Vol.2019-GI-41 No.4
2019/3/8

サッカーエージェントにおける方策勾配法と Q 学習の同時適用

山岸準† 五十嵐治一† 山岸拓海† 入倉雅春†

概要：Robocup サッカーシミュレーション 2D リーグはソフトウェア同士がコンピュータ上でサッカーをするリーグである．オープンソースの agent2d のプレイヤエージェントは「chain action」という枠組みを実装しており，探索木と評価関数を用いてボールを保持した場合の行動決定を行っている．本研究では，評価関数の重みの学習に，エピソードベースの方策勾配法(PGL)と各時刻の行動価値の推定値を用いて学習することができる Q 学習(QL)を併用して，効率的に学習することを試みた．その結果，agent2d に対して PGL,QL それぞれ単独で学習させた勝率は 4%と 11%であったが，PGL と QL を組み合わせた勝率は 43%となり，単独で学習したものと比べて大きく勝率が向上した．

1. はじめに

Robocup サッカーシミュレーション 2D リーグはソフトウェア同士がコンピュータ上でサッカーをするリーグである．サッカーはチェスや将棋などのボードゲームと異なるいくつかの特徴がある [1]．一つ目はチームプレイが要求されることである．サッカーは 11 対 11 の多人数ゲームであるため，協調行動が必要不可欠な要素となる．二つ目は，実時間でゲームが行われるので，瞬時に行動を決定しなければならない．三つめは情報が部分的で不確実なことである．サッカーでは自分の視覚内の情報しか取得できず，その情報もノイズを含んでいる．以上の特徴などから，マルチエージェントシステムや協調行動について研究するためのテストベッドとして用いられている．

このリーグでは多くのチームが agent2d [2]というサンプルチームをベースにしている．agent2d(ver3.0.0)では，「chain action」という枠組みを実装しており，プレイヤが探索木と評価関数を用いてボールを保持した場合の行動決定を行っている．しかし，agent2d で用いられている評価関数はボールの位置のみを考える単純なものであったため，谷川らは評価関数を新たに考案し重みの強化学習を行った [3]．しかし，3000 試合学習しても agent2d に勝ち越すことはできなかった．田川らはこの原因は報酬の質にあると考え，報酬として人間の主観評価を用いるオンライン強化学習システムを開発した [4]．このシステムでは，わずか 10 試合の学習で効果的なスルーパスの発生回数を増加させることができた．一方，大内はレシーバの移動位置の決定に chain action を適用し，強化学習を試みた [5]．また，山岸らは評価関数に状態の他に行動の良さを考慮する項を導入し，教師あり学習を試みた [6]．

上記の強化学習を用いた研究例では，エピソードベースの方策勾配法という手法が用いられてきた．しかし，この手法は報酬が与えられる機会が少ないと十分に学習することができない可能性がある．そこで，本研究では直接与える報酬がなくても各時刻の行動価値の推定値を用いて学習することができる Q 学習を併用して，より効率的に行動を学習させることを目的とした．

2. サッカーシミュレーション 2D リーグ

RoboCup サッカーシミュレーション 2D リーグ [7]は，実機を使わず高さがない 2 次元フィールド上で 11 対 11 のプレイヤがサッカーを行うリーグである．このリーグの試合はサーバクライアント方式でシミュレートされており，以下のような流れでシミュレーションが行われている．

① プレイヤはセンサ情報をサーバプログラム(rcssserver)から取得
② 各自で行動決定を行い，サーバに kick や dash などの行動コマンドを送信
③ 試合終了でなければ①に戻る．

しかし，サーバから受け取るセンサ情報にはノイズが含まれているため，不完全な情報を基に行動決定を行わなければならない．さらに，プレイヤ同士のサーバを介さない直接的な通信はルール上禁止されている．これらの制約があるため，協調行動を実現するための工夫が必要となる．

3. プレイヤの行動決定

3.1 chain action を用いた行動決定

chain action 生成システムはパスやドリブルなどのボール保持者の行動を「枝」とし，行動後の試合局面(状態)を「ノード」とした探索木を作成する．次に評価関数によって全ノードを評価し，最良優先探索によって評価値の最も大きなノードに至るルート直下の行動が選ばれる [8]．図 1 に chain action の例を示す．この図では，*a*,*b*,*c* が選択対象となる行動，S_1～S_8が状態，数値が状態の評価値を示している．この例では，S_7の評価値が最も高いためルートからの次の行動として *b* が選択される．本研究でも，学習時以外はこの行動決定に従う．

