情報処理学会論文誌 Vol.60 No.9 1572–1579 (Sep. 2019)

環境変化をともなう経路選択問題における強化学習

大滝 啓介[1,a)] 西 智樹[1] 吉村 貴克[1]

受付日 2018年10月2日, 採録日 2019年6月11日

概要：強化学習は，ある環境においてエージェントが取るべき行動を経験から学習する手法であり，行動は特定の環境から得られる経験を用いて学習される．そのため環境自体が変化した場合には，新しい環境に対して一から，または以前学習した結果を再利用し，行動の修正が必要な状態に対して学習をやり直す必要がある．我々は経路選択問題において，目的地までの距離の変化に基づいて，再学習が必要となる状態を絞り込むことで学習を高速化する手法を提案する．本稿では格子世界を用いた実験を行い，環境変化の構造的情報を利用することで，再学習が効率的に進むことを確認した．

キーワード：強化学習，環境変化，経路計画問題，重み付きサンプリング

Reinforcement Learning in Routing Problems with Environment Shifts

Keisuke Otaki[1,a)] Tomoki Nishi[1] Takayoshi Yoshimura[1]

Received: October 2, 2018, Accepted: June 11, 2019

Abstract: Reinforcement learning involves learning a policy. The learned policy must be adjusted when the environment shifts from a source domain to another domain. Typical approaches use learned parameters of the policy as initial parameters. We propose to use knowledge of the shifts additionally to adjust the policy. The knowledge is represented by weights on states representing the degree of changes in distances from the states to an absorbing goal. Our method uses these weights to sample states, wherein an agent updates the policy. Numerical experiments on Gridworlds indicate that the knowledge about the shifts is helpful for efficient learning, particularly at an early stage.

Keywords: reinforcement learning, environment shifts, gridworld, routing problem, weighted sampling

1. 背景

強化学習は，与えられた環境において，エージェントが環境とやり取りする際に得られる報酬から，取るべき行動を学習する枠組みである．近年では特に（深層）強化学習の研究が注目されている [1], [2], [3]．様々な状態や行動，遷移/報酬モデルを定義した上で環境を構築することで，ゲームだけではなく材料探索 [4] や経路計画 [5], [6] など，多様な意思決定問題を扱うことができる．

強化学習では，試行錯誤を通してマルコフ決定過程（Markov Decision Process; MDP）で記述される環境における状態価値や政策を学習する．そのため環境自体が変化した場合には，学習済みの情報を，新しい環境に適用できるように更新する必要がある．学習済みの情報を更新するために取りうる手法として，

- 学習を再度一から実行する手法（再学習）
- 過去に学習した知識を利用する手法（転移学習）

が考えられる．前者は，環境が変化する度に学習が必要となる．強化学習が多くの学習リソースを必要とすることを考えると望ましくないため，後者のように既知の情報を再利用する転移学習に注目が集まっている [7], [8]．例として，環境を簡潔に表現する特徴表現を用いた手法が注目されており，近年では Successor Features を用いる手法が研究されており，「目的地が移動する」という変化に追従し，高い性能を示すことが知られている [9], [10]．

環境の変化を大別すると，報酬モデルが変化する場合と，遷移モデルが変化する場合に分けられる．先に述べた SFs

[1] 株式会社豊田中央研究所
Toyota Central R&D Labs., Inc., Bunkyo, Tokyo 112–0004, Japan
[a)] otaki@mosk.tytlabs.co.jp

を用いる手法は，報酬モデルが変化する場合を想定している．一方で，遷移モデル自体が変化する問題は研究例が多くない [11], [12]．たとえば事故や道路保全に起因する通行止めなどの問題は，報酬モデルだけではなく遷移モデル自体も変化するため，これに対応できる手法が望まれる．

本研究では，たとえば交通事情（事故の有無など）が変化する場合においても，適切な経路選択を行えるような意思決定技術を目指して「経路選択問題」を扱う．我々の最終的な目標は，時々刻々と変化する環境において強化学習を運用し，環境の変化に追従可能な意思決定手法を構築することである．本稿ではその一歩として，移動コストが静的に与えられる通常の経路選択問題を仮定し，ある事象の前後で遷移モデルが変化する状況に注目する．強化学習の枠組みにより，時間に依存するコストや，移動や滞在に報酬が与えられる場合の空間経路計画など [13], [14]，背後にマルコフ決定過程を仮定したうえでの学習として，様々な応用問題に対しても同様に議論できる．

上記の前提の下で本稿では，学習済みの状態行動価値を再利用し，新しい遷移モデル（たとえば交通事故が反映されたモデル）が得られる環境に適応する手法を提案する．基本的なアイデアは，状態の集合を「政策が再利用できる可能性の高い状態」の集合と，「再度学習が必要になる可能性の高い状態」の集合に区別し，後者の集合に対して集中して経験を生成して効率的に学習を行うものである．その際，単に誤差が大きいところから再学習するのではなく，ある状態の誤差が残っているが，周辺の誤差は小さくなっているような状態を選択して徐々に学習を進めることで，徐々に誤差が 0 に収束した範囲を増やすように学習する．状態を二分する際に，本稿では経路選択問題において重要な特徴量である目的地までの距離と，その変化を利用する．本手法の有効性を示すため，通行止めが発生するような環境変化型の経路選択問題を想定した数値実験により，従来法との比較結果を示し，再学習に関して議論と考察を行う．

本稿の構成は以下のようになっている．我々の対象とする問題ドメインである「経路選択」と，そのドメインの変化について 2 章で議論する．我々の前提に基づく手法を 3 章で提案し，4 章では数値実験によって提案手法を評価する．最後に 5 章で本稿をまとめ，結論や今後の展望を述べる．

