3F1-1

協調系実現に向けた監視対象の自律的決定モデル

Self-organiging Mutual Surveillance Network to Maintain Cooperation

野村 一平　　荒井 幸代
Nomura Ippei　　Arai Sachiyo

千葉大学大学院工学研究科都市環境システムコース
Course of Urban Environment Systems, Graduate School of Engineering, Chiba University

A mechanism of Meta-Norms is proposed by Axelrod to promote cooperation within the whole system under the situation of Prisoner's Dilemma. In the Meta-Norms game defect agents are punished by other agents, and they are prompted to cooperate. Maintaining cooperation is demonstrated in Axelrod's experiment under the condition where mutual surveillance among all agents is premised. In other words, all agents are inter-connected. Thus, a large amount of costs to monitor other connected agents. Therefore, we propose an autonomous decision of surveillance not to spend excessive cost but to maintain cooperation. This model makes optimal mutual surveillance network in terms of satisfying both criteria, i.e. reducing cost for monitor and increasing the profit from the whole systems. In the proposed model, each agent autonomously selects adversaries to play the Meta-Norm game and puts links to them, then finally the optimal network will be emerged.

1. 研究の目的と背景

n 人囚人のジレンマ解消法の一つに Axelrod のメタ規範ゲーム [1] がある．メタ規範ゲームは，集団内でエージェント同士が行動を監視し裏切り者を罰するモデルであり，計算機実験によって協調を維持できることが示されている．Axelrod のメタ規範ゲームは全エージェント間の相互監視が前提であるので，監視の対象を決めるネットワーク（以後，監視ネットワークとよぶ）は完全グラフである．しかし，監視によるするコストを考えると監視対象数（リンク数）は少ない方がよい．

本研究では監視対象数の削減し，かつ協調維持が可能な監視ネットワーク生成法を提案する．

2. 準備

本章では, 本研究で扱う基礎となる囚人のジレンマゲーム, 強化学習の概略を説明する．

2.1 囚人のジレンマゲーム

他者との利害関係は, 利得行列によって定義される．

本研究で対象とする囚人のジレンマゲームの利得行列を表 1 に示す．ただし，$-1 \leq S < 0$，かつ $1 < T$ を満たす．

各エージェントの行動戦略は協調行動（C: Cooperate）と利己的行動（D: Defect）とする．エージェントの集合を $\mathcal{N} = \{1, 2, \ldots, i, \ldots, n\}$ とする．表 1 に示す利得行列において, エージェント i がとき刻 t の意思決定で C または D を選択することによって得られる利得 $U_{\mathrm{C}}(r_i(t))$, $U_{\mathrm{D}}(r_i(t))$ はそれぞれ式 (1) で表される．ただし, $r_i(t)$ はとき刻 t において, エージェント i の対戦相手のうち C を選択するエージェントの割合を表す．

$$\begin{cases} U_{\mathrm{C}}(r_i(t)) = (1 - S) \cdot r_i(t) + S \\ U_{\mathrm{D}}(r_i(t)) = T \cdot r_i(t) \end{cases} \tag{1}$$

連絡先: 野村一平，千葉大学大学院工学研究科都市環境システムコース，〒 263-8522 千葉市稲毛区弥生町 1-33，043-290-3316，nomura0706@gmail.com

ここで, 各エージェントは合理的選択に基づき意思決定を行うとする．合理的選択では, 式 (2) に従い C または D を選択する．

$$\begin{cases} \text{C} \ \text{ if } \ U_{\mathrm{C}}(r_i(t)) \geq U_{\mathrm{D}}(r_i(t)) \\ \text{D} \ \text{ otherwise} \end{cases} \tag{2}$$

