学習型行動選択ネットワーク: L-ANA

栗原 聡　　菅原 俊治
NTT 光ネットワークシステム研究所

概要

行動選択ネットワークを経験強化型の強化学習に適応した L-ANA を提案する。従来の強化学習法である profit-sharing に比べ、動的な環境の変化に対してよりロバスト性を有し、実世界やインターネットなど複雑で動的な環境を対象とする自律行動主体に適した学習法である。また活性伝幡のためのパラメータを操作することで、容易に学習の特性を制御することができる。今回はシミュレーションにより L-ANA の基本的有効性の確認を行なった。

1　はじめに

近年，実世界やインターネットなど複雑で動的な環境を対象とする自律行動主体のための人工知能に関する研究が盛んである．我々も自律移動ロボットやソフトウェアエージェントなどを念頭に置いたプランニングや、探索アルゴリズムなど、自律行動主体のための行動選択メカニズムに関する研究を行なっている。

動的な環境において有効な行動選択メカニズムとしては，Subsumption Architecture (SA)、実時間リアクティブプランニング，Agent Network Architecture (ANA)[6]、そして我々が提案した MRR-planning[7] などが挙げられ、いずれも動的な変化に即応できることが特徴である。しかしより環境に適応するためには、環境内で動作する自律行動主体が遭遇する種々の状況に個別に対応できるための「学習機能」が必要不可欠である。しかしながら学習機能を組み込んだ行動選択メカニズムの枠組までは提案されていないのが現状である。

学習するためには環境の情報を収集することが必要であるが、実世界やインターネット世界すべての情報をあらかじめ詳細に得るのは不可能である。従ってこれら行動主体はセンサー等を用いてローカルな情報を収集しつつ環境のモデルを構築することになる。しかし環境が動的に変化するため得られたモデルは常に不完全である。このような状況においては、特に動的な環境の変化にロバストな経験強化型の強化学習法が必要である。

従来の強化学習における経験強化型の学習法としては profit-sharing[2] などが有名であり、 環境同定型の学習法である Q-learning[1] などと比べ、動的な環境の変化に対してロバストであることが報告されているが、我々が対象とする環境の動的性を考えるとまだまだ有効とはいえない。

そこで、今回我々は ANA を代表とする活性伝播型の行動選択メカニズムを経験強化型の強化学習に適応した学習型行動選択ネットワーク (L-ANA) を提案する。profit-sharing における強化値の配分を活性伝播で行ない、動的な環境の変化に柔軟に対応する。また学習を活性伝播で行なうことから、活性伝播の持つ特徴を学習に生かせるだけでなく、従来の行動選択の枠組に学習機能を容易に組み込むことも容易である。

以下 2 章では行動選択ネットワーク並びに profit-sharing について簡単に説明し、3 章にて L-ANA を評価するためのシミュレーション環境について説明する。そして 4 章にて今回提案する L-ANA についての説明を行ない、5 章において profit-sharing との比較評価結果について報告する。6 章では L-ANA の特徴である動的環境の変化に対するロバスト性について述べ、その評価結果について報告し、7 章にて議論、8 章にてまとめを述べる。

2 背景

行動選択ネットワーク ANA[6] は比較的単機能なエージェント集合が互いに活性値を伝搬し合うことで協調し、エージェント全体として合目的に機能する枠組である。集中制御が不要で各エージェントがそれぞれ自律的に振舞うことから、特にロバスト性と拡張性を特徴とする。活性伝播には前向きと後向きがあり、それぞれ環境指向と目的指向の行動選択を誘発する。なお、L-ANA では後向きの活性伝播を利用する。

profit-sharing[2] は経験強化型の強化学習法であり、報酬を得た時にそれまでの行動系列を一括して強化する。学習に要する試行回数が少ないこと、また Q-learning[1] と異なり環境が動的に変化しても、変化しなかった部分の強化値が保存されるという意味でロバスト性がある。