2. 経路選択問題と環境変化

本章では以降の章で利用する概念についてまとめる．本稿で利用する代表的な記号を**表 1** に示す．

2.1 マルコフ決定過程と Q 学習

マルコフ決定過程（MDP）を $\langle \mathcal{S}, \mathcal{A}, T, R \rangle$ で表す．ここで $\mathcal{S}$ は状態集合，$\mathcal{A}$ は行動集合であり，$T : \mathcal{S} \times \mathcal{A} \times \mathcal{S} \to [0, 1]$ と $R : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$ はそれぞれ遷移モデルと報酬モデルである．エージェントは状態 s_t において，政策 $\pi : \mathcal{S} \to \mathcal{A}$ に従い $a_t = \pi(s_t)$ を実行し，結果として報酬 $r_t = R(s_t, a_t)$ を得て，環境は状態 s_t から s_{t+1} へと遷移する．強化学習の目的は，T や R が未知の場合に，累積報酬 $R_t = \sum_k \gamma^k r_{k+t}$ を最大化する政策 π を求めることである．時刻 t，状態 $s_t = s$ において行動 $a_t = a$ を選択した際の累積報酬の期待値 $Q(s, a) = \mathbb{E}[R_t \mid s_t = s, a_t = a]$ を**状態行動価値**と呼ぶ．

表 1 本文中の記号一覧
Table 1 Notations.

記号	意味
$X \triangle Y$	集合 X と Y の対称差（$X \cup Y \setminus X \cap Y$）
$[p], p \in \mathbb{N}$	集合 $\{1, 2, \cdots, p\}$
$\langle \mathcal{S}, \mathcal{A}, T, R \rangle$	マルコフ決定過程
$V(s)$	状態価値（$s \in \mathcal{S}$）．$V^\star$ は真値を示す．
$\bar{V}_{\mathrm{error}}(s)$	サンプルベースの TD 誤差
$Q(s, a)$	状態行動価値（$s \in \mathcal{S}$，$a \in \mathcal{A}$）
M	サイズ $n \times m$ の格子世界，格子情報行列（0：目的地，1/2：実行可能/不可能状態）
$\alpha, \gamma \in \mathbb{R}$	Q 学習の学習率と報酬の割引率
$\mathcal{S}_M^{(v)}$	M 上で格子状態が v の状態集合
$\mathcal{M}^k$	M からたかだか k カ所変化した格子世界の集合
$\mathcal{F}_{M,M'}$	環境変化 $M \to M'$ のフロンティア
$\mathcal{N}_M(s)$	状態 $s \in \mathcal{S}$ の格子世界 M における近傍
$\mathcal{H}^{(k)}(F)$	状態集合 $F \subseteq \mathcal{S}$ の k 回拡大集合
$\mathcal{SH}^{(k)}(F)$	状態集合 $F \subseteq \mathcal{S}$ の k 回拡大重み行列

価値ベースの強化学習アルゴリズムは，Q 値を用いて $\pi(s) = \arg\max_{a \in \mathcal{A}} Q(s, a)$ を構築する．代表的な学習アルゴリズムである Q 学習では，状態 s_t において行動 a_t を実行し，即時報酬 r_t を受け取って状態 s_{t+1} に遷移した際に，誤差 $\mathcal{E} = r_t + \gamma \max_{a' \in \mathcal{A}} Q(s_{t+1}, a') - Q(s_t, a_t)$ を用いて $Q(s_t, a_t) \leftarrow Q(s_t, a_t) + \alpha \mathcal{E}$ と状態行動価値を更新する．また $V(s) = \max_{a \in \mathcal{A}} Q(s, a)$ で定義される値は**状態価値**と呼ばれ，学習済みの状態価値 $V^\star$ について**ベルマン方程式**によって以下が成り立つ（詳細は [15], [16] を参照）．

$$V^\star(s) = \max_{a \in \mathcal{A}} \left[R(s, a) + \sum_{s' \in \mathcal{S}} T(s, a, s') V^\star(s') \right] \quad (1)$$

ここで $V^\star(s)$ は，期待報酬と見なすことができる．ある学習中の状態価値 $V(s)$ について，式 (1) の左辺と右辺の間には誤差がある．またある状態 s から報酬 r を受け取って s' に遷移するというサンプルに対して，差分 $r + \gamma V(s') - V(s)$ は **TD** 誤差と呼ばれ，しばしば学習に用いられる．たとえば Q 学習では，TD 誤差を小さくするように勾配を用いて学習することで，期待している政策 π や関数 Q を求める．

2.2 経路選択問題

経路選択問題は，ある位置 $s_0 \in \mathcal{S}$ から目的地 $G \in \mathcal{S}$ まで到達するための行動に関する意思決定問題であり，MDP

によって定義される．本稿では簡潔な記述のために格子世界を利用した経路選択問題を用いるが，以下の議論はグラフ構造 $\mathcal{G}$ 上についても同様に展開される．格子世界が $n \times m$ の大きさであるとき，状態集合はエージェントの位置の集合 $\mathcal{S} = \{(x, y) \mid x \in [n], y \in [m]\}$ であり，唯一の目的地 $G = (x_G, y_G) \in \mathcal{S}$ を仮定する．行動集合 $\mathcal{A}$ は格子世界上の上下左右への移動に対応する．状態 $s \in \mathcal{S}$ において，行動 $a \in \mathcal{A}$ を実行した場合に，a に対応した位置へ移動可能であればつねに移動し，そうでなければ移動しない．本稿では侵入可能/不可能な位置のことを，エージェントにとって実行可能/不可能な状態であると呼ぶ．

実行可能・不可能な状態を形式的に表すため，格子世界と同じ記号を用いた 3 値行列 $M \in \{0, 1, 2\}^{n \times m}$ を定義する．位置 $(x, y) \in \mathcal{S}$ について，$M(x, y) = 0$ であるとき $(x, y) = G$ とする．また $M(x, y) = 1$ および $M(x, y) = 2$ であるとき，状態 (x, y) はそれぞれ実行可能・実行不可能であるとし，集合 $\mathcal{S}_M^{(v)} = \{(x, y) \in \mathcal{S} \mid M(x, y) = v\}$ を定義する．我々の目的は，強化学習を通じて，移動コストが最小になるような最短経路を選択する政策を学習することである．そのため，目的地 G に到達しない限り，移動する度にコスト r（< 0）を受け取るように設定する[*1]．

