3F4-02

罰を考慮したProfit Sharing強化学習法

Profit Sharing Considering Penalty

鈴木淳司[*1] 松井藤五郎[*2] 世木博久[*3]
Atsushi Suzuki Tohgoroh Matsui Hirohisa Seki

[*1]岐阜大学大学院地域科学研究科
Graduate School of Regional Studies, Gifu University

[*2]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

[*3]名古屋工業大学知能情報システム学科
Department of AI and Computer Science, Nagoya Institute of Technology

We propose Flexible Profit Sharing which is based on Profit Sharing. This method guarantees the minimum selection probability drawn from a viewpoint of the expected value of a credit assignment, not to reinforce an invalid rule. This method allows an agent to learn from penalty (negative reward) in which treating was forbidden by the conventional technique. Moreover, we show that Flexible Profit Sharing learns earlier than Profit Sharing.

1. はじめに

強化学習とは，報酬という特別な入力を手がかりに，与えられた目標を達成するための方法を相互作用から学習する機械学習アルゴリズムである．

近年，多くの研究では，目標達成時に正の報酬，制約違反時に罰として負の報酬を付与し，単位行動当たりの期待獲得報酬の最大化，すなわち最適政策の獲得を追求する場合が多い [宮崎 01]．

強化学習アルゴリズムの 1 つである Profit Sharing(PS) は，状態行動対のユニークな価値が特定できない不完全知覚領域においても有効な確率的政策を獲得できる [荒井 98] ことから，マルチエージェント環境などでの利用が可能なことが知られている [荒井 01]．

しかし，PS は負の報酬が扱えないという欠点がある．本論文では，PS で，制約違反時に罰として与えられる負の報酬を扱うための方法を提案する．

2. Profit Sharing とその問題点

PS は，エピソードに含まれる各状態行動対 s_t,a_t をエピソード終了後に一括して次のように強化する．

$$W(s_t, a_t) \leftarrow W(s_t, a_t) + f(t, R_T, T) \tag{1}$$

ここで，W は状態行動対の価値を表す関数，f は強化関数，T はエピソードが終了した時刻である．

強化関数 f として等比減少関数

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

が知られており，よく用いられている [宮崎 94, 荒井 98, Arai 00]．ここで，γ は割引率パラメータ (discount rate)，R_T は時刻 T で得られる報酬である．PS では，行動選択に重み付きルーレット選択を用いる．

連絡先: 鈴木 淳司, 岐阜大学大学院地域科学研究科所属, 〒 501-1112 岐阜県岐阜市柳戸 1-1 岐阜大学地域科学部 宮城研究室, E-mail:a-suzuki@lily.ics.nitech.ac.jp, Tel: 058-293-6107

$$\Pr(s, a) = \frac{W(s, a)}{\sum_{a' \in A(s)} W(s, a')} \tag{2}$$

ここで，$\Pr(s, a)$ は状態 s で行動 a を選択する確率，$A(s)$ は状態 s で実行可能な行動の集合を表す．この式は W 値が負のときは，選択確率が負になることがある．これは，明らかに不適当である．このため，W 値が負の値にならないように，従来の PS では，負の報酬を禁止している．

3. 提案手法

3.1 Flexible Profit Sharing

罰を扱うことができる **Flexible Profit Sharing**(FPS) を提案する．FPS は W 値が強化されたときに負になったとしても，正の値に修正するように，W 値に対して修正値 $\rho(s)$ を加える．

$$W(s, a) \leftarrow W(s, a) + \rho(s) \tag{3}$$

つまり，式 (1) でエピソード中のすべての W 値を強化したあと，式 (3) のようにエピソードに含まれる状態のすべての状態行動対の W 値に一律に $\rho(s)$ を加える．式 (3) で更新された W 値が，常に正であれば，重み付きルーレット選択を用いた行動選択確率は正であることが保証される．しかし，この手法は $\rho(s)$ 値の大小によって，学習の性能に大きな影響を与えてしまう．それゆえ，適切な $\rho(s)$ 値を設定する必要がある．

本論文では，強化値の期待値の観点から，合理的な政策を獲得するために必要最小限の $\rho(s)$ 値を導出する．

Flexible Profit Sharing の手続き的なアルゴリズムを，図 1 示す．

3.2 最も困難な競合構造の拡張

PS の行動選択確率の大小は，強化値の総和の大小によって決定される．FPS では，状態 s のすべての行動優先度 $W(s, a)$ には，同じ $\rho(s)$ 値を加えるため，FPS の行動選択確率の大小は，PS と同様に強化値の総和の大小によって決定される．

