# 歩行者回遊行動モデリングのための最大エントロピー逆強化学習とEMアルゴリズム

山本俊行（名古屋大学）
日髙健・早川敬一郎・西智樹（豊田中央研究所）
薄井智貴（人間環境大学）

# 歩行者の回遊行動

・目的地指向行動
　移動＝不効用，**不効用最小化経路**

・回遊行動
　移動＝正の効用，**効用最大化経路**

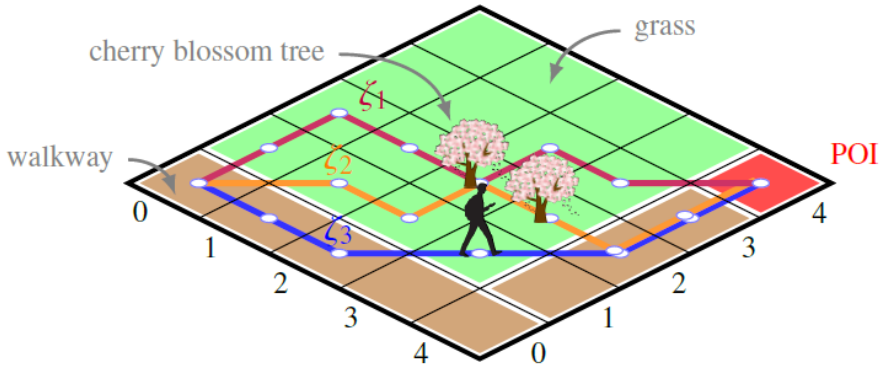
いくらでも効用を大きくできるため
**時間的な制約**が必要


目標

入力：
　行動系列：$\boldsymbol{\zeta} = \{\zeta_1, \zeta_2, \zeta_3\}$
　環境の特徴量：歩道，芝生，桜の木，POI

出力：
　各特徴量の重み：$\theta = [\theta_{\mathrm{ww}}, \theta_{\mathrm{ch}}, \theta_{\mathrm{POI}}]^{\mathrm{T}}$
　報酬関数：$\mathrm{R}(s)$

$$\mathrm{R}(s) = \theta_{\mathrm{ww}} f_{\mathrm{ww}} + \theta_{\mathrm{ch}} f_{\mathrm{ch}} + \theta_{\mathrm{POI}} f_{\mathrm{POI}}$$




Illustrative example of the problem (POI: point of interest)

報酬関数は各特徴量の線形和

# 発表の構成

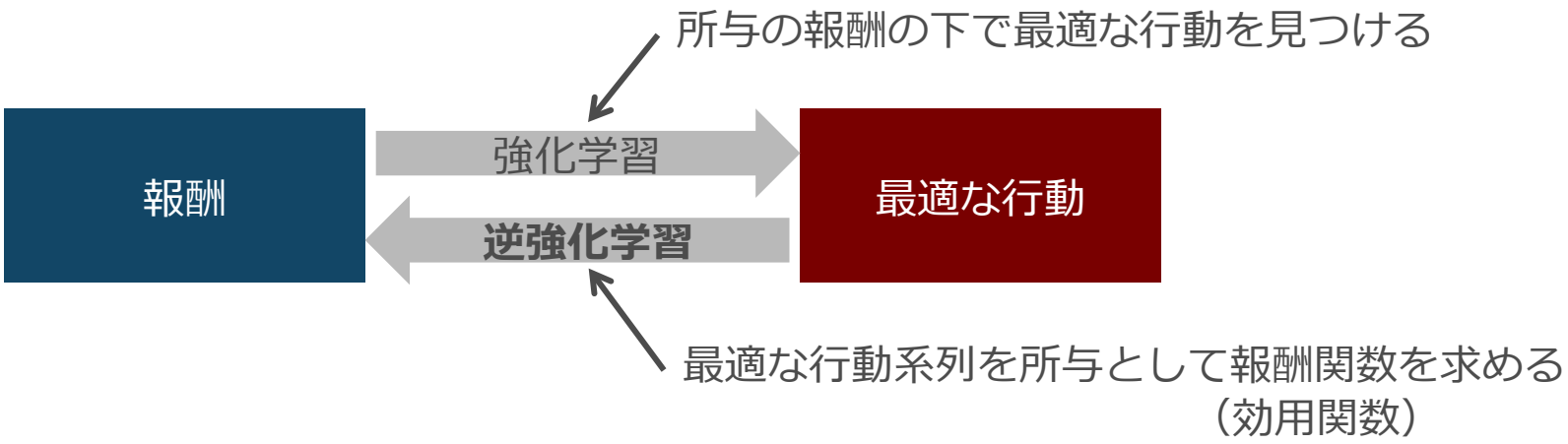
1．逆強化学習
  1-1. 逆強化学習の歴史
  1-2. 最大エントロピー逆強化学習の導出
    1-2-1. マルコフ決定過程（MDP）
    1-2-2. Soft-max MDP
    1-2-3. 最大エントロピー逆強化学習
    1-2-4. 最大エントロピー逆強化学習の推定
    1-2-5. Recursive logitモデルとの比較
  1-3. 時空間制約下の方策

2．EMアルゴリズム
  2-1. EMアルゴリズムとは？
  2-2. EMアルゴリズムの導出
  2-3. EMアルゴリズムのイメージ

3．歩行者回遊行動への適用

# 1．逆強化学習

# 1-1. 逆強化学習の歴史




- 逆強化学習はRussell(1998)によって提案
- 様々な手法が提案されている
  - ✓ 線形計画法を用いた推定（Ng et al., 2000）
  - ✓ 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max-margin法で推定（Abbeel and Ng, 2004）
  - ✓ Abbeel and Ng(2004)の方法は条件を満たす複数の解候補があり，解の不定性を解決するため最大エントロピー原理を導入（Ziebart et al., 2008）
  - ✓ その他にも相対エントロピー逆強化学習（Bouralias et al., 2011），最大エントロピー深層逆強化学習（Wulfmeier et al., 2015)など様々な方法へ発展

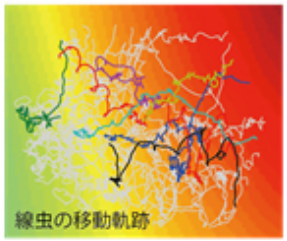
# 1-1. 逆強化学習の歴史

## 逆強化学習の適用例

行動時系列データ → 逆強化学習法 → 行動戦略

線虫

温度勾配

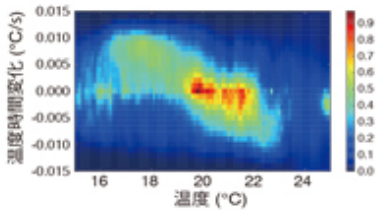



線虫は育成された温度を好むように、また
飢餓を経験した温度を避けるように移動する。

行動戦略を示す数式

$$\pi(s'|s) = \frac{p(s'|s)\exp(v(s'))}{\sum_x p(x|s)\exp(v(x))}$$

推定された価値関数：$v(s)$




動物が行動していて遭遇する各状況が、戦略上
どれくらいの価値があるのかを示している。

機械学習により「動物の行動戦略」を解読 –
動物は何を報酬として行動しているのか–
http://www.kyoto-u.ac.jp/ja/research/research_results/2018/180515_1.html

逆強化学習による人の行動予測
Kitani et al.(2012)

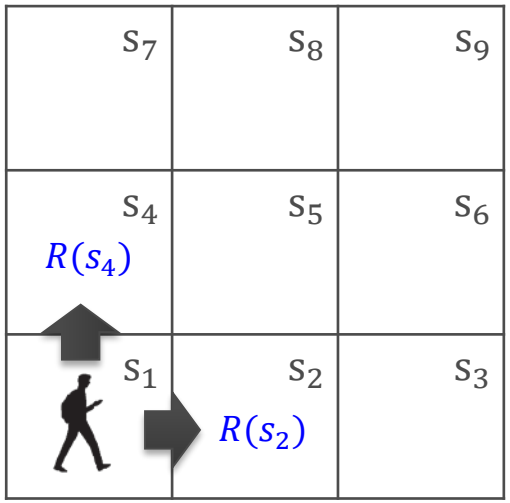
# 1-2-1. マルコフ決定過程（MDP）

$< \mathcal{S}, \mathcal{A}, \mathcal{T}, \mathcal{R} >$:

- $\mathcal{R}$：報酬関数
- $\mathcal{T}$：遷移モデル
- $\mathcal{A}$：行動の集合
- $\mathcal{S}$：状態の集合

マルコフ決定過程（MDP）は
この4つの組で定義される

回遊空間の表現：グリッド




状態の集合：グリッドの全てのセル$s_1$~$s_9$

行動の集合：上下左右

遷移モデル：状態$s$から行動$a$を取ったときに状態$s'$に遷移する確率

報酬関数　：新たな状態に遷移した結果，環境から受け取る値

# 1-2-1. マルコフ決定過程（MDP）

最適な行動　⇒　累積報酬和を最大にする行動の結果として
選択されたものと仮定する

図の例の場合，軌跡$\zeta$の累積報酬和$R(\zeta)$は

$$R(\zeta) = R(s_1) + R(s_2) + R(s_5) + R(s_6) + R(s_9)$$

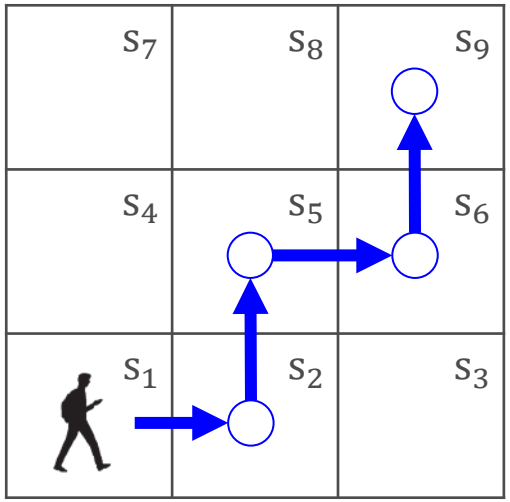
ここでは，単純な累積和の代わりに割引累積報酬和を用いる

$$R(\zeta) = R(s_1) + \gamma R(s_2) + \gamma^2 R(s_5) + \gamma^3 R(s_6) + \gamma^4 R(s_9)$$

ただし，$\gamma \in [0,1]$は割引率
割引率を考慮する理由としては，

・値が発散しない
・将来の不確実性を考慮

逆強化学習は，このような累積報酬を最大にする
行動が観測と合うような報酬関数を見つける方法




グリッド空間上の行動
$\zeta = \{s_1, s_2, s_5, s_6, s_9\}$

# 1-2-1. マルコフ決定過程（MDP）

ここで，状態価値関数V(s)と行動価値関数Q(s, a)を定義する

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

状態価値関数
（**状態s**にいるときの期待累積報酬和）

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

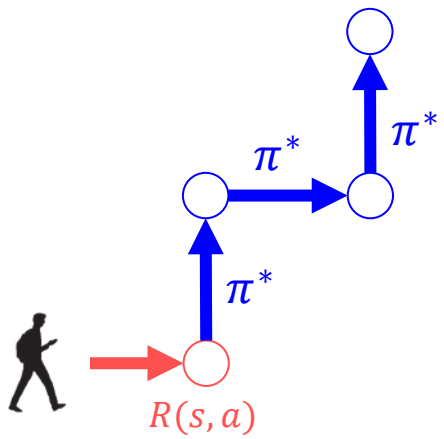
行動価値関数
（**状態s**において**行動a**をとるときの期待累積報酬和）

ベルマンの最適性原理より以下の関係が得られる

次の状態の報酬　　次の状態の価値（＝将来の報酬和）

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

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




ここでのモデルは
決定論的なモデル

このときの最適な行動方策は以下のように書ける

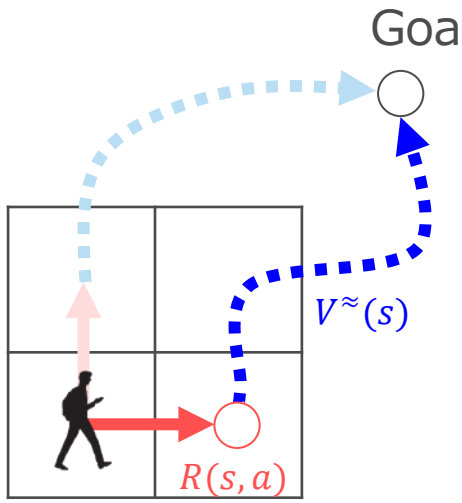
$$\pi^* = \mathrm{argmax}_{a \in \mathcal{A}} Q^*(s,a).$$

最も行動価値が高い行動

# 1-2-2. Soft-max MDP

Ziebart et al.(2009)は意思決定に関わる不確実性を考慮するためにMaxの代わりに期待最大効用（ログサム）を導入

$$Q^{\approx}(s,a) = R(s,a) + \gamma V^{\approx}(\mathcal{T}(s,a)),$$
$$V^{\approx}(s) = \log \sum_{a \in \mathcal{A}} \exp\{Q^{\approx}(s,a)\}.$$




確率論的なモデル
(Recursive logitと同じ)

このとき行動方策は以下のロジットの形で得られる

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

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

次の状態の効用

次の状態にいるときの期待最大効用

将来を考慮した逐次選択のモデル

# 1-2-3. 最大エントロピー逆強化学習

逆強化学習は，$< S, A, T >$とエキスパートの行動から報酬関数$R$を推測する問題
報酬関数は以下の対数尤度関数が最大となるように決定する，すなわち

$$\mathrm{argmax}_R \sum_i \log P(\zeta_i | R)$$

最大エントロピー原理より，観測された軌跡$\zeta_n$が得られる確率は以下の式になる

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

Ziebart et al.(2008)は，報酬関数に線形性を仮定すること，すなわち

$$R := \sum_{s \in \zeta_i} \boldsymbol{\theta}^{\mathrm{T}} \mathbf{f_s}$$

により報酬関数が効率的に推定できることを示した.
ただし，$\theta$はパラメータで，$\mathbf{f}_s$は$k$次元の特徴量ベクトル

# 1-2-4. 最大エントロピー逆強化学習の推定

パラメータ$\theta$は対数尤度関数の最大化によって決定される

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

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

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

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

上の式の勾配を計算すると，以下の式が得られる

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

観測における特徴量の合計の平均値

モデルの特徴量の合計の期待値

$$\mathbb{E}_{P_\theta(\zeta)}[\boldsymbol{f}_s] = \sum_s P(s|\theta) \boldsymbol{f}_s$$

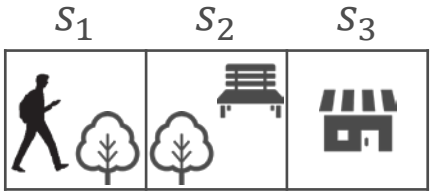
パラメータ$\theta$のモデルで，状態$s$を何回訪れるか
State Visitation Frequency(SVF)と呼ばれる

# 1-2-4. 最大エントロピー逆強化学習の推定

モデルの特徴量の合計の期待値の計算

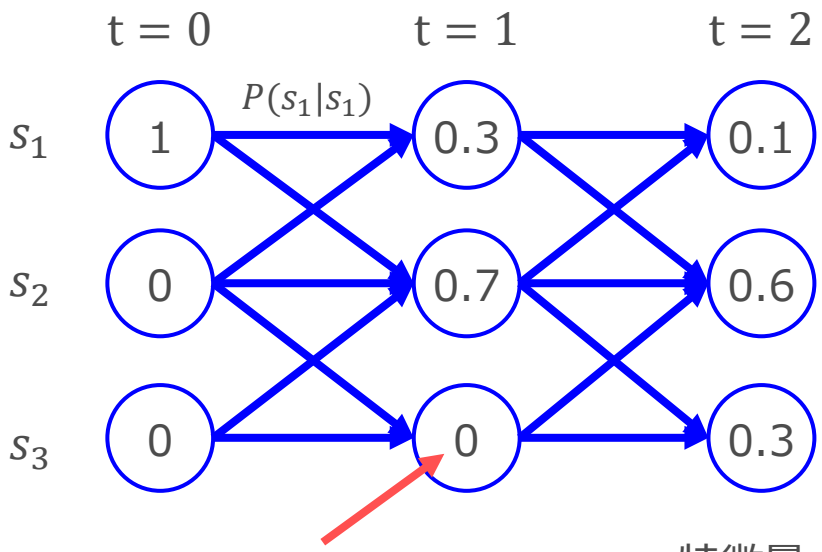
$$\mathbb{E}_{P_\theta(\zeta)}[\boldsymbol{f}_s] = \sum_s P(s|\theta)\boldsymbol{f}_s$$

