の研究課題である.

なお, 以上で構成した反応曲線 $f_L(s)$,$f_P(s)$ は, $s = s^\flat, s^\sharp$ において微分が不連続となっている. また, $s \le s^\flat$ および $s \ge s^\sharp$ では, $f_L'(s) = f_P'(s) = 0$ となっている. このため, 厳密には, 本論文の最初の仮定 (反応曲線が C^2 級, および単調性 (2.2)) が満たされていない. しかし, この f_L,f_P についても次の定理が言える.

定理 5 本節で構成した $f_L(s) = F_L(Q(s))$, $f_P(s) = F_P(Q(s))$ でも, 定理 3と同様の結果が成り立つ. すなわち, 忘却ありの場合も, なしの場合も, 「$s(t) < s^*$ ならロブを出し ($r(t) = 1$), $s(t) > s^*$ ならパスを出す ($r(t) = 0$). $s(t) = s^*$ なら $r(t) = s^*$.」が最適戦略となる. □

証明:

任意に小さい $\delta, \delta' > 0$ に対して, 条件 (2.2)(2.3) (2.4)(2.5) を満たす C^2 級関数 $\tilde{f}_L$,$\tilde{f}_P$ をとることができて,

$$|f_L(s) - \tilde{f}_L(s)| \begin{cases} = 0 & (s^\flat + \delta' < s < s^\sharp - \delta') \\ < \delta & (\text{他}) \end{cases}, \quad |f_P(s) - \tilde{f}_P(s)| \begin{cases} = 0 & (s^\flat + \delta' < s < s^\sharp - \delta') \\ < \delta & (\text{他}) \end{cases}$$

とできる. すると, 任意の戦略 $r(\cdot)$ をとったとき, 反応曲線 f_L,f_P による合計利益 $G_T[r]$ と反応曲線 $\tilde{f}_L$,$\tilde{f}_P$ による合計利益 $\tilde{G}_T[r]$ との差は, たかだか

$$\left|G_T[r] - \tilde{G}_T[r]\right| < T\delta \tag{5.3}$$

とおさえられる. 以上を用いて, 背理法で定理を証明する. 定理で主張されている最適戦略を $r^*(t)$ とおく. 今仮に, 別の戦略 $r(t)$ が, $G_T[r] > G_T[r^*]$ となっていたとしよう. このとき, $\Delta G = G_T[r] - G_T[r^*] > 0$ とおいて, $\delta = \Delta G/(3T)$, $\delta' = 0.5 \min\left(|s^\flat - s^*|, |s^\sharp - s^*|\right)$ と定義

する．ここに，s^* は，反応曲線 (f_L,f_P) のもとでの最適定常ロブ率である．この δ,δ' に対して，上述のような $\tilde{f}_L,\tilde{f}_P$ をとると，反応曲線 $(\tilde{f}_L,\tilde{f}_P)$ もやはりマッチング条件 (4.4) を満たし，最適定常ロブ率は同じ s^* となる．さて，(5.3) から，$\tilde{G}_T[r] > G_T[r] - (1/3)\Delta G$, $\tilde{G}_T[r^*] < G_T[r^*] + (1/3)\Delta G$ という不等式が成り立っている．この 2 つの不等式をあわせると $\tilde{G}_T[r] - \tilde{G}_T[r^*] > (1/3)\Delta G > 0$ という結果が得られる．しかしこれは定理 3と矛盾している．したがって，背理法により，$G_T[r] > G_T[r^*]$ となる戦略 $r(t)$ は存在しないことが証明された． ∎

5.2. オリジナルのロブ - パス問題との比較

本論文の設定には，オリジナルのロブ - パス問題と異なっている点がいくつかある．それらに関して，なぜそのようにしたかの根拠を述べておく．

まず，オリジナルのロブ - パス問題では，相手の状態が s のときプレイヤーが r を選択すれば，確率 $w(r,s)$ で利益 1 を，確率 $(1-w(r,s))$ で利益 0 を得る．そして，合計利益の期待値を戦略の評価基準とする．しかし，本論文では反応曲線 f_L,f_P が既知なので，ゲームの結果からこれを推測する必要はない．そのため，直接 $w(r,s)$ という利益を得る，としても同じことになる．

