重点サンプリングを用いた政策勾配の推定による子個体生成

○土谷　千加夫，木村　元，小林　重信
東京工業大学　大学院総合理工学研究科

Generating Offspring using Estimated Policy Gradient with Importance Sampling

Chikao TSUCHIYA，Hajime KIMURA，Shigenobu KOBAYASHI
Interdisciplinary Graduate School of Science and Eng., Tokyo Institute of Technology

***Abstract*— The most difficult problem of applying GA to a policy search is that evaluation of offspring requires many interactions with an environment. In this paper, we propose a new approach using *importance sampling*. The proposed technique decreases the interactions in evaluating offspring, and improves offspring using estimated policy gradient. It can mitigate the load to the hardware accompanying trial and error and as such, is particularly efficient for a real robot's policy search. The proposed technique was implemented to the crawling robot simulator. The experimental results showed the strong affinity between GA and importance sampling, and confirmed that estimation of policy gradient by importance sampling accelerates a policy search.**

1　はじめに

強化学習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ような政策を学習する．すなわち，状態から行動への写像を求めることが問題とされる．しかし，強化学習は基本的には局所探索であるため，多峰性の景観を持つ政策に適用した際に，局所解に陥る可能性がある．一方，GA は集団で解を探索することから，多峰性の問題に対応できる利点を持つ．しかし，環境とのインタラクションを伴う政策探索では膨大な試行錯誤を必要とすることが大きな障害となっていた．

近年，ある政策のために使われた状態遷移系列のデータを別の政策の学習に再利用しようとの考えから，重点サンプリングの有効性が注目されている [9][6][7]．強化学習の分野では MDP での Q 値を推定するために重点サンプリングを用いる研究がいくつかなされている．一方，GA の分野では政策学習に関する研究はまだ少ないし，また，重点サンプリングに注目した研究は皆無の状況にある．

本論文では，多峰性に対応した政策探索と学習の高速化を達成するために，重点サンプリングを導入した接近法を示す．ナイーブな GA による政策探索では交叉または突然変異によって生成された子個体の政策を評価するために環境とのインタラクションを必要とし，このことが実用上の障害となっている．そこで，著者らは重点サンプリングで子個体の政策評価値を推定する PVIS を提案した [10]．PVIS により環境とのインタラクションを大幅に削減できるが，重点サンプリングの計算量が問題となっていた．そこで，本論文では，政策勾配を用いた局所探索で子個体を改善する PGIS を提案する．

PGIS 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．そして，交叉に加えて重点サンプリングで推定した政策勾配に基づく局所探索で子個体を改善する．勾配情報を用いることで限られた生成子個体数でも効率的な探索が可能となる．これらのプロセスは数値計算のみで可能なため，環境とのインタラクションは削減される．提案手法の有効性は匍匐ロボットを用いた実験を通して確認する．

2　準備

ここでは，強化学習手法と GA それぞれによる政策探索と重点サンプリングを説明する．

2.1　政策探索

環境中に置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．形式的には，エージェントは利得の最大化を目的として，状態観測から行動出力への写像を獲得する．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，政策をパラメータ θ によって記述することができる．そのパラメータを最適化すれば，最適な政策を獲得することができる．ここでは，パラメータを探索する枠組みを政策探索と呼ぶことにする．

2.2　強化学習手法による政策探索

強化学習手法の多くは，状価値や政策価値の勾配を用いて局所探索を行う．ここでは，勾配を用いて局所探索を行う手法を強化学習手法と呼ぶことにする．

本論文では後述するように POMDP を扱う．POMDP を効率的に扱える手法として，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 (SGA) が提案されている [1]．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．しかし，基本的には局所探索法であるので，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，適用が容易であり，数多くの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世代交代モデル MGG[8] を用い，交叉に UNDX[5] を用いる単純な実数値 GA を取り上げる．ここではこれを GA と呼ぶことにする．GA では，生成子個体の評価値を環境とインタラクションにより獲得する．GA は多峰性に対処できることから，多峰性を持つ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が，生成子個体の評価に際して，膨大な時間が必要となる．

