

重点サンプリングを用いたGAによる強化学習

Reinforcement learning by GA using Importance Sampling

土谷 千加夫
Chikao Tsuchiya

東京工業大学大学院総合理工学研究科
Graduate School of Interdisciplinary Science and Engineering, Tokyo Institute of Technology
tsuchiya@fe.dis.titech.ac.jp

木村 元
Hajime Kimura

九州大学大学院工学研究院海洋システム工学部門
Department of Marine Engineering, Graduate School of Engineering, Kyushu University
kimura@nams.kyushu-u.ac.jp

佐久間 淳
Jun Sakuma

東京工業大学大学院総合理工学研究科
Graduate School of Interdisciplinary Science and Engineering, Tokyo Institute of Technology
jun@fe.dis.titech.ac.jp

小林 重信
Shigenobu Kobayashi

(同上)
kobayasi@dis.titech.ac.jp

keywords: GA, reinforcement learning, direct policy search, importance sampling

Summary

Reinforcement Learning (RL) handles policy search problems: searching a mapping from state space to action space. However RL is based on gradient methods and as such, cannot deal with problems with multimodal landscape. In contrast, though Genetic Algorithm (GA) is promising to deal with them, it seems to be unsuitable for policy search problems from the viewpoint of the cost of evaluation. Minimal Generation Gap (MGG), used as a generation-alternation model in GA, generates many offspring from two or more parents selected from a population. Therefore, evaluating policies of generated offspring requires much trial and error (i.e. interaction between an agent and an environment). In this paper, we incorporate importance sampling into the framework of MGG in order to reduce the cost of evaluation on policy search. The proposed techniques are applied to Markov Decision Process (MDP) with multimodal landscape. The experimental results show that these techniques can reduce the number of interaction between an agent and an environment, and also mean that MGG and importance sampling are good for each other.

1. はじめに

強化学習の目的は環境とのインタラクションを通じて平均報酬を最大化する政策を学習することである。ここで、政策 (policy) とは状態から行動への写像をいう。強化学習の多くは基本的には局所探索に基づいているため、多峰性の景観を持つ政策に適用した際に、局所解に陥る危険性がある。

一方、GA は集団で解を探索する多点探索法であることから、多峰性の景観を持つ問題に対応できる利点を持つ。しかし、政策探索問題では、探索の対象となる政策の評価値はその政策下でのエージェントのパフォーマンスとして決定されるため、値が一定とはならない場合が多い。さらに、一定期間のパフォーマンスを平均化するために、環境との膨大なインタラクションが必要とされる。これらの困難を伴うために、GA は政策探索に向かないとされてきた [Tsuchiya 04]。

最近、強化学習の分野で、重点サンプリング (Importance Sampling) [Rubinstein 81] の導入により、ある政

策の下で生成された状態遷移系列を再利用する研究が注目されている [Shelton 01a, Precup 00, Precup 01]。すなわち、重点サンプリングを用いて状態行動対の評価値を推定するもので、これにより環境とのインタラクションを削減する道が拓かれた。しかし、REINFORCE アルゴリズム [Williams 92] に代表される政策勾配法 (policy gradient method) を用いる限り、局所解に陥るという問題を依然抱えている。

通常の GA による政策探索では、個体が政策を表すので、生成子個体のすべてに対して環境とのインタラクションを必要とする。しかし、集団を形成する各個体のインタラクションの履歴を保持しておけば、重点サンプリングの導入により、環境とインタラクションすることなしに生成子個体の評価値の推定を行なうことができる。これにより、大域的最適政策を効率良く発見することが期待できる。

本論文では、GA に重点サンプリングを導入した2つの接近法を提案する。第一の接近法として、重点サンプリングを用いて GA の親集団の経験を再利用することに

より子個体の政策評価を行なう方法を提案する。これにより環境とのインタラクションを大幅に削減することが期待できる。第二の接近法として、子個体の生成に交叉と政策勾配を用いた局所探索を併用することにより、多峰性へ対応するとともに学習を高速化する方法を提案する。

第一の方法では、エージェントは複数の政策を集団として持ち、それが環境中を動くことによって経験を蓄積する。一定期間経過後に、集団からランダムに選択された個体を親として子個体を複数生成する。蓄積された経験を重点サンプリングで処理することにより、異なる政策で獲得された報酬を子個体が獲得すると期待される報酬に変換し、それを子個体の評価値とする。この際、子個体の政策は環境中で試される必要がないので、子個体評価にかかる時間は重点サンプリングの処理時間だけとなり、最適化プロセス自体が高速化される。

第二の方法では、エージェントは複数の政策によって経験を蓄積する。交叉に加えて重点サンプリングで推定した政策勾配を用いた勾配法により子個体を生成する。勾配情報を積極的に用いることで少ない子個体でも効率的な探索が可能となる。第一の方法と同様に、このプロセスは数値計算のみで処理できるので、環境とのインタラクションは大幅に削減される。

本論文は全 6 章からなる。2 章では、強化学習および従来の GA による政策探索と重点サンプリングを説明する。3 章では対象問題を述べた後、接近法を具体的に論じる。4 章では、実験により提案手法と従来手法との比較を行う。5 章では提案手法の特徴と課題について考察する。6 章は本論文のまとめである。

2. 準備

ここでは、3 章以下の準備として、政策探索、強化学習手法による政策探索と GA による政策探索、および重点サンプリングについて説明する。

2.1 政策探索

環境におかれた学習主体をエージェントと呼ぶ。エージェントは試行錯誤を通じて環境に適応する。エージェントは環境と以下のやり取りを繰り返す。

- (1) エージェントは時刻 t において環境の状態観測 o_t に基づいて意思決定を行い、行動 a_t を出力する
- (2) エージェントの行動により、環境は状態遷移し、その遷移に応じた報酬 r_t をエージェントに与える
- (3) 時刻 t を $t+1$ に進めてステップ 1 へ戻る

エージェントは利得 (例えば、報酬の総和) の最大化を目指して、状態観測から行動出力への写像を学習する。この写像を政策と呼ぶ。本論文では、確率的に行動を選択する確率的政策を扱う。

一般に、政策はパラメータ θ によって記述することができる。従って、最適な政策を獲得する問題はパラメータ

θ を最適化する問題に帰着される。ここでは、パラメータを直接的に探索する枠組みを直接的政策探索 (direct policy search) と呼ぶ。

2.2 政策勾配を用いた政策探索

強化学習手法の多くは、状態価値や政策価値などの勾配情報を用いて局所探索を行っている。

本論文ではマルコフ決定過程 (Markov Decision Process: MDP) を扱う。強化学習手法として、単純に観測入力に対する行動出力確率の関数を最適化する確率的傾斜法 (Stochastic Gradient Ascent: SGA) が提案されている [Kimura 98]。SGA は勾配情報を用いるので、高次元の政策を扱う場合でも、高速な最適化が期待できる。また、確率的な行動選択を行うので、ある程度の多峰性に対処できる。しかし、基本的には局所探索法であるため、問題が複雑化し政策空間の多峰性が顕著になると、大域的最適解の発見は困難になることが予想される。