強化学習法としては近年 Q-learning が注目されている。Q-learning は環境同定型の強化学習法であり、Q 値を求めるために環境状態を正確に同定する必要がある。従って profit-sharing に比べて非常に多くの試行回数を要することや、環境が動的に変化してしまうと、それまで得られた学習結果全体に影響が及んでしまうなどの問題点が指摘されている。従って行動主体が環境全体の詳細な知識を獲得できるような状況においては Q-learning は適しているものの、今回我々が対象としている自律行動主体のように環境の不完全な知識しか持つことのできない状況においては、profit-sharing のような学習法の方が適していると考えられる。

profit-sharing が Q-learning に比べて動的環境の変化に強い理由は、profit-sharing では獲得した学習結果が保存され、変化の影響を受けなかった部分の学習効果が持続されるからである (Q-learning ではローカルな環境の変化でもその影響が学習全体に影響を与えてしまう)。しかし profit-sharing において変化の影響を受けてしまった部分の学習は無効とするしかなく、学習効果が動的な環境の変化の規模に大きく依存するという限界がある。

3 シミュレーション環境

まず、L-ANA の説明を行なうためにも今回 L-ANA の性能を評価するためのシミュレーション環境について説明する。図 1のような格子状の状態空間 S 内を移動

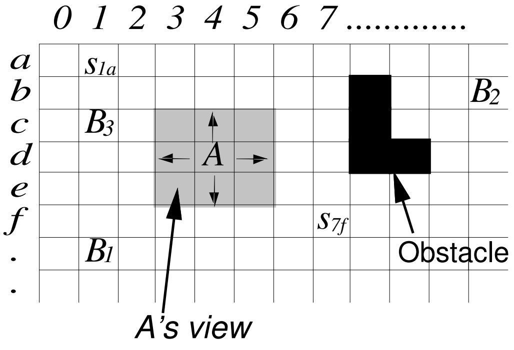


図 1: グリッドワールド

する自律行動主体 A を考える。S を構成する個々の状態はそれぞれ S_{1a}、S_{7f} のように表記する。A は S 内を上下左右に 1 ブロックづつ移動することができ、移動に際しては A に搭載された仮想バッテリーを使用する。バッテリーは移動することにより減少するが、充電ポイント B にて補給することが可能である。B は S 内に数箇所存在し、また補給できるエネルギー量もそれぞれ異なっている。また S 内には障害物も存在し、A は障害物を通過することはできない。B の位置と数は変化しないが、障害物の位置と数は動的に変化するものとする。なお、報酬は補給エネルギー量に比例して与えられる。

A は初期状態では S が格子状の世界で自分が上下左右に移動できること以外、B や障害物の位置に関しての情報は一切持っていない。ただし A には仮想センサーが搭載されており、図 1にある灰色の範囲に関する環境の状況を得ることができる。

A は S に依存して変化する内部状態空間 A_S を有し、この例題では $A_S = \{A_{s_1}$: 充電が必要、A_{s_2} : 充電は必要ない$\}$ という 2 種類の内部状態を考えることができる。内部状態の数は A や S が複雑になるにつれて多くなり、同時に複数の状態が成立する状況となる。

A に要求される学習とは、より多くのエネルギーを補給できる B と、その B への移動経路を学習することである。また、補給エネルギー量は少ないものの、近くの B で補給したほうが効率がよいということも学習できることが望まれる。そして障害物が動的に消滅 / 出現するので、これを察知した時に柔軟に対応でき、学習の効率を低下させないことが何より重要である。

4 L-ANA

L-ANA は profit-sharing と同様に経験強化型の強化学習を行なう。profit-sharing では、エピソードに相当する状態遷移系列を単位として学習を行ない、エピソードとエピソードに割り振られた学習値を記憶する。例えば図 2左のように A が p の状態遷移系列により偶然 B にたどり着けたとすると、まず B の存在する状態に対して報酬が与えられ、さらにエピソード p を構成する各状態に対して強化関数に基づいた強化値が割り振られる。このように 1 つのエピソードを記憶することが profit-sharing における 1 回の学習である。

