行動確率場モデルに基づく強化学習 —拡張Q-学習—

榎田修一[†] 大橋健[†]
吉田隆一[†] 江島俊朗[†]

自律型ロボットの学習による行動獲得は，先見的な知識だけでは補いきれない行動決定の問題に対して有効な手法である．従来，センサ空間を離散化し，有限個の状態上での行動決定問題として定式化され，Q-学習など興味深い学習法が提案されて来た．しかし，離散化に伴う誤差が無視できない状況も多く，そのため誤差の影響を少なくする高精度の方法が研究対象になってきた．

本論文では，Q-学習を拡張した拡張Q-学習を提案する．拡張Q-学習とは，行動確率場モデルに基づき，センサ空間から行動空間への写像を導くものである．本モデルでは写像を表す行動選択確率を規定する行動価値関数は，有限個の基底関数の重み付き和ととして表される．学習は重みを調整する作業に対応し，また，精度を保持しつつより簡潔なモデルで関数近似を行うために基底関数の自律統合を学習アルゴリズムに追加した．

A Reinforcement Learning Based on Stochastic Field Model —Extended Q-learning—

Shuichi Enokida,[†] Takeshi Ohashi,[†] Takaichi Yoshida[†] and Toshiaki Ejima[†]

Reinforcement learning has been used as a method that is useful for an autonomous robot to select an appropriate action in each state with little or no premise knowledge. Typically, even if an autonomous robot has continuous sensor values, sensor space is quantized to reduce learning time. However, the reinforcement learning algorithms including Q-learning suffer from errors due to state space sampling.

To overcome the above, we propose Extended Q-learning (EQ-learning) based on Q-learning creates mapping that maps a continuous sensor space to a descrete action space. Through EQ-learning, action-value function approximation is represented by a summation of weighted base functions, and autonomous robot adjusts only weights of base functions by robot learning. In order to obtain a simpler learning model, other parameters are calculated automatically by unification of two similar base functions.

1. はじめに

自律型ロボットとは，センサ (感覚器) で環境を認識し，直面している状態に対して適切な行動を自ら決定し，エフェクタ (行動器) を通して環境に動作するものである．つまり，自律型ロボットとは，センサ空間から行動空間への適切な写像を知識としてもつものである．自律型ロボットがもつ知識を，表びき等のかたちで先見的に組み込むことも可能である．しかし，組み込まれた知識のみを持つロボットは，環境の変動に弱く，実環境では柔軟に対応できない．

一方，環境との相互作用によりロボットがとるべき行動を自律的に獲得する方法として，強化学習がある．学習による行動獲得は，理想化された環境と実環境とのギャップを埋め，先見的な知識だけでは補いきれない行動決定の問題に対して有効な手法である．学習による行動獲得は，ロボットが直面するあらゆる状態に対する適切な行動を，自らの経験を通して得るものである．強化学習においてロボットは，強化信号 (報酬) により得られる行動価値関数に基づき，直面している状態での行動を決定する．このとき，カメラ画像等の連続量のセンサ空間を持つロボットの学習では，センサ空間をそのまま状態空間と考えると状態が膨大な数

† 九州工業大学情報工学部
Faculty of Computer Science and Systems Engineering, Kyushu Institute of Technology

になり，学習モデルとしては現実的ではない．

そこで，ロボットのセンサ空間を適切に離散化した，状態空間の構築が求められてきた．しかし，遂行する仕事，ロボット，とりまく環境をうまくモデル化し，センサ空間の離散化を人為的に行うことによる人的コストの増大がおきる．また，センサ空間を離散化することにより，行動価値関数の近似に誤差が生じる．結果，学習により得られる行動も誤差を含んだものとなる．そのため，誤差の影響が少なく学習効率の良い手法の研究が盛んになってきた．Asadaらは，ロボットが自らの経験を通し，センサ空間の適切な離散化を行う手法を提案している[1]．また，Onoらは，複数の異なる離散化によって得られたセンサ空間のもとで，それぞれ学習を進め，それらの学習結果を統合して行動を決定する手法を提案している[5]．

本論文ではセンサ空間の離散化を行わずに，センサ空間をそのまま状態空間とする行動確率場モデルにより，センサ空間からK次元確率ベクトル空間への写像を求める手法を提案する（以降，本手法での状態とはセンサ入力を示す）．ここでのK次元ベクトル空間とは，ロボットが取り得るK個の行動それぞれに対する選択確率である．行動選択確率は連続量の行動価値関数により導かれるとする．行動価値関数を基底関数の重み付き和で表し，学習により簡潔で近似精度のよい基底関数の組合せを導き出すことを考える．ここで，近似関数を規定するものは，(1) 基底関数（ガウス関数，シグモイド関数等），(2) 基底関数の次数（総数），(3) 基底関数のパラメータ（重み，中心座標，幅）である．

従来の基底関数の重み付き線形和で関数近似を行う手法として，ニューラルネットワークでの学習[3]等があるが，それらの学習法は，重みの更新学習のために，ロボットの行動毎に強化信号を与えるものである．しかし，実際は，ロボットの行動毎に評価することが難しい状況も多く，強化信号を逐次与えることは困難である．本論文で提案する学習アルゴリズムは，一連の行動結果の善し悪しのみを評価する遅れのある強化信号に対応する報酬伝播の性質をもつ，Q-学習を拡張したものである．

提案する学習アルゴリズムでは，重みを最急降下法により調整する．基底関数としては局所性をもつ方形波関数とガウス関数に注目する．方形波関数を用いると，従来の離散状態での強化学習法と等価となり，ガウス関数を用いると，連続系の状態でのQ-学習となる．ガウス関数を用いることにより，離散化した状態での学習の際に発生していた誤差を吸収し，より高精度の学習，行動獲得が可能となる．

