強化学習による実プラント設備への適用検証

APPLYING REINFORCEMENT LEARNING TO REAL EQUIPMENT SYSTEM FOR PROCESS CONTROL

高見豪 *A)、旭沢仁 A)、松原崇充 B)
Go Takami* A), Hitoshi Asahizawa A), Takamitsu Matsubara B)
A)Yokogawa Electric Corporation
B)Nara Institute of Science and Technology

Abstract

In recent years, the world's expectations for artificial intelligence (AI) are high. Especially, the technology called machine learning is being investigated in many systems. The use of machine learning is also being considered in accelerator systems. There are many tuning parameters in the accelerator system. It takes a long time to adjust these parameters and the finish is often dependent on the system tuner. As such a reason, the attention is focused on reinforcement learning that involves trial and error. In this paper, we introduce an experimental result of applying reinforcement learning to the three-tank level control system and a comparison result with PID control implemented by PLCs used in the accelerator. Moreover, we discuss the future, which uses reinforcement learning for tuning parameters in the accelerator system.

1． はじめに

従来より加速器の制御システムには多くの PLC が使用されてきた。これに加えて、近年ではデータ収集システムにも多く使われるようになっており、例えばビームモニター用途として加速器中のビーム位置を一定周期で計測することにも使用されている。今後はその結果をもとにビームロスを自動調整するなど、単なる制御ではなくリアルタイムに収集したデータをもとにフィードバックをかけるなどの用途が期待されている。リアルタイムに収集した多くのデータを活用する技術として、機械学習の技術が考えられる。

人工知能（AI) やそれを支える機械学習と呼ばれる技術への期待が近年高くなっており、様々な分野での応用が始まっている。加速器のシステムにおいても機械学習の技術を応用することが提案されている [1]。提案手法ではディープニューラルネットワーク (DNN) を使い、調整パラメータと環境パラメータの相関を学習し、最適となる調整パラメータを予測している。DNN は過去のデータを学習することで、過去に得られた環境のデータから考えられる最適なパラメータを導出する。しかし、環境ごとに変わる調整では、環境ごとでの学習が必要になると考えられる。そのため、現地での調整には経験者による試行錯誤が必要になると考えられる。

機械学習では試行錯誤によって最適な行動規則を学習する技術に強化学習がある。一般に強化学習の実化学プラントへの応用可能性があまり検証されていない。強化学習では膨大なサンプル数が必要となるため、従来の強化学習技術では現実の設備に対して適用することが困難であるとされてきた。一方で近年サンプル効率の高い強化学習アルゴリズムが提案されている (KDPP) [2]。我々は強化学習を使ったプラント最適運転の研究を進めており [3–6]、VAM シミュレータでの有効性を確認している。そこで本研究ではシミュレータではなく、現実の設備に強化学習を適用した。サンプル効率の高い KDPP を利用し、実機のプロセス制御のシステムへ適用することで、シミュレーションではなく実際の設備への有効性を確認する。本実験では、横河電機社製の F3RP70 [7] を使い、PID 制御と強化学習による制御をそれぞれ実装し、Fig. 1 で示すプロセス制御装置の 1 つである三段水槽で実験を行った。本稿では 2 章で我々が着目している強化学習を紹介し、 3 章で三段水槽のシステムおよび実験内容について紹介する。そして最後にまとめとして、強化学習の有用性について議論する。

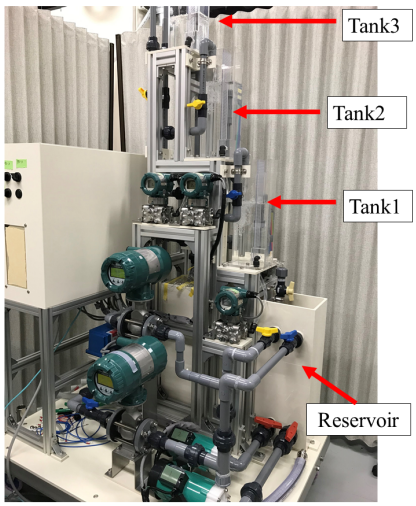


Figure 1: Three tank level control system.