2.4 重点サンプリング

重点サンプリングとは分布の不一致を扱う統計手法である．パラメータ θ で記述され政策 (ビヘイビア政策) とその政策から得られたデータがあるとする．別なパラメータ θ' で記述された政策 (ターゲット政策) があるとき，重点サンプリングを用いると，その政策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 で観測する状態を $o(s)$，状態 s で行動 a を選択する確率を $\pi(o(s),a;\theta)$ と表すと，この政策の下でエピソード $h=\{s_1,a_1,r_1,s_2,a_2,r_2,...,s_M,a_M,r_M,s_{M+1}\}$ が得られる確率は次式で表される：

$$\Pr(h|\theta) = \Pr(s_1)\prod_{i=1}^{M}\pi(o(s_i),a_i;\theta)\Pr(s_{i+1}|s_i,a_i) \quad (1)$$

ここで，$\Pr(s_1)$ 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，ターゲット政策の評価値 $\hat{V}(\theta')$ は次のように推定できる [6]：

$$\hat{V}(\theta') = \frac{\sum_{i=1}^{N} R_i \prod_{j=1}^{M} \frac{\pi(o(s_j),a_j|\theta')}{\pi(o(s_j),a_j|\theta)}}{\sum_{i=1}^{N} \prod_{j=1}^{M} \frac{\pi(o(s_j),a_j|\theta')}{\pi(o(s_j),a_j|\theta)}} \quad (2)$$

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．この式は，評価値の推定値は各政策での行動選択確率の尤度比だけで計算できる．

強化学習の分野では，重点サンプリングを複数タスクの同時学習に適用する研究がなされている [3].

3 問題の定式化と接近法

3.1 問題の定式化

実問題では，ノイズやセンサの能力が不十分なため，状態観測に不確実性や不完全性が存在することが多い．そこで，本論文では，POMDP を対象クラスとし，連続状態・連続行動の政策探索問題を扱う．POMDP は次のように示される．S を状態空間，X を観測状態空間，A を行動空間，R を実数値の集合とする．各離散時間 t において，エージェントは状態 $s_t \in S$ に置かれ，$x_t \in X$ を観測し，行動 $a_t \in A$ を実行し，環境の状態遷移の結果として即時報酬 $r_t \in R$ を受け取る．一般に，報酬と次状態はランダムであるが，その確率分布は 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，報酬 r_t は期待値 $r(s_t,a)$ に従ってランダムに与えられる．

学習の目標は，事前知識がない状況でパフォーマンスを最大化することである．一般に，政策空間は多峰性を持つことが予想される．そのため，学習の高速化に加えて，大域的な最適政策を探索することが学習の目標である．

3.2 重点サンプリングによる子個体の評価

ここでは，著者らが過去に提案した PVIS について簡単に説明する．

3.2.1 重点サンプリングによる政策価値の推定

式 (2) の重点サンプリングは単一の政策下でのサンプルを利用して，他の政策の評価値を求めているが，GA は集団を保持しているため，集団の経験を再利用できれば推定精度が向上すると考えられる．そこで，次のように複数の政策 $\theta_1,\theta_2,...,\theta_N$ でのサンプルを用いる重点サンプリングを用いる [9]：

$$\begin{aligned}\hat{V}(\theta) &= \frac{\sum_{i=1}^{N} R_i \frac{\Pr(h_i|\theta)}{\sum_{j=1}^{N}\Pr(h_i|\theta_j)}}{\sum_{i=1}^{N}\frac{\Pr(h_i|\theta)}{\sum_{j=1}^{N}\Pr(h_i|\theta_j)}} \\ &= \frac{\sum_{i=1}^{N} R_i \frac{\prod_{k=1}^{M}\pi(o(s_k),a_k|\theta)}{\sum_{j=1}^{N}\prod_{k=1}^{M}\pi(o(s_k),a_k|\theta_j)}}{\sum_{i=1}^{N}\frac{\prod_{k=1}^{M}\pi(o(s_k),a_k|\theta)}{\sum_{j=1}^{N}\prod_{k=1}^{M}\pi(o(s_k),a_k|\theta_j)}}\end{aligned} \quad (3)$$

式 (3) 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．

3.2.2 アルゴリズム：PVIS

集団の政策によって得られた経験をデータベースに保持しておく．世代交代の際に集団中から選択された親個体から交叉によって多数の子個体を生成し，それらの評価値を式 (3) で推定する．その推定値を用いてランクベースのルーレット選択を行ない，選択された個体を集団に戻す．

