1D3-03

# ルーレット選択を用いた Profit Sharing 強化学習における合理性についての一考察

The consideration of rationality of Profit Sharing with roulette action selection

河合 宏和 Hirokazu Kawai　上野 敦志 Atsushi Ueno　辰巳 昭治 Shoji Tatsumi

大阪市立大学大学院 工学研究科 電子情報系専攻
Department of Physical Electronics and Informatics, Osaka City University

In this paper, we discuss the rationality of profit sharing (PS) in reinforcement learning (RL) methods with roulette action selection. In PS methods, received rewards are distributed among all selected rules on the way to the rewards. The volume of distribution is fixed as a function of the distance to the rewards. For the rationality of PS methods, a theorem, the Rationality Theorem of Profit Sharing, has been proposed, which enable RL agents to learn routes without loops. Following the theorem, however, makes the function converge to zero very quickly and makes learning inefficient. We propose a theorem for distributing more volume to distant rules through the use of a special feature of roulette action selection. Experimental results have shown that RL agents can learn more efficiently with it.

## 1. 序論

**強化学習法**は，**報酬**という特別な入力を手がかりとして試行錯誤によって行われる教師なし学習法の一種である．エージェントはセンサー等で外界からの入力を知覚し，**状態**として扱い，現在の状態において可能な行動群の中から一つを選んで実行する．この状態行動対を**ルール**，ルールの選択法を**政策**と呼ぶ．

経験強化型学習法の一種である **Profit Sharing 強化学習** [Grefenstette 88]（以下，PS）は，得られた報酬を行動系列に分配し，累積することで適切な行動系列を獲得する学習法である．この分配方法は，**強化関数**という関数によって規定される．一般に，行動選択にはルーレット選択を用いる．PS において報酬獲得が阻害され得る迂回系列の抑制条件を規定する定理として，**合理性定理**[宮崎 94]が提案されている．しかし，この条件を満たす強化関数は分配報酬値の 0 への収束が早く，特に行動選択数や状態数の多い問題環境等では学習効率が悪い．

本研究では，ルーレット選択の特性を利用した迂回系列抑制条件として，**統計的合理性定理**を提案する．統計的合理性定理により迂回系列が十分抑制されていることを示し，従来の合理性定理を満たす学習と比較して学習効率が上昇し得る事を実験により確認する．

## 2. 合理性定理

PS ではエージェントの報酬獲得時，その行動系列中の報酬 $R$ から遡って $x$ 番目のルールの評価値 $w_x$ は，次式のように更新される．

$$w_x \leftarrow w_x + Rf_x \tag{1}$$

ここで，$f_x$ は強化関数である．一般的に，$f_x$ は $x$ のみに依存する非増加関数である．EPS[植村 05]のように，$x$ 以外の値にも依存する強化関数もあるが，本論文では扱わないこととする．PS における学習は，この強化関数の設定に集約される．

迂回系列の抑制条件は，合理性定理によって規定されている．合理性定理を満たす強化関数は，その制約上，分配報酬量の 0 への収束が早く，学習の有効となる距離，すなわち**学習距離**が短い．行動選択数を $L$ とすると，合理性定理を満たす強化関数で，最も学習距離の長い関数は，公比 $1/L$ の等比減少関数である[植村 04].

[宮崎 94]において，**無効ルールの抑制**とは，**無効ルール**が，それと競合する**有効ルール**を差し置いて一番に強化されないこと，と定義されている．ここで，無効ルールとは常に迂回系列上に存在するルール，有効ルールとは無効ルールで無いものを指す．しかし，ルーレット選択を用いる場合，行動決定は評価値 $w$ の比率によって決まるため，ルールが一番に強化されるかどうかは，ε-greedy 行動決定法などの他の行動決定法に比べてさほど重要ではない．

本研究では，上記のルーレット選択の性質を利用した迂回系列抑制条件を提案し，その条件を等比減少強化関数に当てはめたとき，公比 $1/L$ の等比減少関数よりも学習距離が長いものが含まれることを示す．

## 3. 統計的合理性定理

### 3.1 はじめに

本章では，本研究で新しく提案する，ルーレット選択特有の合理性定理について説明する．まずルーレット選択特有の性質を提示する．その性質を考慮した上で，強化関数が満たすべき新しい迂回系列抑制条件を局所的な場合，大局的な場合ともに示し，強化関数に等比減少関数を用いた際，それがどのような条件となるかを示す．

