行動が連続値の強化学習：確率的傾斜法による接近

○木村 元,　　小林 重信
東京工業大学 大学院総合理工学研究科

Reinforcement Learning for Continuous Action using Stochastic Gradient Ascent

○ Hajime Kimura,　　Shigenobu Kobayashi
Interdisciplinary Graduate School of Science and Eng., Tokyo Institute of Technology

Abstract: This paper considers a reinforcement learning (RL) where the set of possible action is continuous and reward is considerably delayed. The proposed method is based on a stochastic gradient ascent with respect to the policy parameter space; it does not require a model of the environment to be given or learned, it does not need to approximate the value function explicitly, and it is incremental, requireing only a constant amount of computation per step. We demonstrate the behavior through a simple linear regulator problem and a cart-pole control problem.

1　はじめに

本論文では，行動出力が連続値で報酬に遅れのある強化学習について考える．強化学習とは，試行錯誤を通じて未知なる環境へ適応していく機械学習の一つである．多くの実環境における応用では連続な行動出力や，しばしば連続値と離散値の混在する行動出力の扱いが求められるが，従来の代表的手法である Q-learning では，そのような行動に対する Q 値の関数近似や Q 値による行動選択は困難だった．本論文中ではエージェントの政策を行動出力の分布と定義し，政策改善アルゴリズムを示すことにより，連続な行動出力への新しい接近法について述べる．提案手法は政策のパラメータ空間における確率的傾斜法に基づいている．環境のモデルをあらかじめ与えたり学習したりする必要がなく，明示的に Value function を推定する必要もなく，1 ステップあたりの計算量が一定の逐次的な処理である．例題として線形の制御問題および倒立振子制御問題を取り上げ，提案手法の特徴や有用性について考察する．

2　従来の連続値行動強化学習

● Q-learning for LQR

[Bradtke 92] や [Baird 94] は線形 2 次形式問題への適用に限定された Q-learning を提案している．線形 2 次形式制御問題の Value function や Q 値は $V(x) = -k_2\, x^2$(ただし k_2 は何らかの正の定数)，$Q(x,a) = -(k_2 + k_1^2 k_3)x^2 - 2k_1 k_3 xa - k_3 a^2$ の 2 次関数形式で与えられる．よってこれらの Value は，2 次関数による関数近似と Q-learning を用いて推定を行い，最適政策は推定された Q 関数の微分がゼロになる点を計算することで容易に求められる．探査戦略として，最適と思われる行動の周辺を正規分布などの確率分布で振らせて出力することにより，連続値の行動を取り扱っている．

● Actor/Critic Algorithms

RFALCON[Lin et al. 96] は，[Barto et al. 83] の Actor/Critic アルゴリズムの Actor へファジィコントローラを適用した手法である．Actor の保持している現在の政策に関して，Critic では状態入力に対する割引報酬の期待値 V_t を TD 法により推定する．現在の政策で示された行動出力の中心値に対してある標準偏差のガウス分布で行動を出力し，その結果得る報酬 r と状態遷移先の割引報酬の見積もり値 $\hat{V}_{t+1}$ が $\hat{V}_t < r + \gamma \hat{V}_{t+1}$ だったら行動出力の中心値を実際に出力した値へ近付けるように Actor の政策を逐次的に改善していく．つまり，明示的に Value function を推定しながら確率的な山登り法で政策を改善する方法である．よって，政策すなわち状態観測から行動出力への写像のための関数近似とは別に，状態観測から Value 推定値への写像を行うための関数近似が必要である．そのため学習の性能は，Actor における政策の関数近似能力だけでなく，Critic における Value function の関数近似能力にも依存する問題点がある．

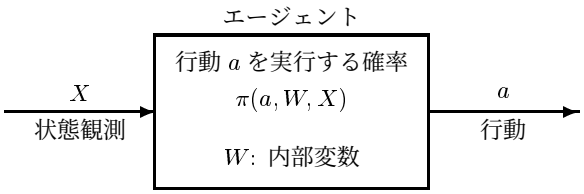

図 1: 確率的政策．エージェントはパラメータ W の調節によって政策 π を改善する．

3 確率的傾斜法による強化学習

エージェントはある時間ステップ t の観測 X_t において行動 a_t を選択する．その結果，環境は状態遷移を行い，エージェントは報酬 r_t を受け取り，次の時間ステップへ進む．エージェントの学習目標は，報酬獲得の合計を最大化すように，それぞれの観測において行動を選択する確率分布を形成することである．エージェントは環境の状態遷移規則や報酬についての知識をあらかじめ持っていないため，上記のような環境とのインタラクションを試行錯誤的に繰り返しながら，行動選択確率の学習を行う．本論文では，エージェントは各時間ステップ t において以下の割引報酬の合計 $V(x_t)$ を最大化するような行動 a_t を学習する．割引率を γ とすると，

