2005－ＩＣＳ－139（２）
2005／3／14

Actor-critic法における 共分散を考慮した多次元正規分布による政策表現

阿部 哲† 上野 敦志†† 木戸出正継†

† 奈良先端科学技術大学院大学 情報科学研究科
〒 630-0192 奈良県生駒市高山町 8916-5
†† 大阪市立大学 大学院 工学研究科
〒 558-8585 大阪府大阪市住吉区杉本 3-3-138
E-mail: †{satosi-a,kidode}@is.naist.jp, ††ueno@info.eng.osaka-cu.ac.jp

あらまし 実世界中での行動学習問題は，入力である状態空間と出力である行動空間が共に連続空間である場合が多い．強化学習の一種である actor-critic 法は，連続状態行動空間を扱う問題にも適用が可能で，いくつかの研究が行われてきた．連続行動空間を扱う場合，行動を選択する確率分布 (政策) として一般に正規分布を用いる．エージェントは，環境との相互作用を通じて，適切な行動を選択できるように，正規分布の平均や標準偏差を調節する．従来手法は簡単化のために，各次元毎に独立した正規分布を用いる．しかしマニピュレータの軌道計画問題やロボットの歩行制御問題などの実問題は，各出力が協調して動かなければならない．従来手法は出力間の相関関係を考慮できないため，協調行動の学習が困難となったり，学習に時間がかかったりする問題が考えられる．そこで本稿では，共分散を考慮した多次元正規分布を政策表現に用いた actor-critic 法を提案し，学習の高速化と性能向上を目指す．本手法の有効性を検証するために，マニピュレータの軌道計画問題を取り上げる．

キーワード 強化学習，actor-critic 法，多次元正規分布，マニピュレータ

Stochastic Policy Representation Using a Multidimensional Normal Distribution for Actor-critic Methods

Satoshi ABE†, Atsushi UENO††, and Masatsugu KIDODE†

† Nara Institute of Science and Technology
8916-5 Takayama, Ikoma, Nara, 630-0192, Japan
†† Osaka City University
3-3-138 Sugimoto, Sumiyoshi, Osaka, Osaka, 558-8585, Japan
E-mail: †{satosi-a,kidode}@is.naist.jp, ††ueno@info.eng.osaka-cu.ac.jp

Abstract Actor-critic methods, which is one of reinforcement learning methods, is applied to that problems easily, and has left many achievements. Generaly, normal distribution has been used as probability distribution on which agent selects action. Agent renews means and standard deviation through policy parameter for selecting appropriate action intercting with environment. Under assumption that output dimensions are individual, conventional methods use normal distribution. Problems, such as trajectory planning of manupulator, and robot walking control etc., every output must cooperate with each other. Conventional methods cannot make consideration correlation, so it takes long time to get policy selecting action cooperately and being high performance. In this paper, we aim that learning speed up and improvement performance by adopting multivariate normal distribution with variance and covariance matrix into probability distribution selecting action. we have some experiments to demonstrate availability of this method by trajectory planning of manipulator.

Key words reinforcement learning, actor-critic methods, multidimensional normal distribution, manipulator

1. はじめに

近年ロボット自身が制御規則を獲得する手法として，強化学習が注目されている．強化学習とは，手本となる出力パターンを予め用意すること無く，問題の解けた度合を報酬として与えることで，エージェント（ロボット）が試行錯誤のうちに，目標とする出力を獲得する学習の枠組みである．エージェントは，報酬を多く獲得できるように，状態観測から行動出力へのマッピングである政策を獲得していく．エージェントは試行錯誤を通じて学習するため，理論的な解法を得ることが困難な問題においても，目標出力を発見する可能性がある．

強化学習が対象としてきた実問題として，ロボットの歩行制御問題やマニュピュレータの軌道計画問題などがある．これらの問題は，入力である状態と出力である行動が共に連続空間である．通常は状態行動空間を離散化してから，Q-learning などの離散空間を扱う手法を適用する．しかし人のような滑らかな動きを実現しようとすると，空間を離散化するのではなく，連続空間のままで扱いたい．このような連続状態行動空間を扱う手法として，actor-critic 法がある．actor-critic 法は，政策に従って行動を選択する actor と，状態の評価値を推定する critic の 2 つの部分で構成される．actor-critic 法はもともと，離散行動空間を扱う手法として提案された．actor の政策関数に連続確率密度関数を用いることで，連続行動空間にも適用が可能となる．連続行動空間を扱う場合，actor の政策関数に正規分布を用いることが一般的である．