### 3.2 ルーレット選択の性質

ルーレット選択では，状態 $s$ で行動 $a_i$ を選択するルール $\overrightarrow{sa_i}$ の行動選択確率 $P(s,a_i)$ は，そのルールの評価値 $w_{\overrightarrow{sa_i}}$ の，状態 $s$ での全評価値に対する割合となる．また $w_{\overrightarrow{sa_i}}$ は，そのルールの期待強化量 $E(f_{\overrightarrow{sa_i}})$ の割合に収束しようとする．すなわち，式(2)のような関係が成り立つ．

$$P(s,a_i) = w_{\overrightarrow{sa_i}} \Big/ \sum_k w_{\overrightarrow{sa_k}} \rightarrow E(f_{\overrightarrow{sa_i}}) \Big/ \sum_k E(f_{\overrightarrow{sa_k}}) \tag{2}$$

ここから，次の式(3)を常に満たす時，$P(s,a_i)$ は常に現在の値より低い収束値を持つ．

$$P(s,a_i) > E(f_{\overrightarrow{sa_i}}) \Big/ \sum_k E(f_{\overrightarrow{sa_k}}) \tag{3}$$

連絡先：河合 宏和，大阪市立大学大学院 工学研究科 電子情報系専攻 知識情報処理工学研究室，〒558-8585 大阪市住吉区杉本，TEL,FAX:06-6605-2778，hiro@kdel.info.eng.osaka-cu.ac.jp

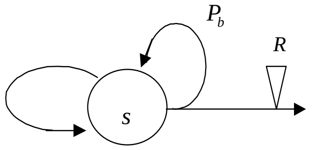


図 1：無効ルールの統計的抑制が最も困難な状態

そのため $P(s,a_i)$ は一時的な増減はあるものの，時間の経過とともに減少し，最終的に 0 へと収束する事が保証される．このような状況を，ルール $\overrightarrow{sa_i}$ の行動選択確率 $P(s,a_i)$ は**常に減少傾向にある**，と定義する．

### 3.3 局所統計的合理性定理

局所的な合理性を満たすための条件として，無効ルールの抑制がある．ここで，無効ルールの抑制とは，

無効ルールが，それと競合する有効ルールを差し置いて一番に強化されないこと

という定義である[宮崎 94]．しかし，ルーレット選択を用いる際は，各ルールの行動選択確率の比はそのルールの評価値の比であるという特性上，無効ルールが一番に強化されているかどうかは，無効ルールによる迂回系列の強化を回避するにあたってさほど重要ではない．

そこで，本研究ではある状態 $s$ で全無効ルールの選択確率が最終的に 0 へと収束する事を，$s$ における**無効ルールの統計的抑制**と定義する．全ての無効ルールの選択確率が 0 に収束する事は，任意の状態においてその状態に存在する無効ルールの選択確率の和が常に減少傾向にある事と同義である．全状態で無効ルールの統計的抑制が行われる条件を導くために，無効ルールの統計的抑制が最も困難な状態を考える．従来の合理性定理では，無効ルールの抑制が最も困難な状態として，

唯一の回帰的無効ルールと $(L-1)$ 本の有効ルールが混在する状態

が挙げられていた．これは，有効ルールの選択･強化が分散するのに対し，無効ルールの選択･強化が分散しない事に起因する．しかし，無効ルールの統計的抑制では，無効ルール選択確率の和が常に減少傾向にあればよいので，無効ルール･有効ルールの本数ではなく，期待強化量･行動選択確率の和に着目すればよい．ただし，強化関数によっては，無効ルールの本数が増えることでその期待強化量の合計値が増加する可能性もある．

なお，非決定的な状態遷移を考慮に入れる場合，有効ルールが一定確率で同一状態に回帰する場合と，異なる状態に遷移する場合が考えられるが，後者の場合，無効ルールと有効ルールの期待強化量に同じ割合で影響するだけで，無効ルールの統計的抑制に関しては決定的な状態遷移環境と変わらない．従って，前者の非決定性のみを考慮すればよい．

