歩行者の報酬関数と潜在的な到着時間制約を同時に推定する逆強化学習法

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 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: 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) 人間環境大学人間環境学部環境科学科 教授（〒 444-3505 愛知県岡崎市本宿町上三本松 6-2）
E-mail: t-usui@uhe.ac.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

近年，広場や街路などの歩行者を中心とした公共空間に対する需要の高まりに伴って，歩行者行動の理解に対する重要性が増してきている．歩行者の回遊行動を理解するためには，移動に関わる正の効用を考慮することが必要である．これまで，移動の正の効用と時空間制約の導入により歩行者の迂回や滞留行動を生成する手法が提案されていた．しかしながら，既存手法では到着時間に制約を持つ人をデータに含む場合に，所要時間の短い経路上の地物の報酬関数（効用関数）を高く評価してしまうことが懸念される．そこで，本稿では，人々が潜在的に持っているであろう到着時間制約と，報酬関数を同時に推定する逆強化学習法を提案する．具体的には，移動時間の制約を考慮した確率的選択肢集合形成モデルを応用し，報酬関数と歩行者の潜在的な到着時間の制約を同時に推定する新たな逆強化学習の問題を提案する．また，この問題は EM algorithm により効率的に学習可能であることを示す．さらに，数値実験を通じて提案手法が既存の手法と比べて報酬関数推定のバイアスが少なく，より正確な推定ができることを示す．本手法を用いることで，時間制約を受けているデータを含めた歩行者回遊行動の分析が可能となり，歩行者行動理解の促進が期待される．

Key Words: 歩行者行動モデル，逆強化学習，確率的な選択肢集合形成モデル，到着時間制約，*EM algorithm*

1. はじめに

近年，都市部を中心として広場や街路などの歩行者を中心とした公共空間創出への需要が高まっている[1)]．これに伴い，歩行者行動の理解に対する重要性が増している．魅力的で，賑わいのある公共空間の創出のために，人々にとって何が魅力的で，また価値があるのか，そして，なぜそこで様々な活動が行われるかを理解する必要がある．例えば，人々はしばしば，美しい景色を見るために迂回路を利用したり，ストリートパフォーマンスを見るために立ち止まったりする．

従来，交通分野において移動は不効用と考えられてきた．つまり，人々は目的地に出来るだけ早く到着しようと行動することが仮定されている．しかしながら，先の例に挙げたように，魅力的な場所を人々が回遊するときは，美しい景色やストリートパフォーマンスのために時間を費やす．これらの行動を理解するためには，移動に関わる正の効用を考慮することが必要である．

予め収集したデータから人々の報酬関数（または効用関数）を推定する方法として逆強化学習がある．日高ら[2)]は，Ziebart et al.[3)]によって提案された逆強化学習の枠組みを用いた歩行者行動モデルを拡張し，移動に関する正の効用と時空間制約を導入することで，歩行者の回遊に特徴的な迂回や滞留などの行動を生成する手法を提案した．しかしながら，日高ら[2)]らの手法は，Ziebart et al.[3)]の手法と同様に，得られたデータには時空間制約がないとみなして報酬関数を学習している．したがって，実際のデータから学習する際に，時空間制約を強く受ける人が混在していると，報酬関数の推定結果にバイアスが生じる恐れがある．例えば，急いでいる人は空間上の地物などの価値とは無関係に，移動時間の短い経路を選ぶはずであり，これにより最短経路上の地物の価値が見かけ上，高く推定される可能性がある．

そこで本稿では，人々が潜在的に持っているであろう到着時間制約と，報酬関数を同時に推定することができる逆強化学習法を提案する．具体的には，Thill and Horowitz[4)]によって提案された移動時間制約を考慮した確率的な選択肢集合形成モデルを逆強化学習の枠組みに導入し，各自の到着時間制約及び報酬関数を同時

に推定する新たな逆強化学習の問題を提案する．さらに，この問題は EM algorithm[5] により効率的に学習可能であることを示す．最後に，簡単な数値実験を通じて，従来の手法では報酬関数の推定にバイアスが生じる場合においても，提案手法を用いることでバイアスの少ない，より正確な推定が可能であることを示す．

本稿の構成は以下に示す通りである．続く第 2 章にて，関連研究のレビューをした後，第 3 章にて最大エントロピー逆強化学習と，時空間制約の下での移動行動軌跡の生成法について概説する．そして，第 4 章にて提案する報酬関数と到着時間制約の同時推定方法について説明を行う．第 5 章では，数値実験について説明を行い，最後に第 6 章で結論を述べる．

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，逆強化学習，確率的な選択肢形成モデル，EM algorithm を用いた推定法について取り上げる．

(1) 逆強化学習

収集したエキスパートの行動系列と環境モデルを所与として報酬関数を推定する代表的な手法として逆強化学習がある[6),7),8)]．なかでも最もよく知られた逆強化学習の手法の一つとして，最大エントロピー逆強化学習がある[8)]．この手法は交通分野でよく知られた Recursive logit（RL）モデル[9)] と基本的に同じモデルであるが，RL モデルでは移動の不効用のみが扱えるのに対し，最大エントロピー逆強化学習は，目的地をベースとしたモデルでないこと，割引率の導入により状態価値の収束を保証していることにより，正の効用を扱うことができる．そのため，歩行者行動への適用として，移動の不効用のみを取り扱った研究[3),10)] の他に，正の効用を扱った研究[2)] も報告されている．

日高ら[2)] は，最大エントロピー逆強化学習の枠組みに，正の効用と時空間制約を導入することにより，歩行者の回遊行動の特徴的な行動である滞留や迂回が表現できることを示した．しかしながら，日高らの方法では，時空間制約をどのようにして与えるかが明らかにされていない．また，推定は過去の研究[3),10)] と同様に得られたデータには時空間制約がないとみなして報酬関数を学習しており，その結果，制約のあるデータを含む場合には，報酬関数の推定にバイアスが生じる可能性がある．それに対し，提案手法では到着時間の制約を考慮して報酬関数の推定を行うため，到着時間の制約を持っているデータも扱うことが可能になる．

(2) 確率的な選択肢集合形成モデル

選択肢集合の誤った指定によってパラメータ推定の結果にバイアスが生じることが過去の研究によって知られている[11),12),13)]．これに対し，常に同じ選択肢集合から行動選択を行うのではなく，何らかの制約によって確率的に変化する選択肢集合から行動を選択するモデルが知られている．Manski[11)] は，確率的な選択肢形成過程を明示的に組み込んだ 2 段階の意思決定プロセスモデルを提案した．確率的な選択肢形成を組み込んだモデルは，交通手段選択[14)]，目的地選択[4),15),16)]，経路選択[17)] など様々な事例へ適用されている．

