小特集「AI トレンド・トップカンファレンス NeurIPS 2018，AAAI 2019 報告会」

AAAI 2019 概要・今年の動向
ゲーム理論のトピックを中心に

Overview of the AAAI 2018 Conference and Interested Topics on Game Theory

奥村　エルネスト　純　株式会社ディー・エヌ・エー
Jun Ernesto Okumura　DeNA Co., Ltd.
paco.sci@gmail.com

Keywords: machine learning, neural network, game theory, reinforcement learning, international conferences.

1. AAAI 2019 概要

AAAI（Association for the Advancement of Artificial Intelligence）Conference は，NeurIPS などと並ぶ人工知能に関する国際会議の一つで，2019 年は第 33 回目の会議[*1]がアメリカ・ハワイで開催された．内容は AI 分野の中でも事業会社における活用のような比較的応用寄りの話題が多いのが特徴で，今年は Emerging Topic として “Artificial Intelligence for Social Impact” が開催されるなど，AI と社会課題についての研究や問題提起も多かった．

今年の会議で特徴的だったのは論文の投稿数である．昨年から 103％増となる 7 095 本を記録しており，次点の NeurIPS 2018 の 4 856 本と比較しても際立っていた．他の国際会議も投稿数の増加傾向があり，大量の投稿に対して信頼性のある査読を担保することがコミュニティの課題となっているように思う．AAAI 2019 でも Summary Reject や Toronto Paper Matching System の導入，Double Blind Review の強化といったさまざまな工夫が行われていたが，まだ模索の段階である．会議の概要を含め，全体的な内容については [奥村 19] でも紹介しているため参照されたい．

本稿では，会議の中から主にゲーム理論に関連した話題を紹介する．2 章ではゲーム理論が社会課題に対してどのように使われているか，近年の研究をまとめた招待講演からいくつかの事例を紹介する．3 章では，不完全情報ゲーム解法の現状として，Outstanding Paper Award で Honorable Mention を受賞した “Discounted Regret Minimization” を紹介する．不完全情報ゲームは，プレーヤのもつ情報が部分的な状況において，利得を最大化する戦略をモデル化する問題系である．現実世界では相手の手の内や戦略がわからない状況で意思決定を下す必要のある場面は多く，交渉や経済活動，安全保障活動といった領域で応用が期待されている．ベンチマークとして使われているポーカーに注目し，過去の研究も踏まえつつ現在のアルゴリズムを概観する．

2. ゲーム理論と実活用の現在

AAAI 2019 では，ゲーム理論の枠組みを用いて社会課題に取り組んでいる事例が多く報告されていた．ここでは長くこのテーマに取り組んでいる Milind Tambe による招待講演 “AI and Multiagent Systems for Social Good” [*2] をもとに，いくつかの例を紹介したい．

[Pita 08] は空港の警備をシュタッケルベルグ競争（Stackelberg competition）モデルで記述し，ロサンゼルス国際空港における重火器や薬物の検知数を大幅に向上させる実績につなげた研究である．シュタッケルベルグ競争とは，ミクロ経済学における寡占モデルとして使われており，寡占者の価格決定の後に追従者が価格決定を行うモデルである．[Pita 08] では，警備スケジュールが既知なうえで敵対者が行動を決定する状況をモデル化し，警備側にとっての利得（犯罪の検知数など）を最大化させるようなスケジュールの最適化に成功している．この事例はさまざまな領域で発展的に適応されており，[Kiekintveld 09] では組合せが 10^{41} に及ぶ航空機監視で使われ，[Pita 11] では湾岸警備における監視船舶のターゲティングを最適化して 350％の利得改善を達成している．

近年では野生動物保護でも同種のアプローチが取られている．[Gholami 18] は，ウガンダの国立公園で密猟者

*1 https://aaai.org/Conferences/AAAI-19/

*2 講演資料は http://teamcore.usc.edu/lectures/AAAI_2019.pdf から取得できる．

が仕掛ける罠を発見するために，限られた警備リソースをどのように振り分ければよいか意思決定する問題を考えた．公園内に設置されている罠の密度をメッシュ単位で予測したうえで，パトロール区域を最適化して罠の検知数を 5 倍に改善させたという．[Xu 16] は，ドローンを使った密猟者への介入を扱った．ここでは監視員が近くにいるというシグナルを発信するかどうか適切に意思決定することで，密猟者のもつ情報量を制限して最適応答戦略をコントロールする手法が提案されている．