† 芝浦工業大学
Shibaura Institute of Technology

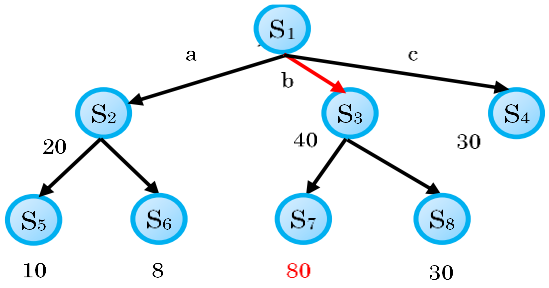

図 1　Chain Action の例

3.2　学習中の行動の抽象化

chain action での行動生成では図 2 のように極端に目標地点が多い行動が候補として生成される場合がある．3.4 で述べるように，学習時には確率的に行動を選択するため，図 2 の場合，候補点の個数が多いプレイヤ 3 へのパスが他のプレイヤのパスに比べて高い確率で選択される．また，前方へのパスやドリブルも候補数が少ないので選択されにくく，後方への安全なパスばかりが選択されてしまう傾向がある．したがって，極端に候補点が多い特定のプレイヤに対してのパス行動ばかり選択され，学習に偏りができる可能性がある．

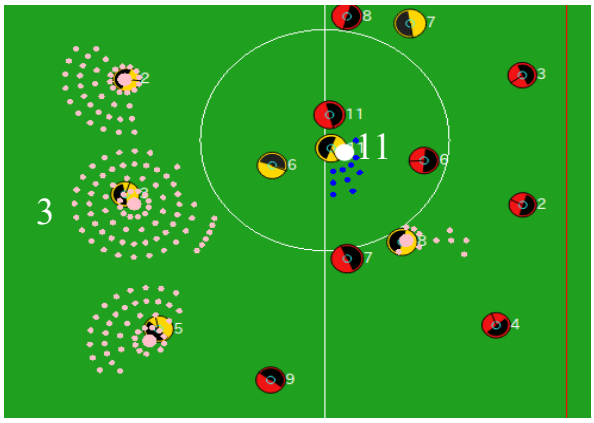

図 2 「chain action」の行動生成例
(プレイヤ 11 番のパスとドリブルの候補点)

そこで，本研究の学習では様々な行動を選択させるために行動生成後に「行動の抽象化」を行った．従来，chain action では一つの方向に対して複数の目標地点を生成していた（例，スルーパスは 16 方向×15）．しかし，この方法では近くの場所に同じような行動が多数生成されてしまう．従って，本研究では一つの方向に対して一つの目標地点を生成するように変更を加えた．また，各行動が生成する方向は最大 8 方向になるように調整した．これは，パスとドリブルで生成する方向の数が異なっていたためである．変更後の行動生成は図 3 のようになる．抽象化後は目標地点が大幅に減少したことにより，変更前に比べ様々な行動を選択することが期待できる．

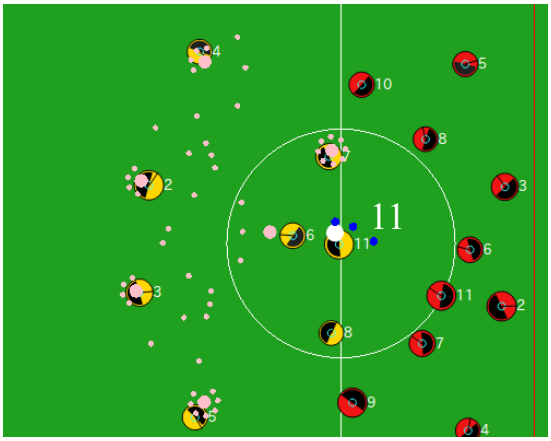

図 3 抽象化を行った行動生成例
(プレイヤ 11 番のパスとドリブルの候補点)