$$V(x_t) = \sum_{t=\tau}^{\infty} \gamma^{t-\tau} r_\tau \ .$$

観測 X においてエージェントが行動 a を選択する確率を政策 π と呼び，関数 $\pi(a, W, X)$ で表す (図 1)．政策 $\pi(a, W, X)$ は行動 a の集合が連続値の場合は確率密度関数である．パラメータ W はエージェントの内部変数ベクトルを表す．エージェントは内部変数 W を調節することにより確率的政策 π を変えることができる．行動選択確率を表す機構が，例えばニューラルネットならば，内部変数 W はリンクの重み変数に相当し，重み付きのルールベースシステムならば，W はルールの重みに相当する．$\pi(a, W, X)$ の関数形については，エージェントに実装できる計算資源の制限など，一般に個別の問題ごとに制約が存在する．すなわちエージェントの構造および制約条件は $\pi(a, W, X)$ の関数形で規定される．

確率的傾斜法を用いた強化学習アルゴリズムの一般形を図 2 に示す [木村 96], [Kimura et al. 97]．W の任意の i 番目の要素を w_i と表す．手順 4 の $e_i(t)$ は適正度 (eligibility)[Williams 92] と呼ばれ，どんな行動をとったのかについての情報である．$D_i(t)$ は適正度の履歴 (eligibility trace) と呼ばれ，今までとった行動の履歴を記憶しているが，過去の行動ほど割引率 γ で減衰／忘却している．エージェントは報酬を受けとると，手順 5,6 にて履歴に記憶されていた行動の確率を高めるように W を更新する．このような処理を繰り返すと，報酬獲得に関係ない行動は打ち消され，報酬獲得に関係する行動だけが強化される．本手法は割引報酬の期待値を最大化する方向へと政策を確率的に逐次改善するものであるが，理論の詳細については [木村 96] を参照されたい．

1. 環境の観測 X_t を受けとる．
2. $\pi(a_t, W, X_t)$ の確率で行動 a_t を実行する．
3. 環境から報酬 r_t を受け取る．
4. 内部変数 W の全ての要素 w_i について以下の $e_i(t)$ と $D_i(t)$ を求める．ただし γ は割引率 $(0 \le \gamma < 1)$ である．

$$\begin{aligned} e_i(t) &= \frac{\partial}{\partial w_i} \ln\Big(\pi(a_t, W, X_t)\Big) \ , \\ D_i(t) &= e_i(t) + \gamma D_i(t-1) \ , \end{aligned}$$

5. 以下の式を用いて $\Delta w_i(t)$ を求める．

$$\Delta w_i(t) = (r_t - b)\, D_i(t) \ ,$$

   ただし b は定数である．
6. 政策の改善: 以下の式で W を更新

$$\begin{aligned} \Delta W(t) &= (\Delta w_1(t), \Delta w_2(t) \cdots \Delta w_i(t) \cdots) \ , \\ W &\leftarrow W + \alpha(1-\gamma)\, \Delta W(t) \ , \end{aligned}$$

   ただし α は非負の学習定数である．
7. 時間ステップ t を $t+1$ へ進めて, 1 へ戻る．

図 2: 確率的傾斜法による強化学習法の一般形

3.1 連続値行動出力のアルゴリズム

政策 $\pi(a, W, X)$ は行動 a の集合が連続値の場合は確率密度関数であることはすでに述べた．正規分布は連続値の確率変数のためのマルチパラメータの分布として単純なためよく知られている．正規分布は平均値 μ と標準偏差 σ の 2 つのパラメータを持つ．政策 π が式 1 に示す正規分布で与えられた場合，パラメータ μ と σ に関する適正度は以下のように計算される．

$$\pi(a, \mu, \sigma) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(\frac{-(a-\mu)^2}{2\sigma^2}\right) \qquad (1)$$

