重点サンプリングを用いたGAの高速化

○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

Speed Up of GA by 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 learning is that interactions with the environment require much time to evaluate the individuals. In this paper, we propose a new approach to estimate the individual's value using *importance sampling*. Importance sampling reuses the experiences obtained by some policy to estimate values of the other policies. The proposed technique cuts down the interactions with the environment in evaluating children, it can speed up optimization. In particular, it is effective in case GA is applied to a real robot's policy learning, because the load to the hardware accompanying trial and error can be mitigated. The proposed technique was implemented to the crawling robot, it was applied to obtain the control rules so that the robot is to walk. The experimental results show the strong affinity between GA and importance sampling, and also mean that GA using importance sampling can be a powerful tool for policy learning.

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要請される政策学習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 [Shelton 2001][Precup 2000][Precup 2001]．強化学習の分野ではマルチタスクの学習に重点サンプリングを適用す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．加えて，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，GA による政策学習を高速化するために重点サンプリングの導入を図る．ナイーブな GA による政策学習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．我々は，重点サンプリングを用いて親集団の経験を再利用することにより子個体の評価を行なう．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．

エージェントは複数の政策を個体集団として持ち，それらに従って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，個体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．重点サンプリングで蓄積された経験を処理することで，別な政策で獲得された報酬を子個体が獲得するであろう報酬に変換し，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実行されることがないので，子個体評価にかかる時間は重点サンプリングの計算処理にかかる時間だけとなり，最適化プロセス自体が高速化される．我々は提案手法を匍匐ロボットに実装し，前進動作規則の獲得に適用した．子個体の政策で環境とインタラクションすることで評価値を得る方法と提案手法の比較実験を通して，提案手法の有効性を検証する．

2　問題の定式化

2.1　対象問題

対象問題はマルコフ決定過程 (MDP) における強化学習問題である．ただし，後述する提案手法は，部分観測マルコフ決定過程 (POMDP) 下でも適用可能である．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 $s_t \in \mathcal{S}$ 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．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エピソード当たりの平均獲得報酬である：

$$V = \frac{1}{M}\sum_{i=1}^{M} r_i$$

2.2 実数値 GA による政策学習

政策パラメータ θ の探索に実数値 GA の交叉オペレータである UNDX[Ono 1997] と多様性維持に優れた世代交代モデル MGG[Satoh 1996] を用いる．UNDX+MGG の枠組みでは，集団からランダムに親個体を選択し (複製選択；Selection for Reproduction)，それらを親として多数の子個体を生成する．そして，子個体の評価値を求め，子個体とその親個体を含む家族から 2 個体を選択し (生存選択；Selection for Survival)，集団に戻す．ナイーブに GA を政策学習に適用すると，生成された子個体の評価値を求めるために，それらの政策を用いて環境とインタラクションする必要がある．Fig.1 はナイーブな GA による政策学習の枠組みを示している．

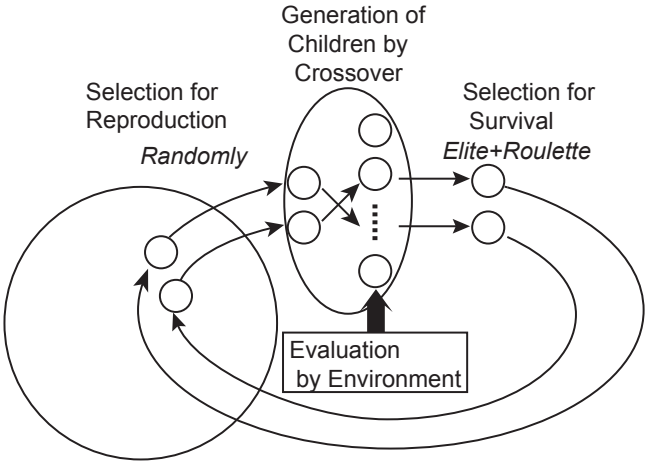

Fig.1: The framework of the policy learning by the naive GA. Evaluating children requires interactions with an given environment. Only the selected individuals require interactions.