上記を踏まえると，非決定的状態遷移も含めた上での，無効ルールの統計的抑制が最も困難な環境は，図 1 に示すような

$(L-1)$ 本の回帰的無効ルールと確率 $P_b$ で回帰する 1 本の有効ルールが競合する状態

であると考えられる．これより，局所統計的な合理性を満たす条件は，次のようになる．

**[定理 1] 局所統計的合理性定理**

強化関数が，任意の無効ルールの統計的抑制を行う必要十分条件は，$(L-1)$ 本の回帰的無効ルールと，確率 $P_b$ で回帰する 1 本の有効ルールが競合する状態 $s$ で，次式が常に成り立つことである．

$$\sum E(f_{\overrightarrow{sa_{ie}}}) \Big/ \sum_k E(f_{\overrightarrow{sa_k}}) < \sum P(s,a_{ie}) \qquad (4)$$

ここで，$\sum E(f_{\overrightarrow{sa_{ie}}})$ は無効ルールの強化量の和，$\sum P(s,a_{ie})$ は無効ルールの選択確率の和を示す．

### 3.4 大局統計的合理性定理

前節では，無効ルールによって構成される迂回系列の強化を回避する条件として，定理 1 に局所統計的合理性定理を提案した．しかし，迂回系列には，無効ルールによって構成される迂回系列以外にも，複数の有効ルールによって構成される**大局的な迂回系列**も存在する．本節では，大局的な迂回系列を統計的に抑制する条件を検討する．

まず，簡単のため，状態遷移が決定的な場合を考える．大局的な迂回系列において，最も基本的な形を考えると，図 2 のようになる．この迂回系列において，ルール $\overrightarrow{s_m a}$（ただし，$m=1,2,...,M,\ \ M\geq 2$）は迂回系列を構成するルールで**再帰ルール**，それ以外のルールはこの迂回系列から脱出するルールで**非再帰ルール**とする．このような，$M$ 個の状態，$M$ 本の再帰ルール，2 本以上の非再帰ルールを持つ迂回系列を，**単一迂回系列**とする．大局的な迂回系列は，単一迂回系列の組み合わせによって構成される．

ここで，大局統計的な合理性について，次のことが言える．

**[補題 1] 大局統計的な合理性**

任意の単一迂回系列において，非再帰ルールと競合する少なくとも一本の再帰ルールが統計的に抑制されれば，大局統計的な合理性は満たされる．

証明は**付録 A** に示す． これは，状態遷移が非決定的な場合も同様である．

次に，状態遷移が決定的な場合で，単一迂回系列中，統計的抑制が最も困難な環境について検討する．

単一迂回系列を統計的に抑制するにあたっては，再帰ルールの選択確率が 0 へと収束する状態が少なくとも一つあれば良いので，非再帰ルールの本数は重要ではなく，その選択確率･期待強化量の合計値にのみ着目すればよい．強化関数によっては，非再帰ルールの本数の増加に従って期待強化量の合計値が増える可能性はあるが，逆に減る事は考えられないので，各状態での非再帰ルールの数が最大で一本の場合が単一迂回系列の統計的抑制にあたっては最も困難な場合となる．
また，各非再帰ルール選択後の報酬値に偏りがあると，多い報酬値を得られる非再帰ルールによる迂回系列の脱出はより容易

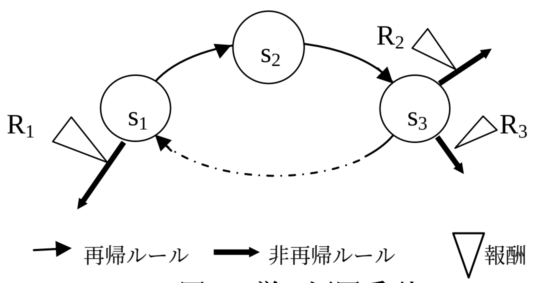


図 2：単一迂回系列

になり，迂回系列の抑制はより簡単になる．すなわち，非再帰ルール選択後の報酬値の偏りや，非再帰ルールを持たない状態が無い単一迂回系列の抑制が最も困難であるといえる．

以上より，決定的な状態遷移の環境中，迂回系列の統計的抑制が最も困難な環境は，$M$ 状態がそれぞれ 1 本の再帰ルールと，同一報酬値を得る 1 本の非再帰ルールを持つ環境であると考えられる．このような環境を**単純迂回系列**と定義する．

