計測自動制御学会 第 50 回知能システムシンポジウム資料，pp.126--131, 2023 年 3 月 28－29 日松江テルサ

強化学習を適用した配管自動設計システム

○稲岡諒朗　木村元　松葉佐哲仁（九州大学）

Automatic Pipeline Design System Using Reinforcement Learning

* A. Inaoka, H. Kimura, A. Matsubasa (Kyushu University)

Abstract－ Prior research has been conducted on automatic pipeline design system to support the piping design field in ship outfitting. In piping design, there is a problem that the final design plan differs depending on the order in which the pipes are placed. In this paper, we investigate the possibility of using the REINFORCE algorithm, one of the reinforcement learning methods, to make AI perform this order determination task, which is currently performed empirically by the designer.

Key Words: Automatic Pipeline Design System, Reinforcement Learning, REINFORCE Algorithm

1　緒言

船舶艤装の配管設計現場では、設計者への時間負担が大きい、若手への技術伝承が難しく人材不足が懸念されるという 2 つの課題がある。配管自動設計システムは設計作業の省力化支援を可能とするため、この課題に対する有望なアプローチである。先行研究 [1]では 1 本の配管経路探索問題を重み付きグラフ上の最小コスト経路探索問題に帰着させダイクストラ法を用いて解く手法や、複数配管問題について干渉を無視した配管を行った後に干渉部のコストを増加させて繰り返し配管を行うことで設計案を生成するタッチアンドクロス（TC）法を応用した手法が提案されている。しかしながら､複数配管問題には配管順決定問題が存在する。ここでの配管順とは、工事現場における施工順ではなく、設計における各配管の配置順のことを指す。これは、配管設計において配管をどの順番で配置するかで最終的な設計案が異なるという問題である。Fig. 1,Fig. 2 に自動設計での例を示す。

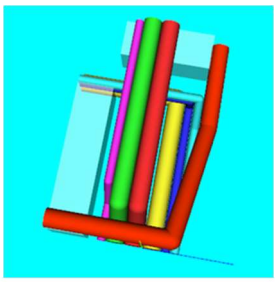

Fig. 1: Example in order 1

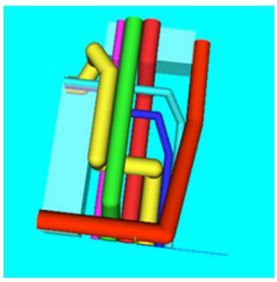

Fig. 2: Example in order 2

これらは同一の配管例題について異なる配管順で設計案を生成した例である。このように配管順をどのように決定するかで最終的に得られる設計案も大きく影響を受ける。現状では、設計に影響するこの配管順決定問題はベテランの経験に頼っており、設計自動化を妨げている。そこで本研究では強化学習手法の一つである REINFORCE（経験強化型強化学習、確率的勾配法など。以下勾配法と呼称）を用いた配管順決定手法を提案し解決を試みる。

2　AIによる配管順決定手法の提案

2.1 配管自動設計システムの概要

本研究で用いる配管自動設計システムの概要について説明する。システムは Java を用いて構成されている。システムの処理の流れは以下のようになっている。

① ユーザーが壁や通路等の障害物情報や配管に関する情報が記された xml ファイルを作成する。(設計案を得たい配管問題の定義)
② ユーザーが設計空間内で各配管が通れる経由候補点について記した xml ファイルを作成する。(経由候補点の定義)
③ システムが用意された xml ファイルを読み込み、設計案を生成する。
④ システムが生成した設計案を xml ファイル及び x3d ファイル（可視化された 3d モデル）にて出力。

今回は手順③にて、配管順決定に関係するクラス、メソッド群を改変・追加することにより検討を行った。

2.2 強化学習・勾配法について

強化学習は学習主体であるエージェントが環境内で自らが行動し、報酬を受け取ることでより良い行動を学習していく手法である。代表的な手法としてはQ-learning がある。この Q-learning は学習が収束すると決定論的に一つの最適な一手を選ぶようになるという特徴があり、近年は深層学習と組み合わせた DQN 等の手法に注目がされている。しかし、Q-learning は環境がマルコフ決定過程という数理モデルで表現されていないと学習の収束が保証されないという性質を持つ。