これに対し、L-ANA では図 2中のようにエピソードを構成する各状態に対して

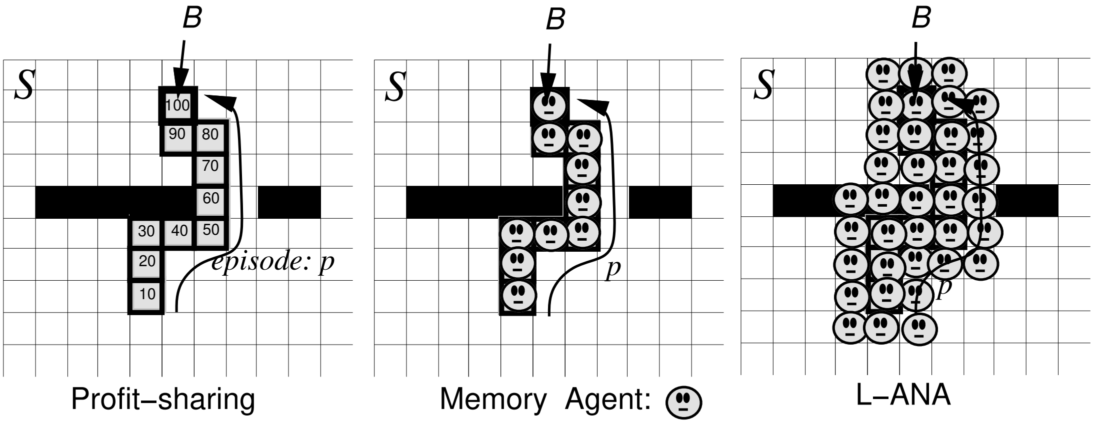


図 2: エピソードと記憶エージェント

一つづつ記憶エージェントを割り付けていく。記憶エージェントはミンスキーが [3] にて述べるエージェントと同類の自律主体であり、後述する活性伝幡は記憶エージェント郡の協調動作として実現される。

L-ANA では報酬を得たことにより変化する内部状態 A_{s_i} が学習の対象となり、この時のエピソードは初期状態あるいはこの A_{s_i} が成立した時から A_{s_i} が不成立となるまでの状態要素の遷移系列である[1]。上述したように内部状態は同時に複数成立している可能性があり、 A が報酬を得られない状況においても他の内部状態は変化することが考えられるがこの変化は A_{s_1} についての学習の対象とはしない。

報酬を得た時に行なわれる強化学習は profit-sharing と同様で、報酬を得た状態に割り当てられた記憶エージェントを s_{ij} とすると、「$s_{ij}(t+1) = s_{ij}(t) - bs_{ij}(t) + bp(t)$、$b$: 学習率、p : 報酬」に従って強化値の更新を行なう。 profit-sharing はエピソードに対して、報酬からそれだけ過去であるかを引数として強化値を返す強化関数に基づいた強化値を割り当る。

profit-sharing においては強化関数の性能が学習効率に大きく影響を与え、この点に関しては [4] [5] においていろいろな試みが行なわれている。 L-ANA では活性伝幡による強化値の伝播を行なう。活性伝幡は報酬を得た状態に割り当てられた記憶エージェント s_{ij} を出発点として、エピソードを構成する各状態に割り当てられた記憶エージェントに向かって強化値を伝幡するように行なわれる。

なお、学習された強化値の利用の仕方は profit-sharing と同様に、より強化値の大きい状態に移動するという方針である。

4.1 活性伝搬

profit-sharing ではエピソードを構成する状態系列にのみ強化値を割り振るわけだが、これは経験以外の不確定要素を排除し経験しか信用できない状況において有効な手段である。しかしながら行動主体が視野を持ちローカルな情報収集が可能で

---

[1] 例えば、 A_{s_1} が成立している状態で充電を行なうことができたとすると、 A_{s_1} は不成立となり代わりに A_{s_2} が成立する。この場合 A_{s_1} についての強化学習が行なわれる。