ナイーブな GA による政策学習では，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，この部分に要する時間が政策学習全体の大部分を占める．そこで，我々は親集団の経験を再利用し，それによって子個体の評価値を環境とインタラクションすることなしに求めることができれば，探索効率の向上を図れると考えた．

3 重点サンプリングを用いた政策の評価値の推定

3.1 重点サンプリングを用いた推定

政策がパラメータ θ で記述されているとする．別な政策 θ' があるとき，政策 θ' が政策 θ とどのくらい類似しているかを知っていれば，政策 θ によって得られた経験に政策の類似度に比例した修正を施すことで，政策 θ' の評価値を求めることができる．

パラメータ θ で記述される政策 π において，状態 s で行動 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 $h = \{s_1, a_1, r_1, s_2, a_2, r_2, ..., s_M, a_M, r_M, s_{M+1}\}$ が得られる確率は次式で表される：

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

ここで，$\Phi(h)$ は行動選択確率，$\Psi(h)$ は状態遷移確率，$\Pr(s_1)$ は初期状態が s_1 となる確率である．$\Phi(h)$ は政策依存，$\Pr(s_1)$ は政策非依存であることに注意されたい．

重点サンプリングを用いると，パラメータ θ' で記述される別な政策 π' の評価値 $\hat{V}(\theta')$ は次のように推定できる [Precup 2000]：

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

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．$\Pr(s_1)$ は政策非依存なので分母・分子で相殺されるので，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．

さらに，[Precup 2000] は重み付き重点サンプリングを提案している．この方法は従来法に比べて推定値の分散がより小さいという特徴を持つ．これは次のように記述される：

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

3.2 子個体の評価値の推定

重点サンプリングは一つの政策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．GA は複数の親個体から多くの子個体を生成する．したがって，複数の政策によって得られた経験を用いれば推定精度を向上させることができる．[Shelton 2001] は複数の政策 $\theta_1, \theta_2, ..., \theta_N$ によって得られた経験を用いる重み付き重点サンプリングを提案している．それによると評価値は次のようになる：

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

これを用いると，親個体の政策を実際に実行し，エピソードを経験として蓄積しておけば，それを使って子個体の評価値を計算することができる．

3.3 重点サンプリングを用いた政策学習アルゴリズム

提案手法はのアルゴリズムは次のように記述される:

1. 初期集団として N 個の政策をランダムに生成し，その政策を用いて経験を蓄積する

2. N 個の政策中から 2+1 個の政策を親個体としてランダム選択し，UNDX によって子個体を C 個生成する

3. 重点サンプリングを用いて，全子個体の評価値を推定する

4. 親個体とそれらの子個体を含む家族から 2 個体を選ぶ：ひとつは最良個体で，もうひとつは C+1 個体からランクベースのルーレット選択 [Goldberg 1989] で選ばれる．その 2 子個体を親と入れ替える．そして，親個体によって得られた経験を破棄する

5. 新たに加わった 2 子個体の政策を用いて環境とインタラクションし経験を蓄積し，ステップ 2 に戻る

Fig.2 に提案手法による政策探索の枠組みを示す．提案手法では，交叉によって生成された多数の子個体の評価値を環境とのインタラクションなしに推定することができる．集団に戻される 2 個体は実際に環境とインタラクションした経験がないので，経験データを蓄積するために環境とインタラクションを行う．

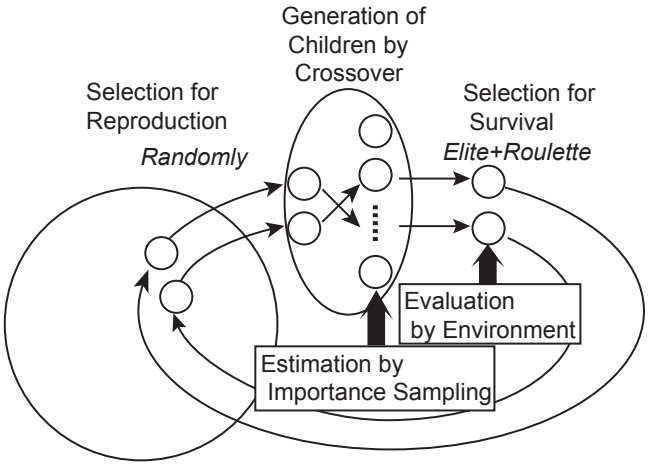

Fig.2: The framework of the policy learning by the proposed technique. `Estimation by IS` requires no interactions with the environment. `Evaluation by Environment` requires only two interactions with the environment.