Manski[11)] のモデルは，選択肢集合の全集合（冪集合）について考慮する必要があり，目的地選択などの選択肢の多い状況においては計算の困難性を伴う．こうした問題に対し，選択肢集合の形成過程で様々な制約を考慮することで現実の問題への適用を目指した研究がある．各選択肢が選択肢集合に含まれる確率に独立性を仮定したモデル[18)] や，Hagärstrand[19)] によって提唱された時空間プリズム制約を取り入れたものがある[4),20)]．時空間プリズム制約は，その行動論的な基盤のためにも有力なアプローチである．Thill and Horowitz[4)] は，選択肢集合が個人の時間予算の制約の中で定義されることを明示的に表現した選択肢–目的地選択（Nested Choice–Set Destination Choice; NCS–DC）モデルを提案した．また，潜在クラスモデル[21),22),23),24),25)] などで一般的に用いられるノンパラメトリックな混合分布モデルで離散近似（ANCS–DC モデルと呼ばれる）し，準ニュートン法によって尤度関数を直接最大化する方法で推定を行った．結果として，一般的な多項ロジットモデルよりもわずかではあるが優れていると結論づけている．

これに対し本稿では，Thill and Horowitz[4)] の時間制約を考慮した確率的選択肢形成モデルを応用し，報酬関数と歩行者の到着時間制約の同時推定を行うための新たな逆強化学習法を提案する．

(3) 推定法（EM algorithm）

モデルの推定法に関しては，選択肢集合形成モデルの文脈ではないものの Bhat[24)] が，メンバシップ関数の概念を取り入れた潜在クラスモデル[22)] に対する推定の安定化を目的として，EM algorithm と直接最大化推定を組み合わせたハイブリッド推定法を提案している．また，Train[26)] は，離散選択モデルの混合分布のノンパラメトリックな推定法として EM algorithm を用いた 3 種類の推定法を提案している．

逆強化学習の文脈の中で EM algorithm を用いて潜在的な変数を扱った研究としては，軌跡が部分的に観測されない状況を潜在変数を用いて表現し，非線形非凸の問題を解くために EM algorithm を用いたもの[27),28),29)]

や，より現実的で複雑な行動を異なる報酬関数と遷移で表現し，その学習に EM alogorithm を用いたもの[30]がある．逆強化学習の推定に EM algorithm を用いている点は，既存研究と同様であるが選択肢集合形成過程を導入した逆強化学習手法は存在しない．

3. 時空間制約下における行動軌跡生成

提案手法を説明する前に，本章では日高ら[2]による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．以降の節では，モデルを理解するための準備として Markov Decision Process（MDP）による確率的な方策の算出方法，最大エントロピー逆強化学習による報酬関数の推定について説明を行った後，時空間制約下における歩行者の行動軌跡生成について説明する．

(1) 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．

$$V(s)=\mathbb{E}\left[\sum_{k=0}^{\infty}\gamma^k R(s_{t+k+1},a_{t+k+1})\middle|s_t=s\right], \quad (1)$$

$$Q(s,a)=\mathbb{E}\left[\sum_{k=0}^{\infty}\gamma^k R(s_{t+k+1},a_{t+k+1})\middle|s_t=s,a_t=a\right], \quad (2)$$

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

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

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

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．

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)[3] 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，

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

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

である．このとき，確率的方策 $\pi(a|s)$ は，

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

と Logit 関数で求められる．

(2) 最大エントロピー逆強化学習を用いた報酬関数の学習

逆強化学習は $<\mathcal{S},\mathcal{A},\mathcal{T}>$ 及び観測された行動軌跡 ζ から報酬関数 $\mathcal{R}$ を推定する手法である．報酬関数 $\mathcal{R}$ は対数尤度を最大化する，すなわち $\mathrm{argmax}_R\sum_i P(\zeta_i|R)$ を満たすように求められる．最大エントロピー原理[31]に基づけば，行動軌跡 ζ_i が得られる確率 $P(\zeta_i|R)$ は

$$P(\zeta_i|R)=\frac{\exp\{R(\zeta_i)\}}{\sum_{\zeta\in\Xi}\exp\{R(\zeta)\}} \quad (8)$$

と求められる．ただし，Ξ は行動軌跡の全集合である．Ziebart et al.[8] は，報酬関数の線形性 $R:=\sum_{s\in\zeta_i}\theta^\top\mathbf{f}_s$ を仮定することで効率的に報酬関数が学習可能であることを示した．ただし，$\mathbf{f}_s$ は状態 s を特徴づける k 次元の特徴量ベクトル，$\theta\in\mathbb{R}^k$ は特徴量の重みを表すパラメータ，$^\top$ は転置の記号である．

したがって，最適なパラメータ θ^* は，対数尤度最大となるよう決定すれば良いので

$$\theta^*=\mathrm{argmax}_\theta\sum_i\log P(\zeta_i|\theta) \quad (9)$$

$$=\mathrm{argmax}_\theta\sum_i\left\{\left(\sum_{s\in\zeta_i}\theta^T\mathbf{f}_s\right)-V^{\approx}(s_{init})\right\} \quad (10)$$

と求めることができる（導出は日高ら[2]を参照のこと）．ただし，s_{init} は行動軌跡 ζ_i の初期状態（初期位置）を指す．

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

$$\nabla_\theta L=\sum_i\left\{\mathbf{f}_{\zeta_i}-\mathbb{E}_{P_\theta(\zeta)}[\mathbf{f}_\zeta]\right\} \quad (11)$$

式 (11) の第 1 項は観測における特徴量の合計である．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である．θ^* は勾配ベースの最適化手法を用いて計算が可能である．詳細は Ziebart et al.[8] に詳しい．

(3) 時空間制約下における行動軌跡生成

本章の (1) 節，(2) 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，日高ら[2]によって提案された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，初期位置 $s_0=s_{init}$ が与えられた下での状態 s_t の存在確率分布（前向き確率と呼ぶ）を $\alpha(s_t)\equiv$

$P(s_t|s_0)$ と表す．また，状態 s_t から到着時間制約 τ に目的位置 s_{goal} に到着する確率（後向き確率と呼ぶ）を $\beta(s_t) \equiv P(s_\tau|s_t)$ と表す．$\alpha(s_t)$，$\beta(s_t)$ は再帰的な関係を持つことが以下の式から確認できる．

$$\alpha(s_t) = \sum_{s_{t-1}} P(s_t|s_{t-1})P(s_{t-1}|s_0) \tag{12}$$

$$= \sum_{s_{t-1}} P(s_t|s_{t-1})\alpha(s_{t-1}) \tag{13}$$

$$\beta(s_t) = \sum_{s_{t+1}} P(s_\tau|s_{t+1})P(s_{t+1}|s_t) \tag{14}$$

$$= \sum_{s_{t+1}} P(s_{t+1}|s_t)\beta(s_{t+1}) \tag{15}$$

ただし， $\alpha(s_0), \beta(s_\tau)$ はそれぞれ初期位置 s_{init} と目的位置 s_{goal} に対応しており，

$$\alpha(s_0) = \begin{cases} 1 & \text{if} \quad s_0 = s_{init} \\ 0 & \text{otherwise} \end{cases}, \tag{16}$$

$$\beta(s_\tau) = \begin{cases} 1 & \text{if} \quad s_\tau = s_{goal} \\ 0 & \text{otherwise} \end{cases}, \tag{17}$$

