The 33rd Annual Conference of the Japanese Society for Artificial Intelligence, 2019

深層強化学習による物流プロセスの全体最適化

Global optimization for supply chain process by deep reinforcement learning

小池和弘 *1
Kazuhiro Koike

*1アスクル株式会社
ASKUL Corporation

The bullwhip effect is known as one of the problems in the supply chain. As a result of demand forecasting and decision-making, demand propagates from downstream to upstream while amplifying. This phenomenon is well reproduced by the Beer Game invented in the 1960's. On the other hand, in online shopping, there is a gap between the information-flow in cyberspace and the object-flow in physical space. This gap can be a factor to promote *the bullwhip effect* , but it is difficult to reproduced with the original Beer Game. Therefore, we set up the new game called "Netshop Game" which extended the rules and the environment. On the new game, by using deep reinforcement learning, we are able to reproduce the local optimum that can occur in net shopping supply chain, and confirmed that it is effective for discovering a global optimum by introducing a meta viewpoint.

1. はじめに

一般的にサプライチェーンで起きうる問題の一つとして *Bullwhip effect* (以下 BE) が知られている．これはサプライチェーンの下流における需要予測と意思決定の結果，需要が拡大しながら下流から上流に向かって伝搬していく現象であり，最初に認識されたのは 1958 年である [Forrester]．この現象は過剰在庫や欠品に繋がるため，発生メカニズムと抑制手法は長年に渡り研究対象となっている [Lee04]．Lee らは BE の要因として、価格表、発注頻度、返品方針、価格販売施策の頻度と深さ、情報共有の程度、需要予測方法、欠品時の配分ルールなどを挙げている [Lee97]．

BE が発生する様子は Beer Game (以下 BG) によってうまく再現される．BG は 1960 年代に MIT で考案されたシミュレーションゲームであり，直列に繋がったビールのサプライチェーンで，4 人の player が Retailer，Wholesaler，Distributor，Manufacturer となって，決められた期間内でのコスト最小化を競う (図 1)．

図 1: Beer Game

ネット通販においては，サイバー空間での情報の流れとフィジカル空間での物の流れの効率の差が顕著になってきており，このズレが BE の新たな要因となる可能性がある．例えば，EC サイトでの高度に効率化された販売施策によって需要の変動が増幅された結果，配送遅延や欠品，過剰在庫などが発生する場合が考えられる．販売施策が BE の要因であることは，前述の通り Lee らによって既に指摘されているが，ネット通販では意図的に過剰に強気な販売施策を取らなくても，構造的に組み込まれていると考えられる．実体と重量を持った Atom の移動は Bit のように容易ではないからである．

このサイバーとフィジカルのズレはオリジナルの BG では再現できないため，本研究ではルールを拡張した Netshop Game を新たに設定し (図 2)，ネット通販の物流プロセスにおける諸問題の再現と全体最適化に深層強化学習を用いる方法について検討と評価を行った．

連絡先: 小池和弘，アスクル株式会社，東京都江東区豊洲 3-2-3，
kazuhiro.koike@askul.com

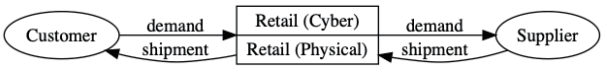

図 2: Netshop Game

2. 関連研究

BG の過程は MDP (*Markov Decision Process*) として知られており，かつ観測できる情報は，隣り合う player との注文と商品のやり取りと自身の在庫レベルのみであるため，POMDP (*Partially Observable Markov Decision Process*) である．各 player は観測可能な部分情報からコストを最小化する行動を選択するのだが，Observation 空間と Action 空間は大きく，非定常な時系列を扱うため複雑な問題である．Mnih らによって提案された DQN (deep Q-network) [Mnih] は，このような複雑な問題を克服する方法として有望である．BG の深層強化学習によるアプローチは，例えば [Oroojlooyjadid] や [Fuji] があり，効果が報告されている．

3. 提案手法

ネット通販の物流プロセス問題を扱う場合，BG そのままでは適用しにくい．ネット通販では，EC サイトやネットによる取引などサイバー空間で完結するプロセスと，物流倉庫や配送センターなどフィジカル空間で行われるプロセスでは特性の違いがある．例えばサイバー空間では商品 100 個はあくまでも数値データ (Bit) であり，100 個を強気な販売施策によって 10 倍にすることについて物理的な制約を受けにくい．一方でフィジカル空間では商品 100 個は体積と重量を持つ実体 (Atom) であり倉庫のキャパシティや出荷能力など物理制約の影響を大きく受ける．また BG では上流 player への注文のリードタイムも設定できるが，ネットによる取引では無視できる．