次に，オリジナルのロブ - パス問題では，時間は離散的 $(t = 1,2,3,\cdots)$ である．それに対し本論文では，離散性ゆえに生じる繁雑さをさけ，本質的な現象に注目するために，時間が連続な場合を扱っている．この設定は，もともと連続時間の現象をモデル化したのだと解釈しても良いが，離散時間の問題の近似として見ることもできる．離散時間の問題で，プレイヤーのロブ率に相手の推定ロブ率が追従する速度が十分遅く，対応してゲームの打ち切り時刻が十分大きいという場合の極限をとると，連続時間問題になる．したがって，相手の戦略が忘却ありだが忘却因子が十分小さい場合や，忘却なしで十分時間が経った後の部分を近似しているのだと解釈する事も出来る．

また，オリジナルのロブ - パス問題では，プレイヤーの選択肢は $r = 1$(ロブ), $r = 0$(パス) の 2 通りだけとされている．それに対し，本論文では混合戦略を許し，プレイヤーは $0 \le r \le 1$ の任意の実数値を「選択肢」として選んでよいとしている．つまり，ある瞬間の手として，ロブ 70%, パス 30% といった選択が可能となっている．その根拠は次の通りである．連続時間にした場合には，$r\epsilon$ の時間ロブを出し $(1-r)\epsilon$ の時間パスを出すという手続きを繰り返せば，ϵ を十分小さく設定することにより，ロブの比率が r という混合戦略をいくらでも良く近似することができるので，はじめから混合戦略を許すのが自然であると考える．定理 1, 定理 2に示したように，混合戦略を許しても最適戦略は bang-bang 制御 ($r = 0,1$ のいずれかしかとらない) の形になる．しかし，$r = 1$ と $r = 0$ の切替え点においては，$0 < r < 1$ の値をとる必要が生じる．

なお，反応曲線 f_L,f_P と打ち切り時刻 T は本論文同様に既知だが，時間が離散で混合戦略を許さないという場合でも，動的計画法のように打ち切り時刻から逆に遡って考えてゆけば，3.1,3.2章に対応した，「ロブを出すべき領域 L」と「パスを出すべき領域 P」を求めることはできる．しかし離散時間の場合には 3.1,3.2章と違って L,P が入り組んだ形 (t を固定した時の切片が非連結) になり，その境界を明示的に書き下すことは困難であると考えられる．

6. おわりに

本論文では，反応曲線および打ち切り時刻が既知な場合のロブ - パス問題を扱い，相手が忘却あり，なしの 2 通りの設定に対してそれぞれ厳密な最適解を求めた．その系として，マッチング条件を仮定すれば最適戦略が打ち切り時刻によらない事を指摘し，これまでの論文 [1][4][6] が採用していた方針の根拠を与えた．マッチング条件自体の意味や妥当性に関する議論も行った．さらに，漸近最適性を定義し，忘却ありなら最適定常戦略が漸近最適となるが，忘却なしの時は漸近最適戦略は存在しない事を示した．

今回仮定した前提条件 (時間が連続, 反応曲線が滑らかで単調, など) が成立していない時の最適戦略は, これからの研究課題である. また, オンライン学習問題としてのロブ - パス問題の戦略を評価するためにも, 忘却なし · 打ち切り時刻未知 · マッチング条件不成立の時, 「最適」戦略をどのように定義すればよいのか, 今後検討する必要がある.

謝辞

本研究に関して有益な議論をして頂いた, 理化学研究所の甘利俊一教授に感謝致します. この研究の一部は, 文部省重点領域研究費 08279102 によって行われました.

参考文献

[1] N. Abe and J.Takeuchi: The 'lob-pass' problem and an on-line learning model of rational choice. *In Proceedings of the 1993 Workshop on Computational Learning Theory*, (1993).

[2] D. A. Berry and B. Fristedt: *Bandit problems* (Chapman and Hall, 1985).

[3] R. J. Herrnstein: Rational choice theory. *American Psychologist*, **45-3** (1990) 356-367.

[4] K. Hiraoka and S. Amari: Strategy under unknown stochastic environment — the nonparametric lob-pass problem. *Algorithmica*, **22** (1998) 138–156.

[5] K. Hiraoka, S. Amari, and S. Yoshizawa: Stochastic game under unknown environment — a strategy for nonparametric lob-pass problem. *In Proceedings of the 1995 International Symposium on Nonlinear Theory and its Applications*, (1995) 171–173.

[6] J. Kilian, K. J. Lang and B. A. Pearlmutter: Playing the matching-shoulder lob-pass game with logarithmic regret. *In Proceedings of the 1994 Workshop on Computational Learning Theory*, (1994).