モデルの善し悪しを測る手法としてAIC[8]があるが，その評価基準はモデルと観測される実際の値との誤差，モデルの簡潔さの二点を持って与えられる．そこで，基底関数の類似に注目し類似するもの同士を統合することにより，学習データを反映した，適切な簡潔さを有するモデルを自律獲得することを目指す．このことにより，学習を行うときに，適切なモデルをタスク毎に構成する労力を削減できる．

本論文の提案手法の性能を評価するため，環境内に散在する餌を集める仕事をロボットに学習させるシミュレーション実験を行なう．ロボットのセンサはカメラを想定しており，連続系のセンサである．まず，連続系のセンサ空間において，基底関数としてガウス関数を用いることにより，方形波関数を用いた学習に比べ，誤差の影響の少ない，精度の高い学習が可能であることを示す．次に，基底関数の自律統合化の実験を行い，与えられた仕事，センサパラメータに対して適切な基底関数を自動獲得可能であることを示す．

2. 強化学習

強化学習とは，ロボットが取り得る様々な行動を選択し，結果として得た報酬なり罰なりのフィードバックを基に環境モデルを獲得し，各状態での最適行動を自動獲得する手法である．

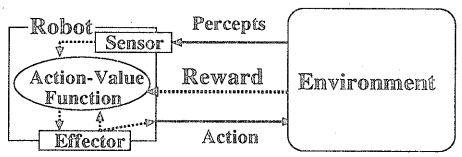


図1 強化学習
Fig. 1 Reinforcement Learning

いま，ロボットが状態$\boldsymbol{x}(t)$において$\boldsymbol{u}(t)$なる行動を取る．すると次の時刻にロボットが観測する状態$\boldsymbol{x}(t+1)$は以下の式に従う．

$$\boldsymbol{x}(t+1) = f(\boldsymbol{x}(t), \boldsymbol{u}(t)) \quad (1)$$

また，ロボットに入力される強化信号（報酬）$r(t)$は状態$\boldsymbol{x}(t)$と行動$\boldsymbol{u}(t)$の関数である．

$$r(t) = r(\boldsymbol{x}(t), \boldsymbol{u}(t)) \quad (2)$$

強化学習が想定する問題では，報酬は逐次与えられるものではなく$(r(t)=0)$ロボットが仕事を遂行した結果の状態に対して与えられる$(r(t)=1)$ことが多い．このとき，政策polが与えられたときある状態変数$\boldsymbol{x}(t)$に対する行動価値関数$V(\boldsymbol{x}(t), pol)$を，現状態から先に得られるであろう期待利得をもとにあらわす．

$$V(\boldsymbol{x}(t), pol) = r(t) + \gamma r(t+1) + \gamma^2 r(t+2) + \cdots + \gamma^n r(t+n) \quad (3)$$

ここで，γは減衰係数で$0 \leq \gamma \leq 1$の値を取り，遠い将来に得られる報酬は割り引いて考える．また，polはロボットがある状態のときにどのような行動を出力するかを各状態毎に決めたものである．つまり，状態から行動への写像ととらえることができる．

$$\boldsymbol{u}(t) = f(\boldsymbol{x}(t)) \quad (4)$$

単位時刻後の行動価値関数$V(\boldsymbol{x}(t+1), pol)$は，

$$V(\boldsymbol{x}(t+1), pol) = r(t+1) + \gamma r(t+2) + \cdots + \gamma^{n-1} r(t+n) \quad (5)$$

となるので，式(3)(4)から以下の関係が導かれる．

$$V(\boldsymbol{x}(t), pol) = r(t) + \gamma(V(\boldsymbol{x}(t+1), pol)) \quad (6)$$

このとき，強化学習の目的は行動価値関数$V(\boldsymbol{x}(t), pol)$を最大にする政策$pol$を獲得することである．一般的に，政策$pol$とはロボットが認識した状態$\boldsymbol{x}$において，ロボットが選択可能な$K$個の行動から適切な行動$a_i$を決定することである．つまり，状態$\boldsymbol{x}$とその状態での適切な行動$\mathcal{Q}$を対応付ける写像$g(\boldsymbol{x})$が学習で求めるべき政策である．

$$g(\boldsymbol{x}) \in \mathcal{Q} = \{a_1, a_2, \cdots, a_i, \cdots, a_K\} \quad (7)$$

一般的に，直接$g(\boldsymbol{x})$を求めるよりは，以下の式で示す行動選択確率$\mathbb{P}(\boldsymbol{x})$を求める方が行動選択にランダム性があり，局所最適からの脱出が可能であり学習時の環境全探査を進めやすく扱いやすい．

$$\mathbb{P}(\boldsymbol{x}) = (p(\boldsymbol{x}, a_1), p(\boldsymbol{x}, a_2), \cdots, p(\boldsymbol{x}, a_K)) \quad (8)$$

$$0 \leq p(\boldsymbol{x}, a_i) \leq 1, \qquad \sum_{i=1}^{K} p(\boldsymbol{x}, a_i) = 1 \quad (9)$$