である．また，遷移確率 $P(s_{t+1}|s_t)$ は

$$P(s_{t+1}|s_t) = \sum_{a_t} P(s_{t+1}|s_t, a_t)\pi(a_t|s_t) \tag{18}$$

で与えられる．

時空間制約下における遷移確率 $P(s_{t+1}|s_t, s_\tau)$ は Beyes の定理から計算できて

$$P(s_{t+1}|s_t, s_\tau) = \frac{\beta(s_{t+1})}{\beta(s_t)} P(s_{t+1}|s_t). \tag{19}$$

と後向き確率 $\beta(s_t)$ を用いて計算することができる．一方で，時空間制約下の時刻 t における状態 s_t の存在確率分布 $P(s_t|s_0, s_\tau)$ は

$$P(s_t|s_0, s_\tau) = \frac{\alpha(s_t)\beta(s_t)}{\beta(s_0)} \tag{20}$$

と前向き確率 $\alpha(s_t)$ と後向き確率 $\beta(s_t)$ の積で求めることができる．

式（19）を用いることで，時空間制約下における行動軌跡を逐次的に生成することが可能である．

4. 到着時間制約と報酬関数の同時推定モデル

前章では，最大エントロピー逆強化学習の枠組みに，時空間制約を導入した場合の行動軌跡の生成法について述べた．この方法では，最大エントロピー逆強化学習で報酬関数を推定した後に，目的位置 s_{goal} と到着時間の制約 τ を与えることで，時空間制約の下での行動軌跡を算出する．しかしながら，報酬関数の学習では時空間制約がないことが暗に仮定されており，時空間制約を含むデータからの学習はできない．

そこで、本章では移動時間制約を考慮した確率的な選択肢形成モデルの考え方を導入することで，潜在的な到着時間の制約と報酬関数を同時に推定する逆強化学習法について詳説する．

(1) 移動時間の制約を考慮した確率的な選択肢形成モデル

Manski[11] によって提案された確率的な選択肢形成モデルは以下の式で表すことができる．

$$P(j) = \sum_{C \in G} P(j|C)P(C) \tag{21}$$

ここで，$P(j)$ は選択肢 j の選択確率，$P(j|C)$ は選択肢集合 C が与えられた下での選択肢 j の選択確率，$P(C)$ は選択肢集合 C が選択される確率である．また，G は考えうる全ての選択肢集合の集合（選択肢集合の冪集合）である．

Thill and Horowitz[4] は，移動時間のみが選択肢集合の決定に寄与する要因であるという仮定の下に以下の定式化を行った．

$$P(j) = \int_{t=0}^{\infty} P(j|C_T) dP_T(t;\phi) \tag{22}$$

ここで，$P_T(t;\theta)$ は，移動時間しきい値 T に関する累積分布関数であり，パラメータ ϕ によって特徴づけられる．また，選択肢集合 C_T は，移動時間がしきい値 T 以下であるという条件を満たす目的地の集合を表す．

一方，本稿で扱うモデルでは離散化された時間ごとに，その制約を満たす選択肢集合を考える．すなわち，

$$P(j) = \sum_{\tau} P(j|C_\tau)p(\tau;\phi) \tag{23}$$

である．ただし，ここでの C_τ は到着時間制約 τ を満たす軌跡の集合（時刻 τ において目的位置にいるという制約），$p(\tau;\phi)$ は到着時間制約に関する確率分布である．

Thill and Horowitz[4] のモデルは，移動時間しきい値の範囲内にあるもの全てが選択肢集合となった．一方，本稿で提案するモデルは，選択肢集合を同一の到着時間のもののみから構成されるものとする．この到着時間制約の分布が，特定の確率分布に従うという仮定の下に定式化を行う．以降の節では，提案モデルについて詳細に説明を行う．

(2) 提案モデル

いま，N 人の移動軌跡の集合 $\boldsymbol{\zeta} = \{\zeta_1, \zeta_2, \cdots, \zeta_N\}$ が観測されているとする．一方，到着時間の制約 $\mathbf{z} = \{\mathbf{z}_1, \mathbf{z}_2, \cdots, \mathbf{z}_N\}$ は観測されないため潜在変数として扱う．個人 n が持つ到着時間の制約を表す潜在変数のベクトル $\mathbf{z}_n$ は要素 $z_{n\tau}$ のうち一つのみが 1 で，残りの要素が全て 0 である 2 値変数で構成される．すなわち，

$$\mathbf{z}_n = (z_{n1}, z_{n2}, \cdots, z_{n\tau}, \cdots)^\top \tag{24}$$

$$\sum_{\tau=1}^{\infty} z_{n\tau} = 1 \tag{25}$$

$$z_{n\tau} = \{0, 1\} \quad \forall\tau \tag{26}$$

である．

観測された移動軌跡の集合が得られる確率 $p(\boldsymbol{\zeta})$ は，式（23）の関係より，到着時間の制約に関する確率 $p(\mathbf{z})$ と到着時間制約の下での軌跡 ζ の選択確率 $p(\boldsymbol{\zeta}|\mathbf{z})$ を用いて表すことができる．すなわち，

$$p(\boldsymbol{\zeta}) = \sum_{\mathbf{z}} p(\boldsymbol{\zeta}|\mathbf{z}) p(\mathbf{z}) \tag{27}$$

である．

個人の到着時間制約 $\mathbf{z}_n$ の事前分布 $p(\mathbf{z}_n)$ は以下のように表されるとする．

$$p(\mathbf{z}_n) = \prod_{\tau=\tau_0}^{\infty} p(\tau)^{z_{n\tau}} \tag{28}$$

ここで，$p(\tau)$ は到着時間制約が従う確率分布を表す．ここでは，最短所要時間以上の定義域で定義される離散確率分布として，以下の式で表される負の二項分布を考える[1]．

$$p(\tau) = \binom{\tau - 1}{\tau_0 - 1} \cdot \mu^{\tau_0} (1 - \mu)^{\tau - \tau_0} \tag{29}$$

ただし，τ_0 は起終点間の最短所要時間を表し，初期位置と目的位置から決まるものである．したがって，ここでのパラメータは μ のみである．このパラメータ μ は，試行の成功確率に相当するパラメータである．

また，移動軌跡 ζ_n の $\mathbf{z}_n$ での条件付き分布は，

$$p(\zeta_n|\mathbf{z}_n) = \prod_{\tau=\tau_0}^{\infty} \left\{ \frac{\exp\{R(\zeta_n)\}}{\sum_{\zeta\in\Xi_\tau} \exp\{R(\zeta)\}} \right\}^{z_{n\tau}} \tag{30}$$

で与えられると仮定する．ただし，Ξ_τ は時間制約 τ を満たす移動軌跡の集合である．観測される移動軌跡 ζ_n は軌跡の報酬和に基づいた Logit 選択により行われる（式 (8)）．また，時間制約を満たす移動軌跡の中からの選択は，前章で示した方法を用いることにより，制約を満たす軌跡の集合の列挙なしに Logit の選択確率と整合的な選択確率を求めることができる（式 (19)）．