今回扱う配管順決定問題はモデル化した際に状態を完全に表現することは難しく、不完全な状態観測が予想される。したがって、そういった非マルコフな環境でも学習により行動改善が可能な強化学習手法の一つである勾配法を今回は用いることとした。勾配法（REINFORCE）はWilliams により提案された手法である [2]。学習が収束した際に Q-learning は決定論的に最適な一手を各状態で取るのに対し、勾配法は確率的に行動を選択する。学習によって行動選択確率が改善されるようになっている。

2.3 配管順決定問題のモデル化

本研究では次のFig. 3のように配管順決定問題をモデル化した。

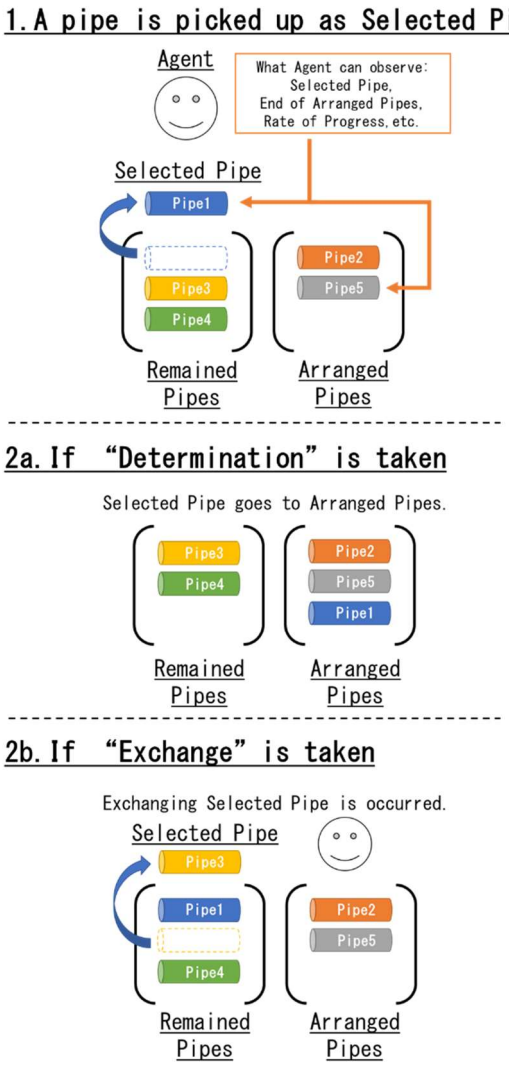

Fig. 3: Problem of determining an order of pipe placement

問題は『行動主体（Agent）』、『判断対象（Selected Pipe）』、『残り配管（Remained Pipes）』、『完了配管（Arranged Pipes）』の4つの要素からなり、行動主体は『決定行動（"Determination"）』、『交換行動（"Exchange"）』の2つの行動を取ることで、残り配管内の配管を完了配管へ移行させていく。最終的には完了配管内の配管の並びから配管順を得られるようになっている。

また、本研究では行動主体は判断対象、完了配管の最後尾、全体の進行度等の限られた範囲しか観測していないため、注意が必要である。

2.4 学習器について

ここからは構築した学習器モデルの説明を行う。状態は $\mathbb{S} = (x_1, x_2, \dots)$ とする。また学習されるパラメタを $\mathbb{W} = (w_1, w_2, \dots)$ とする。勾配法では、ある状態 $\mathbb{S}_t$ においてある行動 a_t が取られる確率は次の式(1)のように表現できると仮定する。

$$P(a_t) = f(\mathbb{W}, \mathbb{S}_t) \tag{1}$$

したがって、今回はこの行動選択確率関数 f を次の式のように定義している。

$$f(\mathbb{S}, \mathbb{W}, a) = \begin{cases} sigmoid(\mathbb{S}, \mathbb{W}) \ (a = a_1) \\ 1 - sigmoid(\mathbb{S}, \mathbb{W}) \ (a = a_2) \end{cases} \tag{2}$$

また状態 $\mathbb{S}_t$ で行動 a_t を取り、報酬 R_t を受取ったとき、パラメタは以下のように更新される。

$$w_i \leftarrow w_i + \alpha \Delta w_i \tag{3}$$

$$\Delta w_i = (R_t - b) \cdot Trace(t) \tag{4}$$

$$Trace(t) = \gamma \cdot Trace(t-1) + \frac{\partial}{\partial w_i} \log f(\mathbb{W}, \mathbb{S}_t) \tag{5}$$