表 1: Payoff Matrix

	C	D
C	1, 1	S, T
D	T, S	0, 0

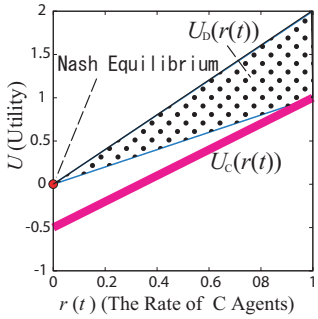


図 1: PD Payoff Structure

$S = -0.5$ の場合を例に, n 人囚人のジレンマゲーム（PD）における $r(t)$ に対する $U_{\mathrm{C}}(r(t))$ と $U_{\mathrm{D}}(r(t))$ の変化を図 1 に示す．太線が $U_{\mathrm{C}}(r(t))$, 点で塗られた領域が $U_{\mathrm{D}}(r(t))$ を表す．囚人のジレンマゲームでは, 常に $U_{\mathrm{D}}(r(t)) \geq U_{\mathrm{C}}(r(t))$ が成り立つため, ナッシュ均衡は全エージェントが D を選択する状態である．一方, 全エージェントが D を選択する状態は全エージェントが C を選択する状態と比較してパレート劣位である $(U_{\mathrm{C}}(1) > U_{\mathrm{D}}(0))$.

2.2 強化学習

強化学習は，未知の環境における最適な行列ルールを試行錯誤的に獲得する手法である [4]．エージェントは獲得する報酬の期待総和を最大化する方策を学習を行う．

環境モデルを $\langle \mathcal{S}, \mathcal{A}, R, \pi \rangle$ と定義する．$\mathcal{S}$ は状態集合，$\mathcal{A}$ は行動集合，R は報酬関数を表し，方策 π は状態 s から可能な行動 a を選択する確率である．エージェントはとき刻 t において，状態 $s_t \in \mathcal{S}$ を観測し，自身の方策 π_t に基づいて行動 $a_t \in \mathcal{A}$ を選択する．その後，とき刻 $(t+1)$ では s_t, a_t によって確率的に次状態 s_{t+1} に遷移し，報酬 r_t を得る．獲得した報酬から価値関数を生成し，その値を用いて方策 π の評価と改善を行う．

a) ϵ-greedy 選択

強化学習では現在の状態から行動を選択し次状態に遷移する．現在の状態 s において，$a = \arg\max_{a \in \mathcal{A}} Q(s,a)$ を選択する方法を greedy 選択という．ϵ-greedy 選択では，確率 ϵ $(0 \le \epsilon \le 1)$ でランダムに行動選択を行い，確率 $1-\epsilon$ で greedy 選択を行う．

b) Sarsa

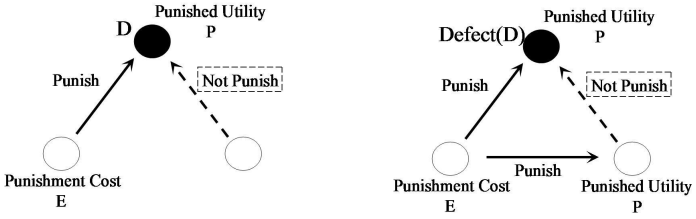
本研究では環境における状態遷移確率を推定しながら学習する環境同定型のアルゴリズム Sarsa[6] を用いる．式 (3) に Sarsa における行動価値関数 $Q(s,a)$ の更新式を示す．

$$Q(s_t,a_t) \leftarrow Q(s_t,a_t) + \alpha[r_{t+1} + \gamma Q(s_{t+1},a_{t+1}) - Q(s_t,a_t)] \quad (3)$$

3. 相互監視による囚人のジレンマ解消

3.1 メタ規範ゲーム

本研究では，メタ規範ゲームを導入し囚人のジレンマ解消を目指す．メタ規範ゲームは Axelrod[1] が提案した相互監視モデルであり，規範ゲームの拡張モデルである．規範ゲームにおける行動戦略は，囚人のジレンマゲームの D 選択エージェントを罰する（PN）と罰しない（UPN) であり，メタ規範ゲームにおける行動戦略は，規範ゲーム PN と UPN に加えて UPN 選択エージェントを罰する（MetaPN）と罰しない（MetaUPN）である．規範ゲームとメタ規範ゲームの概略を図 2(a) と図 2(b)