2． 強化学習

強化学習とは機械学習技術の 1 つであり、マルコフ決定過程でモデル化された環境を想定し、報酬値と呼ばれる値を使って、学習を行うことが特徴である。一般的な強化学習の構成を Fig. 2 に示す。強化学習は「エージェント」と「環境」の 2 つの要素で構成されている。エー

* Gou.Takami@yokogawa.com

ジェントが自身の方策に則り、自律的に行動を起こし、環境に対して変化を起こす。そして、変化した環境を状態としてエージェントが受け取ることで、データサンプリングを実施する。その際に報酬を計算し、より高い報酬が得られるようモデルの更新を進めていく。強化学習は、制御対象の数学的モデルを用意するのではなく、エージェントが収集したデータから制御方策を導き出す手法となる。一般に強化学習では膨大なサンプル数が必要となるためエージェントによる試行回数が多く、サンプルコストが高くなってしまう。現実の設備においてサンプルコストが高いことは非常に大きな課題となるため、少ない経験サンプルでも安定した学習が期待できる手法が求められる。我々は少ない計算サンプルでも学習が期待できる Kernel Dynamic Policy Programming(KDPP) [2] に着目した。

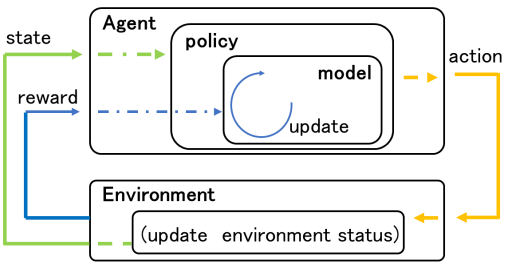


Figure 2: Reinforcement learning.

2.1 Kernel Dynamic Policy Programming

KDPP は強化学習で広く知られている Q 学習などの従来の強化学習手法とは異なり、動的方策計画（Dynamic Policy Programming:DPP) と呼ばれる手法をベースとしている。マルコフ決定過程における環境をモデル化する場合において、有限状態集合を $S = \{s_1, \ldots, s_n\}$、有限行動集合を $A = \{a_1, \ldots, a_m\}$ とする。また状態 s から行動 a を通して、状態 s' へと移る状態遷移確率を $\tau^a_{ss'}$、そのときの報酬を $r^a_{ss'} = R(s, s', a)$、減衰率を $\gamma \in (0, 1)$ とし、状態 s において行動 a を選択する方策を $\pi(a|s)$ と定義する。このとき最適な方策 π^* によって得られる価値関数 V^* は、Eq.(1) で表すことができる。

KDPP ではカルバックライブラーダイバージェンスを用いて更新前後の方策の変化量を定量化する。次に現在の方策と、期待できる報酬値が最大になると予測される方策との差を最小化することで、方策を更新し、サンプルコストを抑えることができる。さらに価値関数の近似にカーネルトリックを採用することで、高次元の状態・行動空間へのスケーラビリティを実現している。

$$V^*(s) = \max_{\pi} \sum_{\substack{a \in A \\ s' \in S}} \pi(a|s) \tau^a_{ss'} (r^a_{ss'} + \gamma V^*(s')), \forall s \in S. \quad (1)$$

3． 三段水槽への適用実験

3.1 システム構成

プロセス制御では温度、水位、流量、圧力など様々なパラメータを制御している。その中で水位制御はプロセス制御において基本的な課題の 1 つである [8]。三段水槽とは基本的な水位制御課題である。横河電機トレーニングセンターにおいても制御トレーニングの実験装置 (Fig. 1) として利用されており、システム図を Fig. 3 に示す。

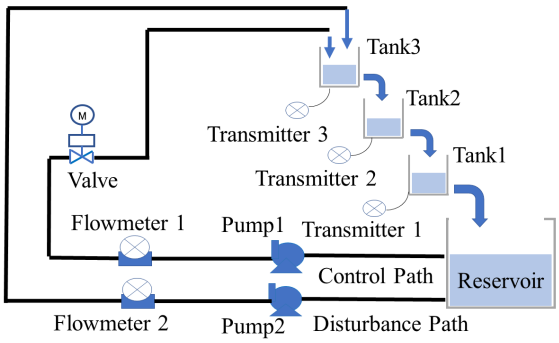


Figure 3: System archtecture.