2.3 GA による政策探索

GA は最適化対象の関数の勾配情報を用いない直接探索法のひとつである。最適化対象の関数形にも依存しないために、適用範囲が広く、数多くの実問題に適用されている。

本論文では、政策パラメータの探索に実数値 GA を用いる。すなわち、多様性維持に優れた世代交代モデル MGG (Minimal Generation Gap) [Sato 96] と単峰性正規分布交叉 UNDX [Ono 97] の組み合わせを採用する。GA による政策探索では政策パラメータをひとつの染色体として扱う。GA では交叉によって生成された子個体の政策を用いて環境と実際にインタラクションすることにより評価値を得る。

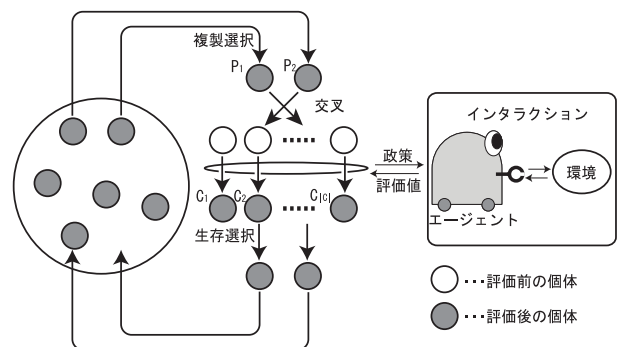


図 1 GA(MGG+UNDX) による強化学習の一般的枠組み

GA は多峰性に対処できる [Ono 97] ことから、景観が複雑な問題においても大域的最適政策を発見できることが期待できる。しかし、上述したように生成子個体の評価に際して、環境とのインタラクションが必要であり、このことが GA による政策探索の大きな障害になっている。

2.4 重点サンプリング

重点サンプリングとは乱数を用いて積分値を近似するモンテカルロ積分の分散を減少させる統計手法として発展したものであり、近年異なる分布の期待値を推定する手法として強化学習における利用可能性が注目されている。パラメータ θ で記述された政策 (挙動政策, behavior policy) とその政策から得られたデータがあるとすると、別のパラメータ θ' で記述された政策 (推定政策, estimation policy) の評価値を推定することができる。

パラメータ θ で記述される政策 π において、エージェントが状態 s を観測し、行動 a を選択する確率を $\pi(s, a; \theta)$ と表すと、この政策の下で長さ M のあるエピソード $h = \{s_1, a_1, r_1, s_2, a_2, r_2, \dots, s_M, a_M, r_M, s_{M+1}\}$ が得られる確率は次式で表される：

$$\begin{aligned} \Pr(h|\theta) &= \Pr(s_1)\Phi(h)\Psi(h) \\ &= \Pr(s_1) \prod_{i=1}^M \pi(s_i, a_i; \theta) \Pr(s_{i+1}|s_i, a_i) \end{aligned} \quad (1)$$

ここで、 $\Phi(h)$ は行動選択確率 (政策依存)、 $\Psi(h)$ は状態遷移確率 (政策非依存)、 $\Pr(s_1)$ は初期状態が s_1 となる確率である。

重点サンプリングの原理を用いると、推定政策の評価値 $\hat{V}(\theta')$ は次のように推定できる [Precup 00]：

$$\begin{aligned} \hat{V}(\theta') &= \frac{1}{N} \sum_{i=1}^N R_i \frac{\Pr(h_i|\theta')}{\Pr(h_i|\theta)} \\ &= \frac{1}{N} \sum_{i=1}^N R_i \frac{\Pr(s_1)\Phi'(h_i)\Psi(h)}{\Pr(s_1)\Phi(h_i)\Psi(h)} \\ &= \frac{1}{N} \sum_{i=1}^N R_i \frac{\Phi'(h_i)}{\Phi(h_i)} \\ &= \frac{1}{N} \sum_{i=1}^N R_i \prod_{j=1}^M \frac{\pi(s_j, a_j|\theta')}{\pi(s_j, a_j|\theta)} \end{aligned} \quad (2)$$

ここで、 R_i はエピソード i の獲得報酬の合計、 N はエピソード数である。式 (2) は、評価値の推定値が各政策での行動選択確率の尤度比だけで計算できることを意味している。

式 (2) はサンプル数が無限大の極限で真の評価値からの偏りが無いが、限られたサンプル数では分散が比較的大きくなることが指摘されている [Precup 00]。これに対して、サンプル数が無限大の極限であっても真の評価値と一致しない (biased) が、限られたサンプル数では式 (2) よりも分散が小さく (low variance) なる重み付き重点サンプリング (Weighted Importance Sampling) が提案されている [Precup 00]：

$$\hat{V}(\theta') = \frac{\sum_{i=1}^N R_i \frac{\Pr(h_i|\theta')}{\Pr(h_i|\theta)}}{\sum_{i=1}^N \frac{\Pr(h_i|\theta')}{\Pr(h_i|\theta)}}$$

$$\begin{aligned} &= \frac{\sum_{i=1}^N R_i \frac{\Phi'(h_i)}{\Phi(h_i)}}{\sum_{i=1}^N \frac{\Phi'(h_i)}{\Phi(h_i)}} \\ &= \frac{\sum_{i=1}^N R_i \prod_{j=1}^M \frac{\pi(s_j, a_j|\theta')}{\pi(s_j, a_j|\theta)}}{\sum_{i=1}^N \prod_{j=1}^M \frac{\pi(s_j, a_j|\theta')}{\pi(s_j, a_j|\theta)}} \end{aligned} \quad (3)$$

エージェントは挙動政策の下での経験 $\{s, a, r\}$ を蓄積しておき、ある時点でその経験データを重点サンプリングで処理すれば、推定政策の評価値を次々に推定することができる。この処理は、環境とのインタラクションを必要としない。

強化学習の分野では、重点サンプリングを Off-Policy 学習に利用する研究がなされている [Peshkin 02, 内部 03]。最近では、この性質を利用して、複数タスクの同時学習の研究がなされている [Kimura 04]。

3. 問題設定と接近法

3.1 問題設定

本論文では、離散状態・離散行動の MDP を扱う。MDP は次のように示される。 $\mathcal{S}$ を状態空間、 $\mathcal{A}$ を行動空間、 $\mathcal{R}$ を報酬値の集合とする。各離散時間 t において、エージェントは状態 $s_t \in \mathcal{S}$ を観測し、行動 $a_t \in \mathcal{A}$ を実行し、環境の状態遷移の結果として即時報酬 $r_t \in \mathcal{R}$ を受け取る。報酬と次状態の確率分布 a_t, s_t のみに依存すると仮定する。次状態 s_{t+1} は状態遷移確率 $\Pr(s_{t+1}|s_t, a_t)$ に従って選択される。

