



交代手番によるスロット・マシン問題 : 熟練プレイヤーと未熟練プレイヤーの例

末廣, 英生

(Citation)

国民経済雑誌, 193(2):51-74

(Issue Date)

2006-02

(Resource Type)

departmental bulletin paper

(Version)

Version of Record

(JaLCD0I)

<https://doi.org/10.24546/00056053>

(URL)

<https://hdl.handle.net/20.500.14094/00056053>



交代手番によるスロット・マシン問題： 熟練プレーヤーと未熟練プレーヤーの例¹⁾

末 廣 英 生

2人のプレーヤーからなるチームが、2つの投資機会に直面し、每期その一方を選択して投資する。1回の投資の結果は成功か失敗のいずれかである。一方の投資機会の成功確率は $\frac{1}{2}$ であることが分かっているが、もう一方の投資機会の成功率は $\frac{2}{3}$ であるか $\frac{1}{3}$ であるかのいずれかが等確率である。2人のプレーヤーの一方はこれ以上の情報を持たないが、もう一方のプレーヤーは成功確率が不明の投資機会の投資試行 M 回分の結果を知っている。2人のプレーヤーが交代に投資選択するとき、それを追加情報を持つプレーヤー1人でする場合に較べて、大きなロスが生まれるだろうか。追加情報を持つプレーヤーが先手番で M が奇数なら、ロスは全く生じない。そうでなければロスは避けられないが、 M が十分大ならば、追加情報を持つプレーヤーが先手番で M が偶数の場合のロスはほぼゼロであり、情報を持たないプレーヤーが先手番の場合のロスとはほぼ1期目の1回分のロスに過ぎない。

キーワード スロット・マシン問題、交代手番

1 問 題

複数の個人からなるチームで1つの意思決定をする場合、個人がそれぞれ持っている、なすべき決定に関わる情報を活用できるか否かが、チームの決定のパフォーマンスを左右する。その特別な場合として、なすべき決定に関する情報がチームの個人の間で偏って持たれている場合がある。例えば、仕事のチームは年長の熟練者と年少の未熟練者から構成される場合がしばしばあり、熟練者はなすべき決定についてよく知っているが未熟練者はそうではない。その場合、その情報は果たしてうまく活用されるのだろうか。²⁾

この問題を、簡単なスロット・マシン問題の例について明らかにする。2人の個人からなるチームが、あらかじめ固定された、リスクを伴う2つの代替的投資機会に直面し、そのいずれかに投資する選択を1期目、2期目という具合に繰り返して行わねばならない。すなわち、1期目の期首に、チームは2つの投資機会のいずれか一方を選択して投資しなければならない。投資の結果は1期目の期末に分かる。2期目の期首に、チームは再び同じ2つの投資機会のいずれか一方を選択して投資し、その結果は2期目の期末に分かる。3期目以後も

同じことの繰り返しである。ここに、あらかじめ固定された投資機会は、2つの投資機会の性質は互いに独立であり、同じ投資機会を異なる期に選択するときその選択結果は確率的に独立である。

投資機会を arm (スロット・マシン) と呼ぶ。投資するとは、arm を1回引くことである。投資の結果は、いずれの投資機会も、成功 (Success) か失敗 (Failure) のいずれかである。チームは、1回の投資によって、それがどちらの投資機会によって実現されたかを問わず、成功ならそれから $u(\text{Success})$ の利得を得、失敗なら $u(\text{Failure})$ の利得を得る。投資結果の評価は Success を Failure より高く評価し、その評価値は、一般性を失うことなく

$$u(\text{Success})=1, u(\text{Failure})=-1$$

と表せる。チームは、無限に続く期間 $t=1, 2, \dots$ にわたって各期にある投資結果が起きるということを、各期 t の投資結果 z_t の利得 $u(z_t)$ を時間割引率 $\delta=0.95$ で現在価値に換算した合計

$$\sum_{t=1}^{\infty} (0.95)^{t-1} u(z_t)$$

で評価する。

2つある投資機会的一方を X-arm, 他方を Y-arm と呼ぶ。Y-arm の成功確率は $\frac{1}{2}$ で、その値はチームの2人には知られており、そのことは2人の間で共有知識である。他方、X-arm の成功確率には、それが $p_H = \frac{2}{3}$ である可能性と $p_L = \frac{1}{3}$ である可能性との2つの可能性があり、チームが実際に直面している X-arm がそのいずれの成功確率を持つ arm であるかは、それを直接知る手だては一切ない³⁾。もし、X-arm の成功確率について全く情報がないなら、人はその2通りの可能性をともに等しいと見なすものとする。すなわち、X-arm の成功確率が $p_H = \frac{2}{3}$ である可能性に付す事前確率は $\rho = \frac{1}{2}$ である。チームの2人のメンバーのうち1人は、そのような、X-arm の成功確率について全く情報を持たない個人である。彼を未熟練プレーヤーと呼ぶ。もう1人のメンバーは、チームが直面している X-arm について、今考えているチームの利得とは関係なく事前にそれを M 回独立に引き、つまりその M 回のベルヌーイ試行を行い、その結果を知っている。彼を M 熟練プレーヤーと呼ぶ。チームの未熟練プレーヤーがそうであることは、チームの2人の間で共有知識である。チームの M 熟練プレーヤーがそうであることも、チームの2人の間で共有知識である。しかし、 M 熟練プレーヤーが知っている事前の M 回のベルヌーイ試行の結果は、それを未熟練プレーヤーに直接教える手だては一切ない。

チームの投資機会の選択は、あらかじめ決まった交代手番で行う。すなわち、1期目は M 熟練プレーヤー、2期目は未熟練プレーヤーという具合に、奇数期目は M 熟練プレーヤーが選択し、偶数期目は未熟練プレーヤーが選択するか、逆に1期目は未熟練プレーヤー、2期

目は M 熟練プレーヤーという具合に、奇数期目は未熟練プレーヤーが選択し、偶数期目は M 熟練プレーヤーが選択するか、のいずれかである。

このチームについて、それが行う投資機会の選択は、それをチームでなく M 熟練プレーヤー 1 人でやる場合に較べてどの程度うまくいくだろうか。

2 分析の準備：熟練プレーヤーが 1 人でプレイする場合

まず、問題のチームが直面する投資機会の選択問題を、チームでなく M 熟練プレーヤー 1 人でやる場合に、彼がどのようにプレイするかを明らかにする。この問題は Bandit Problem と呼ばれ、Berry and Fristedt [1985] が包括的に紹介しているように、確率論で精力的に研究され、多くのことが分かっている。その結果を用いて、我々の M 熟練プレーヤー 1 人でプレイする投資選択問題を具体的に解く。

2.1 熟練プレーヤーが 1 人でプレイする場合の問題の定式化

我々の M 熟練プレーヤー 1 人でプレイする投資選択問題は、特定の数値について考えているパラメーター $\delta=0.95, p_H=\frac{2}{3}, p_L=\frac{1}{3}, \rho=\frac{1}{2}$ を、パラメーターのままとして次のように定式化できる。

問題の X-arm の実際の成功確率が p であるとせよ。 t 期に X-arm を選択したとしたらその結果が何か、を表す確率変数を X_t と書く。すなわち、 X_t は、確率 p で $X_t = \text{Success}$ 、確率 $1-p$ で $X_t = \text{Failure}$ であるような確率変数である。選択をなすべき期が $t=1, 2, \dots$ と無限に続くのに対応して、我々は、そのような確率変数の無限列 $\mathbf{X} = (X_1, X_2, \dots)$ を持つ。同様に、確率 $\frac{1}{2}$ で $Y_t = \text{Success}$ 、確率で $\frac{1}{2}$ で $Y_t = \text{Failure}$ であるような確率変数として、 t 期に Y-arm を選択したとしたらその結果が何か、を表す確率変数 Y_t を定義できて、その無限列 $\mathbf{Y} = (Y_1, Y_2, \dots)$ を考えることができる。

プレーヤーがする t 期の選択 a_t とは、X-arm を選択することかまたは Y-arm を選択すること、つまり $a_t = \text{X-arm}$ または $a_t = \text{Y-arm}$ である。 t 期の選択 a_t の結果が何か、を表す確率変数を Z_t と書く。それは、 $a_t = \text{X-arm}$ なら $Z_t = X_t$ であり、 $a_t = \text{Y-arm}$ なら $Z_t = Y_t$ である。 $t=1, 2, \dots$ の各期にプレーヤーがする選択の無限列 $\mathbf{a} = (a_1, a_2, \dots)$ が定まっていると、それに対応する選択結果を表す確率変数の列 $\mathbf{Z} = (Z_1, Z_2, \dots)$ が定まる。

プレーヤーの選択列 $\mathbf{a}$ が定まると、それが定める選択結果を表す確率変数の列 $\mathbf{Z}$ の測度によって、プレーヤーの期待利得が

$$E_{\mathbf{Z}} \left[\sum_{t=1}^{\infty} \delta^{t-1} u(Z_t) \mid p, \mathbf{a} \right]$$

と計算される。⁴⁾

実際には、プレーヤーは、 t 期の選択 a_t をするのに、 t 期までのプレイ履歴、つまり「1期目には a_1 を選択しその結果が z_1 となり… $t-1$ 期目には a_{t-1} を選択しその結果が z_{t-1} となった」というヒストリー $h_t = ([a_1, z_1], \dots, [a_{t-1}, z_{t-1}])$ を知った上ですることができる。その様な選択の仕方を 1 つ記述したもの、つまりありうるプレイ履歴 h_t ごとにその t 期にする選択 a_t を記述した関数 $s_t: h_t \rightarrow a_t$ を、そのプレーヤーの t 期の戦略と言う。それを、選択をしなければならない全ての $t=1, 2, \dots$ 期に用意したもの $s = (s_1, s_2, \dots)$ がプレーヤーの戦略である。

プレーヤーがその戦略 s に従うときに生成されるプレイ $h = ([a_1, z_1], [a_2, z_2], \dots)$ は、 $t-1$ 期までのプレイ履歴 h_{t-1} が与えられたなら、 t 期までのプレイ履歴 $h_t = (h_{t-1}, [a_{t-1}, z_{t-1}])$ は、 $a_{t-1} = s_{t-1}(h_{t-1})$ が指定する arm の結果を表す確率変数 Z_{t-1} の実現値 z_{t-1} によって定まる、というやり方で定義される確率変数 $H = ([a_1, Z_1], [A_2, Z_2], \dots)$ の実現値である。ここに、 A_t は t 期の選択を表す確率変数で、確率変数 $H_t = ([a_1, Z_1], [A_2, Z_2], \dots, [A_{t-1}, Z_{t-1}])$ から $A_t = s_t(H_t)$ によって誘導される。

