3J-5

# 交渉過程を導入したマルチエージェント強化学習の報酬配分

Reward Distribution of Multi-agent Reinforcement Learning in Negotiation Process

鶴岡　久 (Tsuruoka Hisashi)

福岡工業大学情報工学部(Fukuoka Institute of Technology)

## Ⅰ．はじめに

目標がいずれかのエージェントで達成されればよく、目標の達成により報酬を得たエージェントがその場で報酬を独占すると短期的には得をするが、他のエージェントの貢献を考えて報酬の一部を分配すれば、長期的に得することがありうるシステムを考える。このシステムでは各エピソードの終わりで環境から与えられる正報酬をエージェントに割り振る報酬分配が問題になる。

シングルエージェントで報酬を獲得するまでの行動系列に対して報酬を分配する問題も広義の報酬分配問題であり、profit sharing のような経験強化型では報酬に貢献しない無効ルールの抑制や報酬からどれだけ過去かを引数として強化値を返す強化関数に関する研究がある。これに対してマルチエージェントでの報酬分配や強化関数の研究例は少ない。

## Ⅱ．交渉過程を取り入れた報酬分配

従来、マルチエージェント同士の協調が目標達成に必要とされる典型的問題として追跡問題が考察されてきた。ここでは捕獲条件を一般化し、ハンター1 匹でも、複数のハンターが協力しても、獲物を捕獲できるとする。各エージェントが個別にその場限りの報酬の期待値の最大化を図れば、競合状態が始まり報酬分配は起きず、獲物を獲得したハンターの独占になる。しかし集団（系）の利益の最大化が図るには独占状態が唯一の解ではない。協力か、独占かは獲物を捕獲したエージェントの報酬分配法にかかっている。またエージェント同士の協力が本来有利か、否かは獲物の逃走政策にも依存する。さらにエージェントが個別の報酬関数をもち、ハンター間の能力、動作等に不均衡が存在し、ハンター1 匹でも捕獲可能なヘテロなエージェント同士で獲物を競合する場合の報酬配分は自明ではない。

これらの問題に対処するためここでは報酬分配を交渉過程としてとらえる。一匹のハンターが獲物を捕まえたとき、他のハンターと交渉に入る。交渉では態度、手（例えば独占、半独占、協調（遠慮）の３タイプ）を互いに表明する。提示された態度の組み合わせから決まる利得分配率表から報酬分配する。交渉は戦略学習となっており、互いの報酬の期待値を最大化するようその都度適切な態度を表明する。このような交渉プロセスも追跡学習と同様強化学習（Q学習）で扱う。報酬分配機構は各学習エージェントであるハンターの中に個別に埋め込まれので、各ハンターは追跡の学習と報酬分配の方法を同時に学習することになる。

本実験で利用する利得分配率表を下記に示す。

| 主 | 独占 | 半独占 | 協調 |
|---|---|---|---|
| 独占 | 0.5,0.5 | 0.25,0.75 | 0,1 |
| 半独占 | 0.75,0.25 | 0.5,0.5 | 0.25,0.75 |
| 協調 | 1,0 | 0.75,0.25 | 0.5,0.5 |

追跡問題では学習過程で捕獲ステップが最小値に収束するまで繰り返えされる繰り返しゲームとみなせる。これは捕獲のたびに繰り返し交渉を行い、この交渉プロセスを追跡学習と同様強化学習で扱い、さらに相手行動や相手の報酬分配政策を予測することで、独占という戦略に勝るどのような報酬分配の動的変化が生まれるか、分配率を学習中一定にした場合に比較して最短捕獲ステップを獲得するまでの学習速度は速まるかどうか、を調べる。交渉動作のモデルは有限オートマトンで表現する。

## Ⅲ．実験結果

実験で取り扱う共通のパラメータ設定は下記の通りである。

学習率　α＝０．３、割引率　γ＝０．９、温度Ｔ＝０．１、遅延報酬は－０．１とする。フィールドは５×５（壁あり）、ハンターは同時に動作し、相手ハンターと通信はできないが、動作や位置は観測できるものとする。