2.3 環境変化型の経路選択問題

本章では環境変化について説明する．以降では実行可能・不可能情報が既知なことに加えて，目的地までの距離が計算され，補助情報として与えられているとする．各状態 $(x, y) \in \mathcal{S}$ について，目的地までの距離を $d_M(x, y)$ で示し，まとめて行列 d_M として表す．以下に位置 (x, y) から 1 度の行動で移動できる状態を表す**近傍**を定義する．まず状態 (x, y) の近接位置の集合 $N_M(x, y)$ は $N_M(x, y) = \{(x+1, y), (x-1, y), (x, y+1), (x, y-1)\}$ である．この中で特に実行可能な状態にのみ興味があり，実行可能な位置の集合 $\mathcal{N}_M(x, y)$ を近傍として

$$\mathcal{N}_M(x, y) = \{(x', y') \in N_M(x, y) \mid M(x', y') = 1\}$$

によって定義する．以下に環境変化の直感的な説明を与える．2 つの格子世界 M，M' が与えられ，それぞれ工事前の経路選択問題と，工事後の経路選択問題を考える．我々は変化前の格子世界 M ついて状態行動価値 Q_M が学習済みであるとき，M，M' の変化に関する情報を利用して，新しい $Q_{M'}$ が効率的に学習できるかという問題を考える．

例 1. 図 1 では G に到達するために 2 つの経路が選択できる．あるとき状態 $(x, y) = (3, 0)$ において事故が発生したため，他状態から $(3, 0)$ に侵入することが不可能となり，$M(x, y) = 1$ は $M'(x, y) = 2$ に変化する．

[*1] 1 章で述べたとおり，状態 s と時間 t に依存する報酬 $r(s, t)$ を扱う問題や，正の報酬の和の最大化問題に一般化した場合でも，MDP を学習するという問題として同様の議論が展開できる．

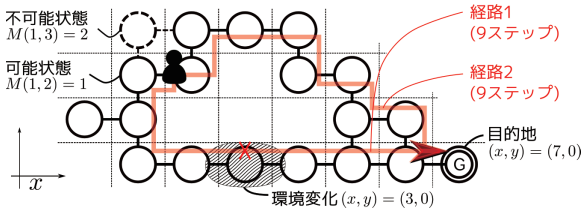


図 1 経路選択問題（2 つの経路が存在）．状態 $(x, y) = (3, 0)$ において障害が発生し，一方は通行不可能になる

Fig. 1 Routing problems in a grid-like graph and an example of obstacles at $(x, y) = (3, 0)$.

本稿ではこのような「観測可能な変化」に注目する．この例では特に $(x, y) = (3, 0)$ の周辺では状態行動価値を更新する必要がある．しかしその一方で，大局的には既存の価値 Q_M が再利用できると想定される．

2.3.1 前提条件

環境変化に関する前提を以下にまとめる．

- **変化が観測可能である．** 文献 [17] で利用されている `lava` など，既存研究は習時とテスト時のドメイン変化に着目した．一方で，たとえば交通事故などの変化はセンサ情報を利用して観測可能であり，M から M' への変化を直接表現できる場合も多い．既存の Jiayu らによる手法はこの変化自体を関数でモデル化するが [12]，本稿では変化前後の環境が与えられるとする．
- **変化は局所的に影響する．** 通行止めなどが発生すると，取るべき最短経路自体が変化する．しかし最短経路の中に環境変化が発生した状態が含まれていなければ，政策は変化しない．本稿では，環境変化によって再学習が必要となる状態が近い距離で固まっていると想定し，この場合に有効な手法を構築する．

2.3.2 形式的定義

定義 1（環境変化）．目的地が $G \in \mathcal{S}$ である M からたかだか k カ所のみ環境が変化した格子世界の集合を $\mathcal{M}^{(k)}$ で表す．$M' \in \mathcal{M}^{(k)}$ な格子世界 M' は，その目的地を G' で表すと

- （目的地が共通）$G = G'$
- （たかだか k カ所のみ変化）$|S_M^{(1)} \triangle S_{M'}^{(2)}| + |S_M^{(2)} \triangle S_{M'}^{(1)}| \leq k$

を同時に満たす．具体的な格子世界の例 $M' \in \mathcal{M}^{(k)}$ が得られることを**経路選択における環境変化**と呼ぶ．

問題 1（環境変化型の経路選択問題）．格子世界 M，その環境変化による格子世界 $M' \in \mathcal{M}^{(k)}$，M における状態行動価値関数 Q_M が与えられる際に，M' において経路選択問題を解く問題（具体的には $Q_{M'}$ を求める）．

3. 環境変化と強化学習

3.1 重み付き初期位置生成を利用した高速学習

Q 学習に代表される強化学習アルゴリズムのいくつかは，サンプルから計算された TD 誤差を用いて政策を更新

Algorithm 1 更新学習フレームワーク W-Copy

1: $Q_{M'}(s,a) = Q_M(s,a)$ とする ▷ (I) Q_M の複製
2: **for** $n = 1$ to N **do**
3: M と M' から重み行列 W_n を計算する
4: $M'(x_0, y_0) = 1$ である初期位置 $s_{n,0} = (x_0, y_0)$ を W_n から重み付きサンプリングする ▷ (II) 重み付きサンプリング
5: 状態 $s_{n,0}$ から有限長のエピソードを発生させる
6: 経験から Q 学習により Q'_M を更新する

する．経路選択において遷移モデルが変化する場合には，距離が大きく変化した状態や，その位置を経由する遠方の位置について誤差が大きくなると予想され，優先的に再学習する必要がある．上記の考察に基づいて本稿では，重み付きサンプリングを利用し再学習を行う．具体的には，変化が局所的という仮定に基づいて，優先的に状態行動価値を再学習すべき状態 $s \in \mathcal{S}$ を選択する．

まずはじめに，格子世界 M と M' の構造変化に着目して構築される優先度を表す重み行列 $W \in \mathbb{R}^{n \times m}$ を導入した更新学習フレームワーク W-Copy の擬似コードを Algorithm 1 に示す．以下に W-Copy 法において，経験を生成するため重み行列を計算する手法と，計算方法の背景を説明する．