1. Generate N policies as an initial population, and obtain the experiences under these policies.
2. Choose parents $\{P_1, P_2\}$ and additional parents from N policies by random sampling, and generate C offspring $\{C_1, ..., C_{|C|}\}$ by a crossover. Let $\{P_1, P_2, C_1, ..., C_{|C|}\}$ be family.
3. Estimate the values of the family by importance sampling.
4. Choose two policies from the family: one is the best and the other is selected by the rank-based roulette wheel selection. Replace the two parents with those two policies.
5. Throw away with the experiences obtained under the parents' policies.
6. Obtain the experiences through interactions with the environment under the two newly added policies.
7. Repeat the above procedures from step 2.

Fig.1: The algorithm of PVIS.

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 (3) の精度は順序を保存する程度であればよい．このアルゴリズムを PVIS と呼ぶことにする (Fig.1，Fig.2).

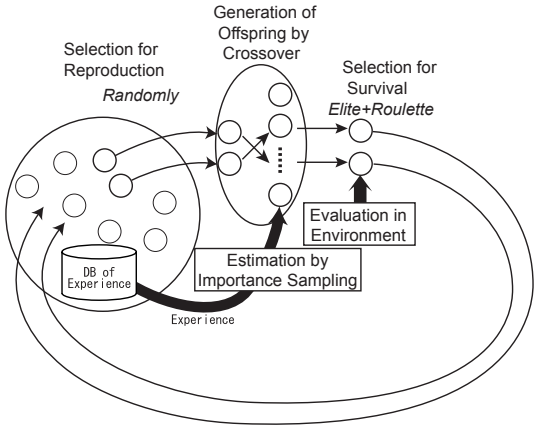

Fig.2: The framework of the direct policy search by PVIS.

3.3 政策勾配を用いた局所探索による子個体の改善

ここでは，本論文の提案手法である PGIS について説明する．MGG において，交叉によって多数の子個体を生成することは，親個体の周辺を局所探索していることに相当する．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指数関数的に増加する．前述の 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．しかし，計算量の増加は深刻である．そこで，交叉に加えて政策勾配に基づいた局所探索を導入する．これを PGIS と呼ぶことにする．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される．

3.3.1 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 (2) で表されるので，式 (2) を i 番目のパラメータ θ_i で微分することでターゲット政策の勾配を以下のように計算できる：

$$\frac{\partial}{\partial \theta_i}\hat{V}(\theta) = \frac{\sum_{i=1}^{N}(R_i - \hat{V})\frac{\frac{\partial}{\partial \theta_i}\Pr(h_i|\theta)}{\sum_{j=1}^{N}\Pr(h_i|\theta_j)}}{\sum_{i=1}^{N}\frac{\Pr(h_i|\theta)}{\sum_{j=1}^{N}\Pr(h_i|\theta_j)}} \quad (4)$$

政策価値の最大化問題の場合，政策を式 (4) の方向へ更新すればよい．式 (4) の妥当性は式 (2) 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．

3.3.2 アルゴリズム：PGIS

MGG の枠組みに PGIS を組み込む．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によって生成された子個体を式 (4) で推定される政策勾配を用いた勾配法で更新する．最後に，それらを家族として生存選択を行う．Fig.3 に PGIS のアルゴリズムの詳細を示し，Fig.4 にこのアルゴリズムの枠組みを示す．

4 実験と考察

提案手法の有効性を確認するために匍匐ロボットを用いた比較実験を行った．

4.1 匍匐ロボット

実験には Fig.5 に示すような匍匐ロボットシミュレータを用いる．匍匐ロボットは 2 つのサーボモータによって間接角度を制御できる N 本のアームとアームの先端が地面に接触したかどうかを調べるタッチセンサーを備えている．実験では 1 本足，2 本足，4 本足の匍匐ロボットを扱う．タスクの目的は，できるだけ高速に前進するような制御規則を獲得することである．

匍匐ロボットは範囲が制限された連続の状態変数と離散の状態変数を持つ．連続な状態変数は各アームの 2 つのジョイントの角度であり，離散の状態変数はアームのタッチセンサーを表している．エージェントはこれらの状態変数を観測する．エージェントの選択する行動は 2 つのモー