それゆえ，FPS における学習が最も困難な競合構造を考えるには，PS と同様に，強化値の大小のみに注目すればよい．

すべての $s \in \mathcal{S}$, $a \in \mathcal{A}$ に対して：
　$W(s,a) = C$　（C は 0 でない任意の定数）
各エピソードに対して繰り返し：
　s を初期化
　エピソード中の各ステップに対して繰り返し：
　　W から導かれる重み付きルーレット選択を用いて，
　　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
　エピソードに含まれるすべての状態行動対に対して：
　　$W(s_t,a_t) \leftarrow W(s_t,a_t) + f(t, r_T, T)$
　エピソードに含まれる状態を含む
　　すべての状態行動対に対して：
　　$W(s_t,a_t) \leftarrow W(s_t,a_t) + \rho(s)$

図 1: Flexible Profit Sharing アルゴリズム.

PS の報酬は非負の値としているが，このとき学習が最も困難な競合構造は，図 2 に示すような，L 本の有効ルールと唯一の回帰的無効ルールが競合している構造である．この構造のとき，回帰的無効ルールの抑制が最も困難である [宮崎 94].

FPS は，罰として報酬に負の値も扱う．しかし，その場合も学習が最も困難な競合構造は，従来のそれと同じ構造である．その証明を付録Aに示す.

また，得られる報酬が最も大きい有効ルールで得られる報酬を $R_{\max}$ とすると，学習が合理的であると言えるのは，$R_{\max} > 0$ のとき，無効ルールの抑制することであり，$R_{\max} \leq 0$ のとき，無効ルールの抑制しない（促進する）ことである.

3.3　強化値の期待値の合理性

学習が最も困難な構造において，最悪のケース (worst case) を考える．すなわち，唯一の回帰的無効ルールが最も強化されていて，かつ，得られる報酬 R が最も大きい有効ルールが最も強化されていない状況を設定する.

このときの無効ルールの選択確率を $\Pr_{\max}$，最も強化されていない有効ルールの選択確率を $\Pr_{\min}$ と表記する．ここで，$\Pr_{\max}$, $\Pr_{\min}$ 条件は，$0 < \Pr_{\min} < \Pr_{\max} < 1$ である.

このような状況であっても，強化値の期待値の観点から合理的な $\rho(s)$ 値の範囲を考える．強化値の期待値が合理的であるというのは，

$R_{\max} > 0$ のとき，

$$\text{最も強化されていない有効ルールの強化値の期待値} \geq \text{無効ルールの強化値の期待値} \quad (4)$$

$R_{\max} \leq 0$ のとき，

$$\text{最も強化されていない有効ルールの強化値の期待値} \leq \text{無効ルールの強化値の期待値} \quad (5)$$

が成り立つことである.

唯一の回帰的無効ルール $(s, a_{(0)})$ に対する強化値の期待値は，次式のようになる.

$$\frac{\Pr_{\max}\gamma}{1-\Pr_{\max}\gamma} E_R \quad (6)$$

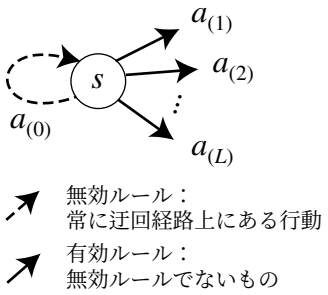

図 2: 最も困難な競合構造

ここで，

$$E_R = \frac{\sum_{i=1}^{L} \Pr_{a_{(i)}} R_{a_{(i)}}}{\sum_{i=1}^{L} \Pr_{a_{(i)}}} \quad (7)$$

である．$\Pr_{a_{(i)}}$ は行動 $a_{(i)}$ を選択する確率，$R_{a_{(i)}}$ は行動 $a_{(i)}$ を選択したときに得られる報酬を表す [松井 03].