3.1.1 TD 誤差に基づくサンプリング

図 1 で示した格子世界について，変化前の環境 M と変化後の M' において，それぞれ状態価値の真値 $V^\star_M$，$V^\star_{M'}$ を計算した結果を**図 2**（上段）に示す．また変化直後を想定して，M で学習した政策 π_M による誤差と，フロンティアの例（3.1.2 項参照）を（下段）に示した．図 2（下段左）の図は，環境変化が実際の状態価値に影響する範囲が局所的である例であり，すべての状態について誤差を小さくするためには，現在誤差が残っている状態を優先的に選択して再学習を行えばよいという発想を裏付けている．

しかし，強化学習を行う環境では真値 $V^\star_{M'}$ は実際に計算できないため，現在の政策 π に基づいて得られるサンプルから，サンプルベースの TD 誤差 $\bar{V}_{\mathrm{error}}(s)$ を計算し，$W(s) = \bar{V}_{\mathrm{error}}(s)$ とすることで，現在の政策 π を用いた場合に誤差が多く残っていると推測される状態から順番にサンプリングして経験を作成し，学習に利用することができる．本稿ではこの手法を Error と呼ぶ．学習においては状態価値 V_M が変化していない状態について，再度学習を試みる必要がないため，誤差情報を利用して経験を生成することを抑制することで，効率的に学習が可能である．図 2 の例では，環境変化にともない誤差が生じている状態数は 5 つ存在する．この状態間の優先順位を用いてサンプリングするという発想は，深層強化学習においても利用されることがある [18]．

3.1.2 フロンティアに基づくサンプリング

本稿では誤差だけではなく環境変化がもたらす距離変化に着目し，状態価値の変化が見込まれる状態を順序付けして徐々に学習を行う手法を提案する．格子世界 M と M'

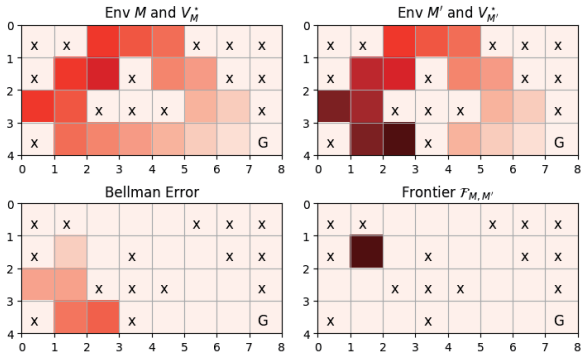


図 2 （上段）状態価値（下段）ベルマン誤差とフロンティア $\mathcal{F}_{M,M'}$．図中の × は実行不可能状態，G はゴールを示す

Fig. 2 (Above) Examples of value functions (Below) the Bellman error, and the picture of frontier $\mathcal{F}_{M,M'}$.

は目的地が共通と仮定しているため，エージェントの取るべき行動が変化する位置は「距離行列 d_M と $d_{M'}$ を比較して大きく変化した位置」である可能性が高い．逆に距離変化が小さければ，少量の経験のみを利用した学習により，変化後の行動を学習することが期待される．これを特徴付けるため，**フロンティア**を定義する．

定義 2（フロンティア $\mathcal{F}_{M,M'}$）．M と M' において，距離変化の絶対値を示す行列 Δ を $\Delta(x,y) = |d_M(x,y) - d_{M'}(x,y)|$ とし，格子世界上の近傍を用いて定まる集合

$$\mathcal{F}_{M,M'} = \left\{ (x,y) \in \mathcal{S} \mid \begin{array}{c} \exists (x',y') \in \mathcal{N}_{M'}(x,y) \\ \text{s.t. } \Delta(x,y) \neq 0, \Delta(x',y') = 0, M(x,y) = 1 \end{array} \right\}$$

を**フロンティア**と呼ぶ．

フロンティア計算において重要な点は，距離の差分に関する情報 $\Delta(x,y)$ である[*2]．図 1 に示した問題例に対して計算した結果の例を図 2 に示す．この例においてフロンティアは $\mathcal{F}_{M,M'} = \{(1,2)\}$ であり，集合 $\mathcal{F}_{M,M'}$ は距離変化に関して二分割された集合の境界に位置する．よって前章の議論より，フロンティアに到達するために必要なステップ数が小さい状態は誤差も小さいと考えられる．そのためフロンティアと TD 誤差の両方を利用することで，Error における再学習の順序を与える基準とすることができると期待される．

3.1.3 Grid k-hop と重み累積情報

先に述べたとおり Q 学習ではサンプルに基づく TD 誤差を利用して学習を行う．そのため，$V(s_{t+1})$ と $V(s_t)$ の誤差がともに大きいときに，TD 誤差による更新量 $\mathcal{E}$ を過小評価する可能性がある．この場合には，本来は $V(s_t)$ や $V(s_{t+1})$ が $V^\star$ の値と異なっていても，学習が進まないか，間違った方向に学習が進む可能性がある．これを考慮し，$\mathcal{F}_{M,M'}$ の周辺から，サンプリング範囲を徐々に広げていき，Q'_M の学習を行うように手法を拡張する．この操作を Grid k-hop として定義する．

[*2] これは仮に最短経路以外の問題を扱う場合には，変化が発生したと思われる地点までのステップ数などで代用してよい．

定義 3（集合 F に対する Grid k-hop）．集合 F の Grid 1-hop である集合を $\mathcal{H}^1(F) = \bigcup_{(x,y)\in F} \mathcal{N}_M(x,y)$ と定義する．2 以上の自然数 k については，k-hop $\mathcal{H}^k(F) = \mathcal{H}(\mathcal{H}^{k-1}(F))$ のように再帰的に定義される．

