深層強化学習とベイズ最適化による渋滞制御を行う搬送制御システム

Congestion Control on Conveyor Lines with Deep Reinforcement Learning and Bayesian Optimization

高　橋　健　吾　株式会社 IHI 物流産業システム　ロジスティクス BU プロジェクト部制御設計グループ
鹿　山　宏　之　株式会社 IHI 物流産業システム　ロジスティクス BU プロジェクト部制御設計グループ　主幹

物流システム内の搬送路におけるワークの渋滞制御は，その問題の特性から古典制御で扱うことが難しい．そこで本研究では，深層強化学習に，パラメータ最適化手法であるベイズ最適化を組み合わせることで渋滞制御問題の解決に取り組んだ．学習を終えた制御器は搬送路の渋滞制御に成功し，古典 PI (Proportional Integral) 制御の性能を上回った．設計者への依存が少ない本手法により，工数やリードタイムの削減，稼働設備のエネルギー効率向上などの付加価値をお客さまへ提供できるようになると期待される．

The characteristics of congestion control on conveyor lines cause difficulty in handling the control with classical control theories. In this study, we addressed it by combining deep reinforcement learning with Bayesian optimization, a method for optimizing parameters. The agent trained with our method successfully controlled the congestion on the conveyor line and outperformed the classical PI control. This method, which is less dependent on the designer, is expected to provide customers with added value such as reduction of person-hours and lead-time, and improvement in energy efficiency of their equipment.

1. 緒　　言

産業設備を稼働させるに当たり，1950 年代に体系化された古典制御は，PID (Proportional Integral Differential) 制御を中心に現在でも有力なアプローチの一つである．PID 制御とはフィードバック制御の一種であり，現在の出力値と目標値との差，その時間積分および時間微分によって入力値を決定する制御手法である．この手法は，パラメータの意味が明快で扱いやすい一方，それらの決定には制御設計者の経験や勘に基づく試行錯誤，問題に対する深い理解が求められる．また，ある種の問題に対しては PID 制御の適用自体が難しい場合もある．

物流システム内の搬送路におけるワークの渋滞制御は，そうした問題の一つである．搬送路における渋滞は，新たなワークの投入を阻む「ドロップ」と呼ばれる事象を引き起こす（詳細は **2. 1 節**にて述べる）．ドロップは搬送効率の低下などにつながるため回避すべきものであるが，これそのものに対して制御を行い抑制することは難しい．なぜなら，例えば後追い型の制御を用いる場合，ドロップが発生してからそれを防ぐ方向へ制御が働くため，原理的にドロップは避けられないからである．したがって，ドロップの要因である搬送路上の渋滞，すなわちワークの分布の仕方を制御する必要があるが，古典制御の枠組みではそうした分布を直接扱うことへの難しさも伴う．例えば，前述の PID 制御では現在の出力値と目標値との差を求める必要があるが，分布に対する差は簡単には定義できない．さらに，目標とする分布自体が必ずしも事前に分かっているとは限らない．

そこで本研究では，強化学習に深層学習を導入した深層強化学習という手法と，最適化手法の一つであるベイズ最適化を組み合わせ，人の手をほとんど借りない形での搬送路の制御の最適化に取り組んだ．深層強化学習において用いられるニューラルネットワークは，搬送路上の分布を直接扱うことができ，またベイズ最適化と組み合わせることで，設計者への依存が少ない制御ロジックの作成が可能となる．

株式会社 IHI 物流産業システムはこれまでにも，ロボットによるピースピッキング・詰め合わせ作業の自動化や，デパレタイズ（荷降ろし）システムへの画像認識 AI (Artificial Intelligence) の導入など，お客さま設備の高効率化や無人化・省人化に貢献する開発を行ってきた．本研究もそのような開発の一環であり，深層強化学習の特長を活かすことで工数やリードタイムの削減，これまでよりもエネルギー効率の高い設備稼働などの付加価値をお客さま