BG では，直列に連結された Retailer，Wholesaler，Dis-

tributor，Manufacturer の 4 player がゲームに参加するのだが，上記のネット通販特有の問題にフォーカスするため，player を Retailer のみとし，上流の 3 player，すなわち Wholesaler, Distributor, Mnufacturer は Supplier として一括りで考えることにする．そして Retailer の中はサイバー空間のプロセスを管理する Cyber player と，フィジカル空間を管理する Physical player に分ける．ゲームのゴールは，Cyber は Customer からの需要に対してどれだけ欠品なくデリバリーできたかを示す *fill rate*（以下 FR) が設定した閾値より大きくなること，Physical は *bullwhip effect index*（以下 BEI) の値が設定した閾値より小さくなることとした．

これを Netshop Game と名付け，この環境を OpenAI Gym で実装し，Cyber player，Physical player それぞれの最適な行動を DQN によって学習し評価する．2 つの player にはそれぞれ偏った報酬の与え方とゴールを設定し，あえて個別最適化行動するようにして，どのような結果になるかを観察する．両者の報酬の与え方とゴール設定はいわば Netshop Game におけるジレンマである．ここに全体最適化のため，ジレンマを抱えながらバランスをとることを目指す高次元な視点を持った Meta player を加えた．

4. player 定義

Netshop Game の player は次の 3 タイプとする．

- CYBER：サイバー空間で主に報酬最大化を狙いとする player である．EC サイト上でセール，ポイント n 倍，などのセールスプロモーションを積極的に行う．在庫レベルの上昇によって生じるコストは無視し，欠品による機会損失を最小化するように行動する．
- PHYSICAL：フィジカル空間で主に物流コスト最小化を狙いとする player である．欠品による機会損失については無視し，倉庫や配送のコストを最小化すべく在庫レベルを低く抑え，BE を抑制するように行動する．
- META：CYBER と PHYSICAL それぞれの報酬とゴール条件のジレンマを抱えてバランスをとるように行動する．

5. 環境設計

5.1 状態変数

タイムステップ t における観測可能な状態変数 o_t を式 1 で定義する．

$$o_t = [IL_t, OO_t, d_t, RS_t, SS_t, a_t] \quad (1)$$

$$ho_t = [o_1, ..., o_t] \quad (2)$$

IL_t は t における在庫数，OO_t は supplier に対して発注したがまだ入荷していない商品数，d_t は customer からの需要，a_t は t における action すなわち supplier への発注数，RS_t は supplier から入荷した商品数，SS_t は customer に対して出荷した商品数である．

ho_t は 1 エピソードの全タイムステップの状態変数を historical observation として記憶する (式 2)．

状態変数については [Oroojlooyjadid] の定義を参考にし，ゴール条件の BEI と FR に関与する SS_t を追加した．

5.2 Action 空間

Netshop Game における action は，supplier に対する商品の発注数であり Action 空間の自由度をどこまで許容するかは慎重に決める必要がある．ここでは大きすぎる空間はメモリ効率と処理時間に悪影響があると考え，player が選択可能な Action 空間は 0 から 20 の離散値集合 $[0, 1, 2, ..., 20]$ としたが，上限について特に強い根拠はない．

5.3 報酬

BG では 式 3 に示す通り，在庫レベルによって報酬が決まる．在庫数が正の場合は在庫数分の在庫コスト c_h を，負の場合は欠品による機会損失 c_p をコストとする [Oroojlooyjadid]．なお右側の Σ は 4 player の総計を求めるためである．

$$\sum_{t=1}^{T}\sum_{i=1}^{4} c_h^i (IL_t^i)^+ + c_p^i (IL_t^i)^- \quad (3)$$

$$(x)^+ \quad : \quad max(0, x)$$

$$(x)^- \quad : \quad max(0, -x)$$

Netshop Game では売値，仕入れ値，販売促進費，配送費を追加した．s_p は売値，c_r は仕入れ値，c_s は販売促進費，c_d は配送費とする．なおこれらの値は環境の中に隠され，各 player からは観測できない．式 4，式 5，式 6 はそれぞれ CYBER, PHYSICAL，META の報酬である．