再学習に利用するエピソードを生成するための初期位置をサンプリングするために，利用する状態の集合 $F \subseteq \mathcal{S}$ を，**影響範囲**と呼ぶ．図 2 中の M，M' について，初期の影響範囲 F_0 をフロンティア $F_0 = \mathcal{F}_{M,M'}$ として，Grid k-hop $F_k = \mathcal{H}^k(F_0)$ を $k \in \{0,1,2,3\}$ を適用して影響範囲を拡大する計算の結果を**図 3** に示す．図 3 では位置 (x,y) について，$(x,y) \in F_k$ であれば 1，そうでなければ 0 として可視化している．この演算と影響範囲の列 $F_0, F_1, \cdots$ を利用することで，フロンティア $\mathcal{F}_{M,M'}$ の周辺からエピソードを生成しつつ，学習を行うことができる．

経路選択問題では，目的地 G までの距離が近い位置の方が，遠い位置に比べて正確な $Q(s,a)$ の値が学習されていると見込まれる．たとえば図 2 に示す例によると，環境変化の影響を（ほとんど）受けていない状態が存在し，これはゴールとフロンティアとの相対位置，特に移動に必要なステップ数に関連付いている．そのため影響範囲 F に含まれているかいないかの 2 値で状態 $s_0 \in F$ を選択するのではなく，フロンティアからの距離や G からの距離を利用した重み付けが学習に有効であると考えられる．本報ではこれを以下の形で実現する．位置 (x,y) における 2 値情報の累積情報 $\mathcal{SH}^k(x,y)$ は，$\mathcal{H}(F)$ を再帰的に適用した際に，位置 (x,y) が影響範囲に含まれた回数を計数して行列 $\mathcal{SH}^k(x,y) = |\{i \in \mathbb{N} \mid 1 \le i \le k, (x,y) \in \mathcal{H}^i(\mathcal{F}_{M,M'})\}|$ として得られる．また上の説明では集合 F に含まれた回数を単純に計数しているが，この $+1$ を距離に比例するような重み関数 $w(x,y)$ を利用することで，距離に比例した優先度を考慮することができる．本稿では $d_{\max} = \max d_M$ とパラメータ δ, α を利用して，

$$w(x,y) = \frac{1}{(d_M(x,y)/d_{\max} + \delta)^{\alpha}} \tag{2}$$

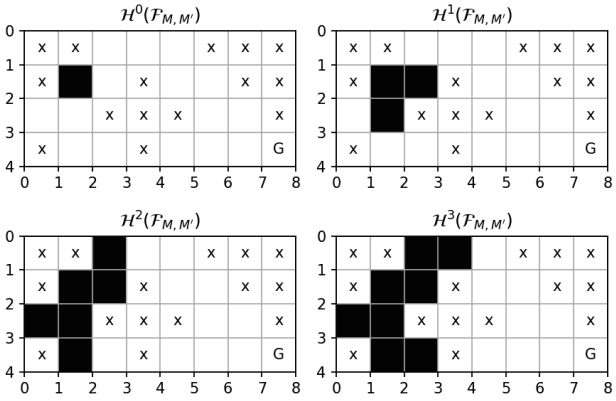


図 3 集合 $\mathcal{F}_{M,M'}$ に Grid k-hop を適用した結果．黒塗りの位置は $(x,y) \in F_k$ を示す

Fig. 3 Frontier $\mathcal{F}_{M,M'}$ and computing k-hops from $\mathcal{F}_{M,M'}$.

の式で計算される重みを利用する．

4. 数値実験

本章では観測可能な環境変化が発生するような経路選択問題に提案手法を適用し，数値実験によって評価する．なお図 1 で利用した格子世界を M_1 と呼ぶ．

4.1 評価基準と比較手法

本稿では以下のスコアを評価に利用する．

定義 4（全位置テスト誤差 ϵ_{error}）．格子世界を M，政策を π とする．すべての実行可能な状態 (x,y) について，政策 π に従った場合に目的地到達までにかかったステップ数を $t^{\pi}_{x,y}$ とし，ϵ_{error} を以下で定義する．

$$\epsilon_{\text{error}} = \sum_{\substack{1\le x\le n, 1\le y\le m \\ \text{s.t. } M'(x,y)=1}} |t^{\pi}_{x,y} - d_M(x,y)|$$

全位置テスト誤差が小さいほど，位置 (x,y) から目的地 G まで取るべき政策が学習できていることを意味する．なお以降では可読性を向上させるため，実験で得られた数値列は移動平均を適用した結果を図示する．

実験の前に，比較を行う学習手法を説明する．

`Init` **法**　転移学習を行わず，変化後の格子世界 M' において一から学習を行う．

`Copy` **法**　学習済みの状態行動価値 Q_M を $Q_{M'}$ の初期値として，その後経験を一様サンプリングから作成して再学習を行う．

`Error` **法**　`Copy` 法と同様の初期値から，TD 誤差を重みとして経験を生成して再学習する．

`W-Copy` **法**　`Error` 法に加えてフロンティアに基づく補正を掛けた重みを用いて経験を生成して再学習する．

これらの手法の特徴を比較したものを**表 2** にまとめる．

4.2 予備実験 (1) 転移学習とサンプリングの効果

格子世界 M_1 について `Init`，`Copy`，および `Error` を適用し，転移学習の効果を検証する．M_1 の再学習を 5 回学習を行った場合の ϵ_{error} の平均と標準偏差の領域を**図 4** に示す．図 4 から分かるとおり，環境変化が発生した場合に，既知の情報を活用しない `Init` に比較して，学習済みの状態行動価値を再利用する `Copy` によって新しい格子世界 M' における学習を高速に誤差を小さく抑えたまま学習

表 2 手法の特徴比較．特徴の (I) と (II) は Algorithm 1 のコメントに対応している

Table 2 Comparisons of methods.

手法	(I) Q_M の複製	(II) 重み付きサンプリング
`Init`	–	–
`Copy`	✓	$\mathcal{S}$ から一様サンプリング
`Error`	✓	✓（TD 誤差を利用）
`W-Copy`$\star$	✓	✓（構造変化と TD 誤差を利用）

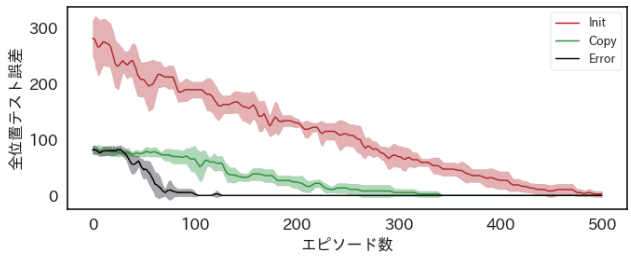


図 4 M_1 における Init/Copy/Error の誤差 ϵ_{error} 比較
Fig. 4 Comparisons of ϵ_{error} for Init, Copy, and Error at M_1.