行動学習問題は，出力間の協調を必要とする問題が非常に多い．しかし従来の actor-critic 法は，出力次元間の関係を考慮せず各次元毎に独立した正規分布を用いる．そのため，協調行動の学習が困難になったり，学習に時間がかかったりする問題が考えられる．

そこで本稿では，共分散を考慮した多次元正規分布を政策表現に用いた actor-critic 法を提案し，学習の高速化と学習の性能向上を目指す．行動学習問題としてマニピュレータの軌道計画問題を取り上げ，手法の有効性を示す．

2. 強化学習

強化学習では，図 1 に示すようなエージェントと環境のやりとりが行われる．エージェントは，累積報酬（将来得られるであろう報酬の総和）の最大化を目的として，状態から行動への確率分布である確率的政策を獲得していく．累積報酬 R_t は，無限時間先まで考慮するために，割引率 $\gamma(0 \leqq \gamma \leqq 1)$ を用いて式 (1) で表される．

$$R_t = \sum_{k=0}^{\infty} \gamma^k r_{t+k} \tag{1}$$

強化学習では，状態 s_t の評価値 $\tilde{V}(s_t)$ を累積報酬の期待値 $E\{R_t\}$ として，式 (2) で定義する．

$$\begin{aligned} \tilde{V}(s_t) &= E\{R_t|s_t\} \\ &= E\{\sum_{n=0}^{\infty} \gamma^n r_{t+n}\} \end{aligned} \tag{2}$$

ステップ 1 エージェントは時刻 t において，環境から状態 s_t を観測し，状態に応じた意志決定を行い，行動 a_t を出力する．
ステップ 2 エージェントの行動により，環境は状態 s_{t+1} へ遷移し，その遷移に応じたスカラー値の報酬 r_t を与える．
ステップ 3 時刻 $t+1$ に進み，ステップ 1 へ戻る．

図 1 強化学習の枠組み

3. Actor-critic 法アルゴリズム

図 2 に一般的な actor-critic 法の概要を示す．

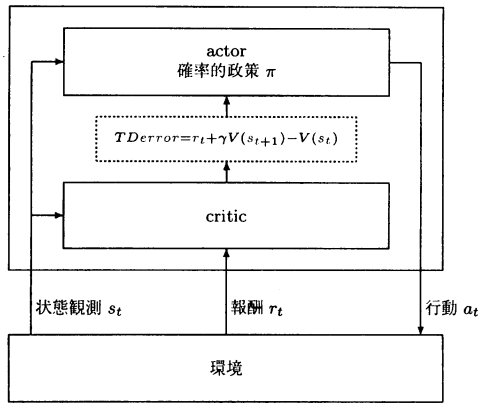


図 2 一般的な actor-critic 法の概要

actor は，状態から行動への確率分布である政策関数 π に従って行動を選択する．critic は，状態の評価値を推定する．エージェントは，報酬 r_t と状態の評価値の推定値 $V(s_t), V(s_{t+1})$ から，式 (3) で計算される $TD\ error$ を手がかりに，actor における政策関数と critic における状態の評価値の推定値を更新する．

$$TDerror = r_{t+1} + \gamma V(s_{t+1}) - V(s_t) \tag{3}$$

$TD\ error$ の値が正の場合，行動 a_t は予想より望ましい行動であったと判断される．逆に負の場合，行動 a_t は予想より望ましくない行動であったと判断される．

actor-critic 法は，もともと離散行動空間を扱う手法として提案された．政策関数に連続確率密度関数を用いることで，連続行動空間にも適用できる．連続行動空間に actor-critic 法を適用する場合，actor の政策関数に式 (4) の正規分布を用いることが一般的である．

$$p_\pi(a) = \frac{1}{\sigma\sqrt{2\pi}} \exp(\frac{-(a_n - \mu)^2}{2\sigma^2}) \tag{4}$$