A. 忘却ありの場合の最適戦略の導出

この付録では, 定理1の証明を与える. $r(\cdot)$ を決定するかわりに, 同じことなので $s(\cdot)$ を決定することを考える. すなわち,

$$s(t_0) = s_0 \quad (0 \leq s_0 \leq 1)$$

$$0 \leq r(t) = s(t) + \frac{d}{dt}s(t) \leq 1 \tag{A.1}$$

の条件のもと, 連続かつ区分 C^1 級関数 $s(t)$ に対する $\mathcal{G}[s] = \int_{t_0}^{T} w(r(t), s(t))dt$ の最大化問題を考える. まず, 終端点 $s(T)$ を固定しての最大化を行い, その後, 最適な終端点を決定する.

A.1. 終端点を固定しての最大化

次の補題は明らかである.

補題 3 二つの関数 $s^{(0)}(t)$,$s^{(1)}(t)$ が, 制約 *(A.1)* を満たしているとする. この時, $0 \leq \alpha \leq 1$ に対して, 関数 $s^{(\alpha)}(t) \equiv (1-\alpha)s^{(0)}(t) + \alpha s^{(1)}(t)$ も制約 *(A.1)* を満たす. □

また, 次の計算も容易である.

補題 4 補題 *3*と同じ設定で, $s^{(0)}(t_0) = s^{(1)}(t_0) = s_0$, $s^{(0)}(T) = s^{(1)}(T) = s_T$, であるとする. このとき,

$$\frac{d}{d\alpha}\mathcal{G}[s^{(\alpha)}] = \int_{t_0}^{T} v'(s^{(\alpha)}(t))\Big(s^{(1)}(t) - s^{(0)}(t)\Big)dt$$

□

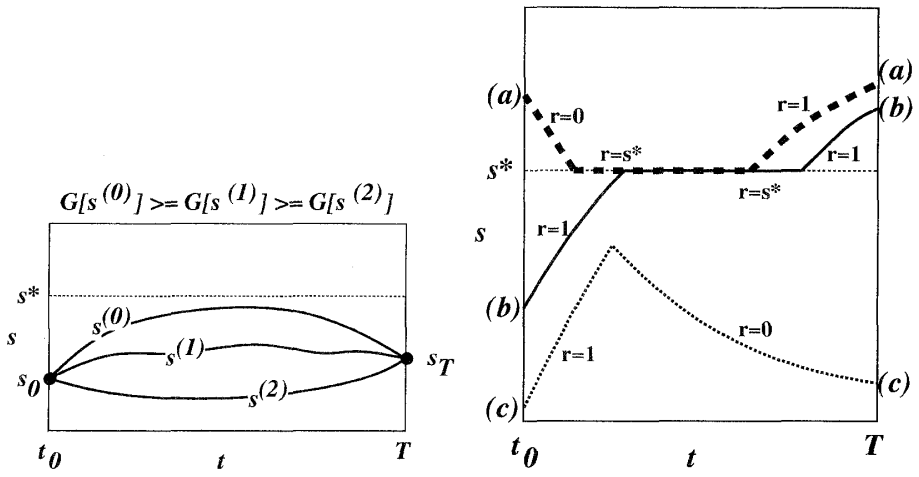

図 7: (左) 両端点を固定した時の $\mathcal{G}$ の比較, (右) 両端点を固定した時の最適戦略

すると, 単峰性 (2.4) から, 次の系が導かれる.

系 1 補題 4と同じ設定で, $s^{(0)}$ と $s^{(1)}$ が s^* に関して同じ側にあり, かつ $s^{(0)}$ の方が $s^{(1)}$ よりも s^* に近いとする. つまり, すべての t で $(s^{(0)}(t)-s^*)(s^{(1)}(t)-s^*) \geq 0, |s^{(0)}(t)-s^*| \leq |s^{(1)}(t)-s^*|$ を仮定する (図 7左). このとき, $\mathcal{G}[s^{(0)}] \geq \mathcal{G}[s^{(1)}]$ である。 □

つまり, この系は次のことを主張している: $s(t_0) = s_0$, $s(T) = s_T$ と両端点を固定した時には, 制約 (A.1) を満たす範囲で, 曲線 $s(t)$ を $s = s^*$ にできるだけ近づけたものが, 最適戦略 ($\mathcal{G}[s]$ を最大化する解) である. 図 7右には, 例として 3 通りの端点 (a)(b)(c) について, この最適戦略が示されている. 図 7右に示されているように, 最適戦略は次の 3 つの段階からなる:

- (ア) $r = 1$ $(s_0 \leq s^*)$ または $r = 0$ $(s_0 > s^*)$ で, $s = s^*$ へ向かう.
- (イ) $s = s^*$ に到達したら, $r = s^*$ で $s(t) = s^*$ を保つ.
- (ウ) 最後に $r = 1$ $(s^* \leq s_T)$ 又は $r = 0$ $(s^* > s_T)$ で, 指定された終端点 $s = s_T$ へ向かう.