パラメータ$\theta$のモデルで，状態$s$を何回訪れるか
State Visitation Frequency(SVF)と呼ばれる



環境モデル

右上の図のSVFの計算例




モデルから計算
される存在確率

| | 特徴量 | | |
|---|---|---|---|
| SVF | 木 | ベンチ | お店 |
| 1.4 | 1 | 0 | 0 |
| 1.3 | 1 | 1 | 0 |
| 0.3 | 0 | 0 | 1 |

特徴量（歩道）の期待値＝内積(SVF,木)＝2.7
特徴量（ベンチ）の期待値＝内積(SVF,ベンチ)＝1.3
特徴量（お店）の期待値＝内積(SVF,お店)＝0.3

# 1-2-4. 最大エントロピー逆強化学習の推定

価値関数の計算：価値反復法
　　数値反復計算によりBellman最適方程式を求める方法
　　強化学習分野では一般的な方法の一つ

**Step1**
全てのsにおいて価値関数$V^{\approx}(s)$の値を初期化
これを$V_0^{\approx}(s)$と表す.

**Step2**
全てのsに対して以下を実行する.

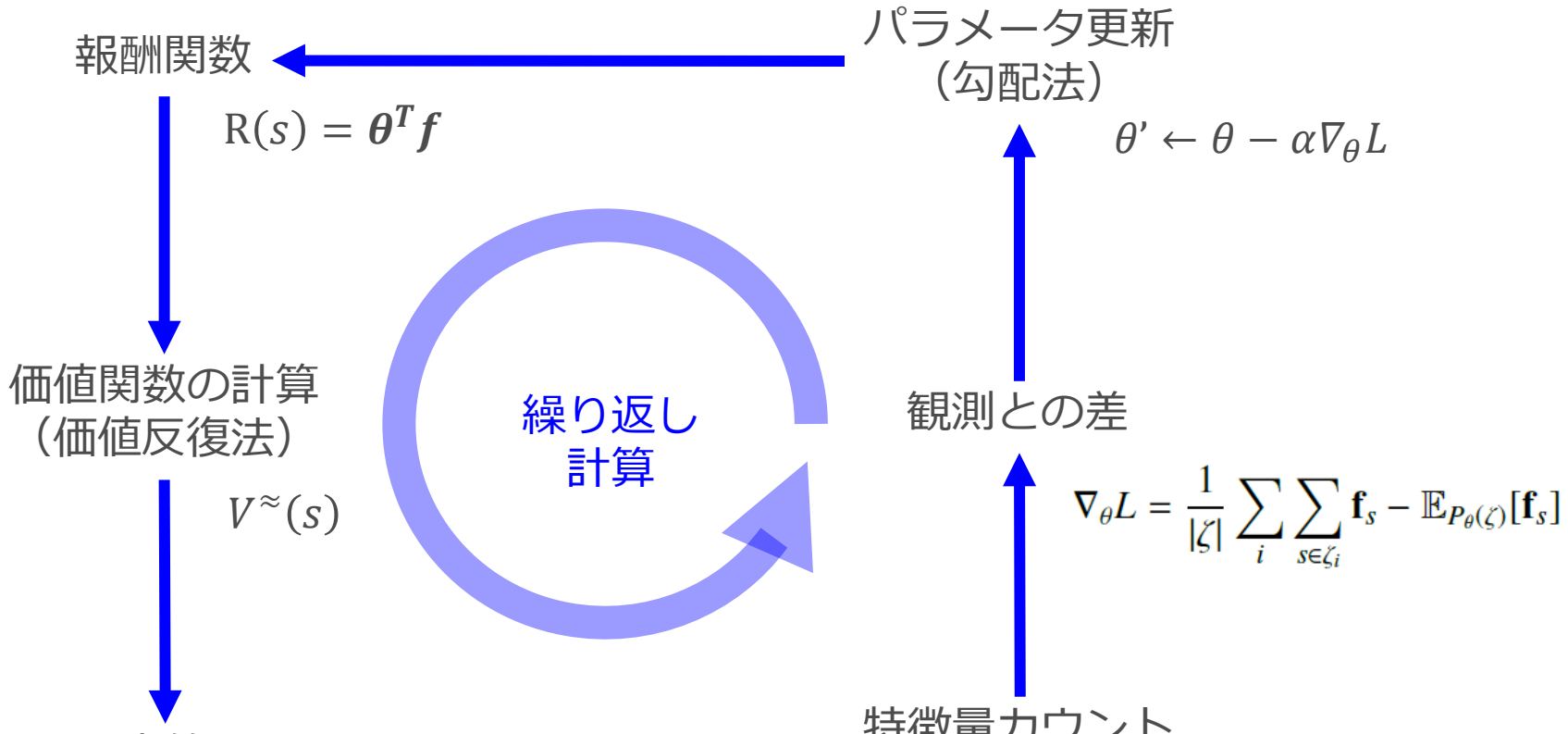
$$Q_k^{\approx}(s,a) = \sum_{s'} P(s|s',a)\{R(s',a) + \gamma V_{k-1}^{\approx}(s')\}$$

$$V_k^{\approx}(s) = \log \sum_{a} \exp\{Q_k^{\approx}(s,a)\}$$ ←――― 通常のMDPならMax

**Step3**
$|V_k^{\approx}(s) - V_{k-1}^{\approx}(s)|$の値が十分小さくなるまでStep2を繰り返す

$P(s|s',a)$：状態$s'$で行動$a$を取ったときに
状態$s$に遷移する確率

# 1-2-4. 最大エントロピー逆強化学習の推定




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

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

# 1-2-5. Recursive logit modelとの違い

- 価値関数の求解方法
  - Recursive logit modelは逆行列計算を利用
    - 逆行列計算ができるよう，負の効用のみを扱う
  - 最大エントロピー逆強化学習は反復計算を利用
    - 正の効用も扱うことができる

ただし，Discounted recursive logit (Oyama and Hato, 2017)では recursive logitに時間割引を導入し，反復計算によって価値関数を計算している.

# 1-3. 時空間制約下の方策

最大エントロピー逆強化学習で得られる方策：$\pi(a_t|s_t)$
状態間の遷移確率は方策と遷移モデルを用いて以下のように表せる.

$$P(s_{t+1}|s_t) = \sum_a \underbrace{P(s_{t+1}|s_t, a_t)}_{\text{遷移モデル}} \pi(a_t|s_t)$$

ここで，目的地に関する時空間制約を導入したい.
ある目的地に到着必須時刻$\boldsymbol{T_{arvl}}$に到着するという制約の下に，時空間制約がないときの遷移確率と整合的な状態遷移確率$\boldsymbol{P\left(s_{t+1}|s_t, s_{T_{arvl}}\right)}$を求めたい.

# 1-3. 時空間制約下の方策

以下の後ろ向き確率を導入する.