(a)Norm Game (b)Meta-Norm Game

図 2: Mutual Surveillance Model by Axelrod

囚人のジレンマゲームにおいて，集団内で裏切り（図 2(a) の黒丸）が発生したとき，裏切り者とリンクで隣接したエージェントが，裏切り者を罰する（図 2(a) の実線矢印 Punish(PN)）ことができる．罰するエージェントは懲罰コスト E を払い，裏切り者に罰 P を与える．さらにその後メタ規範ゲームでは,「裏切り者を罰しなかった者」を同様に罰する（図 2(b) の実線矢印 Punish(MetaPN)）ことができる．罰するエージェントは懲罰コスト MetaE を払い,「裏切者を罰しなかった者」に罰 MetaP を与える．規範ゲームでは，懲罰コスト E がかかるため懲罰にインセンティブが存在せず協調を維持できない．しかしメタ規範ゲームでは,「裏切り者を罰しなかった者」も罰せられるため，このことが裏切り者を罰するインセンティブになっている．Axelrod は計算機実験により，メタ規範ゲームが協調を維持可能であることを示した．

3.2 相互監視ネットワークの最適化

Axelrod のメタ規範ゲームは完全グラフの監視ネットワークを前提としている．しかし，監視に要するコストを考えるとリンク数は小さい方がよく，本研究ではリンク数を抑制し，かつ協調を維持できるネットワークを見つけることを目的とする．そこで本稿では，エージェント自身が自律的に囚人のジレンマゲームとメタ規範ゲームの相手（以後，対戦相手とよぶ）を選択しリンクを張るモデルを提案する．

4. ネットワーク自律的生成モデル

本研究で提案するネットワーク自律的生成モデルではエージェント自身が対戦相手を選択し，囚人のジレンマ状況でメタ規範ゲームを行い { C, D} ，{ PN, UPN} ，{ MetaPN, MetaUPN} の行動を選択する．以下に，監視に要するコストを考慮した自律的ネットワーク生成の手順を示す．

はじめに，本稿で用いる用語を定義する．

- エージェント集合：$N = \{1, ..., n\}$
- エージェント i からエージェント $j_{(i \neq j)}$ に対する行動選択集合：$a_{ij} = \{l_{ij}, CD_{ij}\}$，$l_{ij} \in \{\text{link, not link}\}$，$CD_{ij} \in \{\text{C, D}\}$
- l_{ij} = link：i から j に対する，リンク選択
- l_{ij} = not link：i から j に対する，リンク非選択
- CD_{ij} = C：i から j に対する，C 選択
- CD_{ij} = D：i から j に対する，D 選択
- s_{ij}：j に対する i の状態
- $f_i(a_{ij}, a_{ji})$：a_{ij}，a_{ji} による i の利得

ここで，$S + T = 2$ として { C, D} の各戦略に対する利得行列を表 1，そして，規範ゲーム，メタ規範ゲームにおける利得行列はリンクの有無によって決まるとし，表 2 の利得行列を考える．リンク成立にはコストがかかり，リンク不成立のときはかからないと考えて，ここではリンク成立のとき 0, リンク不成立のとき $\theta \ge 0$ を得ると定義する．

表 2: Payoff Matrix with linking and not linking

$i \setminus j$	link	not link
link	0，0	θ，θ
not link	θ，θ	θ，θ

4.1 エージェントの配置

n 体のエージェントを $\sqrt{n} \times \sqrt{n}$ の 2 次元格子トーラス上に配置する．各エージェントは固定され，上下左右に隣接している 4 エージェントに対して意思決定を行う．

4.2 相互監視ネットワーク生成の手順

本提案モデルは，２つの段階から構成される．第 1 段階では，監視によるコストを考慮しつつ，ネットワークをボトムアプに生成し（Step1,Step2)，生成されたネットワークにおいて，規範ゲーム，メタ規範ッゲームが成立するか否かをチェックする（Step3）．第 2 段階では，実際にメタ規範ゲームを実施し，協調の維持が可能なネットワークを選択する．

4.2.1 第 1 段階：ネットワークの生成

はじめに，第 1 段階として Step1，Step2 の手順を図 3 に示す．Step1 で，各エージェントは上下左右の 4 エージェントに対してリンク選択を行う．お互いにリンク選択を行ったときに限り，リンク成立となる．その後 Step2 で，リンクが成立したエージェントは C または D を選択し，利得を得る．リンク選択と CD 選択は上下左右のエージェントに対してそれぞれ意思決定を行うため，対戦相手ごとに異なる意思決定を行ってもよい．

