RF-007

マルコフ的動的分散制約最適化問題への非厳密解法の適用
Applying Incomplete DCOP Solver to MD-DCOPs

増田清貴 †
Kiyotaka Masuda

松井俊浩 †
Toshihiro Matsui

1. はじめに

複数の自律的な主体が分散・協調的な処理を行う枠組みである，マルチエージェントシステムが研究されている．マルチエージェントシステムの技術は，地理的に分散して配置された自律的なセンサ群による，広域観測システムなどへの応用が期待されている．このようなマルチエージェントシステムにおける協調的問題を解決する枠組みとして，分散制約最適化問題 (DCOP: Distributed Constraint Optimization Problem) [1, 2, 3] が研究されている．DCOP では，エージェントの状態や意思決定が変数として表現され，エージェント間の関係が制約と評価関数として表現される．各エージェントは互いに情報を交換しつつ自身の変数値を決定し，評価関数の評価値を結合した値を最適化する変数値を得る．このような表現は，分散システムにおける資源スケジュールなどに含まれる，協調問題解決の本質的な問題を表すものとして重要である．

マルチエージェントシステムの応用分野として分散センサ網 [4] の研究が行われている．文献 [4] では強化学習を用いて分散センサ網を解く手法が提案されている．分散センサ網に DCOP を適用する場合は，ターゲットが時間とともに移動する動的な問題が動的分散制約最適化問題 (D-DCOP: Dynamic DCOP) [5, 6] としてモデル化される．

しかし，現実的な分散センサ網では，ターゲットの移動がセンサの動作に影響を受けることが考えられる．たとえば，ターゲットが，自身が観測されたことを知り，逃避的な行動を取ることは自然である．そのような逃避的な行動を考慮しなければ，十分な観測が出来ない可能性がある．このように，現在の状態が将来の状態に影響する，マルコフ性を伴う動的な問題は，マルコフ的動的分散制約最適化問題（MD-DCOP: Markovian Dynamic DCOP) [7] としてモデル化される．

MD-DCOP は DCOP と強化学習を統合させたものである．分散センサ網における MD-DCOP では，強化学習により，ターゲットがどのような移動戦略を持っているか学習をする．学習により得られたターゲットの移動戦略に基づいて，センサの適切な割り当てを決定する．マルチエージェント強化学習の一部を DCOP として扱うことにより，行動選択などの強化学習において必要な情報を，分散制約最適化アルゴリズムにより求める．このとき，従来手法 [7] では，DCOP の厳密解法である DPOP [3] を用いる．DPOP は最適解を求めることができるが，変数と制約の数が増え問題の規模が大きくなると，計算量やメッセージのサイズが指数関数的に増加するため，大規模で複雑な問題への適用が困難である．そこで本研究では，非厳密解法である DSA [1] を用いることにより, 大規模かつ複雑な問題においても MD-DCOP を適用できるように改良する手法を提案する．

本論文の構成は以下の通りである．第 2 章では研究背景について説明する．まず，本論文で対象とする分散センサ網について説明する．次に，DCOP と D-DCOP について説明する．そして，本研究の背景となる MD-DCOP と MD-DCOP の解法と関係する強化学習について説明する．また，マルチエージェント強化学習において用いる，DCOP の解法である DPOP と DSA について説明する．第 3 章では提案手法について述べる．提案手法の基本的なアイデア，DSA を用いたときの MD-DCOP のアルゴリズムの概要，期待される効果と影響について説明する．第 4 章では，DPOP アルゴリズムを用いた手法と DSA アルゴリズムを用いた手法を実験により比較し，評価する．第 5 章では本研究のまとめと将来的な展望を示す．

2. 研究背景

2.1. 分散センサ網

本節では本研究で対象にする分散センサ網のモデルについて述べる．分散センサ網は複数のセンサと複数のターゲットからなる．各センサは観測することができる領域をいくつか持つ．センサが観測できる領域は自身と付近の他のセンサとの間にある．センサはいずれかの領域を観測し，ターゲットを捕捉したとき，そのセンサは報酬を得る．ターゲットは，センサが観測できる領域のいずれかに，移動することができる．また，ターゲットは，センサが観測した領域を知ることができる．センサは図 1 のようにグリッド状に配置される．グリッドの 4 近傍にあるセンサを近傍のセンサと呼ぶ．センサが観測できる複数の領域は，近傍センサとの間にある．

2.1.1. 報酬

前述で述べたように，センサはターゲットを捕捉したとき報酬を得る．報酬は次のように設定される．

- いずれか一つのセンサがターゲットを捕捉したとき，そのセンサは報酬を得る．
- 一つのターゲットを二つのセンサで同時に捕捉したとき，それらのセンサは同時に報酬を得る．この報酬はセンサ単独でターゲットを捕捉したときの報酬よりも，大きい．
- 報酬はターゲットを捕捉する領域にも依存する．より重要な領域で捕捉したとき，より大きな報酬を得る．

分散センサ網の目的は，センサのターゲットへの割り当ての評価を表す，報酬の総和を最大にすることである．

†名古屋工業大学，Nagoya Institute of Technology

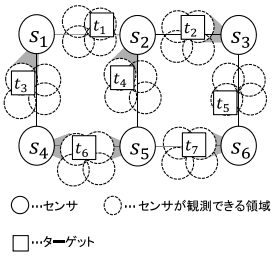

図 1: ターゲットが動かない分散センサ網

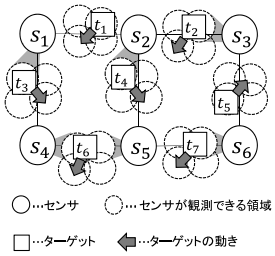

図 2: ターゲットが動く分散センサ網

2.1.2. ターゲットの動作

センサをターゲットへ割り当てる問題は，センサの観測とターゲットの動作の条件にもとづいて設定される．以下では，ターゲットが動く場合と動かない場合に分けて，最適なセンサの割り当てを示す．