α は学習率であり、b はベースラインと呼ばれる定数で各状態での報酬の平均値に近い値を設定するほど性能が向上する。$\frac{\partial}{\partial w_i} \log f(\mathbb{W}, \mathbb{S}_t)$ はeligibilityと呼ばれ、行動に対する適正な更新量を導くために導入される項である。$Trace(t)$ は時刻 t までのeligibilityの履歴であり、割引率 γ に1未満の値を設定することで未来の報酬を割り引いて更新がなされる。

エピソードは残り配管が初期化され、判断対象が1本取り出された状態から開始される。エージェントが行動を行い、全ての配管が完了配管へ移行される、または特定の条件を満たすと終了となる。今回は判断対象が完了配管へ移行されるたびに既存のシステムにその時点での完了配管の自動設計を行わせ、配管の経路が見つかったかどうかという成否を確認するようにしている。特定の条件について、この成否をもとにエピソードの終了条件を設定している。後述の各モデルの状態遷移にてまた言及する。

なお、エピソードの最初では一旦干渉を無視して配管の自動設計をシステムに行わせ、各配管の情報（直径、経路長など）を取得し、状態特徴量の計算を行えるようにしている。

今回はモデル1（試作版）、モデル2（改良版）の2種のモデルを作成した。各モデルで状態特徴量、行動選択確率関数、状態遷移、報酬が異なる。以下に各モデルの詳細を記す。

a. モデル1

- 状態特徴量

次ページのTable. 1を参照のこと。大まかに分類すると $x_1 \sim x_4$ が判断対象固有の特徴量であり、$x_5 \sim x_8$ が判断対象と完了配管の最後尾との相対量、$x_9 \sim x_{11}$ がその他特徴量となっている。

- 行動選択確率

学習パラメタは $\mathbb{W} = [w_1, w_2, \dots, w_{18}]$ の18個である。各行動の選択確率は次の式(6)~(8)で計算される。

Table. 1: Features of model 1

Feature	Explanation	Domain
x_1	Ratio of the diameter of Selected Pipe to the average one	$0 < x_1 < \infty$
x_2	Ratio of the route length of Selected Pipe to the average one	$0 < x_2 < \infty$
x_3	Ratio of the number of transit points of Selected Pipe to the average one	$0 \leq x_3 < \infty$
x_4	Cosine similarity between the vector of Selected Pipe and the average one	$-1 \leq x_4 \leq 1$
x_5	Relative amounts about x1 of the end of Arranged Pipes and Selected Pipe	$-\infty < x_5 < \infty$
x_6	Relative amounts about x2 of the end of Arranged Pipes and Selected Pipe	$-\infty < x_6 < \infty$
x_7	Relative amounts about x3 of the end of Arranged Pipes and Selected Pipe	$-\infty < x_7 < \infty$
x_8	Cosine similarity between the vector of Selected Pipe and the end of Arranged	$-1 \leq x_8 \leq 1$
x_9	Features on the number of consecutive exchanges	$0 \leq x_9 < \infty$
x_{10}	Rate of progress	$0 \leq x_{10} < 1$
x_{11}	First move or not	$x_{11} = (0,1)$

$$P(a) = f(\mathbb{S}, \mathbb{W}, a) = \begin{cases} sigmoid(\mathbb{W}\mathbb{x}) \ (a = a_1) \\ 1 - sigmoid(\mathbb{W}\mathbb{x}) \ (a = a_2) \end{cases} \quad (6)$$

$$\mathbb{W} = [w_1, w_2, \dots, w_{18}] \quad (7)$$

$$\mathbb{x} = [1, x_1, x_2, x_3, x_4, (x_{11} \cdot x_5), (x_{11} \cdot x_6), (x_{11} \cdot x_7), (x_{11} \cdot x_8), x_9, (x_{10} \cdot x_1), (x_{10} \cdot x_2), \dots, (x_{10} \cdot x_8)]^T \quad (8)$$

・ 状態遷移

モデル1の特徴は、判断対象が完了配管に移行された後の自動設計が失敗した場合は完了配管の最後尾の2本の配管を残り配管へ戻すことである（完了配管に1本しかない場合は1本戻す）。また、エピソード中に到達した進行度 x_{10} （完了配管の本数に依存）が更新されないまま10回決定行動が取られて、自動設計に失敗するとエピソード終了となる。