しかし，行動選択確率$\mathbb{P}(\boldsymbol{x})$を学習時の報酬による強化値に基づき直接得るような学習は，確率がもつ性質，制約により窮屈なものとなる．ここで，下記の非線形変換を考え，状態$\boldsymbol{x}$で行動a_iを選択する確率$p(\boldsymbol{x}, a_i)$を求める問題を，状態$\boldsymbol{x}$で行動a_iをとる行動有用度$U(\boldsymbol{x}, a_i)$を求める問題に置き換える．

$$p(\boldsymbol{x}, a_i) = \frac{\exp(U(\boldsymbol{x}, a_i)/T)}{\sum_{j=1}^{K} \exp(U(\boldsymbol{x}, a_j)/T)} \quad (10)$$

$$\mathbb{U}(\boldsymbol{x}) = (U(\boldsymbol{x}, a_1), U(\boldsymbol{x}, a_2), \cdots, U(\boldsymbol{x}, a_K)) \quad (11)$$

ここでTは温度定数である．この確率分布に従ってステップ毎に疑似乱数を発生し，確率的に行動を選択する．任意の状態$\boldsymbol{x}$で次式を満たす行動有用度関数$U(\boldsymbol{x}, a_i)$を求めることができれば，式（10）によって最大確率を与える行動a_iを選択することが最適政策となる．

$$U(\boldsymbol{x}, a_i) = r(\boldsymbol{x}, a_i) + \gamma V(\boldsymbol{x}', pol_{opt}) \quad (12)$$

$$\boldsymbol{x}' = f(\boldsymbol{x}, a_i) \quad (13)$$

ここで，pol_{opt}は最適政策である．

2.1 Q-学　習

強化学習法として広く用いられている，Watkinsにより提案されたQ-学習[6]について述べる．Q-学習は，状態$s \in \mathcal{S}$で行動$a_i \in \mathcal{A}$を取り，報酬$r(s, a_i)$を受け，次状態s'に遷移したときに，行動有用度$Q(s, a_i)$を以下の式で更新する学習法である．

$$Q(s, a_i) \Leftarrow Q(s, a_i) + \alpha\{r(s, a_i) + \gamma M(s') - Q(s, a_i)\} \quad (14)$$

$$M(s) = \max_{j \in A} Q(s, a_j) \quad (15)$$

ここで，αは学習率，γは減衰係数である．

これは，現状態と次状態との行動有用度の差分，すなわち時間的差分を最小にするような更新式である．学習時の報酬は，仕事をうまく遂行できた目的状態（ゴール状態）に到達したときにのみ入力されるため，目的状態に到達する1ステップのみの行動有用度が強化される．そこで，行動有用度の時間的差分を最小にすることにより行動有用度が目的状態から徐々に伝播し，各状態での行動有用度が得られ，行動有用度関数の近似とする．

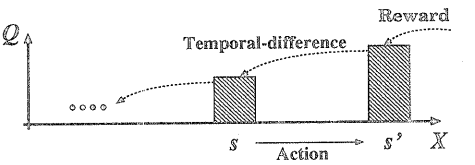


図2　時間的差分の最小化による報酬伝搬
Fig. 2　Minimize a temporal-difference to propagate a reinforcement signal

ここで，連続系のセンサ状態に対して，強化学習を適用することを考える．強化学習により，連続系の状態全てに対し行動有用度を決定することは，学習時間の爆発的な増大をまねき，非常に困難であり現実性に欠ける．そこで，一般的にQ-学習を用いるときは，感覚入力を離散化しラベル付けを行う．そして，学習のときに報酬が得られると同一ラベルのセンサ状態に対して一律に行動有用度の更新を行うことにより，学習時間の爆発を抑えている．

2.2 離散化による問題

従来のQ-学習法は，学習時間の爆発的増加を抑える為に，センサ空間を離散化し，ラベル付けを行なって，有限個の状態での行動決定問題に定式化してきた．

しかし，センサ空間の離散化を行うことで様々な誤差が発生することが考えられる．一つは，学習中に発生する誤差である．図3で示すように，同一ラベルの

状態に対して一様に行動有用度の強化を行うため，非常に偏ったものとなる．また，学習した結果得られる写像も，離散化された状態に依存しており，誤差が含まれたものとなる．

従来の Q-学習法でこの誤差を抑えるには，センサ状態の離散化を慎重に行う必要がある．つまり，十分に離散化を行なわないと行動の獲得精度に悪影響を及ぼす．しかし，強化学習の収束に必要な時間は，状態数に対し指数関数的に増加するため，不必要に細かな離散化は学習時間の増大を招く．

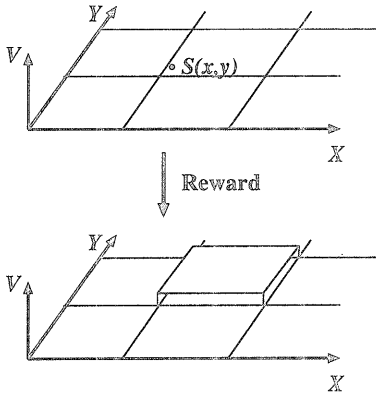


図 3 センサ空間の離散化による強化学習への障害
Fig. 3 Reinforcement Learning algorithms suffer from the quantization of the sensor space

3. 拡張 Q-学習

本論文では，行動確率場モデルに基づく拡張 Q-学習（Extended Q-learning．以降 EQ-学習）を提案する．本モデルではセンサ空間から K次元の確率ベクトルへの写像を学習により導く．行動選択確率 $\mathbb{P}(x)$ は式（10）により決定される．また，式（10）での行動有用度関数をセンサ空間で近似する際に，連続関数の近似に一般的に用いられるモデルである，基底関数の重み付き線形和を用いる[10)11)]．いま，行動価値関数を有限個の基底関数で近似しようとしたとき，モデルを規定するものを以下に示す．