戦略 s が生成するプレイ $H = ([a_1, Z_1], [A_2, Z_2], \dots)$ の測度によって、想定している p の下でプレーヤーが戦略 s によるプレイから期待できる利得、つまり戦略 s の期待利得

$$U(p, s) = E_H \left[\sum_{t=1}^{\infty} \delta^{t-1} u(Z_t) | p, s \right]$$

が定まる。ここで実際には、プレーヤーが現に直面している X-arm の成功確率 p は、それを直接確かめる手だてではない。このプレーヤーが、 $p = p_H$ である可能性に付す確率を μ とすると、プレーヤーが戦略 s によるプレイから期待する利得は

$$U(\mu, s) = \mu U(p_H, s) + (1 - \mu) U(p_L, s)$$

である。

ある戦略が

$$\text{問題 } P[\mu] \cdots \max_s U(\mu, s)$$

の解であるとき、 $p = p_H$ である可能性に付す確率が μ のプレーヤーの最適戦略であると言って、 $s^*[\mu]$ と書く。プレーヤーが、今考察している M 熟練プレーヤーの場合、彼がプレイにさしかかる段で $p = p_H$ である可能性に付す確率 (initial belief) としての μ は、元々の可能性 $p = \frac{1}{2}$ に、彼が X-arm について事前に行った M 回ベルヌーイ試行の結果を加味して形成される。それは、 M 回のうち Success が結果した回数が s_0 なら

$$\mu = \mu(s_0, M) = \frac{\rho(p_H)^{s_0} (1 - p_H)^{M - s_0}}{\rho(p_H)^{s_0} (1 - p_H)^{M - s_0} + (1 - \rho)(p_L)^{s_0} (1 - p_L)^{M - s_0}}$$

となる。⁵⁾従って、実際に M 回のうち Success が結果した回数が s_0 なら、その M 熟練プレーヤーは、戦略 $s^*[\mu(s_0, M)]$ に従ってプレイすると予測される。

2.2 Bandit Problem の既知の結果

Bandit Problem の研究によって、ここに定式化した我々の M 熟練プレーヤー 1 人でプレイする投資選択問題は、多くのことが分かっている。

まず、事前の X -arm の M 回ベルヌーイ試行の特定の s_0 によって生成される initial belief $\mu(s_0, M)$ について最適戦略を問題にする代わりに、区間 $[0, 1]$ の任意の initial belief μ ごとに問題 $P[\mu]$ を考え、 $\mu \in [0, 1]$ でインデックスされた問題群を考察する。すると、定常マルコフ決定問題に対するダイナミック・プログラミングの基本定理により、最適戦略の存在に関する次の事実が成り立つ。

定理 1 (Bellman [1956], [1957], Ross [1983])

任意の $\mu \in [0, 1]$ について、問題 $P[\mu]$ には解 $s^*[\mu]$ が存在する。

この事実は、区間 $[0, 1]$ の initial belief μ ごとに最適戦略からの期待利得

$$V(\mu) = U(\mu, s^*[\mu]) = \max_s U(\mu, s)$$

に対応させる価値関数 (value function) $V: [0, 1] \rightarrow R$ を定める。そして、定理 1 に言う最適戦略は、この value function によって間接的に記述できる。すなわち、initial belief μ の下での投資選択問題 1 期分の結果、 a を選択しその結果が z だったというヒストリーの 1 期追加分 $[a, z]$ が誘導する posterior を考えよ。 $a = X$ -arm の時、それは $z = \text{Success}$ なら

$$\mu^+ = \frac{\mu p_H}{\mu p_H + (1 - \mu) p_L}$$

であり、 $z = \text{Failure}$ なら

$$\mu^- = \frac{\mu(1 - p_H)}{\mu(1 - p_H) + (1 - \mu)(1 - p_L)}$$

である。それらの posteriors が生成される確率は、それぞれ

$$\pi(\mu, \mu^+ | X) = \mu p_H + (1 - \mu) p_L$$

$$\pi(\mu, \mu^- | X) = \mu(1 - p_H) + (1 - \mu)(1 - p_L)$$

である。他方、 $a = Y$ -arm の時、それは $z = \text{Success}$ か $z = \text{Failure}$ かによらずもとの μ を posterior として再生する。つまり、 $a = Y$ -arm の時の posterior の遷移確率は

$$\pi(\mu, \mu | Y) = 1$$

である。すると、定理 1 に随伴して次が成り立つ。

定理 2 (Bellman [1956], [1957], Ross [1983])

任意の $\mu \in [0, 1]$ について、Bellman 方程式

$$V(\mu) = \max \{ E_Z[u(Z) | \mu, a = X\text{-arm}] + \delta(\pi(\mu, \mu^+ | X)V(\mu^+) + \pi(\mu, \mu^- | X)V(\mu^-)), \\ E_Z[u(Z) | \mu, a = Y\text{-arm}] + \delta V(\mu) \}$$

が成り立つ。そして、initial belief μ の下での最適戦略は、1 期目にこの右辺の最大値を

与える arm を指示する。

Bandit Problem では、定常マルコフ決定問題全般について成り立つ事実に由来する定理 1, 定理 2 に加えて、最適戦略がさらに明確に特徴づけられる。すなわち、戦略 $s = (s_1, s_2, \dots)$ が、「全ての $t=1, 2, \dots$ について、もし t で $s_t(h_t) = Y\text{-arm}$ ならば、任意の $t' > t$ と任意の $z_t, \dots, z_{t'-1}$ に対して $s_{t'}(h_t, [Y\text{-arm}, z_t], \dots, [Y\text{-arm}, z_{t'-1}]) = Y\text{-arm}$ である」という性質を満たす時、つまりいったん $Y\text{-arm}$ を選択するならばそれ以後は $Y\text{-arm}$ を選択し続けるという性質を持つ時、言い換えるといったん $X\text{-arm}$ の選択を停止するならば 2 度と $X\text{-arm}$ を選択しないという性質を持つ時、停止戦略 (stopping strategy) であると言う。すると、次が成り立つ。

定理 3 (Bellman [1956], Berry and Fristedt [1979])

任意の $\mu \in (0, 1)$ について、initial belief が μ の下での最適戦略 $s^*[\mu]$ は停止戦略である。従って、最適停止戦略が最適戦略である。

そして、どのようなプレイ履歴で $X\text{-arm}$ の選択を停止するのが最適かについて、次が言える。

定理 4 (Berry and Fristedt [1979])

任意の $\mu \in (0, 1)$ について、ある $\bar{p}(\mu) \in [0, 1]$ が存在し、initial belief が μ の下での最適停止戦略は、 $Y\text{-arm}$ の成功確率 $\frac{1}{2}$ に対して、 $\frac{1}{2} > \bar{p}(\mu)$ なら 1 期目に停止する、 $\frac{1}{2} < \bar{p}(\mu)$ なら 1 期目に停止しない、 $\frac{1}{2} = \bar{p}(\mu)$ なら停止してもしなくても良い、というものになる。

この定理により、各 $\mu \in (0, 1)$ に $\bar{p}(\mu)$ を対応させる関数 $\bar{p}: (0, 1) \rightarrow [0, 1]$ が計算できれば、我々の M 熟練プレーヤー 1 人でプレイする投資選択問題の最適戦略は、彼が事前の M 回ベルヌーイ試行の結果持っている initial belief $\mu = \mu(s_0, M)$ を初期値として、それから生成される posteriors μ^+, μ^- ないし μ の経路を計算することによって、完全に記述できることになる。

関数 $\bar{p}: (0, 1) \rightarrow [0, 1]$ がどのような関数かを間接的に記述する方法がある。それは、次の通りである。

定理 5 (Berry and Fristedt [1979])

任意の $\mu \in (0, 1)$ について、

$$\bar{p}(\mu) = \max_{s \text{ s.t. stopping strategy}} \frac{E_X \left[\sum_{t=1}^T \delta^{t-1} \mathbf{1}(X_t = \text{Success}) \mid \mu, s \right]}{E_X \left[\sum_{t=1}^T \delta^{t-1} \mid \mu, s \right]}$$

ただし、 T は、stopping strategys に従ったプレイで、 $X\text{-arm}$ が選ばれる最後の回を表

す確率変数であり、

$$1(X_t = \text{Success}) = \begin{cases} 1 & \text{if } X_t = \text{Success} \\ 0 & \text{if } X_t = \text{Failure} \end{cases}$$

である。

この間接的な記述は、我々の投資選択問題がそうである $0 < p_L < p_H < 1$ かつ $p_H + p_L = 1$ の場合には、運良く次の明示的な計算結果をもたらす。

定理 6 (Berry and Fristedt [1979])

$0 < p_L < p_H < 1$ かつ $p_H + p_L = 1$ とせよ。ならば、任意の $\mu \in (0, 1)$ について、

$$\bar{p}(\mu) = \frac{\mu g_H}{\mu g_H + (1-\mu)g_L} p_H + \frac{(1-\mu)g_L}{\mu g_H + (1-\mu)g_L} p_L$$

ただし

$$g_H = 1 - \frac{1 - \sqrt{1 - 4\delta^2 p_H p_L}}{2\delta p_H} \quad g_L = 1 - \frac{1 - \sqrt{1 - 4\delta^2 p_H p_L}}{2\delta p_L}$$

こうして、我々の M 熟練プレーヤー 1 人でプレイする投資選択問題の最適戦略は、それを明示的に計算する方法がわかった。

2.3 既知の結果の適用

Bandit Problem の既知の結果を適用して、我々の M 熟練プレーヤー 1 人とする投資選択問題の最適戦略を計算する。定理 3 により、最適停止戦略を求めればよい。つまり、最適戦略 $s = (s_1, s_2, \dots)$ の $T+1$ 期の戦略 s_{T+1} を見いだすのに、考察するプレイ履歴 h_{T+1} として、1 期目から T 期目までは X -arm を選択し続けていたもの

$$h_{T+1} = ([X\text{-arm}, z_1], [X\text{-arm}, z_2], \dots, [X\text{-arm}, z_T])$$

だけを検討すればよい。このプレイ履歴における Success の回数を

$$s(h_{T+1}) = \#\{t \in \{1, \dots, T\} \mid z_t = \text{Success}\}$$

と呼ぶなら、初期値 $\mu = \mu(s_0, M)$ から生成される posteriors μ^+, μ^- の経路で達せられる、 $T+1$ 期首の M 熟練プレーヤーの posterior は

$$\begin{aligned} \mu((s_0, M), (s(h_{T+1}), T)) &= \rho(p_H)^{s_0}(1-p_H)^{M-s_0}(p_H)^{s(h_{T+1})}(1-p_H)^{T-s(h_{T+1})} / \\ &\quad [\rho(p_H)^{s_0}(1-p_H)^{M-s_0}(p_H)^{s(h_{T+1})}(1-p_H)^{T-s(h_{T+1})} \\ &\quad + (1-\rho)(p_L)^{s_0}(1-p_L)^{M-s_0}(p_L)^{s(h_{T+1})}(1-p_L)^{T-s(h_{T+1})}] \\ &= \frac{\rho}{\rho + (1-\rho)} \frac{(p_L)^{s_0+s(h_{T+1})}(1-p_L)^{M+T-s_0-s(h_{T+1})}}{(p_H)^{s_0+s(h_{T+1})}(1-p_H)^{M+T-s_0-s(h_{T+1})}} \end{aligned}$$