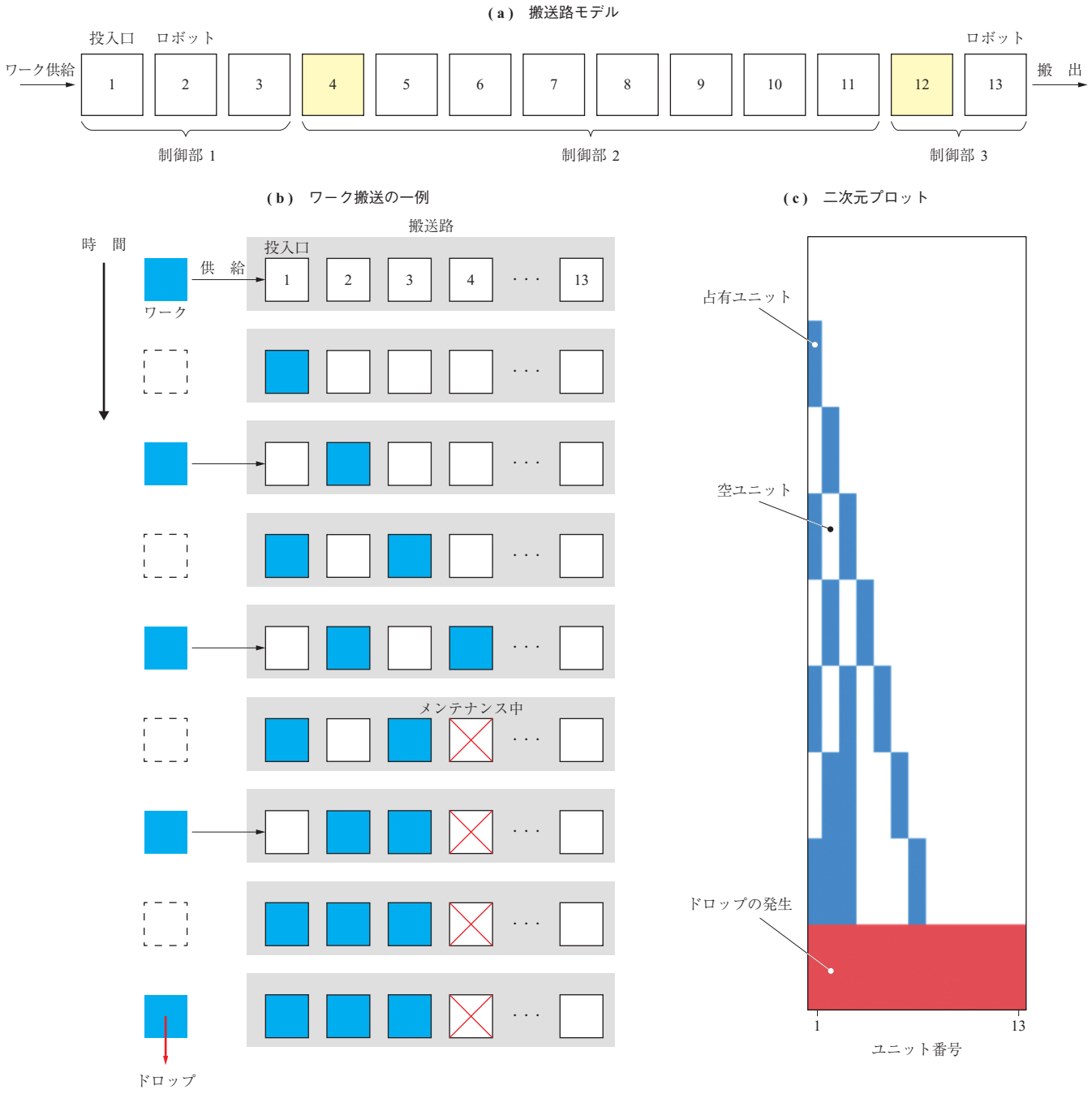
へ提供することを目指すものである．

2. 実 施 方 法

2.1 搬送路モデル

第 1 図に搬送路のモデルおよびワーク搬送の一例を示す．本研究では，**第 1 図 - (a)** に示すような搬送路をシミュレーション上で構成する．1 列に並べられた四角形（ユニット）はそれぞれワークの停止位置を表し，隣り合うユニットの中心間の距離は 1 m とする．ある周期 T (s) で投入口に供給されるワークは，ユニットからユニットへと順々に下流に搬送され，最下流においてロボットにより搬出される．複数のワークを同時に一つのユニットへ置くことはできないものとする．また，色付きのユニット 4 および 12 はそれぞれ，L_4，L_{12} 個のワークを搬送し終えるたびに 60 秒のカウントダウンを開始し，カウントダウン終了後でユニットが空になった時点にて，時間長さ M_4，M_{12} (s) のメンテナンス状態へ移行する．メンテナンス中のユニットにワークを搬送することはできない．

第 1 図 - (b) にはワーク搬送の時刻歴の一例を示した．ここで見られるように，あるユニットでメンテナンスが始



第 1 図 搬送路のモデルおよびワーク搬送の一例
Fig. 1 A model of conveyor line and an example of work transportation

まると，その上流側でワークの搬送が滞り，渋滞が発生する．もし渋滞が最上流に達してしまうと，投入口のユニットが占有されているので，新たなワークの供給を行えない．このような事象を本研究では「ドロップ」と呼ぶ．

第 1 図 - (b) のワーク搬送の時刻歴を，**第 1 図 - (c)** に示すような二次元プロットで表現することにする．横軸は搬送路のユニット番号，縦軸は上から下へ時間の進む方向に対応する．