(a) ターゲットが動かない問題

ターゲットが動かない場合の例を図 1 に示す．この問題では，位置が固定されたターゲットへの，センサの最適な割り当てを求める．したがって，評価値の和が最も大きくなるようにターゲットにセンサを割り当てればよい．

(b) ターゲットが動く問題

ターゲットが動く場合の例を図 2 に示す．このような問題では，ある時間ステップが経過するごとにターゲットの位置が変化する．このとき，センサの観測がターゲットの動作に影響することを考慮するか否かにより，問題の設定が異なる．

1. センサの観測の影響を考慮しない場合

   ターゲットの動作における，センサの観測の影響を考慮しない問題では，その時間ステップではターゲットが動かないものとして，時系列的な問題をそれぞれ解く．すなわち，各時間ステップにおいて，評価値の和が最も大きくなるようにセンサを割り当てる．

2. センサの観測の影響を考慮する場合

   ターゲットの動作における，センサの観測の影響を考慮する問題は，ある時間ステップのセンサの割り当てが，次の時間ステップ以降のターゲットの移動に影響するマルコフ的な問題として表される．センサは各時間ステップにおいて，ある戦略に基づいて観測を行うことを反復する．ターゲットはセンサの割り当てに基づいて，次の時間ステップの移動先を決定する．このような問題では，ターゲットの動きに基づいて，ある期間の評価値の和の平均が最大となるように，ターゲットにセンサを割り当てる．

本研究では，(b) 2 に示した，ターゲットの動作がセンサの観測の影響を受けるような，分散センサ網上の割り当て問題を対象とする．

2.2. 分散制約最適化問題

分散制約最適化問題（DCOP: Distributed Constraint Optimization Problem）は，マルチエージェントシステムにおける協調問題解決のための基礎的な枠組みとして，研究されている．DCOP では，マルチエージェントシステムを，各エージェントに変数と評価関数が分散して配置された離散最適化問題として形式化し，分散協調型の最適化アルゴリズムにより，問題を解決する．本節では，DCOP の形式化と解法，および分散センサ網への適用について説明する．

2.2.1.DCOP の形式的な表現

DCOP はエージェントの集合 A，変数の集合 X，各変数の値域 D，制約の集合 F からなる．各エージェント a_i は，自身の状態や意思決定を表す，変数を持つ．本研究では，一般的かつ簡単な設定として，各エージェント a_i が単一の変数 x_i を持つこととする．各変数 x_i は有限離散集合 D_i に含まれる値を取る．制約および，その評価関数 f_j はエージェント間の関係を表す．各変数は互いに，制約により関係する．各制約についての評価関数 f_j により，変数値の割り当てが評価される．各エージェントが協力し，全ての評価関数の合計値を最大化するような変数値の割り当てを決定することが，分散制約最適化問題の目的である．

2.2.2. 分散制約最適化問題の解法

DCOP の解法は，最適解を得ることが保証されている厳密解法と，最適解を得ることが保証されていない非厳密解法に分類される．

- 厳密解法

  厳密解法には，木探索に基づく解法 ADOPT [2] や動的計画法に基づく解法 DPOP [3] がある．これらの手法は，制約最適化問題のグラフ表現である制約網における，擬似木に基づく．擬似木は，一般にはグラフ上の深さ優先探索木に基づいて生成される，グラフ構造である．この擬似木により問題を分解し，木探索や動的計画法を用いて問題を解く．しかし，擬似木の誘導幅 [3] の増加にしたがって，探索に必要なサイクル数やメッセージの表の大きさが指数的に増加することが問題点である．

- 非厳密解法

  非厳密解法には，Distributed BreakOut Algorithm (DBA) [8] や Distributed Stochstic Search (DSA) [1] がある．これらの手法は，局所探索に基づく反復改善型のアルゴリズムである．DBA は特に制約充足問題のための解法であり，ブレークアウト法により局所解から脱出する．これに対し，DSA は確率的な局所探索を用いる．これらの解法は最適解を得ることを保証しないが，比較的短時間で解を得ることができる．

本研究で注目する MD-DCOP の従来研究 [7] では，解法の一部に DPOP を用いている．これに対し，本研究の提案手法では DSA を用いる．MD-DCOP の解法に用いる DPOP の説明を 2.6 節に示す．また，DSA の説明を 2.7 節に示す．

2.2.3. 分散センサ網の DCOP による表現

ターゲットが動かない静的な分散センサ網を DCOP として形式化する場合は，変数，エージェントはセンサを表し，変数値はセンサが観測する領域を表し，制約・評価関数はセンサのターゲットへの割り当ての評価を表す．DCOP の解法を用いて，評価値の総和が最大となる解を求めることにより，静的な分散センサ網におけるセンサの最適な割り当てが得られる．

