方策勾配法を用いたサッカーエージェントの学習
～フリーキック～

中村浩二[†]，五十嵐治一[†]，石原聖司[‡]

[†]芝浦工業大学大学院電気電子情報工学専攻，[‡]近畿大学工学部電子情報工学科

概要

RoboCup シミュレーションリーグは，マルチエージェント系における協調行動の学習を研究するための題材として知られている．本研究では，協調行動の一事例として，相手ゴール前での味方チームの直接フリーキック時に，キッカーと味方レシーバ１名との連係プレーを学習する問題を取り上げた．本研究では，このような場面において，キッカーがどの地点にパスを出せば，パスが通りやすく，かつ，得点に結びつくかというヒューリスティクスを関数化した．これにより，キックオフの位置や，味方レシーバと守備側の選手の位置の組合せからなる膨大な状態空間を処理することができる．パス先の決定には，この関数を用いたボルツマン選択を採用した．関数中のパラメータは，強化学習の一種である方策勾配法により学習することができる．同様に味方レシーバも走り込む位置を学習する．実際にこれまでに RoboCup の競技会に参加したチームのいくつかを用いて学習実験を行い，提案した学習方式の有効性を確認した．

Learning of Soccer Player Agents Using Policy Gradient Methods
~Free Kicks~

Koji Nakamura[†], Harukazu Igarashi[†], Seiji Ishihara[‡]

[†]Department of Electrical Eng. and Com. Sci., Shibaura Institute of Technology
[‡]Department of Electronic Eng. and Com. Sci., School of Eng., Kinki University

Abstract

RoboCup Simulation League is known to be a test bed for research of multi-agent learning. As an example of multi-agent learning in a soccer game, we dealt with a learning problem between a kicker and a receiver when a direct free-kick is awarded just outside the opponent's penalty area. To which point should the kicker kick the ball in such a situation? We propose a function that expresses heuristics on evaluating how the target point is advantageous in sending/receiving a pass safely and contributing to scoring. The evaluation function makes it possible to handle a large space of states consisting of positions of a kicker, a receiver and the opponent's players. A target point of a free kick is selected by the kicker using Boltzmann selection with the evaluation function. Parameters in the function can be learned by a kind of reinforcement learning called policy-gradient method. A point where a receiver should run to receive the ball is simultaneously learned in the same manner. Experiments with four teams that had participated in the past RoboCup competitions showed the effectiveness of our solution.

1. はじめに

近年，人工知能の分野ではマルチエージェント環境下における協調行動の学習が研究されている[1]. この研究の題材としてロボットサッカーの競技会である RoboCup が提唱されている．この RoboCup の中の一部門であるシミュレーション部門では，実環境でロボットを動かす難しさから解放されるため，協調的な行動に研究の焦点を当てることができる[2]．一般に，マルチエージェント学習には，状態空間の爆発問題，同時学習問題，不完全知覚問題，実時間処理問題等のマルチエージェント系特有の問題があるとされている[1]．RoboCup サッカーシミュレーションでは，これらの中でも，まず第一に「状態空間の爆発問題」が大きな問題と考えられる．

本研究ではこのように多数のエージェントが広い対象領域に集まり状態空間が大きくなる場合でも適用できる行動学習法の開発を目的としている．今回，その一例としてゴール前でのフリーキックの場面を取り上げた．具体的にはキッカーとレシーバのそれぞれ１名が協調するサインプレー（パス行動）において，キッカーのみが学習する場合とキッカーとレシーバが共に学習する場合の実験を行った．協調行動の学習には強化学習法の一種である方策勾配法と，状態数によらない方策表現とを用いた．

2. RoboCup シミュレーションリーグ

2.1 RoboCup とは

RoboCup とは，'92 年に日本の研究者によってロボット工学と人工知能研究の融合・発展のために提唱された自律型移動ロボットによるサッカーを題材とした競技会である．現在では，サッカーだけでなく，大規模災害へのロボットの応用としてレスキュー，次世代の技術の担い手を育てるジュニアリーグなどが組織されている．最終的な目標は 2050 年までにヒューマノイドロボットが，人間のサッカー世界チャンピオンに勝つこととされている．

2.2 シミュレーションリーグのルール

RoboCup にはいくつかの競技部門があるが，シミュレーションリーグはコンピュータ上の二次元の仮想フィールド上で，それぞれ異なる頭脳を持った 22 のプレーヤーが５分ハーフのサッカーを行うリーグである．

シミュレーションリーグでは公式シミュレータとして SoccerServer が用いられている．SoccerServer は電子技術総合研究所の野田らによって開発されたコンピュータ上で動作する仮想サッカーゲームのシミュレータで，オープンソースとして誰でも入手することができる[3].