次に，最も強化されていない有効ルールに対する強化値の期待値について考える．$R_{\max} = R_{a_{(k)}}$ とすると，ルール $(s, a_{(k)})$ が最も強化されていないときの強化値の期待値は次式のようになる.

$$\frac{\Pr_{\min}}{1-\Pr_{\max}} R_{\max} \quad (8)$$

3.4　合理性を満たす $\Pr_{\min}$ の範囲

$R_{\max} \geq 0$ のとき，式 (4) が成り立つとき，合理的である．式 (6), (8) より，式 (4) を常に満たすような $\Pr_{\min}$ の範囲を求める.

$$\frac{\Pr_{\min}}{1-\Pr_{\max}} R_{\max} \geq \frac{\Pr_{\max}\gamma}{1-\Pr_{\max}\gamma} E_R \quad (9)$$

式 (7) より，$E_R \leq R_{\max}$ であり，式 (9) の右辺は高々 $\frac{\Pr_{\max}\gamma}{1-\Pr_{\max}\gamma} R_{\max}$ である．これより，次式を満たせばよい.

$$\Pr_{\min} \geq \frac{\Pr_{\max}\gamma(1-\Pr_{\max})}{1-\Pr_{\max}\gamma} \quad (10)$$

$R_{\max} < 0$ のとき，同様に，式 (6), (8) より，式 (5) を常に満たすような $\Pr_{\min}$ の範囲を求める.

$$\frac{\Pr_{\min}}{1-\Pr_{\max}} R_{\max} \leq \frac{\Pr_{\max}\gamma}{1-\Pr_{\max}\gamma} E_R \quad (11)$$

ここで，$R_{a_{(1)}} = R_{a_{(2)}} = \cdots = R_{a_{(L)}} = R_{\max}$ と仮定すると，$E_R = R_{\max}$ なので，式 (11) は次のように書ける.

$$\begin{aligned} \frac{\Pr_{\min}}{1-\Pr_{\max}} R_{\max} &\leq \frac{\Pr_{\max}\gamma}{1-\Pr_{\max}\gamma} R_{\max} \\ \frac{\Pr_{\min}}{1-\Pr_{\max}} &\geq \frac{\Pr_{\max}\gamma}{1-\Pr_{\max}\gamma} \\ \Pr_{\min} &\geq \frac{\Pr_{\max}\gamma(1-\Pr_{\max})}{1-\Pr_{\max}\gamma} \end{aligned} \quad (12)$$

式 (10) と式 (12) は同値であり，これを満たすとき、強化値の期待値のうえで合理性を満たす.

3.5 合理性を満たす最低遷移確率 $g(\gamma)$

式 (10), (12) の右辺の最大値を考えると，$\Pr_{\max} = (1 - \sqrt{1-\gamma}\,)/\gamma$ のとき，最大値 $(1 - \sqrt{1-\gamma}\,)^2/\gamma$ をとる．また，$1/2 < (1 - \sqrt{1-\gamma}\,)/\gamma \leq 1$ である．

$$g(\gamma) = \frac{(1 - \sqrt{1-\gamma}\,)^2}{\gamma}$$

とすると，式 (10), (12) を満たすためには，$\Pr_{\min} \geq g(\gamma)$ を満たせばよい．

3.6 $\rho(s)$ 値の範囲

常に $\Pr_{\min} \geq g(\gamma)$ を満たす $\rho(s)$ 値の範囲を考える．各状態の行動選択確率は式 (2) で与えられる．最も強化されていないルールの行動優先度を $W_{\min}$ とすると，

$$W_{\min} \geq g(\gamma) \sum_{j=0}^{L} W(s, a_{(j)}) \tag{13}$$

となる．式 (3) のように $\rho(s)$ 値を加えることで，式 (13) を満たしていればよいので，次のように書ける．

$$W_{\min} + \rho(s) \geq g(\gamma) \sum_{j=0}^{L} \{W(s, a_{(j)}) + \rho(s)\}$$

$$W_{\min} + \rho(s) \geq g(\gamma) \left\{ \sum_{j=0}^{L} W(s, a_{(j)}) + (L+1)\rho(s) \right\}$$

$$\rho(s) \geq \frac{g(\gamma) \sum_{j=0}^{L} W(s, a_{(j)}) - W_{\min}}{1 - (L+1)g(\gamma)}$$

宮崎の合理性定理 [宮崎 94] を満たす $\gamma = 1/(L+1)$ を用いると，$(L+1)g(1/L+1)$ は L について単調減少関数で，$\lim_{L \to 1}(L+1)g(1/L+1) = 6 - 4\sqrt{2} \quad 0.34$ である．

これらより，FPS では次式のような値を用いる．

$$\begin{cases} \gamma = 1/(L+1) \\ \rho(s) = \dfrac{g(\gamma) \sum_{j=0}^{L} W(s, a_{(j)}) - W_{\min}}{1 - (L+1)g(\gamma)} \end{cases}$$