$$\sum_{t=1}^{T} s_p SS_t - c_r RS_t - c_p (IL_t)^- \quad -c_s (a_t - d_t)^+ \quad (4)$$

$$\sum_{t=1}^{T} s_p SS_t - c_r RS_t - c_h (IL_t)^+ \quad -c_d SS_t \quad (5)$$

$$\sum_{t=1}^{T} s_p SS_t - c_r RS_t - c_p (IL_t)^- \quad -c_h (IL_t)^+ \quad -c_s (a_t - d_t)^+ - c_d SS_t \quad (6)$$

CYBER は欠品の機会損失コストに加え supplier への発注量が需要より大きければその差分を販売促進費として加算する．PHYSICAL は在庫数分の在庫コストと customer への出荷数分の配送コストを加算する．CYBER は過剰在庫を気にせず，PHYSICAL は欠品を気にしないという偏った報酬の設定は，局所最適に陥る状況を意図的に発生させるためである．META はコストを全て加算するが，これはバランスをとったグローバルな解を求めるためである．

5.4 ゴール条件

BG の場合は決められたタイムステップ期間の総報酬で競うのであるが Netshop Game では二つの指標をゴール条件として設定する．ゴール条件を満たしたら episode 終了となる．指標の一つは *bullwhip effect index* (BEI) 式 7 であり，もう一つは *fill rate*（FR) 式 8 である．なお，demand は t における需要数の直近 p 期間分の配列，shipped は t における出荷数の直近 p 期間分の配列，$Var(x)$ は x の分散，$Mean(x)$ は x の平均である．式 9 は PHYSICAL のゴール条件，式 10 は CYBER のゴール条件，META はその AND である．

$$BEI = Var(shipped)/Var(demand) \quad (7)$$

$$FR = Mean(shipped/demand) \quad (8)$$

$$\texttt{BEI threshold} > BEI \quad (9)$$

$$\texttt{FR threshold} < FR \quad (10)$$

5.5 需要生成

t における需要 D_t は直近 p 期間の平均に変動分として正規分布に従う確率変数 x を加えて生成される．なお初期値は 1 から 10 までの自然数からランダムに選択する．

$$D_t = \frac{\sum_{j=t-p+1}^{t} d_j}{p} + x \qquad x \sim \mathcal{N}(\mu, \sigma^2) \tag{11}$$

5.6 アルゴリズム

Action の結果から得られた経験の蓄積と活用のトレードオフバランスをとる方法として ϵ-greedy algorithm を採用し，状態評価には DNN を採用した．

Listing 1: Netshop Game Algorithm

```
procedure DQN
    for episode = 1 : n do
        reset environment
        for t = 1 : T do

            a_t = { take random action prob. ε
                  { arg min_a Q(s_t, a, θ)(otherwise)

            observe reward r_t and state s_{t+1}
            mini-batch (s_j, a_j, r_j, s_{j+1})

            y_j = { r_j                    (goal)
                  { r_j + min(Q(s, a, θ))  (otherwise)

            loss function (y_j − Q(s_j, a_j, θ))^2
        end for
    end for
end procedure
```

6. 学習

CYBER，PHYSICAL，META それぞれの player について 50,000 steps を上限として学習を行った．図 3 は META player agent の学習が進むにつれて報酬が上昇していく様子を示したものである．

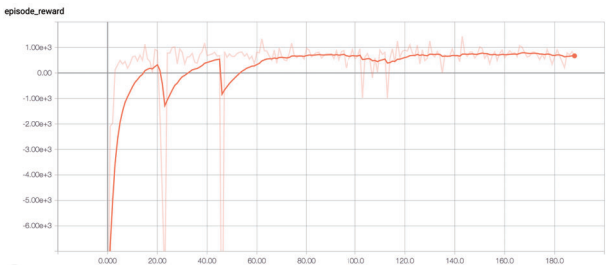

図 3: episode-rewards of META player Agent

7. テスト

学習済みの CYBER，PHYSICAL，META 各 player agent モデルを使って それぞれ 100 episodes のテストを行った．1 episode のタイムステップは上限を 1,000 とし，ゴール条件に達していなくても打ち切ることとした．ゴール条件は学習時と同じである．それぞれ最初の 9 episodes の 最後の 100 steps について，demand，action，stock の推移グラフを示す．それぞれ CYBER:図 4，PHYSICAL: 図 5，META:図 6 である．また，図 7 は各 player の獲得報酬のヒストグラムである．

