3A2-1

Profit Sharingの期待値に基づく合理性

Rationality of Profit Sharing Based on Expected Value

松井 藤五郎　大和田 勇人
Tohgoroh Matsui　Hayato Ohwada

東京理科大学 理工学部 経営工学科
Department of Industrial Administration, Faculty of Science and Technology, Tokyo University of Science

This paper describes the rationality of profit sharing based on expected value. Miyazaki et al. proved the rationality, we call this the complete rationality, of profit sharing by analysing the worst case in all possible situations. However, the strong condition based on the complete rationality reduces the effectiveness of profit sharing. In this paper, we analyse the average case to show the expected rationality of profit sharing.

1. はじめに

Profit sharing [Grefenstette 88] は，それぞれの行動に優先度を割り当てるタイプの強化学習法である．Profit sharing は，クラシファイア・システム [Holland 86] における信用割当の技法として提案され，その後の多くの強化学習アルゴリズムに影響を与えた．Grefenstette の手法は遺伝的アルゴリズムを併用していたが，宮崎らによって profit sharing が一般的な強化学習の枠組みで扱えることが示された [宮崎 94]．

Profit sharing は，環境から報酬を獲得すると，報酬獲得に至った状態行動系列中の状態行動対に信用割当関数に基づいて優先度を増加させる．Profit sharing は，Q 学習のような行動価値を推定するタイプの手法に比べて (1) 学習が速く，(2) 学習中の振る舞いが優れているという大きな利点を持っている．

宮崎らは，[宮崎 94] において，profit sharing を報酬が獲得できない政策に収束させないための条件を示した．本論文では，これを宮崎の完全合理性条件と呼ぶ．その後，これに基づいて profit sharing をマルチエージェント環境，非定常環境，部分観測環境へ profit sharing を適用する研究などが数多く行われている．

Profit sharing の信用割当関数には等比減少関数がよく用いられる．ところが，等比減少信用割当関数を用いたときに宮崎の完全合理性条件に従うと，学習が遅くなってしまい，profit sharing を用いる際に大きな問題となっている．

そこで，本論文では，MDP における等比減少信用割当関数を用いた profit sharing に対する期待値の観点に基づく profit sharing の合理性と割引率の条件について述べる．

2. Profit Sharing とその完全合理性

2.1 対象とする問題

本論文では，標準的なマルコフ決定過程（MDPs）におけるエピソード型タスクを対象とする．

MDPs に関する記号は，[Sutton 98] に倣う．すなわち，エージェントと環境の相互作用を

$$s_0, a_0, r_1, s_1, a_1, r_2, \ldots, s_{T-1}, a_{T-1}, r_T, s_T$$

というように状態 $s \in \mathcal{S}$（$\mathcal{S}$ は状態の集合），行動 $a \in \mathcal{A}$（$\mathcal{A}$ は行動の集合），報酬 $r \in \mathfrak{R}$（$\mathfrak{R}$ は実数の集合）の列で表し，

連絡先: 松井藤五郎, 東京理科大学 理工学部 経営工学科,
matui@ia.noda.tus.ac.jp, http://とうごろう.jp

すべての $s \in \mathcal{S}, a \in \mathcal{A}$ に対して：
　$P(s,a) \leftarrow C$　（C は任意の小さな正の定数）
各エピソードに対して繰り返し：
　状態 s を初期化する
　エピソード中の各ステップに対して繰り返し：
　　P から導かれる行動選択確率の分布に従って，
　　状態 s での行動 a を選択する
　　行動 a を取り，報酬 r と次状態 s' を観測する
　　$s \leftarrow s'$
　s が終端状態ならば繰り返しを終了する
　エピソードに含まれるすべての状態行動対に対して：
　　$P(s_t, a_t) \leftarrow P(s_t, a_t) + f(t, r_T, T)$

図 1: Profit sharing アルゴリズム．

エージェントと環境の相互作用が終端状態によって部分系列に分解できるタスクを扱う．この初期状態から終端状態までの部分系列をエピソードと呼ぶ．

2.2 Profit Sharing アルゴリズム

Profit sharing [Holland 86, Grefenstette 88, 宮崎 94] は，代表的な行動優先度学習型の強化学習アルゴリズムであり，それぞれの状態ごとに行動の優先度を学習する．状態 s における行動 a の優先度を $P(s,a)$ と表し，エージェントは優先度 $P(s,a)$ に基づいて行動を選択する．

