逆強化学習を用いた報酬関数推定と時空間制約下における歩行者の行動軌跡生成

日高 健[1]・早川 敬一郎[2]・西 智樹[3]・薄井 智貴[4]・山本 俊行[5]

[1]正会員 修 (工) 株式会社豊田中央研究所 社会システム研究領域（〒 480-1192 愛知県長久手市横道 41-1）
E-mail: hidaka@mosk.tytlabs.co.jp
[2]正会員 修 (工) 株式会社豊田中央研究所 社会システム研究領域（〒 480-1192 愛知県長久手市横道 41-1）
E-mail: kei-hayakawa@mosk.tytlabs.co.jp
[3]非会員 修 (工) 株式会社豊田中央研究所 データアナリティクス研究領域（〒 480-1192 愛知県長久手市横道 41-1）
E-mail: nishi@mosk.tytlabs.co.jp
[4]正会員 博 (工) 名古屋大学大学院経済学研究科 特任准教授（〒 464-8603 愛知県名古屋市千種区不老町 C1-3（651））
E-mail: tomo.usui@nagoya-u.jp
[5]フェロー会員 博 (工) 名古屋大学未来材料・システム研究所 教授（〒 464-8603 愛知県名古屋市千種区不老町 C1-3（651））
E-mail: yamamoto@civil.nagoya-u.ac.jp

都市部の歩行者空間などの複雑な環境における歩行行動の記述を行うためには，歩行の意思決定に関する異なるレベル（Strategic, Tactical, Operational）を統一的に扱うモデルが必要とされる．これらを統一的に扱う試みは，一部にはあるものの実用的なレベルにはない．我々は歩行者の経路選択に関する行動方策を行動軌跡から明らかにするデータ駆動型の学習手法に着目し，歩行行動の記述を試みる．これまでに提案されているモデルでは，目的地指向の行動しか扱うことができなかったが，時空間制約を考慮することで目的地指向の行動軌跡のみならず，散策的な行動軌跡についても表現可能な手法について提案する．さらに，数値実験を行い，制約の強さを変更することで目的地指向の行動軌跡から散策的な行動軌跡まで多様な軌跡の生成が可能であることを確認した．

Key Words: 歩行者行動モデル，逆強化学習，時空間制約，行動軌跡生成

1. はじめに

歩行行動の理解は，都市部の空間計画，公共空間の設計などへの応用が期待される重要な研究トピックである．Hoogendoorn(2001)[1] によれば，歩行行動は Strategic（活動集合の選択），Tactical（活動スケジューリング，場所，経路の選択），Operational（歩行挙動）の 3 つのレベルに分けられ，それらが相互に依存関係を持つ．しかしながら，既存の歩行モデルの多くは，施設内の限定されたエリアを対象に歩行者の進行方向や速度を決定する Operational なレベルの歩行挙動を扱うことが多い[1]．また，Tactical なレベルの意思決定である経路選択を考えてみると，目的地指向（goal-directed）の行動方策が用いられることが多く，散策的な行動や環境中の対象物からの影響による滞在行動が扱われることは少ない．

都市部の歩行者空間などの複雑な環境における歩行者の回遊行動の記述を行うためには，異なるレベルの歩行の意思決定を統一的に扱うモデルが必要となる．異なるレベルの歩行の意思決定を統一的に扱う試みとして，Hoogendoorn and Bovy(2004)[5] は，歩行者が主観的効用を最大化する（Utility maximizer）という仮定のもと規範的な歩行行動理論を提案した．このモデルでは，Tactical なレベルの意思決定（活動のスケジュール，活動場所，経路）と Operational なレベルの意思決定（方向，速度）を効用最大化により同時に決定する．ここでは，歩行による不効用（例えば，旅行時間や障害物による不快感など）が考慮され，活動の完遂と歩行の不効用のトレードオフを扱うことができる．しかしながら，計算時間や収束性，現実的でない行動の仮定等からシミュレーション目的としては実用的でないという批判がある[6]．

Hoogendoorn and Bovy(2004)[5] のモデルを歩行者の意思決定に関わる様々な要因をモデルとして記述するモデルベースの手法とするならば，その一方で，実際に得られた歩行者の行動軌跡データから経路選択の行動方策を明らかにするデータ駆動型の手法も既に幾つか提案されている[7),8]．これらのモデルでは，歩行者の行動がマルコフ決定過程（Markov Decision Processe; MDP）に従うという比較的単純な仮定のもと，歩行による効用（不効用）を推定しつつ，歩行者の経路選択の行動方策を明らかにするというものである．Ziebart et

[1] Social Force モデル[2)]，Cellular Automata（CA）モデル[3)]，離散選択モデル[4)] など様々な手法が提案されている．

al.(2009)[7] は，Ziebart et al.(2008)[9] で提案された最大エントロピー逆強化学習の枠組みをロボットの経路計画のための歩行者行動予測に適用した．また，Kitani et al.(2012)[8] は，これを発展させ，画像から場所の特徴量（道路，歩道，車，芝生など）を自動でラベリングし，ノイズ環境下でこれら特徴量が観測されることを考慮した上で，人の歩行の特徴量に対する好みを学習する Hidden variable MDP（hMDP）を提案した．しかしながら，Ziebart et al.(2009)[7] や Kitani et al.(2012)[8] は，行動予測の段階では acyclic な経路しか考慮されない目的地指向のモデルになっている．

本稿では，Ziebart et al.(2009)[7] や Kitani et al.(2012)[8] で用いられたデータ駆動型の経路選択行動方策の学習方法を援用しつつ，さらに，時空間制約[10] 下における行動軌跡の生成手法について新たに提案する．これにより，時間的な余裕度合いのなかでの目的地指向の行動と散策行動のトレードオフが表現可能となる．このトレードオフは，与えられた活動場所にいち早く向かい，活動の遂行によって効用を獲得するのか，もしくは，散策的行動によって経路上で効用を獲得するのかという背反関係を表すものである．提案モデルの位置づけを**図–1** に示した．提案モデルは，データ駆動型の学習を採用することにより，より現実に即した歩行者の行動方策を学習し，かつ，時空間制約を考慮することで目的地指向の行動軌跡のみならず，散策的な行動軌跡も生成することができる．