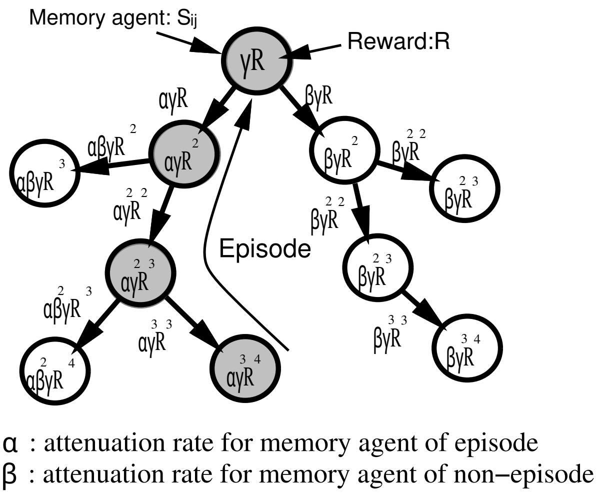


図 3: 活性伝幡

ある状況を考えると、行動主体から見える範囲においてはエピソードに割り振る強化値よりも小さい割合[2]で強化値を割り振ることは学習に柔軟性とロバスト性を与えられる可能性がある。

そこで L-ANA では A が移動する際に視野に入ったエピソード以外の状態に対しても記憶エージェントを割り当てていく。例えば図 2中のようなエピソードが得られたとすると、A が割り当てることができる記憶エージェントは図 2右の部分となる。各記憶エージェントは割当てられる時に、自分がどの記憶エージェントと隣接[3]関係にあるのか、また自分がエピソードを構成しているかどうかを記憶する。割り付けられた時にはエピソードを構成していなくても、その後の試行でエピソードの一部分となる場合もある。

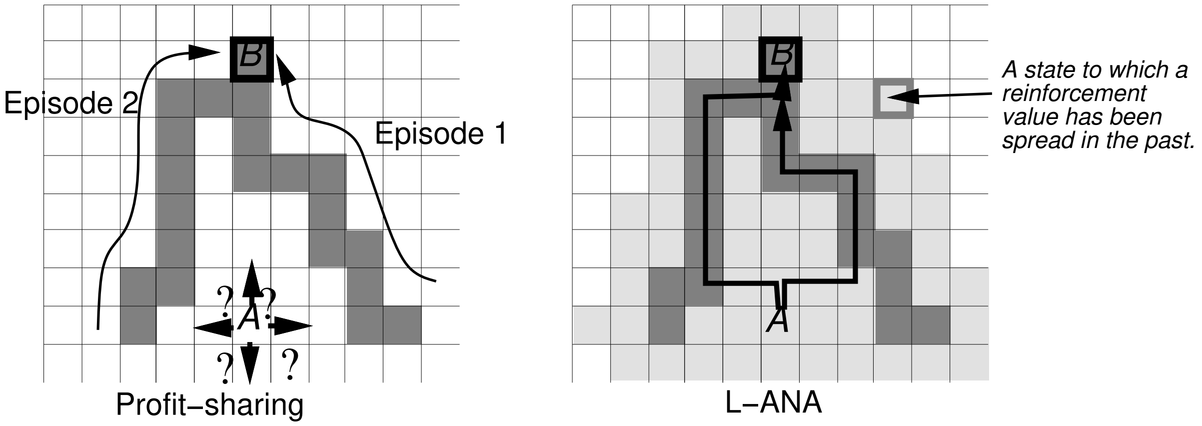
記憶エージェントは隣接する記憶エージェントから強化値が伝幡されると自発的に自分が隣接する他の記憶エージェントに向かって強化値を伝幡する。伝幡の仕方は図 3にあるように、全体としてある減衰を行なった後、エピソードを構成した経験がある記憶エージェントとそれ以外で異なる減衰率により強化値を伝播する。伝幡する強化値がある閾値以下になった時点で活性伝播を終了する。 L-ANA における学習とは、各記憶エージェントが自分がどの隣接する記憶エージェントに対し、それくらいの強化値を伝播するのかを記憶することである。