このシミュレータはクライアント・サーバ方式をとっており，ユーザは 11 人分のプレーヤープログラムを作成し，クライアントとしてサーバに UDP/IP で接続する．サーバは物体の動きをシミュレートして，クライアントに視覚・聴覚・感覚情報を送信し，クライアントはサーバへ kick/dash/turn などの基本コマンドを送信してプレーヤーを動かす．しかし，クライアント同士は say/hear コマンドなどを用いたサーバを介した通信はできるが，直接通信を行うことができない．つまり，チームの制御は独立したクライアントによる分散制御であり，その中で協調的な戦術を駆使した試合を行わなければならない．

2.3 行動のための基本コマンド

クライアントは以下の基本コマンドを用いてサーバ上のプレーヤーを動かす．

- (turn *Angle*) : 自分の体を Angle だけ回転させる．
- (turn_neck *Angle*)：自分の首を Angle だけ回転．
- (dash *Power*) :
  今向いている方向へ Power だけ加速する．
- (kick *Power Dir*) :
  Dir の方向へ Power だけボールを蹴る．
- (catch *Dir*) : Dir の方向でボールをキャッチする．
  ただし，キーパーしか使えない．
- (say *Message*) :
  他のプレーヤーに Message を送る．

なお，ドリブル，パス，シュートに相当するといったコマンドは用意されておらず，上記の基本コマンドを組合せて実現しなければならない．

2.4 センサ情報

一方，サーバは以下のセンサ情報をクライアントに送る．なお，各プレーヤーの視野は限定されており，その範囲内での視覚情報のみが送られる．

- sense_body : プレーヤーのスタミナやスピードなどの身体に関する情報を送る．
- see：視野範囲内での物体の情報を送る．
- hear :
  他のプレーヤーが say した Message を送る．

クライアントは基本的に１シミュレーションステップ（100ms）に一回しかコマンドを送信することを許されていない．また，サーバから送られてくる視覚情報は，通常 150ms に一回しか送られてこない．つまり，コマンドの送信と視覚情報の受信は非

同期である．さらに，実世界の不確実さを反映するために情報にノイズが入っており，これらを考慮した行動決定が必要である．

3. フリーキックにおける協調行動

3.1 シミュレーションリーグにおける協調行動

１章で述べた「状態空間の爆発問題」とは，全エージェントの状態数と各状態におけるエージェントのとる行動の組み合わせが多すぎて，理論的には実行可能だが実際問題として時間的に使いものにならなかったり，メモリオーバーとなったりする問題である．これは状態空間を適切に設計すれば，ある程度，状態空間の爆発を抑えることができる[1].

しかし，少領域内で存在するプレーヤーが少数であれば熊田らの研究[4]のようにパス問題を取り扱うことは可能であるが，文献[4]では課題として「フルゲームへ拡張すると計算量爆発を引き起こす」と述べている. つまり, マルチエージェント学習では, 仮に小さなシステムにおいて適切な状態空間を設計できたとしても系を大きくすると状態空間の爆発が起きてしまいやすい．したがって，状態空間の表現法を工夫し，状態数によらずにマルチエージェント学習が可能な手法を用いることが望ましいと考えられる.

3.2 先行研究

比較的広い領域で多数のプレーヤーが存在する場合のパス問題において，状態数によらない手法を用いて協調行動の獲得を提案した亀島らの研究例がある[5]．そこでは，まずフィールドを格子状の長方形セルの集合に区切り，passer はルーレット方式でパスを出すセルを決めてパスを出す．もしパスが receiver に通ったら，強化学習でパスを出したセルの価値を高めていき，最終的に一番価値の高いセルにパスを出している.

この手法の特徴は，セルの価値を保存しておくだけでよいので，フィールドにおけるプレーヤーの配置の組み合わせ等といった広大な状態空間を保存する必要がないことである．しかし，この研究では敵プレーヤーの位置が固定されていることが前提であり，敵プレーヤーが移動すると学習結果が利用できず実用性は低い.

3.3 本研究の方針

本研究では，比較的広い領域で多数のプレーヤーが存在する場合のパス問題の一つとして，具体的にゴール前でのフリーキックの場面を取り上げる．そこで，先行研究の欠点を克服するために，行動決定（方策）における状態表現においてプレーヤーの絶対的な位置座標を用いるのではなく，配置から計算されるいくつかの特徴量（ヒューリスティクス）を用いる.