1) 基底関数：ガウス関数，方形波関数，シグモイド関数，ウエーブレット関数等
2) 次数：1）で選択した基底関数の個数
3) 自由パラメータ:
   （a） 重み－$\{W_1, W_2, \cdots, W_N\}$ (個々の基底関数の背の高さ)
   （b） 配置－$\{\mu_1, \mu_2, \cdots, \mu_N\}$ (個々の基底関数の中心座標 (位置))
   （c） 幅－$\{\Sigma_1, \Sigma_2, \cdots, \Sigma_N\}$ (個々の基底関数の幅 (粒) の大きさ，粒度)

EQ-学習では，各パラメータの調整法を以下のようにする．

1) ガウス関数，方形波関数を用いる
2), 3)（b）（c） 基底関数の自律統合により求める
3)（a） 強化信号のフィードバック (学習) により調整する

3.1 連続状態での行動有用度関数

連続系での行動有用度関数 $U(x, a_i)$ を以下のように基底関数の重み付き線形和で表せるものとする．

$$U(x, a_i) = \sum_{m=1}^{N} W_m(a_i) B_m(x) \tag{16}$$

つまり，状態 x における行動 a_iの行動有用度 $U(x, a_i)$ を各行動に共通な N個の基底関数 $B_m(x)$ の重みつき線形和で表す．この行動有用度を用いて式（10）に従い確率分布を算出し，確率的に行動選択を行う．

従来の離散的状態空間での学習は，基底関数 $G(x)$ を以下の式で示す方形波関数（図 4）で表されるものを用いてきたと捉えることができる．

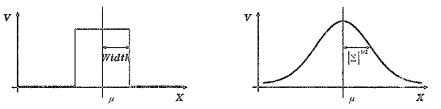
$$B_m(x) = \begin{cases} 1 & |\mu_m - x| \leq W \\ 0 & |\mu_m - x| > W \end{cases} \tag{17}$$

ここで，μ_mとは，基底関数 $B_m(x)$ の中心座標である．

そこで，EQ-学習は，従来の Q-学習の手法を含む一般的な枠組となるものを目指す．また，本論文では，基底関数 $B_m(x)$ をガウス関数で表すものを新たに提案し，従来法との比較を行う．

$$B_m(x) = \exp(-\frac{1}{2}(x - \mu_m)^t \Sigma_m^{-1}(x - \mu_m)) \tag{18}$$

ここで，Σ_mとは，基底関数 $B_m(x)$ の形状を表す幅の値である．



Rectangular function Gaussian function

図 4 行動確率場モデルが用いる基底関数
Fig. 4 Base functions in SFM

3.2 EQ-学習における重みの更新[7)]

本論文では，経験を通して強化信号を伝播し，期待利得が最大となるような行動を獲得する強化学習に基づき，重みの更新学習法について考える．特に，各状

態で個々の行動を選択したときの期待利得を逐次的に求めるQ-学習に焦点を当て，式（10）で得られる行動確率場モデルにより連続センサ空間でも働くように，学習アルゴリズムの拡張を図る．

先に示した，Q-学習による行動有用度の時間的差分を小さくする更新法を，行動確率場モデルに基づき一般化する．状態 $\boldsymbol{x}$ で行動 a_i を取り，報酬 $r(\boldsymbol{x}, a_i)$ を受け，次状態 $\boldsymbol{x}'$ に遷移したときに，行動 a_i がもつ価値の時間的差分 $E(\boldsymbol{x}, a_i)$ は以下の式であらわされる．

$$\begin{aligned} E(\boldsymbol{x}, a_i) &= \{r(\boldsymbol{x}, a_i) + \gamma M(\boldsymbol{x}') - U(\boldsymbol{x}, a_i)\}^2 \\ &= \{r(\boldsymbol{x}, a_i) + \gamma \max_{a_k \in A} U(\boldsymbol{x}', a_k) \\ &\quad - \sum_{m=1}^{N} W_m(a_i) B_m(\boldsymbol{x})\}^2 \end{aligned} \tag{19}$$

Nは基底関数の個数である．ここで，基底関数の持つパラメータを変更し，時間的差分を最小にすることを考える．

行動確率場モデルでの学習更新式を，最急降下法により導く．つまり，各パラメータの誤差微分値より，誤差が最も減少する方向を求める．重みのパラメータ数は基底関数の次数 Nである．また，中心座標$\boldsymbol{\mu}_m$，幅$\boldsymbol{\Sigma}_m$のパラメータ数は，センサ空間が d 次元であるときに，それぞれ $N \times d$ となる．重み・中心座標・幅の値全てを学習で操作すると $(2d+1)N$次のパラメータを更新することとなり，センサ空間の次元数に比例して増加する．しかし，学習時間の増加を抑えるため，学習により変更するパラメータは少ない程良い．そこで，センサ空間の次元数に影響されない重みのみを更新し，行動有用度関数の時間的差分を少なくする．

重み $W_j(a_i)$ の誤差微分値は，

$$\begin{aligned} \frac{\partial E(\boldsymbol{x}, a_i)}{\partial W_j(a_i)} &= -2B_j(\boldsymbol{x})\{r(\boldsymbol{x}, a_i) \\ &\quad + \gamma \max_{a_k \in A} U(\boldsymbol{x}', a_k) \\ &\quad - \sum_{m=1}^{N} W_m(a_i) B_m(\boldsymbol{x})\} \end{aligned} \tag{20}$$