エピソード以外の記憶エージェントに対しても活性伝播を行なうことにより、学習に柔軟性を持たせることができる。例えば図 4のように profit-sharing ではどちらかのエピソードに出会うまではランダムに移動するしかないが、 L-ANA では A が既に活性伝播された状態に位置していれば、近隣のエピソードに最短経路で引き込まれるので効率良く B まで移動することができる。

---

[2] 経験を重視するために差をつける必要がある。

[3] 隣接関係にあるのは上下左右に存在する記憶エージェント。



Profit−sharing　　L−ANA

図 4: 活性伝幡の効果

4.2　活性伝幡の特性について

活性伝幡は、【1】活性伝幡を行なう際に用いる強化値の大きさと、【2】伝幡する時に用いる減衰率、の 2 つのパラメータによりその特性を容易に操作することができ、この 2 つのパラメータの設定の仕方により、以下のように学習の特性を制御することができる。

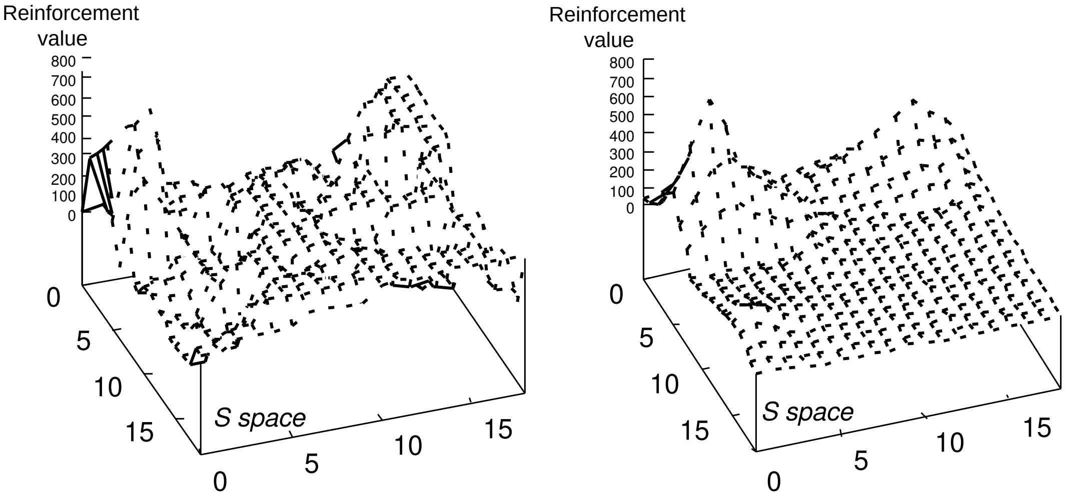
(1)A が報酬を得られる状態の近傍に来た時のみ、学習効果が発揮されるような学習の効果を得たい場合には、強化値を大きく減衰率を高くする。つまり報酬を得た状態を頂点とする強化値の山を考えると、その高度は高くしかも急勾配となる。例えば、S 内に A に対する捕食者を考えてみると、捕食者から逃げるのは捕食者が A の近傍に迫った時のみでよい。このような学習を行なうには (1) の設定が有効である。

(2)A が報酬を得る状態から離れている状態でも学習効果が発揮されるようにするには、強化値を小さく減衰率を低くする。つまり報酬を得た状態を頂点とする強化値の山を考えた時、その高度は低くしかも勾配もなだらかとなる。例えばエネルギーを充電するようなことを学習することを考えた時には (2) のように学習効果が広く行き渡っていた方がよい。強化値を (1) に比べて小さく設定しなければならない理由は、(1) が (2) の強化値の分布に完全に含まれないようにするためである。(2) の学習結果に従って B まで移動中であっても、捕食者が接近してきた時には (1) の学習効果が発揮されなくてはならない。

5　シミュレーション実験

実際に S を用いて profit-sharing と L-ANA の比較評価を行なった。まず、profit-sharing と L-ANA でそれぞれ学習を行なった際の獲得された強化値の分布を図 5に示す。profit-sharing はエピソードしか強化しないので、強化値の分布に乱れが見られるが、L-ANA ではエピソード以外の状態まで活性伝幡を行なうことにより、なだらかな分布が形成されている。