できる．さらに，経験を生成するための初期状態 s_0 を一様サンプリングする Copy と比較して，サンプルベースの TD 誤差（これを $\bar{V}_{\text{error}}$ で表す）を利用する Error は，より高速に誤差が収束し，結果として所望の政策を得られる．以上の結果より，以降の実験では Init および Copy を省略し，Error を基本的な転移学習法として用いる．

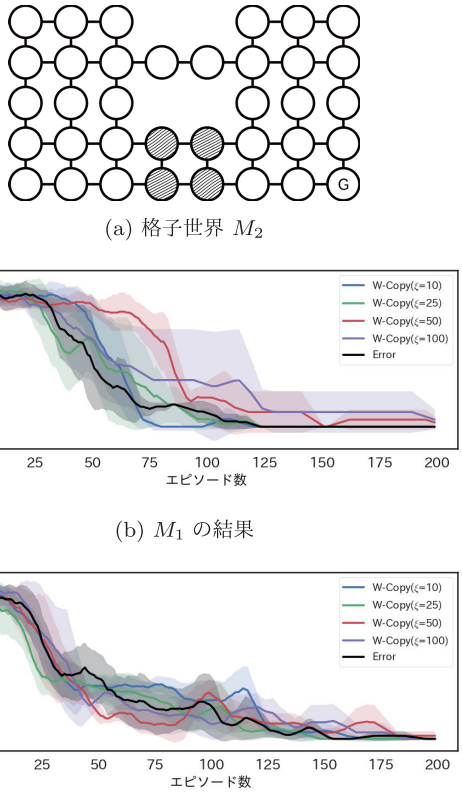
4.3 予備実験 (2) 提案手法の適用方法と検証方法

3.1.3 項で述べたとおり，$\bar{V}_{\text{error}}$ のみを利用した重み付けサンプリングは，TD 誤差が大きい状態の近傍も誤差が大きい可能性が高いため，再学習が停滞すると危惧される．我々が W-Copy を採用する基本的なアイデアは，$\bar{V}_{\text{error}}(s) > 0$ のような状態 $s \in \mathcal{S}$ であり，s の周辺の状態 s' がつ $\bar{V}_{\text{error}}(s') \sim 0$ である場合に着目し，優先的に誤差を小さくするように学習することで，再学習の停滞を防ぐというものである．以上をふまえて W-Copy では，フロンティア $\mathcal{F}_{M,M'}$，Grid k-hop，および $\bar{V}_{\text{error}}$ を同時に利用する．つまり，フロンティアを「状態価値や政策が変化した境界」と見なし，同時に TD 誤差を考慮して効率的にサンプリングを行う．学習が進むと同時に，サンプリングに利用する領域を図 3 で示すように徐々に拡大し，最終的に学習を完了させる．

影響範囲を Grid k-hop で拡大するというアイデアについて考察するため，予備実験を行う．我々の提案手法では，再学習に利用するエピソード数 N のうち，フロンティアを用いて定義される影響範囲を次の形で拡大していく．ある定数 $\xi \in [1, N]$ に対して，N/ξ で N 個のエピソードを区分けし，ある n 番目のエピソードの属する区間 $l = n/\xi$ について，$W_n(x,y) = \mathcal{SH}^{l+2}(\mathcal{F}_{M,M'})(x,y) \times \bar{V}_{\text{error}}(x,y)$ を重みとし，パラメータは $\delta = 1.0$，$\alpha = 0.5$ とする[*3]．予備実験では，M_1 と似た図 5 (a) のような格子世界 M_2 を作成し，$\xi \in \{10, 25, 50, 100\}$ の場合の Error と W-Copy の誤差曲線を比較する．

得られた全位置テスト誤差 ϵ_{error} を図 5 に示す．図 5 (b) および 5 (c) の双方から観察されるとおり，Error の誤差曲線と比較して，学習中の誤差を減らす場合と，逆に増加する場合がある．しかしながら，特に M_2 の再学習序盤（エピソード数が 100 未満）において，W-Copy のアイデアが効果を発揮する可能性が示唆されている．また仮に $F \sim \mathcal{S}$ であり，環境変化に基づいて計算された重みが一様分布に近い場合には W-Copy は Error と同一の手法となる．そのため Error は W-Copy の特殊ケースである．

[*3] すべてのエピソード l について TD 誤差を計算すると計算量が増大するため，本稿では区間を 1 つのバッチと見なし，各バッチごとに TD 誤差を計算して利用する．



(a) 格子世界 M_2

(b) M_1 の結果

(c) M_2 の結果

図 5 格子世界 M_2 と M_1 および M_2 における ξ の影響
Fig. 5 Gridworld M_2 and effects of ξ at M_1 and M_2.

4.4 ランダムな格子世界による実験

予備実験により，W-Copy は Error では利用しない環境変化に起因する重み補正項 $w(x,y)$ を持つため，場合によっては学習を加速させる効果が期待される．これらの手法は利用するエピソードが十分に多ければ誤差 $\epsilon_{\text{error}} \to 0$ となるため，特に我々はエピソード数が少ない序盤に注目する．手法を比較した際に誤差がより小さければ，学習中の政策 π がより良いものになっていることが分かる．

平均順位の検証 誤差を検証するため，10×10 の格子世界と 15×15 の格子世界をランダムに 50 個ずつ作成し，それぞれランダムな環境変化を発生させる．ランダムな環境において，ゴールは右下に固定されている．状態の集合 $\mathcal{S}$ の中から，10×10 の場合には 5 カ所，15×15 の場合には 10 カ所の状態をランダムに変化させた．問題例には通行不可能になる場合と，通行可能になる場合の両方を同数試行した．つまりあるランダムなペア M，M' に対し

表 3 50 個の格子世界における誤差 ϵ_{error} の平均順位比較．（順）は通れなくなる環境変化を，（逆）は通れるようになる環境変化を示す

Table 3 Comparisons of average ranks of ϵ_{error} at 10×10 and 15×15 gridworlds with random environment shifts.