である。これは、我々の投資選択問題がそうである $p_H + p_L = 1$ の場合には、

$$\mu((s_0, M), (s(h_{T+1}), T)) = \frac{\rho}{\rho + (1-\rho)\left(\frac{p_L}{p_H}\right)^{s_0 + s(h_{T+1}) - (M+T-s_0-s(h_{T+1}))}}$$

となる。つまり、事前の M 回ベルヌーイ試行とプレイ履歴 h_{T+1} を合わせた Success 総回数から Failure 総回数をひいた Success 総優越回数

$$n = n((s_0, M), (s(h_{T+1}), T)) = s_0 + s(h_{T+1}) - (M+T-s_0-s(h_{T+1}))$$

で決まる。我々の問題のパラメーター $p_H = \frac{2}{3}$, $p_L = \frac{1}{3}$, $\rho = \frac{1}{2}$ の下で、Success 総優越回数 $n = -4, \dots, 0$ に対応する posteriors μ を計算すると次の表 1 の第 2 列のようになる。⁶⁾

表 1

n	μ	$\bar{p}(\mu)$	$V(\mu)$
0	0.5	0.606682	1.965720
-1	0.333333	0.564993	0.767161
-2	0.2	0.510848	0.091521
-3	0.111111	0.454301	0
-4	0.05882	0.407225	0

表 1 に求めた posteriors μ について、定理 4 に言う $\bar{p}(\mu)$ を、 $\delta = 0.95$ の下で定理 6 の公式に従って計算すると表 1 の第 3 列が得られる。これに定理 4 を適用すると、Success 総優越回数が $n = -2, -1, 0$ なら X-arm を選択し、 $n = -4, -3$ なら Y-arm を選択すべきである。定理 6 の公式によると $\bar{p}(\mu)$ は μ の増加関数であり、 $\mu((s_0, M), (s(h_{T+1}), T))$ の計算式によると μ は n の増加関数であるから、 $n = -2, -1, 0$ のみならず $n \geq -2$ である任意の整数 n の Success 総優越回数の下では X-arm を選択し、 $n \leq -3$ である任意の整数 n の Success 総優越回数の下では Y-arm を選択すべきである、と結論される。こうして、次が分かった。

命題 1

我々の M 熟練プレーヤー 1 人でプレイする投資選択問題の最適戦略は、「Success 総優越回数 n が -2 以上である限り X-arm を選択し続け、それが -3 となった次の選択からは Y-arm を選択し続ける」ことである。

3 問題の定式化と主結果

3.1 チームでプレイする場合の問題の定式化

M 熟練プレーヤー 1 人でプレイする投資選択問題の分析結果を用いて、チームでプレイする元々の投資選択問題を研究する。

我々の投資選択問題は、次のように定式化できる。奇数期目に M 熟練プレーヤーが選択し、偶数期目に未熟練プレーヤーが選択する交代手番を考察しよう。(逆の交代手番も全く同様である。) 奇数期 $t=1, 3, \dots$ に行う選択の仕方を記述する、 M 熟練プレーヤーの戦略 s_t^M は、ありうる h_t にその時にする選択 a_t を対応させる関数 $s_t^M: h_t \mapsto a_t$ である。同様に、偶数期 $t=2, 4, \dots$ に行う選択の仕方を記述する、未熟練プレーヤーの戦略 s_t^o は、ありうる h_t にその時にする選択 a_t を対応させる関数 $s_t^o: h_t \mapsto a_t$ である。 M 熟練プレーヤーの戦略 $s^M = (s_t^M)_{t=1, 3, \dots}$ と未熟練プレーヤーの戦略 $s^o = (s_t^o)_{t=2, 4, \dots}$ を組み合わせてできる $s = (s^M, s^o) = (s_1^M, s_2^o, s_3^M, s_4^o, \dots)$ をこのチームのチーム戦略という。

1人でプレイする場合と全く同様に、X-arm の成功確率が p の下でチーム戦略 s が生成するプレイ $H = ([a_1, Z_1], [A_2, Z_2], \dots)$ の測度が生成される。この測度によって、想定している p の下でチームがその戦略 s によるプレイから期待できる利得、つまりチーム戦略 s の期待利得

$$U(p, s) = E_H \left[\sum_{t=1}^{\infty} \delta^{t-1} u(Z_t) \mid p, s \right]$$

が定まる。

ただし、実際には、2人のプレーヤーが現に直面している X-arm の成功確率 p は、それを直接確かめる手だてではない。そして、プレイにさしかかる段でプレーヤーが $p = p_H$ である可能性に付す確率、つまりプレーヤーの initial belief は、 M 熟練プレーヤーと未熟練プレーヤーとで、そのあり方が異なる。それを反映して、自分の戦略選択の評価方法が異なる。

すなわち、未熟練プレーヤーの戦略を s^o と予想する M 熟練プレーヤーが、事前の M 回ベルヌーイ試行で Success が s_0 回だったときに、自分が戦略 s^M を取ることから期待する利得は

$$U(\mu(s_0, M), (s^M, s^o))$$

である。

他方、未熟練プレーヤーが、事前の M 回ベルヌーイ試行のありうる結果ごとに M 熟練プレーヤーがとる戦略を予想したもの、つまり、事前の M 回ベルヌーイ試行で Success が s_0 回だったときに戦略 $s^M[s_0]$ をとるということを $s_0=0, \dots, M$ について考えておいたもの $\beta^M = (s^M[s_0])_{s_0=0, \dots, M}$ を、 M 熟練プレーヤーのベイズ戦略と言う。今、未熟練プレーヤーが戦略 s^o を取る時、事前の M 回ベルヌーイ試行で Success が s_0 回となり、 M 熟練プレーヤーが戦略 $s^M[s_0]$ をとるなら、未熟練プレーヤーが期待できる利得は、

$$U(\mu(s_0, M), (s^M[s_0], s^o))$$

である。しかし、 M 熟練プレーヤーによる事前の M 回ベルヌーイ試行の結果 s_0 は、未熟練プレーヤーはそれを知ることとはできない。彼は、 s_0 が M 回ベルヌーイ試行の成功回数を表す確

率変数 S_0 の実現値である、ということだけを知っている。従って、 M 熟練プレーヤーのベイズ戦略を $\beta^M = (\beta^M[s_0])_{s_0=0, \dots, M}$ と予想する未熟練プレーヤーが、自分が戦略 s^0 を取ることから期待する利得は、実際には

$$E_{S_0}[U(\mu(S_0, M), (s^M[s_0], s^0))] = \sum_{s_0=0}^M \left[\frac{1}{2} \binom{M}{s_0} (p_H)^{s_0} (1-p_H)^{M-s_0} \right. \\ \left. + \frac{1}{2} \binom{M}{s_0} (p_L)^{s_0} (1-p_L)^{M-s_0} \right] U(\mu(s_0, M), (s^M[s_0], s^0))$$

である。⁷⁾

このように定義される利得の下でのベイズ均衡を $(\hat{\beta}, \hat{s}^0)$ とせよ。ならば、このチームでは、 M 熟練プレーヤーの事前の M 回ベルヌーイ試行で Success が s_0 回だったときには、チーム戦略 $\hat{s}[s_0] = (\hat{s}^M[s_0], \hat{s}^0)$ に従ったプレイが起こると予測される。

3.2 主結果

同じ投資選択問題を、 M 熟練プレーヤーと未熟練プレーヤーのチームがプレイする場合に、それがどの程度効率的であるかを、ベイズ均衡となるチーム戦略の分析を通して明らかにしたい。

チームのパフォーマンスを、 M 熟練プレーヤーによる事前の M 回ベルヌーイ試行が行われる前の時点での、チームの期待利得として定義する。それは、ベイズ均衡 $(\hat{\beta}, \hat{s}^0)$ の下で未熟練プレーヤーが期待する利得に他ならない。

それを、その投資選択問題を M 熟練プレーヤーが 1 人でプレイするなら同じ時点で期待できる利得と比較する。彼は、事前の M 回ベルヌーイ試行で Success が s_0 回となったなら、戦略 $s^*[\mu(s_0, M)]$ に従ってプレイすると予測される。その戦略からの期待利得は

$$U(\mu(s_0, M), s^*[\mu(s_0, M)])$$

である。従って、事前の試行の前の時点で期待できる利得は

$$E_{S_0}[U(\mu(S_0, M), s^*[\mu(S_0, M)])] = \sum_{s_0=0}^M \left[\frac{1}{2} \binom{M}{s_0} (p_H)^{s_0} (1-p_H)^{M-s_0} \right. \\ \left. + \frac{1}{2} \binom{M}{s_0} (p_L)^{s_0} (1-p_L)^{M-s_0} \right] U(\mu(s_0, M), s^*[\mu(s_0, M)])$$

である。

各 $s_0=0, \dots, M$ について、 M 熟練プレーヤーが 1 人でプレイする時の戦略 $s^*[\mu(s_0, M)]$ は問題 $P[\mu(s_0, M)]$ の解だから、

$$U(\mu(s_0, M), s^*[\mu(s_0, M)]) \geq U(\mu(s_0, M), (s^M[s_0], s^0))$$

が成り立つ。ゆえに、 $s_0=0, \dots, M$ の実現について期待値をとると次が成り立つ。

命題 2

我々の投資選択問題を M 熟練プレーヤーと未熟練プレーヤーのチームが交代手番でプレイする場合のパフォーマンスは、それを M 熟練プレーヤー 1 人がプレイする場合のパフォーマンスを上回ることではない、すなわち

$$E_{S_0}[U(\mu(S_0, M), s^*[\mu(S_0, M)])] \geq E_{S_0}[U(\mu(S_0, M), (s^M[S_0], s^0))]$$

この事実から、同じ投資選択問題をチームでプレイすることの非効率性を、 M 熟練プレーヤー 1 人がプレイする場合を基準とした、ベイズ均衡 $(\hat{\beta}, \hat{s}^0)$ によるプレイのパフォーマンスのロス

$$L(M) = E_{S_0}[U(\mu(S_0, M), s^*[\mu(S_0, M)])] - E_{S_0}[U(\mu(S_0, M), (s^M[S_0], \hat{s}^0))]$$