図 6には、S の各状態がどの B までの移動経路 (エピソード) に含まれるたのか



Profit-sharing　　L-ANA

図 5: 強化値分布

を表示したものである。図 6左のように、 profit-sharing では B_1 の近傍であっても B_2 に向かう経路が学習されており効率が悪い。しかしながら図 6右のように L-ANA では B_1 の近傍では B_1 に向かう経路が学習されており、エネルギー補給量と補給ベースまでの移動距離を考慮した学習が行なわれていることが確認できる。

図 7には A がどれだけ最短経路で B まで移動できたか、即ち空腹状態になってから充電を完了するまでに移動した経路の長さと、 A が空腹状態になった地点から充電した地点までの最短距離 (マンハッタン距離) との比率を示したものである。この結果からもわかるように L-ANA は常に最短経路に近い経路で移動できているものの、 profit-sharing ではかなりのばらつきが見られる。

6　動的環境の変化に対して

L-ANA と profit-sharing の性能の差が最も顕著に現れるのは環境が動的に変化した時である。まず profit-sharing と L-ANA における動的環境の変化に対する対応の仕方を整理すると以下のようになる。動的変化としては、〔1〕それまでエピソードだった経路を分断するような位置への障害物の出現と、〔2〕それまで存在した障害物の消滅による新しい移動経路の出現が考えられる。

profit-sharing の場合　上記〔1〕の変化により、エピソードに従って移動を開始してから障害物に行き当たる状況が生じる。この時は行き当たるまでの状態遷移系列の強化値を初期化してこの移動経路に関する学習を無効にする。この作業を繰り返すことで新しい環境に適応する。学習が無効になるのはエピソード単位であり、生き残るエピソードも存在するという点で Q-learning よりもロバスト性がある。

L-ANA の場合　環境の変化を察知すると、まず変化が起こった部分の記憶エージェントが隣接する記憶エージェントに関する情報を更新する。その後、以前報酬を得

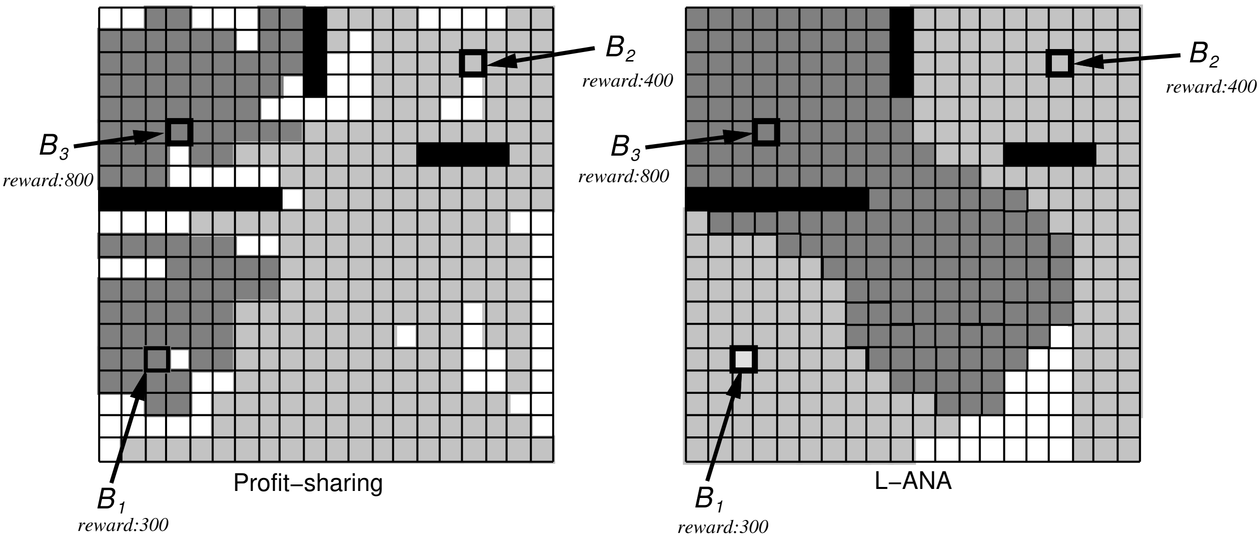


図 6: 学習された強化値の勢力分布

たことを記憶している記憶エージェントが、活性伝幡を開始して強化値の更新を順次行なうことで動的環境の変化に対応する。このしくみにより動的に出現した障害物に行き当たると、次に最適な移動経路が自動的に選択できるようになる。それでも新たな障害物に行き当たった場合には、再び活性伝幡を行ない次に最適な移動経路を選択しようとする。