なお, 図 7右の (c) のような場合 ($s = s_0$ から $s = s^*$ を経由して $s = s_T$ へ到達することが, 制約 (A.1) から不可能な場合) には, 段階 (イ) が飛ばされることになる.

A.2. 最適な終端点の決定

前節で, 開始点および終端点を固定した時の最適解が求められたので, 次に終端点 s_T を動かして最適な終端点を決定する. まず, $r(t) = 0 (r(t) = 1)$ を続けた時の, 微分方程式 (2.1) (忘却ありの方) の解を確認しておく.

補題 5 $t_1 \leq t \leq t_2$ において $r(t) \equiv 0$ のとき, 微分方程式 (2.1)(忘却あり) の解は $s(t) = s(t_1)e^{-(t-t_1)}$ である. 一方, $r(t) \equiv 1$ のときは, $s(t) = 1 - \Big(1 - s(t_1)\Big)e^{-(t-t_1)}$ である.

特に, 段階 (ウ) の開始時刻 (以下, 切り替え時刻と呼ぶ) を τ とおくと, τ と終端点 s_T の間には,

$$s_T = \begin{cases} 1-(1-s(\tau))e^{-(T-\tau)} & (\tau \text{ 以降ロブ } r(t)=1 \text{ に切り替える場合}) \\ s(\tau)e^{-(T-\tau)} & (\tau \text{ 以降パス } r(t)=0 \text{ に切り替える場合}) \end{cases} \tag{A.2}$$

という関係がある. そこで, 終端点 $s(T) = s_T$ を動かすかわりに, 同値なことなので, 切替え時刻 τ の方を調節することにする. 切替え時刻 τ に対応する最適戦略の $s(t)$ を, $s_\tau(t)$ で表す. τ 以降ロブに切り替えるかパスに切り替えるかは, 別途指定する. ここで, 次の補題が成り立つ. (D_L,D_P は, 定理 1で定義されている.)

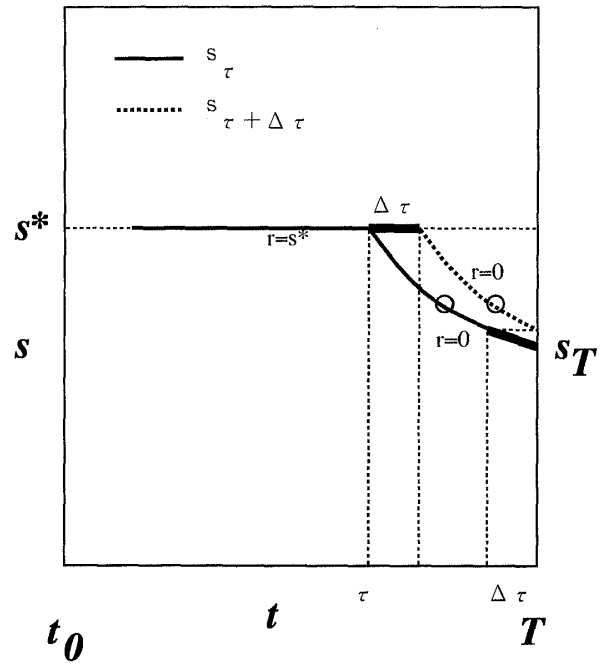

図 8: 切り替え時刻を微小に変化させた時 (段階 (イ) →段階 (ウ))

補題 6 段階 *(イ)* から段階 *(ウ)* へ切り替えるときには,

$$\frac{d}{d\tau}\mathcal{G}[s_\tau] = \begin{cases} D_L(\tau, s_\tau(\tau)) & (\tau \text{ 以降ロブに切り替える場合}) \\ D_P(\tau, s_\tau(\tau)) & (\tau \text{ 以降パスに切り替える場合}) \end{cases}$$