方策に従って行動したときに目的地に時間通りに到着する確率（後向き確率）

$$\beta(s_t) \equiv P\left(s_{T_{arvl}}|s_t\right) = \sum_{s_{t+1}} P\left(s_{T_{arvl}}|s_{t+1}\right)P(s_{t+1}|s_t)$$

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

ベイズの定理より以下の関係が得られる.

$$P\left(s_{t+1}|s_t, s_{T_{arvl}}\right) = \frac{P\left(s_{t+1}, s_{T_{arvl}}|s_t\right)}{P\left(s_{T_{arvl}}|s_t\right)}$$

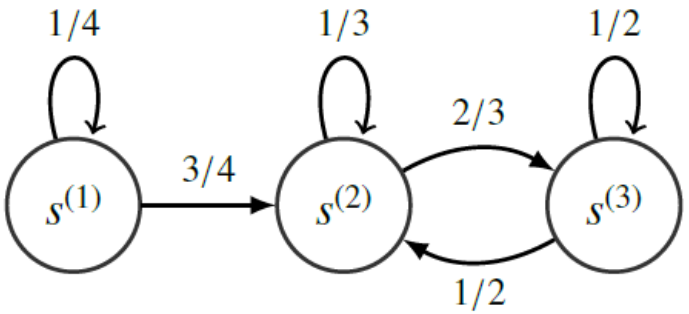
$$= \frac{P\left(s_{T_{arvl}}|s_{t+1}\right)P(s_{t+1}|s_t)}{P\left(s_{T_{arvl}}|s_t\right)}$$

$$= \frac{\beta(s_{t+1})}{\beta(s_t)}P(s_{t+1}|s_t)$$

後ろ向き確率の比を用いて計算できる

# 1-3. 時空間制約下の方策

## 簡易ネットワーク上の計算例




(a) sample network

問題設定：$t = 3$に$s^{(3)}$に到着

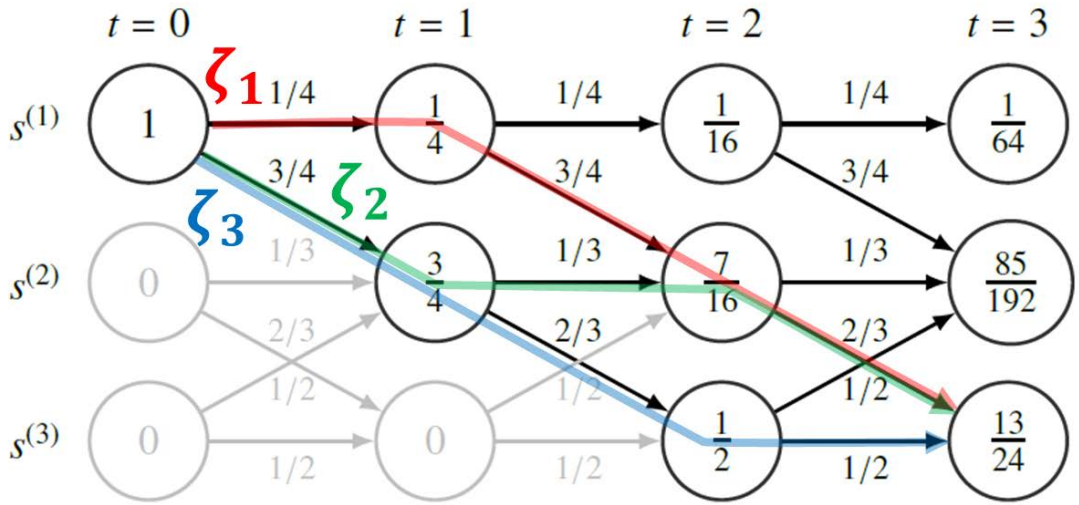
経路は3種類

$$P(\zeta_1) = \frac{1}{4} \cdot \frac{3}{4} \cdot \frac{2}{3} = \frac{1}{8} \quad P(\zeta_2) = \frac{3}{4} \cdot \frac{1}{3} \cdot \frac{2}{3} = \frac{1}{6}$$

$$P(\zeta_3) = \frac{3}{4} \cdot \frac{2}{3} \cdot \frac{1}{2} = \frac{1}{4}$$

それぞれの経路の選択確率は

$$P(\zeta_1|Z) = 3/13,\ P(\zeta_2|Z) = 4/13,\ P(\zeta_3|Z) = 6/13$$

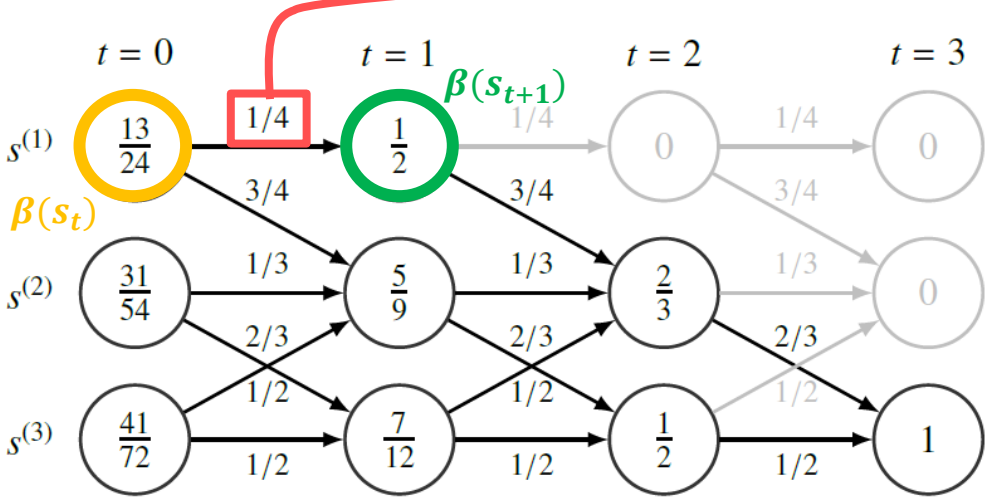


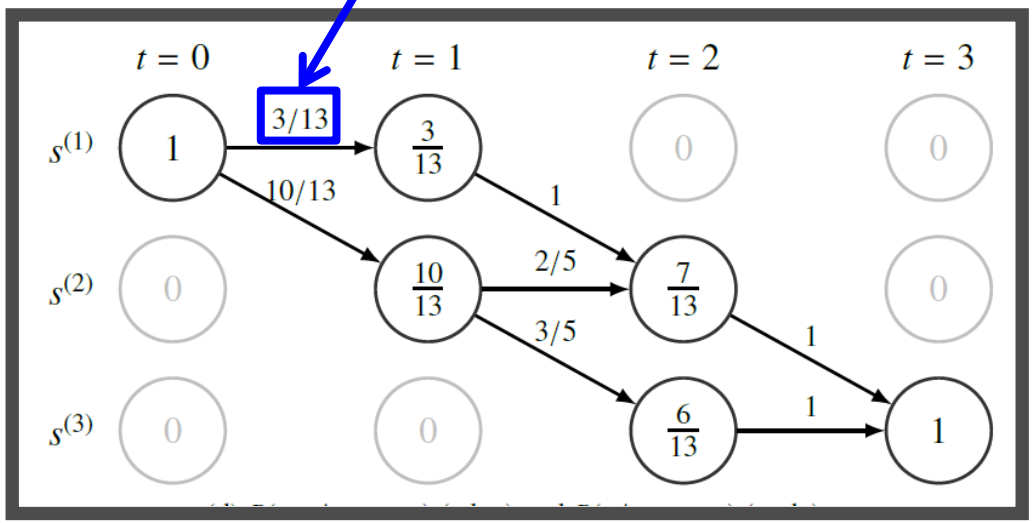

前向き確率の計算
（時空間制約なし）

# 1-3. 時空間制約下の方策

## 簡易ネットワーク上の計算例

$$P(s_{t+1}|s_t, s_{T_{arvl}}) = \frac{\beta(s_{t+1})}{\beta(s_t)} P(s_{t+1}|s_t) = \frac{1/2}{13/24} \cdot \frac{1}{4} = \frac{3}{13}$$