各ユニットには 0 ～ v_{max} の範囲で動作速度 v (m/s) を指示することができる．あるユニットがワークを受け取り，それを一つ下流側のユニットへ搬送し終えるまでに掛かる時間 t_f (s) は v，およびユニットごとに定められた加速度 a (> 0) (m/s^2) を用いて (1) 式のように決まる．

$$t_f = \begin{cases} \dfrac{v}{a} + \dfrac{1}{v}, & v^2 < a \\ \dfrac{2}{\sqrt{a}}, & v^2 \geq a \end{cases} \quad \cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots (1)$$

本研究で用いるモデルでは，搬送路を三つの制御部に大別し（**第 1 図 - (a)**），同じ制御部に属するユニットは速度指示を共有する．すなわち制御部全体を制御するのに三つの指令速度を与えれば十分である．

ドロップを抑制するうえで最も単純な制御方策は，すべてのユニットを最大の搬送速度で動作させるというものであるが，その場合，搬送路上で渋滞が発生していないような状況でもユニットは最大速度で運転される．これはエネルギーの無駄であり，また必要以上に高速での搬送はワークへのダメージも懸念される．そこで本研究は，搬送路におけるドロップの発生回数を最小限に抑えながら，搬送速度の低減も同時に達成することを目的とする．

2. 2　深層強化学習

2. 2. 1　概　　要

ある環境内に置かれたエージェントを考える．エージェントは，環境の状態に基づいて行動を選択することができ，その行動の結果の良しあしに応じて，環境からは報酬と呼ばれる値が与えられる．強化学習とは，このような枠組みの問題において，エージェントが報酬の総和（収益）を最大化するためにどのように行動すればよいかを取り扱う機械学習の手法である．

その代表的なアルゴリズムとして Q 学習が挙げられる．Q 学習の目的は，環境の状態とエージェントの行動のすべての組合せについて，（最善な行動方策のもとでの）収益の期待値を求めることである．この手続きは，列と行にそれぞれ環境の状態，エージェントの行動をもつような期待値の表を作成することに相当する．そのような表をひとたび求めることができれば，問題は解けたも同然である．なぜなら，後は状態が与えられるごとに，その状態に対応する列をたどり，最も期待値の大きい行動を選び続けることが，最善の行動則となるからである．

しかしながら，この手法は環境の状態や行動の選択肢が多数あるような問題に対しては適用することが難しい．なぜなら，そのような問題を扱うためには，多くの列と行から成る表を作成する必要があるが，あまりに大きな表はコンピュータのメモリ空間に収まらないからである [1]（例えば囲碁の場合，可能な盤面の状態は 10^{172} に近いといわれており，もし仮に一つの盤面の状態を 1 バイトで表現することができたとしても，1 列の表を作成するだけで 10^{160} テラバイトもの記憶容量が必要となる）．また，状態や行動が連続的な数値で表されるような問題に対しても適用できない．

そこで近年，関数近似器としてニューラルネットワークを利用する方法 [2] が有力視されている．一般にニューラルネットワークは，単純な構成でも複雑な関数を近似できることが保証されている（普遍性定理）．この利点を活用し，状態と行動からその期待値を出力するような関数や，あるいは期待値を求める過程を省いて，環境の状態から直接最適な行動を出力するような関数を近似的に求める手法である．これならば表を作成することなく問題を扱えるため，膨大な記憶容量は不要であり，また連続値で表される状態や行動にも対応できる．さらに，関数を近似するニューラルネットワークは，誤差逆伝播法および汎用 GPU (Graphics Processing Unit) の利用により効率的に求めることができ，計算時間の観点からも Q 学習に比べて有利なことが多い．強化学習にニューラルネットワーク（深層学習）を取り入れた手法は，特に深層強化学習と呼ばれる．

2. 2. 2　物流搬送問題への適用

2. 2. 1 項で述べたように，ある問題を強化学習で扱うためには，環境とその状態，エージェントとその行動，報酬の計算方法を定義する必要がある．本研究ではこれらを以下のように定義する．

(1)　環境と状態

環境としては，**2. 1 節**で述べた搬送路モデルを用いる．搬送路モデルのパラメータを**第 1 表**に示す．また，環境の状態は以下の要素から成る 19 次元のベク

第 1 表　搬送路モデルのパラメータ
Table 1　Parameters of conveyor line model

パ　ラ　メ　ー　タ	単　位	値
ワークの供給周期 T	s	20
ユニット 4 のメンテナンス間隔 L_4	個	9 ～ 10[*1]
ユニット 12 のメンテナンス間隔 L_{12}	個	47 ～ 50[*1]
ユニット 4 のメンテナンス時間 M_4	s	10
ユニット 12 のメンテナンス時間 M_{12}	s	100
指示速度の最大値 v_{max}	m/s	0.3
加　速　度　a	m/s^2	（ユニット 1，3，8，11，12，13）16.7[*2] （上記以外）0.083 3