非決定的な状態遷移環境も考慮に入れた上での，統計的抑制が最も困難である大局的な迂回系列は，上記に示した単純迂回系列に非決定性を加えたものである．

状態 $s_i$ における再帰ルール，非再帰ルールの状態遷移が，それぞれ一定確率で状態 $s_i$ に非決定的に回帰すると考える．状態 $s_i$ から $s_i$ 以外の状態への非決定性による遷移も考えられるが，同様に $s_i$ 以外の状態から状態 $s_i$ への遷移も考えられるため，対称性を考慮すると，非決定性による遷移先を状態 $s_i$ のみに固定しても一般性は失われない．また，各状態における再帰ルール，非再帰ルールの回帰確率は，状態間で偏りがない場合が迂回系列の抑制は最も困難である．すなわち，非決定的な状態遷移を考慮した大局的な迂回系列の中で統計的抑制が最も困難な環境は，図 3 のように，単純迂回系列中の各状態における再帰ルール，非再帰ルールが，それぞれ同一の回帰確率 $P_A, P_B$ で回帰するような環境となる．このような環境を**非決定的単純迂回系列**と定義する．

これより，非決定的な状態遷移を考慮した上で大局的合理性を統計的に満たすための条件として，次の定理 2 が導き出せる．

**[定理 2]　大局統計的合理性定理**

定理 1 を満たす強化関数が，任意の大局的迂回系列を統計的に抑制するための必要十分条件は，$0 < P(s_m, a_r) < 1$, $0 \le P_A < 1$, $0 \le P_B < 1$, $M \ge 2$（$m = 1,2,...,M$）における任意の非決定的単純迂回系列中の，ある状態 $s_i$ において，次式が常に成り立つことである．

$$E(f_{\overline{s_i a_r}}) \Big/ \sum_k E(f_{\overline{s_i a_k}}) < P(s_i, a_r) \tag{5}$$

ここで，$\overrightarrow{s_i a_r}$ は状態 $s_i$ における再帰ルールを示す．

### 3.5 等比減少強化関数での統計的合理性定理

定理 1 を用いて，等比減少の強化関数における局所統計的合理性を満たす条件を算出すると，次の系 1 が得られる．この証明は省略する．

**[系 1]　割引率 $\gamma$ の条件**

定理 1 を満たす等比減少の強化関数において，割引率 $\gamma$ は次の式(6)を満たす．

$$0 < \gamma < 1 \tag{6}$$

ここで，割引率 $\gamma$ は強化関数の公比を示す．

この条件を満たすものには，明らかに $\gamma = 1/L$ の強化関数より学習距離の長いものが含まれる．

大局的な場合において，状態遷移が決定的である場合については，式(6)を満たせば十分統計的合理性が満たされる事を導き出した．しかし，非決定的である場合については，数式の複雑さ等の理由で現在検討中であり，これは今後の課題となる．

### 3.6 おわりに

本論文で提案する，統計的合理性定理について説明した．ルーレット選択を用いた際，ルールの選択確率が 0 に収束する条件を用いて，迂回系列を統計的に抑制する条件を局所的な場合，大局的な場合ともに示した．その上で一般に使われている等比減少の強化関数において，その条件がどのようなものになるかを検討した．

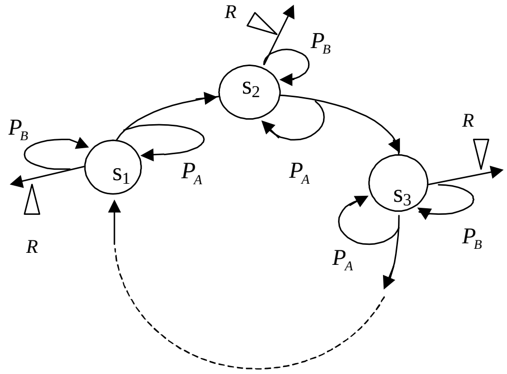


図 3：非決定的単純迂回系列

## 4. 実験

### 4.1 等比減少強化関数の合理性についての実験

等比減少の強化関数において，図 3 の環境で乱数を変えた 100 回のシミュレーション実験を行い，大局統計的合理性を満たす割引率 $\gamma$ の条件を考察する．実験ステップ数は 2500 ステップ，状態は $s_1$ ～ $s_5$ の 5 状態で，非再帰ルール，再帰ルールの数はそれぞれ 1 本ずつとし，各時間ステップで，全状態中最も高い非再帰ルールの選択確率を評価した．学習は状態 $s_1$ ～ $s_5$ のいずれかよりランダムにスタートし，非再帰ルールを選択すると報酬 10 を得て，状態 $s_1$ ～ $s_5$ へとランダムに遷移する．この環境においては，行動選択数 $L = 2$ であるので，従来の合理性定理を満たす割引率は $0 < \gamma \le 0.5$ となる．再帰ルール，非再帰ルールの回帰確率は共に 0.9 とした．