4. Flexible Profit Sharing の利点

4.1 負の報酬が使用可能

エピソードが終了し，強化値を加えた後の $W(s,a)$ 値が下式を満たすとき，$\rho(s) > 0$ となる．

$$\frac{W(s,a)}{\sum_{a' \in A(s)} W(s,a')} = \Pr(s,a) < g(\gamma)$$

特に，PS では禁止されている $W(s,a) < 0$ のときには，$W(s,a) + \rho(s) > 0$ となるような $\rho(s)$ 値をとる．それゆえに，重み付きルーレット選択は正常に働くことになる．

W 値に対する制限 ($W \geq 0$) は必要なくなるので，報酬 R について，従来は禁止していた範囲 ($R < 0$) も制限を加える必要がない．したがって，FPS では，報酬 R について制限はなく，任意の R 値を扱える．

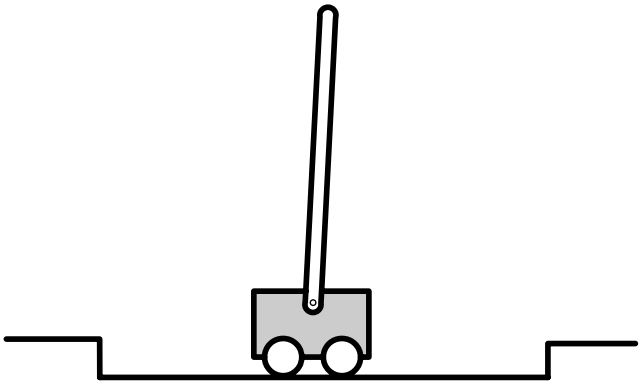

図 3: ポールバランシング問題

表 1: ポールバランシング問題の実験結果

	試行回数	成功回数	平均エピソード数
提案手法 (FPS)	20	20	6530
従来手法 (PS)	- - - -	- - - -	- - - -

4.2 学習速度の向上

一般に，Profit Sharing を用いた学習は，大きな割引率 γ を用いるほど学習が速くなると考えられている．しかし，宮崎の合理性定理で定められた範囲より大きい γ 値を用いると，有効ルールより無効ルールを強化してしまう恐れがある．

FPS は，宮崎の合理性定理を満たした γ 値を用いるので，このようなことなく，また，同じ γ 値を用いた PS より速く学習することができる．

$$\frac{W(s,a)}{\sum_{a' \in A(s)} W(s,a')} = \Pr(s,a) > g(\gamma)$$

なぜなら，状態 s のすべての行動優先度が上式を満たすとき，$\rho(s) < 0$ となり，その $\rho(s)$ 値を同状態の行動優先度に一律に加えるため，強化値の累積で生じた行動優先度の差が強調されるからである．この効果は，エピソードが長くなりがちなタスクにおいて，顕著に現れる．

また，この仕組みのため行動優先度の初期値 W_0 による学習速度の差は生じない．

5. 実験

5.1 ポールバランシング問題

負の報酬からの学習効果を確認するため，ポールバランシング問題 [Sutton 98] を用いて実験を行った（図 3）．この問題は，負の報酬しか存在しないため，Profit Sharing では取り扱えない問題として知られている．

実験では，台車の位置・速度とポールの角度・速度から状態を 162 に分割している．行動は，台車を前・後に動かすの 2 種類である．得られる報酬は，ポールを地面に落としてしまったときに -1 が与えられるのみである．終了条件は 10 万ステップの間，ポールを地面に落とさないことである [RLR]．

本実験では，重み付きルーレット選択を用いて学習させ，終了条件はグリーディー選択を用いて判定する．実験結果を表 1 に示す．実験結果から，FPS が負の報酬しか存在しない問題に順応できることがわかる．