・ 報酬

決定行動の後、自動設計に成功した場合は以下のように与えられる。

$$reward = \left(10 + \left(-\frac{1}{10}r_1 + 1\right) + \left(-\frac{1}{10}r_2 + 1\right)\right) 10^{x_{10}} \quad (9)$$

r_1, r_2 はそれぞれ経路長と経由点数の観点から生成された設計案がどれほどシンプルかを表す指標となっていて、最良で 0 をとるようになっている。また、最後の 1 本を完了配管に移行して成功した場合は $reward$ に 1000 を加算する。

決定行動の後、自動設計に失敗した場合は以下のように与えられる。

$$reward = \left(-10 - (r_1 + r_2)\right) 1.1^{-x_{10}} \cdot n_{unupdate} \quad (10)$$

$n_{unupdate}$ というのは進行度 x_{10} （完了配管の本数に依存）が更新されていない状態で取られた決定行動の回数である。

連続交換回数の特徴量 x_9 が 2 以下で交換行動を取ったときは以下の通り。

$$reward = 0 \quad (11)$$

交換行動を残り配管の本数以上連続で行うと同じ判断対象を繰り返し観測していることになる。x_9 が 1 を超えると 1 周目を終えたことを意味する。

連続交換回数の特徴量 x_9 が 2 より大きいかつ交換行動を取ったときは以下の通り。

$$reward = -10 \quad (12)$$

何回も連続で交換行動を取るとエピソード長が余分に長くなる可能性があるので、負の報酬で抑制を行う。

b. **モデル2**

・ 状態特徴量

モデル2では、扱いやすいように状態特徴量の範囲を0~1に収めるようにした。また取り扱う配管の特徴も一部変更を加えている。Table. 2を参照のこと。

Table. 2: Features of model 2

Feature	Explanation	Domain
x_1	Feature related to the rank on the diameter of Selected Pipe	$0 \leq x_1 \leq 1$
x_2	Feature related to the rank on the route length of Selected Pipe	$0 \leq x_2 \leq 1$
x_3	Feature related to the rank on the number of bends of Selected Pipe	$0 \leq x_3 \leq 1$
x_4	Feature related to the rank on the number of interference of Selected Pipe	$0 \leq x_4 \leq 1$
x_5	Relative amounts about x1 of the end of Arranged Pipes and Selected Pipe	$0 \leq x_5 \leq 1$
x_6	Relative amounts about x2 of the end of Arranged Pipes and Selected Pipe	$0 \leq x_6 \leq 1$
x_7	Relative amounts about x3 of the end of Arranged Pipes and Selected Pipe	$0 \leq x_7 \leq 1$
x_8	Relative amounts about x4 of the end of Arranged Pipes and Selected Pipe	$0 \leq x_8 \leq 1$
x_9	Relative amounts about distance between centers of gravity of the end of Arranged	$0 \leq x_9 \leq 1$
x_{10}	Sine similarity between the vector of Selected Pipe and the end of Arranged	$0 \leq x_{10} \leq 1$
x_{11}	Rate of progress	$0 \leq x_{11} < 1$
x_{12}	First move or not	$x_{12} = (0,1)$

・ 行動選択確率

関数 f は式(6)と同じだがパラメタ $\mathbb{W}$ 及び $\mathbb{x}$ が異なる。

$$\mathbb{W} = [w_1, w_2, \dots, w_{26}] \quad (13)$$

$$\mathbb{x} = [(1-x_{12}), x_1(1-x_{12}), x_2(1-x_{12}), \dots, x_{10}(1-x_{12}), x_1x_{11}(1-x_{12}), x_2x_{11}(1-x_{12}), \dots, x_{10}x_{11}(1-x_{12}), x_{12}, x_1x_{12}, x_2x_{12}, x_3x_{12}, x_4x_{12}]^T \quad (14)$$