1. Generate N policies as an initial population, and obtain the experiences under all these policies.
2. Choose parents $\{P_1, P_2\}$ and additional parent from N policies by random sampling, generate two offspring $\{C_1, C_2\}$ by a crossover.
3. Improve P_1 and P_2 by a gradient method using policy gradient estimated by importance sampling. Let those be $\{C_3, C_4\}$.
4. Improve C_1 and C_2 in the same way. Let $\{P_1, P_2, C_1, C_2, C_3, C_4\}$ be family.
5. Estimate the values of family by importance sampling.
6. Choose two policies from the family: one is the best and the other is selected by the rank-based roulette wheel selection. Replace the two parents with those two policies.
7. Throw away with the experiences obtained under the parents' policies.
8. Obtain the experiences through interactions with the environment under the two newly added policies.
9. Repeat the above procedures from step 2.

Fig.3: The algorithm of PGIS.

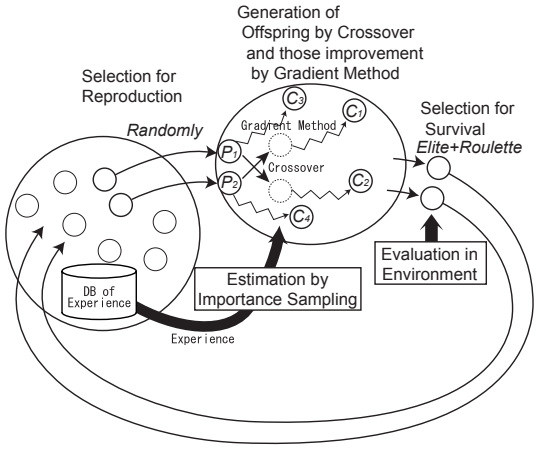

Fig.4: The framework of the direct policy search by PGIS.

タの角度である．これは連続な状態の次元と同じである．エージェントが行動を選択すると，ロボットは指定された位置に向かってモータを動かす．指定された位置までモータが動いたら，遷移の結果として報酬が与えられ，時間ステップは次ステップに進む．匍匐ロボットの行動はモータが指定した角度まで動くかタッチセンサーの値が変化したときに止まる．つまり，アームが地面に接触し続けているか，地面から離れ続けているときは，アームを目標角度まで動かすことができる．そのため，モータの動作中にセンサの値が変化した場合は，モータの角度は選択された目標角度に対応しなくなり，状態遷移の不確実性が存在する．

また，報酬信号は与えられたタスクの達成度を表している．ここでは，即時報酬は各ステップでのボディの速度とした．

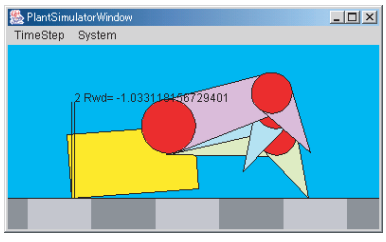

Fig.5: Imaginary crawling robot.

4.2 実験の設定

4.2.1 政策表現

各足毎に行動次元数は 2 次元で，それぞれ [0,1] の区間を持つ．状態観測は各関節の角度 (2 次元ベクトル，各要素は [0,1]) およびタッチセンサの状態 (1 次元ベクトル，各要素は 0 か 1) の 3 次元ベクトル $X = (x_1, x_2, x_3)$ である．

N 本足匍匐ロボットの政策は，入力層と出力層からなる 2 層のニューラルネットで表現される (Fig.6)．各足毎に 3 次元の状態観測ベクトル $X = (x_1, x_2, x_3)$ を基に新たな 3 次元の特徴量ベクトル $(x_4 = 1 - x_1, x_5 = 1 - x_2, x_6 = 1 - x_3)$ を構成し，全体では特徴量ベクトル $F = (x_1, x_2, .., x_{6N}, 1)$ を構成する．最後の要素は常に 1 とする．これに重みベクトル $\Theta = (\theta_{1,i}, \theta_{2,i}, ..., \theta_{6N+1,i})$ を用いて，中心値 $\mu_i = 1/(1 + \exp(-\sum_{k=1}^{6N} \theta_{k,i} x_k))$，標準偏差 $\sigma_i = 1/(1 + \exp(-\theta_{6N+1,i})) + 0.1$ の正規分布 $N(\mu_i, \sigma_i)$ に従って次元 i の行動を選択する．ただし区間 [0,1] の範囲外の場合は区間内になるまで繰り返す [2]．よって，N 本足匍匐ロボットの政策パラメータ数は 2N(6N+1) 個であり，探索空間は 2N(6N+1) 次元となる．1, 2, 4 本足匍匐ロボットでは，探索空間はそれぞれ 14，52，200 次元となる．