の大きさを測ることができる。我々の主結果は、このロスに関する次の事実である。

命題 3

我々の投資選択問題を M 熟練プレーヤーと未熟練プレーヤーのチームでプレイすることの非効率性 $L(M)$ は、次のようになる。

[A] 奇数期目に M 熟練プレーヤーが選択し、偶数期目に未熟練プレーヤーが選択する交代手番では

1. M が奇数なら、 $L(M) = 0$
2. M が偶数なら、 $L(M) > 0$ 、しかし $\lim_{M \rightarrow \infty} L(M) = 0$

[B] 奇数期目に未熟練プレーヤーが選択し、偶数期目に M 熟練プレーヤーが選択する交代手番では $L(M) > 0$ 、そして $\lim_{M \rightarrow \infty} L(M) = \frac{1}{6} > 0$

4 主結果の証明

4.1 命題 3 の [A] の 1. の証明

奇数期目に M 熟練プレーヤーが選択し、偶数期目に未熟練プレーヤーが選択する交代手番で、 M が奇数とせよ。次の、 M 熟練プレーヤーのベイズ戦略と未熟練プレーヤーの戦略の組を考えよ。

$\hat{\beta}^M = (\hat{s}^M[S_0])_{S_0=0, \dots, M}$ ：各奇数期で、 M 熟練プレーヤーが 1 人でプレイする時の最適戦略の対応する奇数期の戦略に従って行動する。つまり、

$$\forall s_0 = 0, \dots, M, \forall t = 1, 3, \dots, \forall h_t : \hat{s}_t^M[s_0](h_t) = s_t^*[\mu(s_0, M)](h_t)$$

$\hat{s}^0$ ：各偶数期で、直前の期に M 熟練プレーヤーが選択したのと同じ arm を選択する。

つまり、

$$\forall t = 2, 4, \dots, \forall h_t = ([a_1, z_1], \dots, [a_{t-1}, z_{t-1}]) : \hat{s}^0(h_t) = a_{t-1}$$

この戦略の組 $(\hat{\beta}, \hat{s}^0)$ は、 $s_0 = 0, \dots, M$ ごとに、それが生成するプレイ $H = ([a_1, Z_1], [A_2, Z_2], \dots)$ の測度が、 M 熟練プレーヤー 1 人でプレイする場合の最適戦略 $s^*[\mu(s_0, M)]$ が生

成するプレイの測度と等しいことを示す。命題 1 により、我々の投資選択問題を M 熟練プレイヤー 1 人でプレイする場合の最適戦略 $s^*[\mu(s_0, M)]$ は、 $T+1$ 期に選択する時、 T 期終了時点での Success 総優越回数が初めて

$$n = n((s_0, M), (s(h_{T+1}), T)) = s_0 + s(h_{T+1}) - (M + T - s_0 - s(h_{T+1})) = -3$$

となった場合に X-arm から Y-arm に切り替えるというものだった。これは、X-arm から Y-arm に切り替える選択が行われる $T+1$ 期が

$$T+1 = 2 + 2(s_0 + s(h_{T+1})) - M$$

であることを意味する。 s_0 と h_{t+1} とにかかわらず、 M が奇数ならば、この $T+1$ も奇数である。ゆえに、 M 熟練プレイヤー 1 人でプレイする場合の最適戦略 $s^*[\mu(s_0, M)]$ は、それに従ったプレイの過程では偶数期には X-arm から Y-arm に切り替える選択はしない、つまり直前の奇数期に自分が選択した arm をその偶数期に選択する。よって、各 $t=2, 4, \dots$ について、それまでのプレイ履歴 $h_t = ([a_1, z_1], \dots, [a_{t-1}, z_{t-1}])$ が $s^*[\mu(s_0, M)]$ によって生成されるプレイ履歴ならば $s_t^*[\mu(s_0, M)](h_t) = a_{t-1}$ が成り立つ。これは、 $s^*[\mu(s_0, M)]$ が生成するプレイの途上では、

$$\forall t=2, 4, \dots, \forall h_t = ([a_1, z_1], \dots, [a_{t-1}, z_{t-1}]) : \hat{s}^o(h_t) = s^*[\mu(s_0, M)](h_t)$$

が成り立つということである。ところが、 $t=1, 3, \dots$ については、プレイ履歴 h_t にかかわらず $\hat{s}_t^M[s_0](h_t) = s_t^*[\mu(s_0, M)](h_t)$ と構成してある。ゆえに、 s_0 によらず、チーム戦略 $\hat{s}[s_0] = (\hat{s}^M[s_0], \hat{s}^o)$ が生成するプレイは、 M 熟練プレイヤー 1 人でプレイする場合の最適戦略 $s^*[\mu(s_0, M)]$ が生成するプレイに限り、しかも両者の測度は同じになる。

各 $s_0=0, \dots, M$ について、生成されるプレイの測度の一致により、パフォーマンスの一致

$$U(\mu(s_0, M), (\hat{s}^M[s_0], \hat{s}^o)) = U(\mu(s_0, M), s^*[\mu(s_0, M)])$$

が成り立つ。

ここに示したパフォーマンスの一致は、戦略 $s^*[\mu(s_0, M)]$ の最適性により、

$$U(\mu(s_0, M), (\hat{s}^M[s_0], \hat{s}^o)) = \max_{s^M} U(\mu(s_0, M), (s^M[s_0], \hat{s}^o))$$

つまり、 $\hat{s}^M[s_0]$ が $\hat{s}^o$ に対するベイズ最適反応であることを保証する。全く同様に、どの $s_0=0, \dots, M$ の下でも $\hat{s}^o$ は $\hat{s}^M[s_0]$ に対するベイズ最適反応であることが保証されるから、実際には s_0 を知ることがなくても

$$E_{S_0}[U(\mu(S_0, M), (\hat{s}^M[S_0], \hat{s}^o))] = \max_{s^M} E_{S_0}[U(\mu(S_0, M), (s^M[S_0], \hat{s}^o))]$$

つまり、 $\hat{s}^o$ は $\hat{\beta}$ に対する最適反応である。ゆえに、戦略の組 $(\hat{\beta}, \hat{s}^o)$ はベイズ均衡である。

$s_0=0, \dots, M$ ごとのパフォーマンスの一致により、ベイズ均衡 $(\hat{\beta}, \hat{s}^o)$ が実現するチームのパフォーマンスも

$$E_{S_0}[U(\mu(S_0, M), (\hat{s}^M[S_0], \hat{s}^o))] = E_{S_0}[U(\mu(S_0, M), s^*[\mu(S_0, M)])]$$

である。よって、 $L(M)=0$ が示された。(証明終)

4.2 命題3の[A]の2.の証明

奇数期目に M 熟練プレーヤーが選択し、偶数期目に未熟練プレーヤーが選択する交代手番で、 M が偶数とせよ。

M 熟練プレーヤーのベイズ戦略と未熟練プレーヤーの戦略のどのような組 (β, s^o) も、 M 熟練プレーヤーが1人でプレイする場合のパフォーマンスを達成できないことを示す。そうできるとせよ。ならば、すべての $s_0 = 0, \dots, M$ について、チーム戦略 $(s^M[s_0], s^o)$ は、最適戦略 $s^*[\mu(s_0, M)]$ が生成するプレイの測度を再現できないからならない。 M 熟練プレーヤーが1人でプレイする場合、命題1により、「 $S_0 = \frac{M}{2}$ が実現し、 M 熟練プレーヤーは1期目に X-arm を選択し、その結果は Failure で、 M 熟練プレーヤーは2期目に再び X-arm を選択する」というプレイと、「 $S_0 = \frac{M}{2} - 1$ が実現し、 M 熟練プレーヤーは1期目に X-arm を選択し、その結果は Failure で、 M 熟練プレーヤーは2期目は Y-arm を選択する」というプレイは、ともに正の確率で起こる。チームでプレイする場合、未熟練プレーヤーは $S_0 = \frac{M}{2}$ と $S_0 = \frac{M}{2} - 1$ を識別できないから、彼のどのような戦略 s^o も、2期目までの同じプレイ履歴 $h_2 = ([a_1, z_1]) = ([X\text{-arm}, \text{Failure}])$ の下での選択 $s_2^o([X\text{-arm}, \text{Failure}])$ が、 M 熟練プレーヤーが1人でプレイする場合の2期目の選択を再現できない。よって矛盾である。こうして、 $L(M) > 0$ が示された。

$\lim_{M \rightarrow \infty} L(M) = 0$ を示す。 M が偶数の時、事前の M 回ベルヌーイ試行における Success の優越回数 $s_0 - (M - s_0) = 2s_0 - M$ は偶数であり、各奇数期で $t = 1, 3, \dots$ の Success 総優越回数 $s_0 + s(h_t) - (M + (t-1) - s_0 - s(h_t)) = 2(s_0 + s(h_t)) - M - (t-1)$ も偶数であることに注意しよう。そして、次の、 M 熟練プレーヤーのベイズ戦略と未熟練プレーヤーの戦略の組を考えよ。

$\hat{\beta}^M = (\hat{s}^M[s_0])_{s_0=0, \dots, M}$ ：事前の M 回ベルヌーイ試行における Success の優越回数が $s_0 - (M - s_0) \leq -4$ ならば、つまり $s_0 \leq \frac{M}{2} - 2$ ならば、すべての奇数期で Y-arm を選択する。つまり、

$$\forall t = 1, 3, \dots, \forall h_t : \hat{s}_t^M[s_0](h_t) = Y\text{-arm}$$

逆に $s_0 - (M - s_0) \geq -2$ ならば、つまり $s_0 \geq \frac{M}{2} - 1$ ならば、 $s_0 = M$ の場合を除いて、各奇数期 $t = 1, 3, \dots$ で Success の総優越回数が -2 となるまでは、つまり

$$s_0 + s(h_t) - (M + (t-1) - s_0 - s(h_t)) \geq 0$$

であり続ける限り X-arm を選択し続ける。そして、それが初めて -2 となった奇数期 $\bar{t}$ では、つまり

$$s_0 + s(h_{\bar{t}}) - (M + (\bar{t}-1) - s_0 - s(h_{\bar{t}})) = -2$$

では Y-arm をとり、 $\bar{t}$ 以後のすべての奇数期 $t = \bar{t} + 2, \bar{t} + 4, \dots$ では M 熟練プレーヤーが1人でプレイする時の最適戦略の対応する奇数期の戦略に従って行動する。つま

り,

$$\forall t = \bar{t} + 2, \bar{t} + 4, \dots, \forall h_t: \hat{s}_t^M[s_0](h_t) = s_t^*[\mu(s_0, M)](h_t)$$

