# 強化学習の転移学習における転移率を用いた再利用方策学習進度の可逆性

河野　仁 *1, 佐藤　弘和 *2

# Reversibility Validation of Learning Progression Using Transfer Rate in Transfer Reinforcement Learning

Hitoshi Kono*1, Hirokazu Sato*2

In recent years, the robot systems with learning algorithms are deployed in the real world situation, for example, automatic driving car, warehouse robots and so on. A reinforcement learning (RL) can be contributed for increasing of intelligence of the robot system, and RL do not need the supervised data. Additionally, RL can explore the optimal solution by itself. However, the robot with reinforcement learning (called RL-agent) has probability to encounter with over fitting caused by reusing obtained policy. A transfer rate has been proposed to reduce the utilization of the policy. Moreover, the transfer rate is thought to have the effect of rolling back the learning progress of the policy to be reused. However, this effectiveness is not validated based on actual reinforcement learning and transfer learning. In this paper, the transfer rate is validated with transfer surface which is visual and quantitative evaluation method of transfer, and the transfer rate is verified the contribution for rolling back of learning progress of reusing policy for transfer learning.

## 1　諸言

近年，自律的・知能的なロボットシステムの実世界応用の研究・開発が活発になされ，学習アルゴリズムが実装されたロボットシステムの研究も多い [1]. さらに，自分で試行錯誤的に解を獲得することができる強化学習は，ヒトによる教示や大量の訓練データを用意する必要がなく，Deep learning などとの技術と融合し実世界応用が活発に議論されている．しかし，強化学習は長時間の学習や環境とのインタラクションが必要であることが課題として指摘されている．そのため，過去に獲得した知識（強化学習の方策）を新たな環境やタスクでの学習に再利用し，学習時間の短縮や高パフォーマンスを獲得しやすくする転移学習が提案されている [2,3].

転移学習は既に獲得した方策をそのまま再利用すると，過学習と呼ばれる環境に過度に適応し，新たな環境や異なる状況に対処できない状態となることがある．これは再利用方策を使用していることに起因していることは自明であり，再利用方策の行動価値を割り引くことで転移時における環境適応性能を向上させるパラメータが提案されている [4,5]. 田代らは，そのパラメータを転移率と呼び転移学習への影響を計算機実験により評価した [6]. 転移率は，再利用方策の学習進度を強制的にロールバックし，転移時の環境適応性能の向上が実験的に示唆されているが，学習進度のロールバックすなわち可逆性については議論が与えられていない．そこで本論文では，河野らが提案した転移曲面という転移学習の効果を可視化する技法 [7] を用いて，転移率による再利用方策への効果を整理し可逆性について議論する．

以下，2 章では方策再利用法の議論に必要な基礎的知識として，強化学習や転移学習などの関連技術を述べる．3 章では，転移率を用いた再利用方策の学習進度をロールバックする手法を述べ，それを検証するための方法を述べる，4 章では実際に強化学習と転移学習を用いて得られた結果を用いて，転移率による効果の評価し，その実験結果と考察し議論する．最後に 5 章では，本稿のまとめと今後の課題を論じる．

## 2　強化学習と転移学習

### 2.1　強化学習

強化学習は，エージェントが試行錯誤的に行動を繰り返し，環境から与えられる報酬を最大化するような行動を学習することで，問題への最適解を獲得するアルゴリズムである [8]．本稿では強化学習に Q 学習を用いている．エージェントは環境の状態空間 $\boldsymbol{S}$ から状態 $s$ を観測でき，それを共に行動空間 $\boldsymbol{A}$ から行動 $a$ を選択し，実行することが可能である．エージェントは目的を達成すると報酬 $r$ が獲得でき，次式により方策 $Q(s,a)$ を獲得する．

$$Q(s_t,a) \leftarrow Q(s_t,a) + \alpha\{r + \gamma \max_{b \in \boldsymbol{A}} Q(s_{t+1},b) - Q(s_t,a)\} \tag{1}$$