3.3　ボール非保持者への chain action の適用

agent2d のレシーバの行動決定では Delaunay Triangulation を使用してレシーバの移動位置を決定する手法を用いている [2]．しかし，この手法はあらかじめ作成したボール位置ごとのプレイヤ配置のサンプルを基に移動先の位置を計算する手法であり，敵プレイヤにマークされてもマークを外す動きをしないという問題点がある．そこで大内ら [5]はレシーバの移動先地点の決定に chain action を適用することを提案した．ただし，レシーバの人数は多いので，計算量の関係で探索木の深さを 1 に制限した．レシーバが作成する探索木の例を図 4 に示す．

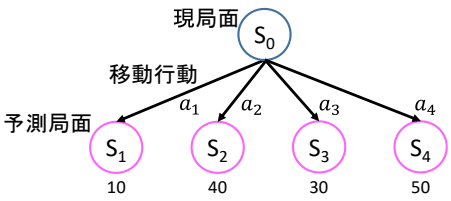

図 4 レシーバの探索木の例 [5]

図 4 では，a_1～a_4が移動行動，S_0～S_4は状態，数値はノードの評価値を表している．この例では，S_4が最も高い評価値であるため次の移動行動はa_4となる．

大内らによると，chain action の適用と強化学習によりレシーバはパサーにとって良い位置取りをするようになり，ゴール前でのパス回しによる得点が増加したことが報告されている [5]．本研究ではレシーバの行動選択としてこの方式を利用する．

3.4　学習中の確率的方策の適用

agent2d では探索木に対して最良優先探索により決定論的に行動を決定していた．しかし，谷川 [3]や，田川ら [4]

の研究では学習を行うために以下のような Boltzmann 分布による確率的な方策を利用している．

$$\pi(a_t|s_t;\omega)\equiv\frac{e^{E(s_t,a_t;\omega)/T}}{\sum_{x\in A(s)}e^{E(s_t,a;\omega)/T}} \quad (1)$$

ただし，$A(s)$は局面 s における行動集合，T は温度パラメータ，ωは評価関数中のパラメータである．

さらに，確率的方策を利用するために，ルート局面 s における行動 a の評価関数$E(s,a;\omega)$を，その行動から派生する全ノード中で最大の局面評価値$E(s_a;\omega)$で置き換える．すなわち(1)式は(2)のようになる [3] [4].

$$\pi(a_t|s_t;\omega)\equiv\frac{e^{E(S_a;\omega)/T}}{\sum_{x\in A(s)}e^{E(S_x;\omega)/T}} \quad (2)$$

ここで，S_aは局面 S において行動 a 以下の部分木での局面評価値$E(S_a;\omega)$が最大の局面（ノード）を表す．

ただし，学習後の重みを使用して試合をする際には T=0 とした 3.1 を用いる．

4. 評価関数

4.1 重みの切り替え

山岸拓海らの研究 [6]ではフィールドの場所により，重みの切り替えを行っていた．本研究でも重みの切り替えを行う．重みを切り替える位置は図 5 重みの切り替えのようになる．重みの切り替えを行う理由は中央にいるときとゴール付近にいる時では望ましい行動が異なるためである．中央にいる時は安全にスルーパスやドリブルで x 座標（フィールド中央を原点とし，原点から敵ゴール方向を x 方向とする）が敵ゴール側に近づく行動などをする必要がある．しかし，ゴール付近にいる時は多少リスクがあっても敵ゴールに向かうような行動をする必要がある．従って，別々の重みで学習を行うほうが良いと考えられるので，両方の重みのセットω^1,ω^2を用意した．

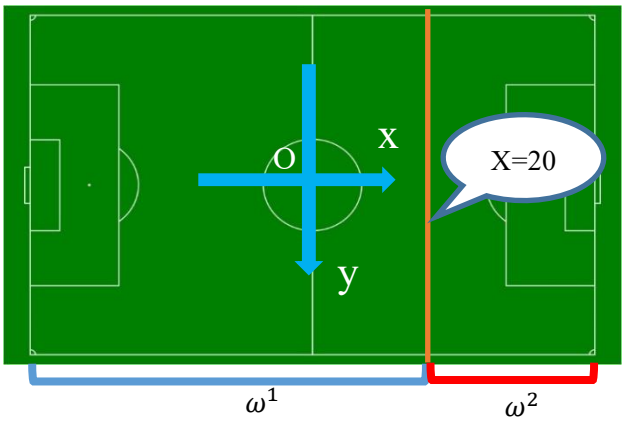

図 5 重みの切り替え（左側が自ゴール）

4.2 評価項目

本研究では山岸拓海らの研究 [6]で考案された評価関数を用いた．この評価関数は(3)に示すような関数で表される．(3)の前半の項では，状態だけではなく，行動の良さを評価する項が含まれている．評価関数の各項の概要を表 1,2 に示す．

$$E(s,a;\omega)=\sum_{i=1}^{n}\omega_i U_i(s,a)+\sum_{j=n+1}^{m}\omega_i U_j(s) \quad (3)$$

$$(0\le U_i\le 10)$$

表 1 ボール保持者の評価内容 [6]