$$e_\mu = \frac{a_t - \mu}{\sigma^2} \qquad (2)$$

$$e_\sigma = \frac{(a_t-\mu)^2-\sigma^2}{\sigma^3}\ . \quad (3)$$

このような行動選択メカニズムは Gaussian unit と呼ばれる [Williams 92]．エージェントはランダムな探査の度合いを自分で制御するという特徴がある．式 2,3 においてパラメータ σ が分母となっていることより，σ が 0 へ近付くと適正度が発散することに注意しなければならない．適正度の発散はアルゴリズムの動作に悪い影響を及ぼす．この問題に対処するための一つの方法として，σ の値に応じて内部変数の更新のステップ幅を制御することが考えられる．更新のステップ幅を σ^2 に比例させると，適正度は以下のように計算される．

$$e_\mu = a_t-\mu \quad (4)$$
$$e_\sigma = \frac{(a_t-\mu)^2-\sigma^2}{\sigma}\ . \quad (5)$$

4 実験

4.1 実験設定: 線形 2 次形式制御問題

ベンチマークとして以下の線形 2 次形式制御問題 (linear quadratic regulator: LQR) を考える．ある時間ステップ t において，環境の状態はある連続値の実数 x_t にある．エージェントは同じく連続値の実数 a_t で表される行動を選択する．環境の状態遷移は以下で与える．

$$x_{t+1} = x_t + a_t + noise\ , \quad (6)$$

ただし $noise$ は標準偏差 $\sigma_{noise}=0.5$ の正規分布で与える．直接報酬は以下のように与える．

$$r_t = -x_t^2 - a_t^2\ . \quad (7)$$

エージェントの学習目標は以下の割引報酬の合計を最大化することである．

$$\sum_{t=0}^{\infty}\gamma^t r_t\ , \quad (8)$$

ただし γ $(0\le\gamma<1)$ は割引率である．本問題は線形 2 次形式制御問題であることより，最適な制御規則を解析的に求めることができる．Riccati 方程式より，最適レギュレータは以下の状態フィードバックで与えられる．

$$a_t = -k_1\,x_t \quad \text{, where}$$
$$k_1 = 1-\frac{2}{1+2\gamma+\sqrt{4\gamma^2+1}}. \quad (9)$$

最適な Value function は $V^*(x_t)=-k_2\,x_t^2$ ただし k_2 は何らかの正の定数の形式で与えられる．本実験では，遷移可能な状態空間は $[-4,4]$ に制限する．式 6 で示される状態遷移が上記の範囲を超える場合は，制限範囲までしか移動できないものとする．エージェントの行動についても同様に $[-4,4]$ の範囲外の行動を実行しても，環境では制限の範囲でしか実行されないものとする．

4.2 エージェントの実装

●確率的傾斜法の実装：

エージェントは μ と σ を何らかの決定的な方法にて計算し，平均値 μ 標準偏差 σ の正規分布に従って行動を出力する．エージェントは 2 つの内部変数 w_1, w_2 を持ち，これを用いて μ と σ を以下のように計算する．

$$\mu = w_1\,x_t \quad , \quad \sigma = \frac{1}{1+\exp(-w_2)}. \quad (10)$$

このとき，w_1 はフィードバックゲインとして見ることができる．σ を上記のように計算するのは，σ の値が負になるのを防ぐためである．エージェント内部変数 w_1 と w_2 に対応する適正度をそれぞれ e_1, e_2 と表す．式 4,5 より，適正度 e_1, e_2 は以下に与えられる．

$$e_1 = e_\mu\frac{\partial}{\partial w_1}\mu = (a_t-\mu)\,x_t \quad (11)$$
$$e_2 = e_\sigma\frac{\partial}{\partial w_2}\sigma$$
$$= ((a_t-\mu)^2-\sigma^2)(1-\sigma)\ . \quad (12)$$

●比較対象の Actor/Critic アルゴリズムの実装：

対等な条件での比較を行うため，Actor には確率的傾斜法と同じように式 10 で示す方法で政策を保持し，行動選択を行う．Critic は TD(0) 法を用いて Value function を推定する．状態観測の空間を格子状に分割して離散化し，それぞれのグリッドに対して Value を学習する．本実験では $-4\le x\le 4$ の状態空間を 3 等分した場合と 10 等分した場合について示す．Critic では観測入力 x_t に対する $V(x_t)$ を推定する．報酬 r_t