問題	Error	W-Copy（α）			
		0.5	1.0	2.0	（平均）
10×10（順）	3.29	**2.63**	2.90	3.31	2.88
10×10（逆）	3.40	2.88	**2.70**	3.16	2.87
15×15（順）	4.20	2.80	2.82	**2.47**	2.73
15×15（逆）	3.21	**2.86**	3.03	2.96	2.96

て，$M \to M'$ と変化する場合と，$M' \to M$ と変化する場合の両方を実験した．各問題インスタンスに対して，スコア ϵ_{error} を比較する．本節の実験では，$\alpha \in \{0.5, 1.0, 2.0\}$ と変化させながら，再学習に利用した全エピソード数 N に対して，$\eta \in \{100, 200, 400, 800\}$ エピソードまで利用した段階での誤差を計測し，平均順位を計算する．このとき `Error` は試行の平均を，`W-Copy` は各パラメータごとの平均と全パラメータの平均を求める．得られた結果の数値を順位付けし，各手法の平均順位を求める．

表 3 に平均順位を計算した結果を示す．結果から示唆されるように，問題ごとに最小の誤差を達成する α は変化するが，平均的な挙動として `W-Copy` の方が誤差が抑えられる．順方向の結果を比較すると，10×10 では $\alpha = 0.5$ が，15×15 では $\alpha = 2.0$ が平均順位の意味で最も優れていた．

スコアの大小関係の検証 次に学習で利用したパラメータ $\xi \in \{100, 200, 400, 800\}$ の各地点における，誤差 ϵ_{error} の大小関係について検証する．各手法が前節と同様にランダムに生成された格子世界に対して適用されるとき，各地点のスコアを比較するため Wilcoxon の順位和検定[*4]を適用した．しかしいずれの場合も $p = 0.05$ に対して中央値に関する有意差は見られなかった．

4.5 考察

本節では表 2 に概要をまとめた手法について，ランダムに作成した格子世界を用いて実験を行った．先に述べたとおり十分な数のエピソードを再学習に用いることで正しい政策に収束する．そのため我々は，変化直後の政策に基づいて評価される誤差 ϵ_{error} の変化を比較することに注目した．

表 3 の結果に示すとおり，誤差 $\bar{V}_{\text{error}}$ に加えて構造変化に起因する情報を利用する `W-Copy` は，平均的には `Error` 単体より有意に優れているわけではないが，平均的に誤差が減少する効果があることが示唆された．今 α が大きくなるにつれてフロンティア $\mathcal{F}_{M,M'}$ の周辺が多くサンプリングされるため，大規模な問題ではフロンティアによって初期位置を偏ってサンプリングすることが重要であると考えられる．また逆方向の問題については，順方向の手法を変更することなく適用することができ，平均的な挙動も似た結果が得られる．一方で，フロンティア $\mathcal{F}_{M,M'}$ は，現在条件付け $M(x, y) = 1$ が付与されているため，逆方向の問題については通行不可能な状態（図 2 などでは $\times$ で表記される）が $\mathcal{F}_{M,M'}$ に含まれないという非対称性があるため，これを拡張することで双方向の変化に対応できる可能性がある．

[*4] Python の scipy.stats より，`mannwhitneyu` 関数を用いて `alternative=less` を指定して p 値を求めた．

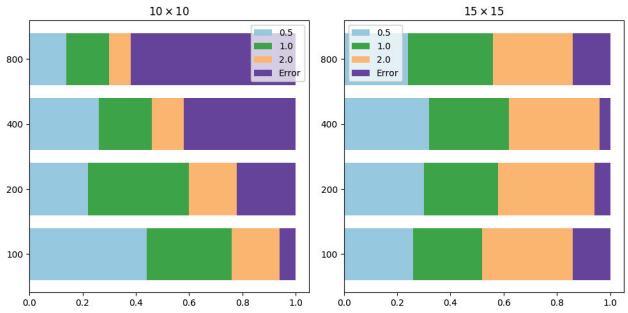


図 6 各 50 個の実験例のうち，$\eta \in \{100, 200, 400, 800\}$ 地点において誤差が最小だった手法の割合比較

Fig. 6 Ratio of methods that achieve the smallest error ϵ_{error} for 50 random instances at $\eta \in \{100, 200, 400, 800\}$ on 10×10 and 15×15 gridworlds.

表 3 は各地点 $\eta \in \{100, 200, 400, 800\}$ の平均的な数値によって順位付けしているため，最後にこれを各地点において詳細に観察した結果を**図 6** に示す．結果から分かるとおり，10×10 では `Error` が最も誤差を減らす場合も η が大きくなるにつれて増加する．しかし 15×15 の場合では，おおむね `W-Copy` のいずれかのパラメータ（$\alpha \in \{0.5, 1.0, 2.0\}$）が最も誤差を最小に保ったまま学習が進行している．この結果より，`W-Copy` は大規模な世界に対して効果的であるとともに，逆に学習が十分に進行した場合では，環境変化による補正重み $w(x, y)$ が不要になっていると考えられる．これは `W-Copy` でフロンティア $\mathcal{F}_{M,M'}$ を大きくした際に，`Error` と同様の重み付きサンプリングになることと性質として同じことになっていると考えられる．

なお経路選択問題は一般のグラフ構造上においても定義されるが，本稿で提案した手法は一般的な問題に対しても適用できる．しかしパラメータ δ，α，ξ を，グラフや格子世界の大きさに応じて調整する必要がある．

