6S-08

入れ子型並列交渉問題のための Deep Deterministic Policy Gradient

荒川 亮太 †　　藤田 桂英 ‡

† 東京農工大学　工学部　知能情報システム工学科　　‡ 東京農工大学大学院　工学研究院

1　はじめに

サプライチェーンマネジメントの商品取引に関する交渉を効率的に実行するために，人間に代わってエージェント同士が交渉する自動交渉の導入が検討されている．例えば自動交渉エージェント競技会 (ANAC) では，サプライチェーン内の工場の運営と工場間の自動交渉を行うエージェントの性能を競う Supply Chain Management League(SCML) が開催されており，様々な交渉エージェントが開発されている．また，自動交渉の分野においては，強化学習を利用して交渉戦略を獲得する研究が進められている．このような研究では 1 対 1 の交渉を取り扱っている．自動交渉をサプライチェーンマネジメントに応用する場合，供給側のエージェントと需要側のエージェントの 2 方向に対して同時に交渉する「入れ子型並列交渉」を行う必要がある．入れ子型並列交渉では，一方のエージェントとの交渉結果に応じてもう一方のエージェントとの交渉戦略を変更する必要が生じるため，単に 1 対 1 の交渉を並列して実行するだけでは不十分である．

本研究では，入れ子型並列交渉の交渉戦略を，Deep Deterministic Policy Gradient(DDPG) [1] によって獲得するためのフレームワークを提案する．

2　問題設定

サプライチェーンは 3 層に分けられ，各層に 1 つのエージェントが配置される．各エージェントは他のエージェントと，売買する商品の価格や個数について交渉を行う．価格や個数に関する提案を相互に送信し，一方のエージェントがもう一方のエージェントの提案に合意すれば交渉が成立する．1 日ごとに交渉を行い，当日中に資金や商品の移動が発生する．中間層のエージェントの目的は，入れ子型並列交渉によって原材料の購入・生産物の売却を行い，より多くの利益を上げることである．商品が売れ残った場合，納品数が不足した場合，生産数が少ない場合にはペナルティコストが発生する．詳細な交渉プロトコルなどは SCML OneShot Track のルール [2] に基づく．

3　提案手法

3.1　環境設計

学習に用いる状態空間は，相手エージェント i ごとに区別される交渉状態 S_{n_i} と，環境状態 S_e から構成される．交渉状態は次の 3 要素のタプルである．

- NT: 現在の交渉ラウンド数
- OQ: 最後に受け取った提案の個数
- OP: 最後に受け取った提案の価格

環境状態は次の 2 要素のタプルである．

- SS: 確保している原材料の個数
- DS: 契約が成立している納品数

行動空間 A は次の 2 要素のタプルである．本エージェントは相手の提案を自発的に受け入れない．

- PQ: 提案する個数
- PP: 提案する価格

当日の交渉が終了したとき，式 (1) で表される報酬 r を与える．u は効用値，p は合意価格，u_h, u_l は効用値のスケーリングのためのパラメータ，p_{min}, p_{max} は提案可能な価格の最小値・最大値を表す．割引率は 1 として，交渉継続時には報酬を与えない．式 (2) で表される報酬 r_p を与えることで学習の高速化を図っている．

$$r = \begin{cases} \max\left(0, \dfrac{u - u_l}{u_h - u_l}\right) + r_p & \text{(Deal)} \\ -1 & \text{(No Deal)} \end{cases} \tag{1}$$

$$r_p = \begin{cases} \dfrac{p - p_{min}}{p_{max} - p_{min}} & (i \text{ is consumer}) \\ 1 - \dfrac{p - p_{min}}{p_{max} - p_{min}} & (i \text{ is supplier}) \end{cases} \tag{2}$$

Deep Deterministic Policy Gradient for Nested Parallel Negotiation Problem

†Department of Electrical Engineering and Computer Science, Faculty of Engineering, Tokyo University of Agriculture and Technology

‡Institute of Engineering, Tokyo University of Agriculture and Technology

3.2 学習アルゴリズム

提案するアーキテクチャの概略図を図 1 に示す．情報抽出層は 2 つの全結合層 υ_0, υ_1 で構成され，それぞれ交渉状態ベクトル s_{n_0}, s_{n_1} の解釈を行う．状態ベクトルはそれぞれの全結合層で変換された後，環境状態ベクトル s_e と結合されて変換後ベクトル s_t として出力される．Actor Network θ と Critic Network ϕ は s_t を入力とする DDPG と同様である．交渉相手の戦略を独立したネットワーク υ_0, υ_1 で学習し，Actor Network と Critic Network で共有することで，パラメータ数の削減を図っている．