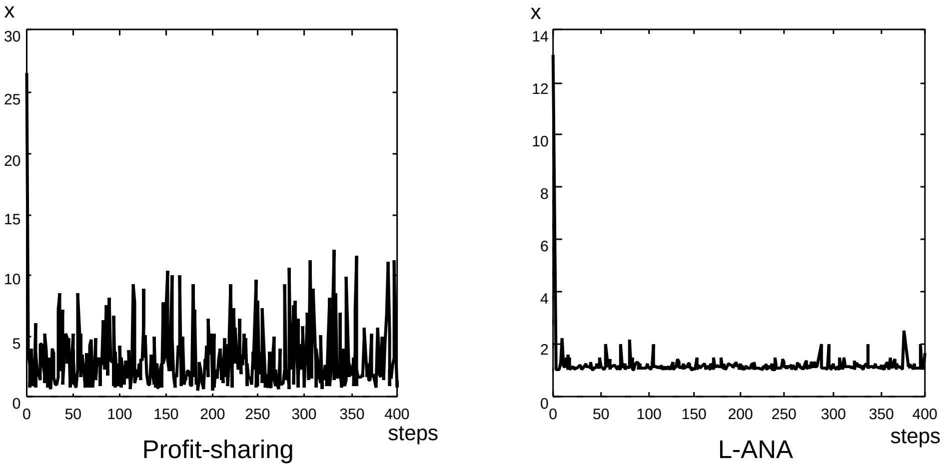
6.1 ロバスト性に関する評価結果

profit-sharing と L-ANA において学習中に環境を動的に変化させた時の A の挙動の変化を計測した結果を図 8に示す。図 7と同様に、 A がどれだけ最短経路で B まで移動できたかを示したものである。 250 ステップ目において環境の変化を意図的に起こした。

L-ANA では変化を察知すると活性伝幡により強化値の更新が行なわれるので、常に安定した経路を選択することができている。しかしながら profit-sharing では変化を察知したエピソードを無効にし、変化の影響を受けなかったエピソードへの経路、または新たな経路を学習するまでの間、学習の効率が下がっている。この結果から L-ANA が profit-sharing に比べてよりロ環境の動的変化に対してロバスト性があることを確認することができる。

7 議論

各記憶エージェントはそれぞれ自律していることから、仮にある記憶エージェントの機能が損なわれたとしても、その記憶エージェントを欠いた状態で活性伝幡が行なわれ、機能が損なわれた部分を迂回するような経路が自動的に選択される。profit-sharing ではエピソードを構成する状態の一つが欠けてしまうと、そのエピソード全体が影響を受けてしまう。このことからも L-ANA はよりロバスト性を有



Profit-sharing L-ANA

図 7: 移動経路の精度

し、実世界などの動的な環境内で動作する自律行動主体のための学習法として適している。

profit-sharing では強化値割り当ての際無効ルールを抑制することが問題となるが、 L-ANA では無効ルールも報酬を得る状態に至る経路として積極的に再利用される。