エージェントは事前に $\Pr(s_{t+1}|s_t, a_t)$ と $r(s_t, a)$ を知らない。この状況で、エージェントのパフォーマンスを最大化する政策を学習することが目的である。自然なパフォーマンスの測定方法は割引期待報酬またはエピソードでの平均獲得報酬である。

3.2 重点サンプリングによる子個体の評価

上述したように、GA による政策探索は多峰性を持つ複雑な政策空間においても大域的最適政策を発見できることが期待できる反面、生成子個体の評価に膨大な時間が必要となるという問題点がある。そこで、本論文では、生成子個体の評価を高速に行うことを主目標におく。GA は集団で解を探索するという性質上、評価の高速化の方法として、集団の複数の政策下での経験を利用して子個体の評価値を推定するのが妥当と思われる。

§1 重点サンプリングによる政策価値の推定

式 (2)、(3) に示される重点サンプリングはエピソード h_i がそれを生成した政策以外の政策からも生成されうることを利用して評価値推定を行なうので、限られたサンプルでも安定した評価ができる。上述したように、式 (2) は unbiased であるが、限られたサンプル数では分散が比較的大きくなる。これに対して、式 (3) は biased ではあるが、限られたサンプル数では式 (2) より low variance である。

一方、複数の挙動政策 $\theta_1, \theta_2, \dots, \theta_N$ の下でサンプルされたエピソードは、それらの挙動政策の混合分布 (式 (4)) からサンプルされたと考えることもできる。挙動政策を混合分布と捉えれば、式 (5) に示される importance sampling [Shelton 01a] になる。これは [Shelton 01b] により、式 (3) よりも分散が小さいことが指摘されており、環境とのインタラクションを削減し限られたサンプル数で推定政策の評価値推定を行おうとする本論文の趣旨と合致する。したがって、本論文では挙動政策を混合分布と捉える式 (5) を評価値推定に用いることにする。

$$\frac{1}{N} \sum_{j=1}^N \Pr(h|\theta_j) \quad (4)$$

$$\begin{aligned} \hat{V}(\theta) &= \frac{\sum_{i=1}^N R_i \frac{\Pr(h_i|\theta)}{\frac{1}{N} \sum_{j=1}^N \Pr(h_i|\theta_j)}}{\sum_{i=1}^N \frac{\Pr(h_i|\theta)}{\frac{1}{N} \sum_{j=1}^N \Pr(h_i|\theta_j)}} \\ &= \frac{\sum_{i=1}^N R_i \frac{\Pr(h_i|\theta)}{\sum_{j=1}^N \Pr(h_i|\theta_j)}}{\sum_{i=1}^N \frac{\Pr(h_i|\theta)}{\sum_{j=1}^N \Pr(h_i|\theta_j)}} \\ &= \frac{\sum_{i=1}^N R_i \frac{\prod_{k=1}^M \pi(o(s_k), a_k|\theta)}{\sum_{j=1}^N \prod_{k=1}^M \pi(o(s_k), a_k|\theta_j)}}{\sum_{i=1}^N \frac{\prod_{k=1}^M \pi(o(s_k), a_k|\theta)}{\sum_{j=1}^N \prod_{k=1}^M \pi(o(s_k), a_k|\theta_j)}} \quad (5) \end{aligned}$$

式 (5) を用いると、集団の政策によってサンプリングされた経験データを保持しておけば、そこから生成子個体の評価値を推定できる。なお、評価値の推定精度については、後述する理由から、それほど高いものは要求されない。

§2 アルゴリズム：PVIS

世代交代モデル MGG に重点サンプリングによる子個体の評価値推定を組み込むことを考える。そのために、集団の政策によって得られた経験 $\{s, a, r\}$ をデータベースに保持しておく。一定期間毎に集団中から選択された親個体から交叉 UNDX によって多数の子個体を生成し、それらの評価値を式 (5) で推定する。推定にはデータベースに蓄積された経験を用いる。その推定値を用いてランクベースのルーレット選択 [Goldberg 89] を行ない、選択された個体を集団に戻す。ランクベースのルーレット選択では、評価値は序数的にしか利用されない。よって、式 (5) の精度は順序を保存する程度であればよい。

新しく集団に加わる個体は環境中での経験を持たないので、ここで環境とのインタラクションが実行される。新たに獲得された経験はデータベース中のそれらの親個体による経験と置換される。この提案アルゴリズムを PVIS (Policy Value estimated by Importance Sampling) と呼ぶ。図 2 に PVIS のアルゴリズムの詳細を、図 3 にアルゴリズムの概念的枠組みをそれぞれ示す。

- (1) 初期集団として N 個の政策を生成し、それらすべての政策の下で環境とのインタラクションを通じて経験を獲得する
- (2) 集団中から親個体 $\{P_1, P_2\}$ をランダムに選び、交叉によって C 個の子個体 $\{C_1, \dots, C_{|C|}\}$ を生成する。ここで、 $\{P_1, P_2, C_1, \dots, C_{|C|}\}$ を家族とする
- (3) 家族の各個体の価値を重点サンプリングで推定する
- (4) 家族中から 2 個体を選ぶ：ひとつは最良個体であり、もうひとつはランクベースルーレット選択で選ばれた個体である。そして、これら 2 個体を親 2 個体と置き換える
- (5) 親個体の政策の下で獲得された経験を破棄する
- (6) 新しく加えられた 2 個体の政策下で、実際に環境とインタラクションすることで経験を獲得する
- (7) 上記の手順をステップ 2 から繰り返す

図 2 PVIS のアルゴリズム

重点サンプリングでは推定政策と挙動政策が政策空間において近くなければ十分な精度で推定することが期待できない。提案手法では、探索序盤は集団が政策空間全体に広がっているため、推定精度はあまり良くないが、探索中盤から終盤になると集団が有望な領域に集まってくるため、十分な推定精度が確保されると見込まれる。

3.3 政策勾配を用いた局所探索による子個体の改善

MGG では交叉によって多数の子個体を生成する。これは親個体の周辺を局所探索していることを意味する。対象とする政策の次元が大きくなるにつれて、探索性能を維持するために必要な生成子個体数は増やさなければならない。上述の PVIS は子個体評価を数値計算だけで行えるので、多数の子個体を生成しても、環境とのインタラクションは増加しないが、計算量の増加は深刻である。そのような問題を回避するために、交叉による多数の子個体生成の代わりに推定された政策勾配に基づいた局所探索を導入する。これを PGIS (Policy Gradient estimated by Importance Sampling) と呼ぶ。これによって、たとえ政策空間が高次元であっても、少ない計算量で的確に解を改善できることが期待できる。また、生成子個体数を抑制しても探索の高速化が期待できる。