であり, 段階 *(イ)* を飛ばして段階 *(ア)* から直接段階 *(ウ)* へ切り替えるときには,

$$\frac{d}{d\tau}\mathcal{G}[s_\tau] = \begin{cases} D_L(\tau, s_\tau(\tau))/s_\tau(\tau) & (\tau \text{ 以降ロブに切り替える場合}) \\ D_P(\tau, s_\tau(\tau))/(1-s_\tau(\tau)) & (\tau \text{ 以降パスに切り替える場合}) \end{cases}$$

である. □

証明:

τ 以降パスに切り替える場合についての証明を示す. 段階 (イ) から段階 (ウ) へ切り替える場合には, $s_\tau(\tau) = s^*$ であり,

$$\begin{aligned} \mathcal{G}[s_{\tau+\Delta\tau}] - \mathcal{G}[s_\tau] &= \int_\tau^{\tau+\Delta\tau} v(s^*)dt - \int_{T-\Delta\tau}^{T} f_P(s_\tau(t))dt \\ &= v(s^*)\Delta\tau - \int_{T-\Delta\tau}^{T} f_P(s_T + O(\Delta\tau))dt \\ &= \Big(v(s^*) - f_P(s_T)\Big)\Delta\tau + O(\Delta\tau^2) \end{aligned}$$

となる (図 8). 図中で○印をつけた 2 つの曲線は合同なため, この部分の利益は s_τ でも $s_{\tau+\Delta\tau}$ でも等しいことに注意されたい. これに (A.2) を代入すれば定理が示される.

一方, 段階 (イ) を飛ばして, 段階 (ア) から直接段階 (ウ) へ切り替える場合には,

$$\mathcal{G}[s_{\tau+\Delta\tau}] - \mathcal{G}[s_\tau] = \int_\tau^{\tau+\Delta\tau} f_L(s_{\tau+\Delta\tau}(t))dt + \int_{\tau+\Delta\tau}^{\tau+\Delta\tau'} f_P(s_{\tau+\Delta\tau}(t))dt - \int_{T-\Delta\tau'}^{T} f_P(s_\tau(t))dt \quad \text{(A.3)}$$

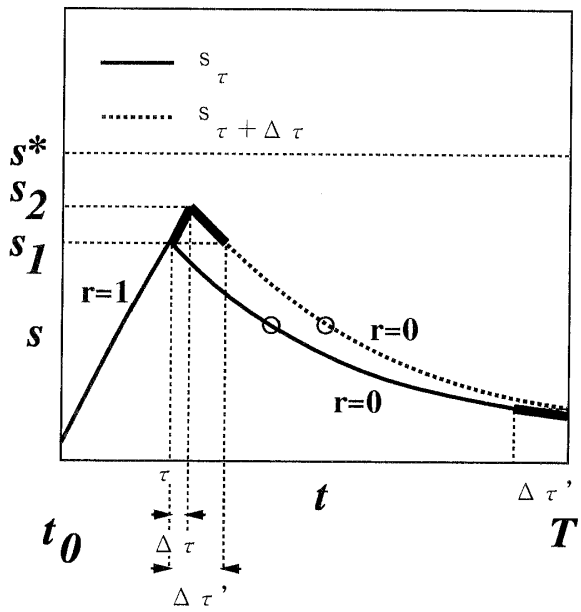

図 9: 切り替え時刻を微小に変化させた時 (段階 (ア) →段階 (ウ))

となる (図 9). ここに $\Delta\tau'$ は, $s_{\tau+\Delta\tau}(\tau+\Delta\tau') = s_\tau(\tau)$ で定義される. 今, $s_1 = s_\tau(\tau)$, $s_2 = s_{\tau+\Delta\tau}(\tau+\Delta\tau)$ とおくと, $s_2 = 1-(1-s_1)e^{-\Delta\tau}$, $s_1 = s_2 e^{-(\Delta\tau'-\Delta\tau)}$ だから, これを解くと

$$\Delta\tau' = \log\Big(s_1 + (e^{\Delta\tau}-1)\Big) - \log s_1 = \frac{1}{s_1}\Delta\tau + O(\Delta\tau^2)$$

という関係が得られる. よって,

$$\begin{aligned}(\text{A.3}) &= f_L(s_1)\Delta\tau + f_P(s_1)(\Delta\tau'-\Delta\tau) - f_P(s_T)\Delta\tau' + O(\Delta\tau^2)\\ &= \Big[s_1 f_L(s_1) + (1-s_1)f_P(s_1) - f_P(s_T)\Big]\Delta\tau' + O(\Delta\tau^2)\\ &= \Big[v(s_1) - f_P(s_T)\Big]\frac{1}{s_1}\Delta\tau + O(\Delta\tau^2)\end{aligned}$$