2.3. 動的分散制約最適化問題

動的分散制約最適化問題 (D-DCOP) は，DCOP を，動的に変化する問題に拡張したものである．

2.3.1. 動的分散制約最適化問題とその解法

簡単な D-DCOP は，DCOP の系列として定義され，その解法は問題の系列を順に解く．系列の要素は，DCOP と同じ要素により定義されるが，各要素は時間とともに変化する．例えば，問題の系列を $P = p_1, \cdots, p_s, \cdots$ とするとき，時刻とともに問題が p_s から p_{s+1} と変化すると，エージェント間の制約および評価関数も変化する．ただし，本研究では，変数とその値域は変化しないこととする．そのため， 2.1.2 節の (b) 1 に示した，ターゲットの動作におけるセンサの観測の影響を考慮しない問題では，各時間ステップでは分散センサ網は静的であり，問題を D-DCOP として表現できる．

2.3.2. 分散センサ網の D-DCOP による表現

ターゲットが動く分散センサ網を，D-DCOP として形式化する場合を考える．本研究で対象とする分散センサ網では，センサの数，センサの配置が途中で変化することはない．変化するのはターゲットの位置のみである．したがって，センサのターゲットへの割り当ての価値を表す評価関数のみが動的に変化する．しかし，2.1.2 節の (b) 2 に示した，ターゲットの動作におけるセンサの観測の影響を考慮する問題では，センサがある戦略に基づいて時系列的に観測をした結果として，評価値の合計の平均値が最大となることが求められる．D-DCOP では，各時間ステップにおいて，評価値の合計が最大になるような解を求めるため，このような問題を表現することができない．たとえば，ある時間ステップのセンサの観測について，次の時間ステップのターゲットの移動が逃避的であるような，動的な分散センサ網における，センサの最適な割り当て戦略を扱えない．

2.4. マルコフ性を伴う動的分散制約最適化問題

2.1.2 節の (b) 2 に示した，ターゲットの動作におけるセンサの観測の影響を考慮する問題は，マルコフ的動的分散制約最適化問題 (MD-DCOP: Markovian Dynamic DCOP) [7] としてモデル化される．

マルコフ性は，次の状態への変化が現在の状態にのみ依存し，過去の状態に依存しない性質である．分散センサ網において考えると，次のターゲットの位置が現在のセンサの状態にのみ依存することを意味する．

2.4.1.MD-DCOP の形式的な表現

MD-DCOP は変数の集合 X，エージェントの集合 A，変数値の集合 D，状態の集合 S，遷移関数の集合 P，制約の集合 F からなる．遷移関数 $P_i(s'_i|s_i, d_i) \in P$ は，状態 $s_i \in S$ において，変数が変数値 $d_i \in D$ をとったとき，次の状態が $s'_i \in S$ となる確率を表す．また，制約 $f_i(s_i, d_i) \in F$ は，状態 $s_i \in S$ において，変数が変数値 $d_i \in D$ をとったときの価値を表す．

2.4.2. 分散センサ網の MD-DCOP による表現

ターゲットの動作におけるセンサの観測の影響を考慮する動的な分散センサ網は，MD-DCOP として以下のように表される．

- 制約: 二つのセンサの観測領域とターゲットの位置の関係
- 状態: ターゲットの位置
- 変数値: 二つのセンサの観測できる領域の組み合わせ
- 遷移関数: センサの観測に対するターゲットの移動戦略
- 評価値: ターゲットの現在の位置におけるセンサの割り当ての価値

この問題の目的は，ターゲットの移動戦略に基づいて，評価値の和の時系列における平均値が最大となるように，センサが観測する領域を決定することである．ここでは，センサである各エージェントはターゲットがどのような移動戦略を持っているかわからないと仮定する．そのため，各エージェントはターゲットの移動戦略を探索し，探索の結果として推定された移動戦略に基づいて，問題を解決する必要がある．このようなアルゴリズムには探査と利用のトレードオフが必要になる．この問題を解決するために強化学習を用いる．

2.5. 強化学習

強化学習は機械学習のひとつである．強化学習では，エージェントは試行錯誤的な行動を通して，どのような状況でどのような行動をするべきか学習する．本節では，MD-DCOP の解法に用いる R 学習について説明する．

2.5.1.MD-DCOP における強化学習

MD-DCOP では Q 学習の一種である R 学習 [9, 10] を用いる．R 学習は以下の式で表される．

$$\begin{aligned} R^{t+1}(s,d) &\leftarrow R^t(s,d)(1-\beta) + \beta[F(s,d) - \rho^t \\ &+ \max_{d' \in D} R^t(s',d')] \quad (1) \\ \rho^{t+1} &\leftarrow \rho^t(1-\alpha) + \alpha[F(s,d) \\ &+ \max_{d' \in D} R^t(s',d') - \max_{d' \in D} R^t(s,d')] \quad (2) \end{aligned}$$

ここで，s は状態を表し，d はエージェントの変数値を表している．$F(s,d))$ は状態 s においてエージェントに変数値 d を割り当てたときの価値を表している．各 R 値を格納する R テーブルが存在し，R テーブルの大きさは，状態数 $|S|\times$ 変数の値域 $|D|$ となる．また，