5. まとめ

本稿では環境の変化が観測可能であることと，変化の影響が局所的であるという 2 つの仮定に基づいて，遷移モデルが変化することを想定した転移学習法について検討した．我々の手法は，学習済みの環境と，変化後の環境が与えられたときに，変化にともなって距離が変化した状態に注目する重み付きサンプリングを利用し，再学習を行う．

実験により，距離変化の境界点として定義したフロンティアに着目し，その周辺から徐々に再学習を行うことで，環境変化に効率的に適用できる可能性が示唆された．

現在の提案手法は環境変化に基づく重みの計算に対称性があるため，本稿で述べた道が通れなくなる場合の変化とは逆に，道が通れるようになる場合の変化に適用した場合でも，同様の結果が得られる．この 2 つの方向の問題が本質的に異なるかどうかについては，詳細な議論や手法の開発を含め，将来検討すべきである．また，より一般的なグラフ構造上での実験による手法検証や，移動コストが時間や周辺環境に依存する動的環境における検証も今後の課題である．たとえばフロンティアの考え方を一般化し，転移学習が効率的である状況と非効率的な状況を区分して効率的な再学習を行う技術，学習環境の切り分けを行い学習を加速させる技術，再学習の進行にともない適用的に再学習を展開する手法の技術，などが将来的な発展として考えられる．

参考文献

[1] Mnih, V., Badia, A.P., Mirza, M., Graves, A., Lillicrap, T., Harley, T., Silver, D. and Kavukcuoglu, K.: Asynchronous methods for deep reinforcement learning, *Proc. ICML 2016*, pp.1928–1937 (2016).

[2] Silver, D., Huang, A., Maddison, C.J., Guez, A., Sifre, L., Van Den Driessche, G., Schrittwieser, J., Antonoglou, I., Panneershelvam, V., Lanctot, M., et al.: Mastering the game of Go with deep neural networks and tree search, *Nature*, Vol.529, No.7587, pp.484–489 (2016).

[3] Silver, D., Schrittwieser, J., Simonyan, K., Antonoglou, I., Huang, A., Guez, A., Hubert, T., Baker, L., Lai, M., Bolton, A., et al.: Mastering the game of Go without human knowledge, *Nature*, Vol.550, No.7676, pp.354–359 (2017).

[4] Olivecrona, M., Blaschke, T., Engkvist, O. and Chen, H.: Molecular de-novo design through deep reinforcement learning, *Journal of Cheminformatics*, Vol.9, No.1, p.48 (2017).

[5] Peshkin, L. and Savova, V.: Reinforcement learning for adaptive routing, *Proc. IJCNN 2002*, Vol.2, pp.1825–1830, IEEE (2002).

[6] Han, M., Senellart, P., Bressan, S. and Wu, H.: Routing an autonomous taxi with reinforcement learning, *Proc. CIKM2016*, pp.2421–2424 (2016).

[7] Taylor, M.E. and Stone, P.: Transfer learning for reinforcement learning domains: A survey, *Journal of Machine Learning Research*, Vol.10, pp.1633–1685 (2009).

[8] 井戸彩華，野津　亮，本多克宏，市橋秀友：Q 学習における転移学習の方法と活用について，日本知能情報ファジィ学会ファジィシステムシンポジウム 講演論文集，Vol.28, pp.229–232 (2012).

[9] Barreto, A., Dabney, W., Munos, R., Hunt, J., Schaul, T., van Hasselt, H.P. and Silver, D.: Successor Features for Transfer in Reinforcement Learning, *Proc. NIPS 2017*, pp.4055–4065 (2017).

[10] Lehnert, L., Tellex, S. and Littman, M.L.: Advantages and Limitations of using Successor Features for Transfer in Reinforcement Learning, arXiv:1708.00102 (2017).

[11] 大滝啓介，西　智樹，吉村貴克：環境変化型経路選択における強化学習のためのサンプリング法，第 45 回知能システムシンポジウム (2017).

[12] Yao, J., Kilian, T., Doshi-Velez, F. and Konidaris, G.: Direct Policy Transfer via Hidden Parameter Markov Decision Processes, *FAIM2018 Workshops on Lifelong Learning: A Reinforcement Learning Approach* (2018).

[13] Ziebart, B.D., Ratliff, N., Gallagher, G., Mertz, C., Peterson, K., Bagnell, J.A., Hebert, M., Dey, A.K. and Srinivasa, S.: Planning-based Prediction for Pedestrians, *Proc. IROS2009*, pp.3931–3936 (2009).

[14] 内田英明，藤井秀樹，吉村　忍，荒井幸代：道路ネットワークの変化に対する経路選択の学習，情報処理学会論文誌，Vol.53, No.11, pp.2409–2418 (2012).

[15] Howard, R.A.: Dynamic Programming, *Management Science*, Vol.12, No.5, pp.317–348 (1966).

[16] Puterman, M.L.: *Markov Decision Processes: Discrete Stochastic Dynamic Programming*, John Wiley & Sons, Inc. (1994).

[17] Leike, J., Martic, M., Krakovna, V., Ortega, P.A., Everitt, T., Lefrancq, A., Orseau, L. and Legg, S.: AI Safety Gridworlds, arXiv:1711.09883 (2017).

[18] Schaul, T., Quan, J., Antonoglou, I. and Silver, D.: Prioritized experience replay, arXiv:1511.05952 (2015).

大滝 啓介 （正会員）

2016 年京都大学大学院情報学研究科知能情報学専攻修了．京都大学博士（情報学）．同年（株）豊田中央研究所入社．意思決定に関するデータ解析と最適化に関する研究に従事．

西 智樹

2005 年大阪大学応用理工学部機械工学科卒業．2007 年同大学大学院修士課程修了．同年（株）豊田中央研究所入社．意思決定の最適化と自動運転および MaaS への応用に関する研究に従事．

吉村 貴克

1997 年名古屋工業大学知能情報システム学科卒業，1999 年名古屋工業大学大学院電気情報工学専攻修士課程修了，2002 年同専攻博士課程修了．博士（工学）．2002 年（株）豊田中央研究所に入社．機械学習を用いた音声認識・対話，車両データに基づいたドライバへの情報提示システムや製品開発の研究に従事．IEEE 会員．