ここで，$\alpha$ は学習率で $(0 < \alpha \leq 1)$，$\gamma$ は割引率 $(0 \leq \gamma \leq 1)$ である．方策 $Q(s,a)$ は，基本的にすべての状態 $s$ とすべての行動 $a$ における行動価値が記された Look–up–table で構成されており，Q–テーブルと呼ばれる．本論文では Q 学習で獲得される Q–テーブルを方策と呼ぶ．

*1 東京工芸大学工学部工学科 助教　*2 東京工芸大学工学部電子機械学科 4 年
2019 年 9 月 25 日受理

### 2.2 転移学習

転移学習は，転移元（Source task）のエージェント（広義の転移学習の場合ヒトも含む）が学習した知識を，転移先（Target task）のエージェントが再利用するフレームワークである．転移学習という用語自体は発達心理学などの分野で用いられているが，工学分野における転移学習は，神嶌による文献 [9] が詳しい．近年は機械学習をはじめ強化学習でも応用され成果を得ている．

強化学習における転移学習では，強化学習エージェントが Source task にて学習を行い方策を獲得する．その後，同一もしくは類似環境である Target task において，同一もしくは類似エージェントが Source task で獲得された方策を再利用し，Target task での解の獲得速度を速めたりすることが可能である [1]．他にも，環境や問題設定によっては強化学習より良い解の獲得が可能であったり，学習全体の探査効率を向上させるなどの効果がある．Taylor らの強化学習における転移学習は，次式のように定義されている [2,3]．

$$Q_c(s,a) = Q_t(s,a) + Q_s(\chi_s(s), \chi_a(a)) \tag{2}$$

ここで，$Q_s(s,a)$ は Source task から転移された方策であり，$Q_t(s,a)$ は Target task にてエージェントが使用する方策であり，さらに新たな環境で学習した行動価値もここに更新される．$Q_c(s,a)$ は転移された方策と，Target task で獲得した方策を結合した方策であり，Target task の行動選択は方策 $Q_c(s,a)$ を用いて行われる．注意が必要なのは，他の文献では明示的に $Q_t$ や $Q_s$ と区別しない場合が多い．

関数 $\chi_x(\cdot)$ は Inter-task mapping とよばれ，Target task における $\boldsymbol{S}$ と $\boldsymbol{A}$ のそれぞれの元を，Source task の $\boldsymbol{S}$ と $\boldsymbol{A}$ のそれぞれの元に対応関係を記述する手法である．すなわち，Source task と Target task のエージェントにおける身体性の違いとその対応関係を定義している．これにより，異なるエージェント間での転移学習が効果的となる場合がある．本論文では，ヘテロジーニアスやホモジーニアスなエージェント間での転移学習は議論の対象でないため，式（2）を簡略化し次式のように表記することとする．

$$Q_c(s,a) = Q_t(s,a) + Q_s(s,a) \tag{3}$$

### 2.3 学習進度調整パラメータ

転移学習を実行する場合，式（3）の様に過去に獲得した方策を再利用すると強化学習の過学習状態に陥ることが考えられる．そのため，再利用方策の状態に対する行動価値を割り引いて転移学習する技法が提案されている [4,5]．転移率のメリットとして，転移学習を行う場合，過学習が発生し方策の再利用が難しい場合に転移元での強化学習のし直しをせず，既に持ち合わせている方策を調整することで少しでも使用可能な方策に変更できることが上げられる．田代らは，Takano ら提案したパラメータを転移率と呼び，効果検証を行っている [6]．ここで注意が必要なのが，Taylor らは転移学習の効果を評価するために Transfer ratio という指標を提案しているが，転移率とは別の手法である．Transfer ratio と区別して表現するために，転移率は Transfer rate と訳される．転移率の調整により，新たな環境が再利用方策を獲得した環境と異なる場合においても適応し，再学習による転移学習の効果が発現するという知見を得ている．転移率を導入した転移学習は次式のように形式的に定義が可能である．