$0 \leq \beta \leq 1$ は R 値の学習率を，$0 \leq \alpha \leq 1$ は ρ 値の学習率を表す．

この R 学習を MD-DCOP で用いることができるように拡張する．MD-DCOP の解法では，制約で関係する各センサの組ごとに学習を行う．すなわち，R 学習を制約ごとに分割して，学習する．そのため，R 値と ρ 値を以下のように分割する．

$$R^t(s,d) \leftarrow \sum_{i=1}^{m} R^t(s_i,d_i) \quad (3)$$

$$\rho^t \leftarrow \sum_{i=1}^{m} {\rho^t}_i \quad (4)$$

ここで，それぞれの R 値と ρ 値はそれぞれの制約の評価関数 f_i と関連しているため，以下のように分割される．

$$\begin{aligned} R_i^{t+1}(s_i,d_i) &\leftarrow R_i^t(s_i,d_i)(1-\beta)+\beta[f_i(s_i,d_i)-\rho_i^t \\ &\quad +R_i^t(s_i',d_i'|d_i' \in \arg\max_{d' \in D} R^t(s',d'))] \quad (5) \\ \rho_i^{t+1} &\leftarrow \rho_i^t(1-\alpha)+\alpha[f_i(s_i,d_i) \\ &\quad +R_i^t(s_i',d_i'|d_i' \in \arg\max_{d' \in D} R^t(s',d')) \\ &\quad -R_i^t(s_i,d_i'|d_i' \in \arg\max_{d' \in D} R^t(s,d'))] \quad (6) \end{aligned}$$

MD-DCOP の解法は反復的に学習する．MD-DCOP における学習サイクル t での処理は以下の通りである．

1. ターゲットの現在の位置において，R 値の総和に基づいてセンサの観測の行動を選択する．
2. センサは各制約について報酬を獲得し，ターゲットは遷移関数に基づいて次の位置に移動する．
3. ターゲットの次の位置における R 値の総和を最大にするセンサの行動を求める．
4. 獲得した報酬と，求められた行動をもとに更新式より R テーブルを更新する．

学習サイクルの手順 1 と 3 において，R 値の総和が最大になるようにエージェントの行動を計算する必要がある．ここで，学習サイクル 1 においては，

$$f_i(d_i) = R_i^t(s_i,d_i) + \sqrt{2 \ln t \frac{|D_i|}{n^t(d_i)}} \quad (7)$$

とする．ここで，$n^t(d_i)$ は前の時間ステップまでに変数値 d_i が選ばれた回数を表す．

また，学習サイクル 3 においては

$$f_i(d_i) = R_i^t(s_i,d_i) \quad (8)$$

とする．これら，$f_i(d_i)$ において

$$d_i \in \arg\max_{d' \in D} \sum_{i=1}^{m} f_i(d_i'|d_i' \in d') \quad (9)$$

となる変数値 d_i を求めなければならない．この問題を DCOP として解くことにより，R 値の総和が最大になるエージェントの行動を計算することができる．

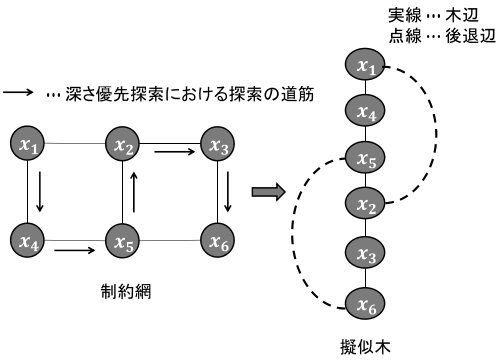

図 3: 擬似木の生成

2.6.DPOP アルゴリズム

MD-DCOP の従来解法 [7] では DCOP の解法として DPOP [3] を用いる．DPOP は問題に対する擬似木に基づく動的計画法により最適解を得る．DPOP アルゴリズムは 1) 擬似木の生成，2) UTIL メッセージの伝搬，3) Value メッセージの伝搬の 3 つのフェーズで構成される．以降では，これらのフェーズについて概説する．

2.6.1. アルゴリズムの概要

(1) 擬似木の生成

後の二つのフェーズで行うメッセージの交換の経路は，擬似木 PT に基づく．擬似木 PT は問題の制約網 G を深さ優先探索により探索することにより得られる生成木に基づいて作成される．制約網 $G=(V,E)$ から生成される擬似木は，G と同じ頂点 V，辺 E により構成され，辺 E は木辺と後退辺に分類される．深さ優先探索において通った辺を木辺と呼び，通らなかった辺が後退辺と呼ぶ．3×2 サイズのグリッドの制約網から，擬似木を生成する例を図 3 に示す．

(2) UTIL メッセージの伝搬

各ノードは，自身を根とする部分木と辺で関係する，上位ノードが持つ変数値の組み合わせについて評価値を最大化し，その組み合わせと評価値からなる表を UTIL メッセージとして，親ノードに伝搬する．図 3 における擬似木の葉ノードでの UTIL メッセージの伝搬の例を図 4 に示す．図 4 において，ノード x_6 は親にノード x_3 を持ち，制約で関係する祖先である擬似親にノード x_5 を持つ．これらのノードとの間の評価関数を合計して集計する．集約された表に基づいて，上位ノードが持つ変数値の各組み合わせにおいて，親と擬似親の変数値の組について，最大化の評価値を求めるこのような評価値からなる表を UTIL メッセージにより，親ノードである x_3 に伝搬する．

(3) VALUE メッセージの伝搬

根ノードがすべての子ノードから評価値を受け取った後，最も評価値が大きくなる根ノードの変数値を決定し，その変数値をすべての子ノードへ伝搬

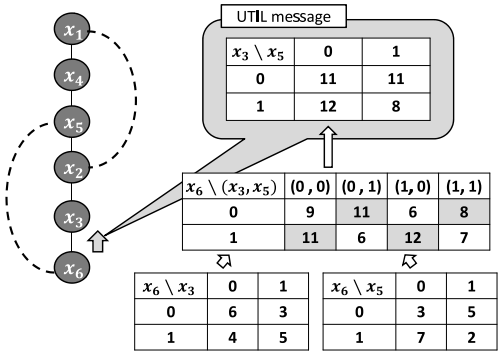

図 4: UTIL メッセージの伝搬

する．根ノード以外のすべてのノードは親から受け取った祖先の変数値に基づいて，自身の最適な変数値を決める．そして，祖先の変数値とともに，自身の変数値を子ノードに伝搬する．