を受け取ると，TD-error $=r_t+\gamma\hat{V}(x_{t+1})-\hat{V}(x_t)$ を計算する．ただし $\hat{V(x)}$ は Critic が出力した Value の推定値である．Critic は TD-error を用いて TD(0) 法により Value を更新する．Actor は TD-error と式 11, 12 を用いて以下のように更新する．

$$\begin{aligned} w_1 &= w_1+\alpha\times(\text{TD-error})\times e_1 \\ w_2 &= w_2+\alpha\times(\text{TD-error})\times e_2 \end{aligned}$$

エージェントのパラメータは，学習係数 $\alpha=0.01$, 報酬基底 $b=0$ に設定し，w_1 の初期値は 0.35 ± 0.15 の範囲内でランダムに初期化，w_2 は 0 すなわち $\sigma=0.5$ に初期化した．Actor/Critic の TD(0) の学習率は 0.2 とした．

4.3 実験結果

図 3,4 は LQR 問題において割引率が 0.9 の場合の 100 試行の結果を示す．
フィードバックゲインに相当する w_1 の値は最適解の周辺で浮遊する傾向がある．
標準偏差 σ は学習の結果，初期値よりは減少するが，0.2 のあたりで停滞する．以下に考察するが，これらはそれほど悲観的な結果ではない．

収束時の定常的な挙動の考察:
図 5 は式 7,8 で定義される Value function を μ と σ で張られるパラメータ空間で表示したものである．最適解の周辺ではほとんど平らになっていることより，確率的傾斜法のエージェントは 1 次元の本 LQR 問題において，だいたいよい政策を得たと結論できる．

状態入力を 3 分割する TD(0) 法を Critic として用いた Actor/Critic では，分散は小さいが最適値へは収束できなかった．状態入力を 10 分割した場合には，分散はかなり小さく，最適点付近まで近付いているが，最適点から外れた値へ収束している．これより，Actor/Critic では Actor による政策関数の近似能力だけでなく，十分な Critic の関数近似能力も求められることが分かる．

学習途中の挙動の考察:
図 6 はフィードバックゲインの初期値を 1.0 として学習した場合の 10 試行それぞれのフィードバックゲインの変化の様子を表す．系が発散するダイナミクスとなるような学習の初期値となっており，試行のうちの半数近くがランダムウォークによって最適解