検証実験として，1)キッカーのみがパス先の地点を決定するために学習を行い，レシーバはキッカーから指示されたセルへ走りこむといった条件での学習実験（実験 1）と，2)キッカーはパス先の地点を，レシーバも走り込む先の地点をそれぞれ独立して両方が学習する実験（実験 2）とを行う.

4. 提案手法

4.1 方策勾配法

本研究では強化学習の一種である方策勾配法を用いて学習を行う．方策勾配法とは，報酬の期待値が最大になるように方策パラメータを更新する学習法である．このときの最大化の手段として確率的勾配法を用いる．数学的な基礎がはっきりしており，理論的に取り扱いやすい．また，方策として if－then 型のルールや，ポテンシャルなどの様々な関数が利用できるので，方策への知識表現が容易であるという長所がある．元々は，Williams により提案された手法[6]であるが，筆者らも追跡問題やカーリングゲームに適用し，有効性を確認している[7][8].

4.2 行動決定のための確率的な方策

本稿では，次章で述べるフリーキックの問題のために，以下のようなキッカーの行動決定方式を提案する．まず，一般的なパス行動を考える．フィールドを[5]と同様に格子状の長方形セルの集合に区切る. キッカーがセル k へパスを出すという行動 a_k の価値を次のような目的関数で表す.

$$E(a_k;\omega) = -\sum_i \omega_i \cdot U_i(a_k) \qquad (1)$$

この関数は，パス先のセル k を決める上で有効と思われる状態の特徴量（ヒューリスティクス）$\{U_i\}$ の線形和である．ただし，目的関数 E の値が小さい方が，行動としての価値が高くなるように設計する.

この目的関数を用いて，キッカーは次のボルツマン選択による確率的な方策を用いてパス先のセル k を決定する.

$$\pi(a_k;s) \equiv \frac{e^{-E(a_k;s)/T}}{\sum_x e^{-E(x;s)/T}} \qquad (2)$$

ボルツマン選択は温度パラメータ T を大きくするほどランダムに行動を選択するようになり，小さくするほど最も大きい価値の行動を選択しやすくなる. 特に, T→0 では決定論的な行動決定となる. なお,

s は全系の状態（i.e. 全プレーヤー，ボールの位置）を表している．

式(1)では先行研究[5]と違い，敵プレーヤーの数や配置，レシーバの位置など，その場面に応じて各セルへのパスを出す行動の価値を計算しており，プレー中の環境の変化も考慮している．勿論，広大な状態空間の保存を必要としていない．

式(1)のパラメータ $\{\omega_i\}$ の学習則を次節以下に示す．この学習則を用いてキッカーの方策(2)を学習することができる．

4.3 自律分散的な行動決定とその学習則

一般に，マルチエージェント系全体の状態を s，行動を a とする．それぞれ，各エージェントの状態と行動とを要素とするベクトルであることに注意する必要がある．行動 a に対する目的関数を E(a;s, θ) とし，方策が式(2)のようなボルツマン選択である場合に，そのまま方策勾配法を適用すると，パラメータ θ に関する学習則は，

$$\Delta\theta = \varepsilon \cdot r \sum_{t=0}^{L-1} e_\theta(t) \qquad (3)$$

で表される[7]．ε は学習係数，r はエピソード終了時に与えられる報酬，Lはエピソード長である．e_θ(t)は，離散時刻 t における適正度(eligibility)で，次の式で定義されている．

$$e_\theta(t) \equiv \frac{\partial}{\partial\theta} \ln \pi(a;s,\theta) \qquad (4)$$

次に，行動決定と学習とを自律分散的に行うことを考える．そこで，系全体の方策を，各エージェント i の方策関数 π_i の積で近似する[7]．すなわち，

$$\pi(a;s,\theta) \approx \prod_i \pi_i(a_i;s,\theta_{ij}) \qquad (5)$$

と仮定する．ここで，a_iはエージェント i の行動であり，θ_{ij}はπ_iに含まれる j 番目のパラメータである．さらに，各π_iは，各エージェントごとに定義される目的関数 E_i (a_i;s, θ_{ij})を用いたボルツマン分布とする．したがって，4.2 の式(2)中の右辺の目的関数Eは，厳密にはエージェントごとに定義された目的関数 E_i である．