2.6.2.DPOP の問題点

DPOP では，変数の数に対する制約の密度が増えると，擬似木の後退辺が増加し，メッセージの表の大きさが指数関数的に増加する．グリッド状に配置された分散センサ網の場合，グリッドのサイズが大きくなると後退辺で結ばれるノードの数が増え，また後退辺で結ばれるノード間の距離が大きくなる．このため，メッセージの表の次元は増加し続け，解くことができなくなる．

2.7.DSA アルゴリズム

DSA [1] は確率的アプローチを用いた反復改善型のアルゴリズムである．本研究では，比較的短時間で収束することを期待し，簡単な確率的山登り法の性質を持つ DSA-A [1] を用いる．DSA-A では，制約で関係する各エージェントが互いに変数値を交換しつつ解を探索する．エージェント a_i は制約で関係するエージェントの変数値に基づき，自身の各変数値についての，自身の変数が関係する関数の評価値の合計を計算する．そして，最も良くなる変数値を，次の変数値の候補とする．評価値の改善量 Δ が $\Delta > 0$ の場合は，確率 p に基づいて，評価値が最も良くなる変数値に変更する．変数値を変更した場合には，現在の変数値を制約で関係するエージェントにメッセージにより送信する．変数値を受信したエージェントは，更新された変数値に基づいて，同様の計算を反復する．DSA アルゴリズムは非常にシンプルであり，エージェントの状態に関する情報のみをメッセージとして送信しているため，通信コストを小さくできる．したがって，比較的大規模な問題に適している．しかし，各エージェントはその近傍エージェントの状態のみに基づいて状態を決定するため，局所最適解に陥りやすい．

3. 提案手法

本研究では，大規模で複雑な問題に MD-DCOP を適用できるように従来解法を改良するために，非厳密解法である DSA を適用する手法を提案する．

3.1. 基本的なアイデア

MD-DCOP の学習手法では，R 値の総和に基づいて，エージェントの行動を選択する．ここで，ターゲットの次の位置における R 値の総和を最大にする最適な行動を求めるために，厳密解法である DPOP を用いている．この DPOP を DSA に置き換える．非厳密解法は最適解を求めることが保証されないが，大規模で複雑な問題においても比較的短時間で解を得ることができる．

3.2. アルゴリズムの概要

MD-DCOP は 2.5.1 節で説明した学習サイクルを繰り返すことにより，評価値の和の平均が最大になるように，センサの変数に値を割り当てる．MD-DCOP に DSA を導入したアルゴリズムの流れは次の通りである．

1. 各制約におけるターゲットの現在の位置を把握する．
2. ターゲットの現在の位置に対して，各センサの行動選択を確率的に行う．本研究では，ここで，ϵ - *greedy* 法を用いる．
   - **(a)** 確率 ϵ でランダムにセンサに値を割り当てる．
   - **(b)** 確率 1 - ϵ で R 値の総和が最大になるように値を割り当てる．R 値の総和が最大になるように値を割り当てるときは，ターゲットの現在の位置に基づいて，各制約が持つ R テーブルからターゲットの現在の位置における R 値のみを取り出す．このとき，取り出した R 値に対し，式 (7) の計算をする．これらの R 値を，二つのセンサの変数値に関する評価関数とする．R テーブルから評価関数数を生成する例を図 5 に示す．そして，ここで，DSA アルゴリズムを用いて，R 値の総和が最大になるように値をセンサに割り当てる．

   ϵ の値は学習の初期では比較的大きく，学習が進むにつれ，値が小さくなるようにする．これは，学習の初期では探索を多めにし，学習後半では学習によって得られた情報の利用を多めにするためである．
3. 選択された行動に基づいて，センサは各制約について報酬を獲得し，ターゲットはセンサの行動に基づいて，次の位置に移動する．
4. ターゲットの次の位置とターゲットが移動する前の位置における，センサの最適な観測の行動を求める．最適な行動は 2 (b) と同様に求める．
5. 得られた報酬と求めた最適な行動に基づいて，式 (5)，(6) にしたがって R テーブルと ρ 値を更新する．
6. 1 ～ 5 を繰り返す．

上記のアルゴリズムを実行することにより，MD-DCOP を解くことができる．

3.3.MD-DCOP における DSA

本節では，分散センサ網に適用した MD-DCOP における DSA アルゴリズムの動きについて説明する．

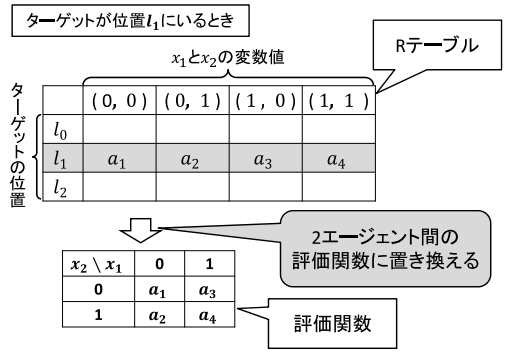

図 5: R テーブルから評価関数を生成

1. 各センサはランダムに自身の変数値を決める．
2. 近傍センサに自身の変数値を伝える．
3. すべての近傍センサから変数値を伝えられたセンサは，近傍センサの変数値に基づいて，自身の各変数値についての，自身の変数が関係する評価関数値の合計を計算する．このとき，R テーブルより生成した，評価関数に基づいて，計算する．
4. 最も評価値の合計が良くなる値を，新しい変数値の候補とする．
5. 評価値の改善量 Δ が $\Delta > 0$ のとき，確率 p に基づいて，先述の変数値の候補に変数値を変える．
6. 変数値を変更した場合，近傍センサに変数値を知らせる．
7. 終了条件を満たしているか判定する．
8. 3 ～ 7 を繰り返す．