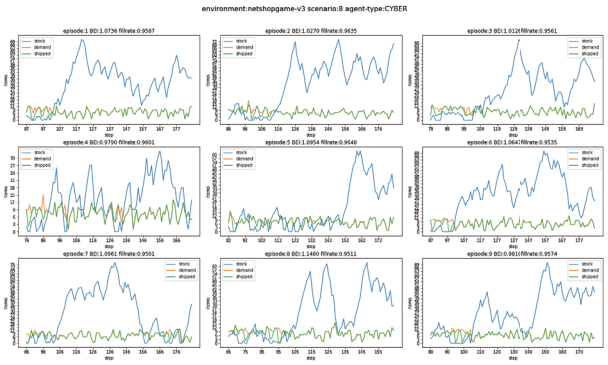

図 4: CYBER player agent

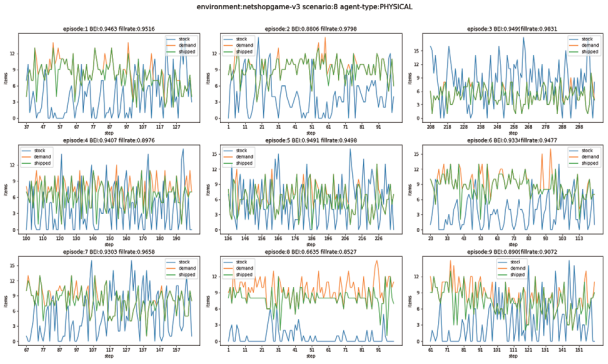

図 5: PHYSICAL player agent

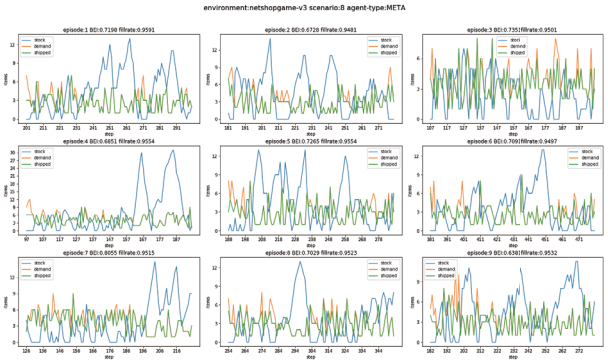

図 6: META player agent

8. 考察

表 1 はテスト結果である．steps はゴールに至るまでにかかった step 数であり，小さいほど良い．BEI は 1.0 より小さければ BE が抑制されたと考えられる．FR は customer からの需要に対して出荷できた割合を示しており 1.0 に近いほど良い．rewards は episode で得た報酬合計である．

報酬の最大化が目的であれば CYBER が一番良いが，CYBER player の episode の推移を見ると，全体的に在庫が高い

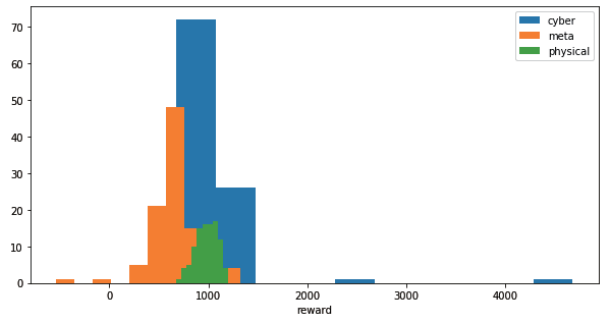

図 7: reward histgram

レベルであり続けるという現象が起きている（図 4 参照）．これは過剰在庫に対して抑制する要素を与えていないため当然予想された結果であるが，在庫レベルが抑制されている episode 4 が興味深い．BEI をゴール条件に入れていないにも関わらず他の episode に比べて低い値になった．発注数 a と 需要 d の差によって生じる販売促進費の抑制効果が働いていると考えられる．