4.2.2 手法の設定

性能比較のため，提案手法 PGIS を GA，SGA，PVIS と比較する．PGIS の局所探索法として，単純なリニアサーチを採用する．式 (4) で示される勾配方向へ L 個の探索点を S 間隔で配置し，各点の評価値を推定し，最良点を受理する．放物線補間や黄金分割法などの高価なリニアサーチも考えられるが，これらは最良点の囲い込み処理および単峰な景観を前提にしているので，複雑な景観が予想される

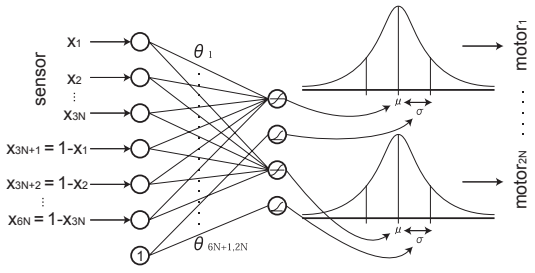

Fig.6: Representation of Policy for N-legged crawling robot.

本問題には適さない.

この実験における各手法の設定をTABLE1 に示す．SGA と GA の細かい設定はそれぞれ [1]，[4] に基づく.

Table1: The configuration of each method.

	PVIS	PGIS	GA
N	30	30	30
M	20	20	20
C	10	—	10
S	—	1.0	—
L	—	5	—

4.3 実験の結果

1，2，4 本足匍匐ロボットの実験結果をそれぞれ Fig.7, Fig.8，Fig.9 に示す.

Fig.7 によると，学習は，SGA，PVIS，PGIS，GA の順に高速である．しかしながら，SGA だけは 5000 ステップ付近で急に不安定になり，性能が低下している．それ以外の手法は安定して学習している．一本足匍匐ロボットの最大平均報酬は経験的に約 4.0 であることがわかっている．平均報酬が 4.0 に達するのに要するステップ数は，PVIS で約 14000 ステップ，PGIS で約 20000 ステップ，GA で約 46000 ステップ (図には示されていない) である．したがって，この実験における PVIS の GA に対する高速化は約 3.3 倍，PGIS の GA に対する高速化は約 2.3 倍である.

Fig.8，Fig.9 によると，学習の進行は，SGA，PGIS，PVIS，GA の順に高速である．1 本足匍匐ロボットの場合と対照的に，SGA は安定している．SGA と PGIS に着目すると，2，4 本足匍匐ロボットの両方で，SGA は PGIS の約 2 倍高速である．また，PVIS と PGIS は順位が逆転している.

4.4 考察

1 本足匍匐ロボットでは SGA は平均報酬 2.0 にしか到達できない試行がいくつか確認されたが，2，4 本足匍匐ロボットでは安定している．このことは，1 本足匍匐ロボットの政策空間は多峰性を有するが，2，4 本足匍匐ロボットの政策空間は単峰に近いことを示している．1 本足匍匐ロボットと 2，4 本足匍匐ロボットで PVIS と PGIS の順位が逆転していることを考えると，PVIS は PGIS よりも多峰性に対処でき，政策勾配を用いる PGIS は PGIS よりも高次元空間に対処できると言える．まとめると，多峰性への適応性は，PVIS，PGIS，SGA の順に優れている．PGIS が多峰性にもある程度対処できるのは，局所探索だけでなく，交叉によっても子個体を生成している効果であると考えられる．PGIS における交叉による効果を調べるために，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一本足匍匐ロボットの実験

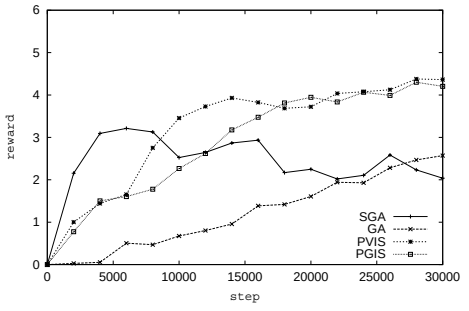

Fig.7: The performance of learned policy in 1-legged crawling robot, averaged over 10 trials.