公衆衛生の文脈では，例えば [Yadav 16] が HIV の感染症拡大を抑制するためにホームレスの若者に介入する方策を研究している．こうした啓蒙活動は，彼らのつながりの中で最も拡散効果のある人物に対して優先的に行うのが効果的であるが，リアルグラフの構造や各エッジで情報が伝わる確率は不明であることが多い．このように情報が不完全な状況で，どのノード（人物）に介入するのが最も利得（拡散人数など）が高くなるかを考える．[Yadav 16] は，リアルグラフ上の情報伝達を POMDP（Partial Observable Markov Decision Process; 部分観測マルコフ決定過程）としてモデル化して解く手法を提案した．[Killian 19] はインドの結核患者に対して治療継続を促す介入行動を扱っており，[Wilder 19] は「結核患者の治療離脱予測」と「介入戦略の最適化」を End to End に統合することで，介入効果の改善に成功している．

このように，さまざまな社会課題を不完全情報ゲームとして見立てることによって介入を最適化するアプローチはこれからも活用の模索が続きそうである．

3. 不完全情報ゲーム解法とポーカー AI

ここまではゲーム理論の実活用について，特に社会課題に対する事例を紹介したが，本章では不完全情報ゲームの解法に着目し，主にポーカーを例としながら最近の論文を紹介したい．ポーカーでは，相手の手札がわからない状況のもとで，ゲームを降りるか，ベッド額を増やすべきか，といった意思決定を下す必要があり，不完全情報ゲーム研究ではベンチマークとしてよく使われるゲームである．

本章では，次のように記号を統一する．あるプレーヤ i にとっての情報集合を $\mathcal{I}_i$ と表現する．各要素 $I \in \mathcal{I}_i$ は，例えば特定の状態 h, h' がゲーム状態としては異なっていても（情報が不完全という理由で）プレーヤ i にとって区別できなければ同じ状態として扱う（$h, h' \in I$）．プレーヤ i の戦略 σ_i とは，例えばじゃんけんで {グー, チョキ, パー} をそれぞれ {1/2, 1/4, 1/4} の割合で出す，といったように各選択を確率的に表現したもの（混合戦略），プレーヤ i が戦略 σ に従った場合に得られる利得を $u_i(\sigma)$ とする．特に，情報ノード I に到達したら必ず行動 a を選択する戦略を $\sigma|_{I\to a}$ と表す．さらに，複数プレーヤが戦略 σ に従って行動することである情報ノード I に到達する確率 $\pi^{\sigma}(I)$ を導入し，この確率へのプレーヤ i 以外の寄与度を $\pi^{\sigma}_{-i}(I)$ と表現する．

3·1 ε - ナッシュ均衡

一般的にゲーム理論では，ナッシュ均衡を考えることがある．ナッシュ均衡とは，ゲームに参加しているプレーヤがおのおのの戦略を変更してもそれ以上高い利得が得られない状態のことである．より具体的には，2 名のプレーヤ 1, 2 を考えた場合のナッシュ均衡戦略は式（1），式（2）を満たす戦略と表現される．この式では，おのおのが取り得る戦略集合（Σ_1, Σ_2）からどのように戦略 σ_1, σ_2 を選んでも，戦略 σ を取った場合の利得を超えられないことを意味している．

$$u_1(\sigma) \geq \max_{\sigma_1' \in \Sigma_1} u_1(\sigma_1', \sigma_2) \tag{1}$$

$$u_2(\sigma) \geq \max_{\sigma_2' \in \Sigma_2} u_2(\sigma_1, \sigma_2') \tag{2}$$