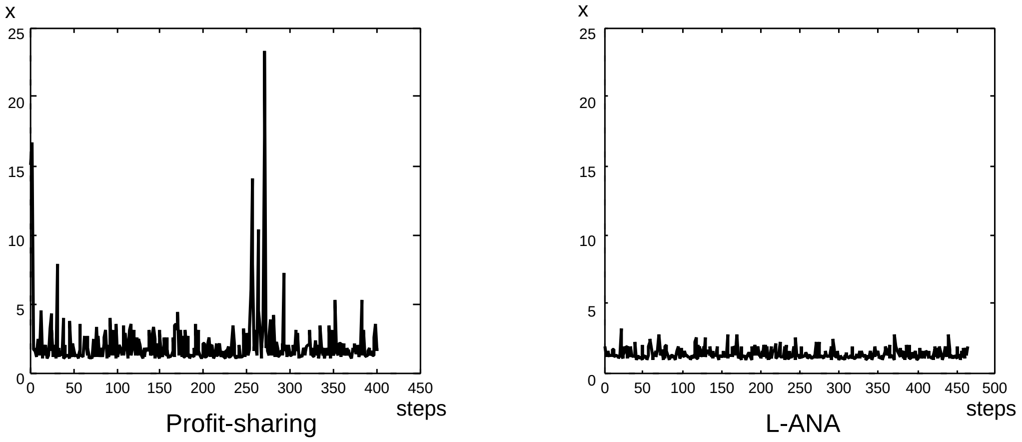
学習は A_{s_i} 単位で行なわれ、 A_{s_i} 単位で独自の活性伝播図形が学習されることになる。従って例えば A_{s_1} と A_{s_2} が成立している状況では両方の強化値分布を重ねた図形を用いて A は行動選択を行なう。行動選択の仕方は、例えば MRR-planning と同様の方法を用いることで容易に実現することができる。今後の課題としては、複数の強化値分布の重ね方に対する考察が必要である。

また A が複数存在し A 同士で協調する枠組を考えると、異なる行動主体の学習した強化値分布を共有することで互いの学習結果を利用し合うことが実現でき、この点に関しても今後研究を続ける予定である。

これまではすべて正の強化学習について述べてきたが、 L-ANA では活性値を吸収する逆の活性伝播を行なうことで、負の強化学習も容易に実現することができる。 A は常により強化値の大きい状態に移動することで学習結果を利用するわけだが、ここで S 内に落し穴の設定を加えてみる。この場合落し穴に近付かないようにするには、落し穴を中心として活性値を吸収する活性伝播を行なえば良い。そしてエネルギーを充電するための正の強化学習のための強化値分布と重ねることで両方の学習効果を容易に統合することができ、落し穴を避けつつ最適に充電ポイントに至る経路を選択することができる。

8　まとめ

行動選択ネットワークを経験強化型の強化学習に適応した L-ANA アルゴリズムを提案した。従来の強化学習法である profit-sharing に比べよりロバスト性を有し、実世界やインターネットなど複雑で動的に変化する環境で動作する自律移動ロボットやソフトウェアエージェントなどの自律行動主体に適した学習アルゴリズム



Profit-sharing

L-ANA

図 8: 動的変化に対する移動経路の精度の変化

である。活性伝幡の特性を制御することで学習の特性を容易に操作することができ、従来の実時間リアクティブプランニングなどの組合せることも容易である。今回はシミュレーションによりその基本的有効性の確認を行なった。

参考文献

[1] C. J. C. Watkins and P. Dayan: Technical Note: Q-Learning, *Machine Learning*, Vol. 8, pp. 55 – 68 (1992).

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

[3] M. Minsky: The Society of Mind, Simon & Schuster, 1986.

[4] J. J. Grefenstette: Credit Assignment in Rule Discovery Systems Based on Generic Algorithms, *Machine Learning*, Vol.3, pp. 225 – 245, 1988.

[5] J. H. Holland, and J. S. Reightman: Cognitive Systems Based on Adaptive Algorithms, *Pattern-Directed Inference System*, D. A. Waterman, and F. Hayes-Roth (eds.), Academic Press, 1978.

[6] P. Maes: The Agent Network Architecture (ANA), *SIGART Bulletin*, Vol.2, No.4, pp. 115 – 120, 1991.

[7] S. Kurihara and R. Onai: Adaptive Selection of Reactive/Deliberate Planning for the Dynamic Environment, *Lecture Notes in Artificial Intelligence 1237, Proceedings of the MAAMAW'97*, 1997.