ここで $n = 1, \cdots, N$ であり，N は行動空間の次元数を表す．actor は観測した状態 s_t と政策パラメータ $W = (w_1, \cdots, w_n)$ を用いて，正規分布の平均 μ と標準偏差 σ を求める．政策パラメータを調節することで，適切な行動を選択できるように平

(1) エージェントは環境において状態 s_t を観測する．
(2)actor は確率的政策 π に従い，行動 a_t を選択，実行する．
(3) 環境は現状態 s_t において実行された行動 a_t についての報酬 r_t を計算し，次状態 s_{t+1} へ遷移する．
(4) エージェントは環境より報酬 r_t を受け取り，critic へ渡す．
(5) エージェントは次状態 s_{t+1} を観測する．
(6)critic は現状態と次状態の評価値 $V(s_t), V(s_{t+1})$ を推定し，以下の $TDerror$ を計算する．

$$TDerror = r_t + \gamma V(s_{t+1}) - V(s_t)$$

ただし $\gamma(0 \leqq \gamma \leqq 1)$ は割引率を表す．
(7) エージェントは政策パラメータ w_i の適正度 e_{w_i} を計算する．

$$e_{w_i} = \frac{\partial}{\partial w_i} \ln\{\pi(a_t, s_t, W)\}$$

ただし $W = (w_1, \cdots, w_n)$ は政策パラメータを表す．
(8) エージェントは政策パラメータ w_i について適正度履歴 $D_{w_i}(t)$ を計算する．

$$D_{w_i}(t) = e_{w_i} + \beta D_{w_i}(t-1)$$

ただし $\beta(0 \leqq \beta < 1)$ は適正度の割引率を表す．
(9) 政策パラメータ w_i を更新する．

$$\Delta w_i = (TDerror) D_{w_i}(t)$$
$$w_i = w_i + \alpha_p \Delta w_i$$

ただし α_p は actor の学習定数を表す．
(10)$TDerror$ を用いて critic の評価値の推定値を更新する．
(11) ステップ (1) から繰り返す．

図 3　適正度履歴を用いた actor-critic 法アルゴリズム

均と標準偏差を調節することができる．

木村らは，actor の政策改善に適正度履歴 D_{w_i} を用いる手法を提案している [4]．式 (5) で計算される適正度 e_{w_i} は，π の政策パラメータ $w_i \in W$ を微小増加させた時に，行動 a_t を選択する確率が変化する向きと度合を表す値である．

$$e_{w_i}(t) = \frac{\partial}{\partial w_i} \ln\{\pi(a_t, s_t, W)\} \tag{5}$$

適正度 e_{w_i} の値が正の場合，政策パラメータ w_i を微小増加させると，行動 a_t の選択確率が高まることを表し，逆に負の場合，政策パラメータ w_i を微小増加させると，行動 a_t の選択率が下がることを表している．

適正度 e_{w_i} は，報酬 r_t を得た直前の行動 a_t の選択確率にのみ着目している．実際は行動 a_t を実行するまでの行動系列 $a_t, a_{t-1}, a_{t-2}, \cdots$ の選択確率の変化にも着目すべきである．そこで式 (6) で計算される適正度履歴 D_{w_i} を利用する．

$$D_{w_i}(t) = e_i(t) + \beta D_{w_i}(t-1) \tag{6}$$

適正度履歴は過去の行動に関する情報を圧縮したものと考えることが出来る．適正度履歴を用いた actor-critic 法のアルゴリズムを図 3 に示す．

従来の actor-critic 法は，出力次元間の関係を考慮せず，行動を選択する確率分布として各次元毎に独立した正規分布を用いてきた．しかしロボットの歩行制御問題やマニピュレータの軌道計画問題など，出力間の協調が重要である問題は数多くある．出力間の協調関係を考慮するためには，出力間の相関関係を考慮できる政策表現が望ましいと考えられる．そこで本稿では，共分散を考慮した多次元正規分布を政策表現に用いる actor-critic 法を提案する．次章で提案手法の説明をする．