である．よって，誤差が急激に減少する方向へ，各重みを少しずつ移動させる更新則は，

$$\begin{aligned} W_m(a_i) \Leftarrow W_m(a_i) &+ N_m(\boldsymbol{x})\{r(\boldsymbol{x}, a_i) \\ &+ \gamma M(\boldsymbol{x}') - U(\boldsymbol{x}, a_i)\} \end{aligned} \tag{21}$$

$$N_m(\boldsymbol{x}) = \alpha \frac{B_m(\boldsymbol{x})}{\sum_j B_j(\boldsymbol{x})} \tag{22}$$

$$M(\boldsymbol{x}) = \max_{k \in A} U(\boldsymbol{x}, k) \tag{23}$$

となる．この更新則に従って各重みを更新する．ここで，αは学習率，γは減衰係数である．基底関数として方形波関数を用いれば，先に示したQ-学習の更新規則と等価となる．

3.3 基底関数の自律統合

基底関数の絶対数を増加させることにより，行動価値関数の近似精度は良くなることが考えられるが，不必要にパラメータの多いモデルでは学習時間の増大を招く．また，モデルの善し悪しを評価するAICにおいても，同じ近似精度を持つモデルであれば，そのモデルの持つ自由パラメータが少ない程良いと評価される．そこで，EQ-学習では基底関数の自律統合により近似精度を保持しつつ，モデル内の自由パラメータ数の削減を試みる．

多数の基底関数を均一に配置する初期状態から学習を開始し，類似度の高い基底関数同士を統合し，最適な配置を自律的に獲得する．方形波関数は局所性の強い関数であるので，初期状態である基底関数を多数配置しての学習では，センサ空間全体に均一に入力が入るまでは基底関数の評価が困難である．そこで局所性の緩いガウス関数をもって類似度による統合を行う．自律統合を行うときの，類似度が高いとは以下に示す2つの要素が近いとする．

(1) 基底関数の中心座標での行動選択確率
(2) 基底関数の中心座標の位置

3.3.1 行動選択確率の類似度

確率分布の類似性をはかる基準として用いられている，Kullback-Leibler 情報量 (K-L 情報量) を用いる．いま，K-L 情報量は二つの離散確率分布モデル

$$\boldsymbol{p} = \{p_1, p_2, \cdots, p_N\} \tag{24}$$

$$\boldsymbol{q} = \{q_1, q_2, \cdots, q_N\} \tag{25}$$

の違いをあらわし，以下の式で計算される．

$$I(\boldsymbol{p}; \boldsymbol{q}) = E \log \frac{\boldsymbol{p}}{\boldsymbol{q}} = \sum_{i=1}^{N} p_i \log \frac{p_i}{q_i} \tag{26}$$

この値がモデル$\boldsymbol{q}$に関するモデル$\boldsymbol{p}$のK-L情報量である．また，K-L情報量は以下の式を満たす．

$$I(\boldsymbol{p}; \boldsymbol{q}) \geq 0 \tag{27}$$

ここで，$I(\boldsymbol{p}; \boldsymbol{q}) = 0$ となるのは，確率分布$\boldsymbol{p}$，$\boldsymbol{q}$が等しいときであり，0に近い程，確率分布の類似度が高い．

K-L 情報量を用いて行動選択確率の類似度とする．ここで，ある基底関数 $B_n(\boldsymbol{x})$，$B_m(\boldsymbol{x})$ の中心座標での行動選択確率は，式（10）からそれぞれ，

$$P(\boldsymbol{\mu}_n) = \{p(\boldsymbol{\mu}_n, a_1), \cdots, p(\boldsymbol{\mu}_n, a_K)\} \tag{28}$$

$$P(\boldsymbol{\mu}_m) = \{p(\boldsymbol{\mu}_m, a_1), \cdots, p(\boldsymbol{\mu}_m, a_K)\} \tag{29}$$

であるから，$P(\boldsymbol{\mu}_n)$ の $P(\boldsymbol{\mu}_m)$ に関するK-L情報量は，

$$I(P(\boldsymbol{\mu}_n); P(\boldsymbol{\mu}_m)) =$$

$$\sum_{i=1}^{K} p(\mu_n, a_i) \log \frac{p(\mu_n, a_i)}{p(\mu_m, a_i)} \quad (30)$$

となる．$I(P(\mu_n); P(\mu_m))$ が 0 に近い程，行動選択確率分布の類似度が高いとする．

3.3.2 中心座標の類似度

マハラノビス距離[9]を用いて，中心座標の類似度とする．マハラノビス距離とは，ある二点間の距離を，分散（本モデルにおける基底関数の幅）を考慮して算出したものである．ここで，ある基底関数 $B_n(\boldsymbol{x})$ の分散を考慮して，中心座標 μ_n から $B_m(\boldsymbol{x})$ の中心座標 μ_n までのマハラノビス距離 $d_n(\mu_m)$ を求めると，

$$d_n^2(\mu_m) = (\mu_m - \mu_n)^t \Sigma_n^{-1} (\mu_m - \mu_n) \quad (31)$$

となる．このマハラノビス距離が 0 に近い程，中心座標の類似度が高いとする．

マハラノビス距離を類似度に用いることにより，分散の大きな基底関数が一方的に高い類似度を見出し，小さな基底関数を吸収するように統合し，自律統合が進む．

3.3.3 基底関数の類似度

行動選択確率と中心座標の類似度より基底関数の類似度とする．今，基底関数 $B_n(\boldsymbol{x})$ の $B_m(\boldsymbol{x})$ に対する類似度 $s(B_n(\boldsymbol{x}), B_m(\boldsymbol{x}))$ を

$$s(B_n(\boldsymbol{x}), B_m(\boldsymbol{x})) = \exp(-aI(P(\mu_n); P(\mu_m)) - bd_n^2(\mu_m)) \quad (32)$$