4 匍匐ロボットへの応用

4.1 匍匐ロボット

本論文では，Fig.3 に示すような匍匐ロボットを用いる．匍匐ロボットは 2 つのサーボモータによって関節角度を制御できるアームを持ち，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．匍匐ロボットと外部の PC は TCP/IP を用いたネットワークで接続されており，PC からロボットに関節角度が送られ，ロボットから PC に遷移後の状態と即時報酬が送られる．

我々は匍匐ロボットが試行錯誤を通して前進するような制御規則を獲得することを目的とする．しかし，実機を用いてアルゴリズムの比較実験を行なうことは困難である．そこで，匍匐ロボットのシミュレータ (Fig.4) を作成し，このシミュレータを用いた実験から提案手法を評価する．このシミュレータも実機と同じように TCP/IP を用いたネットワークで外部の PC と接続されている．

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 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の状態を表している．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状態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，モータの動作中にタッチセンサーの状態が変化した場合は，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．タスクの目的は匍匐ロボットができるだけ高速に前進することであるので，即時報酬は各ステップでのボディの速度とした．

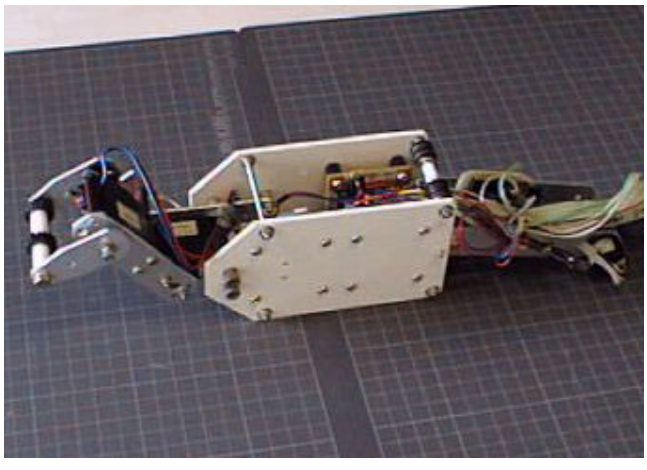

Fig.3: Crawling robot. The robot has an arm with two servo motors and a touch sensor. The arm is controlled by two servo motors that react to angular-position commands. The touch sensor investigates whether the tip of the arm is touching the ground or not.

4.2 ロボットへの実装

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．

政策表現は次の方法で行う．3 次元の状態観測ベクトル $X = (x_1, x_2, x_3)$ を基に 7 次元の特徴量ベクトル $F = (x_1, x_2, x_3, x_4(= 1 - x_1), x_5(= 1 - x_2), x_6(= 1 - x_3), x_7(= 1))$ を構成する．7 番目の要素は常に 1 とする．これに重みベクトル $\Theta = (\theta_{1,i}, \theta_{2,i}, \theta_{3,i}, \theta_{4,i}, \theta_{5,i}, \theta_{6,i}, \theta_{7,i})$ を用いて，中心値 $\mu_i = 1/(1 + \exp(-\sum_{k=1}^{6} \theta_k x_{k,i}))$，標準偏差 $\sigma_i = 1/(1 + \exp(-\theta_{7,i} x_7)) + 0.1$ 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．標準偏差に 0.1 を足しているのは，標準偏差が 0 になって探索行動 (exploration) がとられなくなるのを防ぐためである．また，選択された行動が区間 [0,1] の範囲外の場合は区間内になるまで選択が繰り返される [Kimura 2001]．この政策表現では政策パラメータ数は 14(=7x2) 個であり，GA の探索空間は 14 次元となる．

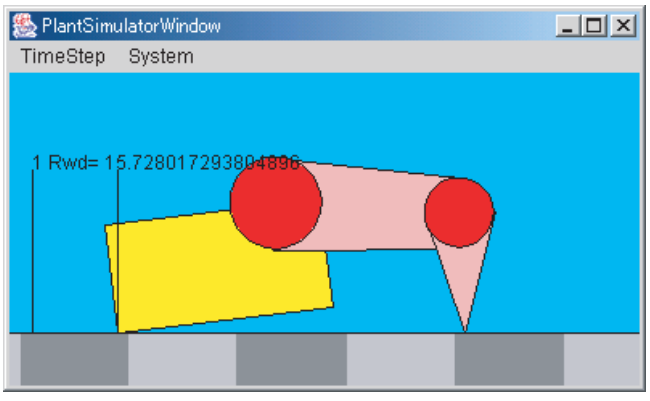

Fig.4: Crawling robot simulator.