本研究では，反復改善を一定回数 T^{DSA} 繰り返したとき，DSA アルゴリズムを終了することとした．

3.4. 予想される効果と影響

大規模な問題は解くことのできない DPOP の代わりに，DSA を用いることにより，より大規模な問題に対しても MD-DCOP を適用できるようになると予想される．しかし，最適解を求める DPOP に対して，DSA が求めることができるのは準最適解を求める場合がある．そのため，DSA を用いる MD-DCOP の解品質は DPOP を用いる MD-DCOP の解品質より劣る場合があると予想される．

4. 実験と評価

MD-DCOP において，DPOP アルゴリズムを用いた手法と DSA アルゴリズムを用いた手法を実験により比較した．ここでは，以下の三つの観点から評価した．(1) DPOP アルゴリズムを用いる手法と，DSA アルゴリズムを用いる手法の解品質の差．(2) 変数の個数と制約の密度が比較的大きい問題における影響．(3) ターゲットの移動場所の数などのパラメータを変化させたときの影響．また，DSA アルゴリズムを用いた手法において，変数値を変更する確率 p を変え比較した．実験に用いた問題は，2.1 節に示した分散センサ網とした．ターゲットはセンサが観測している領域を知ることができ，センサの動作はターゲットの次の位置への移動に影響を与える．評価値の設定は 2.1.1 節のようにした．評価値は $0, 60, 80, 100, 120, 150$ とした．DSA の

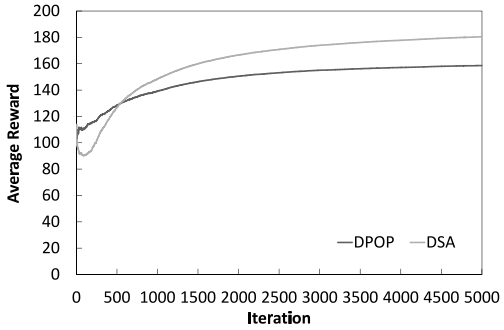

図 6: 3×1 サイズのグリッドにおける解品質の推移

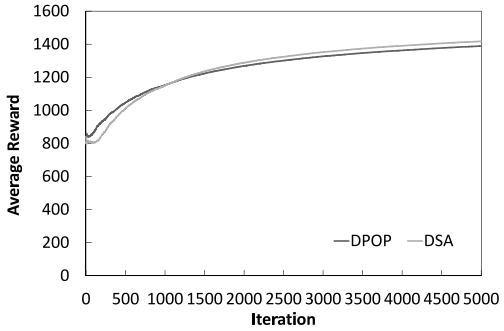

図 7: 3×4 サイズのグリッドにおける解品質の推移

反復改善の回数 T^{DSA} を 20 とした．ϵ - $greedy$ 法における ϵ の値は予備実験により設定した．

4.1. 解品質

DPOP アルゴリズムを用いる手法と，DSA アルゴリズムを用いる手法を比較するために，学習の過程における解品質の推移を評価した．

グリッドのサイズを 3×1 と 3×4，ターゲットの移動場所の数 $|S_i|$ を 3，センサの観測領域の数 $|D_i|$ を 4，イテレーション回数を 5000，DSA において自身の変数値を変更する確率 p を 0.7 としたときの，結果を図 6 と図 7 に示す．