（注）*1 ：メンテナンスの間隔は，これらの範囲からランダムに決定される．
*2 ：ユニット 1，3，11，12，13 はロボット，メンテナンスの行われるユニット，搬出口のいずれかのユニットの直前にあるため，加速度を大きく設定する．またユニット 8 は何らかの動作が行われることを想定して，これも同じ加速度とする．

トルとして定義する．

- 搬送路のユニット 1 ～ 13 の在席フラグ
- ユニット 4，12 のカウントダウンの秒数
- ユニット 4，12 のメンテナンス経過時間
- ユニット 4，12 がメンテナンス中かどうかのフラグ

(2)　エージェントとその行動

本研究では，エージェントの最適化アルゴリズムとして PPO (Proximal Policy Optimization) [3] を採用する．この手法では，エージェントは価値推定ネットワーク（Critic ネットワーク）と方策ネットワーク（Actor ネットワーク）という二つのニューラルネットワークを内部に有し，これらを同時に最適化していく．

これらのネットワークは入力として前述の状態ベクトルを受け取り，出力として価値推定ネットワークは収益の推定値を（これは後にネットワークのパラメータ更新に用いられる），方策ネットワークは制御部 1 ～ 3 への三つの速度指示値を送出する．この速度指示値がエージェントから環境へ渡される行動に当たる．

(3)　報酬の計算方法

ある時刻 t においてワークを最下流まで搬送したかそうでないかに応じて 1 か 0 を取る変数を $x_{t,catch}$，ワークのドロップが発生したかそうでないかについての変数を $x_{t,drop}$，i 番目のユニットに与えられた速度指令を $v_{t,i}$ ($i = 1 \sim 13$) とする．時刻 t における報酬 r_t を (2) 式により定義する．

$$r_t = Ax_{t,catch} - Bx_{t,drop} - C\sum_{i=1}^{13} v_{t,i}^2 \quad \cdots\cdots\cdots\cdots (2)$$

式中の係数 A，B，C (> 0) はハイパーパラメータである．

報酬をこのように設計した理由は次のとおりである．(2) 式の第 1 項は，ワークを搬送することができるたびに与えられる正の報酬であり，これは本研究で作成する搬送路モデルが正しく搬送路としての役割を果たすために必要な項である．また，本研究はドロップの発生回数を最少にしながらも同時に動作速度（エネルギー消費）を抑える制御の獲得を目的としており，その達成のために，ドロップが発生するたびに負の報酬が与えられるように（第 2 項），また動作速度を大きくするほどに大きな負の報酬が与えられるよう設計した（第 3 項）．

2. 2. 3　学習の流れ

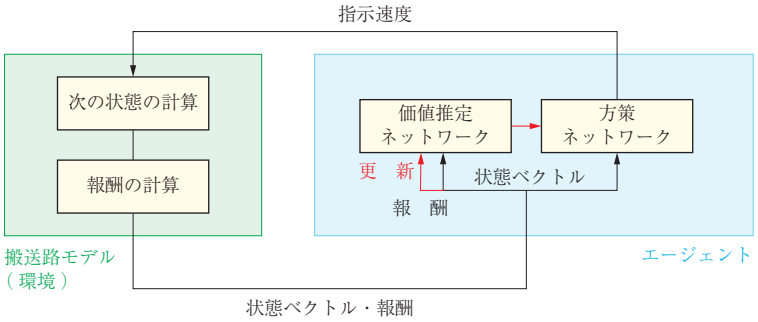
搬送路における深層強化学習の流れを**第 2 図**に示す．

まず初めに，エージェントのニューラルネットワークおよび搬送路モデルは適切に初期化される．次に，搬送路の初期状態がエージェントに対して与えられ，エージェントは受け取った情報からニューラルネットワークにより収益の推定値と速度指示値を計算する．これらのうち，速度指示値は行動として搬送路モデルへと渡される．搬送路モデルはその値に基づいて単位時間後の状態を計算し，さらにその状態の変化に伴う報酬を計算する．得られた状態と報酬はエージェントに返される．

このようなやり取りを一定回数繰り返すごとに，PPO アルゴリズムに従い，価値推定ネットワークおよび方策ネットワークのパラメータ更新が行われる．最適なネットワークが得られるまで以上の手続きを繰り返し行う．

2. 2. 4　評　　価

学習済みエージェントの評価は，搬送路モデルをシミュレーション上で 1 時間運転し，その際に発生したドロップの発生回数と，下記の (3) 式で定まる最大速度の平均



第 2 図　搬送路における深層強化学習の流れ
Fig. 2　Schematic diagram of deep reinforcement learning process on conveyor line

値 $\bar{u}$（以下，平均最大速度）により行われる．

$$\bar{u} = \frac{1}{13N}\sum_{j=1}^{N}\sum_{i=1}^{13} u_{i,j} \quad \cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots (3)$$