本稿の構成は以下の通りである．続く第 2 章にて関連する研究をレビューした後，第 3 章で提案モデルについて詳述する．具体的には，逆強化学習を用いた報酬関数及び確率的な方策の学習方法について言及し，その後に時空間制約下における行動軌跡の生成法について説明を行う．第 4 章にて提案モデルの特性を明らかにするために簡単な数値実験を行い，最後に，第 5 章にて研究のまとめを行う．

2. 関連研究

本章では，関連する研究のレビューを行う．ここでは，目的地指向でない経路選択行動，逆強化学習（Inverse Reinforcement Learning; IRL），空間表現について取り上げる．

(1) 目的地指向でない経路選択行動

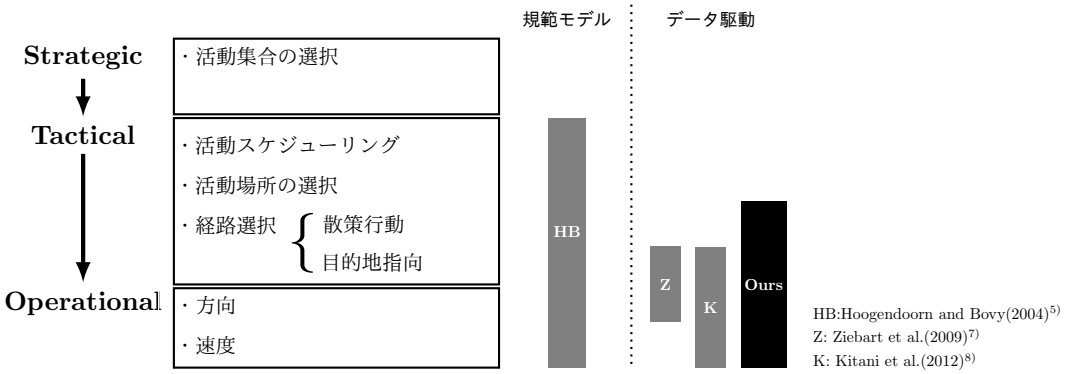
目的地指向でない経路選択行動を扱った研究として，視覚情報と歩行行動の関係を直接モデリングする方法がある[11),12)]．そこでは，Visibility Graph Analysis（VGA）を利用して歩行行動が決定されるが，視覚的な情報によるため探索的な行動しか表現されず，また，広いスペースなど特定の環境において制限がある．これに対して，提案手法では視覚的な情報に関わらず，場所の価値をエキスパートの行動軌跡から学習することで，環境による制限を受けることなく行動方策を学習することが可能である．

視覚情報を利用したモデルは静的な環境情報から決定されるモデルであるが，動的な表現のモデルとして扇型の視野を模した行動空間のなかから速度と方向を選択するモデルが提案されている[4),13),14)]．さらに，Wang et al.(2014)[15] は扇形の視野のモデルに，視野内に入る attractor の影響を考慮した探索的な行動モデルを提案した．これらとは異なるアプローチとして，Borgers and Timmermans(1986)[16] は逐次的意思決定による計画的なトリップとそれに付随して起こる衝動的な立ち寄り行動（impulse stop）によって都心部の消費回遊行動モデルを表現した．しかし，これらのモデルでは環境側の魅力度の大きさによる歩行行動への影響しか考えられておらず，時空間的な制約によって生じうる目的地指向の行動と散策的な行動のトレードオフは一切考えられていない．また，これらの行動は視野に入るものなど近視眼的（myopic）な行動を前提としたモデルとなっている．これに対し，提案モデルでは割引率を考慮することで，歩行者が完全な情報を持つ場合の行動（e.g. Hoogendoorn and Boby(2004)[5]）から近視眼的な行動（e.g. Wang et al.(2014)[15]，Borgers and Timmermans(1986)[16]）まで幅広い表現を可能とする．

(2) 逆強化学習

最適な行動系列と環境モデルを所与として報酬関数[2] を求める問題は，報酬関数を所与として最適な行動系列を決定する強化学習の逆問題として逆強化学習と呼ばれる[17]．逆強化学習には，各状態における最適な行動を所与とし，線形計画問題を解くことで報酬関数を推定する手法[18] や，エキスパートの行動軌跡を所与とし，エキスパートと同じような行動軌跡が得られる報酬関数を学習する手法が存在する．エキスパートの行動軌跡から学習する手法には，Abbeel and Ng(2004)[19] によって提案されたものや Ziebart et al.(2008)[9] による最大エントロピー逆強化学習などがある．Abbeel and Ng(2004)[19] の方法では，条件を満たす解が複数ある曖昧性の問題があることが指摘されている[9] 一方で，最大エントロピー逆強化学習を用いた報酬推定では解が一意に決まるという特徴を持つ．さらに，Ziebart et al.(2009)[7] は，確率的な行動系列の出力が可能なように MDP を緩和し，観測途中の行動系列を入力として Bayes 推定を行うことで，将来の経路の予測を確率的

[2] どのような条件のときにどの程度の報酬を環境から得ることができるかを定めた関数のこと．



図–1: 提案モデルの位置づけ

に求める手法を提案した．また，Kitani et al.(2012)[8)] は，特徴量自体の不確実性も考慮した上で将来の人の行動予測を行う枠組みを提案した．しかしながら，これらのモデルの経路予測では同じ状態は 1 回までしか含まないとの仮定がされており，同じ状態（場所）に留まる滞在や循環的な経路を求めることができない．これに対して提案手法ではこれらを考慮した経路行動方策を求めることが可能である．