$$Q_c(s,a) = Q_t(s,a) + \tau Q_s(s,a) \tag{4}$$

$\tau$ $(0 < \tau \leq 1)$ が転移率と呼ばれているパラメータであり，基本的にはヒトが試行錯誤・経験的に転移率を決定する．転移率は転移元の方策の学習進度を疑似的にロールバックし，すなわち学習途中の方策の状態に変換することが可能であると示唆されている．なお，$\tau = 0$ に設定すると再利用方策をキャンセルしてしまうため，転移の効果が得られない．これはゼロから学習する転移学習，すなわち強化学習であることを意味する．一方 $\tau = 1$ であれば，十分に学習した方策を再利用し，言い換えると再利用方策を最大限活用して転移学習を来なうことを意味する．そのため 0 から 1 の間の転移率は再利用方策の活用度合いが中間的であることから，転移率は学習進度をロールバックできるという仮説が得られている．

### 2.4 転移曲面

通常，強化学習は獲得方策の推移や問題解決速度の推移を現す学習曲線により学習を直感的に評価することが可能である（Fig. 1）．学習曲線は，横軸に学習の繰り返し回数すなわち学習進度，縦軸にパフォーマンスを現す．Fig. 1 の例では，縦軸は行動選択回数を表し，すなわち問題解決速度であるため値が低いほうがパフォーマンスが高い．学習曲線は，ある 1 つの学習において性能や効果を評価するために有用であるが，転移学習の場合転移元でどれくらい学習を継続し，獲得した方策を転移すれば良いのかはヒトの経験と直感に委ねられ転移学習の効果もそれに依存する．そのため，様々な学習条件で転移を行い，転移学習の効果を直感的に把握しやすくする手法が河野らにより提案されている [7]．河野らは Fig. 2 のような転移曲面を提案した．転移曲面は，転移元における各学習進度で転移学習を来ない，その結果である学習曲線を 1 つの曲面として可視化した曲面である．

転移曲面は，転移元での学習進度と転移先での学習進度，転移先でのパフォーマンスの 3 つの軸で構成されており，面の勾配を見ることで転移に最適な転移元の学習進度を直感的に把握することが可能でありる．例えば Fig. 2 においては，転移元の学習進度（エピソード数）を 100 で終了し，転移学習により方策を再利用することで，転移先で高いパフォーマンスが得られる．この転移曲面でのパフォーマンスは問題解決速度であるため，値が低いほうが良い．

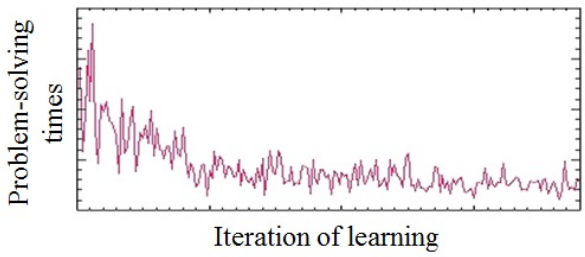



Fig. 1 Simplified example of learning curve.

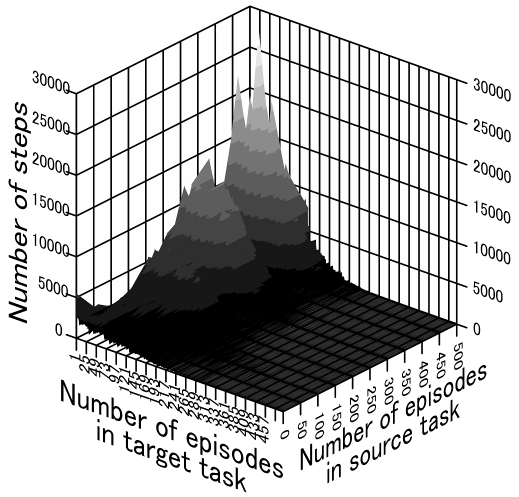



Fig. 2 Example of transfer surface of transfer learning in reinforcemnet learning.