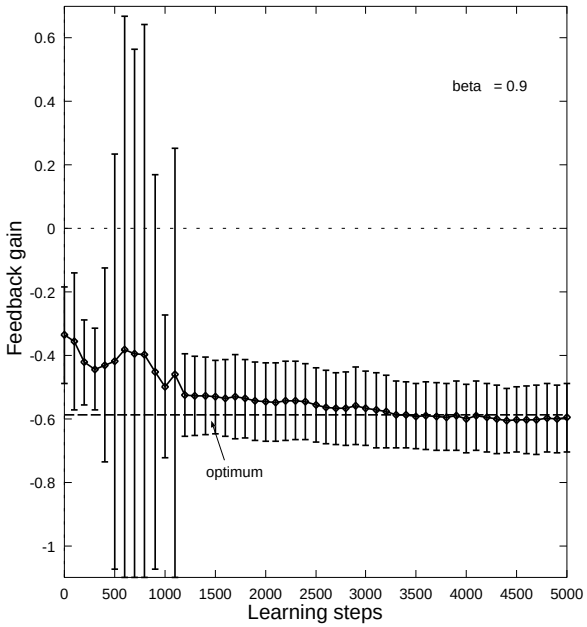

図 3: 確率的傾斜法による LQR 環境の学習結果．割引率 γ =0.9，横軸は学習ステップ，縦軸は獲得したフィードバックゲイン w_1 をあらわす．

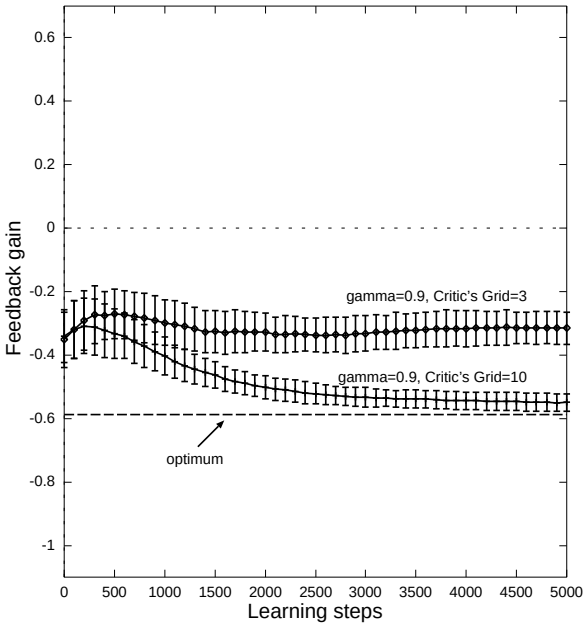

図 4: Actor/Critic による LQR 環境の学習結果．Critic の関数近似方法として，3 分割と 10 分割の 2 種類について示す．割引率 γ =0.9，横軸は学習ステップ，縦軸は獲得したフィードバックゲイン w_1 をあらわす．

から遠ざかる方向へと学習してしまうことが分かる．図5を見るとフィードバックゲインが1.0の周辺から先ではValue functionがほぼ完全に平らになっており，最適点方向への傾斜がないためにランダムウォークしてしまうと考えられる．よって確率的傾斜法を用いる場合にはエキスパートの知識を併用してある程度最適点の近くから学習を開始することが必要と考えられる．これに対して，Actor/Criticでは分散が小さいという特徴がある．

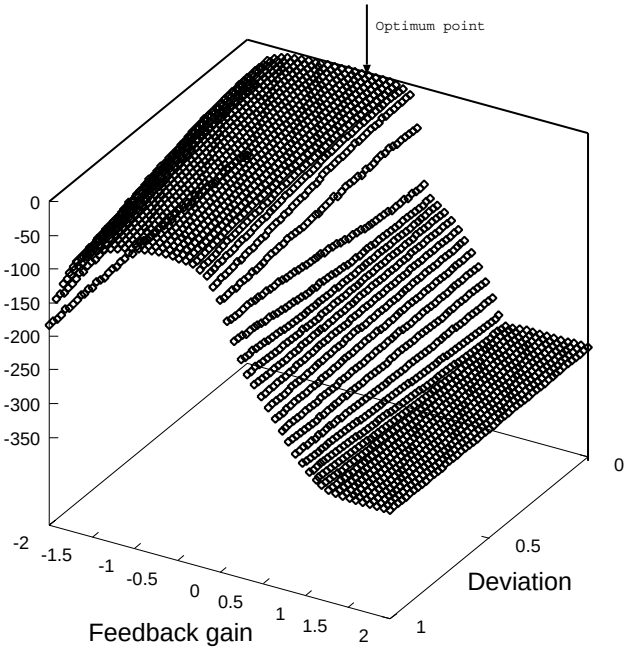

図5: LQR環境のvalue functionの形状．割引率$\gamma = 0.9$．最適点$\mu = -0.5884$, $\sigma = 0$付近はほとんど平らであることが分かる．

5　倒立振子制御問題への適用

提案手法の有用性について示すため，多次元の状態観測を必要とする非線形・非2次形式の倒立振子制御問題(図7)へ適用する．[Barto et al. 83]の実験設定を参考にしたが，彼らは行動を離散値としていたので，いくつか修正を加え，行動を連続値として計算機シミュレーションを行った．

5.1　実験設定

台車の質量M，ポールの質量m，ポールの長さ2ℓ，重力加速度g，台車に加えられる力F，台車の摩擦係数μ_c，ポールの摩擦係数μ_pとすると，図7の倒立振子のダイナミクスは以下で表される．

$$\ddot{\theta} = \frac{g\sin\theta + \cos\theta\left(\frac{-F - m\ell\ddot{\theta}^2\sin\theta + \mu_c sgn(\dot{x})}{M+m}\right) - \frac{\mu_p\dot{\theta}}{m\ell}}{\ell\left(\frac{4}{3} - \frac{m\cos^2\theta}{M+m}\right)}$$

$$\ddot{x} = \frac{F + m\ell\left(\dot{\theta}^2\sin\theta - \ddot{\theta}\cos\theta\right) - \mu_c sgn(\dot{x})}{M+m}$$