(3) 空間表現

ネットワーク表現のモデルには将来に渡る期待効用を考慮した上で経路を逐次的に決定する方法として，recursive logit（RL）モデル[20)] や，RL モデルに割引率の概念を導入した discounted reucursive logit（DRL）モデル[21)] が提案されている．しかしながら，DRL モデルで導入される割引率はリンクに対する割引率であり，意思決定のポイントであるノード毎に割引率が乗算されるモデルである．したがって，経済学分野などで一般的に用いられる時間に対する割引率とは必ずしも整合しない．これに対して提案手法の利点は，均質な時空間表現（グリッド，等時間間隔）を用いることにより，時間に対する割引率を考慮できるようにし，時空間制約を単純な形で表現できるようにした点にある．

3. 時空間制約を考慮した行動軌跡生成モデル

(1) モデル概要

提案モデルは，大きく 2 つのサブモデルから構成される．最初のモデルは，与えられた環境モデルと行動軌跡を入力として報酬関数の推定と確率的な方策を決定するモデルである．このモデルは，Ziebart et al.(2009)[7)] によって提案された確率的な方策を出力に持つ MDP（soft-max MDP と呼ばれる）をもとに最大エントロピー逆強化学習を実行するモデルである．これにより場所ごとの報酬関数の推定と確率的方策を得ることができる．もう一つのモデルは，目的地と到着必須時間を所与とし，推定された報酬関数及び確率的方策をもとに，時空間制約下における行動軌跡を生成するモデルである．

次節以降では，確率的な方策を出力とする MDP について説明した後，2 つのサブモデル，すなわち，最大エントロピー逆強化学習を用いた報酬関数の推定モデル及び，時空間制約下における行動軌跡生成モデルについて説明を行う．

(2) MDP による確率的方策の算出

MDP は $< \mathcal{S}, \mathcal{A}, \mathcal{T}, \mathcal{R} >$ の要素の組で表現することができる．$\mathcal{S}$ は状態空間，$\mathcal{A}$ は行動空間，$\mathcal{T} : \mathcal{S} \times \mathcal{A} \to \mathcal{S}$ は状態遷移モデル，$\mathcal{R} : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$ は報酬モデルを表す．ここで，以下に示される状態価値関数 $V(s)$，行動価値関数 $Q(s,a)$ を導入する．

$$V(s) = \mathbb{E}\left[\sum_{t=0}^{\infty} \gamma^t R(s_t, a_t)\right] \tag{1}$$

$$Q(s,a) = R(s_t, a_t) + \gamma V(\mathcal{T}(s_t, a_t)) \tag{2}$$

ただし，$\gamma \in [0,1]$ は報酬の割引率を表す．いま求めたい最適な方策（報酬和を最大にする方策）$\pi^* : \mathcal{S} \to \mathcal{A}$ は，以下の Bellman 最適方程式

$$Q^*(s,a) = R(s,a) + \gamma V^*(\mathcal{T}(s,a)) \tag{3}$$

$$V^*(s) = \max_{a \in \mathcal{A}} Q^*(s,a) \tag{4}$$

を解くことで求めることができる．すなわち，最適方策 $\pi^* = \mathrm{argmax}_{a \in \mathcal{A}} Q^*(s,a)$ である．

しかしながら，行動軌跡は意思決定に関わる不確実性を伴い，一貫して最適な軌跡を取るわけではない．Ziebart et al.(2009)[7)] は，意思決定に関わる不確実性を取り込めるよう Bellman 方程式の中の最大値関数を

Log-sum 関数に置き換えることでこれを実現した．すなわち，

$$Q^{\approx}(s,a) = R(s,a) + \gamma V^{\approx}(\mathcal{T}(s,a)) \tag{5}$$

$$V^{\approx}(s) = \log \sum_{a\in\mathcal{A}} \exp\{Q^{\approx}(s,a)\} \tag{6}$$

である．このとき，確率的方策 $\pi(a|s)$ は，

$$\pi(a|s) = \frac{\exp\{Q^{\approx}(s,a)\}}{\sum_{a\in\mathcal{A}} \exp\{Q^{\approx}(s,a)\}} \tag{7}$$

と Logit 関数で求められる．

(3) 最大エントロピー逆強化学習を用いた報酬関数の学習

以下では，Ziebart et al.(2008)[9] によって提案された最大エントロピー逆強化学習について説明する．いま，観測された行動軌跡 ζ_i が状態と行動の組の系列，すなわち，$\zeta_i = \{(s,a)\}$ で表されるとする．また，行動軌跡の集合 $\{\zeta_i\}$ は全て同一の起終点を持つものとする．我々は，与えられた行動軌跡の集合 $\{\zeta_i\}$ に対して，それらの挙動を尤も説明するような報酬関数を推定したい．ここでは，状態 s における報酬関数 $R(s)$ がその状態を特徴づける k 次元の特徴量ベクトル $\mathbf{f}_s$ の線形和で表現されると仮定する．したがって，$R(s) = \theta^\top \mathbf{f}_s$ である．ただし，$\theta \in \mathbb{R}^k$ は報酬関数のパラメータ，$^\top$ は転置の記号である．行動軌跡 ζ_i によって得られる報酬 $R(\zeta_i)$ は，各状態を訪れることで得る報酬の総和で表されるので $R(\zeta_i) = \sum_{s\in\zeta_i} \theta^\top \mathbf{f}_s$ である．

最大エントロピーの原理から観測される行動軌跡 ζ_i が得られる確率 $P(\zeta_i|\theta)$ は

$$P(\zeta_i|\theta) = \frac{\exp\{R(\zeta_i)\}}{\sum_{\zeta\in Z} \exp\{R(\zeta)\}} \tag{8}$$

と求めることができる．ただし，Z は同一の起終点を持つ全ての経路集合である．パラメータ θ は，行動軌跡集合 $\{\zeta_i\}$ の対数尤度 $L(\theta) = \sum_i \log P(\zeta_i|\theta)$ が最大となるように決定するとすれば，

$$\theta^* = \underset{\theta}{\operatorname{argmax}} \sum_i \log P(\zeta_i|\theta) \tag{9}$$

$$= \underset{\theta}{\operatorname{argmax}} \sum_i \left\{ \left(\sum_{s\in\zeta_i} \theta^T \mathbf{f}_s \right) - V^{\approx}(s_{t=0}) \right\} \tag{10}$$