4. 共分散を考慮した政策関数

提案手法では，政策関数に式 (7) で示される共分散を考慮した多次元正規分布を用いる．

$$f(x;\mu,\Sigma) = \frac{1}{(2\pi)^{\frac{n}{2}}|\Sigma|^{\frac{1}{2}}} \exp[-\frac{1}{2}(x-\mu)^T\Sigma^{-1}(x-\mu)] \tag{7}$$

ここで，μ は平均ベクトル，Σ は分散共分散行列を表す．分散共分散行列は正定値行列でなければならない．正定値行列とは，任意の x に関して $x^TAx \geqq 0$ を満たす行列 A のことである．行列 Σ が正定値行列であるための必要十分条件は，行列 Σ が $\Sigma = L \cdot L^T$ で表されることである．ただし，L は対角要素非零の下三角行列，L^T は行列 L の転置行列である．ここで新たに政策パラメータ $Z = (z_1, \cdots, z_m)$ を定義する．エージェントは分散共分散行列の要素を直接求めるのではなく，式 (8) のように，政策パラメータ Z を用いて下三角行列 L の各要素を求め，間接的に分散共分散行列を求めることとする．

$$\Sigma = \begin{pmatrix} \sigma_{11} & \cdots & \sigma_{n1} \\ \vdots & \ddots & \vdots \\ \sigma_{n1} & \cdots & \sigma_{nn} \end{pmatrix} = \begin{pmatrix} l_{11} & 0 & 0 \\ \vdots & \ddots & 0 \\ l_{n1} & \cdots & l_{nn} \end{pmatrix} \cdot \begin{pmatrix} l_{11} & \cdots & l_{n1} \\ 0 & \ddots & \vdots \\ 0 & 0 & l_{nn} \end{pmatrix} \tag{8}$$

平均ベクトル μ の要素 μ_i と下三角行列 L の要素 l_{ij} は，それぞれ政策パラメータ $w_{\mu_i,1}, \cdots, w_{\mu_i,u}$ ，$z_{ij,1}, \cdots, z_{ij,v}$ と状態 s_t から，式 (9) で計算される．

$$\begin{aligned} \mu_i &= f(s_t, w_{\mu_i,1}, \cdots, w_{\mu_i,u}) \\ l_{ij} &= g(s_t, z_{ij,1}, \cdots, z_{ij,v}) \end{aligned} \tag{9}$$

ここで関数 f，g には線形関数やシグモイド関数などを用いる．政策関数の更新には適正度履歴を用いる．適正度の算出法は式 (10) である．

$$\begin{aligned} e_{w_{\mu_i,j}} &= \frac{\partial \ln(\pi)}{\partial w_{\mu_i,j}} \\ &= \frac{\partial \ln(\pi)}{\partial \mu_i} \cdot \frac{\partial \mu_i}{\partial w_{\mu_i,j}} \\ e_{z_{ij,k}} &= \frac{\partial \ln(\pi)}{\partial z_{ij,k}} \\ &= \frac{\partial \ln(\pi)}{\partial l_{ij}} \cdot \frac{\partial l_{ij}}{\partial z_{ij,k}} \\ &= \Big(\sum_{1 \leqq l \leqq n, 1 \leqq m \leqq l} \frac{\partial \ln(\pi)}{\partial \sigma_{lm}} \cdot \frac{\partial \sigma_{lm}}{\partial l_{ij}} \Big) \cdot \frac{\partial l_{ij}}{\partial z_{ij,k}} \end{aligned} \tag{10}$$

行動を選択する確率分布に共分散を考慮したことにより，次の利点が得られると期待できる．

- 学習の高速化

出力間の相関関係を考慮することで，高い報酬が得られる行動

を高確率で選択可能な，確率分布を表現できるようになる．このように行動選択の分布の表現力が増えたこで，探索に指向性を持たせることができ，良好な解の発見が速まると期待できる．

- 性能の向上

共分散を考慮した多次元正規分布の導入により，行動の評価値が高い領域に行動選択の分布を大きく重ねることができるようになる．良好な行動を高い確率で選択できるようになることで，タスク成功率が高まると期待できる．