なお，(5)の近似は，エージェント間の行動選択の相関を無視していることに相当している．また，他のプレーヤーの位置情報も分かっているという前提に立っているので，各π_iには系全体の状態 s が使用されている．

一方，報酬 r の期待値E[r]をパラメータについて微分すると，形式的には，

$$\frac{\partial E[r]}{\partial\theta_{ij}} = E\left[r \sum_{t=0}^{L-1} e_{\theta_{ij}}(t) \right] \qquad (6)$$

となる．(5)の近似を用いて(6)の右辺の適正度を計算すると，

$$e_{\theta_{ij}}(t) \equiv \frac{\partial}{\partial\theta_{ij}} \ln \pi(a;s) \qquad (7)$$

$$\approx \sum_i \frac{\partial}{\partial\theta_{ij}} \ln \pi_i(a_i;s,\theta_{ij}) \qquad (8)$$

さらに，各エージェントの目的関数中のパラメータは互いに独立であると仮定すると(7),(8)は，

$$e_{\theta_{ij}}(t) \approx -\frac{1}{T}\left[\frac{\partial E_i}{\partial\theta_{ij}} - \left\langle \frac{\partial E_i}{\partial\theta_{ij}} \right\rangle \right] \qquad (9)$$

と表される．一方，パラメータ θ_{ij} の学習則は(3)の導出と同様にして，(6)を用いて確率的勾配法を適用すると，

$$\Delta\theta_{ij} = \varepsilon \cdot r \sum_{t=0}^{L-1} e_{\theta_{ij}}(t) \qquad (10)$$

となる．ただし，(5)の近似により(10)の右辺の適正度は(9)で与えられる．(3)が系全体の目的関数E(a;s)に含まれるパラメータに関する学習則であるのに対し，(10)はエージェント i が自己の行動 a_i を選択するために使用する目的関数 E_i の中に含まれるパラメータの学習則であり，かつ，自己の目的関数 E_i だけを用いている点で，自律分散的な学習であると言える．

5. 問題設定

5.1 実験１：キッカーの学習

実験１としてキッカーのみが直接フリーキック時にどのセルへ向けて蹴るべきかを学習する．一方，レシーバはキッカーからのパス先セルの情報を受けて，そのセルに走りこみ，パスを受けて直ちにシュートするといった行動をとる．つまり，レシーバはキッカーが決定した方策に従うだけであり，学習を行うのはキッカーだけである．

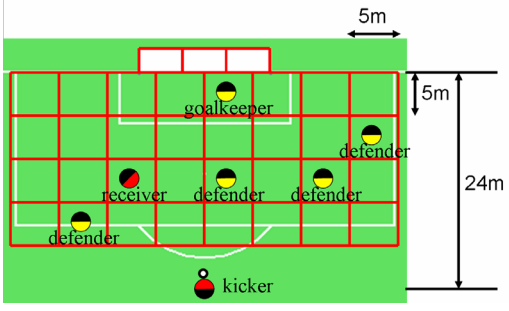


Fig. 1 Arrangement of players

実験では図１に示すようにペナルティエリア付近を一辺 5m の正方形セルで４×８に分割したフィールドを用いる．さらに均等にゴールも３分割する．これにより，「ゴールにパスをする」ことが「直接シュートをする」ことを実現している．

使用するエージェントと配置場所を以下に，その配置例を図１に示す．なお，SoccerServer ではフィールドの中心（センターマーク）を原点とし，タッチラインに平行に X 軸，ゴールラインに平行に Y 軸をとる．プレーヤーは４種類で，キッカー１人，レシーバ１人，敵チームのディフェンダー４人とゴールキーパー１人である．

キッカーにはフリーキック時にボールを蹴るだけの役割を与える．配置場所の X 座標はゴールラインから 24m の位置で固定，Y 座標はランダムな位置とする．レシーバはキッカーが指示したセルに走りこみ，パスを受けたらその場で直ちにシュートをする．配置場所はゴールの３セルを除くセルの中でオフサイドポジションにならないランダムな位置とする．

敵ディフェンダーはゴールを決めさせないためにパスカットやプレスを行う．配置場所はゴールの３セルを除くセルの中でランダムな位置とする．ゴールキーパーはゴールエリア内のランダムな位置に配置する．

ディフェンダーとゴールキーパーは Web 上で公開されているチームを用いる．実験では，以下の異なる強さを持った３チームを対戦相手に採用した．まず，弱いチームとして K チームのプログラムを採用した．このチームのディフェンダーはボールに向かうのみで協調的な守備は行わない．ゴールキーパーは前に出て行くだけである．