この実験の結果を図 4 に示す．この実験結果から，局所統計的合理性を満たす $0 < \gamma < 1$ の範囲において，時間の経過とともに迂回系列は抑制されてゆくことが分かる．なお，実験値は 2500 ステップ以降も増加してゆき，一定の値で学習が止まらない事を確認している．ただし，実験結果は，割引率が 1 に近づくにつれ，迂回系列の抑制に時間が掛かるようになる事も示している．

### 4.2 ランダムウォーク問題

統計的合理性定理に従う強化関数による学習の効果を，ランダムウォーク問題(図 6 参照)のシミュレーション実験で確認する．

問題環境は以下の通りである．状態 $s_n$（$0 \le n \le 250$）は前後一列に並んでおり，状態間の距離は 1 である．行動選択数 $L$ は 40，行動値は $-19$～$+20$ の整数値を取り，その行動値の距離だけエージェントは前後に状態遷移する．実験ステップ数は

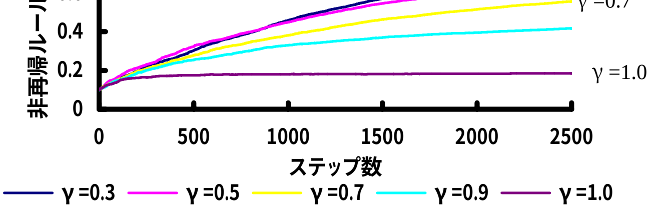


図 4：非決定的単純迂回系列における実験結果

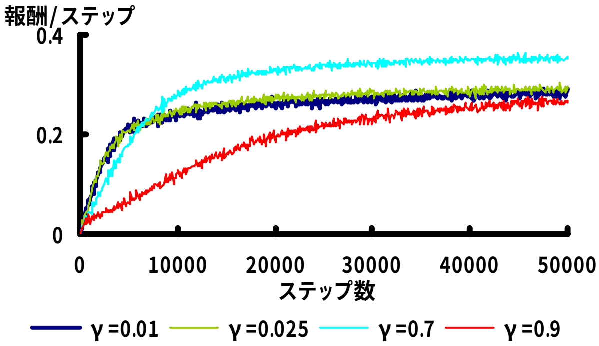

図 5：ランダムウォーク問題の実験結果

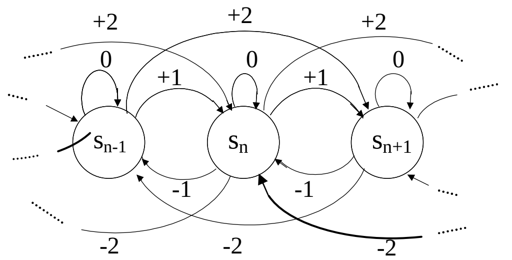

図 6：ランダムウォーク問題

5 万ステップを取る．終点 $s_G$ $(G=250)$に到達すると 10 の報酬値が得られ，再び始点 $s_0$ から出発する．この始点 $s_0$ から終点 $s_G$ に至るまでの行動系列が 1 エピソードとなる．遷移先の状態が 0 より小さい場合は状態 $s_0$ へと，$G$ より大きい場合は状態 $s_G$ へと遷移する．また，状態遷移に非決定性を加え，それぞれのルールは 0.2 の確率で回帰するようにした．

乱数を変えた 100 回の実験を行い， 100 ステップごとの平均報酬獲得値を性能の評価として用いる．最適解は最短経路のステップ数が 13 であることより，$10/13 \cong 0.769$ であるが，各ルールは 0.2 の確率で回帰するため，その分を考慮する必要があり，事実上の最適解は，$0.769 \times 0.8 \cong 0.615$ と考えられる．

この実験の結果を図 5 に示す．図 5 のグラフにおいて，従来の合理性定理を満たす割引率は $\gamma=0.01, \gamma=0.025$ である．$\gamma=0.7$ の割引率は，この環境において迂回系列抑制と学習距離のバランスが比較的取れているため，$\gamma=0.01, \gamma=0.025$ よりも学習効果の上昇が見られる．

## 5. 結論

本論文では，行動選択にルーレット選択を用いた PS の迂回系列抑制条件について，局所的，大局的な場合ともに考察した．

ルーレット選択の特性を用いて，等比減少の強化関数で，従来の合理性定理を満たさなくても迂回系列が抑制される事を示し，より学習距離の長い割引率が適用可能な事を示した．