4.3 実験設定

この実験では，提案手法 (GA-IS) は 30 個の政策を保持し ($N = 30$)，各エピソードで 20 ステップ行動する ($M = 20$)．世代交代の際には，UNDX によって子個体を 10 個体 ($C = 10$) 生成する．UNDX のパラメータは喜多ら [Kita 1998] による理論解析および小野ら [Ono 1997] による実験的推奨値に基づき $\alpha = 0.5, \beta = 0.35$ を用いた．

性能比較のため，重点サンプリングを用いない 2 種類のナイーブな GA と比較する．1 つ目のナイーブな GA(naive-GA-1) は提案手法と同じ設定である ($N = 30, M = 20, C = 10$)．だが，重点サンプリングを用いた推定を行なわないので，子個体の評価値を決定するのに全子個体に関して環境とのインタラクションが必要である．この設定では，naive-GA-1 は GA-IS の 5 倍 ($= C/2$) の環境とのインタラクションが必要となる．2 つ目のナイーブな GA(naive-GA-2) は各エピソードで 600 ステップ行動する ($N = 30, M = 600, C = 10$)．この設定は，GA-IS が重点サンプリングの処理に仮想的な 600 ステップ分の経験を用いていることに対応する．

4.4 実験結果

提案手法 (GA-IS) と重点サンプリングを用いないナイーブな GA(naive-GA-1, naive-GA-2) を 30000 ステップ実行した．10 試行分の各手法の獲得報酬を平均したものがが Fig.5 である．GA-IS とnaive-GA-1 のエラーバーは 10 試行中の最大値・最小値を表している．

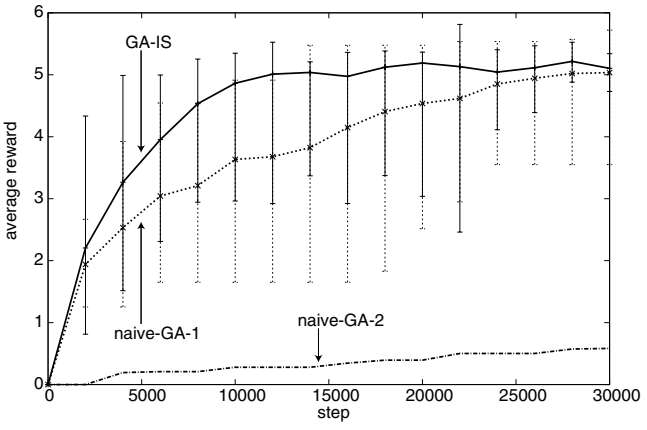

Fig.5: The perfoemance of learned policy averaged over 10 trials. GA-IS is the proposed method, naive-GA-1 and naive-GA-2 are the naive GA for comparison.

5 考察

Fig.5 によると，GA-IS，naive-GA-1，naive-GA-2 の最終性能は共に平均報酬が約 5.3 であり，最終性能に差はない．これは各手法の政策表現能力および探索能力が同じであるために当然の結果である．

学習に要する時間は，平均報酬が 5.0 に達するのに要するステップ数がGA-IS で約 10000 ステップ，naive-GA-1 で約 30000 ステップである．したがって，提案手法による高速化は約 3 倍である．一方，GA-IS のインタラクションはnaive-GA-1 の 5 分の 1 になるので，それによって学習は 5 倍早くなるはずである．これは重点サンプリングによる評価値の推定が完全には正しくないことが原因と考えられる．このタスクにおける重点サンプリングの推定精度を調べるための実験を行なった．Fig.6 は前章の実験と同じ設定でGA-IS によって 3000 ステップ学習した集団から子個体を生成し，その子個体の評価値を最尤推定と重点サンプリングで求めたものである．エラーバーは最尤推定値の 30 エピソードの平均・最小値・最大値を表している．棒グラフは重点サンプリングによる推定値を表している．最尤推定で 30 エピソードの平均を用いたのは，提案手法では重点サンプリングが親個体によって蓄えられた 30 エピソードの経験から子個体の評価値を計算していることに対応している．

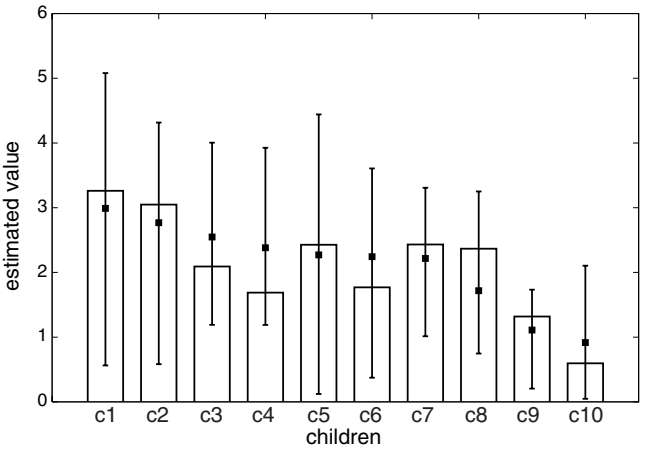

Fig.6: The children's values estimated by importance sampling and by maximum likelihood estimate method. The error-bars show the average rewards and the max/min rewards in 30 episodes, the boxes show the rewards estimated by importance sampling.