PHYSICAL player は 全体的に低い在庫レベルを維持し BEI も低い水準であるが，episode 8，9 のようにしばしば欠品することがあり，結果として FR も低い値になっている（図 5 参照）．これも報酬とゴール設定から予想通りの結果である．

META player の BEI と FR は良い成績であるが獲得報酬が低く，ゴール達成までに他の Agent に比べて倍以上の steps がかかっている．バランスを見出すことが簡単では無いことがわかる．episode 2，5 のように一部であるが周期と高さが一定の波形になっている部分があり，在庫を安定させるための適正な数量管理を学習する可能性を示した．(図 6 参照)

表 1: テスト結果 (100 episodes の平均値)

player	steps	BEI	FR	rewards
CYBER	182.5200	1.0658	0.9562	1051.5985
PHYSICAL	171.4000	0.8388	0.9083	982.1282
META	369.4100	0.7025	0.9471	649.0225

9. 結論

本研究ではネット通販における物流プロセスを抽象化して Netshop Game という形をとったが，その目的は次の通りであった．

1. player 間で共有できない情報を環境の中に隠した POMDP においても解がみつかることを確認する．
2. player に偏った報酬の与え方とゴール条件設定をすることで局所最適に陥る様子を再現する．
3. 報酬とゴール条件設定のジレンマに対して，メタな視点を導入することでバランスの良い解が得られるかどうか確認する．

目的 1 については，BE の抑制に player 間の情報共有が有効であることが分かっている [Lee97]．ビジネス的に共有できない仕入れ値などの情報を隠しても解は見つけられることが確認できた．目的 2 についてはほぼ予想通りの結果が再現でき，報酬とゴール設計の重要さを確認できた．目的 3 のジレンマとは，BEI と FR のことであり，BEI を低くするには FR を下げないとならないという関係になっている．逆も同様であり FR を上げて 1.0 に近づけるためには BEI が高くなる．META player が目指したのはこのトレードオフのバランスであるが，バランスは取れたものの獲得報酬は低いことが課題である．

10. おわりに

BG が考案された 1960 年代とネット通販隆盛の現在の違いは，サプライチェーンの情報の流れが著しく進化し効率化された点である．情流に比べると物流については比較的進化の速度は緩やかであるため，EC サイトやネット取引などのサイバー空間ではデジタル化と AI によるデータ活用が進み，倉庫や配送などフィジカル空間との間で効率化においてズレが起きていると考えられる．そのような状況においては，例えば過度に強気な販売施策が極めて効率的に行われた場合，通常起こりうる需要の変動以上の変動が発生し，BE の影響は更に大きくなると考えられる．

サイバー側の player とフィジカル側の player が個別に指標を定めて最大化を追求するということは，機能ごとに分かれた組織においては合理的にみえるので，個別最適な状況は構造的に起こりうる．しかしそれでは全体として利益は期待できないため，全体最適化には高次元な，つまりメタな視点が必要なのである．

物流問題においては教師データが無い問題が多く，player 間で共有しづらい情報やそもそも観測が困難な情報がある．物理的な設備の変更はコストが掛かるため様々な施策を試すことは容易では無い．そのため仮想的な環境内で相互作用から学習することができる深層強化学習が有効であると考えている．本研究ではその有効性を示すことができたと考えている．

参考文献

[Forrester] J. W. Forrester. Industrial dynamics: A major breakthrough for decision makers. Harvard Bus. Rev. (July/August 1958) 36 3766.

[Fuji] Taiki Fuji et al. Deep Multi-Agent Reinforcement Learning using DNN-Weight Evolution to Optimize Supply Chain Performance. doi 10.24251/HICSS.2018.157 (2018)

[Lee97] H. L. Lee et al. Information distortion in a supply chain: The bullwhip effect. Management Science, 43(4):546-558, (1997).

[Lee04] H. L. Lee et al. Comments on “Information Distortion in a Supply Chain: The Bullwhip Effect”. Management Science 50 (12 supplement). 1887-1893 (2004).

[Mnih] V. Mnih et al. Human-level control through deep reinforcementlearning, doi 10.1038/nature14236 (2015).

[Oroojlooyjadid] Afshin Oroojlooyjadid et al. A Deep Q-Network for the Beer Game: A Reinforcement Learning Algorithm to Solve Inventory Optimization Problems, arXiv:1708.05924v2 (2018).