最後に, $s_0 = M$ の場合は, すべての奇数期で, M 熟練プレーヤーが1人でプレイする時の最適戦略の対応する奇数期の戦略に従って行動する。つまり,

$$\forall t = 1, 3, \dots, \forall h_t: \hat{s}_t^M[M](h_t) = s_t^*[\mu(M, M)](h_t)$$

$\hat{s}^0$: 1期目に M 熟練プレーヤーが Y-arm を選択した場合, 2期目は Y-arm を選択する。3期目に M 熟練プレーヤーが X-arm を選択した場合, その時点で事前の M 回ベルヌーイ試行の S_0 の実現値はちょうど $s_0 - (M - s_0) = -2$ を満たす s_0 , つまり $s_0 = \frac{M}{2} - 1$ だったと想定し, 引き続くすべての偶数期 $t = 4, 6, \dots$ で, M 熟練プレーヤーが1人でプレイする時の最適戦略の対応する偶数期の戦略に従って行動する。つまり,

$$\forall t = 4, 6, \dots, \forall h_t = ([Y - \text{arm}, z_1], [Y - \text{arm}, z_2], [X - \text{arm}, z_3], \dots, [a_{t-1}, z_{t-1}])$$

$$: \hat{s}^0(h_t) = s_t^* \left[\mu \left(\frac{M}{2} - 1, M \right) \right](h_t)$$

他方, 3期目以降も M 熟練プレーヤーが Y-arm を選択し続ける限り, その次の偶数期で Y-arm を選択する。⁹⁾ 1期目に M 熟練プレーヤーが X-arm を選択した場合, M 熟練プレーヤーが X-arm を選択し続ける限り, その次の偶数期で X-arm を選択する。そして, それまでのプレイ履歴 h_i が $\frac{\bar{t}-1}{2} \geq s(h_i) \geq \frac{\bar{t}-1}{2} - \frac{M}{2}$ を満たすある奇数期 $\bar{t}$ で M 熟練プレーヤーが初めて Y-arm を選んだなら, その時点で事前の M 回ベルヌーイ試行の S_0 の実現値は

$$s_0 + s(h_{\bar{t}}) - (M + (\bar{t} - 1) - s_0 - s(h_{\bar{t}})) = -2$$

を満たす s_0 , つまり $s_0 = \frac{M}{2} + \frac{\bar{t}-1}{2} - (s(h_{\bar{t}}) + 1)$ だったと想定し, 引き続くすべての偶数期 $t = \bar{t} + 1, \bar{t} + 3, \dots$ で, M 熟練プレーヤーが1人でプレイする時の最適戦略の対応する偶数期の戦略に従って行動する。つまり,

$$\forall t = \bar{t} + 1, \bar{t} + 3, \dots,$$

$$\forall h_t = ([X - \text{arm}, z_1], \dots, [X - \text{arm}, z_{\bar{t}-1}], [Y - \text{arm}, z_{\bar{t}}], \dots, [a_{t-1}, z_{t-1}])$$

$$: \hat{s}^0(h_t) = s_t^* \left[\mu \left(\frac{M}{2} + \frac{\bar{t}-1}{2} - (s(h_{\bar{t}}) + 1), M \right) \right](h_t)$$

他方, プレイ履歴がある奇数期 $\bar{t}$ で初めて $s(h_i) = \frac{\bar{t}-1}{2} - \frac{M}{2}$ に達してもなおその $\bar{t}$ 期に M 熟練プレーヤーが X-arm を選んだなら, その時点で事前の M 回ベルヌーイ試行 S_0 の実現値は $s_0 = M$ だったと想定し, 引き続くすべての偶数期 $t = \bar{t} + 1, \bar{t} + 3, \dots$ で M 熟練プレーヤーが1人でプレイする時の最適戦略の対応する偶数期の戦略に従って行動する。つまり,

$$\forall t = \bar{t} + 1, \bar{t} + 3, \dots, \forall h_t = ([X - \text{arm}, z_1], \dots, [X - \text{arm}, z_{\bar{t}}], \dots, [a_{t-1}, z_{t-1}])$$

$$: \hat{s}^o(h_i) = s_i^* [\mu(M, M)](h_i)$$

この組 $(\hat{\beta}, \hat{s}^o)$ は、 M が十分大の時、ベイズ均衡をなすことを示す。まず、 $(\hat{\beta}, \hat{s}^o)$ が生成するプレイ $H = ([a_1, Z_1], [A_2, Z_2], \dots)$ の上で、事前の M 回ベルヌーイ試行の S_0 の実現値は、 $s_0 = \frac{M}{2}, \frac{M}{2} + 1, \dots, M$ に関して

$$s_0 + s(h_{\bar{t}}) - (M + (\bar{t} - 1) - s_0 - s(h_{\bar{t}})) = -2$$

が初めて成り立つ奇数期 $\bar{t}$ で fully revealed である。すなわち、 M 熟練プレーヤーが 1 期目に X-arm を選択した場合の Y-arm への切り替えは、 s_0 の値に応じて

$$s_0 = \frac{M}{2} \text{ ならば, プレイ履歴 } h_{\bar{t}} \text{ が初めて } s(h_{\bar{t}}) - ((\bar{t} - 1) - s(h_{\bar{t}})) = -2$$

$$s_0 = \frac{M}{2} + 1 \text{ ならば, プレイ履歴 } h_{\bar{t}} \text{ が初めて } s(h_{\bar{t}}) - ((\bar{t} - 1) - s(h_{\bar{t}})) = -4$$

$\vdots$

$$s_0 = M - 1 \text{ ならば, プレイ履歴 } h_{\bar{t}} \text{ が初めて } s(h_{\bar{t}}) - ((\bar{t} - 1) - s(h_{\bar{t}})) = -M$$

となる奇数期 $\bar{t}$ であり、 $s_0 = M$ ならばプレイ履歴 $h_{\bar{t}}$ が初めて

$$s(h_{\bar{t}}) - ((\bar{t} - 1) - s(h_{\bar{t}})) = -M$$

となる奇数期 $\bar{t}$ でも X-arm が選択される。また、 $s_0 = \frac{M}{2} - 1$ も、1 期目に Y-arm、2 期目に Y-arm が選択され、3 期目に X-arm が選ばれた時点で fully revealed である。

この $s_0 = \frac{M}{2}, \frac{M}{2} + 1, \dots, M$ と $s_0 = \frac{M}{2} - 1$ の full revelation の性質により、[A] の 1. の証明と同様に、命題 1 から、未熟練プレーヤーの戦略 s^o は、1 期目に M 熟練プレーヤーが Y-arm を選択した場合の 2 期目の選択が Y-arm であることを除いて、 $\hat{\beta}^M$ への最適反応であると結論される。

1 期目に M 熟練プレーヤーが Y-arm を選択した場合、 $\hat{\beta}^M = (\hat{s}^M[s_0])_{s_0=0, \dots, M}$ の構成より、事前の M 回ベルヌーイ試行の結果 S_0 の実現値は、 $s_0 = 0, \dots, \frac{M}{2} - 2, \frac{M}{2} - 1$ のいずれかである。このうち $s_0 = 0, \dots, \frac{M}{2} - 2$ のいずれかであることの条件付き確率は、

$$\begin{aligned} & \frac{\text{Prob}\left(S_0 = 0, \dots, \frac{M}{2} - 2\right)}{\text{Prob}\left(S_0 = 0, \dots, \frac{M}{2} - 2, \frac{M}{2} - 1\right)} \\ &= \left[\text{Prob}\left(S_0 = 0, \dots, \frac{M}{2} - 2, \frac{M}{2} - 1\right) + \frac{1}{2} \text{prob}\left(S_0 = \frac{M}{2}\right) \right] - \text{Prob}\left(S_0 = \frac{M}{2} - 1\right) \\ & \quad - \frac{1}{2} \text{Prob}\left(S_0 = \frac{M}{2}\right) / \left[\text{Prob}\left(S_0 = 0, \dots, \frac{M}{2} - 2, \frac{M}{2} - 1\right) + \frac{1}{2} \text{Prob}\left(S_0 = \frac{M}{2}\right) \right] \\ & \quad - \frac{1}{2} \text{Prob}\left(S_0 = \frac{M}{2}\right) \end{aligned}$$

である。この値は、

$$\lim_{M \rightarrow \infty} \left[\text{Prob} \left(S_0 = 0, \dots, \frac{M}{2} - 2, \frac{M}{2} - 1 \right) + \frac{1}{2} \text{Prob} \left(S_0 = \frac{M}{2} \right) \right] > 0$$

$$\lim_{M \rightarrow \infty} \text{Prob} \left(S_0 = \frac{M}{2} - 1 \right) = \lim_{M \rightarrow \infty} \text{Prob} \left(S_0 = \frac{M}{2} \right) = 0$$

より、 M が大きくなるにつれて 1 に収束する¹⁰⁾。命題 1 により、 $s_0 = 0, \dots, \frac{M}{2} - 2$ のいずれであっても、Y-arm は X-arm より厳密に望ましい。ゆえに、1 期目に M 熟練プレーヤーが Y-arm を選択した場合の未熟練プレーヤーによる 2 期目の Y-arm の選択は、 M が十分大の時は $\hat{\beta}^M$ への最適反応である。

今度は、 $\hat{\beta}^M$ が s° への最適反応であることを示そう。 s_0 の値で分類して、 $\hat{s}^M[s_0]$ が s° へのベイズ最適反応であることを示す。 $s_0 = 0, \dots, \frac{M}{2} - 2$ を考えよ。ならば、その事前の M 回ベルヌーイ試行における Success の優越回数は、 $s_0 - (M - s_0) \leq -4$ である。この時、もし M 熟練プレーヤーが 1 人でプレイするとしたら、命題 1 により、 $t = 1, 2, \dots$ のすべての期で Y-arm を選択することが最適である。 $\hat{s}^M[s_0]$ は、すべての奇数期 $t = 1, 3, \dots$ で Y-arm を指示し、そのとき s° もまたすべての偶数期 $t = 2, 4, \dots$ で Y-arm を指示するから、チーム戦略 $(\hat{s}^M[s_0], s^\circ)$ は、 M 熟練プレーヤーが 1 人でプレイする時の最適戦略 $s^*[\mu(s_0, M)]$ によるプレイを再現できる。よって、 $\hat{s}^M[s_0]$ は s° へのベイズ最適反応である。