じゃんけんのような簡単なケースであれば，すべての可能性を列挙して最適なナッシュ均衡解を求めることもできるが，ポーカーのように複雑なゲームではすべての可能性を網羅することは現実的ではなく，式（3），式（4）のようにナッシュ均衡を緩和した ε- ナッシュ均衡解を求めることが一般的である．

$$u_1(\sigma) + \varepsilon \geq \max_{\sigma_1' \in \Sigma_1} u_1(\sigma_1', \sigma_2) \tag{3}$$

$$u_2(\sigma) + \varepsilon \geq \max_{\sigma_2' \in \Sigma_2} u_2(\sigma_1, \sigma_2') \tag{4}$$

3·2 Counterfactual Regret Minimization（CFR）

特に不完全情報ゲームにおいて ε- ナッシュ均衡解を求めるアルゴリズムとしては，[Zinkevich 08] が提案した Counterfactual Regret Minimization（CFR）と呼ばれる手法が有名である．CFR では式（5）で定義される counterfactual regret と呼ばれる量を逐次最小化する．この式の意味するところは，プレーヤ i がある情報ノード I に到達したときに「現在の戦略 σ^t を取った場合」と「あえて行動 a を選択した場合」とでどの程度の期待利得の差があったかを表す量と考えられる．感覚的には“後悔”の度合いを表していると思ってよい．R が正の場合は，より適切な行動 a を選択しておくべきだったことを意味する．

$$R_i^T(I, a) = \frac{1}{T} \sum_{t=1}^{T} \pi^{\sigma^t}_{-i}(I) \left(u_i(\sigma^t|_{I\to a}, I) - u_i(\sigma^t, I) \right) \tag{5}$$

CFR では regret の量に比例して，式（6）のように戦略が更新される．これは regret matching（[Hart 00]）と呼ばれる手法で，$R_i^T \leq \Delta_{u,i} |\mathcal{I}_i| \sqrt{|A_i|} / \sqrt{T}$ の上限をもつことから，更新を繰り返す（$T \to \infty$）ことで ε- ナッシュ

均衡に収束することが保証されている．

$$\sigma_i^{T+1}(I,a)=\begin{cases}\dfrac{R_i^{T,+}(I,a)}{\sum_{a\in A(I)}R_i^{T,+}(I,a)} & \text{if } \sum_{a\in A(I)}R_i^{T,+}(I,a)>0\\ \dfrac{1}{|A(I)|} & \text{otherwise}\end{cases} \tag{6}$$

[Tammelin 15] は CFR を発展させた CFR＋と呼ばれるアルゴリズムによって初めて Heads-Up Limit Texas Hold'em（HULHE）を人間プレーヤと同等のレベルまで解くことに成功した．HULHE はベッド額に制約のある二人対戦ポーカーの一種で 10^{14} 程度のサイズの情報集合からなるゲームである．CFR＋では，戦略を $\bar{\sigma}_i^T=2/(T^2+T)\sum_{t=1}^T t\sigma_i^t$ と重み付けすることに加えて，regret matching の際に負の累積 regret は 0 にする regret matching＋を使うことでより高い収束効率を達成した．

3·3　ポーカー AI “Libratus” の登場

HULHE 程度のサイズのゲームであれば CFR＋を使って解くことができるが，フルサイズの二人対戦ポーカーである Heads-Up No-Limit Texas Hold'em（HUNL）となると，情報ノード数が 10^{161} まで増加するため現実的に解くことができない．そこで登場したのがポーカー AI “Libratus”（[Brown 18]）である．Libratus は，（1）似た状態を結合してゲーム木を抽象化することによってノード数を 10^{12} 程度まで縮減したうえで，CFR+ を適用して大局的な戦略を獲得する，（2）抽象化した木を部分的に展開してより精緻な戦略を獲得する，（3）ゲーム木を自己改良していく，という 3 段階の学習によって HUNL の攻略に成功した．Libratus は実際に 4 名のプロプレーヤと 12 万回に及ぶ対戦を行っており，$1.7M の賞金を勝ち越している．詳細は [小宮山 18] にも解説があるためそちらも参考にされたい．また，同時期には局面評価に CNN を用いたプログラム “DeepStack” もプロプレーヤに勝利した報告があり [Moravčík 17]，二人対戦であればポーカー AI は人間に勝てるようになってきている．ここでは詳細に立ち入らないが，多人数ポーカーへの拡張も模索は続いている（例えば [河村 16]）．