ここで N は 1 時間の運転中に供給されたワークの総数，また添え字 j はそれらワークを識別するものであり，シミュレーション開始から供給された順番に 1，2，3，…，N と割り振る．$u_{i,j}$ は i 番目のユニットをワーク j が通過したときの最大速度を表す．

ドロップの発生回数は少ないほどよく，同数であればより小さい平均最大速度で運転した制御器の方が優れている．

2. 3　ベイズ最適化

搬送路を適切に動作させるに当たり，(2) 式中の報酬パラメータ A，B，C を適切に設定することが必要である．例えば極端な例として，第 1 項と第 2 項が第 3 項に比べ非常に大きい場合，エージェントにとって速度をできるだけ抑えることで得られる報酬は微々たるものであり，常に最大速度を指示するよう学習するかもしれない．逆に，第 3 項が第 1 項，第 2 項を大きく上回る状況では，ワークを搬送することやドロップの発生回数を減らすことで得られる報酬よりも，速度を大きくすることで与えられる罰（負の報酬）の方が大きいために，エージェントはワークを搬送しないことを選択するかもしれない．

所望の動作を得るための A，B，C の値は自明でないため，さまざまな値を試す必要があるが，一般に深層強化学習に掛かる時間コストは大きく，なるべく少ない試行回数で良いパラメータを発見したい．

そこで本研究では，最適化手法の一つであるベイズ最適化を用いた．ベイズ最適化とは，形状の分からない関数の最大値（または最小値）を効率的に求めることができる手法であり，例えば一次元関数 $f(x)$ に対しては，以下のような反復計算によって最適化がなされる [4]．

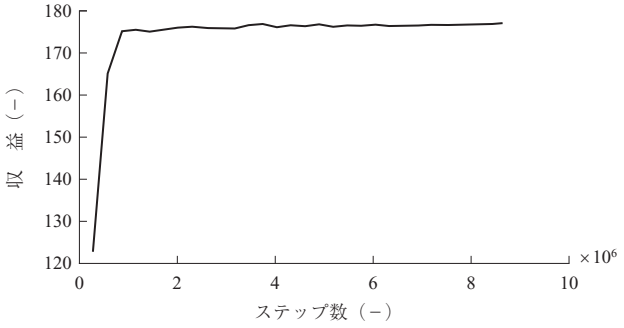
① 初めはランダムに x を決める．
② 前手順で決められた x に対して $f(x)$ の値を調べ，$(x, f(x))$ の組をデータとして保持する．
③ これまでに得られたデータから $f(x)$ の形状を予測するような統計モデルを作成する．
④ 統計モデルから次に調べる x を決定する．
⑤ 手順② に戻る．

本研究においては，x をパラメータ A，B，C に，関数 $f(x)$ を，「ある固定された A，B，C のもとで深層強化学習を行って得られたエージェントの性能」に置き換えて上記手順を実施することで，パラメータを決定した．