後向き確率の計算

時空間制約下についても後ろ向き確率を用いることで逐次選択，経路列挙必要なしで整合的な経路を与えることができる

# 2．EMアルゴリズム

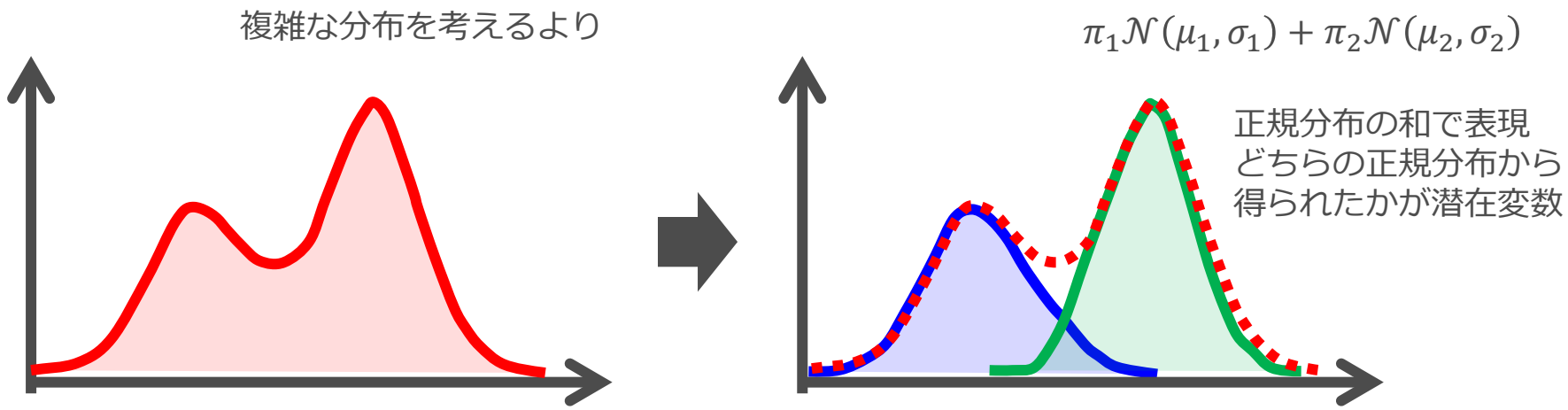
# 2-1. EMアルゴリズムとは？

EMアルゴリズムとは

- 潜在変数を持つ確率モデルの最尤解を求めるための一般的な手法
- 観測変数$\boldsymbol{X}$の尤度関数$P(\boldsymbol{X}|\boldsymbol{\theta})$を最大化するパラメータ$\boldsymbol{\theta}$を見つける際に，潜在変数$\boldsymbol{Z}$を導入した以下の式の方が扱いやすい場合がある

$$p(\boldsymbol{X}|\theta) = \sum_{\boldsymbol{Z}} p(\boldsymbol{X}, \boldsymbol{Z}|\theta)$$

- 混合分布の推定や潜在クラスモデルの推定などによく用いられる



# 2-2. EMアルゴリズムの導出

$$\ln p(\boldsymbol{X}|\theta) = \ln \sum_{\boldsymbol{Z}} p(\boldsymbol{X}, \boldsymbol{Z}|\theta)$$

**最大化したい対数尤度**

$$= \ln \sum_{\boldsymbol{Z}} q(\boldsymbol{Z}) \frac{p(\boldsymbol{X}, \boldsymbol{Z}|\theta)}{q(\boldsymbol{Z})}$$

$$\geq \sum_{\boldsymbol{Z}} q(\boldsymbol{Z}) \ln \frac{p(\boldsymbol{X}, \boldsymbol{Z}|\theta)}{q(\boldsymbol{Z})}$$

Jensenの不等式

$$= \mathcal{L}(q, \theta)$$

変分下限と呼ぶ

EMアルゴリズムの理解には以下のページがおすすめ
**「EMアルゴリズム徹底解説」**
https://qiita.com/kenmatsu4/items/59ea3e5dfa3d4c161efb

$\ln p(\boldsymbol{X}|\theta) - \mathcal{L}(q, \theta)$を計算すると

$$\ln p(\boldsymbol{X}|\theta) - \mathcal{L}(q, \theta) = \text{（中略）}$$

$$= KL[q(\boldsymbol{Z})||p(\boldsymbol{Z}|\boldsymbol{X}, \theta)]$$

KLダイバージェンス
(2つの確率分布間の距離)

$q(\boldsymbol{Z}) = p(\boldsymbol{Z}|\boldsymbol{X}, \theta)$を選ぶことで$\ln p(\boldsymbol{X}|\theta) = \mathcal{L}(q, \theta)$になる

# 2-2. EMアルゴリズムの導出

EMアルゴリズムでは，$\ln p(\boldsymbol{X}|\theta)$の代わりに変分下限$\mathcal{L}(q,\theta)$の最大化を行う．
$\mathcal{L}(q,\theta)$の引数$q$と$\theta$を交互に最適化しながら，$\mathcal{L}(q,\theta)$の最大化を行う．

E-step: $q$の最適化

$\theta$を固定して， $\ln p(\boldsymbol{X}|\theta)$最大となる$q$を選ぶ．
すなわち，以下を求める．

$$q(\boldsymbol{Z}) = p(\boldsymbol{Z}|\boldsymbol{X},\theta)$$

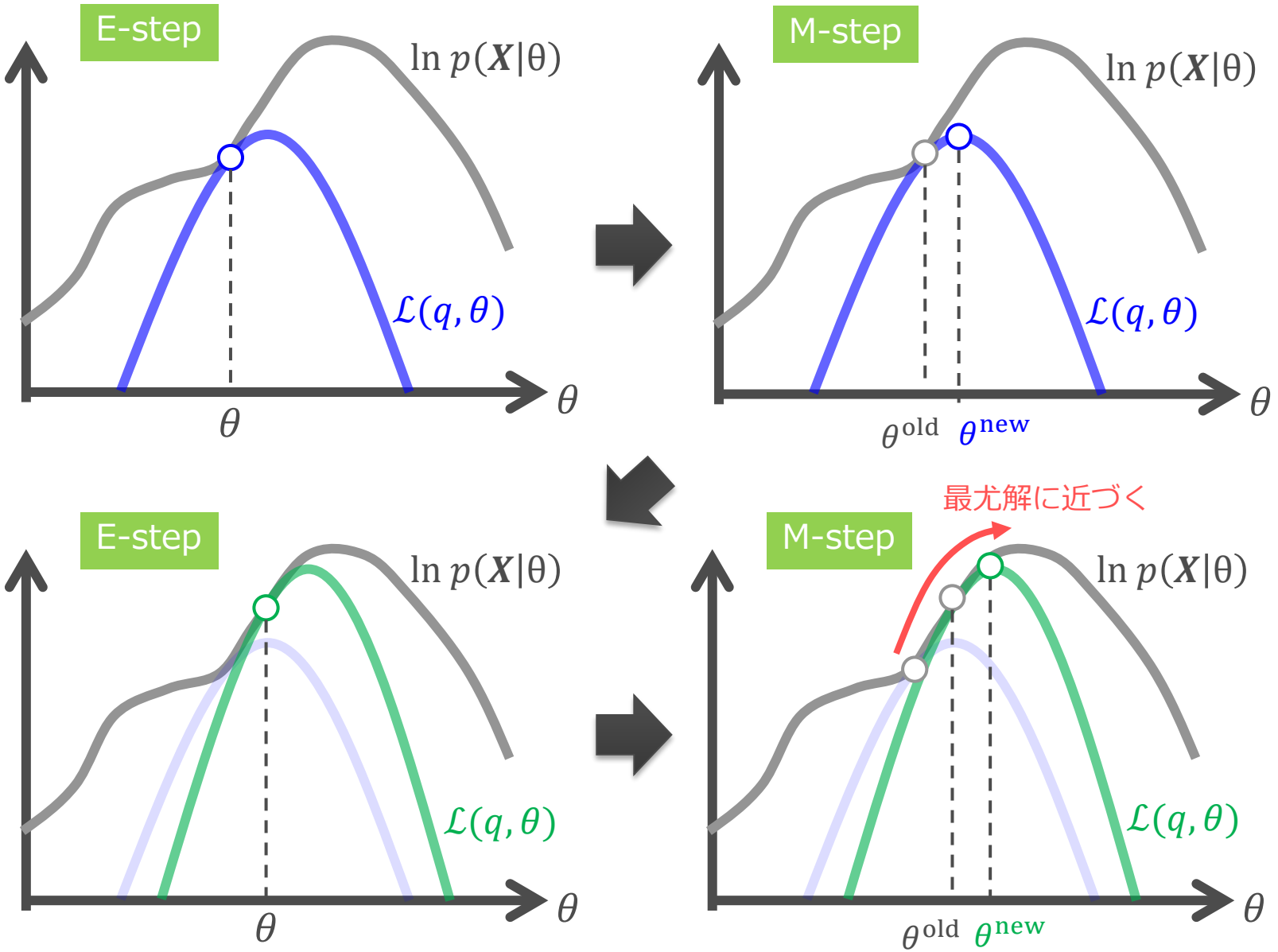
M-step: $\theta$の最適化

$q(\boldsymbol{Z})$を固定して，θを最大化する．
その際に事後分布の算出に用いたθは$\theta^{\mathrm{old}}$として固定したまま数値計算

$$\mathcal{L}(q,\theta) = \sum_{\boldsymbol{Z}} q\left(\boldsymbol{Z}|\boldsymbol{X},\theta^{\mathrm{old}}\right)\underbrace{\ln p(\boldsymbol{X},\boldsymbol{Z}|\theta)}_{\text{完全データの対数尤度関数}} + \mathrm{const}$$
$$= Q\left(\theta,\theta^{\mathrm{old}}\right) + \mathrm{const}$$