本実験では$\Delta t = 0.02$secの離散時間システムとして近似し，$M = 1.0$(kg)，$m = 0.1$(kg)，$2\ell = 1.0$(m)，$g = 9.8(m/sec^2)$，$\mu_c = 0.0005$，$\mu_p = 0.000002$とした．エージェントは$(x, \dot{x}, \theta, \dot{\theta})$を観測し，行動として台車に加える力$F$を出力する．エージェントは行動出力として任意の連続値をとることができるが，$F = \pm 20$(N)の範囲を超える場合には，環境はこの制限を超える分を無視して行動を実行する．環境は$(x, \dot{x}, \theta, \dot{\theta}) = (0, 0, 0, 0)$の初期状態から始まる．ポールの角度$\theta$が$\pm 12$度を超えるか，または台車の中心$x$が$\pm 2.4$の範囲からはみ出すと，環境からエージェントへ$-1$の報酬が与えられ，環境は初期状態へ戻る．

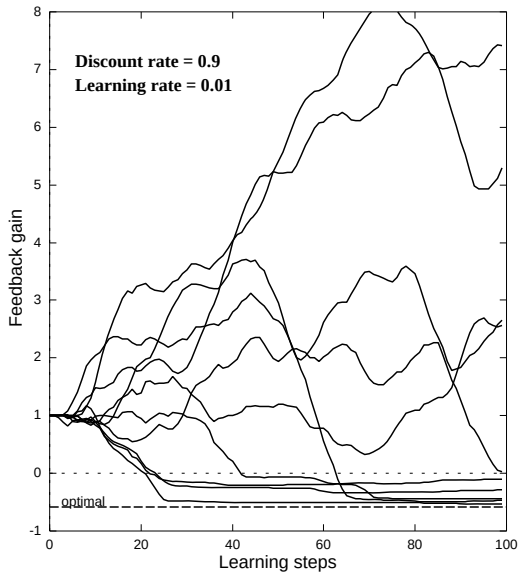

図6: 割引率γ=0.9で学習した場合の10試行それぞれのフィードバックゲインの変化．初期値は1.0から学習を始めた．

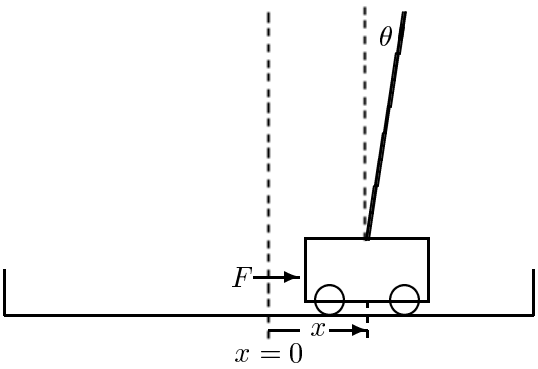

図 7: 倒立振子制御問題.

●確率的傾斜法の実装：

本実験では $(x, \dot{x}, \theta, \dot{\theta}) = (\pm 2.4 \text{ m}, \pm 2 \text{ m/sec}, \pm\pi \times 12/180 \text{ rad}, \pm 1.5 \text{ rad/sec})$ の状態空間を定義して正規化する．エージェントは5つの内部変数 $w_1 \cdots w_5$ を持ち，これを用いて μ と σ を以下のように計算する．

$$\begin{aligned} \mu &= w_1 \frac{x_t}{2.4} + w_2 \frac{\dot{x}_t}{2} + w_3 \frac{\theta_t}{12\pi/180} + w_4 \frac{\dot{\theta}_t}{1.5} \quad , \\ \sigma &= 0.1 + \frac{1}{1+\exp(-w_5)}. \end{aligned} \tag{13}$$

LQR の例題の場合と同様にして，適正度 $e_1, \cdots, e_5$ は以下に与えられる．

$$\begin{aligned} e_1 &= (a_t - \mu)\, x_t \ , \ e_2 = (a_t - \mu)\, \dot{x}_t \\ e_3 &= (a_t - \mu)\, \theta_t \ , \ e_4 = (a_t - \mu)\, \dot{\theta}_t \\ e_5 &= ((a_t - \mu)^2 - \sigma^2)(1 + 0.1 - \sigma) \ . \end{aligned}$$