Fig.6 によると，重点サンプリングによる推定値は最尤推定値の最大値と最小値の間にあるので，真の評価値をうまく推定できていると言える．提案手法では生存選択に子個体の推定評価値のランクに対してルーレット選択を行なっている．したがって，たとえ多少の誤差が重点サンプリングの推定値に含まれていたとしても，それに起因した提案手法の著しい性能低下はないと考えられる．

また，naive-GA-2 は 600 ステップの経験から評価値を決定しているので，その精度はGA-IS と同じかそれ以上に良い．しかし，Fig.5 から，naive-GA-2 は 30000 ステップ以内で収束できないことが分かる．30000 ステップ以降も実験を続けた結果，naive-GA-2 は約 1300000 ステップで収束した．これは提案手法の約 130 倍の時間を要していることになる．GA-IS はnaive-GA-2 と同じ量の経験を仮想的にではあるが利用できる．この点は提案手法の最も注目すべき特徴である．

集団サイズ (N)，生成子個体数 (C)，エピソード長 (M) について考える．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．集団サイズが大きいほど，重点サンプリングの推定精度は向上するが，初期の全親集団による環境とのインタラクションにより多くの時間を要する．

C は GA の探索能力に影響する．生成子個体数が大きいほど，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 (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．

M は個体の評価値の精度に影響する．一般に重点サンプリングでは，エピソード長が大きいほど，個体の評価値の精度は向上するが，エピソードの生起確率の計算が不安定になる．実際，匍匐ロボットの前進タスクでは，予備実験から約 30 ステップがエピソード長の限界であることを確認している．より長いエピソードに対応するには，経験をエピソード単位で扱うのではなく，より細かい単位で扱うなどの工夫が必要である．

6 おわりに

本論文では，GA による政策学習において，重点サンプリングを用いて集団の経験を再利用することで，環境とのインタラクションを行なうことなく子個体の評価値を推定し，学習を高速化する手法を提案した．多点探索法である GA は複数の政策を保持しているために，従来から政策学習には不向きとされてきた．しかし，提案手法では複数政策下での経験を用いる重点サンプリングと組み合せることで，子個体の評価が高速になり，多峰性に対応できるという GA のロバスト性を生かした現実的な時間での政策学習が可能になった．

提案手法を匍匐ロボットに実装し，前進動作の学習に適用した結果，ナイーブな GA と比べて最終性能を低下させずに約 3 倍高速な学習を達成した．これは提案手法の有効性を示すものであり，提案手法により従来困難であった GA による政策学習の敷居が低くなったと言える．

実験で用いたタスクでは，重点サンプリングの計算上の制限からエピソード長を 20 ステップとしたが，実問題に適用するためには，より長いエピソードへの対応は必須である．また，より高次元のタスクへの対応も必須である．今後はこれらの点に対応することを考えている．

参考文献

[Shelton 2001] Shelton, C.R.: Policy Improvement for POMDPs using Normalized Importance Sampling, Proceedings of the 17th Conference on Uncertainty in Artificial Intelligence (UAI2001), 496 ／ 503 (2001).

[Precup 2000] Precup, D., Sutton, R.S., and Singh, S.: Eligibility Traces for Off-Policy Policy Evaluation, Proc. 17th International Conf. on Machine Learning (ICML2000), 759 ／ 766 (2000).

[Precup 2001] Precup, D., Sutton, R.S., and Dasgupta, S.: Off-Policy Temporal-Difference Learning with Function Approximation, Proc. 18th International Conf. on Machine Learning (ICML2001), 417 ／ 424 (2001).

[Goldberg 1989] Goldberg, D.E.: Genetic Algorithms in Search, Optimization, and Machine Learning, AddisonWsley Publishing Company Inc. (1989).

[Satoh 1996] Satoh, H., Yamamura, M. and Kobayashi, S.: Minimal Generation Gap Model for GAs considering Both Exploration and Exploitation, Proceedings of IIZUKA'96, 494 ／ 497 (1996).

[Kimura 2001] 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), 411 ／ 416 (2001).

[Kita 1998] Kita, H., Ono, I. and Kobayashi, S.: Theoretical Analysis of the Unimodal Normal Distribution Crossover for Real-coded Genetic Algorithms, Proc. 1998 IEEE ICEC, 529 ／ 534 (1998).

[Ono 1997] Ono, I. and Kobayashi, S.: A Real-coded Genetic Algorithm for Function Optimization Using Unimodal Normal Distribution Crossover, in Proc. 7th ICGA, 246 ／ 253 (1997).