$Q\left(\theta,\theta^{\mathrm{old}}\right)$を$\theta$に関して最大化すれば良い．
完全データの対数尤度が最適化可能であれば
$Q\left(\theta,\theta^{\mathrm{old}}\right)$は最大化できる

# 2-3. EMアルゴリズムのイメージ




E-stepとM-stepを収束するまで繰り返す

# 3．歩行者回遊行動への適用

# 3-1. 歩行者回遊行動への適用

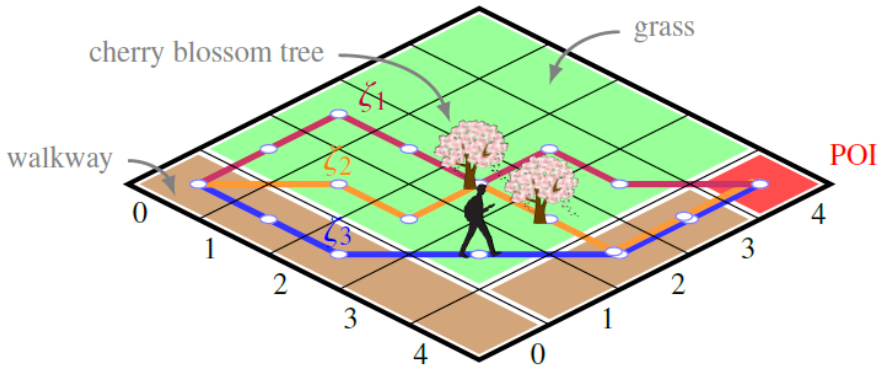
適用例

入力：
行動系列：$\boldsymbol{\zeta} = \{\zeta_1, \zeta_2, \zeta_3\}$
環境の特徴量：歩道，芝生，桜の木，POI

出力：
各特徴量の重み：$\theta = [\theta_{\mathrm{ww}}, \theta_{\mathrm{ch}}, \theta_{\mathrm{POI}}]^{\mathrm{T}}$
報酬関数：$\mathrm{R}(s)$

$$\mathrm{R}(s) = \theta_{\mathrm{ww}} f_{\mathrm{ww}} + \theta_{\mathrm{ch}} f_{\mathrm{ch}} + \theta_{\mathrm{POI}} f_{\mathrm{POI}}$$

報酬関数は各特徴量の線形和




Illustrative example of the problem (POI: point of interest)

推定されたパラメータは，特徴の持つ**価値（魅力度）**を表す
歩行者の軌跡は**効用最大化行動**の結果とみなす

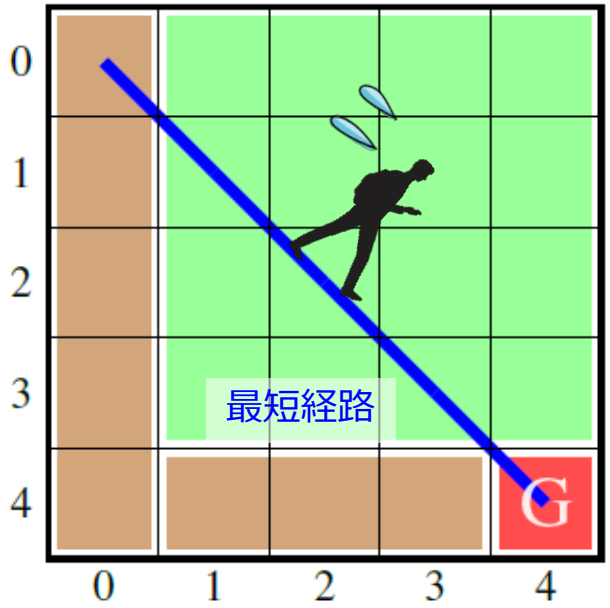
# 3-1. 歩行者回遊行動への適用

正の効用を扱う場合，同じ場所に滞在することで
効用をいくらでも大きくできてしまう
⇒　**時空間制約が必要**

しかしながら，実際の観測データからの推定を考えた場合，
**時間制約は観測できない**

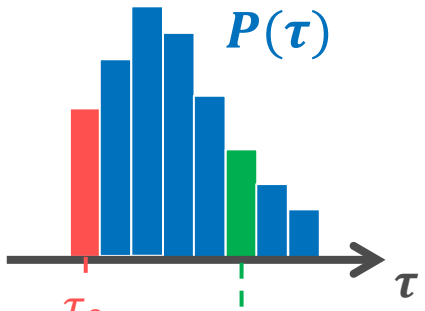
図の例のように最短経路の軌跡が観測された場合，
Goal地点の報酬が高いのか，単に急いでいるか区別がつかない.