次に，普通の強さとして，大阪府立大学チーム hana を採用した．このチームは'03 年 RoboCup ジャパンオープンで総合ベスト８に入ったチームで，ディフェンダーはよく動き，パスカットなどを行う．ゴールキーパーもよく動き，セービング能力も高い．

最後に，強いチームとしてアムステルダム大学チーム UvA Trilearn 2003 を採用した．このチームは'03 年 RoboCup 世界大会で優勝したチームでディフェンダーは hana に比べるとあまり動かないが，レシーバへのマークやパスカット，シュートコースを消す動きなどの協調的な守備を行う．ゴールキーパーもあまり動かないが，セービング能力が非常に高く，ディフェンダーと協力してシュートコースを消す動きもする．

プレーヤーの他に trainer というエージェントも使用する．trainer はプレイモードを kick_off から free_kick に変更したり，エージェントとボールを移動したりすることができる．そして，フィールドを監視し，パスは通ったのか，シュートは決まったのかをキッカーに伝えることができる．

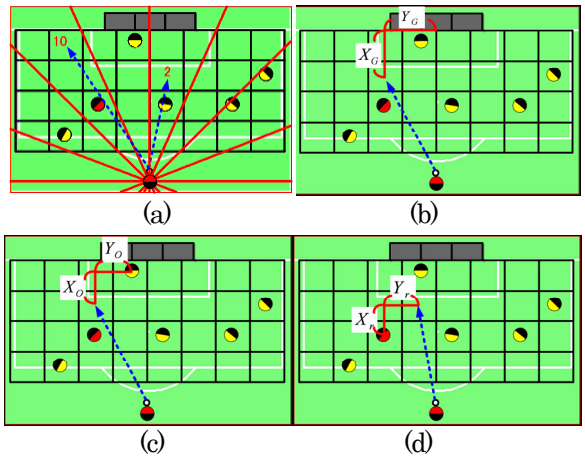


Fig. 2 Heuristics $\{U_j\}(j=1,2,3,4)$ used in $E(a_k;s,\theta_{ij})$

5.2　目的関数に用いたヒューリスティクス

実験ではゴール前でのフリーキックにおいて有効と思われる以下の４つのヒューリスティクスを用いる．なお，どの関数も大きいほど価値がある．

① パスコースにおける敵の有無：

$$U_1=\begin{cases}10.0 & \text{敵がいない場合}\\ 2.0 & \text{敵がいる場合}\end{cases} \qquad (11)$$

実際にパスをするにあたって，パスコースに敵がいない方がパスは通りやすい．図２(a)のようにキッカーの正面 180 度を 22.5 度ずつ８領域に分割し，敵がいれば２，いなければ 10 を返す．

② パス先とゴールとの距離：

$$U_2=(-(X_G+Y_G)+37.5)/3.5 \qquad (12)$$

パス先とゴールとの距離が近い方が，パスを受け取った時にゴールに結びつきやすい．式(12)の X_G と Y_G は図２(b)のようにそれぞれパス先とゴールとの距離の X 成分，Y 成分である．なお，式(12)は $0.5\le U_2\le 10.0$ と正規化してある．

③ パス先と最近接の敵との距離：

$$U_3=(X_O+Y_O)/5.0 \qquad (13)$$

パス先と敵との距離が遠いほど，そのパス先にはスペースがあるということである．つまり，フリー

でシュートを打ちやすくなる．式(13)のX_OとY_Oは図 2(c)のようにそれぞれパス先と最近接の敵との距離の X 成分，Y 成分である．なお，式(13)は$0.0 \le U_3 \le 10.0$と正規化してある．

④ パス先とレシーバとの距離：

$$U_4 = \left(-\left(X_r + Y_r\right) + 50.0\right)/5.0 \qquad (14)$$

パス先とレシーバとの距離が近いほど，パスは通りやすくなる．式(14)のX_rとY_rは図２(d)のようにそれぞれパス先とレシーバとの距離の X 成分，Y 成分である．なお，式(14)は$0.0 \le U_4 \le 10.0$と正規化してある．

5.3 報酬 r

報酬 r の値は以下のように設定した．１エピソード（後述）終了後にキッカーに与える．

- パス失敗時 ・・・・・・・・・・・・・・・・・・・・ -1.5
- パス成功∧シュート失敗時 ・・・・・・ 0.5
- パス成功∧シュート成功時 ・・・・・・ 5.0