と求めることができる（導出は付録 I を参照のこと）．ただし，$s_{t=0}$ は行動軌跡 ζ_i の初期状態（初期位置）を指す．上式は凸関数であり，勾配法を用いることにより，最適解を求めることができる．

上記尤度の勾配 $\nabla_\theta L$ は以下のように求められる．

$$\nabla_\theta L = \frac{1}{|\zeta|} \sum_i \sum_{s\in\zeta_i} \mathbf{f}_s - \mathbb{E}_{P_\theta(\zeta)}[\mathbf{f}_s] \tag{11}$$

ただし，$|\zeta|$ は観測された軌跡の本数であり，したがって，式 (11) の第 1 項は観測における特徴量の合計の期待値（軌跡あたりの平均値）を表す．一方で，$P_\theta(\zeta)$ はパラメータ θ のもとに行動軌跡 ζ が得られる確率を示し，したがって，第 2 項は MDP のもとに得られる特徴量の合計の期待値を表す．パラメータ推定では，推定中のパラメータ θ を用いて状態 s の報酬を求め，価値反復法により方策を計算し，特徴量の合計の期待値を計算する．計算された期待値と観測における特徴量の期待値の差を計算し，パラメータを更新していくことで最終的に θ^* を得ることができる．

(4) 時空間制約下における行動軌跡生成

本章の (2) 節，(3) 節の方法により，与えられた行動軌跡の集合 $\{\zeta_i\}$ から確率的な方策 $\pi(a|s)$ を得ることができた．本節では，時空間制約を考慮した行動軌跡生成モデルについて説明する．

いま，確率的方策 $\pi(a|s)$ が所与のもと，時刻 t において状態 s に存在する確率を $\alpha_t(s)$ で表し，これを前向き確率と呼ぶ[3]．この前向き確率は前の時刻における確率 $\alpha_t(s')$ から再帰的に計算でき，以下のように表すことができる．

$$\alpha_t(s) = \sum_s P(s|s')\alpha_{t-1}(s') \tag{12}$$

ただし，時刻 0 における存在確率は初期位置に相当し，

$$\alpha_0(s) = \begin{cases} 1 & \text{if} \quad s = s_0 \\ 0 & \text{otherwise} \end{cases} \tag{13}$$

である．また，状態 s' から s への遷移確率は状態 s' における確率的方策 $\pi(a|s')$ と，方策 π により状態 s に遷移する確率 $P(s|s',a)$ を用いて

$$P(s|s') = \sum_a P(s|s',a)\pi(a|s') \tag{14}$$

と表される．

ここで，目的地 s_g と到着必須時間 T_{arvl} が与えられるとする．この目的地は提案モデルとは別に目的地選択モデル等によって与えることができる．これらによる時空間制約の下での遷移確率 $P(s_{t+1} = s|s_t = s', s_{T_{arvl}} = s_g)$ を求めたい．

[3] 本節における考え方は，自然言語処理分野で Forward-Backward アルゴリズム[22]，あるいは Baum-Welch アルゴリズム[23] として知られる考え方を用いたものである．したがって，本稿に出てくるこれらに関連する用語や記法（前向き確率 $\alpha_t(s)$，後向き確率 $\beta_t(s)$）は基本的にこれらの分野で用いられるものを参考にしている．

ここで新たに後向き確率 $\beta_t(s)$ を導入する．この後向き確率 $\beta_t(s)$ は目的地に到達する確率を表すもので，以下の再帰式により計算することができる．

$$\beta_t(s) = \sum_{s'} P(s'|s)\beta_{t+1}(s') \tag{15}$$

ただし，時刻 T_{arvl} においては目的地 s_g に到着していなければならないので

$$\beta_{T_{arvl}}(s) = \begin{cases} 1 & \text{if} \quad s = s_g \\ 0 & \text{otherwise} \end{cases} \tag{16}$$

である．さらに，以下では目的地 s_g への到着は吸収状態であると仮定する．すなわち，

$$P(s|s_g) = \begin{cases} 1 & \text{if} \quad s = s_g \\ 0 & \text{otherwise} \end{cases} \tag{17}$$

である．

時空間制約がない場合の遷移確率 $P(s_{t+1} = s|s_t = s')$ に対し，時空間制約下の遷移確率 $P(s_{t+1} = s|s_t = s', s_{T_{arvl}} = s_g)$ は次の時刻における目的地への到着確率である $\beta_{t+1}(s)$ を重みとして以下のように与えることができる．

$$\begin{aligned} &P(s_{t+1} = s|s_t = s', s_{T_{arvl}} = s_g) \\ &\quad = \frac{\beta_{t+1}(s)P(s_{t+1} = s|s_t = s')}{\sum_s \beta_{t+1}(s)P(s_{t+1} = s|s_t = s')} \\ &\quad = \frac{\beta_{t+1}(s)}{\beta_t(s')}P(s_{t+1} = s|s_t = s') \end{aligned} \tag{18}$$

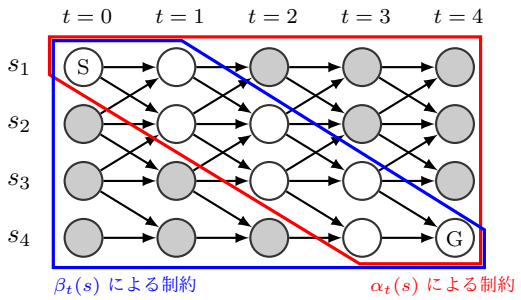
したがって，時空間制約下の遷移確率はもとの遷移確率 $P(s_{t+1} = s|s_t = s')$ に対して，その時点での目的地到達確率 $\beta_t(s')$ と，次時刻での目的地到達確率 $\beta_{t+1}(s)$ の比を用いて計算できる．