3·4　Libratus 以降のアルゴリズムの発展

Libratus の登場以降は，ポーカー以外の不完全情報ゲームへの拡張やより収束の早いアルゴリズムが研究されている．例えば [Brown 19a] は，ゲーム木の抽象化にはドメイン知識の導入が不可欠である課題に着目し，深層学習を使って抽象化と CFR を同時に扱う deep CFR を提案した．CFR アルゴリズム自体も改良が繰り返されており，以下では，regret の重み付けによって CFR＋を拡張すると同時に収束効率を高めた Discounted Regret Minimization（[Brown 19b]）を紹介する．この論文は AAAI 2019 で Outstanding Paper Award を受賞している．

3·5　Discounted Regret Minimization

CFR＋によって HUNL のような巨大な不完全情報ゲームも解けるようになったが，一方でこのアルゴリズムには大きな負の利得や準最適な選択肢がある場合に収束が遅くなることが知られている．例えば図 1 のように，終端ノードの利息がそれぞれ {0, 1, −1 000 000} となっている場合を考えてみよう．regret を計算すると，行動可能手に対する regret はそれぞれ {333 333, 333 334, 0} となる．regret matching で戦略を更新すれば，大きな負の利得が発生する三つ目のノードは選択されずに，エージェントは一つ目と二つ目のノードをほぼ 50％の確率で選択するようになる．ここまではいいが，利得が最大となる二つ目のノードを選択させるためにはどの程度の更新を繰り返せばいいだろうか．再び更新を行うと，一つ目と二つ目のノードの regret は {333 332.5, 333 334.5} となり戦略はほとんど改善しない．計算を続けていくと，二つ目のノードを選択するようになるまでには 471 407 回に及ぶ更新が必要なことがわかる．この例は極端に思えるかもしれないが，ゲームによってはこのような大きなペナルティが発生することは不自然ではない．

この課題への対策として考えられるのは，割引率を導入することで過去のイテレーションの影響を引きずらないようにすることである．[Brown 19b] では，t 番目の更新に重み w_t を導入して，dicounted regret を式（7）と定義，戦略を式（8）のように更新する手法を提案をした．

$$R_i^{w,T}(I,a)=\max_{a\in A}\frac{\sum_{t=1}^T w_t R^t(I,a)}{\sum_{t=1}^T w_t} \tag{7}$$

$$\sigma_i^{w,T}(I,a)=\frac{\sum_{t=1}^T w_t \pi_i^{\sigma^t}(I)\sigma_i^t(I,a)}{\sum_{t=1}^T w_t \pi_i^{\sigma^t}(I)} \tag{8}$$

例えば

$$w_t=\frac{t}{t+1}$$

とすれば（論文では linear CFR と命名されている），図 1 のタスクで最適戦略を選択するまでにかかる更新数は 970 まで減じることができる．彼らはこの考えを一般化して，

正の regret には　$\dfrac{t^\alpha}{t^\alpha+1}$，

負の regret には　$\dfrac{t^\beta}{t^\beta+1}$，

戦略更新には　$\left(\dfrac{t}{t+1}\right)^\gamma$

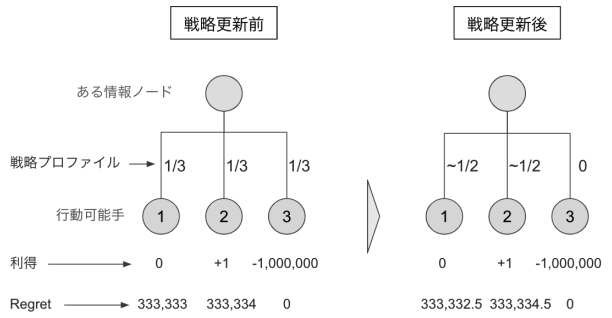

図1 CFRアルゴリズムの収束が遅くなる例

をそれぞれの重みとしたDiscounted CFR（$\mathrm{DCFR}_{\alpha,\beta,\gamma}$）を提案し，ハイパーパラメータ$\alpha,\beta,\gamma$の影響を考察している．このフレームでは，linear CFRは$\mathrm{DCFR}_{1,1,1}$に，CFR+は$\mathrm{DCFR}_{\infty,-\infty,2}$と統一的に表現することができる．