ここで，１エピソードを 70 シミュレーションサイクル（７秒）とし，１エピソード終了前にゴールを決めることができれば，エピソードはそこで終了とする．

以上の準備の下で，学習は次のような流れで 500 エピソード行う：

① trainer がプレーヤーを適切に配置
② キッカーは目的関数を計算し，パス先のセルを決定
③ キッカーはパス先をレシーバへ伝える
④ キッカーはパスを出し，レシーバは走りこむ
⑤ パスを受けたらその場でシュート
⑥ 結果に対して報酬を与え，重みを更新

なお，⑤でレシーバがパスを受けたその場ですぐにシュートをしているのは，本手法ではパス行動だけを評価したいので，ドリブルなどを行うとパス先がよかったのかが分からなくなるからである．

5.4 実験２：キッカーとレシーバの学習

2 つのエージェントが独立に学習を行っても，共通の方策を獲得する「合意形成」[5]が 4 章で述べた本方式の枠組みで可能であるかを検証するため，実験２として次のキッカーとレシーバの同時学習を行った．

基本的に実験１と同じ環境で実験を行うが，いくつか異なる点がある．まず，キッカー1 人，レシーバ 1 人，ディフェンダー1 人，ゴールキーパー1 人の 2 対２であることと使用する敵エージェントが Trilearn Base であることである．Trilearn Base とはアムステルダム大学チーム UvA Trilearn 2003 から高度な行動決定や戦略を除いてソースコード形式配布されているチームである．したがって，基本的な行動はしっかりしているものの UvA Trilearn 2003 より弱くなっている．

また，報酬 r の値は以下のとおりである．

- パス失敗時 ・・・・・・・・・・・・・・・・・・・・ -0.5
- パス成功∧シュート失敗時 ・・・・・・ 3.0
- パス成功∧シュート成功時 ・・・・・・ 10.0

同時学習をするに伴い，学習の流れを次のように変更した(②)．学習は 1000 エピソード行う．

① trainer がプレーヤーを適切に配置
② キッカーとレシーバはそれぞれ目的関数を計算し，パス先または走りこむ先セルを独自に決定
③ キッカーはパスを出し，レシーバは走りこむ
④ パスを受けたらその場でシュート
⑤ 結果に対して報酬を与え，重みを更新

6. 実験結果と考察

6.1 実験１の結果

5.1 で述べた３つのチームとそれぞれ学習させた後，各々の重みを用いて，改めて３チームと対戦させた．さらに比較対象として，全く学習を行っていないチームでも３チームと対戦させた．実験に使用した重みを表１に，パス成功率と得点成功率の結果

Table 1 Values of $\{\omega_i\}$ obtained in Experiment 1

	学習前	K	hana	UvA
ω_1	1.00	4.68	1.29	0.64
ω_2	1.00	1.08	0.50	0.26
ω_3	1.00	2.23	1.15	1.03
ω_4	1.00	5.03	1.57	0.78

Table 2 Pass success rate

		対戦チーム		
		K	hana	UvA
学習に用いたチーム	K	73.4%	21.3%	39.0%
	hana	74.6%	23.1%	34.8%
	UvA	77.4%	19.0%	40.0%
学習なし		45.5%	5.0%	19.8%

Table 3 Goal success rate

		対戦チーム		
		K	hana	UvA
学習に用いたチーム	K	73.4%	21.3%	39.0%
	hana	74.6%	23.1%	34.8%
	UvA	77.4%	19.0%	40.0%
学習なし		45.5%	5.0%	19.8%

を表２と表３に示す．なお，各成功率の値は 100 エピソードを 10 セット行った平均値である．

表２からＫチームよりそれより強いhanaやUvAを学習対象とした重みを用いて K チームと対戦した方が，若干優れた結果が出た．しかし，弱いＫチームの重みを用いて hana や UvA と対戦しても決して悪い結果が出たわけではない．さらに，個別学習が特によかったわけでもない．しかし，表３より得点に関しては強いチームで学習したほうが良いといえる．

それでも，強いチーム相手にはほとんどゴールを決めることができなかった．その理由として，味方の数的不利，強いチームのゴールキーパーの能力が高いこと，レシーバのシュート技術不足などが考えられる．