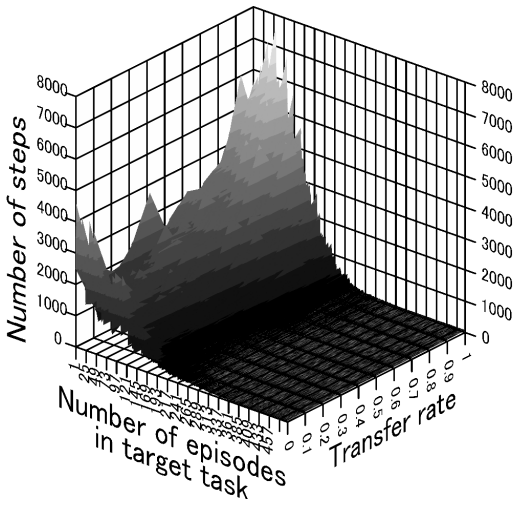



Fig. 3 Example of transfer surface with transfer rate.

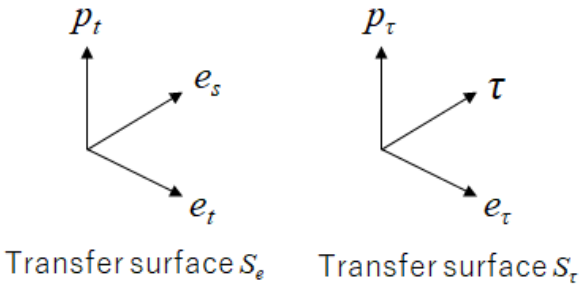



Fig. 4 Axes of transfer surface.

また，転移曲面において転移元のエピソード数が 0 である時が，転移先では再利用方策なしでの転移学習，すなわち単なる強化学習のパフォーマンスとなるため，転移曲面を評価する際の基準となる．転移曲面は，転移率を用いた転移学習においても有効であり，例えば Fig. 3 のように，転移もとでの学習進度を転移率に置き換えることで転移曲面を構成可能である．Fig. 3 では，転移率を 0.2 に設定すると転移先で高いパフォーマンスが得られることが見て取れる．

Fig. 2 と Fig. 3 で共通して読み取れる結果であるが，転移先の学習回数を進めエピソード数を増やしていくと，収束する値がどれも同様である．しかし，転移先での学習初期では明らかなパフォーマンスの違いが発現しており，この違いが転移先での学習効率の良し悪しを判断することが出来る．

## 3 転移率における学習進度調整の検証方法

### 3.1 転移曲面の設定

上述した通り，転移曲面は転移元の学習進度における転移の効果を表し，転移率へ応用した場合は，各転移率における転移学習の効果を評価することが可能となる．この 2 つの転移曲面の形状を比較することで，転移率により学習進度がロールバック可能であるかが評価することが可能であると考える．

転移元の学習進度をもとにした転移曲面を $S_e$ とする．転移元の学習エピソード数 $e_s$，転移先の学習エピソード数 $e_t$ とし，転移曲面を次式のような関数として定義する．

$$S_e(e_s, e_t) = p_t \tag{5}$$

ここで $p_t$ は転移先でのパフォーマンス（例えば Number of steps）である．なお，転移曲面 $S_e$ を $e_t$ と $p_t$ の軸で読み取ると，転移先での学習曲線となる．また，転移曲面 $S_e$ の $e_s$ を転移率 $\tau$ $(0 < \tau \leq 1)$ に置き換えた転移曲面 $S_t$ を次式のように定義する．

$$S_t(\tau, e_\tau) = p_\tau \tag{6}$$

ここで $e_\tau$ は転移率を用いて転移した転移先でのエピソード数である．それぞれのラベルを各軸で整理した転移曲面軸を Fig. 4 に示す．

### 3.2 転移曲面の比較

転移曲面と $S_e$ と $S_t$ が与えられたとき，軸 $e_s$ と $\tau$ が同等で，さらに軸 $e_t$ と $e_\tau$ が同等に設定される時，軸 $p_t$ と $p_\tau$ の比較が可能となる．軸 $e_s$ と $\tau$ を同等に設定するためには，転移元でエピソード $e_s$ まで強化学習を行い，獲得した方策を転移している状態を $\tau = 1$ での転移とする．