また，時空間制約下での時刻 t における状態 s の存在確率 $P(s_t = s|s_{T_{arvl}} = s_g)$ は前向き確率 $\alpha_t(s)$ と後向き確率 $\beta_t(s)$ の積で与えられる．すなわち，

$$P(s_t = s|s_{T_{arvl}} = s_g) = \frac{\alpha_t(s)\beta_t(s)}{Z_{\alpha\beta}} \tag{19}$$

ただし，$Z_{\alpha\beta}$ は規格化定数である．

前向き確率 $\alpha_t(s)$ と後向き確率 $\beta_t(s)$ を同時に考慮することは，**図–2** 中に示される出発側からの制約（赤枠）と到着側からの制約（青枠）を同時に考えることに相当し，これがいわゆる時空間プリズム制約になっている．これらを同時に満たす状態（白抜きノード）が活動可能な領域である．次状態の選択確率は式 (18) で示したように，確率的方策 $\pi(a|s)$ と目的地への到着確率 $\beta_t(s)$ から計算される．



図–2: 提案手法における時空間プリズム制約表現

さらに，時空間制約下における経路の報酬和の期待値は，状態の存在確率と報酬関数を用いて以下のように計算ができる．すなわち，

$$\mathbb{E}_{P_\theta(\zeta|s_{T_{arvl}}=s_g)}[R(\zeta)] = \sum_t \sum_s \frac{\alpha_t(s)\beta_t(s)}{Z_{\alpha\beta}} \cdot R(s) \tag{20}$$

と求めることができる．ただし，$P_\theta(\zeta|s_{T_{arvl}} = s_g)$ は，目的地の時空間制約の下で経路 ζ が得られる確率である．

最後に，提案モデルにおける前向き確率及び後向き確率は，均質な時空間表現の下では行列を用いて簡単に計算が可能である．1 時間ステップ毎に状態遷移が起こる（同状態への遷移を含む）ような状態表現を考える．ここで，以下の $N \times N$ 次元の遷移確率行列 M を定義する．

$$M = \begin{pmatrix} P(s_1|s_1) & P(s_1|s_2) & \cdots & P(s_1|s_N) \\ P(s_2|s_1) & P(s_2|s_2) & \cdots & P(s_2|s_N) \\ \vdots & \vdots & \ddots & \vdots \\ P(s_N|s_1) & P(s_N|s_2) & \cdots & P(s_N|s_N) \end{pmatrix} \tag{21}$$

各要素の定義は式 (14)，(17) に示した通りである．このとき，前向き確率 $\alpha_t(\mathbf{s})$ 及び後向き確率 $\beta_t(\mathbf{s})$ は，

$$\alpha_t(\mathbf{s}) = M^t \alpha_0(\mathbf{s}) \tag{22}$$

$$\beta_t(\mathbf{s}) = \left(M^\top\right)^{T_{arvl}-t} \beta_{T_{arvl}}(\mathbf{s}) \tag{23}$$

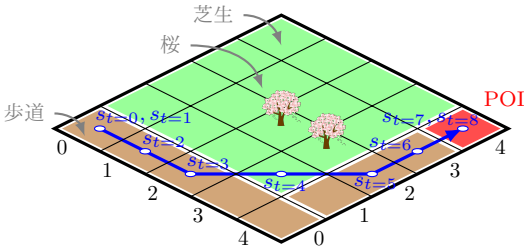
と表すことができる．

4. 数値実験

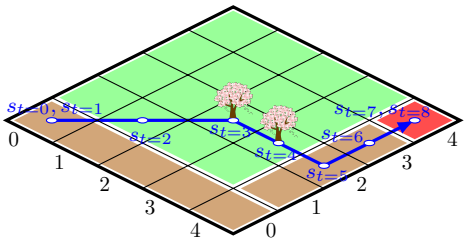
本章では，簡単な問題設定に対する数値実験を通して，提案モデルの特性を明らかにする．

(1) 問題設定

5×5 のグリッドに空間を離散化した環境モデルを用いて実験を行う．観測された行動軌跡として，**図–3** に



(a) ζ_1：軌跡 1



(b) ζ_2：軌跡 2

図–3: 問題設定

示された 2 種類の行動軌跡 ζ_1, ζ_2 がそれぞれ 50%ずつ得られているとする．観測された行動軌跡をもとに報酬関数の推定及び確率的方策の学習を行う．実験環境は，**図–3** に示されているように歩道，芝生，桜，POI（Point of Interests）の 4 つの特徴によって特徴づけられていると仮定する．また，行動空間 $\mathcal{A}(s)$ は 8 方位と滞在の 9 種類と仮定する．ただし，エリアの外への行動は含まれない．

(2) 報酬関数の推定

第 3 章の (3) 節に示した方法を用いて報酬関数の推定を行った．歩道，芝生，POI に関しては状態（グリッド）に該当する特徴量が存在するかどうかをダミー変数として扱った．桜に関しては空間的に離れていても視認することにより，報酬を得ることができると仮定した．桜の特徴量 $f_{cherry}(s)$ は，桜の存在するグリッド s_{cherry} までの最短距離 $\min|d(s, s_{cherry})|$ を用いて

$$f_{cherry}(s) = \exp\{-\min|d(s, s_{cherry})|\} \tag{24}$$

と表現した．また，推定に用いた割引率は $\gamma = 0.9$ である．

パラメータの推定結果を**表–1** に示す．また，報酬関数と状態価値関数の推定結果を**図–4** に示す．図中の座標値 (0,0) のグリッド（左上）が初期位置，座標値 (4,4) のグリッド（右下）が POI の位置に相当する．方策 π は式 (7) に示される通り，行動価値関数 $Q^{\approx}(s,a)$ が高くなるよう行動する．$Q^{\approx}(s,a) = R(s,\pi(s)) + \gamma V^{\approx}(s,\pi(s))$ なので，方策 π に従って行動すると POI のグリッドに向かう軌跡が得られる（**図–6d**）．

表–1: パラメータ推定結果

特徴量	推定値
歩道	6.02
芝生	1.62
桜	6.23
POI	8.14

(3) 時空間制約下における生成軌跡の比較

第 3 章の (4) 節に示した提案モデルを用いた場合の生成軌跡の違いについて比較を行う．今回は所与の目的地として以下の 2 種類：

- 最も状態価値の高いグリッド　座標値 (4,4)
- 最も状態価値の低いグリッド　座標値 (4,0)