・ 状態遷移

モデル1では、決定行動後の自動設計が失敗した場合巻き戻るような状態遷移を行っていたが、モデル2では失敗した場合即エピソードを終了するようにした。

・ 報酬

決定行動後、自動設計に成功した場合は以下の通り。

$$reward = \frac{1}{C} \quad (15)$$

C は既存の自動設計システムで定義されているコストという経路長やベンド数等と連動した設計案の品質を表す指標をもとに決まる値である。設計案の配管の経路長が長かったり、ベンド数が多かったりすると報酬が小さくなるようにしている。式(15)は最大で 1、最小で 0 である。

決定行動後、自動設計に失敗した場合は以下の通り。

$$reward = \left(\frac{0.5}{r_1} - 0.5\right) + \left(\frac{0.5}{r_2} - 0.5\right) \quad (16)$$

r_1, r_2はモデル 1 と異なるため注意。それぞれ配管の体積とベンド数に関する指標である。失敗時は制約上コストを計算出来ないため、代替で用いている。報酬の考え方は成功時と同様に設計案の品質が反映されることを期待して定義している。式(16)は最大で 0、最小で-1 となる。

連続交換行動が 1000 周を下回るときに交換行動を取った場合は式(11)と同様に報酬は 0 とした。1000 周というのはエピソード全体で共有される。モデル 1 よりも連続交換の制限を緩和している。

連続交換行動が 1000 周を超える場合に交換行動を取ると以下の負の報酬が与えられる。

$$reward = -0.01 \quad (17)$$

3 学習結果

以上の学習器モデル 1,2 を学習させた結果を示す。また学習は 1 種の例題を用いて行った。

3.1 成功率について

学習中の各エピソードでの獲得報酬から配管順を最後まで決定でき、最終的に有効な設計案が得られた割合を成功率として結果を示す。

a. モデル 1

学習エピソード-エピソードあたりの獲得報酬についての 1000 エピソード移動平均グラフを Fig. 4 に示す。初期では成功率が 47.3%であるのに対し、末期では 88.5%まで上昇している。