5. マニピュレータの軌道計画問題

本章では，共分散を考慮した政策表現の効果を確認するために，提案手法をマニピュレータの軌道計画問題に適用する．

マニピュレータの軌道計画問題は，式 (11) の拘束のもと，マニピュレータの先端を初期地点からゴール地点まで移動させるための，関節モータの制御規則を獲得する問題である．

$$M(\theta)\ddot{\theta} + h(\theta,\dot{\theta}) + g(\theta) + C\dot{\theta} = u \tag{11}$$

ここで，$M(\theta)$ は $n \times n$ 慣性行列，$h(\theta,\dot{\theta})$，$g(\theta)$ はそれぞれ，コリオリ・遠心力，および重力を表す n 次ベクトルである．また C は各関節の粘性摩擦係数を要素に持つ対角行列であり，u は n 次の関節駆動トルクベクトルである．式 (11) は $2 \times n$ 次の非線形状態方程式である式 (12) に書き換えることができる．

$$\frac{d}{dt}\begin{pmatrix}\theta\\ \dot{\theta}\end{pmatrix} = \begin{bmatrix}\dot{\theta}\\ -M^{-1}(\theta)(h(\theta,\dot{\theta})+g(\theta)+C\dot{\theta})\end{bmatrix} + \begin{bmatrix}0\\ M^{-1}(\theta)\end{bmatrix}u \tag{12}$$

提案手法の有効性を検証するために，5.1 節の実験 1 では 3 リンクマニピュレータの軌道計画問題，5.2 節の実験 2 では 2 リンクマニピュレータと 3 リンクマニピュレータの軌道計画問題を取り上げる．

5.1 実 験 1

5.1.1 問 題 設 定

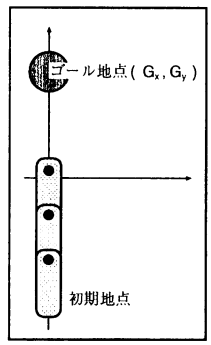
出力間の相関関係があまり強くない問題として，3 リンクマニピュレータの先端をゴールまで到達させる問題を取り上げる．環境からエージェントに与えられる報酬は式 (13) とする．

$$r_t = -\frac{\sqrt{(G_x - X(t))^2 + (G_y - Y(t))^2}}{MAX\_DISTANCE} \tag{13}$$

図 4 のように，(G_x, G_y) はゴール地点の中心座標，$(X(t), Y(t))$ は時刻 t におけるマニピュレータの先端の座標を表している．$MAX\_DISTANCE$ は，ゴール地点からマニピュレータの先端が到達できる最も遠い地点までの距離を表す．エージェントは毎ステップ報酬を得ることができる．

3 リンクマニピュレータの軌道計画問題において，下三角行列 L の各要素 l_{ij} は，式 (15) に $N = 3$ を代入し，式 (14) から算出する．

$$l_{ij} = \frac{3.0}{1.0 + \exp(-sum_{ij})} + 0.1 \quad (i = j)$$
$$l_{ij} = \frac{1.8}{1.0 + \exp(-sum_{ij})} - 0.9 \quad (i \neq j) \tag{14}$$



実験1

図 4 実験 1 のイメージ図

$$sum_{ij} = \sum_{l=1}^{N}\{z_{ij\,3l-2}\sin\theta_l + z_{ij\,3l-1}\cos\theta_l + z_{ij\,3l}\dot{\theta}_l\} \tag{15}$$

5.1.2 エージェントの実装

3 リンクマニピュレータにおいて，エージェントは $s = (\theta_1,\theta_2,\theta_3,\dot{\theta}_1,\dot{\theta}_2,\dot{\theta}_3)^T$ を観測し，出力として目標角 $a = (\tilde{\theta}_1,\tilde{\theta}_2,\tilde{\theta}_3)^T$ を出力する．目標角から，モータにかけるトルク値は式 (16) で計算される．

$$\tau_i = k_\tau(\tilde{\theta}_i - \theta_i) - b_\tau\dot{\theta}_i \tag{16}$$

5.1.3 結　　果

提案手法と比較する手法は，共分散を考慮しない actor-critic 法とする．式 (8) の下三角行列 L の対角要素以外を 0 にすることで，共分散を考慮できないようにする．