また，時間制約として以下の 6 種類：

- 制約なし（$T_{arvl} = \infty$ に相当）
- $T_{arvl} = 4$（最短経路），7，10，15，20

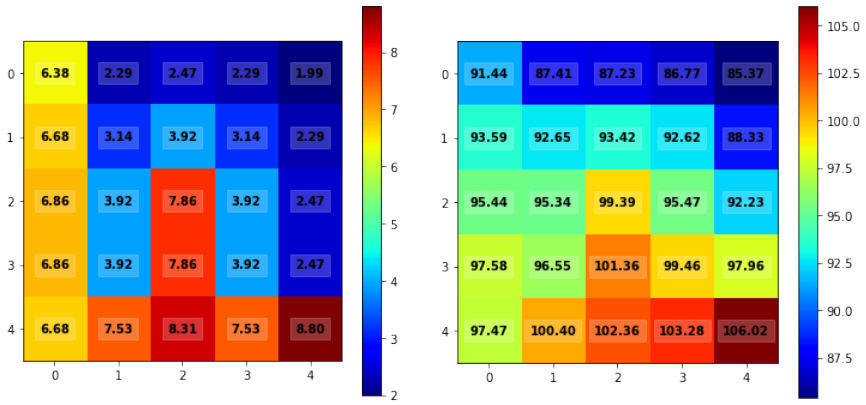
のもとに実験を行った．

まずはじめに，与えられた目的地へ到達する時間の違いを見るために，時刻 t までに目的地に到達する確率を到達率と定義し，時空間制約の有無による到達率の違いを比較する．到着率は到着必須時間が短くなればなるほど立ち上がりが急になるはずであり，最も極端なケース，すなわち到着必須時間が最短経路時間と同じ場合には，到達率は到着必須時間以降に 1 になる階段関数の形になる．一方，到着必須時間が長い場合には，緩やかな立ち上がりになるはずであり，余分にかかる時間が迂回や滞在の行動として費やされた時間に相当する．

実験によって得られた目的地への到達率を**図–5** に示す．**図–5a** のケースでは，状態価値の高いグリッドが目的地であるため，時空間制約がない場合でも $t = 20$ の付近には到達率がほぼ 1 に達した．一方で，時空間制約を考慮した場合には到着必須時間を増やすごとに最短経路を選択する目的地指向型の行動軌跡から，より散策的な行動軌跡に変化していく様子が確認できた．

図–5b のケースでは状態価値の低いグリッドが目的地であるため，時空間制約がない場合には，到達率がほぼ 0 のままであった．一方で，時空間制約を考慮した場合には，**図–5a** のケースと同様に散策的な行動軌跡への変化が確認できた．しかしながら，到着必須時間 $T_{arvl} = 20$ の場合でも $t = 10$ 付近には到達率がほぼ 1 に達し，**図–5a** に比べると目的地指向の行動軌跡となった．

次に，時間制約が行動軌跡の生成に与える影響を可視化し，到着必須時間の増加がどのような行動軌跡の生成につながるかを調べた．**図–6** は，到着必須時間が $T_{arvl} = 4, 7, 10$ のときの行動軌跡と時間制約がない場



(a) 報酬関数　　(b) 状態価値関数

図–4: 推定された報酬関数と状態価値関数

合の行動軌跡である．図中のリンクの太さは状態遷移確率の時間累積値，ノードの大きさは滞在確率の時間累積値を表現している．提案モデルでは循環経路を許容したモデルとなっているため，状態遷移確率の累積値が 1 を超えることがあることに注意されたい．$T_{arvl} = 4$ の場合は，目的地に到達する解が 1 つしか存在しないため，真っ直ぐに目的地を目指した行動軌跡となり，途中に滞在も行われない（**図–6a**）．$T_{arvl} = 7$ の場合は，目的地に到達する解が複数存在し，**図–6b** に示されるように最短経路よりは価値の高い歩道や桜を通る経路が選ばれやすくなっている様子が確認できるが，滞在行動はあまり行われていない．$T_{arvl} = 10$ の場合は，さらに多様な経路が表れており（**図–6b**），制約のない場合の自由な行動軌跡（**図–6d**）に近い行動軌跡生成が行われていることが確認できる．この場合には，目的地に続いて報酬が最も高いグリッド (2,4)（**図–4a**）での滞在が行われており，軌跡の多様性と合わせて散策的な行動軌跡の生成が行われている．

最後に，到着必須時間と経路の報酬和の期待値の関係を**図–7** に示す．どちらの目的地に対しても到着必須時間に対して報酬和の期待値が時間に対して逓減する様子が確認できた．**図–7a** と**図–7b** を比較すると，**図–7a** では，$T_{arvl} = 20$ でもまだ効用の増加が見られるのに対し，**図–7b** では，$T_{arvl} = 10$ 前後で飽和している様子が分かる．このことが原因となり，**図–5b** では目的地指向の行動軌跡になっていると推察される．

5. おわりに

本稿では，Ziebart et al.(2009)[7] によって提案されたデータ駆動型の経路選択行動方策の学習手法を用い，さらに，時空間制約を考慮することで目的地指向の行動軌跡のみならず，散策的な行動軌跡についても生成する手法について提案した．簡易的な数値実験により，任意の目的地に対して時空間制約を考慮した行動軌跡が生成可能であることを確認した．また，到着必須時間が増えると軌跡の多様化と滞在行動が起こり，より散策的な行動軌跡に変化する様子が確認できた．

複雑な環境における歩行者の回遊行動の記述を行うためには，異なるレベルの歩行意思決定の統一化が求められる．提案モデルにより，Tactical なレベルの意思決定の一つである経路選択について散策的な行動も含めた経路の生成が可能となった．今後は活動場所の選択やスケジューリングといった他の Tactical なレベルの意思決定やより上位の Strategic な意思決定を統一的に扱うことが求められる．