上述の条件が満たすことが出来れば，転移率の調整に対するパフォーマンスの変化が，転移元における学習進度が途中の状態で転移を行っている状態と対応を取ることが可能となり，転移率を小さくしていくことが学習進度のロールバックであることの評価が可能となる．すなわち，実際に各学習進度で転移学習を行い，そのパフォーマンスを表す転移曲面 $S_e$ が真値となり，転移率を用いた転移曲面 $S_t$ がどの程度の誤差が表れるかという問題となる．したがって，転移曲面同士の形状の違いを評価することで「転移率調整（値を下げる）＝学習進度のロールバック」を検証することが可能となる．

ここで，理想的な $S_e = S_t$ とは $p_t = p_\tau$ であり線形関係が成立する．曲面形状の比較を行う技術は多く存在するが，本研究の曲面比較では単純に得られた 2 つの転移曲面から $p_t = p_\tau$ における直線へのフィッティングを評価すればよい．そのため，本論文では転移曲面 $S_e$ と $S_t$ の比較に，1 次式へのフィッティングを評価する相関係数（correlation coefficient: CC）と二乗平均平方根誤差（root mean squared error: RMSE）を用いて評価を試みる．

相関係数である CC は式 (7) で定義し，RMSE は観測値と予測値において回帰モデルの誤差を評価するが，本論文においては真値と言える転移曲面が存在するため，RMSE は式（8）で定義する．

$$\mathrm{CC} = \frac{\frac{1}{n}\sum_{i=1}^{n}(p_{e,i} - \overline{p_e})(p_{\tau,i} - \overline{p_\tau})}{\sqrt{\frac{1}{n}\sum_{i=1}^{n}(p_{e,i} - \overline{p_e})^2} \times \sqrt{\frac{1}{n}\sum_{i=1}^{n}(p_{\tau,i} - \overline{p_\tau})^2}} \tag{7}$$

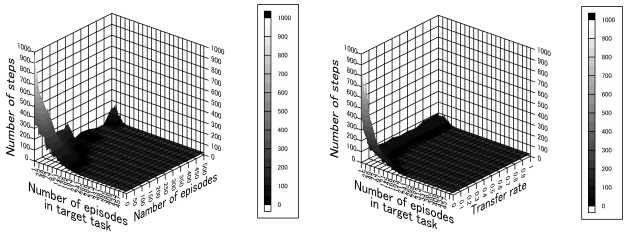
$$\mathrm{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(p_{\tau,i} - p_{e,i})^2} \tag{8}$$

ここで，$n$ は転移曲面における $p_e, p_\tau$ の総数であり，$p_{\tau,i}, p_{e,i}$ の添え字 $i$ はそれぞれの転移曲面で対応する点 $p_e, p_\tau$ の番号である．$\overline{p_e}, \overline{p_\tau}$ は $p_e, p_\tau$ それぞれの平均である．