評価項	評価内容
$U_1(s,a)$	パスコースと敵の最短距離
$U_2(s,a)$	ボールの移動距離
$U_3(s)$	ボールと敵ゴールの距離
$U_4(s)$	ボールに最も近い敵との距離
$U_5(s)$	ボールより敵ゴール側にいる敵人数

表 2 ボール非保持者の評価内容 [6]

評価項	評価内容
$U_1(s,a)$	パスコースと敵の最短距離
$U_2(s)$	自身に最も近い味方との距離
$U_3(s)$	自身と敵ゴールの距離
$U_4(s)$	自身に最も近い敵の距離
$U_5(s)$	自身より敵ゴール側にいる敵人数
$U_6(s)$	自身とオフサイドラインの距離

5. 評価関数の強化学習

本章では本研究で使用する方策勾配法と Q 学習について述べる．

5.1 方策勾配法の学習則

学習するエピソード（σとする）を定義し，エピソード終了時にその時点の状態やエピソード全体に対して評価し，報酬を与える [9]．エピソードあたりの報酬の期待値を最大化するために，確率的勾配法を用いて評価関数のωを更新する．

学習則は以下の(4),(5)のように表される．学習中は Boltzmann 分布による確率的な方策(2)を用いる．

$$\Delta\omega_{PGL}(\sigma)=\varepsilon\cdot r\cdot\sum_{t=0}^{L-1}e_\omega(t) \quad (4)$$

$$e_\omega(t)\equiv\frac{\partial}{\partial\omega}\ln\pi(a_t|s_t;\omega) \quad (5)$$

ただし，s_tは時刻tにおける局面，a_tは選択された行動，L

はエピソード長，εは学習係数である．

5.2 Q 学習の学習則

Q 学習は状態-行動対(s,a)が多くなるほどテーブルが巨大になる．従って，本研究ではテーブルを関数近似する手法を用いる [10]．近似関数$Q(s,a;\omega)$を学習する際，(6)に表される最適行動価値関数$Q^*(s,a)$と$Q(s,a;\omega)$の誤差$V(t)$を最急降下法により，最小化する．

$$V(t)=\frac{1}{2}[Q^*(s_t,a_t)-Q(s_t,a_t;\omega)]^2 \quad (6)$$

ここで，$Q^*(s,a)$を$r+\gamma \max_a Q(s_{t+1},a;\omega)$で近似し，$Q(s_t,a_t;\omega)$の近似としては(3)の評価関数$E(s,a;\omega)$を用いる．学習則は(7)のようになる．

$$\Delta\omega_{QL}(t)=\alpha\cdot[r(t)+\gamma \max_a E(s_{t+1},a;\omega)-E(s_t,a_t;\omega)] \cdot \nabla_\omega E(s_t,a_t;\omega) \quad (7)$$

ただしαは学習率，γは割引率である．

5.3 報酬関数

攻撃時のプレイヤに対して，エピソード(σとする)に対する報酬$r_{PGL}(\sigma)$を表 3 に示すr_1～r_3の和として与えた [11]．一方，Q 学習で与える報酬は方策勾配法と違い，マルコフ性を有する必要がある．そこで，報酬$r_{QL}(t)$を表 4 に示すr_1～r_3の和として各時刻 t ごとに与えた [11]．

表 3 方策勾配法で利用する報酬関数$r_{PGL}(\sigma)$

評価項	評価内容
$r_1(\sigma)$	エピソード最初と最後のボールとゴールまでの距離の差
$r_2(\sigma)$	最後にペナルティエリア内でシュートができた角度
$r_3(\sigma)$	エピソード最初と最後のボールとディフェンスラインの距離の差