図 5，6 は，1 実験 100 トライアル（1 トライアル 3000 ステップ）の実験を 50 回行った結果の平均を示す．横軸はトライアル数，図 5 の縦軸はゴールまでの平均ステップ数，図 6 の縦軸はタスク成功率を表す．

図 5 から，提案手法は共分散を考慮していない場合より，学習に要したトライアル数が少ないことがわかる．また図 6 から，提案手法は共分散を考慮していない場合より，タスク成功率を高めることができたとわかる．actor に共分散を考慮した政策関数を用いたことで，行動選択の分布に指向性を持たせること

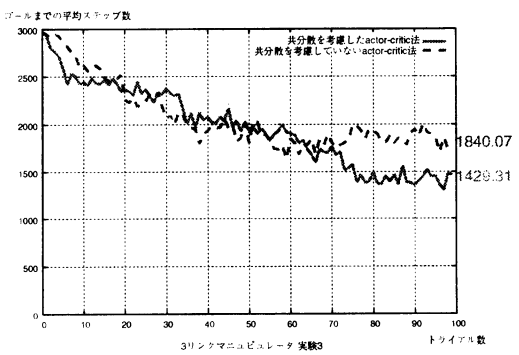


図 5 3 リンクマニピュレータ：ゴールまでの平均ステップ数

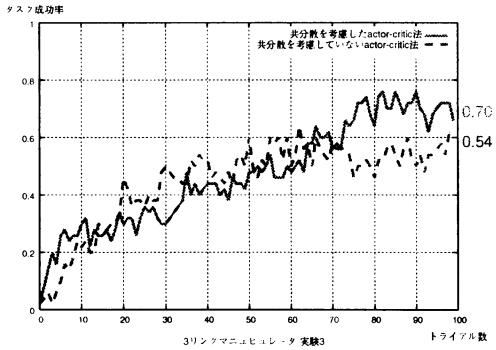


図 6　3 リンクマニピュレータ：タスク成功率

ができ，学習の高速化が実現できたと考えられる．また行動選択の分布の形状を変化させることができたため，評価が高い行動を高確率で選択できるようになり，性能が向上したと考えられる．

5.2 実　験　2

5.2.1 問題設定

出力間の相関関係が強い問題として，マニピュレータの先端を指定した軌道を通過させながらゴールへ到達させる問題を取り上げる．環境からエージェントに与えられる報酬は式 (17) とする．

$$\begin{array}{lll} if & \text{先端が指定軌道上} \rightarrow r_t = REWARD \\ else & \rightarrow r_t = 0 \end{array} \quad (17)$$

ただし $REWARD$ は以下の式とする．

$$REWARD = -\frac{\sqrt{(G_x - X(t))^2 + (G_y - Y(t))^2}}{MAX\_DISTANCE}$$

マニピュレータの先端は，図 7 のような軌道を通過しながら，ゴール地点へと向かう ((a) は 2 リンクマニピュレータ，(b) は 3 リンクマニピュレータの軌道を表す)．マニピュレータの先端が指定軌道上にある時にしか，報酬が与えられない．各出力が協調し合わなければ，マニピュレータの先端は指定軌道を通過することができない．

2 リンクマニピュレータの軌道計画問題において，下三角行列 L の各要素 l_{ij} は，式 (15) に $N = 2$ を代入し，式 (14) から算出する．3 リンクマニピュレータの軌道計画問題において，下三角行列 L の各要素 l_{ij} は，実験 1 と同様に算出する．

5.2.2 エージェントの実装

2 リンクマニピュレータにおいて，エージェントは $s = (\theta_1, \theta_2, \dot{\theta}_1, \dot{\theta}_2)^T$ を観測し，出力として目標角 $a = (\tilde{\theta}_1, \tilde{\theta}_2)^T$ を出力する．3 リンクマニピュレータにおいては，5.1.2 と同じである．目標角からモータにかけるトルク値は式 (16) と同じである．

5.2.3 結　　果

提案手法と比較する手法は，共分散を考慮しない actor-critic 法とする．式 (8) の下三角行列 L の対角要素以外を 0 にすることで，共分散を考慮できないようにする．