§1 重点サンプリングによる政策勾配の推定

パラメータ θ で記述される政策の価値は式 (5) で表されるので、式 (5) を k 番目のパラメータ $\theta^{(k)}$ で微分する

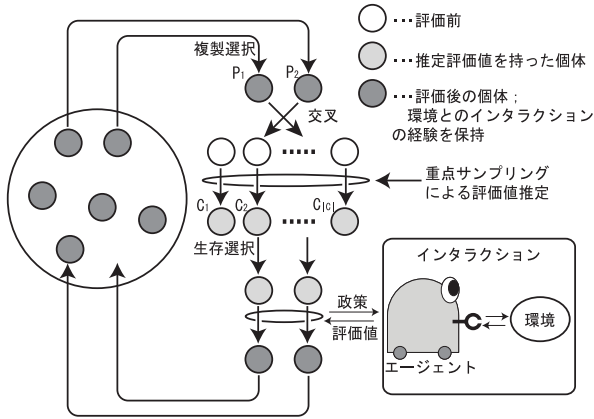


図3 PVISによる強化学習の枠組み

ことで推定政策の勾配を以下のように計算できる：

$$\frac{\partial}{\partial \theta^{(k)}} \hat{V}(\theta) = \frac{\sum_{i=1}^N (R_i - \hat{V}) \frac{\frac{\partial}{\partial \theta^{(k)}} \Pr(h_i | \theta)}{\Pr(h_i | \theta_j)}}{\sum_{i=1}^N \frac{\Pr(h_i | \theta)}{\sum_{j=1}^N \Pr(h_i | \theta_j)}} \quad (6)$$

政策価値の最大化問題の場合、政策を式(6)の方向へ更新すればよい。式(6)の妥当性は式(5)の評価値推定の妥当性に依存する。推定される勾配の方向ベクトルが真の勾配の方向ベクトルと一致する場合に最大の改善がなされるが、それらが一致しなくても内積が正である限り勾配法による改善が見込めるので、十分利用可能と判断される。

§2 アルゴリズム：PGIS

MGGの枠組みにPGISを組み込むことを考える。最初に、集団中から親個体を選択する。次に、選択された親個体および親個体から交叉UNDXによって生成された子個体を式(6)で推定される政策勾配を用いた勾配法で改善する。最後に、それらを家族として生存選択を行う。図4にPGISのアルゴリズムの詳細を、図5にアルゴリズムの概念的枠組みを示す。

勾配法による子個体の改善は集団の経験だけを用いるため、それらは並列処理が可能である。したがって、アルゴリズムを4並列で実装することにより4倍の高速化が可能である。

3.4 提案手法の特徴

SGAおよびGAと比較することにより、提案手法PVISおよびPGISの特徴をまとめる。

GAを政策探索に適用する利点は、GAの持つ潜在的な能力、つまり、多峰性に対するロバスト性を引き出すことである。探索に十分な時間をかければ、得られる最終性能は強化学習手法をも凌ぐ。だが、生成子個体ひとつひとつを環境に適用すること、および集団で並列に探索することが、学習速度の低下を招き、探索効率は良くない。

- (1) 初期集団として N 個の政策を生成し、それらすべての政策の下で環境とのインタラクションを通じて経験を獲得する
- (2) 集団中から親個体 $\{P_1, P_2\}$ をランダムに選び、交叉によって2つの子個体 $\{C_1, C_2\}$ を生成する。
- (3) 重点サンプリングで推定される政策勾配を用いた勾配法によって $\{P_1, P_2\}$ を改善する。これらを $\{C_3, C_4\}$ とする
- (4) 同様に $\{C_1, C_2\}$ を改善し $\{C'_1, C'_2\}$ を生成し、 $\{P_1, P_2, C'_1, C'_2, C_3, C_4\}$ を家族とする
- (5) 家族の各個体の価値を重点サンプリングで推定する
- (6) 家族中から2個体を選ぶ：ひとつは最良個体であり、もうひとつはランクベースルーレット選択で選ばれた個体である。そして、これら2個体を親2個体と置き換える
- (7) 親個体の政策の下で獲得された経験を破棄する
- (8) 新しく加えられた2個体の政策下で、実際に環境とインタラクションすることで経験を獲得する
- (9) 上記の手順をステップ2から繰り返す

図4 PGISのアルゴリズム

PVISは、生成子個体ひとつひとつを環境に適用しなければならないGAの欠点を補うために、重点サンプリングによる推定をもって子個体の評価とする。実際に環境に適用されるのは、2個体の政策だけとなり、単純なGAに比べて学習速度は(生成子個体数)/2倍になる。しかし、交叉による局所探索には多数の子個体を生成することが求められ、高次元政策空間ではそのための計算量が無視できなくなるという欠点がある。

一方、PGISは限られた生成子個体数で効率的に局所探索を行なう目的で提案されたものである。PGISは強化学習手法で使われている政策勾配を重点サンプリングにより推定し、局所探索に勾配法を適用する。高次元政策空間においては、PVISに比べてさらなる高速化が期待できる。

4. 実験

提案手法の有用性を検討するために、多峰性の景観を有する問題をベンチマークとして既存の他手法との比較を行なう。

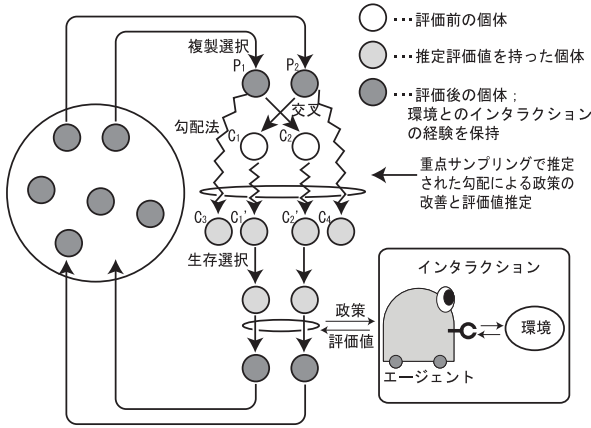


図 5 PGIS による強化学習の枠組み

4.1 2 分木問題

図 6 や図 7 に示される 2 分木状の単純なエピソードタスクを考える. このような問題を 2 分木問題と呼ぶ. 2 分木問題は騙しの強い局所解を有する強化学習問題を単純化した問題となっている. 図 6 は深さ 2 の 2 分木問題である. 初期状態は 2 分木の根ノードにあたる状態 s_0 であり, 行動 a_L で左, a_R で右の状態へ移動し, 葉ノードにあたる状態に遷移したときに報酬を得て終了する. 一般に, 深さ n の 2 分木問題は図 7 のように表される. 深さ n の問題では, 状態数は $N_S = 2^n - 1$ となる. 報酬は状態 s_1 以下の部分木では $2/3$, 状態 s_2 以下の部分木では最右の状態遷移のみ 1 , それ以外は 0 である. 学習初期の自然な行動選択規則は a_L, a_R を等確率で選択するというものである. この規則の下では, 深さ n の問題における s_1 以下の部分木での期待報酬は $2/3$ となり, s_2 以下の部分木での期待報酬は $1/2^{n-1}$ となる. そのため, greedy な学習を行なう強化学習手法では状態 s_1 以下の部分木に進むほうが, 状態 s_2 以下の部分木に進むよりも良く見えてしまい, 状態 s_2 以下の部分木の探索が困難になる. このことは, 深さの増加に伴い顕著になる. このように 2 分木問題は小さな状態空間にも関わらず, 騙しの強い局所解を有するために, 単純な勾配法では大域的最適政策の学習が困難な問題となっている. なお, 著者らは [Tsuchiya 04] において, 多峰性に起因して SGA では局所解に陥りやすい匍匐ロボットに対して提案手法の適用を試みており, 大域的に最適政策を安定して獲得できることを示している. 2 分木問題は, 匍匐ロボットでの経験を踏まえて, 強化学習における多峰性問題のベンチマークのひとつとして今回考案したものである. この実験では, 深さ 2, 4, 5 の問題を扱う. 状態数 (探索次元数) は, それぞれ 3, 15, 31 となる.