HUNLのサブゲーム4種類に加え，Goofspielと呼ばれる不完全情報のカードゲームをテストベッドとして評価した結果，$\alpha=3/2$，$\beta=0$，$\gamma=2$の組合せがどの既存手法よりも高い収束性を示すことが確認されている．ゲームによって適切なパラメータを選択する必要はあるが，これまで提案されてきたCFRの改善手法を統一的に評価した点，また理論的な収束条件も丁寧にフォローしている点から評価が高かったのだろうと思われる．

また，関連する話題として，近年はMOBA（Multiplayer Online Battle Arena）と呼ばれる複数人対戦の不完全情報ゲームにおいて，人間のトップチームにも勝てるAIが登場している．『DotA2』ではプログラム“OpenAI Five”が[*3]，『StarCraft II』ではプログラム“AlphaStar”が[*4]，それぞれゲームの攻略を達成した．これらは「できる限りゲームの事前知識を排除した人間レベルの知能の再現」を目指し，分散型の深層強化学習を活用している．それに対して本稿で紹介したポーカーAIは「モデル（ゲーム木）が所与の前提で不完全情報問題をゲーム理論で解く」ことの効率化を目指しており，研究目的とアプローチが異なる．一方で，近年は深層強化学習もマルチエージェント課題を積極的に扱うようになり，自己対戦の文脈でゲーム理論を援用している．今後，ポーカーAIなどに用いられている“DeepStack”やDeep CFRといった深層学習による状態価値の抽象化手法が発展することで，事前知識に依存しすぎずにより高度な状態表現の学習が可能になると考える．そうなれば，両者の技術はマルチエージェント強化学習の文脈でより近く密接に混合されていくだろう．今後も，強化学習とゲーム理論を使ったゲーム理論の発展には注目が集まりそうだ．

*3 https://openai.com/five/
*4 https://deepmind.com/blog/alphastar-mastering-real-time-strategy-game-starcraft-ii/

4. ま と め

本稿では，AAAI 2019の概要とゲーム理論を中心としたトピックの紹介を行った．紹介した話題は，ポーカーのようなゲームの攻略だけでなく，1章で触れたようにさまざまな現実課題を解決するための活用が期待されている．企業でゲームAIを研究開発している立場としても，このような応用の場を知ることができたのはとても意味のある経験であった．ゲーム理論やゲームAIの現状について，読者に興味をもってもらえたら著者としても幸いである．