非厳密解法である DSA は厳密解法である DPOP に比べると，学習初期こそ劣ったものの，最終的には，DPOP と同等の結果となった．

また，グリッドのサイズを $3 \times 1 \sim 3 \times 8$ と変化させ，最終的な評価値の総和の平均を求め，比較した．

二つのアルゴリズムにおける解品質を図 8 に示す．DSA アルゴリズムを用いた手法の解品質は，DPOP アルゴリズムを用いた手法の解品質と同等であることが

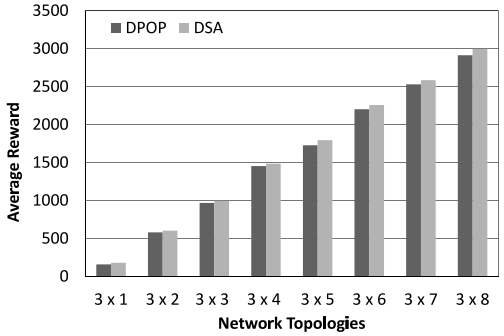

図 8: $3 \times 1 \sim 3 \times 8$ サイズのグリッドにおける解品質

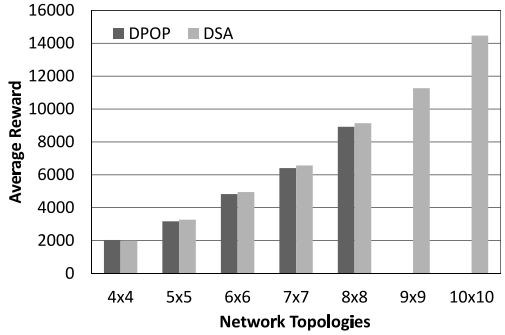

図 9: 4×4 から 10×10 サイズのグリッドにおける解品質

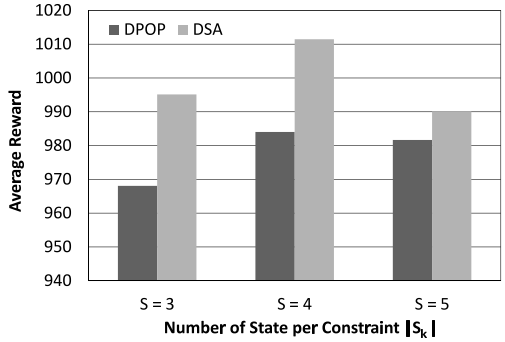

図 10: ターゲットの移動可能場所の数の違いによる影響

示された．MD-DCOP において，非厳密解法を用いることが可能であることが示された．

4.2. 問題の規模の影響

二つのアルゴリズムにおいて，より大規模な問題の影響を評価するために，4×4 から 10×10 までのサイズのグリッドを用いて評価した．問題の各パラメータはグリッドのサイズ以外は 4.1 節と同じである．

実験の結果を図 9 に示す．DPOP の実験は 8×8 のサイズで打ち切ったが，解品質は同等であることが示された．DPOP ではグリッドのサイズが大きくなると深さは全ノード数となる．そのため，後退辺で結ばれるノード間の距離が大きくなり，メッセージの表の次元の数が徐々に増加していく．10×10 サイズのグリッドでは，擬似木の幅は最大で 10 となり，メッセージの次元数は 10 まで増加する．本実験では，変数の値域の大きさは 4 であるため，メッセージの表の大きさは 4 の 10 乗となる．一方で DSA は，自分が変数値を変えた場合のみ，自分の変数値を近傍ノードに送るため，メッセージのサイズは小さい．また，グリッド状に配置された場合，近傍ノードの数は最大で 4 である．そのため，計算コストは抑制される．

4.3. 他のパラメータを変化させたときの影響

各ターゲット k の移動場所の数 $|S_k|$，各センサの観測領域の数 $|D_i|$ を変化させたときの影響について比較した．グリッドのサイズを 3×3，イテレーション回数を 5000，DSA において自身の変数値を変更する確率 p を 0.7 とした．ターゲットの移動場所の数の違いによる影響を評価するために，センサの観測領域の数 $|D_i|$ を 4 とし，ターゲット k の移動場所の数 $|S_k|$ を $3 \sim 5$ と変化させ実験した．また，センサの観測領域の数の違いによる影響を評価するために，ターゲットの移動場所の数 $|S_k|$ を 3 とし，センサの観測領域の数 $|D_i|$ を $4 \sim 6$ と変化させ実験した．

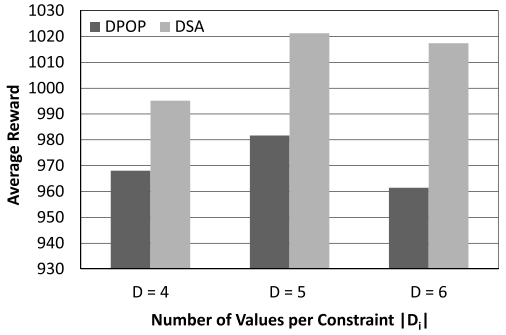

図 11: センサの観測領域の数の違いによる影響

表 1: 最終的な R テーブルにおいて未更新である R 値の平均数

分散制約最適化アルゴリズム	DSA	DPOP
未更新であるR値の平均数	5.18	19.33

グリッドのサイズ:3 × 3　各ターゲットの移動場所の数:3　各センサの観測領域の数:4

結果を図 10，11 に示す．これらの結果では，3.4 節の予想に反し，全体的に，DPOP を用いた手法に比べ，DSA を用いた手法の方が高かった．本研究では，分散制約最適化アルゴリズムをセンサの最適な行動を求めるために用いている．このため，強化学習の探査次第では，DSA が DPOP の結果を上回ることはあり得る．ここで，これら 2 つのアルゴリズムにおける探査の程度を調べるために，最終的な R テーブルの中身を比較する．表 1 は 4.1 節における実験でのグリッドのサイズが 3×3 のときの最終的な R テーブルにおける未更新の R 値の平均数を表している．表 1 から未更新の R 値の数，すなわち一度も選択されていない行動の数が DPOP に比べて DSA の方が少ないことが確認できる．DPOP により，現時点の学習結果において，最適な行動が選択されるが，DSA では必ずしも最適な行動が選択されない．この揺動により，DSA は DPOP よりもセンサの行動選択における偏りが小さく，より探査が行われたと考えられる．このことが，DSA により良い政策を得る場合がある一因と考えられる．

4.4.DSA における変数値を変更する確率を変えたときの影響

DSA (DSA-A [1]) において変数値を変更する確率 p の値を異なる値としたときの影響を評価した．グリッドのサイズを 3×3，ターゲットの移動場所の数 $|S_i|$ を 3，センサの観測領域の数 $|D_i|$ を 4，イテレーション回数を 5000 とし，DSA において自身の変数値を変更する確率 p を $0.7 \sim 1.0$ と変化させ評価した．