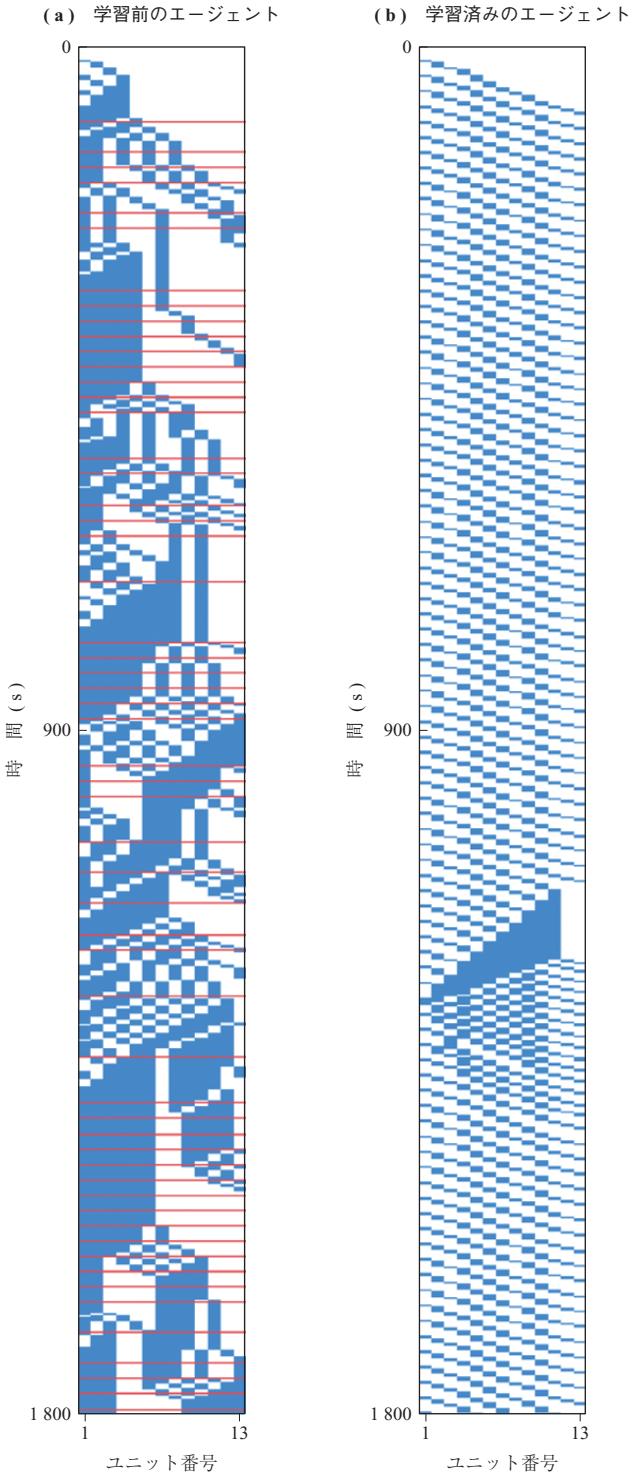
3.　結　　果

3. 1　エージェントの学習

典型的なエージェントの学習曲線を**第 3 図**に示す．学習のステップ数を進めるほどに収益は増加しており，エージェントの学習が安定して進展していることが分かる．**第 4 図**に学習前のエージェントと学習済みのエージェントによる搬送路制御の比較を示す．30 分間のワーク搬送の時刻歴を二次元プロットにより示している．学習前のエージェントによる搬送では，ワークの搬送が滞り，ドロップが多数発生している一方で，学習済みのエージェントで



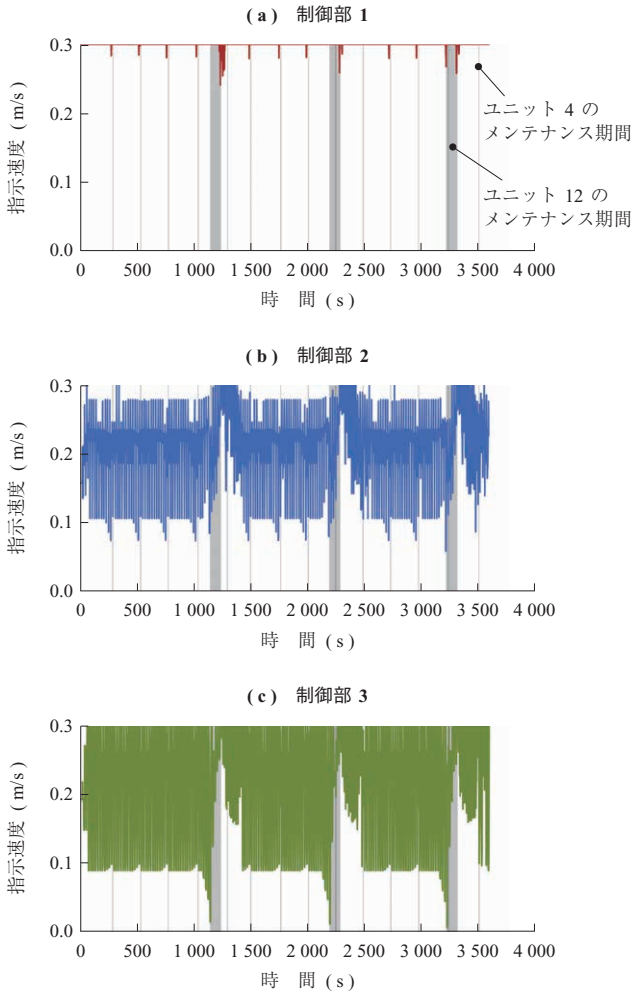
第 3 図　学習曲線
Fig. 3　Learning curve



第 4 図　学習前のエージェントと学習済みのエージェントによる搬送路制御の比較
Fig. 4　Comparison of conveyor line control by agent before learning and after learning

は，ワークはなめらかに搬送され，ドロップの発生回数はゼロに抑えられた．

第 5 図に学習済みエージェントによる指示速度の時間変化を示す．時間経過を 1 時間分プロットしたものである．**第 5 図 - (a) 〜 (c)** の 3 図はそれぞれ制御部 1 〜 3 に対応し，またグラフ中の灰色の領域はユニット 4 または 12 でメンテナンスが行われている時間帯を示す．この図から，エージェントが渋滞の発生しやすいメンテナンスの前後で指示速度を調整し，ワークのドロップを避けながら，効率的な搬送を実現できていることが分かる．