図中の実線はレイヤー間で誤差逆伝播が行われることを示し，点線は誤差逆伝播が行われないことを示す．$\upsilon_0, \upsilon_1, \phi$ のパラメータは r から逆伝播を行って同時に更新される．θ の学習は独立して行われ，情報抽出層への逆伝播は発生しない．

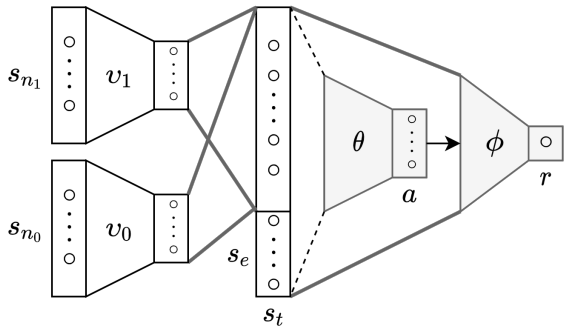

図 1　提案するアーキテクチャ

4 実験

提案したフレームワークの評価を行うため，交渉のシミュレーションを行った．供給側エージェントは TimeDependent 戦略，需要側エージェントは FootInTheDoor 戦略で提案を行った．通常の DDPG を使用し，報酬 r_p を与えないエージェント (DDPG w/o r_p), 通常の DDPG と報酬 r_p を使用するエージェント (DDPG w/ r_p)，提案アーキテクチャと報酬 r_p を使用するエージェント (Proposed) を比較した．エージェントの実装には PyTorch と自動交渉プラットフォーム NegMAS [3] を用いた．

4000 エピソードで学習した後の各エージェントの個別効用値を図 2 に示す．図中の水平線は，個数を考慮せずに交渉するルールベースのエージェントの効用値 (-646) を表す．報酬 r_p を与えて学習を行うことにより，ルールベースのエージェントよりも高い効用値を得られた．また，提案したアーキテクチャによって学習した場合の効用値の平均は DDPG w/ r_p と同程度であり，その分散はより小さかった．提案したアーキテクチャのパラメータ数は 1663, 通常の DDPG のパラメータ数は 2747 であり，約 6 割のパラメータ数で同程度の効用値を実現した．

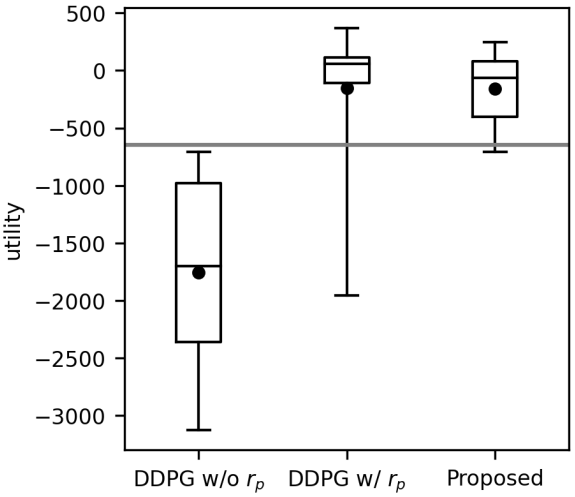

図 2　効用値の比較

5 まとめ

本研究では，入れ子型並列交渉の交渉戦略を学習するためのフレームワークを提案した．交渉のシミュレーションにより，合意価格に基づく報酬 r_p の付与が有効であること，提案したアーキテクチャがより少ないパラメータ数で通常の DDPG と同等の学習性能を持つことを示した．

参考文献

[1] Timothy P Lillicrap, Jonathan J Hunt, Alexander Pritzel, Nicolas Heess, Tom Erez, Yuval Tassa, David Silver, and Daan Wierstra. Continuous control with deep reinforcement learning. *arXiv preprint arXiv:1509.02971*, 2015.

[2] Yasser Mohammad, Amy Greenwald, Katsuhide Fujita, Mark Klein, Satoshi Morinaga, and Shinji Nakadai. Supply chain management league (oneshot). `http://www.yasserm.com/scml/scml2022oneshot.pdf`, 2022.

[3] Yasser Mohammad, Shinji Nakadai, and Amy Greenwald. Negmas: A platform for situated negotiations. In *Recent Advances in Agent-based Negotiation*, pages 57–75, 2021.