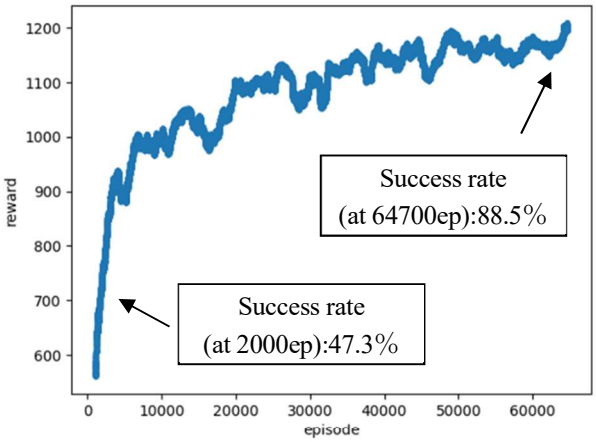

Fig. 4: Episode-Reward moving average of model 1

b. モデル 2

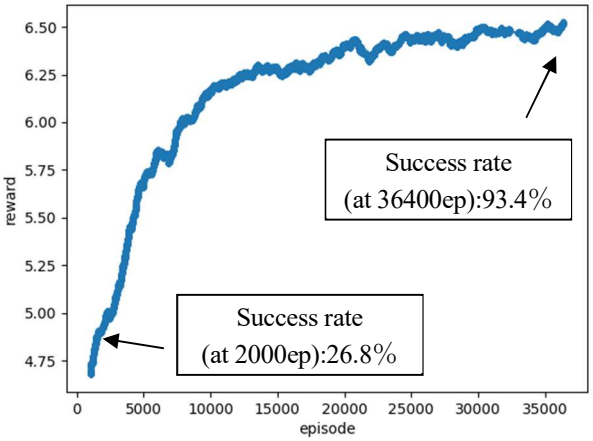

Fig. 5: Episode-Reward moving average of model 2

モデル 1 と同様に Fig. 5 に結果を示す。こちらも成功率は初期の 26.8%から末期では 93.4%まで上昇している。

3.2 エピソード長について

学習中のエピソード長についての変化を示す。

a. モデル 1

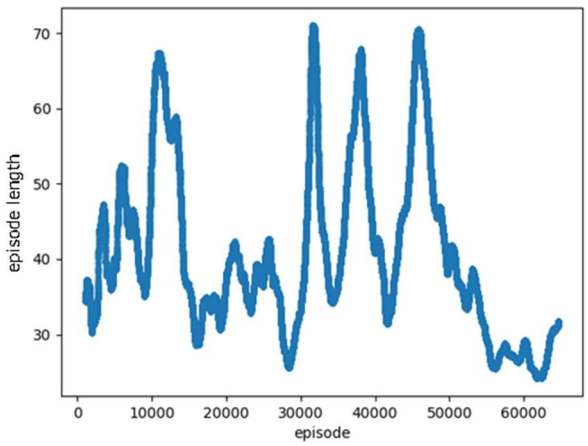

Fig. 6: Episode-Episode length moving average of model 1

学習エピソード-エピソードあたりの長さについての 1000 エピソード移動平均グラフを Fig. 6 に示す。エピソード長は最大で 70 ほどの幅で収まっている。

b. モデル 2

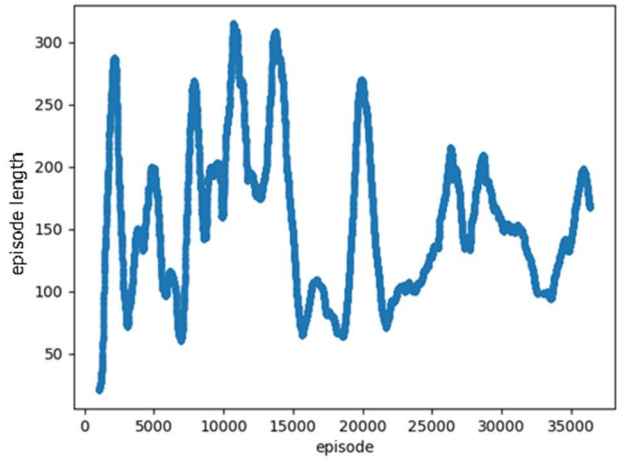

Fig. 7: Episode-Episode length moving average of model 2

モデル 1 と同様に Fig. 7 に結果を示す。こちらも同様にエピソード長は最大で 320 ほどの幅に収まっている。

参考のため、冗長な連続交換行動の抑制のための負の報酬を与えずに学習させた結果を Fig. 8 に示す。

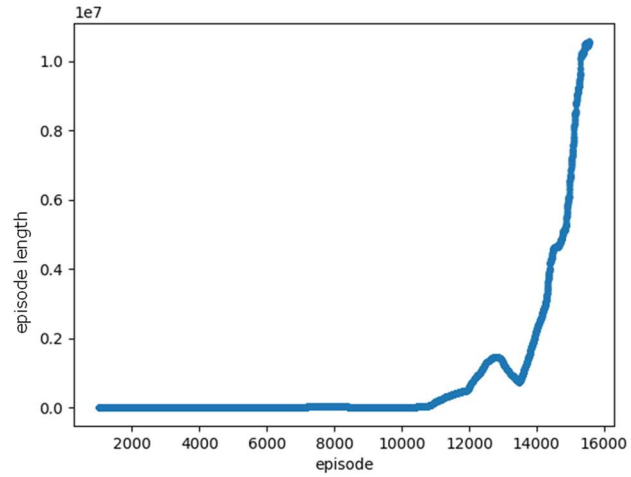

Fig. 8: Example of no penalty for consecutive exchanges

このように学習の末期では初期に比べエピソード長が 10^7 倍ほどになっている。したがって、適切な連続行動の抑制が必要である。

4 評価実験

以上の学習させたモデル 1,2 について性能の評価実験を行った。学習例題を含む 3 種の例題について、自動設計手法（TC 法と normal な手法）と配管順決定手法（モデル 1,2 と従来手法 2 種）の各条件下において実験を行った。Fig. 9,Fig. 10,Fig. 11 に例を示す。各試行は 100 回ずつ行った。評価は成功率と成功時のコスト平均(設計案の品質)の 2 つの観点から行っている。Table. 3,Table. 4 に結果を示す。

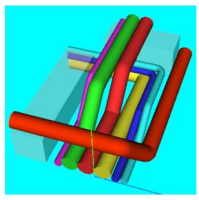

Fig. 9: Case 1 (learned)