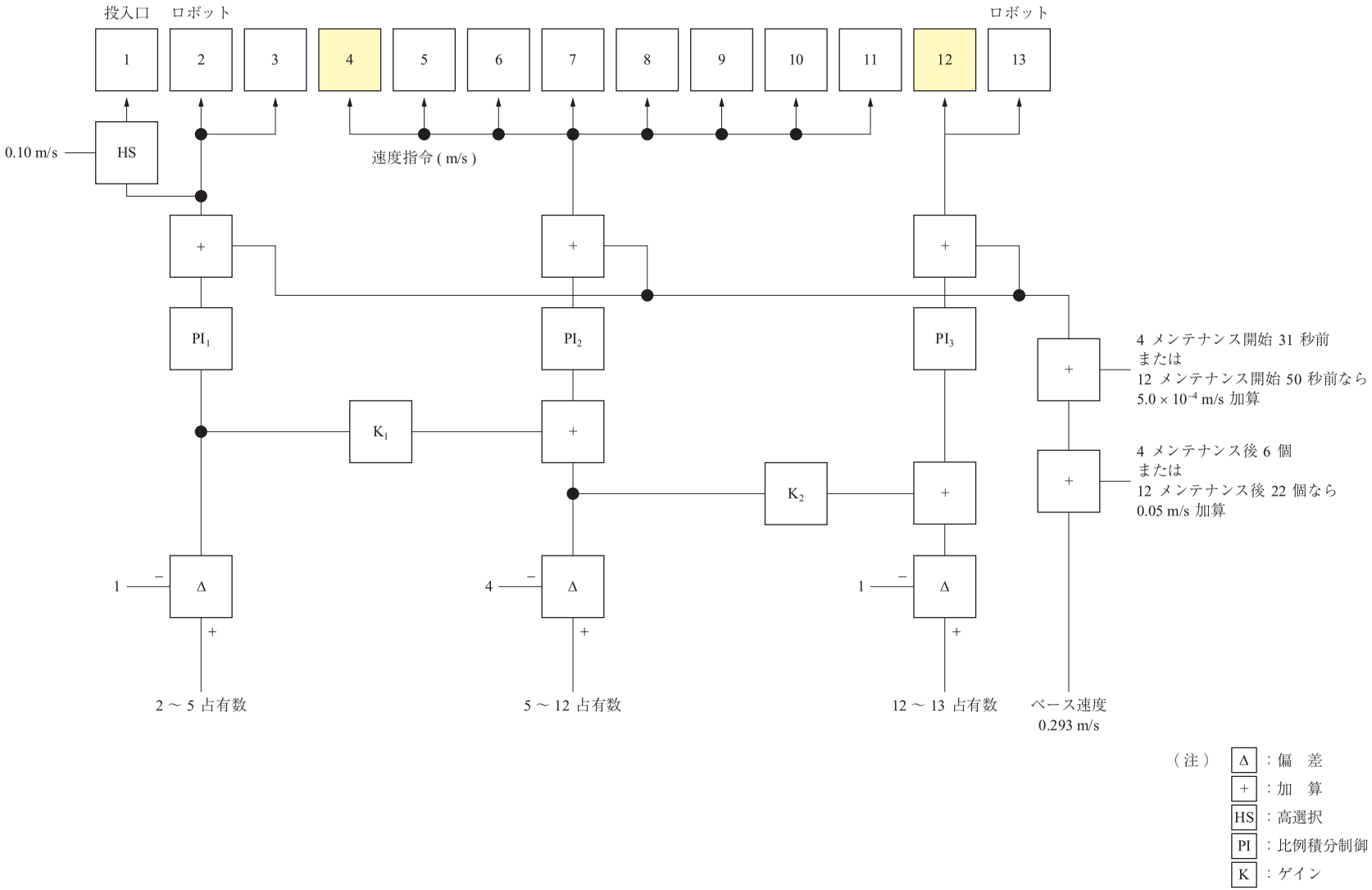


第 5 図　学習済みエージェントによる指示速度の時間変化
Fig. 5　Time-dependent change of speed order values given by trained agent

3. 2　PI 制御との比較

深層強化学習の性能を検討するため，PI 制御（PID 制御から時間微分を除いた制御手法）を用いた搬送路制御のシミュレーションを実施した．ここでは「搬送路の占有率が 50%を超えると渋滞が発生する」という渋滞学からの知見 [5] に基づき，占有率を制御量とするような PI 制御を構成した．搬送路における PI 制御のブロック線図を**第 6 図**に示す．

PI 制御においてもワークのドロップの発生回数をゼロに抑えることができたが，平均最大速度は 0.270 m/s となった．深層強化学習による学習済みエージェントでは 0.257 m/s となり，搬送速度の観点では深層強化学習が上回った．



第 6 図　搬送路における PI 制御のブロック線図
Fig. 6　Block diagram of PI control on conveyor line

また，**第 2 表**に学習時と異なる環境における本手法と PI 制御の性能の比較を示す．これは言わば，二つの制御器が未知の環境に対してどの程度対応できるかを調べたものである．深層強化学習による制御は PI 制御に比べ，平均最大速度を抑えながら，同時にドロップの平均発生回数も 4.5 倍小さくすることに成功した．この結果は，深層強化学習と PI 制御との，パラメータ変化に対する頑強性の差を示すものである．