とする．ここで，a, b はともに任意の定数である．類似度の最大値は 1 であり，類似度が低い程 0 に近付く．

また，類似度を評価する時には，学習が進み，各基底関数の自己組織化が進むなかで，高い類似度を持ち続ける基底関数こそ高い類似度を持つとすべきある．そこで，類似度 $S_{nm}(t)$ を新たに定義する．

$$S_{nm}(0) = 0 \quad (33)$$

$$S_{nm}(t+1) = (1-\beta) S_{nm}(t) + \beta s(B_n(\boldsymbol{x}), B_m(\boldsymbol{x})) \quad (34)$$

ここで，βは $0 \leq \beta \leq 1$ の値をとる定数である．

また，このときの単位時刻 t は，報酬が入力されたときに進むものであるとする．このことにより，学習の初期段階で報酬入力回数が少なく，各基底関数が自己組織化されていないときの類似度を少なく評価し，報酬入力が増える学習収束時の類似度の評価を重くおくことが可能である．

3.3.4 基底関数の統合

任意の閾値 th 以上の類似度を持つ基底関数 $B_n(\boldsymbol{x})$ と $B_m(\boldsymbol{x})$ とを統合する．統合の結果，新たに生成される基底関数 $B_{new}(\boldsymbol{x})$ のパラメータ μ_{new}，Σ_{new} を以下のように決定する．

$$\mu_{new} = \frac{d_m^2(\mu_n)}{d_n^2(\mu_m) + d_m^2(\mu_n)} \mu_n + \frac{d_n^2(\mu_m)}{d_n^2(\mu_m) + d_m^2(\mu_n)} \mu_m \quad (35)$$

$$\Sigma_{new} = \Sigma_n + diag(\Delta x_i^2) \quad (36)$$

$$\Delta x_i = \mu_{new}(i) - \mu_n(i) \quad (37)$$

ここで，任意の重み $W_{new}(a_i)$ は，

$$W_{new}(a_i) = W_n(a_i) B_n(\boldsymbol{x}) + W_m(a_i) B_m(\boldsymbol{x}) \quad (38)$$

とする．これは，新たに置き換えられる基底関数の中心座標における行動価値関数の値を保持するためである．統合の結果生成された新たな基底関数に関する類似度は，すべて 0 に初期化する．

4. 実　　験

EQ-学習の評価をシミュレータ実験により行う．実験対象のロボット（大きさ 100×100）は視覚機能を持ち，餌（半径 50）を認識可能であると想定する．具体的には餌までの距離，視界端からの角度の 2 次元ベクトルが入力される．また，選択可能な行動は，前進・左右前進・左右旋回の 5 通りである（図 5）．

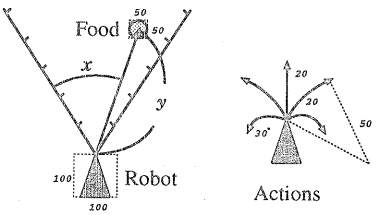


図 5 ロボットパラメータ
Fig. 5 Robot parameters

ロボットは，2500×2500 の広さを持つフィールドに散在する餌を拾った時のみ報酬を獲得する．

$$r(\boldsymbol{x}, a_i) = \begin{cases} 1 & \text{if the robot gets a ball} \\ 0 & \text{otherwise} \end{cases} \quad (39)$$

また，学習時の各パラメータは以下のように定める．

$$\alpha = 0.25 \quad \gamma = 0.99 \quad T = 0.1 \quad (40)$$

各学習回数（報酬入力 $r(\boldsymbol{x}, a_i) = 1$）毎に学習を停止し，各有用度関数を用いて求餌行動に適用し，餌を拾うまでにかかったステップ数を評価する．評価は，各々学習し，個性を持つ 30 台のロボットが環境内に散在する 3 つの餌を拾い終えるまでにかかったステップ数を 1000 回ずつ記録し，その平均ステップ数により行う．また，評価を行う際に 10000 ステップ以上必要であったもののステップ数は 10000 とした．

4.1 実験結果

EQ-学習の評価実験を行う．各行動に同一の基底関数をセンサ空間上に格子状に配置し，初期状態とする．基底関数の配置は図6に示す5通りがあり，4×4の配置に対してのみ自律統合を行う．方形波関数を基底関数とすると，同一色の状態に対して一様に強化することとなる．また，基底関数がガウス関数であるときは重みの値を無視し，最大値を返す関数によってセンサ空間を色分けをしている（自律統合結果の図12の下段も同様）．

性能評価のために本実験では以下の実験を行う．ここで，図1における実験1は従来のQ-学習に対応し，それに対する学習の収束の具合を評価し，仕事効率に基づき行動価値の近似精度の評価とする．また，基底関数の自律統合を行わないときは図6で示す5通りの基底関数を適用する．図7～図11に実験1と実験2の結果を，また，図12には実験3の結果を自律統合される基底関数の例とともに示す．自律統合の各パラメータは以下のように定めた．

$$a=1 \quad b=0.5 \quad \beta=0.1 \quad th=0.25 \quad (41)$$

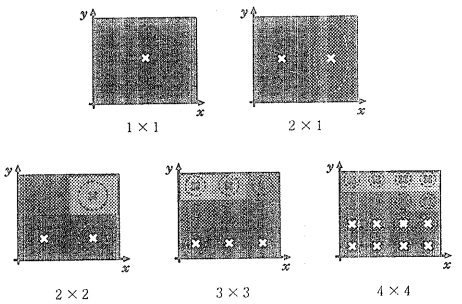


図6 基底関数次数の違い
Fig. 6 Difference which is the number of base functions

4.2 考察

図7，図8の方形波関数の結果より，基底関数の次数が少なすぎると学習が収束できないことが確認された．また，基底関数の次数（学習により調整するパラメータ数）が適切な個数を越えると，学習が収束するのに必要な時間が増大することも確認された．しかし，図8においてガウス関数のときは収束することにより，ノイズに対する頑健性が確認された．つまり，有