結果を図 12，13 に示す．図 12 の縦軸は平均報酬値，図 13 の縦軸はターゲットの捕捉率，各図の横軸は DSA において変数値を変更する確率を表している．図 12 において，$p = 0.7 \sim 0.9$ では，解品質に大きな差は見られないが，$p = 1.0$ のときでは，解品質が低下した．ま

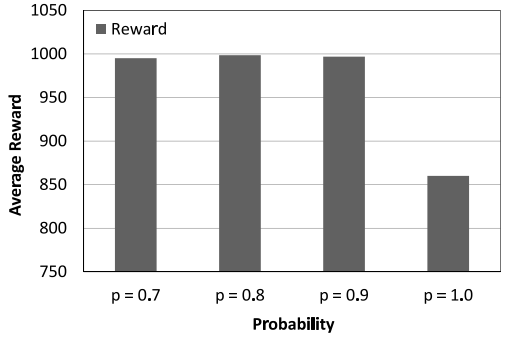

図 12: DSA において変数値を変更する確率を変えたときの平均報酬への影響

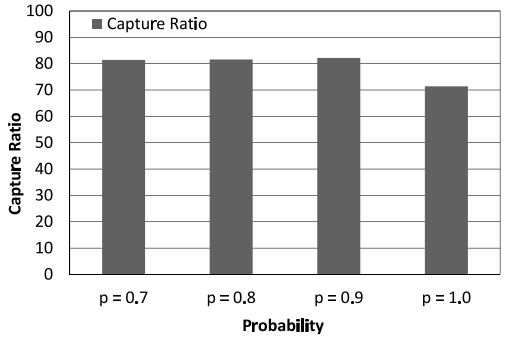

図 13: DSA において変数値を変更する確率を変えたときのターゲットの捕捉率への影響

た，同様に図 13 においても，$p = 1.0$ のみターゲットの捕捉率が 10% ほど低下している．これは，$p = 1.0$ では局所的最適解に陥ったためであると考えられる．すなわち，DSA-A の確率的な探索は必要であると考えられる．

5. まとめ

本研究では，マルコフ的動的分散制約最適化問題 (MD-DCOP) の解法において，DSA を用いることにより，MD-DCOP を適用できる問題の規模を拡大した．提案手法を実験により評価し，DSA を用いた場合においても，DPOP を用いた場合と同程度であるという十分な解品質が得られることが示された．また，問題の規模を拡張した場合でも，MD-DCOP を適用できることが示された．

今後の研究課題として，局所解から確率的に脱出する DSA の他のバージョンの効果を評価することが挙げられる．また，本研究ではターゲットの行動が完全観測であるが，現実的な問題では，ターゲットの行動を完全に観測できるわけではない．こうした現実的な問題に適用するために，部分観測下で行える強化学習を組み込む手法の検討が必要である．また，本研究は従来研究の問題設定に従ったが，マルコフ性が満足されない場合についての拡張も今後の課題である．

謝辞

本研究の一部は，科研費基盤研究 (C)25330257 による．

参考文献

[1] Weixiong Zhang, Ong Wang, and Lars Wittenburg. Distributed stochastic search for constraint satisfaction and optimization: Parallelism, phase transitions and performance. In *in PAS*, pp. 53–59, 2002.

[2] Pragnesh Jay Modi, Wei-Min Shen, Milind Tambe, and Makoto Yokoo. Adopt: Asynchronous distributed constraint optimization with quality guarantees. *Artificial Intelligence*, Vol. 161, No. 1, pp. 149–180, 2005.

[3] Adrian Petcu and Boi Faltings. A scalable method for multiagent constraint optimization. In *Proceedings of the 19th International Joint Conference on Artificial Intelligence*, IJCAI'05, pp. 266–271, San Francisco, CA, USA, 2005. Morgan Kaufmann Publishers Inc.

[4] Chongjie Zhang and Victor R Lesser. Coordinated multi-agent reinforcement learning in networked distributed pomdps. In *AAAI*, 2011.

[5] 松井俊浩, 松尾啓志. 動的な分散制約最適化問題のための基本的な枠組みの提案. 人工知能学会全国大会論文集, No. 0, pp. 271–271, 2005.

[6] William Yeoh, Pradeep Varakantham, Xiaoxun Sun, and Sven Koenig. Incremental dcop search algorithms for solving dynamic dcops. In *The 10th International Conference on Autonomous Agents and Multiagent Systems-Volume 3*, pp. 1069–1070. International Foundation for Autonomous Agents and Multiagent Systems, 2011.

[7] Duc Thien Nguyen, William Yeoh, Hoong Chuin Lau, Shlomo Zilberstein, and Chongjie Zhang. Decentralized multi-agent reinforcement learning in average-reward dynamic dcops. In *Proceedings of the 2014 international conference on Autonomous agents and multi-agent systems*, pp. 1341–1342. International Foundation for Autonomous Agents and Multiagent Systems, 2014.

[8] 横尾真, 平山勝敏. 分散 breakout：反復改善型分散制約充足アルゴリズム (特集: 並列処理). 情報処理学会論文誌, Vol. 39, No. 6, pp. 1889–1897, jun 1998.

[9] Anton Schwartz. A reinforcement learning method for maximizing undiscounted rewards. In *Proceedings of the tenth international conference on machine learning*, Vol. 298, pp. 298–305, 1993.

[10] Sridhar Mahadevan. Average reward reinforcement learning: Foundations, algorithms, and empirical results. *Machine learning*, Vol. 22, No. 1-3, pp. 159–195, 1996.