## 4 転移曲面を用いた転移率評価

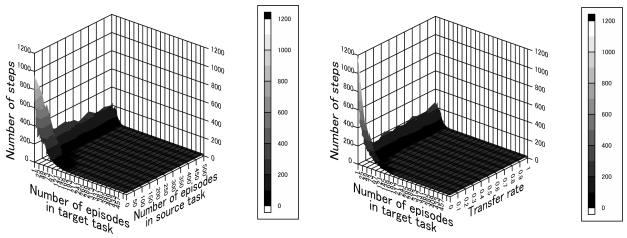
### 4.1 条件

本論文では，転移曲面の生成に実際の強化学習と転移学習で得られた結果を用いる．問題としてはグリッドワールドの最短経路問題を採用し，転移元では最大 500 エピソードの学習を行う．またグリッドワールドでは障害物が配置されており，その配置を変更することで計 5 種類の転移曲面を生成する．真値となる転移曲面 $S_e$ は，転移元の 50 エピソードずつで転移学習を行い，転移先では 500 エピソードの転移学習を行い学習曲線を取得，統合することで生成する．評価対象となる転移曲面 $S_t$ は，転移元で 500 エピソードの学習を行った状態の方策を転移し，その方策の再利用時に式（4）を用いる．そのとき，転移率 $\tau$ を 1 から 0 まで 0.1 ずつ値を変更して学習曲線を取得し，転移曲面を構成する．転移率 $\tau$ は $(0 < \tau \leq 1)$ であり，$\tau \in \mathbb{N}$ であるが，本実験では $\tau = 0$ で転移しない状況，すなわち強化学習で一から転移先の環境を学習する学習曲線も基準として評価したいため，$\tau = 0$ も用いる．なお，本実験でも用いた最短経路問題では，転移元で 500 エピソードの強化学習を行えば十分に最短経路を探査・学習可能であることを事前に確認している．また，転移元と転移先では環境の障害物配置は異なっている，しかし，転移学習が可能である環境に事前に調整されている．




(a) With episode axis (b) With transfer rate axis

Fig. 5 Transfer surfaces of condition 1.



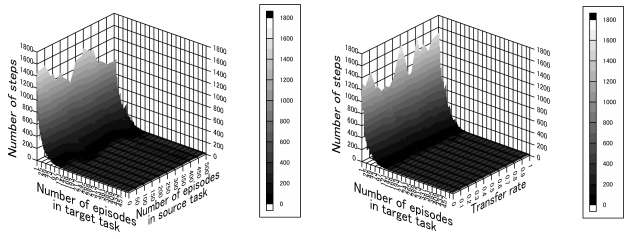

(a) With episode axis (b) With transfer rate axis

Fig. 6 Transfer surfaces of condition 2.

### 4.2 評価に用いる転移曲面

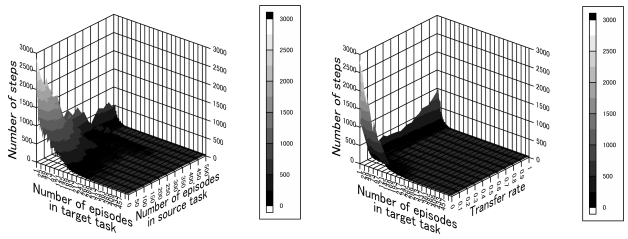
上述の実験条件により生成した 5 種類，計 10 個の転移曲面を Fig. 5 から Fig. 9 に示す．それぞれの図の転移曲面において，(a) With episode axis は転移曲面 $S_e$ であり，(b) With transfer rate axis は転移曲面 $S_\tau$ である．

また，明らかに $S_e = S_\tau$ が成立しない，言い換えれば転移学習の効果が無い状態の CC と RMSE も，疑似的に条件を作成し評価する．この評価には，転移曲面 $S_e$ には Fig. 9(a) を用い，転移曲面 $S_\tau$ には Fig. 5(b) を用いる．この条件では，明らかに形状が異なり，Condition 1 から 5 の評価値を比較する指標の一つとなる．これを Condition



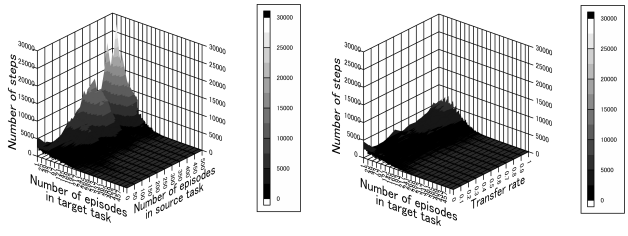

(a) With episode axis (b) With transfer rate axis

Fig. 7 Transfer surfaces of condition 3.