付録 I　式 (10) の導出

式 (9) から式 (10) の導出を行う．

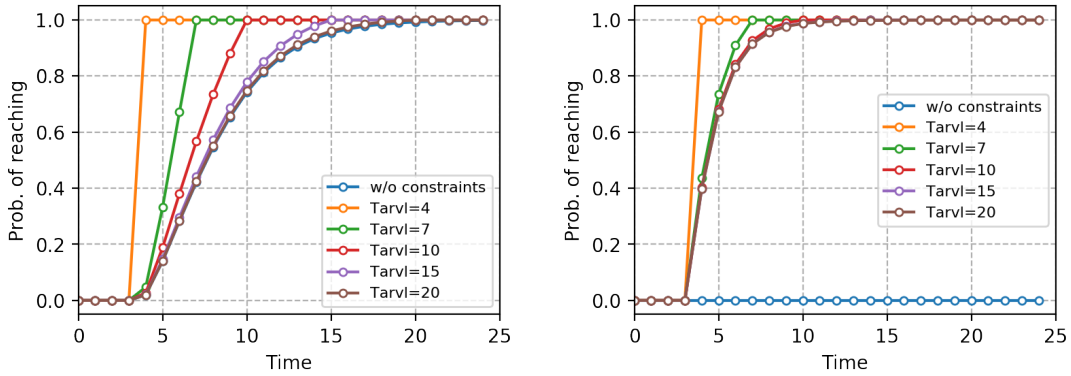
$$\theta^* = \underset{\theta}{\operatorname{argmax}} \sum_i \log P(\zeta_i|\theta) \tag{I.1}$$

$$= \underset{\theta}{\operatorname{argmax}} \sum_i \log \frac{\exp\{R(\zeta_i)\}}{\sum_{\zeta \in Z} \exp\{R(\zeta)\}} \tag{I.2}$$

$$= \underset{\theta}{\operatorname{argmax}} \sum_i \left\{ R(\zeta_i) - \log \sum_{\zeta \in Z} \exp\{R(\zeta)\} \right\} \tag{I.3}$$

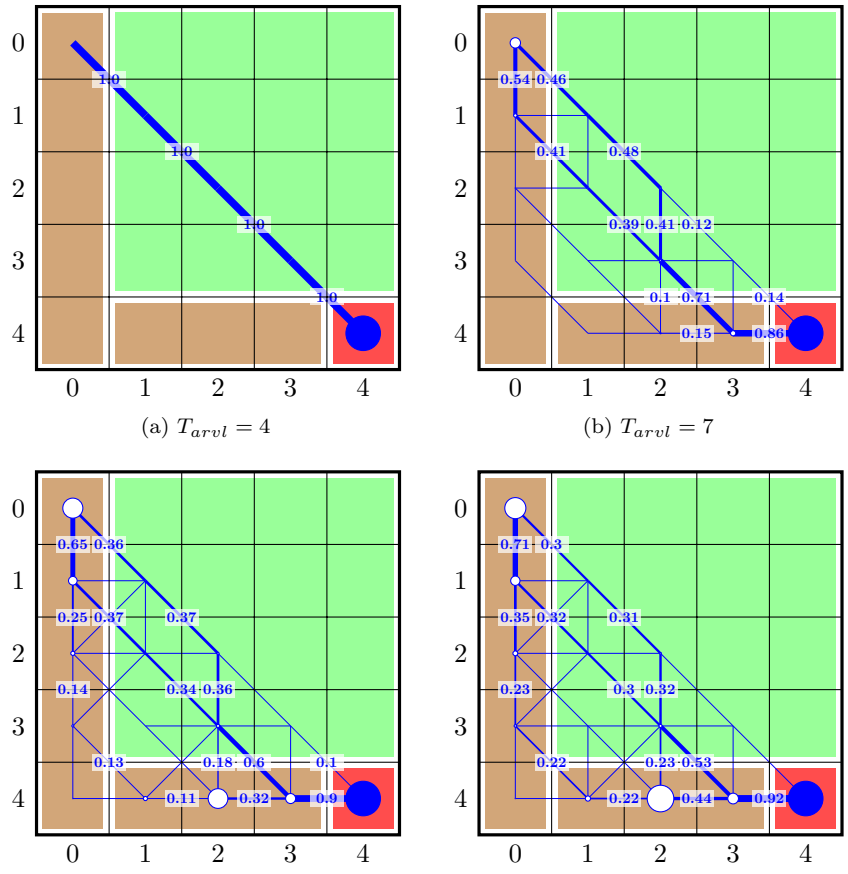
式 (I.3) の第一項は，既出の通り $R(\zeta_i) = \sum_{s \in \zeta_i} \theta^\top \mathbf{f_s}$ である．第二項は，軌跡の報酬の Log-sum 項になっているので，これは正に初期位置における状態価値関数 $V^{\approx}(s_{t=0})$ そのものである．したがって，

$$\theta^* = \underset{\theta}{\operatorname{argmax}} \sum_i \left\{ \left(\sum_{s \in \zeta_i} \theta^T \mathbf{f}_s \right) - V^{\approx}(s_{t=0}) \right\} \tag{I.4}$$



(a) グリッド (4,4)　　(b) グリッド (4,0)