4.2 手法の設定

提案手法を GA, SGA と比較する. 全手法で共通して用いられる政策は図 8 に示す 2 層ニューラルネットでは表現

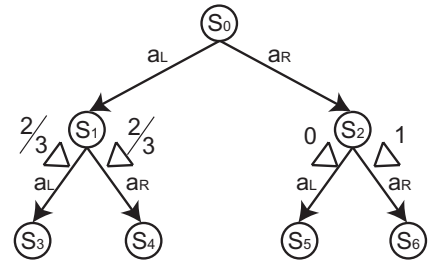
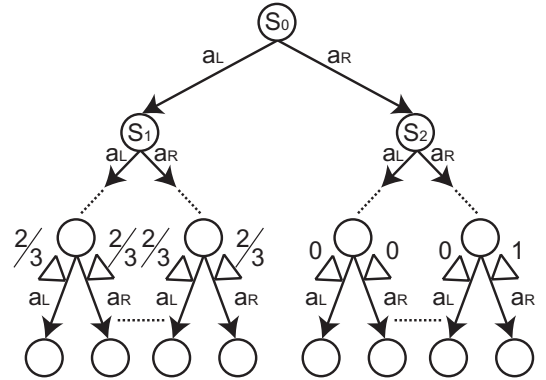


図 6 深さ 2 の 2 分木問題

図 7 深さ n の 2 分木問題

される. 状態 s_i で特徴量ベクトル $\mathbf{x} = (0, \dots, 0, \overset{i}{1}, 0, \dots, 0)$ を入力し, 政策パラメータである重みベクトル θ を用いて行動 a_L, a_R の選択確率を次のように決定し, 行動を出力する.

$$\pi(a_L, \mathbf{x}, \theta) = \frac{1}{1 + \exp(-\sum_{i=1}^{N_S} \theta_i x_i)} \quad (7)$$

$$\begin{aligned} \pi(a_R, \mathbf{x}, \theta) &= 1 - \pi(a_L, \mathbf{x}, \theta) \\ &= \frac{\exp(-\sum_{i=1}^{N_S} \theta_i x_i)}{1 + \exp(-\sum_{i=1}^{N_S} \theta_i x_i)} \end{aligned} \quad (8)$$

重点サンプリングで用いる式 (6) 中のエピソードの生起確率の微分 $\frac{\partial}{\partial \theta^{(k)}} \Pr(h_i | \theta)$ は次のように計算できる:

$$\begin{aligned} \frac{\partial}{\partial \theta^{(k)}} \Pr(h_i | \theta) &= \Pr(h_i | \theta) \frac{\partial}{\partial \theta^{(k)}} \ln \Pr(h_i | \theta) \\ &= \Pr(h_i | \theta) \frac{\partial}{\partial \theta^{(k)}} \ln (\pi(s_1, a_1, \theta) \cdots \pi(s_M, a_M, \theta)) \\ &= \Pr(h_i | \theta) \frac{\partial}{\partial \theta^{(k)}} (\ln \pi(s_1, a_1, \theta) + \cdots + \ln \pi(s_M, a_M, \theta)) \\ &= \Pr(h_i | \theta) \left(\frac{\partial}{\partial \theta^{(k)}} \ln \pi(s_1, a_1, \theta) + \cdots + \frac{\partial}{\partial \theta^{(k)}} \ln \pi(s_M, a_M, \theta) \right) \\ &= \Pr(h_i | \theta) \left(\sum_{m=1}^M \frac{\partial}{\partial \theta^{(k)}} \ln \pi(s_m, a_m, \theta) \right) \end{aligned} \quad (9)$$

ここで,

$$\begin{aligned}
 & i = j \text{ のとき} \\
 & \quad \frac{\partial}{\partial \theta(j)} \ln \pi(s_i, a_L, \theta) = 1 - \pi(s_i, a_L, \theta) \\
 & i \neq j \text{ のとき} \\
 & \quad \frac{\partial}{\partial \theta(j)} \ln \pi(s_i, a_L, \theta) = 0 \\
 & i = j \text{ のとき} \\
 & \quad \frac{\partial}{\partial \theta(j)} \ln \pi(s_i, a_R, \theta) = \pi(s_i, a_R, \theta) - 1 \\
 & i \neq j \text{ のとき} \\
 & \quad \frac{\partial}{\partial \theta(j)} \ln \pi(s_i, a_R, \theta) = 0
 \end{aligned} \tag{10}$$

よって, 経験データ中の各行動の選択確率の微分を足し合わせることでエピソードの生起確率の微分を計算できる.

また, PGIS の子個体生成における局所探索は, 式 (6) で示される勾配方向へ黄金分割法による直線探索を適用することで行なう. 直線探索は始点から距離 10(深さ 2), 20(深さ 4,5) の区間で行なう.

PVIS, PGIS, GA における政策の評価期間であるエピソード長は問題の深さと等しく設定した. つまり, 初期状態 s_0 から最下段の状態に達するまでの獲得報酬を政策の評価とする. 集団サイズは 30, PVIS と GA における生成子個体数は 10 とした. SGA と GA のその他の細かい設定はそれぞれ文献 [Kimura 98], [Kita 98] に基づく.

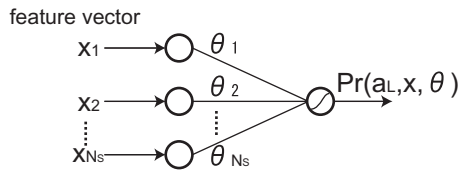


図 8 2 分木問題の政策表現

4.3 実験の結果

各手法の学習曲線を図 9, 図 10, 図 11 に示す. 学習速度はエージェントと環境とのインタラクション数によって規定されるとの考えから横軸は学習ステップ数 (環境とのインタラクション数) とした. 縦軸は獲得報酬の 100 試行平均である. PVIS と PGIS は初期集団が環境とインタラクションして経験を蓄積するまでに深さ 2 で $MN = 60$ ステップ, 深さ 4 で $MN = 120$ ステップ, 深さ 5 で $MN = 150$ ステップを要するため, 経験の蓄積が完了するまでは性能の改善がない同一の学習曲線となる. その後は, PGIS と PVIS が急激に最適政策に収束し, SGA と GA に対して有意に高速である. GA は政策の評価の分散が大きいため, エリート保存に失敗することが確認された. そのため, 集団が最適政策に完全に収束しきれず, 深さ 2 の場合ではさらに学習を続けても最終性能は約 0.96 にしか達しない. 深さ 4, 5 の問題では, 探索空間の次元の増加に伴い, PVIS と PGIS においても, 獲得報酬 2/3 の局所解に陥る場合が観察された.