表 4 Q 学習で利用する報酬関数$r_{QL}(t)$

評価項	評価内容
$r_1(t)$	行動前と行動後のボールとゴールまでの距離の差
$r_2(t)$	シュート可能なゴールエリアの角度
$r_3(t)$	行動前と行動後のボールとディフェンスラインの距離の差

また，それぞれの評価項の例を図 6～図 8 に示す．

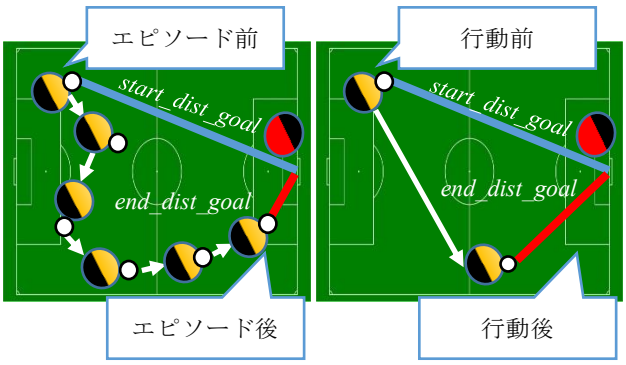

図 6 $r_1(\sigma)$と$r_1(t)$の例

$r_1(\sigma)$はエピソード全体でボールがゴールに使づいた距離を評価する項であり，$r_1(t)$は 1 行動でボールがゴールに近づいた距離を評価する項である．*start_dist_goal* が長く，*end_dist_goal* が短いほど評価が高くなる．

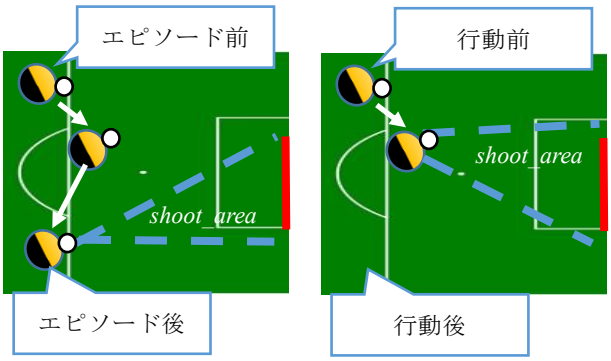

図 7 $r_2(\sigma)$と$r_2(t)$の例

$r_2(\sigma)$はエピソード中のシュートチャンスを評価する項であり，$r_2(t)$は行動後のシュートチャンスを評価する項である．シュートできる角度，すなわちシュート可能なエリアの大きさ(*shoot_area*)が大きいほど評価が高くなる．

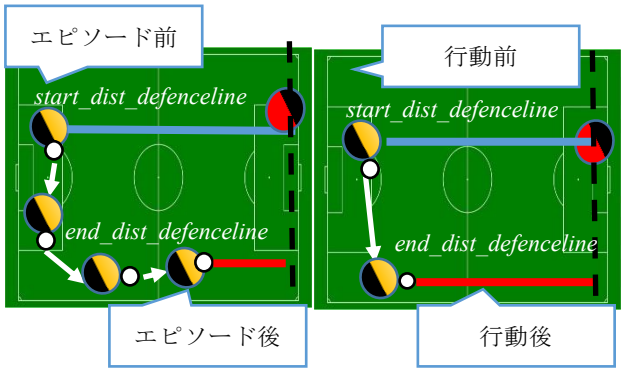

図 8 $r_3(\sigma)$と$r_3(t)$の例

$r_3(\sigma)$はエピソード全体でボールがディフェンスラインに使づいた距離を評価する項であり，$r_3(t)$は 1 行動でボー

ルがディフェンスラインに近づいた距離を評価する項である．*start_dist_defenceline* が長く，*end_dist_defenceline* が短いほど評価が高くなる．

5.4 エピソードの定義

報酬関数により報酬を与えていた谷川 [3]は味方がボールを持ってから相手にボールを取られるまでを 1 エピソードと定義していた．しかし，この研究ではフィールドの全体で同じ重みを使用していた．しかし，本研究では 4.1 で述べたように重みの切り替えを行っているため，新しくボール保持者のエピソードの終了条件を図 9 のように定義した．