Profit sharing のアルゴリズムを図 1 に示す．Profit sharing は，エピソード $s_0, a_0, r_1, \ldots, s_{T-1}, a_{T-1}, r_T$ に含まれる各状態行動対 s_t, a_t に対する優先度をエピソード終了後に一括して次のように強化する．

$$P(s_t, a_t) \leftarrow P(s_t, a_t) + f(t, r_T, T)$$

ここで，f は信用割当関数と呼ばれる関数である．

ある一つの状態行動対 s, a について，エピソードに出現した信用割当関数の値を合計したものを s, a に対する強化値といい，$\Delta P(s,a)$ と表す．

2.3 Profit Sharing にとって学習が最も困難な状態

ある状態行動対が現在までのすべてのエピソードにおいて常に迂回経路上に出現するとき，その状態行動対は無効であるという．状態行動対が無効でないとき，有効であるという．

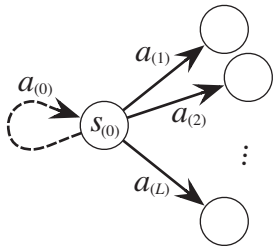

図 2: Profit sharing にとって学習が最も困難な状態 [宮崎 94].

宮崎らは，[宮崎 94] の中で，profit sharing にとって学習が最も困難な状態を示した．

補題 1 (学習が最も困難な状態) 図 2 に示されたような，ひとつ以上の有効な行動と唯一の回帰的で無効な行動が競合している状態が，*profit sharing* にとって学習が最も困難な状態である．

ここでは，有効な行動を実行したときの遷移先の状態はすべて終端状態かつ目標状態であるとし，正の報酬を与える．すなわち，$\mathcal{R}^{a_{(i)}}_{s_{(0)}} > 0$（$i = 1, 2, \ldots, L$）である．ここで，$\mathcal{R}^{a}_{s}$ は状態 s で行動 a を行ったときに得られる報酬の期待値を表し，L は有効な行動の数を表す．Profit sharing ではエピソード途中の報酬を 0 とみなしているので，唯一存在する回帰的で無効な行動 $a_{(0)}$ に対する報酬は 0——すなわち，$\mathcal{R}^{a_{(0)}}_{s_{(0)}} = 0$ である．

2.4 宮崎の合理性定理と合理性条件

宮崎らは，[宮崎 94] において，profit sharing にとって学習が最も困難な状態における強化値を分析することによって，無効な行動を抑制するための条件を示した．

定理 1 (Profit sharing の完全合理性) *Profit sharing* の信用割当関数 f が，

$$f(t, r_T, T) > L \sum_{i=0}^{t-1} f(i, r_T, T)$$

を満たすならば，任意の状態において，任意の無効な行動が抑制される．ここで，L は，同一の入力状態に存在する有効な行動の最大個数である．

これを，本論文では，宮崎の完全合理性定理と呼ぶ（証明は [宮崎 94] を参照）．

Profit sharing の信用割当関数 f には，等比減少関数

$$f(t, r_T, T) = \gamma^{T-t-1} r_T \quad (0 \leq \gamma \leq 1)$$

がよく用いられている．ここで，γ は割引率パラメータである．

等比減少関数を用いると，優先度 $P(s_t, a_t)$ の増分は時間をさかのぼるにつれて減っていく．つまり，前に使われた状態行動対ほど，その優先度の増分が小さくなる．等比減少信用割当関数は，割引率が

$$\gamma \leq \frac{1}{L}$$

のときに宮崎の完全合理性定理を満たす．この条件を，ここでは，宮崎の完全合理性条件と呼ぶ．

宮崎の完全合理性条件に従うと，取り得る行動の数が大きいときには信用割当が非常に小さくなり，学習の進行が遅くなってしまう．したがって，宮崎の完全合理性条件は実用的ではない．

3. Profit Sharing の期待合理性

宮崎の完全合理性定理と完全合理性条件は，profit sharing が直面する可能性がある最悪の場合について解析し，無効な行動を抑制する——すなわち，合理性が保たれるようにしている．しかしながら，実際には，そのようなケースは非常に稀である．