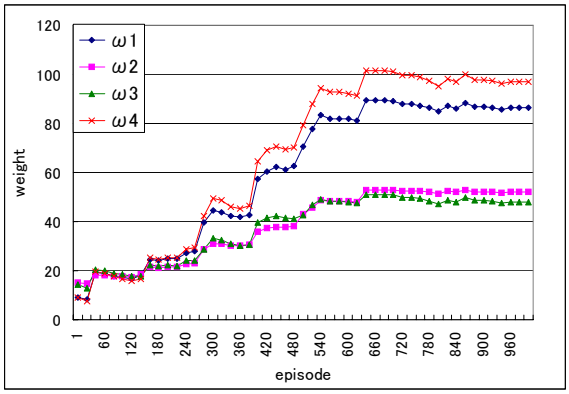
しかし，全く学習をしていないときと比べるとパス成功率と得点成功率はかなり向上している．したがって本方式による学習効果があったことがわかる．

6.2　実験２の結果

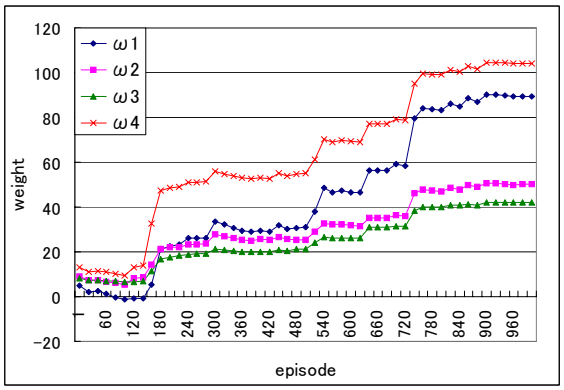
Trilearn Base を相手にしてキッカーとレシーバを独立に学習を行った．重み$\{\omega_i\}$の初期値はランダムとし，学習中の変化を図３に示す．パス成功率と得点成功率の結果を図４に，報酬の期待値を図５に示す．なお，２つの成功率の値はその学習回数における重みの値を用いて 100 エピソードを 10 セット行った平均値である．

図３より 1000 エピソードの学習における重みの初期値が異なっていても，キッカーとレシーバの重みの値は学習が進むに連れてほぼ一致していく．これによりキッカーとレシーバが共通の知識(方策)を獲得できたことがわかる．

図４よりパス成功率は 36.2 ポイント，得点成功率は 12.1 ポイント上昇した．どちらも学習前の 2 倍以上の値である．しかし，パス成功率の大幅な上昇のわりに得点成功率が伸びていないのは，図３より「パス先とゴールとの距離」の重みが小さく「パス先とレシーバとの距離」の重みが大きいからである．つまり，パス先を決定する上でレシーバとの距離が優先されるので，パスはよりたくさん通るようになる．一方，ゴールとの距離はあまり優先されなくなるので，ゴールから遠いところでボールをもらうことがたくさんあり，得点成功率はそれほど上がらない．それでも得点成功率が改善されたのは，ゴール付近へのパスでパスが通れば得点，通らなければキーパーに取られるといったギャンブル性のパスよりも，ゴールから遠くてもたくさんパスを通してたくさんシュートを打ったほうがより得点を奪うことができることをエージェントが選択したと思われる．