本来は到着時間の制約を持っているにも関わらず，
制約がないとみなして推定を行うと，推定にバイアスが生じる

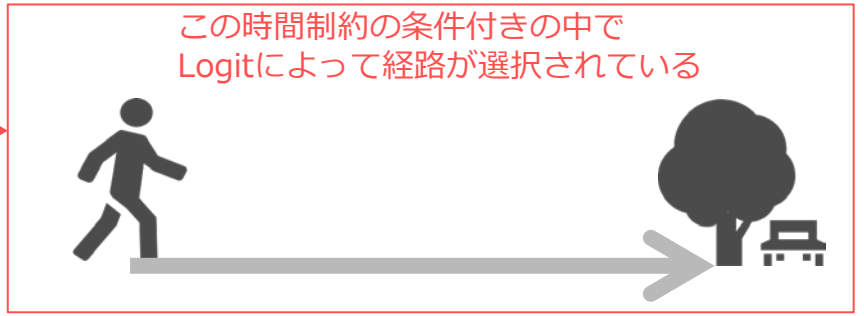


# 3-2. 潜在変数導入の考え方

潜在的な到着時間制約
の確率分布（ex. 負の二項分布）




時間制約：強い




時間制約：弱い


$$P(\zeta) = \sum_{\tau} P(\tau) P(\zeta|\tau)$$

軌跡$\zeta$の得られる確率

到着時間制約が$\tau$である確率
（何らかの確率分布に従うと仮定）

到着時間制約$\tau$を満たす軌跡の集合の
中からの軌跡$\zeta$の選択確率

基本的な考え方は，確率的な選択肢集合形成
Manski(1977)と同じ

# 3-3. 定式化

軌跡の観測数：N

$\zeta_n = \{s_0, s_1, \cdots, s_T\}$
状態の系列

観測された軌跡 $\boldsymbol{\zeta} = \{\zeta_1, \zeta_2, \ldots, \zeta_N\}$

到着時間制約（潜在変数） $\mathbf{z} = \{\mathbf{z}_1, \mathbf{z}_2, \ldots, \mathbf{z}_N\}$

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

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

$$z_{n\tau} = \{0, 1\} \quad \forall \tau \in \mathbb{N}.$$

到着時間制約に対応するτのみ1になる2値変数

潜在変数の導入により，軌跡の得られる確率$p(\zeta)$は以下のように表される

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

観測された軌跡集合ζの得られる確率

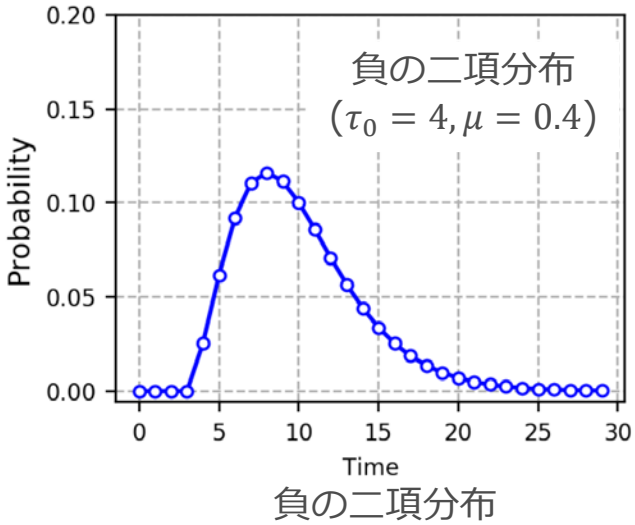
到着時間制約を満たす軌跡の集合の中からの軌跡ζの選択確率

到着時間制約の確率分布

# 3-3. 定式化

個々の到着時間制約の事前分布は以下の式で表される

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



負の二項分布

ここで，到着時間制約が従う確率分布として，以下の負の二項分布を仮定

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

$\mu$ ：パラメータ（試行の成功確率に相当）
$\tau_0$：最短所要時間（ODから決まる）

また，到着時間制約の下の軌跡の選択は最大エントロピー原理より以下に従うと仮定

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

# 3-3. 定式化

したがって，完全データの対数尤度関数は

$R(\zeta) = \theta^T \mathbf{f}_\zeta$を仮定

$$
\begin{aligned}
&\ln p(\boldsymbol{\zeta}, \mathbf{z}|\mu, \theta) \\
&= \ln \left\{ \prod_{n=1}^{N} \prod_{\tau=\tau_0}^{\infty} p(\tau|\mu)^{z_{n\tau}} \left\{ \frac{\exp\left\{\theta^\top \mathbf{f}_\zeta\right\}}{\sum_{\zeta \in \Xi_\tau} \exp\left\{\theta^\top \mathbf{f}_\zeta\right\}} \right\}^{z_{n\tau}} \right\} \\
&= \sum_{n=1}^{N} \sum_{\tau=\tau_0}^{\infty} z_{n\tau} \left( \ln p(\tau|\mu) + \ln \frac{\exp\left\{\theta^\top \mathbf{f}_\zeta\right\}}{\sum_{\zeta \in \Xi_\tau} \exp\left\{\theta^\top \mathbf{f}_\zeta\right\}} \right).
\end{aligned}
$$

完全データの対数尤度関数が最適化可能であれば**EMアルゴリズム**の適用が可能

# 3-4. EMアルゴリズムによる推定

E-step

$z_{n\tau}$の$\mathbf{z}$の事後分布による平均 $\mathbb{E}_{\mathbf{z}|\boldsymbol{\zeta}}[z_{n\tau}]$ を求める.

$\mathbf{z}$の事後分布は

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

$$= \frac{\prod_{n=1}^{N}\prod_{\tau=\tau_0}^{\infty} p(\tau|\mu)^{z_{n\tau}} \left\{ \frac{\exp\{\theta^\top \mathbf{f}_\zeta\}}{\sum_{\zeta\in\Xi_\tau}\exp\{\theta^\top \mathbf{f}_\zeta\}} \right\}^{z_{n\tau}}}{\sum_{\mathbf{z}_n}\prod_{n=1}^{N}\prod_{\tau'=\tau_0}^{\infty} p(\tau'|\mu)^{z_{n\tau'}} \left\{ \frac{\exp\{\theta^\top \mathbf{f}_\zeta\}}{\sum_{\zeta\in\Xi_{\tau'}}\exp\{\theta^\top \mathbf{f}_\zeta\}} \right\}^{z_{n\tau'}}}.$$

なので，$\mathbb{E}_{\mathbf{z}|\boldsymbol{\zeta}}[z_{n\tau}]$ は

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

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

$$\equiv \gamma(z_{n\tau}),$$

$\gamma(z_{n\tau})$：負担率(responsibility)

各到着時間制約τが経路ζの観測を説明する度合

# 3-4. EMアルゴリズムによる推定

M-step

E-stepで求めた事後分布を用いて対数尤度の期待値を求める

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

Q関数と呼ばれる

M-stepではQ関数を微分して，最尤解を求めれば良い

μの最尤解は$\frac{\partial Q}{\partial \mu} = 0$を解くことで求まる.

$\theta$の最尤解は$\frac{\partial Q}{\partial \theta} = 0$を解くことで求まる.

# 3-4. EMアルゴリズムによる推定

到着時間制約分布のパラメータ：$\mu$の最尤解

以下の式を解けば良い

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

したがって，

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

$\mu$について解くと

$$\mu = \frac{\sum_{n=1}^{N} \sum_{\tau=\tau_0}^{\infty} \tau_0 \gamma(z_{n\tau})}{\sum_{n=1}^{N} \sum_{\tau=\tau_0}^{\infty} \tau \gamma(z_{n\tau})} \quad \begin{matrix} = N \\ = N_\tau \end{matrix}$$

$$= \frac{\tau_0 N}{\sum_{\tau=\tau_0}^{\infty} \tau N_\tau}$$

$$= \frac{\tau_0}{\bar{\tau}}, \quad = \frac{\text{最短所要時間}}{\begin{matrix}\text{到着時間制約の} \\ \text{平均値}\end{matrix}} = \frac{1}{\text{平均迂回係数}}$$

$\mu$は最短所要時間に対する
迂回の度合の逆数として求まる

# 3-4. EMアルゴリズムによる推定

到着時間制約分布のパラメータ：$\theta$の最尤解

以下の式を解けば良い

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

したがって，

観測の特徴量の平均値

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

負担率による重みづけ

各到着時間制約τの下の特徴量の期待値（求め方については，付録を参照）

強い制約を受けている場合，経路の選択肢が少ないため
パラメータの大小に関わらず，モデルと観測が一致しやすい
⇒**自由に回遊している（と思われる）データほどパラメータ推定に寄与する**

# 3-4. EMアルゴリズムによる推定

**Algorithm**

1. 負担率$\gamma(z_{n\tau})$を設定
2. 収束するまで以下を繰り返す
    1. パラメータの更新
        1. $\mu$の更新
        2. $\theta$の更新
    2. 負担率の更新

# 参考文献

関連発表

[1] 日高健，早川敬一郎，西智樹，薄井智貴，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，CD-ROM，2018.

[2] 日高健，早川敬一郎，西智樹，薄井智貴，山本俊行，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，第59回土木計画学研究発表会，CD-ROM，2019.

参考文献

Abbeel, P., Ng, A. Y., Apprenticeship learning via inverse reinforcement learning, In: Proceedings of the 21st international conference on Machine learning (ICML), 2004.

Boularias, A., Kober, J., Peters, J. Relative entropy inverse reinforcement learning. In Proceedings of the 14th International Conference on Artificial Intelligence and Statistics (AISTATS), 182-189, 2011.

Fosgerau, M., Frejinger, E., Karlstrom, A., A link based network route choice model with unrestricted choice set. Transp. Res. Part B: Methodol. 56, 70–80, 2013.

Kitani, K. M., Ziebart, B. D., Bagnell, J. A., Hebert, M., Activity forecasting. In: European Conference on 418 Computer Vision (ECCV), 201–214, 2012.

Manski, C. F., The structure of random utility models. Theory and decision 8 (3), 229–254, 1977.

Ng, A. Y., Russell, S. J., Algorithms for inverse reinforcement learning. In: Proceedings of the 17th International Conference on Machine Learning (ICML), 663–670, 2000.

Oyama, Y., Hato, E., A discounted recursive logit model for dynamic gridlock network analysis. Transp. Res. Part C: Emerg. Technol. 85, 509–527, 2017.

Oyama, Y., Hato, E., Prism-based path set restriction for solving Markovian traffic assignment problem. Transp. Res. Part B: Methodol. 122, 528-546, 2019.

Russell, S.: Learning agents for uncertain environments, In: Proceedings of the 11th annual conference on computational learning theory, 101-103, 1998.

Wulfmeier, M., Ondruska, P., Posner, I. Maximum entropy deep inverse reinforcement learning. arXiv preprint arXiv:1507.04888, 2015.