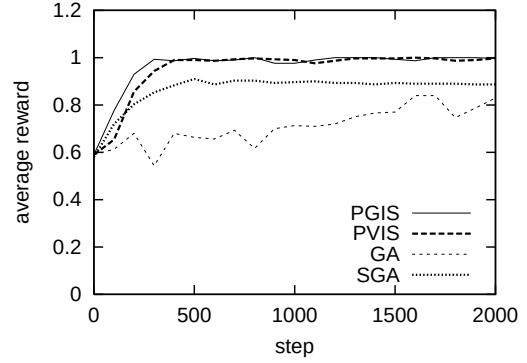


図 9 深さ 2 の 2 分木問題の学習曲線 (100 試行の平均)

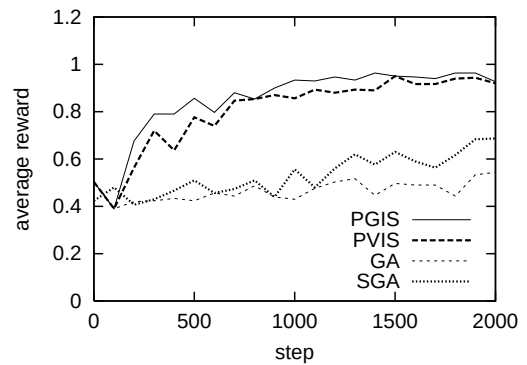


図 10 深さ 4 の 2 分木問題の学習曲線 (100 試行の平均)

ところで, 重点サンプリングの推定精度は PVIS と PGIS の探索能力に大きく影響する. そこで, 解析的に求めた期待報酬と重点サンプリングによって推定された期待報酬を比較する. ここで, 解析のしやすさのために, 規模の小さい深さ 2 の 2 分木問題を考える. 図 12 は最適政策に関係しないパラメータ θ_1 を 0 に固定したときの範囲 $[-5, 5]$ の (a) 真の期待報酬と (b) 重点サンプリングによって 30 個の経験データから推定された期待報酬を示したものである. (a) と (b) の間には有意な差は見られず, この問題における重点サンプリングの推定精度は十分に高いと判断される.

2 分木問題は MDP であることから, 参考のために, MDP における最適政策への収束が保証される Q-learning を同じ問題に適用した. Q-learning は, 全状態行動対の価値を保持する Q-table を構築し, 確率 $\epsilon = 0.05$ でランダム行動, 確率 $1 - \epsilon$ で最大価値の行動を選択する ϵ -greedy 行動選択規則を用いた.

図 13 は深さ 2, 4, 5 の 2 分木問題での 2000 ステップまでの Q-learning の学習曲線である. 図に示す範囲内では, Q-learning は最適政策に収束していない. ϵ を大きくすれば, 局所解に陥りにくくなるが, 適切なスケジューリングを行なわないと, 最終性能が悪化するという問題点がある.

表 1 は 2000 ステップまでに各手法によって 100 試行

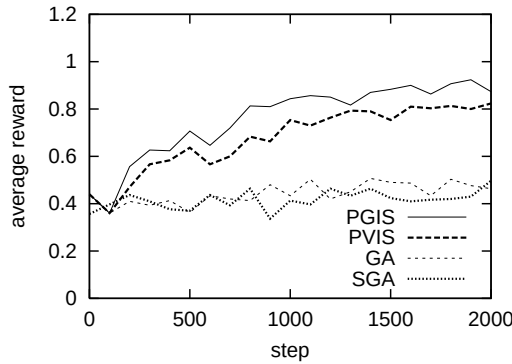


図 11 深さ 5 の 2 分木問題の学習曲線 (100 試行の平均)

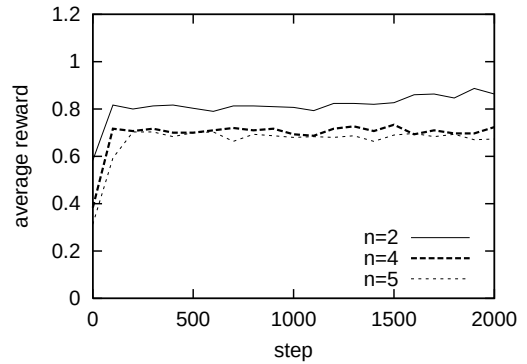


図 13 深さ 2.4,5 の 2 分木問題における Q-learning の学習曲線 (100 試行の平均)

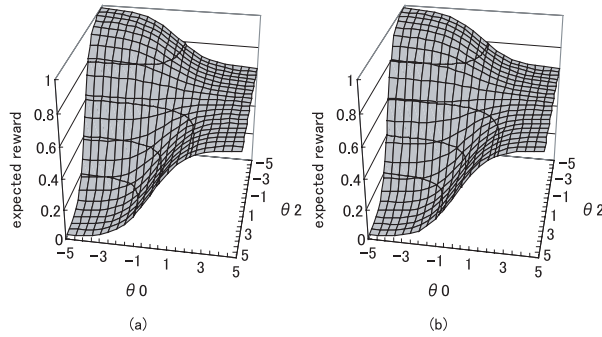


図 12 深さ 2 の 2 分木問題における真の期待値と推定された期待値の比較 (a) 真の期待値, (b) 重点サンプリングによって推定された期待値

中最適政策が獲得された試行の回数を表す。図 12 から $\theta = -3$ で真の期待値がほぼ 1 となることが分かる。そこで、SGA については報酬 1 の獲得に影響するすべての分岐点 (s_{2i-2}) における政策パラメータ θ が $\theta < -3$ となる時に収束と判断し、PVIS, PGIS, GA については集団内の最良個体について上記の条件が満足されるときに収束と判断した。表 1 から PGIS と PVIS が他に比べて高い割合で最適政策を獲得していることが分かる。表 2 は Q-learning が最適政策に収束するまでに要したステップ数を示している。ここで、 $\epsilon = 0.05$ の ϵ -greedy を用いる Q-learning では、深さ 2 の場合の期待報酬の最適値は、 $2/3 \cdot 0.025 + 1 \cdot 0.975^2 = 0.9673$ となる。深さ 4, 5 の場合も同様に、0.9204, 0.8978 となる。そこで、平均獲得報酬がこれらの値に達したときに Q-learning の収束と判断する。表 2 から分かるように、Q-learning は最適政策への収束が保証されているが、深さ 4 以上では現実的なステップ数での収束は実現されない。この点において、確率的ではあるが高速に最適解に収束する提案手法のほうが実用的である。