図–5: 与えられた目的地への到達率



(a) $T_{arvl} = 4$　　(b) $T_{arvl} = 7$

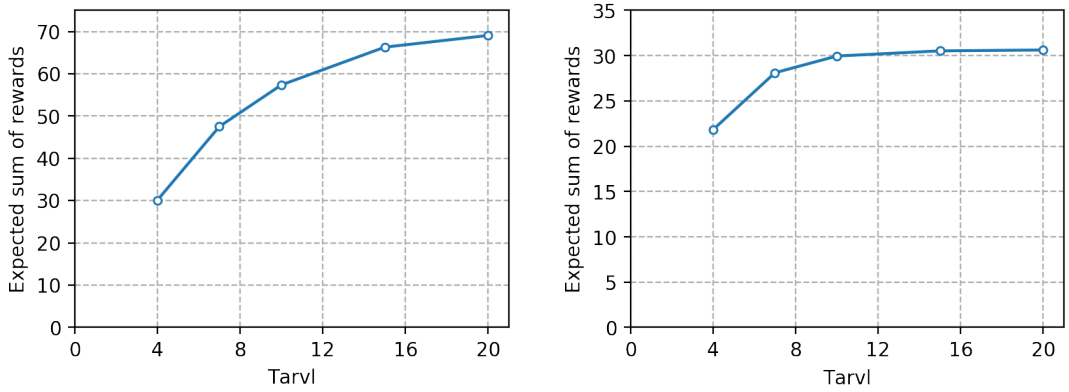
(c) $T_{arvl} = 10$　　(d) 制約なし

図–6: 生成される行動軌跡の比較
（リンクの太さは状態遷移確率の時間累積値，ノードの大きさは滞在確率の時間累積値を表現）

が得られる．

参考文献

1) Hoogendoorn, S. P.: Normative pedestrian flow behavior theory and applications, *LVV rapport, VK 2001.002*, 2001.
2) Helbing, D., Farkas, I., and Vicsek, T.: Simulating dynamical features of escape panic, *Nature*, Vol.407, No.6803, pp.487, 2000.
3) Blue, V. J. and Adler, J. L.: Cellular automata microsimulation for modeling bi-directional pedestrian walkways, *Transportation Research Part B: Methodological*, Vol.35, No.3, pp.293–312, 2001.
4) Antonini, G., Bierlaire, M., and Weber, M.: Discrete choice models of pedestrian walking behavior, *Transportation Research Part B: Methodological*, Vol.40, No.8, pp.667–687, 2006.
5) Hoogendoorn, S. P. and Bovy, P. H.: Pedestrian route-choice and activity scheduling theory and models, *Transportation Research Part B: Methodological*,



(a) グリッド (4,4)　　(b) グリッド (4,0)

図–7: 軌跡の報酬和の期待値

Vol.38, No.2, pp.169–190, 2004.
6) Hoogendoorn, S. P., van Wageningen-Kessels, F., Daamen, W., Duives, D. C., and Sarvi, M.: Continuum theory for pedestrian traffic flow: Local route choice modelling and its implications, *Transportation Research Procedia*, Vol.7, pp.381–397, 2015.
7) 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, *Intelligent Robots and Systems, 2009. IROS 2009. IEEE/RSJ International Conference on*, pp. 3931–3936, IEEE, 2009.
8) Kitani, K. M., Ziebart, B. D., Bagnell, J. A., and Hebert, M.: Activity forecasting, *European Conference on Computer Vision*, pp. 201–214, Springer, 2012.
9) Ziebart, B. D., Maas, A. L., Bagnell, J. A., and Dey, A. K.: Maximum entropy inverse reinforcement learning., *AAAI*, Vol. 8, pp. 1433–1438, Chicago, IL, USA, 2008.
10) Hägerstraand, T.: What about people in regional science?, *Papers in regional science*, Vol.24, No.1, pp.7–24, 1970.
11) Turner, A. and Penn, A.: Encoding natural movement as an agent-based system: an investigation into human pedestrian behaviour in the built environment, *Environment and planning B: Planning and Design*, Vol.29, No.4, pp.473–490, 2002.
12) Penn, A. and Turner, A.: Space syntax based agent simulation, Springer-Verlag, 2002.
13) Robin, T., Antonini, G., Bierlaire, M., and Cruz, J.: Specification, estimation and validation of a pedestrian walking behavior model, *Transportation Research Part B: Methodological*, Vol.43, No.1, pp.36–56, 2009.
14) Asano, M., Iryo, T., and Kuwahara, M.: A pedestrian model considering anticipatory behaviour for capacity evaluation, *Transportation and traffic theory 2009: Golden jubilee*, pp. 559–581, Springer, 2009.
15) Wang, W., Lo, S., Liu, S., and Kuang, H.: Microscopic modeling of pedestrian movement behavior: Interacting with visual attractors in the environment, *Transportation Research Part C: Emerging Technologies*, Vol.44, pp.21–33, 2014.
16) Borgers, A. and Timmermans, H.: A model of pedestrian route choice and demand for retail facilities within inner-city shopping areas, *Geographical analysis*, Vol.18, No.2, pp.115–128, 1986.
17) Russell, S.: Learning agents for uncertain environments, *Proceedings of the eleventh annual conference on Computational learning theory*, pp. 101–103, ACM, 1998.
18) Ng, A. Y., Russell, S. J., et al.: Algorithms for inverse reinforcement learning., *Icml*, pp. 663–670, 2000.
19) Abbeel, P. and Ng, A. Y.: Apprenticeship learning via inverse reinforcement learning, *Proceedings of the twenty-first international conference on Machine learning*, p. 1, ACM, 2004.
20) Fosgerau, M., Frejinger, E., and Karlstrom, A.: A link based network route choice model with unrestricted choice set, *Transportation Research Part B: Methodological*, Vol.56, pp.70–80, 2013.
21) Oyama, Y. and Hato, E.: A discounted recursive logit model for dynamic gridlock network analysis, *Transportation Research Part C: Emerging Technologies*, Vol.85, pp.509–527, 2017.
22) Rabiner, L. R.: A tutorial on hidden markov models and selected applications in speech recognition, *Proceedings of the IEEE*, Vol.77, No.2, pp.257–286, 1989.
23) Baum, L.: An inequality and associated maximization technique in statistical estimation of probabilistic functions of a markov process, *Inequalities*, Vol.3, pp.1–8, 1972.

(2018. 7. 31 受付)

LEARNING REWARD FUNCTION USING INVERSE REINFORCEMENT LEARNING AND PEDESTRIAN ACTIVITY TRAJECTORY GENERATION UNDER SPACE-TIME CONSTRAINTS

Ken HIDAKA, Keiichiro HAYAKAWA, Tomoki NISHI, Tomoki USUI and Toshiyuki YAMAMOTO