図 9 ボール保持者のエピソード終了条件

①で切る理由は重みが切り替わるためである．$x \leq 20$でエピソードが開始した場合，ゴール付近で無意味な行動をとり続けても，一連の行動を考えるとゴールに近づいているため高報酬が与えられる．これが原因で$x>20$の重みが無意味な行動を良い行動だと学習してしまう恐れがある．従って，①をエピソードの終了条件としている．また，ペナルティエリアでもエピソードを終了させている．これは，シュートチャンスになるペナルティエリア内に侵入する行動を学習させたいと考えたためである．変更前と変更後のエピソード例は図 10 のようになる．

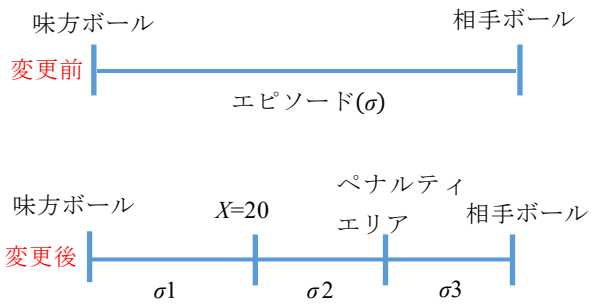

図 10 変更前と変更後のエピソード例

一方，ボール非保持者のエピソードの定義は大内らの研究 [5]と同様である．

6. 方策勾配法と Q 学習の同時適用

本研究ではボール保持者に対してはエピソードσが終了した時点($t=L$)で方策勾配法と Q 学習によるパラメータの更新を行う．従って，方策勾配法と Q 学習の同時適用時の学習則は(8)のようになる．

$$\Delta\omega(\sigma)=\Delta\omega_{PGL}(\sigma)+\sum_{t=0}^{L-1}\Delta\omega_{QL}(t) \tag{8}$$

また，ボール非保持者には Q 学習は適用しない．なぜならば，移動行動 a は完了するまでに別の行動が選択されてしまい，行動 a による遷移先の状態を得ることができないからである．

7. 学習実験

本研究ではディフェンシブハーフ(DH)1 人，オフェンシブハーフ(OH)2 人，サイドフォワード(SF)2 人，センターフォワード(CF)1 人のボール保持者に対して方策勾配法と Q 学習を行った．一方，ボール非保持者は DH,OH,SF,CF に対しては，大内 [5]と同様な方策勾配法のみを行った．対戦相手は agent2d，学習数は 100 試合，学習率εとαはそれぞれ 0.01 と 0.001 である．これは方策勾配法と Q 学習の更新を同程度進行させるように値を調整した結果である．温度 T は 10，割引率γは 0.9 であり，重みの初期値はすべて 1 に設定した．

学習後，他に比べて特に大きくなった重みと小さくなった重みは表 5,6 のようになった．表 5,6 の「+」は最も大きくなった重みを表している．また，最大値×0.9 以上の値があった場合にも「+」の記号を付けている．一方，「-」は最も小さくなった重みを表している．また，最小値×1.1 以下の値があった場合にも「-」の記号を付けている．

表 5 学習後の重みω^lの特徴（$x>20$ の場合）

	ω_1			ω_2			ω_3			ω_4			ω_5		
	P	Q	PQ	P	Q	PQ	P	Q	PQ	P	Q	PQ	P	Q	PQ
CF	+			+			+	+		+		-	+	-	+
SF	-	-	-	+				+	+						+
OH				-			-	+	+			-	+	-	
DH				+	-			+	+	-		-		+	

※P：方策勾配法，Q：Q 学習，
+：特に大きくなった重み，-：特に小さくなった重み

表 6　学習後の重みω^2の特徴（x≦20の場合）

	ω_1			ω_2			ω_3			ω_4			ω_5		
	P	Q	PQ	P	Q	PQ	P	Q	PQ	P	Q	PQ	P	Q	PQ
CF				-				+	+	+	-	-	+		
SF		-				-			+	-			+	+	
OH	-			-				+		+		-		+	+
DH	+			-	-			+	+			-			+

※P：方策勾配法，Q：Q 学習，
＋：特に大きくなった重み，-：特に小さくなった重み

表 5,6 から，方策勾配法の結果と Q 学習の結果が異なる重みがあったことが分かる．これは，方策勾配法と Q 学習では$\Delta\omega$の更新方向が異なるためだと考えられる．また，方策勾配法と Q 学習を同時適用したチームは，方策勾配法で小さかったものが Q学習によって大きな値に修正されるなど，お互いの学習結果に影響を与えていた．従って，一つの学習則のみを適用したチームとは違う行動が学習できたと考えられる．