4. 結　　言

古典制御で扱うことの難しい搬送路における渋滞制御問題について，深層強化学習とベイズ最適化を用い，ドロップの発生回数を最少に，かつ動作速度を最小にする制御ロジックの開発を行った．

深層強化学習のアルゴリズムとして PPO という手法を採用し，パラメータ調整にはベイズ最適化を活用することで，人手を排しながらもエージェントの学習を安定して進めることに成功した．学習済みエージェントによる搬送路シミュレーションではドロップの発生回数をゼロに抑えることができ，エネルギー効率も PI 制御の結果を上回るものとなった．また，深層強化学習によって得られた制御器は環境の変化に対してより頑強であることが分かった．これは同一ロジックの流用を行う際にパラメータの再調整が簡単であることを示唆するものである．

これらの結果から，お客さまへ工数やリードタイムの削減，エネルギー効率の優れた設備稼働といった付加価値を新たに提供できるようになると期待される．

本研究で用いた，深層強化学習とベイズ最適化を組み合わせた枠組みは搬送路以外へも適用可能であり，特にこれまで古典制御では扱いづらかった問題に対しても最適な制御ロジックを与えられるようになる．今後，本研究の実設備への早期実装を目指すと共に，深層強化学習・ベイズ最適化の適用先を広げ，お客さまの価値を最大化することに注力していく所存である．

第 2 表　学習時と異なる環境における本手法と PI 制御の性能の比較 *1

Table 2　Performance comparison between this method and PI control with different parameters from those used for training *1

	ドロップの平均発生回数	平均最大速度の平均値 (m/s)
本 手 法	0.06	0.257
PI 制 御	0.27	0.270

（注）*1：メンテナンス間隔 $L_4 = 6 \sim 10$，$L_{12} = 30 \sim 50$ である搬送路モデル上で，1 時間の運転シミュレーションを 100 回行った際の，ドロップの発生回数，平均最大速度の平均値の比較

── 謝　辞 ──

本研究を進めるに当たり，東京大学先端科学技術研究センターの西成活裕教授からご助言をいただきました．ここに記して謝意を表します．

参 考 文 献

(1)　中井悦司：IT エンジニアのための強化学習理論入門，技術評論社，2020 年

(2)　V. Mnih, K. Kavukcuoglu, D. Silver, A. A. Rusu, J. Veness, M. G. Bellemare, A. Graves, M. Riedmiller, A. K. Fidjeland, G. Ostrovski, S. Petersen, C. Beattie, A. Sadik, I. Antonoglou, H. King, D. Kumaran, D. Wierstra, S. Legg and D. Hassabis：Human-level control through deep reinforcement learning，Nature，Vol. 518，Iss. 7 540，(2015. 2)，pp. 529 － 533

(3)　J. Schulman, F. Wolski, P. Dhariwal, A. Radford and O. Klimov：Proximal Policy Optimization Algorithms，https://arxiv.org/abs/1707.06347，（参照 2021. 8. 23 ）

(4)　B. Shahriari, K. Swersky, Z. Wang, R. P. Adams and N. de Freitas：Taking the Human Out of the Loop: A Review of Bayesian Optimization，Proceedings of the IEEE，Vol. 104，Iss. 1，(2016. 1)，pp. 148 － 175

(5)　西成活裕：渋滞学，新潮社，2006 年

【ご案内】

IHI 技報をご覧頂きありがとうございます。
是非、関連する他の記事・論文もご一読ください。

IHI 技報 WEB サイト

IHI 技報（日本語）

IHI ENGINEERING REVIEW（英語）

Vol. 61　No. 3　特集　産業インフラの新しい価値の創出を目指して

◆特集　産業インフラの新しい価値の創出を目指して
- デザイン思考と本当のユーザーを意識した技術開発
- お客さまへの価値をデジタルで創造
- IHI エアロスペース　衛星打上げビジネスへ参入！
- EV 船へ向けた Z ペラ®　電気推進システムの開発
- 再生部品を利用した車両用ターボのリマニュファクチャリング
- 振動のモニタリングサービス
- 深層強化学習とベイズ最適化による渋滞制御を行う搬送制御システム

◆箸休め
- 土光氏の言葉に触れ，いまを考える

◆記事
- 進化する固体ロケットブースタ
- バイオマス発電所の営業運転開始から安定運転へ

Vol. 61 No. 3（2021 年 12 月）

インタビュー・特集外の記事も閲覧できます。

WEB サイトでは、社会と向き合い、社会とともに進化する IHI の技術・製品・サービスもご紹介しております。関連する技報も掲載しておりますので、ぜひご覧ください。

IHI 技報を通じて IHI グループのイノベーションを知る

IHI 製品を支える技術