◇ 参 考 文 献 ◇

[Brown 18] Brown, N. and Sandholm, T.: Superhuman AI for headsup no-limit poker: Libratus beats top professionals, *Science,* Vol. 359, No. 6374, pp. 418-424（2018）
[Brown 19a] Brown, N., Lerer, A., Gross, S. and Sandholm, T.: Deep counterfactual regret minimization, Chaudhuri, K. and Salakhutdinov, R., eds., *Proc. 36th Int. Conf. on Machine Learning,* Vol. 97 *of Proc. Machine Learning Research,* pp. 793-802, Long Beach, California, USA, PMLR（2019）
[Brown 19b] Brown, N. and Sandholm, T.: Solving imperfect information games via discounted regret minimization, *33rd AAAI Conf. on Artificial Intelligence*（2019）
[Gholami 18] Gholami, S., Mc Carthy, S., Dilkina, B., Plumptre, A., Tambe, M., Driciru, M., Wanyama, F., Rwetsiba, A., Nsubaga, M. and Mabonga, J., et al.: Adversary models account for imperfect crime data: Forecasting and planning against real-world poachers, *Proc. 17th Int. Conf. on Autonomous Agents and MultiAgent Systems,* pp. 823-831, International Foundation for Autonomous Agents and Multiagent Systems（2018）
[Hart 00] Hart, S. and Mas-Colell, A.: A simple adaptive procedure leading to correlated equilibrium, *Econometrica,* Vol. 68, No. 5, pp. 1127-1150（2000）
[河村 16] 河村圭悟，直紀水上，慶雅鶴岡：多人数不完全情報ゲームにおける仮想自己対戦を用いた強化学習，ゲームプログラミングワークショップ2016論文集, 第2016巻, pp. 188-195（2016）
[Kiekintveld 09] Kiekintveld, C., Jain, M., Tsai, J., Pita, J., Ordóñez, F. and Tambe, M.: Computing optimal randomized resource allocations for massive security games, *Proc. 8th Int. Conf. on Autonomous Agents and Multiagent Systems,* Vol. 1, pp. 689-696, International Foundation for Autonomous Agents and Multiagent Systems（2009）
[Killian 19] Killian, J. A., Wilder, B., Sharma, A., Choudhary, V., Dilkina, B. and Tambe, M.: Improving tuberculosis treatment by integrating optimization and learning, *36th Int. Conf. on Machine Learning*（2019）
[小宮山 18] 小宮山純平：機械学習による意思決定，人工知能, Vol. 33, No. 5, pp. 637-640（2018）
[Moravčík 17] Moravčík, M., Schmid, M., Burch, N., Lisỳ, V., Morrill, D., Bard, N., Davis, T., Waugh, K., Johanson, M. and Bowling, M.: Deepstack: Expert-level artificial intelligence in heads-up no-limit poker, *Science,* Vol. 356, No. 6337, pp. 508-513（2017）
[奥村 19] 奥村エルネスト純：会議報告：The 33rd AAAI Conf. on Artificial Intelligence（AAAI 2019），人工知能，Vol. 34, No. 3, pp. 427-428（2019）
[Pita 08] Pita, J., Jain, M., Ordónez, F., Portway, C., Tambe, M.,Western, C., Paruchuri, P. and Kraus, S.: ARMOR Security for Los Angeles International Airport, *AAAI,* pp. 1884-1885（2008）

[Pita 11] Pita, J., Tambe, M., Kiekintveld, C., Cullen, S. and Steigerwald, E.: GUARDS: Game theoretic security allocation on a national scale, *10th Int. Conf. on Autonomous Agents and Multiagent Systems*, Vol. 1, pp. 37-44, International Foundation for Autonomous Agents and Multiagent Systems（2011）

[Tammelin 15] Tammelin, O., Burch, N., Johanson, M. and Bowling, M.: Solving heads-up limit Texas Hold'em, *24th Int. Joint Conf. on Artificial Intelligence*（2015）

[Wilder 19] Wilder, B., Dilkina, B. N. and Tambe, M.: Melding the data-decisions pipeline: Decision-focused learning for combinatorial optimization, *CoRR,* Vol. abs/1809.05504（2019）

[Xu 16] Xu, H., Freeman, R., Conitzer, V., Dughmi, S. and Tambe, M.: Signaling in Bayesian stackelberg games, *Proc. 2016 Int. Conf. on Autonomous Agents & Multiagent Systems,* pp. 150-158, International Foundation for Autonomous Agents and Multiagent Systems（2016）

[Yadav 16] Yadav, A., Chan, H., Jiang, A., Rice, E., Kamar, E., Grosz, B. and Tambe, M.: Pomdps for assisting homeless shelters-computational and deployment challenges, *Int. Conf. on Autonomous Agents and Multiagent Systems,* pp. 67-87, Springer（2016）

[Zinkevich 08] Zinkevich, M., Johanson, M., Bowling, M. and Piccione, C.: Regret Minimization in Games with Incomplete Information, Platt, J. C., Koller, D., Singer, Y. and Roweis, S. T., eds., *Advances in Neural Information Processing Systems,* Vol. 20, pp. 1729-1736, Curran Associates, Inc.（2008）

2019 年 7 月 2 日　受理

著　者　紹　介

奥村　エルネスト　純

京都大学，東京大学，ローレンス・バークレー国立研究所にて宇宙物理学の研究に従事し，2014 年株式会社ディー・エヌ・エー入社．データアナリストとしてゲーム事業やオートモーティブ事業のデータ分析に携わり，2016 年末より AI エンジニアに転身．強化学習，深層学習を活用した Game AI 研究開発プロジェクトをリード．強化学習コミュニティの運営や書籍の執筆など対外活動も積極的に行っている．共著に「データサイエンティスト養成読本ビジネス活用編」（技術評論社，2018）がある．