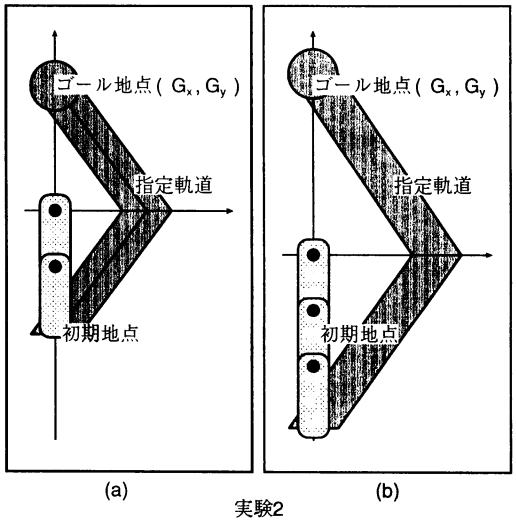


図 7　実験 2 のイメージ図

図 8, 9, 10, 11, 12 は，1 実験 100 トライアル (2 リンクの場合 1 トライアル 2000 ステップ，3 リンクの場合 1 トライアル 3000 ステップ) の実験を 50 回行った結果の平均を示す．横軸はトライアル数，図 8, 9 の縦軸はゴールまでの平均ステップ数，図 10, 11 の縦軸はタスク成功率，図 12 の縦軸はステップあたりの平均報酬を表す．

図 8, 9 から，提案手法は共分散を考慮していない場合より，学習に要したトライアル数が少ないことがわかる．また図 10, 11 から，提案手法は共分散を考慮していない場合より，タスク成功率を高めることができたとわかる．

実験 1 と同様に，学習の高速化と性能の向上が実現できた．

実験 1 に比べて実験 2 は出力間の相関関係が強い．実験 1 と実験 2 における提案手法と共分散を考慮していない場合との差が広がっていることから，提案手法は出力間に相関関係がある問題に特に有効であることがわかる．