表 1 2000 ステップまでに 100 試行中最適政策が獲得された試行の回数

	$n = 2$	$n = 4$	$n = 5$
PGIS	100	93	66
PVIS	100	95	81
GA	66	12	2
SGA	65	0	0

5. 考 察

i. 学習速度について

学習速度について考察する。環境とのインタラクション数が学習時間を規定するとの考えから、本論文では、収束に要するインタラクション数の逆数をもって学習速度とする。

PVIS が世代交代に要するインタラクション数は GA の $2/C$ 倍であり、 $C/2$ 倍の高速化が期待できる。本論文で用いた実験設定では $C/2 = 5$ 倍の高速化となる。しかし、実験ではそれ以上の高速化が達成されている。提案手法では、重点サンプリングにより集団の経験を再利用できるので、評価値の分散が小さくなり生存選択が安定することが高速化に寄与しているためと考えられる。

また、政策空間が高次元になると、PGIS が PVIS よりも性能が良くなる。これは PGIS が勾配情報を用いて政策を改善しているためである。政策空間に景観上の困難がない限り、この手法は学習速度の面で有利である。さらに、PGIS の勾配法による子個体生成は並列実装が可能であり、並列化により子個体生成に要する計算時間を $1/4$ にすることができる。

表 2 深さ n の 2 分木問題に対する Q-learning の収束に要するステップ数

	最適値 ($\epsilon = 0.05$)	平均ステップ数
$n = 2$	0.9673	2700
$n = 4$	0.9204	13000
$n = 5$	0.8978	20000

PGIS と PVIS は Q-learning と比べても有意に高速である。Q-learning は MDP において最適政策への収束が保証されていることが利点であるが、実応用を考慮した場合に、現実的なステップ数では収束しないことは大きな欠点であり、その点で提案手法は優位であると言える。

ii. 大域的収束性について

多峰性の強い問題では、政策勾配だけを利用する SGA は局所解に陥ることが多いのに対し、PVIS, PGIS, GA は大域的最適政策を発見できる能力を持つ。これは多点探索法の利点である。

しかし、深さ 4, 5 の 2 分木問題では、PVIS や PGIS であっても獲得報酬 2/3 の局所解に陥ることが観察されている。これは最適政策の解が政策空間の端に存在するために、集団が中央付近に存在する局所解の方向に引き寄せられてしまうことが原因と考えられる。そこで、集団サイズを大きくして同様の実験を行なった結果、集団サイズを大きくすることで、探索時間は増加するものの、大域的最適政策を発見しやすくなることが確認されている。

次に、PGIS が勾配法以外に交叉によっても子個体を生成している効果を考察する。PGIS から交叉を取り除き、 $\{P_1, P_2, C_3, C_4\}$ の 4 個体だけで家族を構成する方法で同様の実験を行なった結果、交叉をしない PGIS は頻繁に局所解に陥ることが観察された。このことから、PGIS の交叉は多峰性への対処に寄与していると言える。

Q-learning は MDP において最適政策への収束が保証されているのに対し、PGIS と PVIS は確率的な収束しか保証されていない。しかし、提案手法は POMDP に対応可能である。これは提案手法が重点サンプリングによる評価値推定をエピソード単位で行なっているためである。

iii. 提案手法のパラメータについて

提案手法のパラメータである集団サイズ (N)、エピソード長 (M)、PVIS における生成子個体数 (C) について議論する。

集団サイズ N は重点サンプリングに用いるサンプル数に関係するので、重点サンプリングの推定精度に影響する。 N が大きいほど、重点サンプリングの推定精度は向上する。

エピソード長 M は親個体の評価値の精度に影響する。 M が大きいほど、親個体の評価値の精度は向上するが、提案手法ではエピソードの生起確率の計算が不安定になる。匍匐ロボットを用いた実験では、おおそ 30 ステップが限界であることが分かっている [Tsuchiya 04]。長いエピソードに対応することは今後の課題である。

生成子個体数 C は GA の探索能力に影響する。 C が大きいほど、GA の探索能力は向上する。それに伴って、重点サンプリングによって推定しなければならない個体数も増えるが、推定処理は数値計算のみで可能である。すなわち、環境とのインタラクションは増えない。これは、多数の子個体を生成する MGG の枠組みと重点サンプリ

ングによる評価値推定の親和性が高いことを意味する。

iv. 政策表現について

本論文の実験では、政策表現にニューラルネットを用いた。ニューラルネットによる政策表現は強化学習手法でよく使われる方法であるが、GA や重点サンプリングに適しているとは言い切れない。また、ニューラルネットによる政策は内部がブラックボックスとなり、ユーザにとって分かりにくいという欠点がある。本論文では、先行研究である SGA と同じ設定にするためにこの政策表現を用いたが、政策表現に関しては今後更なる調査研究を行う必要があると考えている。

6. お わ り に

本論文では、GA の世代交代モデル MGG と集団の経験を再利用する重点サンプリングが強い親和性を持つことに着目し、重点サンプリングを用いて子個体の評価値を推定する PVIS を提案した。さらに、MGG の枠組みを用いて並列で勾配法を適用する PGIS を提案した。PGIS は、勾配情報を基に高速な学習を実現する強化学習手法の長所と、集団で解を探索することで多峰性に対処する GA の長所を取り入れた手法である。これらの提案手法の有効性は実験を通じて確認された。

今後の研究課題として、以下の 4 つに取り組む予定である。

- POMDP への適用： 提案手法はエピソード単位で政策の評価値を推定しているの、観測状態しか利用でいない POMDP への拡張は自然に行なえる。
- 連続的タスクの扱い： エピソード長が大きくなると重点サンプリングの計算が不安定になる問題を緩和するために、タスクをある長さに分割して扱う必要がある。そのことが政策評価に悪影響を及ぼさないようにする工夫が必要である。
- 政策表現： 考察で述べたように、ニューラルネットによる政策表現は強化学習手法でよく使われる方法であるが、提案手法に適した政策表現を調査研究する必要がある。
- 他の探索手法との組み合わせ： GA と同程度に多峰性に対処できると考えられるマルチスタート・ローカルサーチと重点サンプリングの組み合わせも本論文に示した方法と同様の方法で構成可能である。与えられた問題に対し探索手法と重点サンプリングをどのように組み合わせたらよいのかという設計指針を確立することは今後の重要な課題のひとつである。

◇ 参 考 文 献 ◇

- [Aoki 04] Aoki, K., Kimura, H., Kobayashi, S.: Distributed Reinforcement Learning using Bi-directional Decision Making for Multi-criteria Control of Multi-Stage Flow Systems, The 8th Conference on Intelligent Autonomous