いま求めたい尤度関数 $p(\boldsymbol{\zeta}, \mathbf{z}|\mu, \theta)$ は，上の定義より

$$p(\boldsymbol{\zeta}, \mathbf{z}|\mu, \theta) = p(\boldsymbol{\zeta}|\mathbf{z}, \theta) p(\mathbf{z}|\mu) \tag{31}$$

と表すことができる．したがって，対数尤度関数は以下の式で表すことができる．

$$\ln p(\boldsymbol{\zeta}, \mathbf{z}|\mu, \theta) = \ln \left\{ \prod_{n=1}^{N} \prod_{\tau=\tau_0}^{\infty} p(\tau|\mu)^{z_{n\tau}} \left\{ \frac{\exp\{\theta^T \mathbf{f}_\zeta\}}{\sum_{\zeta\in\Xi_\tau} \exp\{\theta^T \mathbf{f}_\zeta\}} \right\}^{z_{n\tau}} \right\} \tag{32}$$

$$= \sum_{n=1}^{N} \sum_{\tau=\tau_0}^{\infty} z_{n\tau} \left(\ln p(\tau|\mu) + \ln \frac{\exp\{\theta^T \mathbf{f}_\zeta\}}{\sum_{\zeta\in\Xi_\tau} \exp\{\theta^T \mathbf{f}_\zeta\}} \right) \tag{33}$$

$\mathbf{z}$ の事後分布 $p(\mathbf{z}|\boldsymbol{\zeta}, \mu, \theta)$ は，

$$p(\mathbf{z}|\boldsymbol{\zeta}, \mu, \theta) = \frac{p(\boldsymbol{\zeta}, \mathbf{z}|\mu, \theta)}{p(\boldsymbol{\zeta}|\mu, \theta)} \tag{34}$$

$$= \frac{\prod_{n=1}^{N} \prod_{\tau=\tau_0}^{\infty} p(\tau|\mu)^{z_{n\tau}} \left\{ \frac{\exp\{\theta^T \mathbf{f}_\zeta\}}{\sum_{\zeta\in\Xi_\tau} \exp\{\theta^T \mathbf{f}_\zeta\}} \right\}^{z_{n\tau}}}{\sum_{\mathbf{z}_n} \prod_{n=1}^{N} \prod_{\tau'=\tau_0}^{\infty} p(\tau'|\mu)^{z_{n\tau'}} \left\{ \frac{\exp\{\theta^T \mathbf{f}_\zeta\}}{\sum_{\zeta\in\Xi_{\tau'}} \exp\{\theta^T \mathbf{f}_\zeta\}} \right\}^{z_{n\tau'}}} \tag{35}$$

と表すことができる．したがって，$\mathbf{z}$ の事後分布による $z_{n\tau}$ の期待値 $\mathbb{E}_{\mathbf{z}|\boldsymbol{\zeta}}[z_{n\tau}]$ は

$$\mathbb{E}_{\mathbf{z}|\boldsymbol{\zeta}}[z_{n\tau}] = \sum_{\mathbf{z}_n} z_{n\tau} p(z_{n\tau}|\zeta_n, \mu, \theta) \tag{36}$$

$$= \frac{p(\tau|\mu) \left\{ \frac{\exp\{\theta^T \mathbf{f}_\zeta\}}{\sum_{\zeta\in\Xi_\tau} \exp\{\theta^T \mathbf{f}_\zeta\}} \right\}^{z_{n\tau}}}{\sum_{\tau=\tau_0}^{\infty} p(\tau|\mu) \left\{ \frac{\exp\{\theta^T \mathbf{f}_\zeta\}}{\sum_{\zeta\in\Xi_\tau} \exp\{\theta^T \mathbf{f}_\zeta\}} \right\}^{z_{n\tau}}} \tag{37}$$

$$\equiv \gamma(z_{n\tau}) \tag{38}$$

となる（E ステップ）．ここで定義した $\gamma(z_{n\tau})$ は，各到着時間 τ が移動軌跡 ζ_n の観測を説明する度合いを表す，いわゆる負担率（responsibility）である．

上記で計算した $\mathbf{z}$ の事後分布を用いて完全データの対数尤度の期待値を求める．

[1] 以下の定式化は，$p(\tau)$ を負の二項分布に限るものではなく，適切な確率分布を選択すれば良い．

$$\mathbb{E}_{\mathbf{z}|\boldsymbol{\zeta}}[\ln p(\zeta,\mathbf{z}|\mu,\theta)]$$
$$= \mathbb{E}_{\mathbf{z}|\boldsymbol{\zeta}}\left[\sum_{n=1}^{N}\sum_{\tau=\tau_0}^{\infty} z_{n\tau}\left(\ln p(\tau|\mu) + \ln\frac{\exp\left\{\theta^T\mathbf{f}_\zeta\right\}}{\sum_{\zeta\in\Xi_\tau}\exp\left\{\theta^T\mathbf{f}_\zeta\right\}}\right)\right] \quad (39)$$
$$= \sum_{n=1}^{N}\sum_{\tau=\tau_0}^{\infty}\gamma(z_{n\tau})\left(\ln p(\tau|\mu) + \ln\frac{\exp\left\{\theta^T\mathbf{f}_\zeta\right\}}{\sum_{\zeta\in\Xi_\tau}\exp\left\{\theta^T\mathbf{f}_\zeta\right\}}\right) \quad (40)$$
$$= Q(\mu,\theta) \quad (41)$$

これが，Q 関数に相当するものである．この Q 関数を用いてパラメータの更新を行う（M ステップ）．

μ の最尤解は，Q 関数の μ 微分が 0 になる点として求めることができるので，

$$\frac{\partial Q}{\partial\mu} = \sum_{n=1}^{N}\sum_{\tau=\tau_0}^{\infty}\gamma(z_{n\tau})\left(\tau_0\frac{1}{\mu} - \frac{\tau-\tau_0}{1-\mu}\right) = 0 \quad (42)$$

である．したがって，

$$\sum_{n=1}^{N}\sum_{\tau=\tau_0}^{\infty}\gamma(z_{n\tau})\left(\tau_0-\mu\tau\right) = 0 \quad (43)$$

これを μ について解けば，

$$\mu = \frac{\sum_{n=1}^{N}\sum_{\tau=\tau_0}^{\infty}\tau_0\gamma(z_{n\tau})}{\sum_{n=1}^{N}\sum_{\tau=\tau_0}^{\infty}\tau\gamma(z_{n\tau})} \quad (44)$$
$$= \frac{\tau_0 N}{\sum_{\tau=\tau_0}^{\infty}\tau N_\tau} \quad (45)$$
$$= \frac{\tau_0}{\bar{\tau}} \quad (46)$$