表1 実験
Table 1 experiments

	基底関数	自律統合化
実験1	方形波関数	無し
実験2	ガウス関数	無し
実験3	ガウス関数	あり

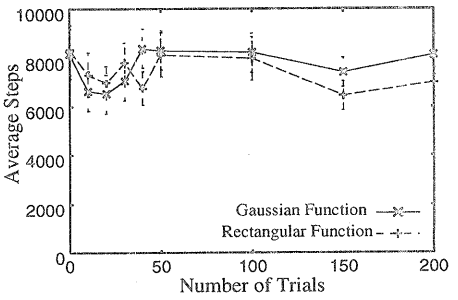


図7 次数1×1の基底関数による学習結果
Fig. 7 Result of the reinforcement learning with 1 × 1 base function

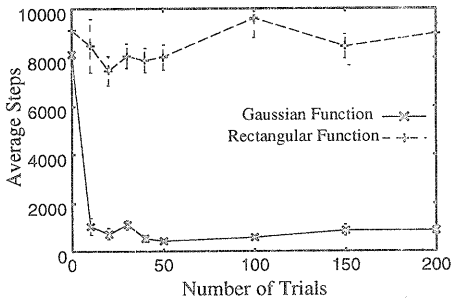


図8 次数2×1の基底関数による学習結果
Fig. 8 Result of the reinforcement learning with 2 × 1 base functions

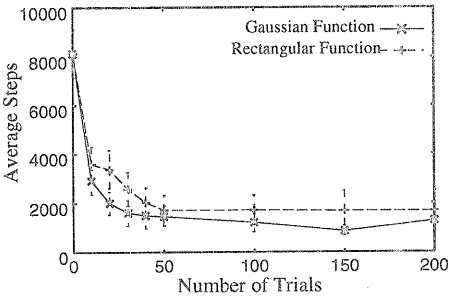


図9 次数2×2の基底関数による学習結果
Fig. 9 Result of the reinforcement learning with 2 × 2 base functions

用度関数を近似する基底関数としてガウス関数を用いた場合，従来法に対して次数の変動に頑健であり，より簡潔なモデルでの学習が可能であることが示されている．

また，図11，図12の結果が示すように，基底関数が過剰に存在するとしても，EQ-学習により，早い学習の収束，効率の良い行動獲得が確認され，従来法をしのぐ結果が得られた．これは，類似度の高い基底関数同士を統合することにより，適切な基底関数を自動獲得し，学習パラメータ数の適正化が行われた結果で

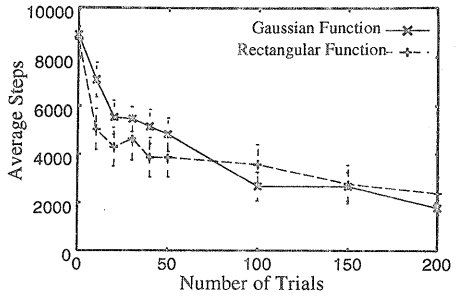


図 10 次数 3 × 3 の基底関数による学習結果
Fig. 10 Result of the reinforcement learning with 3 × 3 base functions

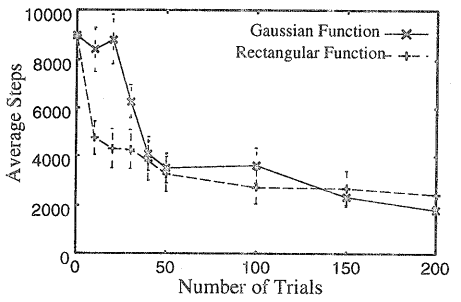


図 11 次数 4 × 4 の基底関数による学習結果
Fig. 11 Result of the reinforcement learning with 4 × 4 base functions

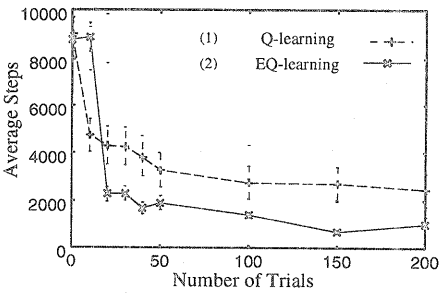
ある.

5. まとめと課題

強化学習とは，あらゆる状態に対する最適な行動を，自らの経験により獲得するものである．従来，ロボットが連続系のセンサ空間をもつときは，センサ空間を離散化し，有限個の状態上での行動決定問題に帰着して，学習時間の増大を抑え，強化学習に適用してきた．しかし，離散化を行うことにより，様々な誤差が発生しているのも事実である．

そこで，本論文では，強化学習の一つとして Q-学習を拡張した EQ-学習を提案した．EQ-学習は，行動有用度関数を基底関数の重み付き和であらわすことにより，連続系のセンサ空間をそのまま扱うことが可能である．また，基底関数の自律統合により，近似精度をある程度保ちつつ，より簡潔なモデルを自律獲得可能となった．これにより，学習を行うときに，基底関数の配置をタスク毎に構成する必要がなく，タスク，ロボットパラメータに対して適切，かつ簡潔なモデルを自律的に獲得可能である．