(a) With episode axis (b) With transfer rate axis

Fig. 8 Transfer surfaces of condition 4.




(a) With episode axis (b) With transfer rate axis

Fig. 9 Transfer surfaces of condition 5.

x と呼ぶこととする．

### 4.3 結果と考察

Condition 1 から Condition 5，Condition x までの CC と RMSE を Table. 1 に示す．相関係数で転移曲面同士の評価を考えた場合，Conditon 1 から 5 のすべてで 0.6 以上であり $p_t$ と $p_\tau$ の間には線形関係が成立していると考えられる．一方，Condition x においては，相関係数が 0.2 を下回ることから相関が無く，すくなくとも $p_t$ と $p_\tau$ は成立していないと言える．しかし相関係数の場合，評価している直線とのフィッティングは $p_t = ap_\tau + b$ の一次方程式における相関関係であるため，完全な $p_t = p_\tau$ ではない．例えば，Fig. 9 の転移曲面では，明らかに $p_t = p_\tau$ ではないが比例関係があると考えられ，$p_t = ap_\tau + b$ としては比較的高い相関係数が得られたと考えられる．そのため，Condition 5 においては RMSE が Condition x と同様に高い値となっている．他の条件においては，Condition x と比較すると RMSE も値が低く，相関係数の値からも $p_t = p_\tau$ の関係が成り立っていると考えることが出来る．

相関係数が高く RMSE も高い場合における $p_\tau$ の散布図を Fig. 10 に示す．Fig. 10 から真値とする $p_e$ の誤差はあるものの，$p_t = p_\tau$ となるような相関関係が表れていると考えられる．Conditon x の散布図を Fig. 11 に示す．Fig. 11 から，転移元と転移先において転移曲面の形状が極端に異なる場合は，$p_t = p_\tau$ をはじめ $p_t = ap_\tau + b$ とのフィッティングもできない．

これらの結果から，転移学習が行えるという前提条件のもと転移元と転移先の転移曲面形状が類似傾向にあることがわかり，すなわち転移先における転移率の変更は再利用方策の学習進度を疑似的にロールバックすることが可能であると考えられる．Condition x の相関係数と RMSE からもわかる通り，転移曲面の形状が似ていない場合，すなわち転移が成功しないような条件においては転移率の変更は意味をなさないことも明らかである．この場合，転移率を $\tau = 0$ とすることで方策を再利用しない状況となるので，方策の忘却のような応用も可能である．

Table. 1 Results of CC and RMSE in each condition

| Condition | CC | RMSE |
|---|---|---|
| Condition 1 | 0.81 | 34.98 |
| Condition 2 | 0.83 | 38.71 |
| Condition 3 | 0.91 | 77.54 |
| Condition 4 | 0.62 | 231.63 |
| Condition 5 | 0.74 | 2256.36 |
| Condition x | 0.16 | 3094.72 |

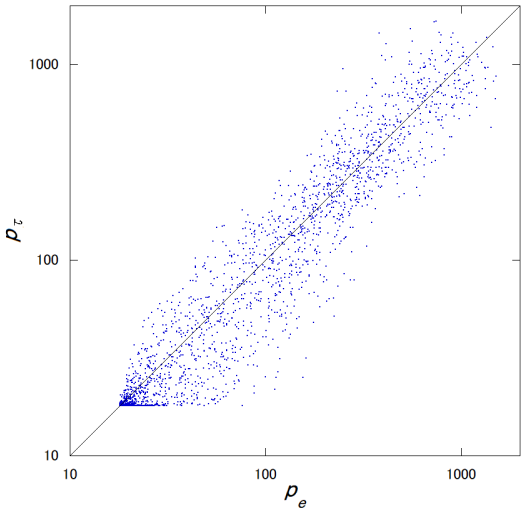



Fig. 10 Scatter plot between $p_e$ and $p_\tau$ with regression slope in condition 5.