また 2 リンクマニピュレータと 3 リンクマニピュレータの結

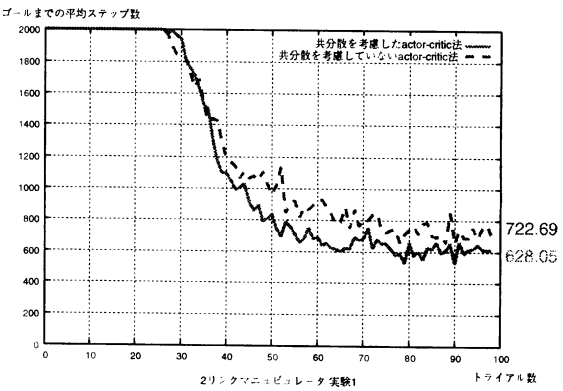


図 8　2 リンクマニピュレータ：ゴールまでの平均ステップ数

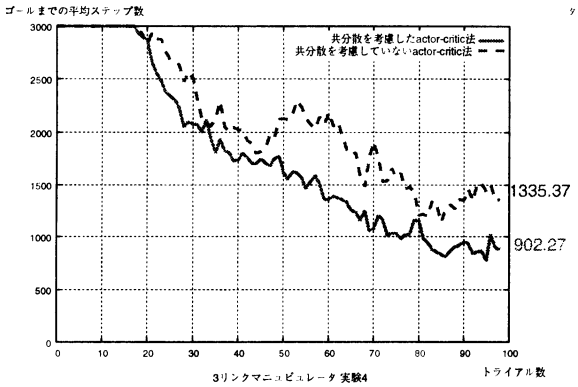


図 9　3 リンクマニピュレータ：ゴールまでの平均ステップ数

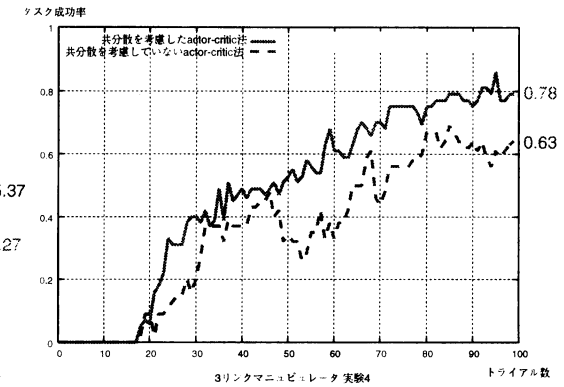


図 11　3 リンクマニピュレータ：タスク成功率

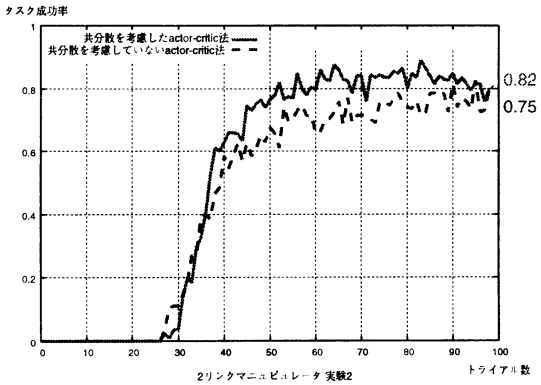


図 10　2 リンクマニピュレータ：タスク成功率

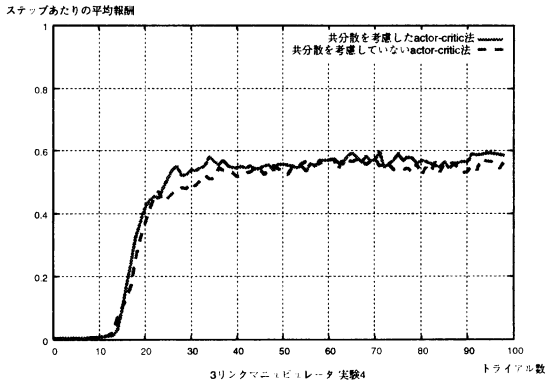


図 12　3 リンクマニピュレータ：ステップあたりの平均報酬

果から，行動空間の次元が増加すると，提案手法と共分散を考慮していない場合との差が大きくなっている．このことから共分散を考慮した政策関数は，行動空間の次元が大きくなった場合にも有効に働くことが期待できる．

図 12 は，3 リンクマニピュレータにおけるステップあたりの平均報酬を示す．学習の中盤以降，提案手法と従来手法ともに獲得する平均報酬には差がほとんど無い．しかし図 9 ではゴールまでに到達するステップ数に明らかな差が現れている．共分散を考慮していない場合，マニピュレータの先端を，そこそこ良い報酬が得られる領域に長くとどめようとする傾向が強いと考えられる．しかしながら提案手法は，更に速くマニピュレータの先端をゴールへ到達できるように，政策を更新しようとする傾向が強いと考えられる．

6. おわりに

本稿では，共分散を考慮した多次元正規分布を政策表現に用いた actor-critic 法を提案し，共分散を考慮しない場合と比較し有効性を確認した．出力間の協調が必要なマニピュレータの軌道計画問題に適用し，学習の高速化と性能の向上を実現できた．

今後の課題として，更に出力次元が大きくなった場合についての解析が必要である．また提案手法では，actor の政策関数に焦点を当てているが，critic における状態の評価値を推定する能力を高めることも重要である．actor と critic の両方を改善することで，従来手法では解くことが困難であった問題にも適用が可能になるか調べる必要がある．

文　献

[1] Gullapalli, V.: Associative Reinforcement Learning of Real-Valued Functions, COINS Technical Report 90-129(1990)

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

[3] 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)

[4] 木村 元, 小林重信: Actor に適正度の履歴を用いた Actor-Critic アルゴリズム, 人工知能学会論文誌, Vol.15, No.2, pp.267-275(2000)

[5] 森本 淳, 銅谷 賢治: 強化学習を用いた高次元連続状態空間における系列運動学習:起き上がり運動の獲得, 電子情報通信学会論文誌, D-II, Vol.J82-D-II, No.11, pp.2118-2131(1999)

[6] Pedro Martin, J. R. Millan: Reinforcement Learning of Sensor-based Reaching Strategies for a Two-Link Manipulator, IEEE

[7] Richard S. Sutton, Andrew G. Barto: 三上 貞芳・皆川 雅章 共訳: 強化学習, 森北出版株式会社 (1998)