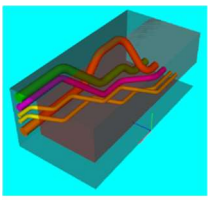

Fig. 10: Case 2

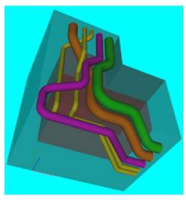

Fig. 11: Case 3

Table. 3: Success rate

		Case 1	Case 2	Case 3	Success rate(%)
Model 2	normal	94	0	0	41.0
	TC	89	43	20	
Model 1	normal	74	0	0	33.7
	TC	93	20	15	
Random sort	normal	28	0	2	22.2
	TC	68	16	19	
Cost sort	normal	100	0	0	31.2
	TC	67	20	0	

Table. 4: Average cost of success

		Case 1	Case 2	Case 3
Model 2	normal	986515		
	TC	1058642	594652	5437442
Model 1	normal	989617		
	TC	1081467	587986	544970
Random sort	normal	1011028		527468
	TC	1058110	596617	24376442
Cost sort	normal	921717		
	TC	1056790	598220	

行に配管順決定手法、列に例題を並べている。列について赤字が主席、青字が次席の成績を表している。

成功率の観点からは Table. 3 を見ると、配管順決定手法ごとに全てのシチュエーションを合算した場合の成功率はモデル 2 が 41.0%とやや高かったため多少の汎用性を獲得していることが示唆された。しかし個別に見ると例題 1（Case 1）では従来手法のコスト順（Cost sort）の方が提案手法より良い成績であり、また例題 3（Case 3）では従来手法と有意な差が見られなかった。

設計案の品質の観点からは Table. 4 を見ると、提案手法は従来手法に対して優位性は認められなかった。

5 考察

5.1 例題毎の学習

3.1 節の Fig. 4,Fig. 5 での学習初期から末期への成功率の上昇や 4 章の Table. 3 でのモデル 1,2 のランダムソートに対する成功率の優位性から、今回作成した学習器は例題毎には成功しやすい配管順を学習することが分かった。

5.2 重視される特徴量

モデル 2 の学習済パラメタについて分析し、どの特徴量が配管順決定問題において重視されているかを考

察する。まず、最初の 1 本を完了配管へ移す際(以下初手と呼称)の行動確率に影響を与えている特徴量を見る。初手の行動確率は次の Table. 5 にあるパラメタと x の積の和が式(6)の $\mathbb{W}\mathbb{x}$ に入力されて計算される。

Table. 5: Parameters used in the first move

Parameta	Result of learning	x	Related
w_{23}	-5.067906984	x_1x_{12}	Diameter
w_{24}	-3.88894835	x_2x_{12}	Route length
w_{26}	-3.58914133	x_4x_{12}	Number of interference
w_{22}	-0.866940911	x_{12}	(Bias)
w_{25}	-0.152338889	x_3x_{12}	Number of bends

絶対値の大きさから、初手では直径（Diameter）や経路長（Route length）の特徴量が重視されている。エージェントの初手での方針は直径が大きいもの、経路長が長いものを避けるということになる。

次に初手以降の行動確率に影響を与えている状態特徴量について見る。初手以降で式(6)の $\mathbb{W}\mathbb{x}$ に入力されているのは以下の各式とバイアス項の和である（以下バイアス項は省略）。

$$\begin{matrix}(w_{12}x_{11}+w_2)x_1(1-x_{12})\\ \vdots \\ (w_{21}x_{11}+w_{11})x_{10}(1-x_{12})\end{matrix} \tag{18}$$

この形式で整理されたパラメタの表が Table. 6 である。

Table. 6: Parameters used after the first move

Coefficient of x	Result of learning		x	Related
$w_{12}x_{11}, w_2$	1.376804544	-5.430115912	$x_1(1-x_{12})$	Diameter
$w_{13}x_{11}, w_3$	0.732070727	-5.344375882	$x_2(1-x_{12})$	Route length
$w_{14}x_{11}, w_4$	4.131766525	2.02951501	$x_3(1-x_{12})$	Number of bends
$w_{15}x_{11}, w_5$	0.522330178	-5.858365592	$x_4(1-x_{12})$	Number of interference
$w_{16}x_{11}, w_6$	2.308283243	2.393934183	$x_5(1-x_{12})$	Relative amounts about diameter
$w_{17}x_{11}, w_7$	1.513081022	1.747820264	$x_6(1-x_{12})$	Relative amounts about route length
$w_{18}x_{11}, w_8$	0.760428926	-3.599611204	$x_7(1-x_{12})$	Relative amounts about num of bends
$w_{19}x_{11}, w_9$	2.060989128	2.31719798	$x_8(1-x_{12})$	Relative amounts about num of interference
$w_{20}x_{11}, w_{10}$	0.520499967	0.223748185	$x_9(1-x_{12})$	Relative amounts about center of gravity
$w_{21}x_{11}, w_{11}$	2.567853726	2.420949492	$x_{10}(1-x_{12})$	Relative amounts about sine similarity