本実験で使用した三段水槽は、通常経路の制御経路と、外乱を模擬した外乱用経路を持つ。貯水タンクからポンプ 1 で水をくみ上げ、一番上の水槽 3 に水を注ぎ、水槽 2、水槽 1 へと流れ、再び貯水タンクにもどる循環システムである。三段水槽ではバルブ開閉により水槽 3 へ流れる流量を調整し、水槽 1 の水位を制御することを目標としている。バルブ開閉のアクションが水槽 1 の水位に反映されるまでにタイムラグがあるため、マニュアル操作で水槽 1 の水位を目標値に保つことが難しいシステムである。

3.2 実験内容

本実験では次の 3 つの実験を行った。

- PID 制御
- 強化学習による制御 (Setting1)
- 強化学習による制御 (Setting2)

PID 制御の実験では、水槽 1 の水位をフィードバックするループを作成し、限界感度法 [9] を用いてパラメータチューニングを実施した。

次に強化学習を適用する実験について紹介する。本実験中、エージェントは F3RP70 上で学習・判断を行う演算処理部分のことであり、環境は三段水槽となる。状態として取得するデータは、Table 1 に示す 6 つのパラメータとした。行動はバルブ V001 を操作するものとし、バルブの現在値から加算する値として次の 5 通りの選択肢から選択されるものとした [-3%, -1%, 0%, 1%, 3%]。今回の実験では、水槽 1 の水位を 30% に制御することを目的とし、それに合わせて報酬値 R を Eq.(2) と定めた。

$$R = 0.1 * (30 - |LI001 - 30|) \quad (2)$$

実験では 1 試行を 200 ステップと定義し、1 試行あたりに 200 回エージェントにバルブの操作をさせる。1 ステップは 2 秒と定義し、各ステップで行動を起こし、その状態を観測することとした。強化学習の学習中は、試行の開始時に一度バルブを閉じて水槽 1 の水をすべて抜く処理を入れ、必ず水槽 1 が空になったところからスタートさせた。200 ステップの操作を行ったところで試行を終了させてモデルを更新し、試行回数が 30 回に達するまで繰り返し学習を進めた。強化学習による制御の実験では、1 ステップ中に収集するデータ数を変更した 2 つの条件で実施した。Setting1 では、1 ステップ中に 1 回状態を観測し 6 次元のベクトルとして状態を入力する。Setting2 では、1 ステップ中に 4 回、0.5 秒に 1 回ずつ状態を観測し、24 次元のベクトルとして状態を入力した [10]。

Table 1: Observed Values

Tag Name	Description
LI001	Tank1 Level (%)
LI002	Tank2 Level (%)
LI003	Tank3 Level (%)
FI001	Control Path Flow (%)
FI002	Disturbance Path Flow (%)
V001	Valve actuater (%)

Table 2: RL Experiment Setting

Name	Description
Setting1	6-dimensional input. Read the status once per step.
Setting2	24-dimensional input. Read the status 4 times per step.

3.3 実験結果

Figure 4 に PID 制御時の水位を示す。三段水槽はバルブを操作してから水槽 1 の水位が変化するまでに時間がかかる時定数が長い制御となるため、(A) に示すようにオーバーシュートが発生し、目標値の ±5% 以内とする制定値に入るまでに 200 秒以上かかっていることが分かる。