となり, 定理を得る. τ 以降ロブに切り替える場合についても証明は同様である. ∎

今, $s^* \geq s^\circ$ であった場合を考えよう (図 1). この時, D_L, D_P の符号は, 図 10, 図 11のようになっている. $D_L(\tau, s_\tau(\tau))$ または $D_P(\tau, s_\tau(\tau))$ が正 (負) なら, 切り替え時刻 τ をもっと遅く (早く) した方が合計利益が大きくなるのであった. 図 10, 図 11中の太い矢印は, この方向を示している. したがって, 初期条件 $s(t_0)$ に応じて, 最適解は図 3(左) のような 5 通りの型になる. $s^* < s^\circ$ であった場合も, 同様にして, 図 3(右) のようになることがわかる. これらはすべて, 「図 3の, 領域 L でロブ, 領域 P でパスを出す」という形をしている. これが定理 1の主張であった.

B. 忘却なしの場合の最適戦略の導出

この付録では, 定理 2の証明を与える. 忘却ありの時 (付録 A) と全く同じように考えていく.

B.1. 終端点を固定しての最大化

A.1節の補題 4と同様の計算を行うと,

$$\frac{d}{d\alpha}\mathcal{G}[s^{(\alpha)}] = \int_{t_0}^{T}[s^{(\alpha)}f_L'(s^{(\alpha)}) + (1-s^{(\alpha)})f_P'(s^{(\alpha)})](s^{(1)}-s^{(0)})dt$$

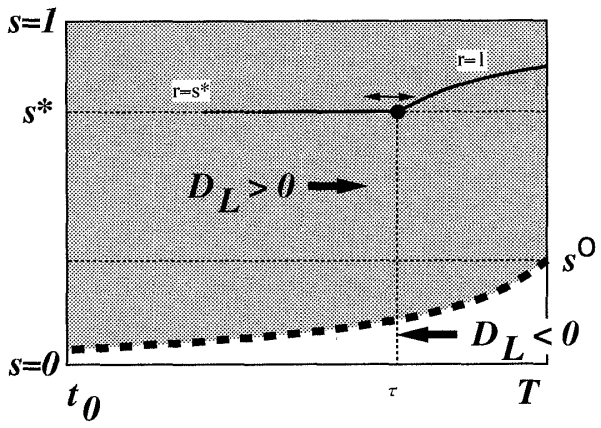

図 10: ロブに切り替える場合

となる. この $sf_L'(s)+(1-s)f_P'(s)$ の零点が $s=s^\times$ である. したがって, A.1節で述べた「忘却ありの最適戦略」の s^* を $s^\times$ でおきかえたものが, 忘却なしの最適戦略になる.

B.2. 最適な終端点の決定

今, $u=\log t, u_0=\log t_0, U=\log T$ と変数変換し, $R(u)=r(t)\big|_{t=e^u}, S(u)=s(t)\big|_{t=e^u}$ とおくと,

$$\frac{d}{du}S(u)=R(u)-S(u) \tag{B.1}$$

であり, 最大化すべきものは, $\mathcal{G}[s]$ のかわりに $\mathcal{H}[S]=\int_{u_0}^{U}w(R(u),S(u))e^u du$ となる. すなわち, 相手が忘却ありで, 未来を重視するような重みのかかった合計利益を最大化する問題に書きかえられる. 特に, $S(u)$ のダイナミクス (B.1) は忘却ありの $s(t)$ と同じだから, 補題5の (t,r,s) を (u,R,S) でおきかえた命題が成り立つ. A.2節と同様に, 段階 (ウ) の開始時刻を $u=\sigma$ とし, この σ に対応する (終端点固定の時の) 最適戦略を $S(u)=S_\sigma(u)$ であらわすことにする. このとき, 補題6に対応する命題は, 次のようになる.