と求めることができる．ただし，ここで $N = \sum_{n=1}^{N}\sum_{\tau=\tau_0}^{\infty}\gamma(z_{n\tau})$，$N_\tau = \sum_{n=1}^{N}\gamma(z_{n\tau})$ を用いた．また，$\bar{\tau}$ は到着時間制約の平均値であり，$\bar{\tau} = \sum_{\tau=\tau_0}^{\infty}\tau N_\tau/N$ で求められる．到着時間制約の平均値を平均所要時間と見做せば，μ の逆数は最短所要時間に対する平均所要時間の比になり，いわゆる平均迂回因子[16),32),33)]（Mean detour factors）に相当するものと解釈することができる．

また，θ の最尤解も同様に以下の計算によって求めることができる．

$$\frac{\partial Q}{\partial\theta} = \sum_{n=1}^{N}\sum_{\tau=\tau_0}^{\infty}\gamma(z_{n\tau})\frac{\partial}{\partial\theta}\left\{\ln\frac{\exp\left\{\theta^T\mathbf{f}_\zeta\right\}}{\sum_{\zeta\in\Xi_\tau}\exp\left\{\theta^T\mathbf{f}_\zeta\right\}}\right\} = 0 \quad (47)$$

$\{\cdot\}$ 内の微分は最大エントロピー逆強化学習のパラメータ推定と同じものである．したがって，式 (11) の勾配と同様に計算すると

$$\sum_{n=1}^{N}\sum_{\tau=\tau_0}^{\infty}\gamma(z_{n\tau})\left\{\mathbf{f}_{\zeta_n} - \mathbb{E}_{p_\theta(\zeta|\tau)}[\mathbf{f}_\zeta]\right\} = 0. \quad (48)$$

が得られる．ただし，$p_\theta(\zeta|\tau)$ は到着時間 τ の制約の下での移動軌跡 ζ の生成確率である．上式は，一般的な最大エントロピー逆強化学習と同様に勾配法などを用いることで推定が可能である．モデルから期待される特徴量の平均 $\mathbb{E}_{p_\theta(\zeta|\tau)}[\mathbf{f}_\zeta]$ は τ の時間制約の下の各状態の訪問数の期待値（State Visitation Frequency; SVF）と各状態の特徴量の内積により計算できる．時間制約下における各状態の訪問数の期待値 SVF_τ は，各時刻における状態ごとの存在確率を時刻について和を取ることで求めることができる．時間制約下における存在確率は前述の式 (20) で表されるので，結局のところ SVF_τ は

$$SVF_\tau = \sum_{t=0}^{\tau}\frac{\alpha(s_t)\beta_\tau(s_t)}{\beta_\tau(s_0)} \quad (49)$$

と計算することができる．

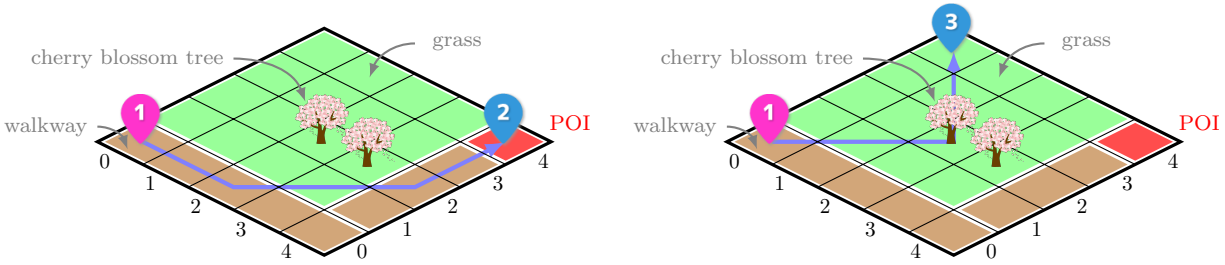
一般的な最大エントロピー逆強化学習との違いは，負担率 $\gamma(z_{n\tau})$ による重みづけがされていることにある．例えば，$\gamma(z_{n\tau_0}) = 1$ で，かつ最短所要時間経路が 1 種類しか存在しない場合，パラメータ θ によらず，観測とモデルの特徴量の期待値は必ず一致する．考えうる経路選択肢数は到着時刻制約 τ と共に増加する．したがって，到着時間の制約を強く受けているデータは経路の選択肢自体が少ないことにより，観測とモデルの期待値が一致しやすい．一方で，到着時間の制約をあまり受けていないデータでは経路の選択肢が膨大になったことにより，パラメータ値がより正確でないと観測とモデルの期待値が一致しづらくなる．その結果として，到着時間の制約をあまり受けていないデータの方が報酬関数の推定に対する寄与が大きくなる．つまり，場所や地物の価値（魅力）を計算する際には，制約を強く受けていると思われる人よりも，自由に歩行（散策）を行っていると思われる人を重視されることを示している．これにより，急いでいる人の存在によって報酬関数の推定にバイアスが生じる場合にも，バイアスの少ない，より正確な推定が可能となる．

5. 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じて提案モデルの検証を行う．

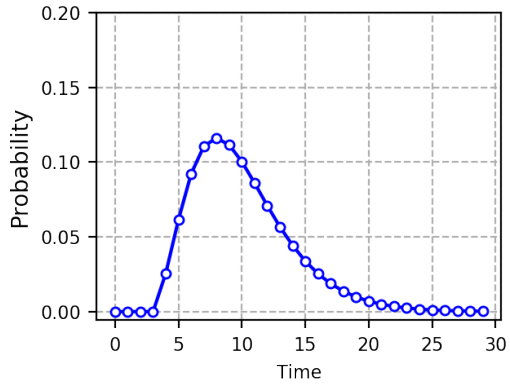
(1) 問題設定

実験は，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行う．実験環境は，図 1 に示されるよう



(a) 実験 1（OD:1 → 2）　(b) 実験 2（OD:1 → 3）

図–1: 問題設定



図–2: 設定した到着時間制約分布（$\tau_0 = 4, \mu = 0.4$）

に，歩道，芝生，桜，PoI（Point of Interests）の 4 つの特徴により特徴づけられているとする．また，行動空間 $\mathcal{A}(s)$ は 8 方位と滞在の 9 種類とする．ただし，エリアの外への行動は含まれない．

歩道，芝生，桜，PoI の特徴量をそれぞれ f_{ww}，f_{gr}，f_{ch}，f_{PoI} と表す．このうち，歩道，芝生，PoI は特徴量の存在可否をダミー変数で表す特徴量である．すなわち，$f_{ww}, f_{gr}, f_{PoI} \in \{0, 1\}$ である．また，桜は空間的に離れていても視認できることから，桜の存在するグリッド s_{ch} までの最短距離 $\min |d(s, s_{ch})|$ を用いて $f_{ch} = \exp\{-\min |d(s, s_{ch})|\}$ で表されるとする．

歩道，芝生，PoI の間には排他的な関係がある．すなわち，$f_{ww} + f_{gr} + f_{PoI} = 1$ の関係が成り立っている．この従属関係から，推定すべきパラメータはこの 3 つの特徴量の中の 2 つである．ここでは，歩道と PoI を推定するパラメータとした．