ここでは，profit sharing にいくつかの自然な制限を加えることにより，profit sharing にとって学習が最も困難な状態 [宮崎 94] において profit sharing が直面する平均的な場合について——すなわち，期待値の観点に基づいて profit sharing の合理性を示す．

3.1 非決定的単調増加行動選択法

期待値の観点に基づいた profit sharing の合理性を示すために，本論文では，行動選択法を次に定義されるような非決定的単調増加行動選択法に限定する．

定義 1 (非決定的行動選択法) 非決定的行動選択法は，任意の状態行動対 s, a に対して，行動選択確率が次の式を満たす．

$$\Pr(a_t = a \,|\, s_t = s) < 1$$

定義 2 (単調増加行動選択法) 単調増加行動選択法は，優先度が $P(s, a) < P(s, a')$ となる任意の状態行動対 s, a と s, a' に対して，行動選択確率が次の式を満たす．

$$\Pr(a_t = a \,|\, s_t = s) \leq \Pr(a_t = a' \,|\, s_t = s)$$

強化学習では，一般的に，非決定的単調増加行動選択法がよく用いられる．たとえば，Boltzmann 分布を用いたソフトマックス選択，ϵ-グリーディ選択，一様選択は，いずれも非決定的単調増加行動選択法である．ルーレット選択も，profit sharing においては非決定的単調増加行動選択法である．

したがって，強化学習において行動選択法を非決定的単調増加行動選択法に限定することは厳しい制約ではない．

3.2 期待強化値

Profit sharing において，ひとつのエピソードから得られる強化値は，そのエピソードに出現した状態行動対と最後に得られた報酬によって確率的に決まる．すなわち，profit sharing における強化値は，確率変数として表すことができる．

行動選択法を非決定的なものに限定すると，profit sharing にとって学習が最も困難——すなわち，ひとつ以上の有効な行動 $a_{(i)}$（$i = 1, 2, \ldots$）と唯一の回帰的で無効な行動 $a_{(0)}$ が競合している状態 $s_{(0)}$ において，profit sharing が n 回目のエピソードから得る有効な行動 $a_{(i)}$ に対する強化値 $\Delta P_n(s_{(0)}, a_{(i)})$ の期待値は，行動選択確率 $p_{n(0)}, p_{n(i)}$ と期待報酬 $\mathcal{R}^{a_{(i)}}_{s_{(0)}}$ によって表すことができる．ここで，$p_{n(i)}$ は，n 回目のエピソードにおける状態 $s_{(0)}$ での行動 $a_{(i)}$ の選択確率 $p_{n(i)} = \Pr_n(a_t = a_{(i)} \,|\, s_t = s_{(0)})$ を表す．

補題 2 (有効な行動の期待強化値) ひとつ以上の有効な行動 $a_{(i)}$（$i = 1, 2, \ldots$）と唯一の回帰的で無効な行動 $a_{(0)}$ が競合している状態 $s_{(0)}$ において，*profit sharing* が非決定的行動選択法に従って行動を選択するならば，*profit sharing* が n 回目のエピソードから得る有効な行動 $a_{(i)}$ に対する強化値 $\Delta P_n(s_{(0)}, a_{(i)})$ の期待値は，

$$\mathrm{E}\left[\Delta P_n(s_{(0)}, a_{(i)})\right] = \mathrm{E}\left[\frac{p_{n(i)}}{1 - p_{n(0)}} \mathcal{R}_{s_{(0)}}^{a_{(i)}}\right]$$

である．

証明は紙面の都合により省略する．

また，信用割当関数を等比減少関数に限定すると，同様にして，唯一の回帰的で無効な行動 $a_{(0)}$ に対する強化値 $\Delta P_n(s_{(0)}, a_{(0)})$ の期待値を行動選択確率 $p_{n(i)}$ と期待報酬 $\mathcal{R}_{s_{(0)}}^{a_{(i)}}$（$i = 0, \ldots, L$）によって表すことができる．