$s_0 = M$ を考えよ。この時、もし M 熟練プレーヤーが 1 人でプレイするとしたら、命題 1 により、プレイ履歴 h_t が

$$M + s(h_t) - (M + (t-1) - M - s(h_t)) \geq -2$$

つまり $s(h_t) \geq \frac{t-1}{2} - \frac{M}{2} - 1$ である限り X-arm を選択し続け、初めて $s(h_t) = \frac{t-1}{2} - \frac{M}{2} - \frac{3}{2}$ となった t 期以後は Y-arm に切り替える。 $s(h_t) = \frac{t-1}{2} - \frac{M}{2} - \frac{3}{2}$ が起こるのは、ある奇数期 $\bar{t}$ で初めて $s(h_t) = \frac{\bar{t}-1}{2} - \frac{M}{2} - 1$ が成立し、引き続いて X-arm が選択され続けた結果 $\bar{t}$ 以後の Success 優越回数が -1 となる t である。 $\hat{s}^M[M]$ は、すべての奇数期で、この通りの行動を指示する。 s° は、 $s(h_t) = \frac{\bar{t}-1}{2} - \frac{M}{2} - 1$ が成立するまですべての偶数期で X-arm を選択し続け、 $s(h_t) = \frac{\bar{t}-1}{2} - \frac{M}{2} - 1$ が成立した奇数期 $\bar{t}$ で M 熟練プレーヤーが X-arm を選択した後のすべての偶数期で、あたかも M 熟練プレーヤーが 1 人でプレイする場合の行動と同じ行動をとる。ゆえに、チーム戦略 $(\hat{s}^M[M], s^\circ)$ は、 M 熟練プレーヤーが 1 人でプレイする時の最適戦略 $s^*[\mu(M, M)]$ によるプレイを再現できる。よって、 $\hat{s}^M[M]$ は s° へのベイズ最適反応である。

$s_0 = \frac{M}{2}, \dots, M-1$ を考えよ。 $s_0 = M$ の場合と同様の理由で、プレイ履歴 h_t が初めて

$$s_0 + s(h_t) - (M + (t-1) - s_0 - s(h_t)) = -2$$

となる奇数期 $\bar{t}$ よりも後の任意の奇数期で、 $\hat{s}^M[s_0]$ が指示する行動は s° へのベイズ最適反応である。 $\bar{t}$ までは X-arm を選択するという $\hat{s}^M[s_0]$ の指示も、また、命題 1 により最適であ

る。よって、 $\bar{i}$ で Y-arm を選択することが最適であることを示せばよい。

その為に、まず、 $\bar{i}$ で M 熟練プレーヤーが Y-arm を選択し、その後引き続いてチーム戦略 $(\hat{s}^M[s_0], s^0)$ に従ったプレイが行われることの、 $\bar{i}$ から見た利得を計算する。 $\bar{i}$ 期の利得は Y-arm の利得

$$E_Z[u(Z) | \mu, a = Y\text{-arm}] = \frac{1}{2} \cdot 1 + \frac{1}{2} \cdot (-1) = 0$$

である。 $\bar{i}$ 期の Y-arm の選択は、 M 熟練プレーヤーと未熟練プレーヤーの双方に同じ posterior $\mu((s_0, M), (s(h_{\bar{i}}), \bar{i}-1))$ をもたらす。それは、Success 総優越回数が

$$n = n((s_0, M), (s(h_{\bar{i}}), \bar{i}-1)) = s_0 + s(h_{\bar{i}}) - (M + (\bar{i}-1) - s_0 - s(h_{\bar{i}})) = -2$$

である時の posterior である。 $\bar{i}$ 以降は、この posterior から出発する時の、 M 熟練プレーヤーが 1 人でプレイする場合の行動が起こるから、 $\bar{i}$ 以降の継続利得は、Success 総優越回数が $n = -2$ の時の value function $V(\mu((s_0, M), (s(h_{\bar{i}}), \bar{i}-1)))$ の値に他ならない。value function の値 $V(\mu)$ を、Success 総優越回数が $n = -100, \dots, 100$ の 201 通りについて、逐次近似法で計算して求める。定理 2 の Bellman 方程式の右辺にある μ^+, μ^-, μ について、それが $n = -100, \dots, 100$ のすべての場合について $V(\mu) = 0$ を代入する。その右辺の最大化問題を解いて左辺の $V(\mu)$ とする。ただし、左辺が $n = 100$ の時の右辺の μ^+ は $n = 101$ に当たるはずだが、トランケートして $n = 100$ の場合と見なす。左辺が $n = -100$ の時の μ^- の扱いも同様である。左辺で求めた $V(\mu)$ を、再び右辺に代入して、この新しい右辺について、その右辺の最大化問題を解いて左辺の $V(\mu)$ とする。この手続きを 1000 回繰り返した。その結果を $n = -4, \dots, 0$ について示したものが表 1 の第 4 列である。示した有効数字 6 桁について、反復計算の 1000 回目でその数字の変動が起こらないので、逐次近似は有効である。それによると、 $V(n = -2) = 0.091521$ である。ゆえに、 $\bar{i}$ で M 熟練プレーヤーが Y-arm を選択し、その後引き続いてチーム戦略 $(\hat{s}^M[s_0], s^0)$ に従ったプレイが行われることの、 $\bar{i}$ から見た利得は

$$E_Z[u(Z) | \mu, a = Y\text{-arm}] + \delta V(n = -2) = 0 + (0.95) \cdot (0.091521) = 0.086945$$

である。

次に、 $\bar{i}$ で M 熟練プレーヤーが X-arm の選択へ deviate し、その後引き続いて s^0 への最適反応戦略をとることからの、 $\bar{i}$ から見た利得を評価する。 $\bar{i}$ で M 熟練プレーヤーが X-arm を選択すると、 s^0 の構成方法より、その次の期 $\bar{i}+1$ で未熟練プレーヤーは X-arm を選択する。従って、 M 熟練プレーヤーの最適 deviation とは、 $\bar{i}$ と $\bar{i}+1$ で続けて X-arm が選択された後に s^0 に対する最適反応戦略をとることである。その利得は、 $\bar{i}$ と $\bar{i}+1$ で続けて X-arm が選択された後は M 熟練プレーヤーが 1 人でプレイする場合の利得を越えることはない。後者の利得を計算しよう。今 $\bar{i}$ で、Success の総優越回数が $n = -2$ となっている。 $\bar{i}$ と

$\bar{i}+1$ で続けて X-arm を選択することがもたらす、 $\bar{i}+2$ までのプレイ履歴 $h_{\bar{i}+2}$ の下での Success の総優越回数は、 $X_{\bar{i}}$ と $X_{\bar{i}+1}$ の実現値に応じて、それがともに Success なら $n=0$ 、一方が Success でもう一方が Failure なら $n=-2$ 、ともに Failure なら $n=-4$ となる。 $n=0$ となるなら $\bar{i}+2$ 以降の継続利得は表 1 の $V(n=0)=1.96572$ 、 $n=-2$ なら $V(n=-2)=0.091521$ 、 $n=-4$ なら $V(n=-4)=0$ である。Success の総優越回数が $n=-2$ である $\bar{i}$ での M 熟練プレーヤーの posterior は、表 1 の第 2 列より $\mu(n=-2)=0.2$ だから、 $\bar{i}$ と $\bar{i}+1$ で X-arm を選択する時の 1 回あたりの成功確率は $0.2 \times \frac{2}{3} + 0.8 \times \frac{1}{3} = \frac{2}{5}$ と評価されていることになる。 $\bar{i}$ と $\bar{i}+1$ の X-arm からの期待利得は、この評価確率で計算される。そして、 $\bar{i}+2$ 以降の継続利得を計算する上での、 $\bar{i}+2$ での Success の総優越回数は、 $n=0$ が確率 $\left(\frac{2}{5}\right)^2$ で、 $n=-2$ が確率 $2 \times \left(\frac{2}{5}\right)\left(\frac{3}{5}\right)$ で、 $n=-4$ が確率 $\left(\frac{3}{5}\right)^2$ で起こると見積もられる。よって、求める利得は

$$\begin{aligned} & \frac{2}{5} \cdot 1 + \frac{3}{5} \cdot (-1) + \delta \left(\frac{2}{5} \cdot 1 + \frac{3}{5} \cdot (-1) \right) + \delta^2 \left[\left(\frac{2}{5} \right)^2 V(n=0) + 2 \left(\frac{2}{5} \right) \left(\frac{3}{5} \right) V(n=-2) \right. \\ & \quad \left. + \left(\frac{3}{5} \right)^2 V(n=-4) \right] \\ &= \frac{2}{5} \cdot 1 + \frac{3}{5} \cdot (-1) + (0.95) \left(\frac{2}{5} \cdot 1 + \frac{3}{5} \cdot (-1) \right) \\ & \quad + (0.95)^2 \left[\left(\frac{2}{5} \right)^2 \cdot (1.96572) + 2 \left(\frac{2}{5} \right) \left(\frac{3}{5} \right) \cdot (0.091521) + \left(\frac{3}{5} \right)^2 \cdot (0) \right] \\ &= -0.066503 \end{aligned}$$

である。

以上の 3 つの利得を比較すると、

$\bar{i}$ とそれ以後で $\mathbf{s}^M[s_0]$ に従うことからの利得 (0.086945)

$> \bar{i}$ で X-arm に deviate し、 $\bar{i}+2$ 以後は M 熟練プレーヤーが 1 人でプレイする場合の利得 (-0.066503)

$\geq \bar{i}$ で X-arm に deviate し、 $\bar{i}+2$ 以後は「 $\bar{i}$ と $\bar{i}+1$ で続けて X-arm が選択された後の $\mathbf{s}^0$ に対する最適反応戦略」をとる場合の利得

がわかる。これは、 $\bar{i}$ で Y-arm を選択することが最適である、ということである。

最後に、 $s_0 = \frac{M}{2} - 1$ を考えよ。その時の M 熟練プレーヤーの戦略 $\mathbf{s}^M\left[\frac{M}{2} - 1\right]$ が指示する「1 期目に Y-arm、2 期目の Y-arm の後の 3 期目に X-arm」という行動が $\mathbf{s}^0$ へのベイズ最適反応であることは、 $s_0 = \frac{M}{2}, \dots, M-1$ の場合と同様に確かめられる。すなわち、 $s_0 = \frac{M}{2} - 1$ の場合とは、

$$s_0 + s(h_{\bar{i}}) - (M + (\bar{i} - 1) - s_0 - s(h_{\bar{i}})) = -2$$

が $\bar{i}=1$ で成り立っている場合である。ただし、 $s_0=\frac{M}{2}, \dots, M-1$ の場合 $\hat{s}^M[s_0]$ に従って奇数期 $\bar{i}$ で Y-arm を選択することはその次の $\bar{i}+1$ 期の未熟練プレーヤーの番から直ちに $s^*[\mu(s_0, M)]$ に従うチーム・プレイを実現したが、 $s_0=\frac{M}{2}-1$ の場合 $\hat{s}^M\left[\frac{M}{2}-1\right]$ に従って $\bar{i}=1$ で Y-arm を選択することは、次の $\bar{i}+1=2$ 期の未熟練プレーヤーによる Y-arm の選択に続く $\bar{i}+2=3$ 期の M 熟練プレーヤー自身による X-arm の選択からである。従って、その利得は

$$0 + \delta \cdot 0 + \delta^2 V(n=-2) = 0 + (0.95) \cdot 0 + (0.95)^2 \cdot (0.091521) = 0.082598$$