次に強化学習の結果を Fig. 5、Fig. 6 に示す。強化学習の結果は 30 回の試行を行った実験を 10 回実施した時の結果を示している。Figure 5a は横軸を試行回数とし、縦軸に試行中に得られた報酬値の合計値をプロットしている。Fig. 5b は横軸にステップ数、縦軸に水槽 1 の水位を示している。Figure 5c は横軸にステップ数、縦軸にバルブの開度を示している。Setting1 の結果を見ると、Fig. 5a からエージェントによる試行回数が増えるごとに徐々に取得できる報酬値が上がっている。また Fig. 5b からは制御対象とする水槽 1 の水位を 30% に制御している様子がうかがえる。しかし、目標値の ±5% 以内で安定していることはできず、10 回の実験においてもばらつきが多い。一方で Setting2 の結果を見てみると、報酬値の収束が向上していることを確認することができる。また水槽 1 の水位についても 10 回の実験におけるばらつきがほとんどなく、安定して 30% の水位へと制御できている。さらに、注目したいのは安定した制御になるまでの時間である。PID 制御の場合は Fig 4 の (A) で示したようにオーバーシュートがあるため、安定するまでに 200 秒以上の時間がかかっていた。しかし、強化学習で学習した後の制御では、オーバーシュートがほとんどなく、目標値まで滑らかに立ち上がり、100 秒 (50step) 以内で安定した制御ができている。これは、立ち上がりまでの時間を約半分にまで改善できたことを示しており、PID 制御よりも効率的にバルブ操作を行うことができていることを示している。バルブ操作を表す Fig. 6c を見ると、試行がスタートしてからすぐにバルブを全開まで開放し、水位が 30% に達する前にバルブを閉じ始めていることがわかる。強化学習のエージェントからすると、より多くの報酬値を得るためには、水位 30% にいる時間をできるだけ長くいられることがベストであり、学習過程の中で、現在のバルブ値と水位の関係を学習し、どのタイミングでバルブを閉じはめるとより多くの報酬がもらえるのかを探索した結果、バルブを閉じ始めるタイミングを学習できたものと考えられる。

Table 3: Experiment Result

Experiment	Time (s) until stable level	Level average (%) after stable status
PID	208	29.96
RL setting1	-	26.56 ± 11.95
RL setting2	82	31.36 ± 1.49

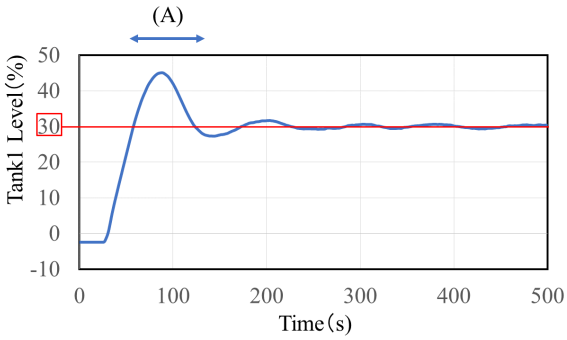
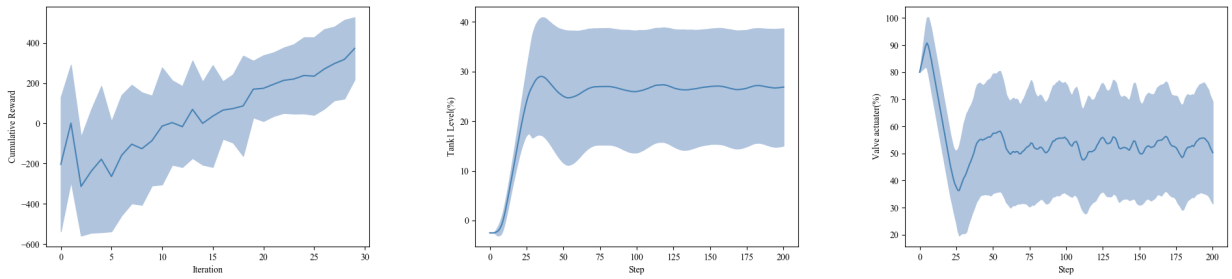


Figure 4: PID control level result.