またこれを進行度 x_{11} を横軸として可視化したグラフが Fig. 12 である。Table. 6 と Fig. 12 から、まず傾きが全て正であるから、進行するにつれ決定行動自体が取られやすくなることが分かる。また絶対値の大きさから、初手以降では干渉回数（Number of interference）や経路長の特徴量が重視されていることが分かった。最も傾きが大きかったベンド数の特徴量についても、エピソード後半では重視されていた。初手以降の行動方針としては、基本的には干渉回数が多い、経路長が長い配管を前半では避けるようにし、後半ではベンド数が多いものを選ぶようになった。

ここまで考察したが、特に初手で直径が大きい配管を避ける方針は、一般的な配管設計ノウハウと逆だっ

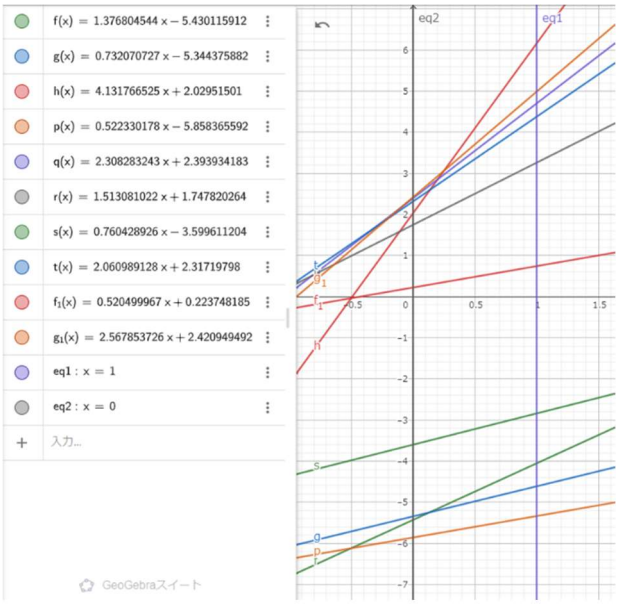

Fig. 12: Visualized parameters

たことから、学習例題が 1 種類だったことによるオーバーフィッティングの可能性がある。検討課題としてより多くの例題で学習させた汎用的な学習器について再度重視された特徴量を分析するのが適当と考える。

5.3 勾配法の有効性

学習が収束していると考えられるモデル 2(36400 エピソード)において、初手で計算される各配管の決定行動の確率から各配管の選択割合を調査した結果を Table. 7 に示す。例題 1 を用いた。

Table. 7: Percentage of selection for each pipe at the first move

	Pipe 1	Pipe 2	Pipe 3	Pipe 4	Pipe 5	Pipe 6	Pipe 7
Percentage	6.704	1.125	0.138	0.764	0.05	91.212	0.007

これより、パイプ 6 が 91.2%ほどで選ばれるが他のパイプも選ばれることが分かる。これは確率的な行動選択が行われていると考えられるので、今回デザインした環境は状態観測が不完全という非マルコフな環境であり、勾配法が有効であったと考えられる。

6 結言

本研究で提案した強化学習を適用した配管順決定手法は成功率の観点では従来手法に対し多少の汎用性が見られた。例題毎については成功率の観点でより良い配管順を学習していた。また学習された方針では最初の 1 本は直径、それ以降は干渉回数の特徴量等が重視されていたが、オーバーフィッティングの可能性があり今後の検討課題である。勾配法についても、学習結果から間接的に有効であったと考えられる。

参考文献

1) 木村元：パイプサポートや曲がり船殻に対応した配管自動設計に関する研究, 日本船舶海洋工学会講演会論文集, 22 号, 371/376 (2016)
2) Ronald J. Williams：Simple statistical gradient-following algorithms for connectionist reinforcement learning, Machine Learning, vol.8, 229/256 (1992)