(a) Kicker



(b) Receiver

Fig. 3　Change of weight

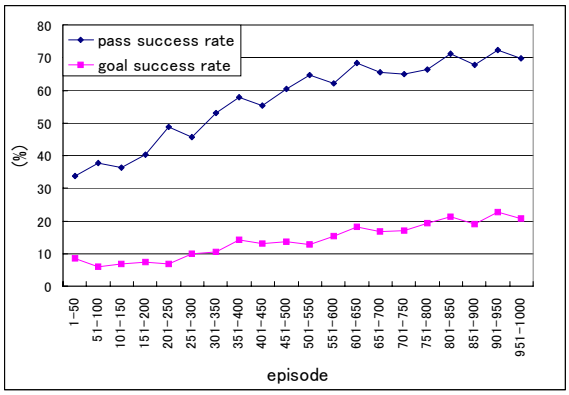


Fig. 4　Pass/goal success rate

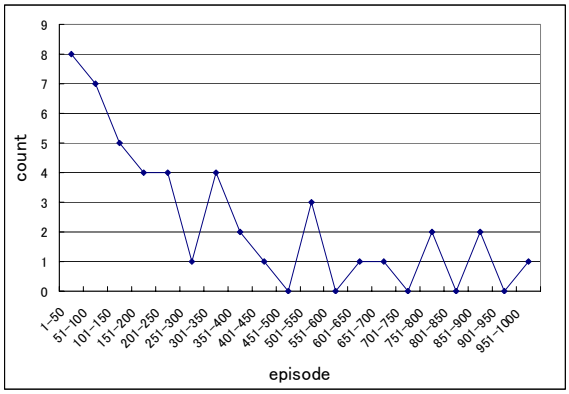


Fig. 5　Number of direct shots

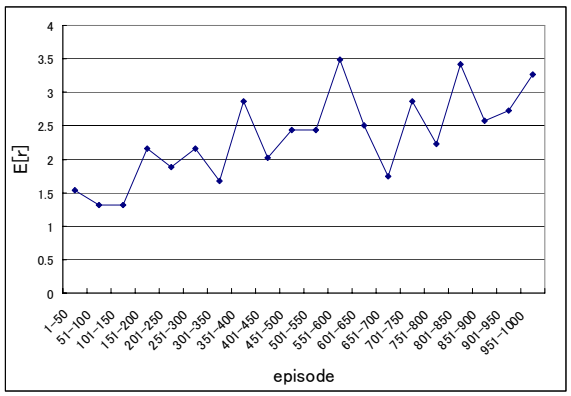


Fig. 6　Expected value of reward

また，学習の初期段階（50～150）で得点成功率が一度下がりまた上昇するといった現象が見受けられた．図６の報酬の期待値の変化からも同様なことが見て取れる．これは重みの初期値がランダムであるが故に，図５の直接シュート数の変化からわかるとおり初めは直接シュートを狙うことが比較的多く，シュート数が増えることにより得点を取ることができたためである．その後，協調的なパス行動を探索するため一時的得点成功率が下がる（報酬の期待値も下がる）段階を経た後，パス成功率の上昇と共に得点成功率が上がっていく過程を踏んでいるのが見て取れる．

7.　まとめ

本研究では，マルチエージェント系における行動学習の一例として RoboCup シミュレーションリーグにおけるゴール前のフリーキックの場面に適用し，1)キッカーのみの学習(実験 1)と，2)キッカーとレシーバの双方が学習する問題(実験 2)を取り上げた．また，「状態空間の爆発」を抑えるため，状態数によらない方策表現と方策勾配法とを用いた協調行動獲得法を提案した．その結果，２つの実験共にパス成功率と得点成功率が学習前と比較して２倍以上となり，提案手法による学習が有効であることを確認することができた．また，2)の実験よりキッカーとレシーバが共通の知識（方策）を獲得でき，提案手法の枠組みで合意形成が行われることを確認した．なお，本実験では学習するエージェントはキッカーとレシーバの２名だけであったが，学習が自律分散的であるため，学習するエージェントの数を増やしても学習時間が爆発的に増えることはない．

今後の予定としては，キッカーとレシーバとでヒューリスティクスが異なる場合や，得られる視覚情報が不完全であったり互いに異なる場合にどれだけ学習能力があるのかを調べたい．

また，各エージェントがパス行動とレシーブ行動との２つの方策を切り替えることにより，複数プレーヤー間での複数回のパス交換を実現することを考えている．さらには，通常の試合中での一般的な協調行動の学習へと研究を進めて行きたい．

参考文献

[1] 高玉圭樹：マルチエージェント学習ー相互作用の謎に迫るー，コロナ社，2003

[2] 野田五十樹：シミュレーションリーグとインフラ技術の技術的課題と展望，日本ロボット学会誌，Vol.20，No.1，pp.7-10，2002

[3] The RoboCup Soccer Simulator 〈http://sserver.sourceforge.net/〉

[4] 熊田陽一郎，上田一博：予測能力を持つサッカーエージェントによる協調戦術の獲得，人工知能学会論文誌，Vol.16，No.1，pp.120-127，2001

[5] 亀島力，遠藤聡志，山田孝治：強化学習を用いた共同注視点に基づく合意形成の獲得，電子情報通信学会・信学技報，Vol.99，No.131，pp.29-35，1999

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

[7] 石原聖司，五十嵐治一：マルチエージェント系における行動学習への方策こう配法の適用ー追跡問題ー，電子情報通信学会論文誌 D-I，Vol.J87-D-I，No.3，pp.390-397，2004

[8] 五十嵐治一，石原聖司，野原 勉：方策勾配法を用いた運動方程式中のパラメータ学習~２ストーン系のカーリングゲーム~，ロボティクス・メカトロニクス講演会’05(ROBOMEC’05)講演論文集，1A1-N-028(pp.1-4)，（2005.6.10-11，神戸，主催：日本機械学会ロボティクス・メカトロニクス部門）