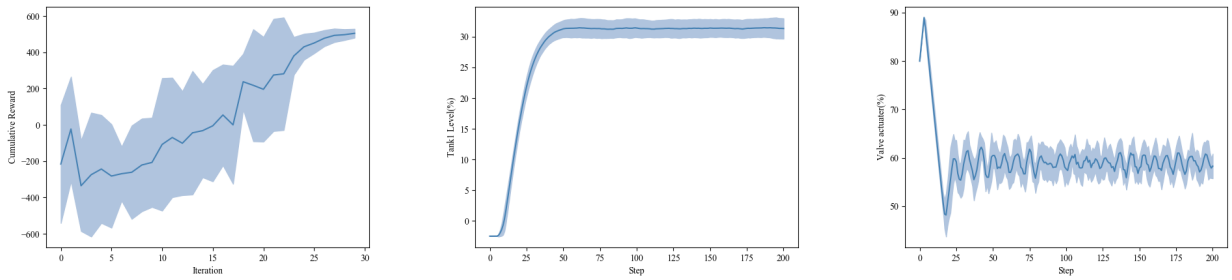
4. まとめ

本研究では、プロセス制御装置である三段水槽の制御に対して強化学習を適用し、制御を行った実験を紹介した。従来の数理ベースを基にした制御形態と異なり、データサンプリングによる経験則に基づいた制御方式として、強化学習を適用することで、目標値までの制定時間を短縮できる効果があることを示すことができた。ま



(a) reward result (b) level result (c) valve result

Figure 5: RL setting1 result. (a) : x axis means iteration count and y axis means each iteration's reward sum total. (b), (c) : x axis means step count and y axis means level or valve value(%) after 30 iteration's learned.



(a) reward result (b) level result (c) valve result

Figure 6: RL setting2 result. (a) : x axis means iteration count and y axis means each iteration's reward sum total. (b), (c) : x axis means step count and y axis means level or valve value(%) after 30 iteration's learned.

たシミュレーションではなく、現実設備の制御を行えたことは大きな進歩であると考えている。また学習の時間として約 4 時間で達成できたことも大きな成果になるといえる。知識伝承や人手不足などの社会的な課題がある中で、自動で学習し、制御できるようになることも強化学習を利用するうえでのアドバンテージになると考えられる。今後は実設備への応用例を増やしていきつつ、安全性や汎用性についても技術開発を進めていきたい。

謝辞

本研究にあたり、参考ソースコードを提供いただいた奈良先端大の崔允端博士に感謝申し上げます。

参考文献

[1] 城庵颯ら.,"機械学習を使用した KEK Linac 加速運転調整システムの開発", 第 16 回 日本加速器学会年会, 2019 pp.600-603.

[2] Y. Cui, T. Matsubara, and K. Sugimoto., "Kernel Dynamic Policy Programming: Practical Reinforcement Learning for High-dimensional Robots", Neural networks, vol. 94, 2017 pp. 13-23.

[3] L.Zhu, Y.Cui, *et al*., "Factorial kernel dynamic policy programming for vinyl acetate monomer plant model control", IEEE international conference on automation science and engineering 2018 pp.304-309.

[4] L.Zhu, Y.Cui *et al*., "Scalable reinforcement learning for plant-wide control of vinyl acetate monomer process", Control Engineering Practice, vol. 97 2020.

[5] 松原崇充，鹿子木宏明，"強化学習によるプラント自動最適化操業への試み ～ 酢酸ビニルモノマー製造プラントモデルへの適用 ～"，化学工学会会誌，Vol. 83，No. 4，2019，p. 1-3.

[6] 横河電機と NAIST が化学プラント向けに強化学習　少ない試行回数で高度な制御を実現，日経 Robotics 3 月号，No. 44，2019.

[7] `https://www.yokogawa.co.jp/solutions/products-platforms/control-system/real-time-os-controller/rtos-cpu/rtos-osfree-cpu/`

[8] E. Govinda Kumar, M.Sankar, "Detection of Oscillation in Three Tank Process for Interacting and Non-Interacting Cases", International Mutli-Conference on Automation, Computing, Communication, Control and Compressed Sensing (iMac4s) 2013 pp. 252-257.

[9] J.G. Ziegler and N.B. Nichols: "Optimum Settings for Automatic Controllers", Trans. ASME, 64-8, 759/768 1942.

[10] Volodymyr Mnih *et al*., "Playing Atari with Deep Reinforcement Learning", NIPS 2013.