- Systems(IAS-8), pp.281–290, 2004.
- [Bellman 57] Bellman, R. E.: Dynamic Programming. Princeton University Press, Princeton, 1957.
- [Boyan 93] Boyan, J., Littman, M.: Packet routing in dynamically changing networks: A reinforcement learning approach, Advances in Neural Information Processing Systems 6, pp.671–678, 1993.
- [Crites 95] Crites, R., Barto, A.: Improving elevator performance using reinforcement learning, Advances in Neural Information Processing Systems 8, pp.1017–1023, 1995.
- [Goldberg 89] Goldberg, D.E.: Genetic Algorithms in Search, Optimization, and Machine Learning, Addison-Wesley Publishing Company Inc., 1989.
- [Kimura 97] Kimura, H., Miyazaki, K., and Kobayashi, S.: Reinforcement Learning in POMDPs with Function Approximation, Proc. 14th International Conf. on Machine Learning (ICML1997), pp.152–160, 1997.
- [Kimura 98] Kimura, H. and Kobayashi, S.: Reinforcement Learning for Continuous Action using Stochastic Gradient Ascent, Intelligent Autonomous Systems (IAS-5), pp.288–295, 1998.
- [Kimura 01] Kimura, H., Yamashita, T. and Kobayashi, S.: Reinforcement Learning of Walking Behavior for a Four-Legged Robot, 40th IEEE Conference on Decision and Control (CDC2001), pp.411–416, 2001.
- [Kimura 04] Kimura, H., Kobayashi, S.: Reinforcement Learning by Policy Improvement Making Use of Experiences of The Other Tasks, The 8th Conference on Intelligent Autonomous Systems(IAS-8), pp.413–421, 2004.
- [Kita 98] Kita, H., Ono, I. and Kobayashi, S.: Theoretical Analysis of the Unimodal Normal Distribution Crossover for Real-coded Genetic Algorithms, Proc. 1998 IEEE ICEC, pp.529–534, 1998.
- [Kumar 99] Kumar, S., Miikkulainen, R.: Confidence based dual reinforcement Q-routing: An adaptive online network routing algorithm, Proc. of 16th international joint conference on artificial intelligence, pp.758–763, 1999.
- [Ono 97] Ono, I. and Kobayashi, S.: A Real-coded Genetic Algorithm for Function Optimization Using Unimodal Normal Distribution Crossover, in Proc. 7th ICGA, pp.246–253, 1997.
- [Pepyne 96] Pepyne, D., Looze, D., Cassandras, C., Djaferis, T.: Application of Q-learning to elevator dispatching, 13th Triennial World Congress, pp.317–322, 1996.
- [Peshkin 02] Peshkin, L., Shelton, C.R.: Learning from Scarce Experience, Proc. 19th International Conf. on Machine Learning (ICML2002), pp.498–505, 2002.
- [Precup 00] Precup, D., Sutton, R.S., and Singh, S.: Eligibility Traces for Off-Policy Policy Evaluation, Proc. 17th International Conf. on Machine Learning (ICML2000), pp.759–766, 2000.
- [Precup 01] Precup, D., Sutton, R.S., and Dasgupta, S.: Off-Policy Temporal-Difference Learning with Function Approximation, Proc. 18th International Conf. on Machine Learning (ICML2001), pp.417–424, 2001.
- [Rubinstein 81] Rubinstein, R.Y.: Simulation and the monte carlo method. Wiley, New York, 1981.
- [Satoh 96] Satoh, H., Yamamura, M. and Kobayashi, S.: Minimal Generation Gap Model for GAs considering Both Exploration and Exploitation, Proceedings of IIZUKA'96, pp.494–497, 1996.
- [Shelton 01a] Shelton, C.R.: Policy Improvement for POMDPs using Normalized Importance Sampling, Proceedings of the 17th Conference on Uncertainty in Artificial Intelligence (UAI2001), pp.496–503, 2001.
- [Shelton 01b] Shelton, C.R.: Policy improvement for POMDPs using normalized importance sampling, Technical Report AI-MEMO 2001-002, MIT, AI Lab, 2001.
- [Singh 97] Singh, S., Bertsekas, D.: Reinforcement learning for dynamic channel allocation in cellular telephone systems, Advances in Neural Information Processing Systems 9, pp.974–980, 1997.
- [Tsuchiya 04] Tsuchiya, C., Kimura, H., Kobayashi, S.: Policy Learning by GA using Importance Sampling, The 8th Conference on Intelligent Autonomous Systems(IAS-8), pp.385–394, 2004.
- [内部 03] 内部 英治, 銅谷 賢治: 重点サンプリングを用いた複数強化学習器の同時学習, 信学技報 (ニューロコンピューティング研究会), pp.179–184, 2003.
- [匹田 02] 匹田 志朗, 阿部 茂: エレベータ群管理制御における AI 技術の応用, 人工知能学会誌, Vol.17, No.1, pp.57–62, 2002.
- [Williams 92] Williams, R. J.: Simple Statistical Gradient Following Algorithms for Connectionist Reinforcement Learning, Machine Learning 8, pp.229–256, 1992.

(担当委員: 山川 宏)

2004 年 3 月 31 日 受理

著者紹介

土谷 千加夫(学生会員)



2002 年東京農工大学工学部情報コミュニケーション学科卒業。2004 年東京工業大学大学院総合理工学研究科知能システム科学専攻修士課程修了, 同年 4 月博士後期課程進修と共に(株)東芝に入社, 現在同社研究開発センター知識メディアラボラトリーにてエージェント技術の研究開発に従事。大学では強化学習の研究に従事。

木村 元(正会員)



1992 年東京工業大学工学部制御工学科卒業。1994 年同大学大学院総合理工学研究科知能システム科学専攻修士課程修了。1997 年同大学大学院博士課程修了。同年 4 月日本学術振興会 PD 研究員, 1998 年 4 月, 東京工業大学大学院総合理工学研究科助手。2004 年 4 月, 九州大学大学院工学研究院海洋システム工学部門助教授, 現在に至る。人工知能, 特に強化学習に関する研究を行なっている。計測自動制御学会, 電気学会, 日本ロボット学会各会員。

佐久間 淳(正会員)



1997 年東京工業大学生命理工学部生物工学科卒業, 2003 年 3 月, 同大学院総合理工学研究科知能システム科学専攻博士後期課程修了。工学博士。同年 4 月, 日本アイ・ビー・エム株式会社入社, 東京基礎研究所に勤務。2004 年 7 月, 東京工業大学総合理工学研究科助手, 現在に至る。最適化アルゴリズム, 機械学習, データマイニング, テキストマイニングの研究に従事。

小林 重信(正会員)



1974 年東京工業大学大学院博士過程経営工学専攻終了。同年 4 月, 同大学工学部制御工学科助手。1981 年 8 月, 同大学大学院総合理工学研究科助教授。1990 年 8 月, 教授。現在に至る。創発的設計論・制御論の確立を目的に進化計算, 強化学習などの適応システムの研究に従事。計測自動制御学会, 情報処理学会各会員。