これらを用いて図 2 の手順で政策を学習する．

●比較対象の Actor/Critic アルゴリズムの実装：

対等な条件での比較を行うため，Actor には確率的傾斜法と同様に式 13 で示す方法で政策を保持し，行動選択を行う．Critic は状態観測の空間を格子状に分割して離散化し，それぞれのグリッドに対して TD(0) 法を用いて Value を推定する．本実験では正規化された状態空間を各軸に対して 3 等分，つまり $3^4 = 81$ 個の矩形に分割した場合について実験を行う．割引率 $\gamma = 0.95$ に設定，それ以外のパラメータ等は LQR の実験と同じである．

実験結果を図 8 に示す．1 trial は初期状態からポールや台車が許容範囲をはみ出すまでを表す．Actor/Critic では全く学習できないのに対し，政策を保持するための関数近似が全く同じである確率的傾斜法では学習できた．

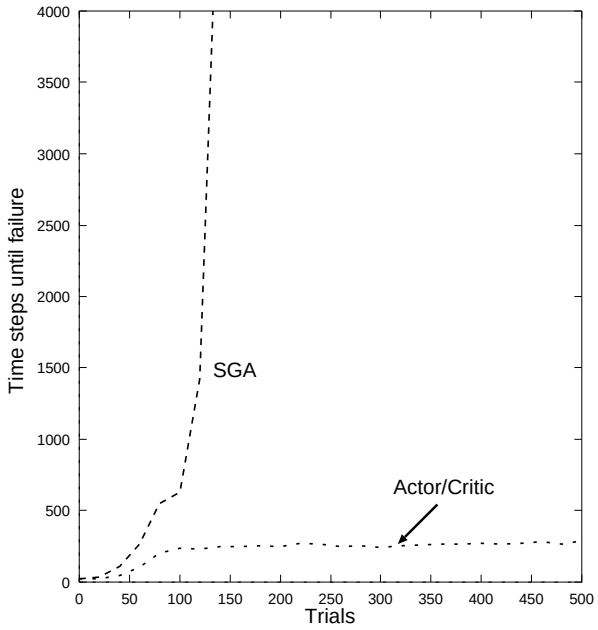

図 8: 倒立振子問題への適用結果．100 試行の平均.

6 おわりに

本論文では，行動出力が連続値で報酬に遅れのある強化学習問題に対し，確率的傾斜法を用いた強化学習アルゴリズムによる接近法を示した．連続値の行動出力を扱える政策関数としてガウス分布を取り上げ，確率的傾斜法による学習アルゴリズムを実装し，実験により特徴と有効性を示した．

参考文献

[Baird 94] Baird, L. C.: Reinforcement Learning in Continuous Time: Advantage Updating, *Proceedings of IEEE International Conference on Neural Networks*, Vol. IV, pp. 2448-2453 (1994).

[Barto et al. 83] Barto, A. G., Sutton, R. S. and Anderson, C. W.: Neuronlike Adaptive Elements That Can Solve Difficult Learning Control Problems,

[Bradtke 92] Bradtke, S. J.: Reinforcement Learning Applied to Linear Quadratic Regulation, *Advances in Neural Information Processing Systems 5*, (1992).

[木村 96] 木村 元, 山村 雅幸, 小林 重信: 部分観測マルコフ決定過程下での強化学習：確率的傾斜法による接近，人工知能学会誌, Vol.11, No.5, pp.761-768 (1996).

[Kimura et al. 97] Kimura, H., Miyazaki, K. and Kobayashi, S.: Reinforcement Learning in POMDPs with Function Approximation, *Proceedings of the 14th International Conference on Machine Learning*, pp. 152–160 (1997).

[Lin et al. 96] Lin, C. J. and Lin, C. T.: Reinforcement Learning for An ART-Based Fuzzy Adaptive Learning

Control Network, *IEEE Transactions on Neural Networks*, Vol.7, No. 3, pp. 709-731 (1996).

[Sutton 88] Sutton, R. S.: Learning to Predict by the Methods of Temporal Differences, *Machine Learning 3*, pp. 9-44 (1988).

[Watkins et.al 92] Watkins, C. J. C. H., & Dayan, P.: Technical Note: *Q*-Learning, *Machine Learning 8*, pp. 55-68 (1992).

[Williams 92] Williams, R. J.: Simple Statistical Gradient Following Algorithms for Connectionist Reinforcement Learning, *Machine Learning 8*, pp. 229-256 (1992).