**実験１：**　ハンターを均質エージェントとする。獲物の行動は、いずれかのハンターと獲物の距

離＞＝L2　のとき　ｙ＝３上をランダム行動し、そうでなければ、獲物は２匹のハンターの方向ベクトルから９０度もしくは１８０度曲がった方向を選択する。

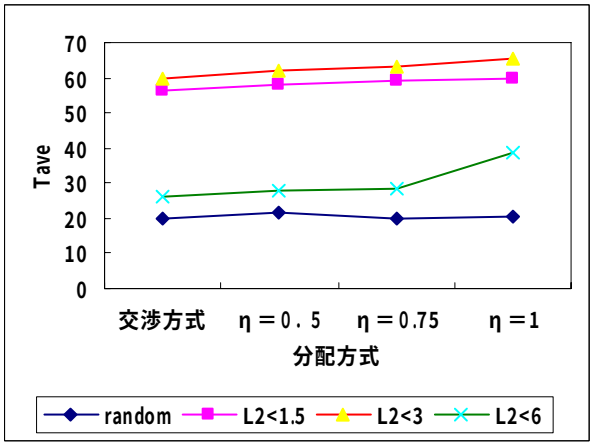



図１　追跡問題における学習時間の比較

図１は捕獲学習所要時間の１００回平均を示し、所要ステップが７以下になる試行が５回以上継続することをもって収束と判定している。ランダム行動では分配率 η を固定した場合、報酬を等分配することは不利であることを示している。

L2＜1.5 の場合は交渉方式は固定分配方式に比べて有利性がはっきりしている。これは獲物とハンターの距離が遠い場合、獲物はランダム行動、近い場合は両ハンターの動きに対応して動くので、ハンターとしては競合と協調の戦略を織り交ぜる必要があり、この点が交渉方式に有利に働いていると考えられる。L2<6 ではほぼフィールド全域で獲物は両ハンターの動きに呼応した逃げ方をした場合であり、固定的に捕獲者に独占分配させるのは不利であることを示している。このように獲物が両ハンターの動きに対応して逃走する場合、ランダム行動をとった場合の、いずれに対しても交渉方式は固定分配方式に比べて学習時間で有利であることが示された。

**実験２．** ハンターにヘテロな能力を与えるためハンター１は獲物と同一座標、ハンター２は獲物との距離の２乗が３以下で捕らえられるとする。獲物は１ステップ左、停止、右の動作を繰り返す。図８に学習時間と分配方式の関係を示す。交渉方式の学習が収束する過程ではハンターが取っている態度（手）は、下記のプロセスのいずれかが観察される。

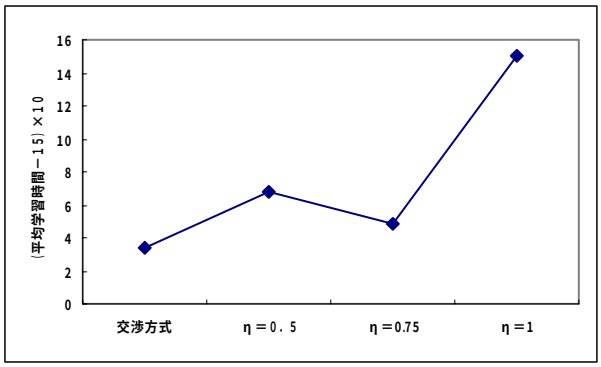



図８ヘテロエージェントにおける学習時間

（半独占、協調）→（独占、独占）→(協調、協調)の繰り返し
（独占、協調）→（半独占、半独占）の繰り返し
(半独占、半独占)　（独占、半独占）の繰り返し
ハンター２は楽に捕まえられるため、学習過程で捕まえる頻度は多いにも拘わらず、独占報酬はあまり要求していない。もし η＝１の固定方式とすると、ハンター２の行動のみが強化され、η＝０．５とするとハンター２の行動強化は弱い。交渉方式はそのあたりのバランスがとれているように見える。

## Ⅳ　結言

交渉方式による報酬分配を追跡問題に適用し、報酬分配率を固定化する場合に比べて、最短捕獲ステップに到達する学習時間が短縮されることを明らかにした。獲物の逃走戦略が両ハンターの動きを意識した場合には協調が有利となるが、そうでなければハンターは獲物を独占する方向が一般には有利になる。ところが獲物がこれらの戦略を織り交ぜていると分配率が固定的に設定されるよりも交渉方式が有利となることを示した。

またヘテロな捕獲能力をもった追跡問題でも、提案した交渉方式によれば、交渉過程を通して動的に合理的に報酬が分配されていくことを示した。

## Ⅴ　参考文献

１）伊藤他；マルチエージェント強化学習による交渉問題へのアプローチ、信学技報、AI2003-84、２００４年３月