## 5 結言

本論文では，強化学習の転移学習における方策再利用度合い調整パラメータである転移率 $\tau$ が，再利用方策の学習進度をロールバックする効果があるか確認した．最短経路問題を用いた実際の強化学習と転移学習の結果から得られ

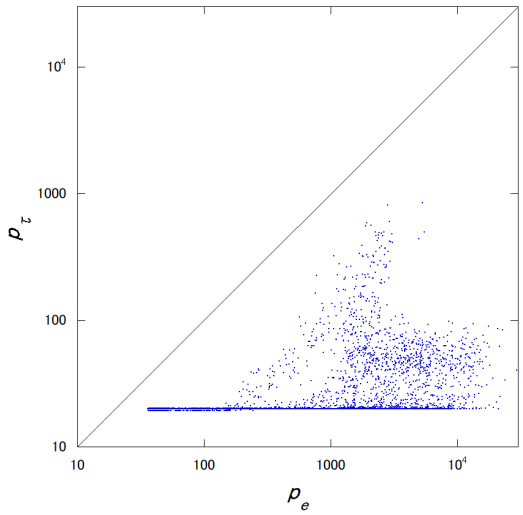



Fig. 11 Scatter plot between $p_e$ and $p_\tau$ with regression slope in condition x.

る転移曲面を用いて，真値となる転移曲面 $S_e$ と転移率の調整により得られた転移曲面 $S_\tau$ の線形関係 $p_e = p_\tau$ を相関係数と RMSE により評価した．評価結果から，転移学習が成功するという前提条件において，2 つの転移曲面は $p_e = p_\tau$ の関係があると考えられ，すなわち転移学習時の転移率調整は再利用方策の学習進度を疑似的にロールバックすることが可能であると考えられる．

しかし，転移曲面の形状は強化学習や転移学習が解く問題の難易度やエージェントの自由度等にも左右され，本論文の結論を一般化するには評価が十分ではない．そのため，今後は最短経路問題以外や様々な転移曲面が表れる実験条件においても転移率の検証が必要であると考えられる．

## 謝辞

本研究の一部は JSPS 科研費 JP18K18133, JP19K12173, 公益財団法人スズキ財団の助成を受けた．

## 参考文献

[1] A. Lazaric, “Transfer in Reinforcement Learning: a Framework and a Survey”, Reinforcement Learning–State of the art, vol.12, Springer, pp.143-173, 2012.

[2] M. E. Taylor and P. Stone, “Transfer learning for reinforcement learning domains: A survey”, Journal of Machine Learning Research, vol.10, pp.1633–1685, 2009.

[3] M. E. Taylor, ”Transfer in Reinforcement Learning Domains.” Studies in Computational Intelligence, vol.216, Springer, 2009.

[4] T. Takano, H. Takase, H. Kawanaka and S. Tsuruoka, “Effective Reuse Method for Transfer Learning in Actor-critic”, Proceedings of the SCIS & ISIS 2010, pp.137–141, 2010.

[5] T. Takano, H. Takase, H. Kawanaka, H. Kita, T. Hayashi, and S. Tsuruoka, “Transfer Learning based on Forbidden Rule Set in Actor-Critic Method”, International Journal of Innovative Computing, Information and Control, vol.7, no.5(B), pp.2907–2917, 2011.

[6] 田代淳史, 河野仁, 神村明哉, 富田康治, 鈴木剛, “ヘテロジーニアス間転移学習のための知識再利用法の検討”, JSME Conference on Robotics andMechatronics, 2A1-L06(CD-ROM), 2015.

[7] 河野仁, 三浦昇三, 温文, 鈴木剛, “強化学習における方策転移度合い決定のための転移曲面の検討”, 24 回画像センシングシンポジウム（SSII2018）, IS2-28, 2018.

[8] R. S. Sutton, A. G. Barto, “Reinforcement Learning: An Introduction”, The MIT Press, 1998.

[9] 神嶌敏弘, “転移学習”, 人工知能学会誌 vol.25, no.4, pp.572–580, 2010.