今後の課題としてまず挙げられるのは，基底関数の配置法で統合だけの一方通行的な手法を用いていることである．統合を行った後に誤差が大きくなり，仕事遂行の効率が著しく落ちたときには再度基底関数を分割するなどの対処が必要である．次に取り組むべき課題は，拡張 Q-学習の実機への実装がある．実ロボットのセンサはノイズを多分に含むが，従来の学習法と比較して，ノイズに対する頑健性が確認されているので，よりよい行動獲得が期待できる．現在，人工知能と知能ロボットの標準問題となっている RoboCup[4] のサッカーロボットに適用することを考えている．サッカー行動は，本論文で実験を行った餌拾い行動と比べると，味方や敵，ボールなどの環境が動的に変化し，獲得すべき行動も，その複雑さが予想される．そこで，新たに，行動の階層性[2] を組み込んだ拡張 Q-学習を検討している．



Number of trials	0	50	100	200
Average numbers of base functions	16	4.2	2.8	2.5
Example of the unification				

図 12 Q-学習と (次数 4 × 4) との比較 (上) と EQ-学習の基底関数の統合例 (下)
Fig. 12 Comparison (above) Q-learning and EQ-learning whose unification (below).

謝 辞

本研究の一部は，文部省科学研究費 (基盤 (c)，課題番号 10680408) の補助を受けて行なわれた．

参 考 文 献

1) Minoru Asada, Shoichi Noda, and Koh Hosoda : Non-physical intervention in robot learning based on KfE method, *Proc. of Machine Learning Conferene Workshop on Learning from Examples vs. Programming by Demonstration*, pp. 25–31 (1995).
2) Digney, B. L. : Learning hierarchical control structures for multiple tasks and changing environments, *From animals to animats* 5:SAB

98 (1998).
3) Andrew G. Barto, Richard S. Sutton, and Charles W. Anderson: Neuronlike adaptive elements that can solve difficult learning control problems, *IEEE Transaction on systems, man, and cybernetics*, Vol.SMC-13 No.5, pp.834-846 (1983).
4) Hiroaki Kitano, Minoru Asada, Yasuo Kuniyoshi, Itsuki Noda, Eiichi Osawa, and Hitoshi Matsubara : RoboCup:a challenge problem for AI and robotics, *Lecture Notes in Artificial Intelligence*, Vol.1395, pp. 1–19 (1998).
5) Norihiko Ono, Yoshihiro Fukuta : Learning coordinated behavior in a continuous environment, *Lecture Notes in Artificial Intelligence*, Vol.1221, pp. 73–81 (1997).
6) C.J.C.H. Watkins : *Learning from Delayed Rewards*, PhD thesis, University of Cambridge (1989).
7) 榎田 修一，大橋 健，吉田 隆一，江島 俊朗：行動確率場モデルによる行動獲得手法の高性能化，第16回日本ロボット学会学術講演会予稿集，pp. 405–406 (1998).
8) 坂元 慶行，石黒 真木夫，北川 源四郎：情報量統計学，共立出版株式会社，pp. 42–64（第4章）(1983).
9) 鳥脇 純一郎：認識工学，コロナ社，pp. 23–61（第2章）(1993).
10) 村川 正宏，米井 友浩，樋口 哲也，吉澤 修治：RBF ネットワークを用いた時変環境における Q-learning，電子情報通信学会論文誌 D-II Vol.J81-D-II No.12，pp. 2828-2840 (1998).
11) 森谷 淳，銅谷 賢治：強化学習による起き上がりパターンの獲得，電子情報通信学会信学技報 TECHNICAL REPORT OF IEICEL NC97-28, pp. 25–32 (1997).

(平成11年4月16日受付)
(平成11年6月 4 日再受付)
(平成11年6月 5 日採録)

榎田 修一

1997年九州工業大学情報工学部知能情報工学科卒業．1999年九州工業大学大学院情報工学研究科情報科学専攻博士前期課程修了．現在，同大学院博士後期課程在学中．自律型ロボットの行動獲得に興味を持つ．日本ロボット学会学生会員．

大橋 健（正会員）

1989年長岡技術科学大学・工学部・電気電子システム工学課程卒業．1991年同大大学院修士課程修了．同年九州工業大学情報工学部知能情報工学科助手，現在に至る．マルチモーダルインタフェース，画像や音声の認識理解に興味を持つ．博士（情報工学）．電子情報通信学会，日本ソフトウェア科学会，IEEE 各会員．

吉田 隆一（正会員）

1982年慶應義塾大学工学部電気工学科卒業．1987年慶應義塾大学大学院工学研究科博士後期課程電気工学専攻修了．工学博士．同年九州工業大学情報工学部知能情報工学科助手．1990年同助教授．1993年から翌年にかけてオレゴン科学技術大学院大学において客員研究員．オブジェクト指向計算，分散計算，分散処理システム，オブジェクト指向データベースに興味を持つ．日本ソフトウェア科学会，人工知能学会，IEEE, ACM 各会員．

江島 俊朗（正会員）

1973年東北大・工・通信卒．1978年同大大学院博士課程了．同年東北大・工・通信工学科助手．1985年同大情報助教授．同年長岡技術科学大学電気系助教授．1990年九州工業大学情報工学部知能情報工学科教授．1995年から96年までカリフォルニア大学デイビス校客員教授．文字・図形，音声の認識およびヒューマンインタフェースに興味を持つ．工博．電子情報通信学会，人工知能学会，IEEE 各会員．