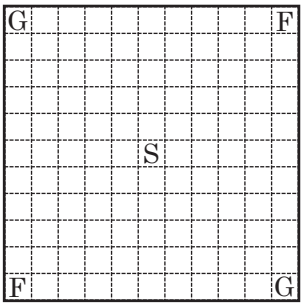

図 4: 11×11 の迷路問題

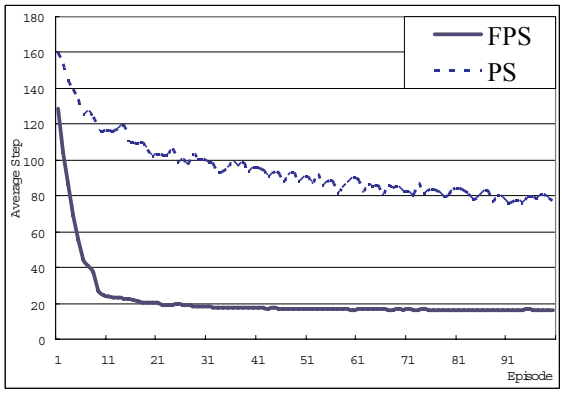

図 5: 11×11 の迷路問題の実験結果

5.2 11×11 の迷路問題

本手法の有効性を確認するために，迷路問題を用いて実験を行った．図 4 に示すような 11×11 の格子状のマス目を用意する．中央のマス目 (S) をスタート地点とする．左上と右下のマス目 (G) を成功ゴールとし，到達したときは報酬 1.0 を与え，左下と右上のマス目 (F) は失敗ゴールとし，到達したときには報酬を与えない．エージェントの行動は（上，下，左，右，その場に留まる）の 5 種類とする．性能比較には PS を用い，パラメータは $\gamma = 1/5$, $W_0 = 0.1$ とした．

各手法における，乱数の種を変えて行った 20 回の実験でのゴール到達までの平均ステップ数を測定した（図 5）．縦軸が平均ステップ数，横軸がエピソード数を意味する．

この問題は成功ゴールに到達するまでの最小ステップ数は 10 ステップである．FPS と PS は同じ γ 値を用いたにもかかわらず，図 5 は FPS が PS よりも速く学習したことを示している．これは，4.2 節で述べたことが有効に機能しているからである．

6. まとめ

本論文で提案した Flexible Profit Sharing は，従来の Profit Sharing に比べて，次のような利点を持つ．

- 負の報酬が扱える
- Profit Sharing より学習が速い
- 行動優先度の初期値 W_0 のパラメータ設定が不要

今後は，Q-Learning や Sarsa(λ) など他の類似手法との比較，マルチエージェントタスクでの学習効果の確認などが課題である．

A 負の報酬が存在する場合の最も困難な構造の証明

行動選択確率の大小関係は，強化値の累積の大小関係のみで決定する．したがって，無効ルールの強化値のみに注意すればよい．

1. 正の報酬と負の報酬が混在しているとき

   無効ルールを抑制することが合理的であり，2 に比べて，明らかに無効ルールを抑制することは簡単である．

2. 正の報酬と負の報酬が混在していないとき

   (a) 正の報酬のみが存在するとき，無効ルールを抑制することが合理的である．明らかに無効ルールが強化される回数が多いほど，無効ルールを抑制することは困難になる．

   (b) 負の報酬のみが存在するとき，無効ルールを抑制しない（促進する）ことが合理的である．明らかに無効ルールが強化される回数が多いほど，無効ルールを抑制しない（促進する）ことはより困難になる．

   (a), (b) より，強化される回数の大小のみを考えれば，十分である．以降，[宮崎 94] の付録Aと同じ．

ゆえに，無効ルールを抑制することが最も困難となる構造は，L 本の有効ルールと唯一の回帰的無効ルールと競合している構造である．

参考文献

[荒井 98] 荒井, 宮崎, 小林：マルチエージェント強化学習の方法論－ Q-learning と Profit Sharing による接近－, 人工知能学会誌, Vol.13, No.4, pp.609-618(1998)

[Arai 00] Arai, S. , Sycara, K. ,and Payne, T. R. : Experience-Based Reinforcement Learning to Acquire Effective Behavior in a Multi-agent Domain, in Mizoguchi, R. and Slaney, J. eds. , *Proceedings of the 6th Pacific Rim International Conference on Artificial Intelligence(PRICAI2000)*, pp.125-135, (2000).

[荒井 01] 荒井幸代：マルチエージェント強化学習－実用化に向けての課題・理論・諸技術との融合－, 人工知能学会誌, Vol.16, No.4, pp.467-481(2001)

[松井 03] 松井藤五郎：自律型エージェントの行動学習に関する研究, 名古屋工業大学学位審査論文 (2003)

[宮崎 94] 宮崎, 山村, 小林: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)

[宮崎 01] 宮崎, 坪井, 小林：罰を回避する合理的政策の学習, 人工知能学会誌, Vol.16, No.2, pp.148-156(2001)

[RLR] Reinforcement Learning Repository, University of Massachusetts, Amherst, http://www-anw.cs.umass.edu/rlr/

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