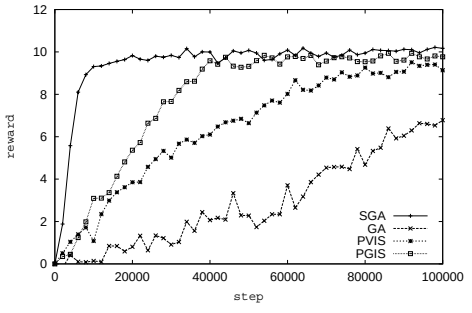

Fig.8: The performance of learned policy in 2-legged crawling robot, averaged over 10 trials.

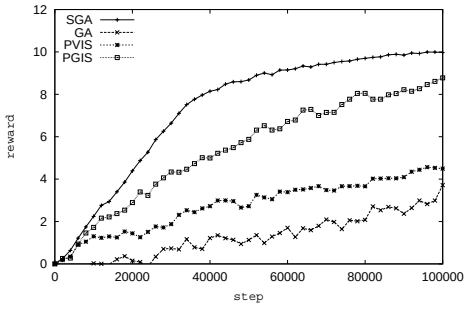

Fig.9: The performance of learned policy in 4-legged crawling robot, averaged over 10 trials.

を行なった．その結果では，交叉をしない PGIS の性能は GA と同程度であった．交叉をしない PGIS は明らかに局所最適政策に陥っている．このことから，PGIS の交叉は多峰性への対処に貢献していると言える．

また，2，4 本足匍匐ロボットの場合，学習速度は SGA, PGIS，PVIS の順に優れていた．勾配情報を用いない著者らの以前の提案手法 PVIS よりも PGIS が学習速度の点で優れていたことから，MGG の枠組みに政策勾配を導入した効果があったと言える．

最後に，PGIS におけるサンプリング間隔 (S) とサンプル数 (L) について考える．PGIS の S と L は，政策のパラメータ数の他，対象問題の性質にも依存する．この実験では，予備実験での試行錯誤によりパラメータを設定した．現時点では S と L の設計指針はない．ステップサイズを適応的に設定できれば効率の良い局所探索が可能になるだろう．ステップサイズを十分に小さな定数に設定し，勾配方向へパラメータの更新を繰り返す方法は効果的かもしれない．我々はパラメータを推定される政策勾配方向へステップサイズ 2.0 で更新することを繰り返す方法を 1，2，4 本足匍匐ロボットに適用した．一本足匍匐ロボットでは，SGA に迫る高速な学習を達成できた．これは多峰の政策空間の尾根に沿ってパラメータを更新できたためと考えられる．一方，2，4 本足匍匐ロボットでは GA と同程度の性能であった．これは高次元空間での重点サンプリングの不安定さが原因であると考えられる．これらの結果から，PGIS の勾配法には改善の余地が多分にあると言える．

5 おわりに

本論文では，集団で解を探索する MGG と集団の経験を再利用できる重点サンプリングが強い親和性を持つことに着目し，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匍匐ロボットの実験を通じて確認された．

PGIS の性能は推定された政策勾配に基づいた局所探索に依存する．高次元政策空間での効率的な局所探索の実現は今後の課題である．

REFERENCES

[1] Kimura, H. and Kobayashi, S.: Reinforcement Learning for Continuous Action using Stochastic Gradient Ascent, Intelligent Autonomous Systems (IAS-5), pp.288–295, 1998.

[2] 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.

[3] 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), to appear.

[4] 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.

[5] 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.

Peshkin, L., Shelton, C.R.: Learning from Scarce Experience, Proc. 19th International Conf. on Machine Learning (ICML2002), pp.498–505, 2002.

[6] 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.

[7] 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.

[8] 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.

[9] 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.

[10] Tsuchiya, C., Kimura, H., Kobayashi, S.: Policy Learning by GA using Importance Sampling, The 8th Conference on Intelligent Autonomous Systems(IAS-8), to appear.