補題 7 段階 *(イ)* から段階 *(ウ)* へ切り替えるときには,

$$\frac{d}{d\sigma}\mathcal{H}[S_\sigma] = \begin{cases} D_L(\sigma,S_\sigma(\sigma)) & (\sigma \text{ 以降ロブに切り替える場合}) \\ D_P(\sigma,S_\sigma(\sigma)) & (\sigma \text{ 以降パスに切り替える場合}) \end{cases}$$

であり, 段階 *(イ)* を飛ばして段階 *(ア)* から直接段階 *(ウ)* へ切り替えるときには,

$$\frac{d}{d\sigma}\mathcal{H}[S_\sigma] = \begin{cases} D_L(\sigma,S_\sigma(\sigma))/S_\sigma(\sigma) & (\sigma \text{ 以降ロブに切り替える場合}) \\ D_P(\sigma,S_\sigma(\sigma))/(1-S_\sigma(\sigma)) & (\sigma \text{ 以降パスに切り替える場合}) \end{cases}$$

である. ここに,

$$D_L(u,S) = e^u\Big[v(S)-f_L(1-(1-S)e^{-(U-u)})e^{U-u} + \int_0^{U-u} f_L(1-(1-S)e^{-\xi})e^{\xi}d\xi\Big]$$

$$D_P(u,S) = e^u\Big[v(S)-f_P(Se^{-(U-u)})e^{U-u} + \int_0^{U-u} f_P(Se^{-\xi})e^{\xi}d\xi\Big]$$

と定義する. □

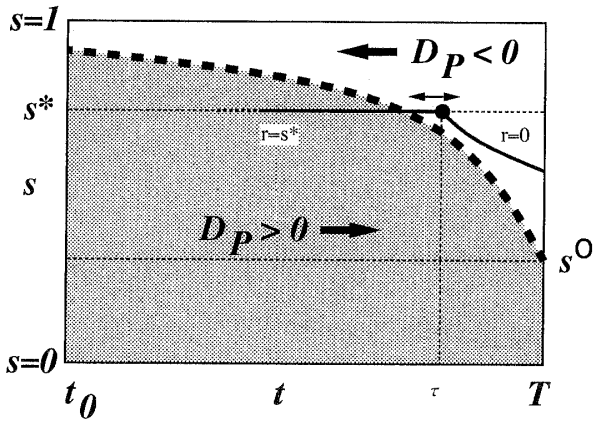

図 11: パスに切り替える場合

証明:

σ 以降パスに切り替える場合についての証明を示す．段階 (イ) から段階 (ウ) へ切り替える場合には，$S_\sigma(\sigma) = s^\times$ である．(u, S) の図は，図 8の (t, r, s, τ, s^*) を $(u, R, S, \sigma, s^\times)$ でおきかえたものになる．s^* のかわりに $s^\times$ となることを注意しておく．2 つの曲線の○印をつけた部分が合同なことも補題 6のときと同様である．しかし，最大化すべき汎関数 $\mathcal{G}$ と $\mathcal{H}$ には，違いがある．すなわち，$\mathcal{H}$ には未来を重視するような重み e^u が入っている．このため，補題 6のときとは異なり，○印をつけた部分での利益は相殺されない．この理由で，D_L,D_P の形に差がでてくる．具体的には，

$$
\begin{aligned}
&\mathcal{H}[S_{\sigma+\Delta\sigma}] - \mathcal{H}[S_\sigma] \\
&= \left[\int_\sigma^{\sigma+\Delta\sigma} v(s^\times)e^u du + \int_{\sigma+\Delta\sigma}^{U} f_P(S_{\sigma+\Delta\sigma}(u))e^u du\right] - \left[\int_\sigma^U f_P(S_\sigma(u))e^u du\right] \\
&= \left[\int_\sigma^{\sigma+\Delta\sigma} v(s^\times)e^u du + e^{\Delta\sigma}\int_\sigma^{U-\Delta\sigma} f_P(S_\sigma(u))e^u du\right] \\
&\quad - \left[\int_\sigma^{U-\Delta\sigma} f_P(S_\sigma(u))e^u du + \int_{U-\Delta\sigma}^{U} f_P(S_\sigma(u))e^u du\right] \\
&= e^\sigma\left[\int_\sigma^{\sigma+\Delta\sigma} v(s^\times)e^{u-\sigma} du + (e^{\Delta\sigma}-1)\int_\sigma^{U-\Delta\sigma} f_P(S_\sigma(u))e^{u-\sigma} du - \int_{U-\Delta\sigma}^{U} f_P(S_\sigma(u))e^{u-\sigma} du\right] \\
&= e^\sigma\left[v(s^\times) + \int_\sigma^U f_P(S_\sigma(u))e^{u-\sigma} du - f_P(s_T)e^{U-\sigma}\right]\Delta\sigma + O(\Delta\sigma^2)
\end{aligned}
$$

一方，段階 (イ) を飛ばして，段階 (ア) から直接段階 (ウ) へ切り替える場合には，補題 6の証明と同じ事情[4]で，$1/S_\sigma(\sigma)$ という因子がかかることになる．σ 以降ロブに切り替える場合についても，証明は同様である． ∎