パラメータ θ の設定値は，$[\theta_{ww}, \theta_{ch}, \theta_{PoI}]^\top = [2, 2, 4]^\top$ とした．ここで，θ_{ww}，θ_{ch}，θ_{PoI} はそれぞれ歩道，桜，PoI に関するパラメータである．θ_{ww}，θ_{PoI} は $\theta_{gr} = 0$ からの相対値であり，したがって，生成されるデータは芝生よりも歩道や PoI を通りやすい傾向にあるような設定としている．

実験では，2 種類の起終点（Origin-Destination；OD）を設定し，上記で設定したパラメータに従ってそれぞれ 1,000 件ずつのデータを生成する．データの生成は，まず到着時間制約の分布に従って，到着時間の制約値をサンプリングし，その到着時間を制約として軌跡を生成する．生成された 1,000 件の移動軌跡データを入力としてパラメータ θ, μ の設定値が再現できるか検証を行う．設定した OD は図 1a，1b に示す通りである．いずれの OD も最短所要時間 $\tau_0 = 4$ である．また，到着時間制約の分布のパラメータ μ は 0.4 に設定した（図 2）．したがって，到着時間制約の平均値は $\tau = \tau_0/\mu = 10$ となる．また，データの観測時間長は全て 30 で統一し，到着時間制約以降のデータは，目的位置で留まるものとした．これは，実際の GPS などの軌跡の観測でいえば，移動と活動が区別されないことに相当し，目的位置でそのまま活動が行われることに対応している．

(2) 推定結果

以下では，2 種類の実験について，それぞれパラメータを推定した結果について示す．推定では，負担率の初期値として $\gamma_{n,30} = 1$ を用いた．したがって，初期値では時間制約がない[2]自由な回遊行動を仮定しており，EM algorithm の最初の反復計算で計算されるパラメータ θ は，一般的な最大エントロピー逆強化学習で推定されたパラメータに相当する．

a) 実験 1

まずはじめに，実験 1 でのパラメータ推定の結果を図 3 に示す．図 3a は，反復回数と θ の推定値，図 3b は，反復回数と μ の推定値を示している．どちらの推定値も約 10 回前後の反復で，設定値 $[\theta_{ww}, \theta_{ch}, \theta_{PoI}]^\top = [2, 2, 4]^\top$，$\mu = 0.4$ に収束している様子が確認できた．一方で，1 回目の反復における結果をみると全てのパラメータで過大に推計されていることが分かる．これは，従来の方法[2)] では目的位置での滞在を含めて全てのデータが

[2] 正確にいえば $\tau = 30$ の制約であるが，最短所要時間に比べて十分大きいと見なせる．

用いられているためだと考えられる．

また，個々の到着時間の制約について設定値と推定値の比較を行った（図 4a）．図中の推定値は負担率が最大となる τ，すなわち $\mathrm{argmax}_\tau\, \gamma(z_{n\tau})$ を用いている．多くのデータは対角上に位置しており，精度良く推定ができている様子が分かる．また，推定された負担率の例を図 4b に示す．図より軌跡を生成することができる時間の中では時間が短いほど負担率の値が大きくなる傾向にあることが分かる．実際，PoI のグリッドの報酬関数は高く設定されており，軌跡はできるだけ PoI のグリッドに早く向かう傾向にあることから整合的な結果が得られていると考えられる．

b) 実験 2

続いて，実験 2 でのパラメータ推定の結果を図 5 に示す．図 5a は，反復回数と θ の推定値，図 5b は，反復回数と μ の推定値を示している．実験 2 の推定は収束が早く，3 回目の反復で設定値に収束している様子が確認できた．実験 2 の問題設定は，価値が高くない芝生上のグリッドを目的地に設定しており，この設定は，従来の最大エントロピー逆強化学習の枠組みでは正しい推定ができない問題である．最大エントロピー逆強化学習では，特徴量の合計の期待値をモデルと観測で一致するようにパラメータを推定するため，基本的には目的地などの訪問が多い場所や特徴の価値が高く推定される．実際，今回の推定結果の 1 回目を見てみると全てのパラメータで過小に評価されており，これは芝生の価値が本来よりも過大に推定されているという結果が得られたことを意味している[3]．一方で，提案手法では時間の制約を強く受けているデータの寄与が小さくなるため，制約によって通らざるを得ない場所での特徴（今回の場合，芝生）の価値が補正され，逆に時間的制約をあまり受けていないデータから歩道や桜，PoI の正しい価値が推定できたと考えられる．

実験 2 の到着時間制約の結果（図 6a）は，極めて高い一致を示していることが分かる．実験 1 とは反対に，目的とする場所の価値が低い場合，移動軌跡は時間の限り報酬の高い場所に滞在し，到着時間ちょうどに到着するような行動になると考えられる．したがって，提案手法では到着時間の推定がしやすく，結果として θ も含めて早い収束となったと考えられる．図 6b に示される負担率の例も，1 つの到着時間制約に高い確率を示している．

最後に，それぞれの実験におけるパラメータ推定値の比較を図 7 にまとめた．図中の青色は設定値，橙色は反復の 1 回目の値（日高ら[2] に相当），緑色が提案手法である．どちらの実験においても，既存手法では正確な推定ができていない一方で，提案手法では設定値の通りに推定ができている様子が確認できる．

[3] θ_{ww} や θ_{PoI} は芝生の価値からの相対値で表されると仮定したため．

6. おわりに

本稿では，人々が潜在的に持っているであろう到着時間制約と，報酬関数を同時に推定する逆強化学習法を提案した．具体的には，移動時間の制約を考慮した確率的選択肢集合形成モデルを応用し，報酬関数と歩行者の潜在的な到着時間の制約を同時に推定する新たな逆強化学習の問題を提案した．また，EM algorithm により効率的な学習が可能であることを示した．さらに，数値実験を通じて提案手法が既存手法と比べて報酬関数推定のバイアスが少なく，より正確な推定ができることを示した．

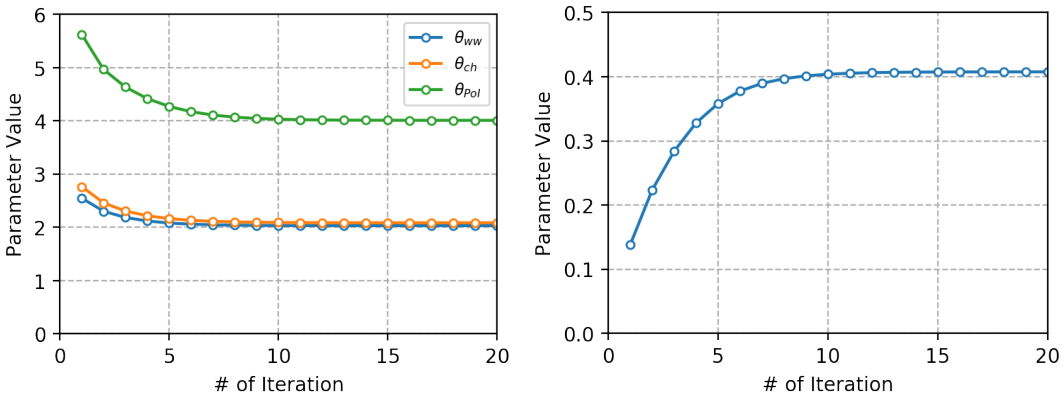
以下，本稿の分析の限界について触れておく．本稿では，到着時間の制約分布に負の二項分布を仮定した．これは，2 つの理由によるものであった．一つは，最短所要時間以上の定義域で定義することができる点，もう一つは，離散確率分布である点である．負の二項分布を用いることで推定するパラメータが 1 つになり計算が簡便になった一方，分布の形状が 1 つのパラメータ（平均迂回因子の逆数）で決定されてしまう．実際には，平均は同じでもばらつきが違うケースが考えられるため，それらが報酬関数の推定にどのような影響を与えうるか調べる必要がある．また，提案手法は分布を限るものではないため，適切な確率分布を選択する方法も課題に挙げられる．