リンク選択，CD 選択をそれぞれ 1 ステップとし，これら 2 ステップを 1 エピソードし，十分なエピソード数を繰り返す．

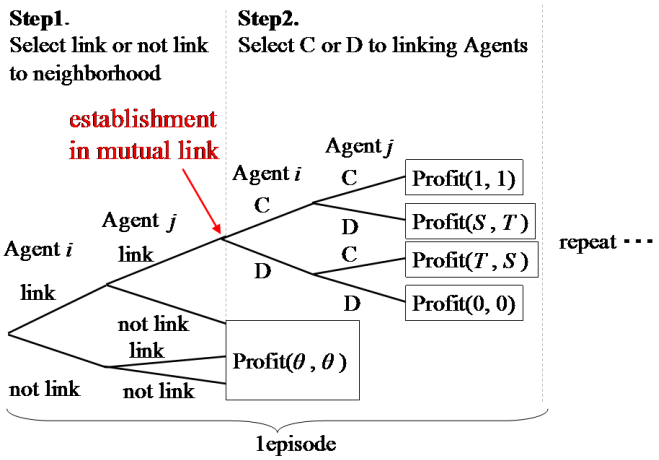


図 3: Game Process

【強化学習を用いた意思決定】：上述のリンク選択や，CD 選択に対するエージェントの意思決定に強化学習を用いる．以下に強化学習の設定を示す．

- 行動集合：リンク選択または CD 選択とする．とき刻 t において，自分 i から対戦相手 j に対する行動を $a_{ij}(t)$ と表現する．1 とき刻を 1 ステップとする．
- 状態表現：対戦相手との過去 2 ステップの行動履歴とする．とき刻 t における状態は以下で表される．
  $s_{ij}(t) = [a_{ij}(t-1), a_{ji}(t-1), a_{ij}(t-2), a_{ji}(t-2)]$
- 報酬：表 1，表 2 における利得を報酬 r し，i の利得 r_i は $r_i = f(a_{ij}, a_{ji})$ で表される．
- 方策：ϵ-greedy 選択を用いる．
- 学習アルゴリズム：Sarsa を用いる．

【メタ規範ゲーム成立の可否】：生成されるネットワークの目的は，少ない監視コストでメタ規範ゲームによって協調を維持することである．Step3 として，以下の 2 つの評価基準から Step1,Step2 で生成したネットワークを選定する．

条件 1：生成されたネットワークのリンク数 ≥ 1，かつ，全結合でない．
条件 2：エージェントの行動がパレート最適ではない．

条件 1 は監視対象数（リンク数）の抑制に関与する．リンクコスト θ によってリンクの要/不要を判定する上で，θ の値が過小であれば全結合となり，θ が過大であればリンクが無くなる．したがって，適切な θ を用いなければ，監視対象数を抑えたネットワークが生成されない．生成されたネットワークには，リンクが少なくとも 1 つ存在し，かつ，全結合ではないという規範 1 を満たす必要がある．

また，条件 2 はメタ規範ゲームの導入に関与する．集団がパレート最適な行動を選択すれば，メタ規範ゲームは不要になる．したがって，メタ規範ゲームが有効となるためには規範 2 を満たす必要がある．これらの条件を満たせなければ，θ を変更し Step1 に戻る．

本稿では，S と θ を変化させ実験を行い，Step3 のにおける条件 1，条件 2 を満たすネットワークを得ることを目的とする．

4.2.2 第 2 段階：メタ規範ゲームの適用

第 2 段階では，Step3 の条件を満たしたネットワークに対してメタ規範ゲームを行う．リンクが成立したエージェント間で囚人のジレンマゲームとメタ規範ゲームを行い，{ C, D} ，{ PN, UPN} ，{ MetaPN, MetaUPN} を選択し利得を獲得する．集団の協調を維持できればその監視ネットワークを解とし，維持できなければ θ を変更し Step1 へ戻る．

4.3 関連研究

本研究の特徴の一つとして，メタ規範ゲームにおけるネットワークの導入がある．Galan ら [2] はメタ規範ゲームをスケールフリーネットワーク上で，Newth[3] はスモールワールドネットワーク上でそれぞれメタ規範ゲームを行い，平均次数，クラスタ係数，クラスタ数が協調に影響することを示した．また，ネットワーク生成モデルの既存研究として，Aljaz[5] は，エージェントがリンク生成の意思決定を行うモデルを提案し，生成されるネットワークと意思決定のナッシュ均衡を示している谷本 [8] は対戦相手入替えモデルを提案し協調行動の優位性を示した．ネットワーク生成モデルにメタ規範ゲームを導入した例はない．