しかし，等比減少の強化関数で割引率の値を 1 に近づけると，学習距離が伸びる反面，迂回系列の抑制効率は下がる．両者のバランスが取れた割引率の設定法を検討する必要がある．

また，等比減少の強化関数において，状態遷移が非決定的である場合の大局統計的な合理性を満たすための条件は，その計算の複雑さや時間の都合上，理論的に求めることが出来なかったので，今後その条件を求める必要がある．

その他，従来の合理性定理と本定理の間の因果関係について今後調べる予定である．以上 3 点が，今後の課題となる．

## 6. 謝辞

この研究を進めるにあたり，ご協力を頂いた，大阪市立大学の植村 渉氏に感謝を申し上げます．

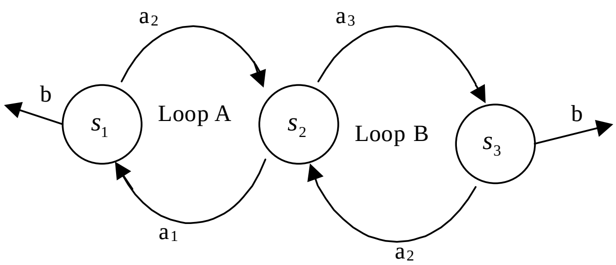

図 A：単一迂回系列の組み合わせ

## 付録A. 補題 1 の証明

単一迂回系列内のある一つの再帰ルールが統計的に抑制されれば，大局統計的な合理性が満たされる事を証明する [1]．図 A に二つの単一迂回系列が組み合わさった環境を示す．三つ以上の単一迂回系列が組み合わさった場合でも，対象とする迂回系列を二つずつに分ければ同様に考えることが出来る．

図 A に示した環境において，行動 $a_n$ は状態 $s_n$ へと遷移する行動，状態 $s_2$ は二つの迂回系列 Loop A と Loop B に共通する状態となる．ここで，状態 $s_2$ において両方の迂回系列での非再帰ルールが存在する場合は，単にそのルールが強化されれば良いだけになるので，ここでは一方の迂回系列における再帰ルールが，もう一方の迂回系列における非再帰ルールとなるような環境のみを考える．

ここで，一つの迂回系列につき一つの再帰ルールが統計的に抑制される場合を考える．Loop A，Loop B の抑制をそれぞれ状態 $s_1$，$s_3$ で行う場合，それぞれ，その状態で迂回系列を抜け出す経路が学習される．Loop A，Loop B のどちらか片方の迂回系列の抑制を状態 $s_2$ で行う場合，Loop A を状態 $s_2$ で抑制した場合は状態 $s_3$ で，Loop B を状態 $s_2$ で抑制した場合は状態 $s_1$ で迂回系列を抜け出す経路が学習される．Loop A，Loop B の抑制をどちらも状態 $s_2$ で行った場合は，状態 $s_1$，$s_3$ ともに迂回系列を抜け出す経路を学習しなくなるが，この場合ルール $\overrightarrow{s_2 a_1}$，$\overrightarrow{s_2 a_3}$ がともに統計的に抑制される，すなわち行動選択確率が 0 へと収束する，ということになる．そのようなことは起こりえないので，Loop A，Loop B の抑制をどちらも状態 $s_2$ で行うということはありえない．

以上より，単一迂回系列において再帰ルールを統計的に抑制できれば大局統計的な合理性は満たされると言える．

## 参考文献

[Grefenstette 88] Grefenstette J.J. : Credit Assignment in Rule Discovery Systems Based on Genetic Algorithms, *Machine Learning*, Vol.3, pp.225-245 (1988).

[宮崎 94] 宮崎 和光，山村 雅幸，小林 重信：強化学習における報酬割当ての理論的考察，人工知能誌，Vol.9，No.4，pp.580-587 (1994).

[植村 04] 植村 渉，辰巳 昭治：Profit Sharing 法における強化関数に関する一考察，人工知能論文誌，Vol.19，No.4，pp.193-203 (2004).

[植村 05] 植村 渉，上野 敦志，辰巳 昭治：POMDPs 環境のためのエピソード強化型強化学習法，電子情報通信学会論文誌，Vol.J88-A，No.6 (2005)掲載予定.

[1] この証明は，[植村 05]での証明を一般的な強化関数及び統計的合理性の場合に拡張したものである．