8. 評価実験

①未学習チームと②方策勾配法のみの学習チーム，③Q 学習のみの学習チーム，④方策勾配法と Q 学習の同時学習チームそれぞれが agent2d と 500 試合行った結果を表 5 に示す．

表 7 agent2d との対戦結果(500 試合)

	勝率	勝-負-分	平均得点	平均失点
①	1.9%	8 -414- 78	0.12	1.85
②	3.9%	16 -398- 86	0.22	1.97
③	10.6%	44 -371- 85	0.61	2.23
④	42.8%	166 -222- 112	1.73	1.97

※勝率は引き分けを除く

表 7 より，①の未学習チームと②の方策勾配法のみを行ったチームは約 2～4%の勝率であった．一方，③の Q 学習のみを行ったチームは約 11 パーセントの勝率であり，②の勝率を上回った．これは，Q 学習が方策勾配法と比べてより細かく行動に対して報酬を与えるためだと考えられる．

次に，④の方策勾配法と Q 学習を同時適用したチームは約 43%の勝率となり最も高かった．特に，③に比べて④は得点力が約 3 倍に上がっている．これは，方策勾配法によるエピソード全体に対する報酬と Q学習による各行動に対する報酬がうまく組み合わさることにより，より多くの価値基準でお互いを補い合うような学習ができたからだと考えられる．

9. 結論

本研究では，方策勾配法によるエピソード全体を考慮した学習にQ学習による行動単体の学習を組み合わせた．方策勾配法単体のものは勝率約3%，Q学習単体のものは勝率約11%であったのに対し，同時学習を適用したチームは勝率約43%となり，勝率を大きく上昇させることができた．

今後は，ボール非保持者にも同時学習を適用することでより勝率を上げることができる可能性がある．また，本研究で提案した報酬関数にも改善の余地があり，より良いヒューリスティクスを取り入れることが考えられる．さらに，本研究では Q 関数の関数近似や，行動決定の際に行動の良さを評価する評価関数に，行動や状態の特徴量の線形関数を使用したが，ニューラルネットワークのようなより豊富な表現が可能な非線形の関数を使用することも今後は必要だと考えている．

参考文献

[1] 松原仁, 竹内郁雄, 沼田寛, ”ロボットの情報学 2050 年ワールドカップ，人間に勝つ?”,NTT 出版,2001.
[2] Hidehisa Akiyama,Tomoharu Nakashima,”HELIOS Base : An Open Source Package for the RoboCup Soccer 2D Simulation”, RoboCup2013 : Robot World Cup XVⅡ,pp.528-535,2013.
[3] 谷川俊策, 五十嵐治一, 石原聖司, ”RoboCup サッカーシミュレーションリーグ 2D における局面評価関数の学習”, GPW2013 論文集, pp.106-109, 2013.
[4] 田川諒, 五十嵐治一, ”サッカーエージェントにおけるスルーパスの強化学習”, FIT2016, F-42, 2016.
[5] 大内斉, 五十嵐治一, ”局面評価関数を用いたサッカーエージェントの移動先決定”, GPW2016 論文集, pp.49-56, 2016.
[6] 山岸拓海, 五十嵐治一, 山岸準, 入倉雅春, ”サッカーエージェントの攻撃時における評価関数：方策勾配法を用いた教師あり学習”, 第 34 回ファジィシンポジウム講演論文集, pp.682-687, 2018.
[7] 秋山英久, “ロボカップサッカーシミュレーション 2D リーグ必勝ガイド”, 秀和システム, 2006.
[8] 秋山英久, ”連続行動空間での木探索によるオンライン協調行動プランニング”, 情報処理学会研究報告, Vols.2012-GI-27, No.11, pp.1-8, 2012.
[9] 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勾配法の適用-追跡問題-“, 電子情報通信学会論文誌(D-I), Vol.J87-D1, No.3, pp.390-397, 2004.
[10] Richard S.Sutton, Andrew G.Barto, ”強化学習”, 三上貞芳, 皆川雅章訳, 森北出版, pp.209-227, 2000.
[11] 山岸準, ”サッカーエージェントにおける方策勾配法と Q 学習の同時適用”, 芝浦工業大学大学院修士論文, 2019