本研究における二つ目の特徴はエージェントの学習ダイナミクスである．山本 [7] は，メタ規範ゲームのエージェントの意思決を遺伝的アルゴリズムを用いて更新した．本研究における意思決定の更新に強化学習を用いている．これはネットワーク生成モデルでは，環境の不確実性や報酬の遅れが存在する環境には強化学習が適していると考えたためである．

5. 計算機実験

集団数 $n = 64$，学習率 $\alpha = 0.03$，割義気率 $\gamma = 0.9$，探索確率 $\epsilon = 0.1$，エピソード数 100,000 とする

5.1 実験設定と結果

S を -0.9 から -0.1 まで 0.2 ずつ変化させ，それぞれの S に対し，θ を -0.5 から 1.5 まで 0.1 ずつ変化させ実験を行う．実験結果を図 4 に示す．各 S 毎に，左のグラフの横軸は θ，縦軸は最終エピソードにおける成立したリンク数を表し，右のグラフの横軸は θ，縦軸は最終エピソードにおけるリンクが成立したエージェントの CD 選択組み合わせの割合を表している．左のグラフの成立したリンク数が 0 のとき，右のグラフにはプロットされない．
実験により，メタ規範ゲーム導入が可能な条件は $-0.9 \leq S \leq -0.3$ における $\theta = 0$ である．このとき，メタ規範導入が可能となる 2 つの条件を満たす．

5.2 考察

$-0.9 \leq S \leq -0.1$ において，全エージェントがナッシュ均衡に陥り D を選択した．これにより，リンク選択は $f_i(\mathrm{D}, \mathrm{D})$ と θ の大小関係により決定する．$\theta < 0$ のとき，式 (4) を満たすため，（リンク選択，D 選択）が（リンク非選択）より利得が高い．

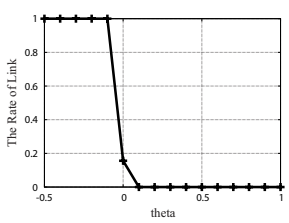
$$f_i(\mathrm{link}, \mathrm{link}) + f_i(\mathrm{D}, \mathrm{D}) > \theta \tag{4}$$

よって全リンクが成立する．

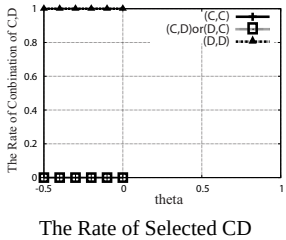
$\theta = 0$ のとき，式 (5) を満たすため，（リンク選択，D）が（リンク非選択）と利得が等しい．

$$f_i(\mathrm{link}, \mathrm{link}) + f_i(\mathrm{D}, \mathrm{D}) = \theta \tag{5}$$

よってリンクを選択するエージェントと選択しないエージェントが混在する．

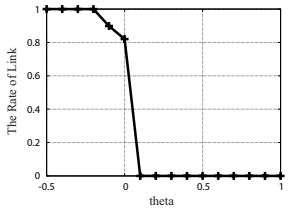


The Rate of Established Links

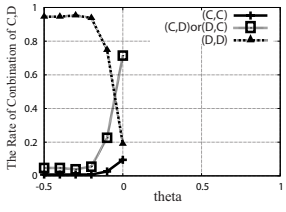


The Rate of Selected CD between Linking Agents

Result(1) $-0.9 \leq S \leq -0.3$



The Rate of Established Links



The Rate of Selecting CD between Linking Agents

Result(2) $S = -0.1$

図 4: Final Episode's Behavior on Prisoner's Dilemma

$\theta > 0$ のとき，式 (6) を満たすため,（リンク選択，D）が（リンク非選択）より利得が低い．

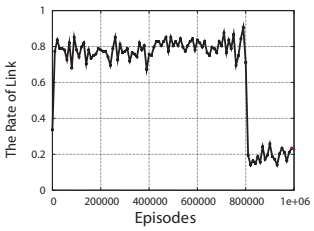
$$f_i(\text{link}, \text{link}) + f_i(\text{D}, \text{D}) < \theta \quad (6)$$