補題 3 (唯一の回帰的で無効な行動の期待強化値) L 個の有効な行動 $a_{(i)}$（$i = 1, \ldots, L$）と唯一の回帰的で無効な行動 $a_{(0)}$ が競合している状態 $s_{(0)}$ において，割引率 γ の等比減少信用割当関数を用いた *profit sharing* が非決定的行動選択法に従って行動を選択するならば，n 回目のエピソードから得る唯一の回帰的で無効な行動 $a_{(0)}$ に対する強化値 $\Delta P_n(s_{(0)}, a_{(0)})$ の期待値は，

$$\mathrm{E}\left[\Delta P_n(s_{(0)}, a_{(0)})\right] = \mathrm{E}\left[\frac{\gamma p_{n(0)}}{\left(1 - p_{n(0)}\right)\left(1 - \gamma p_{n(0)}\right)} \sum_{i=1}^{L} p_{n(i)} \mathcal{R}_{s_{(0)}}^{a_{(i)}}\right]$$

である．

この証明も紙面の都合により省略する．

3.3 期待合理性定理

Profit sharing において，確率的な行動選択法を用いると，行動選択確率 $\Pr(a_t = a \mid s_t = s)$ は，それまでに経験した状態行動対と獲得した報酬によって確率的に決まる確率変数として表される．そこで，この確率変数の期待値について考える．

補題 2 と補題 3 から，profit sharing にとって学習が最も困難な状態 $s_{(0)}$ において，profit sharing が等比減少信用割当関数を用い，非決定的単調増加行動選択法に従って独立に行動を選択するときの有効な行動に対する強化値の期待値と無効な行動に対する強化値の期待値の大小関係を調べ，補題 1 と合わせてこれを一般的な状態に拡張することによって次の定理を導くことができる．

定理 2 (Profit sharing の期待合理性) *Profit sharing* が，等比減少信用割当関数を用い，かつ，非決定的単調増加行動選択法に従って独立に行動を選択するならば，有効な状態が存在する任意の状態 s において，任意の行動 a に対して

$$\mathrm{E}\left[\Pr(a_t = a' \mid s_t = s)\right] \geq \mathrm{E}\left[\Pr(a_t = a \mid s_t = s)\right]$$

を満たす有効な行動 a' が存在する．

この証明も紙面の都合により省略する．本論文では，これを profit sharing の期待合理性定理と呼ぶ．

この定理では，信用割当関数については，関数を等比減少なものに制限するだけでその割引率 γ を制限していない．つまり，期待合理性定理における γ の条件は

$$0 \leq \gamma \leq 1$$

である．本論文では，これを期待合理性条件と呼ぶ．

4. 考察

宮崎の完全合理性定理が最も学習が困難な状態における最悪ケースの分析であるのに対し，本論文が示した期待合理性定理は同状態における平均的ケースの分析である．

期待合理性定理は，profit sharing が，等比減少信用割当関数を用い，非決定的単調増加行動選択法に従って独立に行動を選択するならば，有効な行動が存在するどのような状態においても，選択確率の期待値が最大となる有効な行動が存在することを示している．

期待合理性条件は，上の条件を満たしている profit sharing を用いるならば，平均的には，割引率 γ を上限 1 までの範囲内で大きくしてもいいことを示している．また，エージェントが取り得る行動の数に関係なく割引率を設定することができるという点で，宮崎の完全合理性条件よりも優れている．

5. まとめ

本論文では，期待値の観点に基づく profit sharing の合理性について述べた．

本論文に示した期待合理性定理は，profit sharing が (1) 等比減少信用割当関数を用い，(2) 非決定的単調増加行動選択法に従い，(3) 独立に行動を選択するならば，有効な行動の選択確率の期待値が最大となることを示している．すなわち，割引率 γ は，等比減少関数となるための条件 $0 \leq \gamma \leq 1$ を満たしていればいい．この期待合理性条件は，(1) より速く学習でき，(2) 取り得る行動の数に依存しないという点で宮崎の完全合理性条件よりも優れている．

紙面の都合により省略した証明については，近いうちに人工知能学会論文誌に発表したい．

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)

[Holland 86] Holland, J. H.: Escaping Brittleness: The Possibilities of General-Purpose Learning Algorithms Applied to Parallel Rule-Based Systems, in Michalski, R. S., Carbonell, J. G., and Mitchell, T. M. eds., *Machine Learning: An Artificial Intelligence Approach*, Vol. 2, Morgan Kaufmann Publishers (1986)

[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, 人工知能学会誌, Vol. 9, No. 4, pp. 580–587 (1994)

[Sutton 98] Sutton, R. S. and Barto, A. G.: *Reinforcement Learning: An Introduction*, The MIT Press (1998), 三上貞芳, 皆川雅章 共訳. 強化学習. 森北出版, 2000