である。しかし、これは、なお $\bar{i}=1$ で X-arm に deviate し、 $\bar{i}+2=3$ 以後は M 熟練プレーヤーが 1 人でプレイする場合の利得である -0.066503 を上回る。よって、 $\hat{s}^M\left[\frac{M}{2}-1\right]$ も s^0 へのベイズ最適反応である。こうして、 $(\hat{\beta}, \hat{S}^0)$ は十分大きな M の下でベイズ均衡であることが分かった。

チーム戦略 $(\hat{\beta}, \hat{S}^0)$ のパフォーマンスを調べる。この戦略の組に代えて、次の M 熟練プレーヤーのベイズ戦略と未熟練プレーヤーの戦略の組を考えよ。

$\tilde{\beta}^M = (\tilde{s}^M[s_0])_{s_0=0, \dots, M}$ ：事前の M 回ベルヌーイ試行における Success の優越回数が $s_0 - (M - s_0) \leq -4$ ならば、つまり $s_0 \leq \frac{M}{2} - 2$ ならば、すべての奇数期で Y-arm を選択する。つまり、

$$\forall t = 1, 3, \dots, \forall h_t : \tilde{s}_t^M[s_0](h_t) = \text{Y-arm}$$

逆に $s_0 - (M - s_0) \geq -2$ ならば、つまり $s_0 \geq \frac{M}{2} - 1$ ならば、各奇数期 $t=1, 3, \dots$ で Success の総優越回数が -2 となるまでは、つまり

$$s_0 + s(h_t) - (M + (t-1) - s_0 - s(h_t)) \geq 0$$

であり続ける限り X-arm を選択し続け、それが初めて -2 となった奇数期 $\bar{t}$ 、つまり

$$s_0 + s(h_{\bar{t}}) - (M + (\bar{t}-1) - s_0 - s(h_{\bar{t}})) = -2$$

となる $\bar{t}$ とそれ以後のすべての奇数期 $t=\bar{t}, \bar{t}+2, \bar{t}+4, \dots$ で Y-arm を選択する。

$\tilde{s}^0$ ：各偶数期で、直前の期に M 熟練プレーヤーが選択したのと同じ arm を選択する。つまり、

$$\forall t = 2, 4, \dots, \forall h_t = ([a_1, z_1], \dots, [a_{t-1}, z_{t-1}]) : \tilde{s}_t^0(h_t) = a_{t-1}$$

$(\tilde{\beta}, \tilde{S}^0)$ の分析より、 $(\tilde{\beta}, \tilde{S}^0)$ はベイズ均衡をなさない、しかしもしそれがプレイされるなら $(\tilde{\beta}, \tilde{S}^0)$ のパフォーマンスは $(\hat{\beta}, \hat{S}^0)$ のパフォーマンスを厳密に下回る、ということがわかる。この $(\tilde{\beta}, \tilde{S}^0)$ のパフォーマンスを、 M 熟練プレーヤーが 1 人でプレイする場合の最適戦略 $(s^*[\mu(s_0, M)])_{s_0=0, \dots, M}$ のパフォーマンスと比較する。

$(s^*[\mu(s_0, M)])_{s_0=0, \dots, M}$ によるプレイと $(\tilde{\beta}, \tilde{S}^0)$ によるプレイの違いは、前者が

$$s_0 + s(h_{\bar{t}+1}) - (M + \bar{t} - s_0 - s(h_{\bar{t}+1})) = -3$$

が初めて成立する奇数期 $\bar{t}+1$ で Y-arm への切り替えをするのに対し、後者はその 1 期前の

$$s_0 + s(h_{\bar{t}}) - (M + (\bar{t} - 1) - s_0 - s(h_{\bar{t}})) = -2$$

が成立する偶数期 $\bar{t}$ で Y-arm への切り替えをすることであり、それ以外にはない。問題の偶数期 $\bar{t}$ で Y-arm への切り替えをすると、その $\bar{t}$ での最適戦略からの継続利得 $V(n=-2) = 0.091521$ を失って継続利得 0 を得る。ゆえに、求めるパフォーマンスの差は

$$\begin{aligned} & E_{S_0}[U(\mu(S_0, M), s^*[\mu(S_0, M)])] - E_{S_0}[U(\mu(S_0, M), (\tilde{s}^M[S_0], \tilde{s}^0))] \\ &= E_{S_0}\left[\sum_{\bar{t}=1,3,\dots} \delta^{\bar{t}-1} V(n=-2) \times \text{Prob}(\bar{t} \text{ で初めて } S_0 + s(h_{\bar{t}}) - (M + (\bar{t} - 1) - S_0 - s(h_{\bar{t}})) = -2)\right] \\ &< E_{S_0}\left[\sum_{\bar{t}=1,3,\dots} \delta^{\bar{t}-1} V(n=-2) \times \text{Prob}(\exists t < \infty \text{ s.t. } S_0 + s(h_t) - (M + (t - 1) - S_0 - s(h_t)) = -2)\right] \\ &= E_{S_0}[\text{Prob}(\exists t < \infty \text{ s.t. } S_0 + s(h_t) - (M + (t - 1) - S_0 - s(h_t)) = -2)] \\ &\quad \times \sum_{\bar{t}=1,3,\dots} \delta^{\bar{t}-1} V(n=-2) \\ &= E_{S_0}[\text{Prob}(\exists t < \infty \text{ s.t. } S_0 + s(h_t) - (M + (t - 1) - S_0 - s(h_t)) = -2)] \times \frac{V(n=-2)}{1 - \delta^2} \end{aligned}$$

この第 1 項を計算する為に、確率論で良く知られている次の事実を応用する。

補題 (Feller [1957] 第 XIV 章, 邦訳 440 ページ)

数直線上に取られた整数点上で、 t 期にある整数点 w_t にある時、次の $t+1$ 期に確率 p で整数点 w_t+1 にあり、確率 $1-p$ で整数点 w_t-1 にある、というランダム・ウォークを考えよ。ならば、任意の整数 Δw について、初期に整数点 w_1 にある点が、有限期間の後

にその Δw 左に至る確率は、 $p \leq \frac{1}{2}$ なら 1 であり、 $p > \frac{1}{2}$ なら $\left(\frac{1-p}{p}\right)^{\Delta w}$ である。

ここで、 $s_0 = \frac{M}{2} - 1, \frac{M}{2}, \dots, M$ によって事前の M 回ベルヌーイ試行における Success 優越回数を $\Delta w = s_0 - (M - s_0)$ とおくと、この補題によって

$$\begin{aligned} & \text{Prob}(\exists t < \infty \text{ s.t. } s_0 + s(h_t) - (M + (t - 1) - s_0 - s(h_t)) = -2 | p) \\ &= \begin{cases} \left(\frac{1-p_H}{p_H}\right)^{s_0 - (M - s_0) + 2} & \text{if } p = p_H = \frac{2}{3} \\ 1 & \text{if } p = p_L = \frac{1}{3} \end{cases} \end{aligned}$$

であることが分かる。ゆえに、

$$E_{S_0}[\text{Prob}(\exists t < \infty \text{ s.t. } S_0 + s(h_t) - (M + (t - 1) - S_0 - s(h_t)) = -2) | p = p_H]$$

$$\begin{aligned}
&= \sum_{s_0=\frac{M}{2}-1}^M \binom{M}{s_0} (p_H)^{s_0} (1-p_H)^{M-s_0} \\
&\quad \times \text{Prob}(\exists t < \infty \text{ s.t. } s_0 + s(h_t) - (M + (t-1) - s_0 - s(h_t)) = -2 \mid p = p_H) \\
&= \sum_{s_0=\frac{M}{2}-1}^M \binom{M}{s_0} (p_H)^{s_0} (1-p_H)^{M-s_0} \left(\frac{1-p_H}{p_H} \right)^{s_0 - (M-s_0) + 2} \\
&= \left(\frac{1-p_H}{p_H} \right)^2 \sum_{s_0=\frac{M}{2}-1}^M \binom{M}{s_0} (p_H)^{M-s_0} (1-p_H)^{s_0} \\
&= \left(\frac{1-p_H}{p_H} \right)^2 \sum_{s_0=0}^{\frac{M}{2}+1} \binom{M}{s_0} (p_H)^{s_0} (1-p_H)^{M-s_0}
\end{aligned}$$

および $p_H + p_L = 1$ の下で

$$\begin{aligned}
&E_{S_0}[\text{Prob}(\exists t < \infty \text{ s.t. } S_0 + s(h_t) - (M + (t-1) - S_0 - s(h_t)) = -2) \mid p = p_L] \\
&= \sum_{s_0=\frac{M}{2}-1}^M \binom{M}{s_0} (p_L)^{s_0} (1-p_L)^{M-s_0} \\
&\quad \times \text{Prob}(\exists t < \infty \text{ s.t. } s_0 + s(h_t) - (M + (t-1) - s_0 - s(h_t)) = -2 \mid p = p_L) \\
&= \sum_{s_0=\frac{M}{2}-1}^M \binom{M}{s_0} (p_L)^{s_0} (1-p_L)^{M-s_0} \cdot 1 \\
&= \sum_{s_0=0}^{\frac{M}{2}+1} \binom{M}{s_0} (p_L)^{M-s_0} (1-p_L)^{s_0} \\
&= \sum_{s_0=0}^{\frac{M}{2}+1} \binom{M}{s_0} (p_H)^{s_0} (1-p_H)^{M-s_0}
\end{aligned}$$

ゆえに求めたい第1項は

$$\begin{aligned}
&E_{S_0}[\text{Prob}(\exists t < \infty \text{ s.t. } S_0 + s(h_t) - (M + (t-1) - S_0 - s(h_t)) = -2)] \\
&= \frac{1}{2} \left(\frac{1-p_H}{p_H} \right)^2 \sum_{s_0=0}^{\frac{M}{2}+1} \binom{M}{s_0} (p_H)^{s_0} (1-p_H)^{M-s_0} + \frac{1}{2} \sum_{s_0=0}^{\frac{M}{2}+1} \binom{M}{s_0} (p_H)^{s_0} (1-p_H)^{M-s_0} \\
&= \frac{1}{2} \left[\left(\frac{1-p_H}{p_H} \right)^2 + 1 \right] \sum_{s_0=0}^{\frac{M}{2}+1} \binom{M}{s_0} (p_H)^{s_0} (1-p_H)^{M-s_0}
\end{aligned}$$