よって全リンクが不成立となる．

一方，$S = -0.1$ における CD 選択組み合わせの割合は，囚人のジレンマ環境にもかかわらず θ が 0 に近づくにつれ，(D, D) の割合が減少し，C を選択するエージェントの割合が増加した．この原因として，S が 0 に近いため $f_i(\text{D}, \text{D})$ と $f_i(\text{C}, \text{D})$ の差が小さい事が挙げられる．これにより C を選択する確率が上がる．さらに，θ が 0 に近づくと D を多く選択するエージェントにはリンク選択しなくなり，(D, D) の割合が減少する．

5.3 メタ規範ゲーム導入に可能なネットワーク

メタ規範ゲームが導入可能な利得構造の一例として，$S = -0.5$ かつ $\theta = 0$ における最終エピソードの挙動を図 5 に示す．



(a)The Rate of Link ($\epsilon = 0$ after 800000 Episode)

(b)Generated Network in the Last Episode

図 5: Link Behavior in each Episode on $S = 0.5$, $\theta = 0$

図 5(a) より $\epsilon = 0$ となる 800,000 エピソード以降も各エージェントの行動選択は収束しない．この理由はリンク選択とリンク非選択における期待報酬が等しいためである．リンク選択では囚人のジレンマゲームのナッシュ均衡となる利得 0 を得，リンク非選択では $\theta = 0$ を得る．図 5(b) は最終エピソードにおいて生成されるネットワークを表す．リンク数が全結合の約 2 割に抑えることができた．発表では，このネットワークでメタ規範ゲームを行い協調維持実現の可否を議論する．

6. まとめ

本研究では，集団のジレンマ解消のアプローチとして，メタ規範ゲームに基づいた相互監視ネットーワークを導入した．既存手法では監視ネットワークが完全グラフであり，監視コストが多く発生する．そこで本研究では，ネットワーク自律的生成モデルを提案し，リンク数が最小となる最適なネットワークの創発を考えた．このモデルでは，エージェント自身が対戦相手を選択し囚人のジレンマゲームを繰り返し，ネットワークを生成する．次に，生成されたネットワークが相互監視可能であれば，メタ規範ゲームを導入して協調維持を確認する．

本稿では，リンク選択と囚人のジレンマゲームによるネットワーク生成の実験を行い，リンク選択におけるコストと囚人のジレンマゲームにおける「裏切り」の利得が等しいとき，メタ規範ゲームを導入可能なネットワークが生成された．今後，提案モデルによってリンク数が最小かつ協調維持可能な監視ネットワークの生成を目指す．

参考文献

[1] Axelrod, R: Cooperation: Agent-Based Models of Competition and Collaboration, Chapter 3:Promoting Norms, Princeton University Press, pp40–68, (1997)

[2] Galan,J. M., Latek, M. M., Rizi,S. M. M.: Axelrod's Metanorm Games on Networks，PLoS ONE 6(5) : e20474. doi : 10.1371/journal.pone.0020474, (2011)

[3] Newth, D.: Altruistic Punishment, Social Structure and the Enforcement of Social Norms, KES 2005, LNAI 3683, Volume 3683, pp806–812, (2005)

[4] Sutton, R. S., Barto A. G.: Reinforcement Learning : An Introduction, *MIT Press*, Cambridge, MA, (1998)(邦訳: 強化学習，三上 貞芳，皆川 雅章訳，森山出版, (2000)).

[5] Ule, A.: Network formation and cooperation in finitely repeated games, (2006)

[6] Watkins, C. J. C. H., Dayan, P. D.: Q-learning, *Machine Learning*, Vol.8, pp279–292, (1992)

[7] 山本仁志，岡田勇: 社会的ワクチンメタ規範ゲームにおける裏切りの効果，第 17 回社会情報システム学シンポジウム 講演論文集,pp149–154, (2011)

[8] 谷本潤: 2 × 2 対称ゲームにおけるネットワークと戦略共進化に基づくジレンマの解消について，日本応用数理学会分誌 Vol.18,No.1,pp17–27, (2008)