本稿では，数値実験での分析に留まったが，実データへの適用は今後の課題である．提案手法は，歩行者の GPS などの移動軌跡データを想定したものであるが，実環境のモデル化（場所の特徴量表現）が課題となる．しかしながら，提案手法を適用により，どんな場所や地物の報酬が高い傾向にあるのか，また，それらは歩行者のどのような属性によって異なるのか，時間帯によって到着時間の制約に違いがあるのかなどの分析が可能になるであろう．

本稿で提案した手法は，特定の一つの目的地への移動を対象とした手法になっているが，一般的な回遊行動を考える上では，複数の目的地を考慮する必要がある．そのようなモデルへの拡張も取り組むべき課題である．

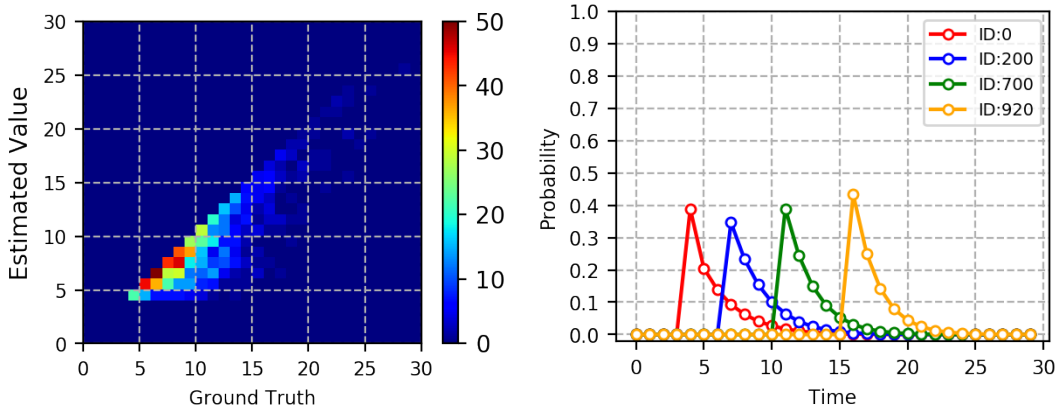
参考文献

1) Gehl, J.: *Life between buildings: using public space*, Island press, 2011.
2) 日高健, 早川敬一郎, 西智樹, 薄井智貴, and 山本俊行: 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, **第 *58* 回土木計画学研究発表会・講演集**, 2018.



(a) θ の推定値　　(b) μ の推定値

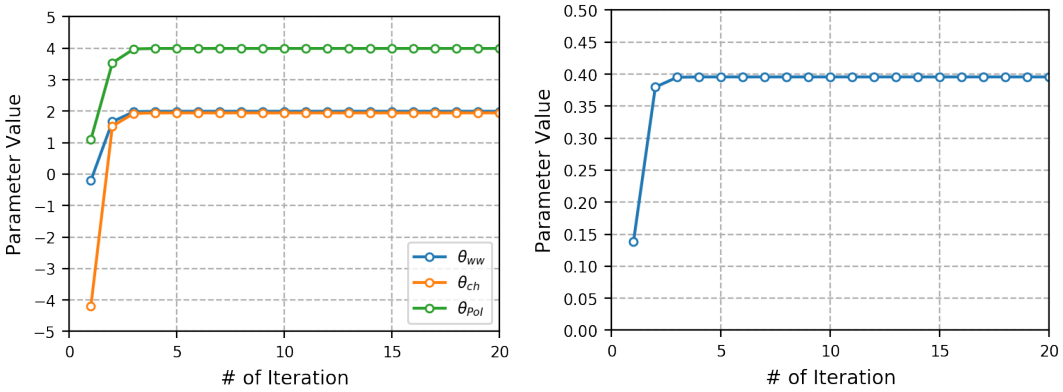
図–3: 反復回数とパラメータ推定値（実験 1）



(a) 到着時間制約の設定値と $\mathrm{argmax}_\tau\, \gamma(z_{n\tau})$ の比較　　(b) 負担率 $\gamma(z_{n\tau})$ の例

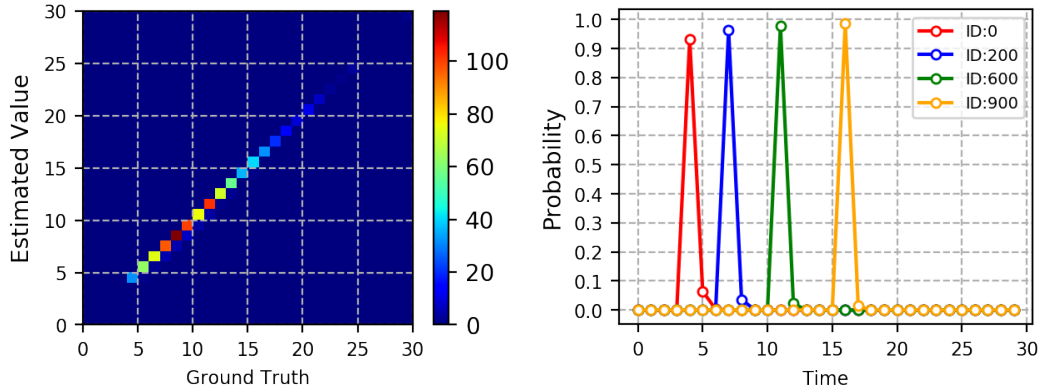
図–4: 到着時間制約の推定結果（実験 1）

3) 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.
4) Thill, J.-C. and Horowitz, J. L.: Modelling non-work destination choices with choice sets defined by travel-time constraints, *Recent Developments in Spatial Analysis*, pp. 186–208, Springer, 1997.
5) Dempster, A. P., Laird, N. M., and Rubin, D. B.: Maximum likelihood from incomplete data via the em algorithm, *Journal of the royal statistical society. Series B (methodological)*, pp. 1–38, 1977.
6) Ng, A. Y., Russell, S. J., et al.: Algorithms for inverse reinforcement learning., *Icml*, pp. 663–670, 2000.
7) 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.
8) 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.
9) 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.
10) Kitani, K. M., Ziebart, B. D., Bagnell, J. A., and Hebert, M.: Activity forecasting, *European Conference on Computer Vision*, pp. 201–214, Springer, 2012.
11) Manski, C. F.: The structure of random utility models, *Theory and decision*, Vol.8, No.3, pp.229–254, 1977.
12) McFadden, D.: Modeling the choice of residential location, *Transportation Research Record*, No.673, 1978.
13) Williams, H. and Ortúzar, J. d. D.: Behavioural theories of dispersion and the mis-specification of travel demand models, *Transportation Research Part B: Methodological*, Vol.16, No.3, pp.167–219, 1982.
14) Ben-Akiva, M. and Boccara, B.: Discrete choice models with latent choice sets, *International journal of*