である。そして、大数の弱法則により、 $p_H = \frac{2}{3} > \frac{1}{2}$ は

$$\lim_{M \rightarrow \infty} \sum_{s_0=0}^{\frac{M}{2}+1} \binom{M}{s_0} (p_H)^{s_0} (1-p_H)^{M-s_0} = \lim_{M \rightarrow \infty} \text{Prob}\left(S_0 \leq \frac{M}{2} + 1\right) = 0$$

を与える。

ゆえに

$$\lim_{M \rightarrow \infty} E_{S_0} [U(\mu(S_0, M), s^*[\mu(S_0, M)])] - E_{S_0} [U(\mu(S_0, M), (\hat{s}^M[S_0], \hat{s}^0))] = 0$$

と結論される。すでに述べたところから、任意の M について

$$\begin{aligned} L(M) &= E_{S_0} [U(\mu(S_0, M), s^*[\mu(S_0, M)])] - E_{S_0} [U(\mu(S_0, M), (\hat{s}^M[S_0], \hat{s}^0))] \\ &< E_{S_0} [U(\mu(S_0, M), s^*[\mu(S_0, M)])] - E_{S_0} [U(\mu(S_0, M), (\hat{s}^M[S_0], \hat{s}^0))] \end{aligned}$$

だから、これは $\lim_{M \rightarrow \infty} L(M) = 0$ を保証する。(証明終)

4.3 命題 3 の [B] の証明

奇数期目に未熟練プレーヤーが選択し、偶数期目に M 熟練プレーヤーが選択する交代手番とせよ。

$L(M) > 0$ は、命題 3 の [A] の 2. と同様に示される。すなわち、 M 熟練プレーヤーが 1 人でプレイする場合には、事前の M 回ベルヌーイ試行の結果 S_0 の実現値が $s_0 \geq \frac{M}{2} - 1$ であるか $s_0 \leq \frac{M}{2} - \frac{3}{2}$ であるかに応じて 1 期目に X-arm を選択する場合と Y-arm を選択する場合とがともに正の確率で起こるが、未熟練プレーヤーではその識別はできないからこの 1 期目の選択を再現できない。

$\lim_{M \rightarrow \infty} L(M) = \frac{1}{6}$ は、命題 3 の [A] から出る。 M が偶数とせよ。この時、1 期目に未熟練プレーヤーが X-arm を選択するとせよ。ならば、2 期目の先頭で、チームは、あたかも事前に $M+1$ 回のベルヌーイ試行を行った M 熟練プレーヤー先手番の投資選択問題と、1 つの点を除いて同じ状況におかれる。1 つの違いとは、未熟練プレーヤーは、 $M+1$ 回のベルヌーイ試行の最後の $M+1$ 回目の試行結果を見ているということである。したがって、2 期目以降、チームのパフォーマンスは、 $M+1$ 回のベルヌーイ試行の下での M 熟練プレーヤー先手番の投資選択問題のパフォーマンスより下がることはない。 M が偶数なら $M+1$ は奇数である。この時、命題 3 の [A] の 1. は $L(M+1) = 0$ を保証する。よって、1 期目に未熟練プレーヤーが X-arm を選択するなら、その 2 期目からの継続利得は、2 期目以降 M 熟練プレーヤーが 1 人でプレイする場合のそれと同じになる。全く同様に、1 期目に未熟練プレーヤーが Y-arm を選択しても、2 期目以降の継続利得は、2 期目以降 M 熟練プレーヤーが 1 人でプレイする場合のそれと同じになる。ただし、1 期目に X-arm を選択する方が、この議論で言う 2 期目以降のプレイにとって 1 回分事前のベルヌーイ試行が増えることを意味するから、2 期目からの継続利得は高くなる。未熟練プレーヤーにとって、1 期目の期待利得は、X-arm と Y-arm で同じである。よって、未熟練プレーヤーは 1 期目に X-arm を選択する。こうして、 M が偶数の場合、ベイズ均衡で、1 期目に未熟練プレーヤーが X-arm を選択し、2 期目以降は M 熟練プレーヤーが 1 人でプレイするのと同じプレイが起こる。よって、チームの

パフォーマンス・ロスは、1期目に必ず X-arm を選択することによるロスに等しい。ここで、1期目も M 熟練プレーヤーが選択した場合、 M が十分大きな偶数なら、X-arm の成功確率が $p_H = \frac{2}{3}$ のときはほぼ確率 1 で X-arm を選択し、 $p_L = \frac{1}{3}$ のときはほぼ確率 1 で Y-arm を選択する。¹²⁾ したがって、チームが 1 期目に必ず X-arm を選択することによるロスは、 M が十分大きな偶数なら、X-arm の成功確率が $p_L = \frac{1}{3}$ の時に、 M 熟練プレーヤーなら 1 期目に Y-arm を選択するところを X-arm を選択することによるロスにはほぼ等しい。それは、 $\frac{1}{2}(0 - (p_L \cdot 1 + (1 - p_L) \cdot (-1))) = \frac{1}{6}$ である。

M が奇数の場合、2 期目の先頭でチームがおかれる状況は、偶数 $M+1$ 回のベルヌーイ試行を行った M 熟練プレーヤー先手番の投資選択問題に対応する。しかし、命題 3 の [A] の 2. により $\lim_{M \rightarrow \infty} L(M) = 0$ だから、 M が十分大きな奇数なら、2 回目以降のプレイは 2 回目から M 熟練プレーヤーが 1 人でプレイする場合のパフォーマンスをほぼ達成し、従って 1 回目の未熟練プレーヤーの選択は X-arm で、そのことに由来するロスは $\frac{1}{6}$ である。(証明終)

5 結 論

スロット・マシンについて事前の試行結果を知っているプレーヤーと知らないプレーヤーが交代でプレイする時、前者が持つ情報はそのプレイにどの程度生かされるかを、1 つの例について分析した。その主結果は、次のことを意味している。第 1 に、情報を持つプレーヤーが先手番で、事前の試行が一定の条件（考察した例では、試行が奇数回）を満たす場合には、情報は完全に生かされるが、そうでなければロスが生じる。第 2 に、しかし、一方のプレーヤーが持つ情報の量が多ければ、ロスはあったとしても十分小さい。すなわち、十分な熟練を持つ者と未熟練者とのチームは、2 人が交代でプレイしたとしても、あたかも熟練者が 1 人でそれをするのとはほぼ同じプレイを実現する。

注

- 1) 本稿をゼミで発表した際、天谷研一講師（神戸大学経済経営研究所）から主結果の証明の一部にあった誤りを指摘して頂いた。記して感謝します。
- 2) 末廣 [2002] は、同じ問題を、チームの各メンバーがそれぞれ同程度に情報を持つ例について研究した。
- 3) $t=1, 2, \dots$ の全期間にわたって、チームは最初に固定された同じ投資機会に直面する。つまり、チームが直面している X-arm は、 $t=1, 2, \dots$ の全期間にわたって成功確率 $\frac{2}{3}$ の arm であるか、 $t=1, 2, \dots$ の全期間にわたって成功確率 $\frac{1}{3}$ の arm であるか、のいずれかである。
- 4) E の下付の Z は、その期待値が定義されている測度の対象が Z であることを表す。 E の条件付きの a は、 Z の測度がそれによって条件付けられて定まっていることを示す。
- 5) M 熟練プレーヤーは、この M 回ベルヌーイ試行を直接見ており、従って s_0 回の Success が M 回のうちの何回目の試行で生じたかを見ているから、このベイズ・ルールの計算式が得られる。

- ただし, s_0 値のみが分かっており, s_0 回の Success の配置が分からない場合の μ も同じ値になる。
- 6) 表 1 に掲げられている数字は, すべて小数点以下 7 桁目を四捨五入してある。
- 7) $\binom{M}{s_0}$ は M 個の対象から s_0 個の対象の組を取り出す取り出し方の数 $\frac{M!}{s_0!(M-s_0)!}$ を表す。
- 8) 特に $s_0 = \frac{M}{2} - 1$ の時, これは, 1 期目に Y-arm を選択し, 2 期目に未熟練プレーヤーが Y-arm を選択するなら 3 期目には X-arm を選択する, ということを意味する。
- 9) M 熟練プレーヤーが 1 期目と 3 期目に Y-arm を選択し, 5 期目以降で X-arm を選択することは, 均衡外プレイである。均衡外プレイに対する未熟練プレーヤーの行動は, 以下の証明に影響しないので省略する。
- 10) $\rho = \frac{1}{2}$, $p_H + p_L = 1$ の下では, $\text{Prob}\left(S_0 = 0, \dots, \frac{M}{2} - 2, \frac{M}{2} - 1\right) + \frac{1}{2} \text{Prob}\left(S_0 = \frac{M}{2}\right) = \frac{1}{2}$ である。
- 11) value function の値 $V(\mu)$ が $V(\mu) > 0$ であることは, 右辺の最大化問題の解が X-arm の選択であることを意味し, $V(\mu) = 0$ であることは, 右辺の最大化問題の解が Y-arm の選択であることを意味する。表 1 に示した $V(\mu)$ の計算結果は, $V(\mu) > 0$ と $V(\mu) = 0$ の切り替えが $n = -2$ と $n = -3$ の間でおこること, つまり X-arm と Y-arm の切り替えが $n = -2$ と $n = -3$ の間でおこることを示している。それは, 解析的に導かれた結果である命題 1 に符合する。
- 12) 正確に言うと, X-arm の成功確率が $p_H = \frac{2}{3}$ のときに Y-arm を選択する過誤の確率と, X-arm の成功確率が $p_L = \frac{1}{3}$ のときに X-arm を選択する過誤の確率とは, ともに正で残る。そして, 命題 1 により, 後者の過誤の確率は, 前者のそれよりも大きい。

参 考 文 献

- Bellman, R. [1956], "A problem in the sequential design of experiments," *Sankhya*, A 16, pp. 221-229.
- Bellman, R. [1957], *Dynamic Programming*, Princeton University Press.
- Berry, D. A. and B. Fristedt [1979], "Bernoulli one-armed bandits—arbitrary discount sequences," *Annals of Statistics*, Vol. 7, pp. 1086-1105.
- Berry, D. A. and B. Fristedt [1985], *Bandit Problems: Sequential Allocation of Experiments*, Chapman and Hall.
- Feller, W. [1957], *An Introduction To Probability Theory and Its Applications Volume I*, John Wiley and Sons. (河田龍夫監訳, ト部舜一, 矢部真, 池守昌幸, 大平坦, 阿部俊一訳 [1961], 『確率論とその応用 I』, 紀伊国屋書店。)
- Ross, S. M. [1983], *Introduction to Stochastic Dynamic Programming*, Academic Press.
- 末廣英生 [2002], 「あるスロットマシン問題のグループによる解決」, 『国民経済雑誌』, 第186巻, 第5号, 29-50ページ。