A.2節と同様の考察を行えば，

$$P = \{(u, S) | s \le s^\times \text{かつ} D_P(u, S) \ge 0\}, \qquad L = P^c \ (P \text{ の補集合})$$

[4]補題 6の証明と同様に，$S_{\sigma+\Delta\sigma}(\sigma + \Delta\sigma') = S_\sigma(\sigma)$ で $\Delta\sigma'$ を定義すると，$\Delta\sigma' = \Delta\sigma/S_\sigma(\sigma) + O(\Delta\sigma^2)$ となる．

とおいて, 領域 L ではロブを, 領域 P ではパスを出すのが最適であることがわかる. 以下, 領域 L,P の境界曲線 $S_c(u)$ について考える. この曲線は, ある時刻 $u^\times$ を境にして,

$$S_c(u) = s^\times \qquad (u \le u^\times) \tag{B.2}$$

$$D_P(u, S_c(u)) = 0 \qquad (u > u^\times) \tag{B.3}$$

という形をしている. 部分積分により, (B.3) は

$$\int_0^{U-u} f_P'(S_c(u)e^{-\xi})d\xi = f_P(S_c(u)) - f_L(S_c(u)) \tag{B.4}$$

と変形される. (B.3) の両辺を u で微分して (B.4) を代入すると,

$$\frac{dS_c}{du} = \frac{f_P'(S_c(u)e^{-(U-u)})}{\dfrac{\partial w(r,s)}{\partial s}\Big|_{r=s=S_c(u)} - f_P'(S_c(u)e^{-(U-u)})} S_c(u)$$

という微分方程式に直される. 境界条件は, (B.3) に $u = U$ を代入した $f_P(S_c(U)) - v(S_c(U)) = 0$, すなわち, $S_c(U) = s^\circ$ である. $u^\times$ は, $u = u^\times$, $S_c(u^\times) = s^\times$ を (B.4) に代入して,

$$\int_0^{U-u^\times} f_P'(s^\times e^{-\xi})d\xi = -\Big(f_L(s^\times) - f_P(s^\times)\Big)$$

という条件から定められる. 変数をもとの t に戻すと,

$$\frac{ds_c}{dt} = \frac{f_P'(s_c(t)t/T)}{\dfrac{\partial w(r,s)}{\partial s}\Big|_{r=s=s_c(t)} - f_P'(s_c(t)t/T)} \frac{s_c(t)}{t}, \qquad s_c(T) = s^\circ$$

$$\int_{t^\times/T}^1 f_P'(s^\times\zeta)\frac{d\zeta}{\zeta} = f_P(s^\times) - f_L(s^\times)$$

となり, 定理 2を得る.

Kazuyuki Hiraoka and Shuji Yoshizawa
Department of Information Engineering
University of Tokyo, 7-3-1, Hongo, Bunkyo-ku
Tokyo 113, Japan
E-mail: hiraoka@bios.t.u-tokyo.ac.jp

ABSTRACT

THE OPTIMAL SOLUTION OF THE LOB-PASS PROBLEM WITH KNOWN REACTION CURVES

Kazuyuki Hiraoka Shuji Yoshizawa
The University of Tokyo

The "lob-pass problem" is a model which is used in the psychology. It describes the phenomena that the same choices decrease the effect, like the experience or the weariness. Abe and Takeuchi formulated it as an on-line learning problem, and pointed out that it is an extension of the multi-armed bandit problem. In the lob-pass problem, the player's choices will change the environment itself. This is the difference from the multi-armed bandit problems.

The all proposed strategies for the lob-pass problem repeat the following procedures: (i) observe the reaction from the unknown environment (ii) estimate the environment (iii) find the optimal "stationary" strategy for the estimated environment (iv) determine the choice according to the strategy. Moreover, the criteria for the strategies in these studies are the loss due to uncertainness of the environment, compared with the optimal "stationary" strategy for the known-environment case.

To judge whether such policies are appropriate or not, we have to know the optimal strategy, which may not be "stationary", for the known-environment case. It is calculated in the present paper. It is also shown that the "matching condition" assumed in the past studies is the necessary and sufficient condition that the optimal strategy doesn't depend on the stopping time of the game. The meaning and the appropriateness of the matching condition are discussed. Finally, the asymptotically optimality is defined. We prove that the stationary strategy can be asymptotically optimal for the opponent with the forgetting factor, but no strategy is asymptotically optimal for the opponent without the forgetting factor.