(a) θ の推定値　　(b) μ の推定値

図–5: 反復回数とパラメータ推定値（実験 2）

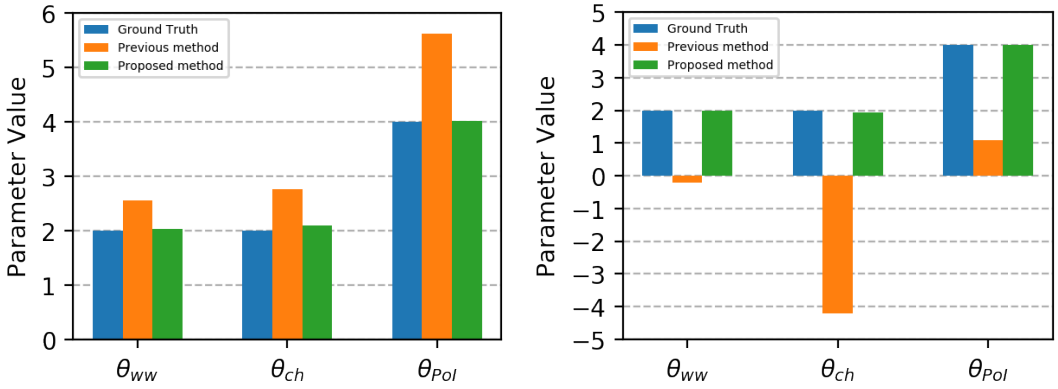


(a) 到着時間制約の設定値と $\mathrm{argmax}_\tau\, \gamma(z_{n\tau})$ の比較　　(b) 負担率 $\gamma(z_{n\tau})$ の例

図–6: 到着時間制約の推定結果（実験 2）

Research in Marketing, Vol.12, No.1, pp.9–24, 1995.

15) Scott, D. M. and He, S. Y.: Modeling constrained destination choice for shopping: a gis-based, time-geographic approach, *Journal of Transport Geography*, Vol.23, pp.60–71, 2012.
16) Mariante, G. L., Ma, T.-Y., and Van Acker, V.: Modeling discretionary activity location choice using detour factors and sampling of alternatives for mixed logit models, *Journal of Transport Geography*, Vol.72, pp.151–165, 2018.
17) Kaplan, S. and Prato, C. G.: Closing the gap between behavior and models in route choice: The role of spatiotemporal constraints and latent traits in choice set formation, *Transportation Research Part F: traffic psychology and behaviour*, Vol.15, No.1, pp.9–24, 2012.
18) Swait, J. and Ben-Akiva, M.: Incorporating random constraints in discrete models of choice set generation, *Transportation Research Part B: Methodological*, Vol.21, No.2, pp.91–102, 1987.
19) Hägerstraand, T.: What about people in regional science?, *Papers in regional science*, Vol.24, No.1, pp.7–24, 1970.
20) Thill, J.-C. and Horowitz, J. L.: Travel-time constraints on destination-choice sets, *Geographical Analysis*, Vol.29, No.2, pp.108–123, 1997.
21) Kamakura, W. A. and Russell, G. J.: A probabilistic choice model for market segmentation and elasticity structure, *Journal of marketing research*, pp. 379–390, 1989.
22) Gupta, S. and Chintagunta, P. K.: On using demographic variables to determine segment membership in logit mixture models, *Journal of Marketing Research*, pp. 128–136, 1994.
23) Greene, W. H. and Hensher, D. A.: A latent class model for discrete choice analysis: contrasts with mixed logit, *Transportation Research Part B: Methodological*, Vol.37, No.8, pp.681–698, 2003.
24) Bhat, C. R.: An endogenous segmentation mode choice model with an application to intercity travel, *Transportation science*, Vol.31, No.1, pp.34–48, 1997.
25) Hess, S., Bierlaire, M., and Polak, J.: A systematic comparison of continuous and discrete mixture models, *European Transport\Trasporti Europei*, No.37, pp.35–61, 2007.
26) Train, K. E.: Em algorithms for nonparametric es-



(a) 実験 1　　(b) 実験 2

図–7: パラメータ推定値の比較（青：設定値，橙：日高ら[2] の方法，青：提案手法）

timation of mixing distributions, *Journal of Choice Modelling*, Vol.1, No.1, pp.40–69, 2008.
27) Bogert, K. and Doshi, P.: Scaling expectation-maximization for inverse reinforcement learning to multiple robots under occlusion, *Proceedings of the 16th Conference on Autonomous Agents and Multi-Agent Systems*, pp. 522–529, International Foundation for Autonomous Agents and Multiagent Systems, 2017.
28) Bogert, K., Lin, J. F.-S., Doshi, P., and Kulic, D.: Expectation-maximization for inverse reinforcement learning with hidden data, *Proceedings of the 2016 International Conference on Autonomous Agents & Multiagent Systems*, pp. 1034–1042, International Foundation for Autonomous Agents and Multiagent Systems, 2016.
29) Shahryari, S. and Doshi, P.: Inverse reinforcement learning under noisy observations, *Proceedings of the 16th Conference on Autonomous Agents and Multi-Agent Systems*, pp. 1733–1735, International Foundation for Autonomous Agents and Multiagent Systems, 2017.
30) Nguyen, Q. P., Low, B. K. H., and Jaillet, P.: Inverse reinforcement learning with locally consistent reward functions, *Advances in Neural Information Processing Systems*, pp. 1747–1755, 2015.
31) Jaynes, E. T.: Information theory and statistical mechanics, *Physical review*, Vol.106, No.4, pp.620, 1957.
32) Wiedemann, C.: External evaluation of road networks, *International Archives of Photogrammetry Remote Sensing and Spatial Information Sciences*, Vol.34, No.3/W8, pp.93–98, 2003.
33) Witlox, F.: Evaluating the reliability of reported distance data in urban travel behaviour analysis, *Journal of Transport Geography*, Vol.15, No.3, pp.172–183, 2007.

(2019. 3. 10 受付)

INVERSE REINFORCEMENT LEARNING FOR SIMULTINEOUS ESTIMATION OF PEDESTRIAN REWARD FUNCTION AND LATENT ARRIVAL TIME CONSTRAINTS

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