Ziebart, B. D., Maas, A. L., Bagnell, J. A., Dey, A. K., Maximum entropy inverse reinforcement learning. In: Proceedings of the Association for the Advancement of Artificial Intelingence (AAAI), 1433–1438, 2008.

Ziebart , B. D., Ratliff, N., Gallagher, G., Mertz, C., Peterson, K., Bagnell, J. A., Hebert, M., Dey, A. K., Srinivasa, S., Planning-based prediction for pedestrians. In IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), 3931–3936, 2009.

# 3-5. 数値実験

## 実験による検証①

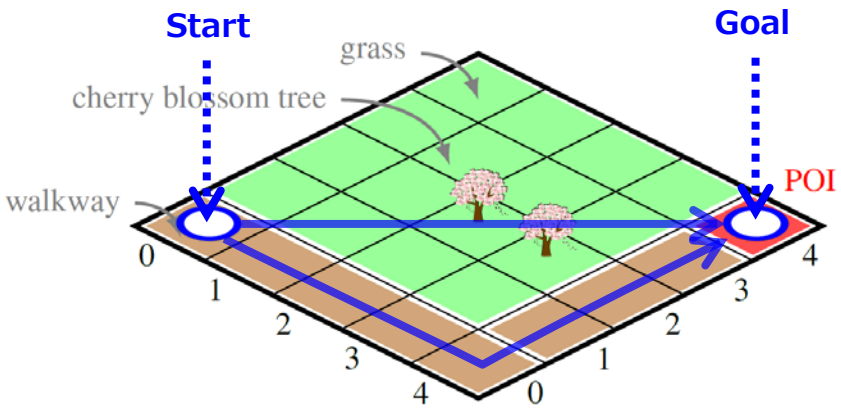
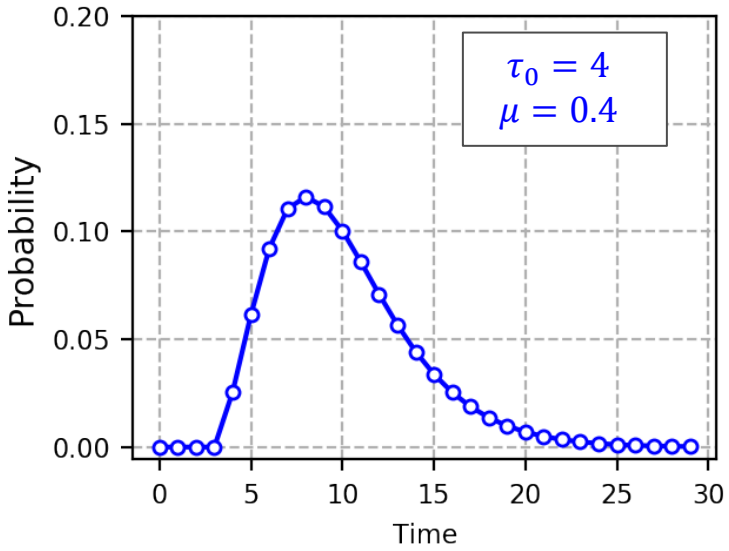



Figure 1: Problem settings




設定した到着時間制約分布

$$[\theta_{\text{walkway}}, \theta_{\text{cherry}}, \theta_{\text{PoI}}] = [2,2,4]$$

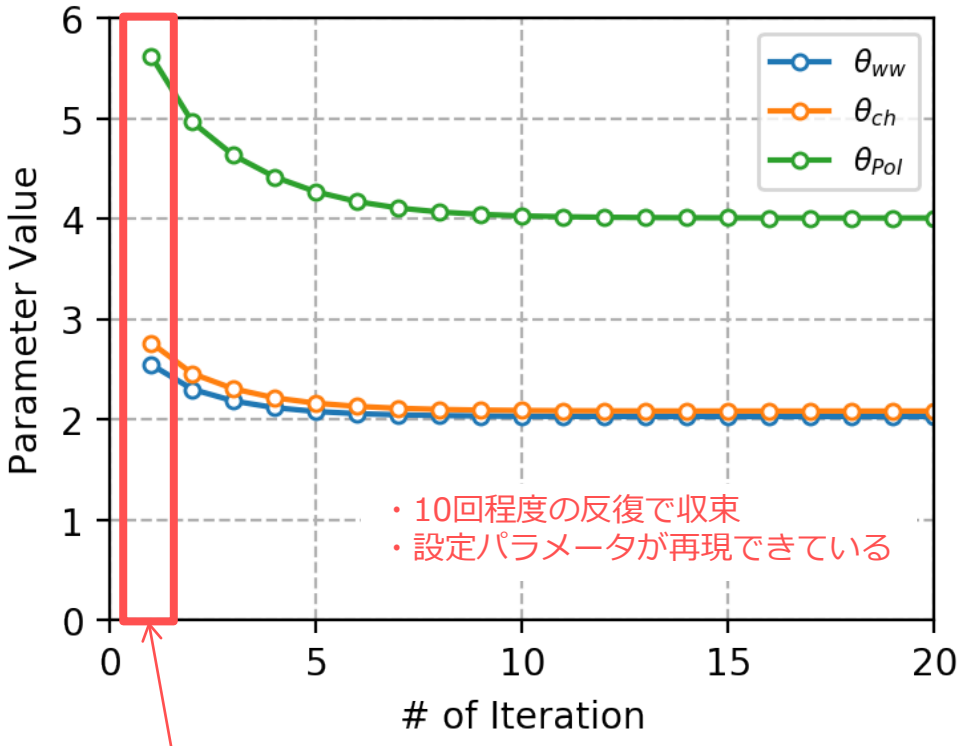
$$\bar{\tau} = \frac{\tau_0}{\mu} = 10$$

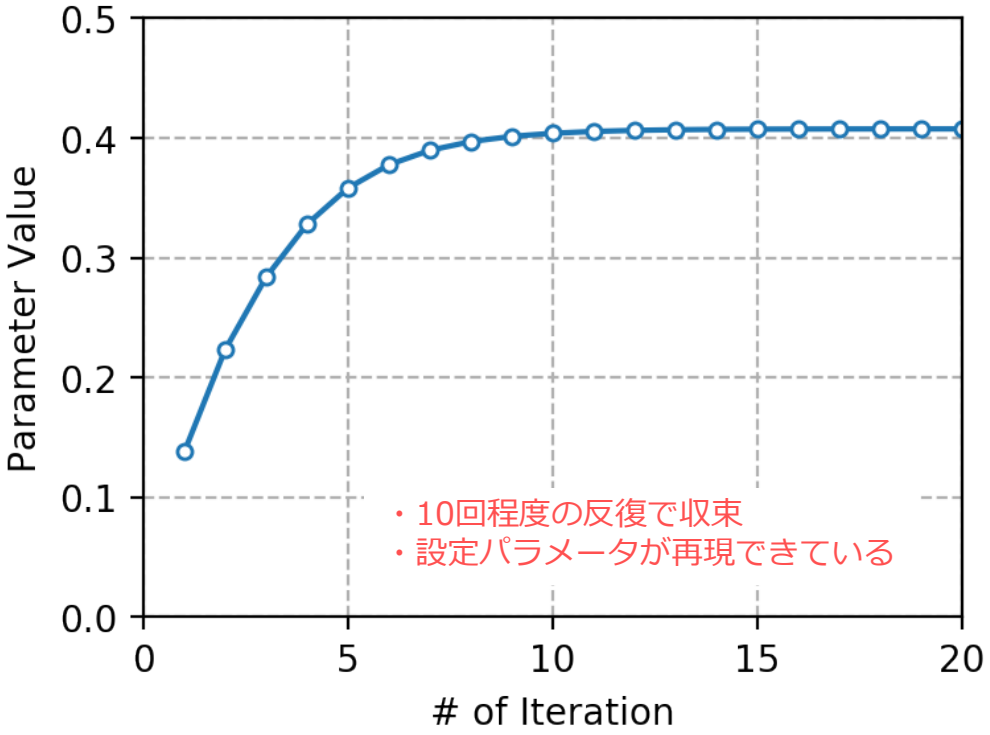
到着時間制約の平均
10ステップ

経路データのみから到着時間制約分布のパラメータμと特徴量の重みθを推定できるか検証

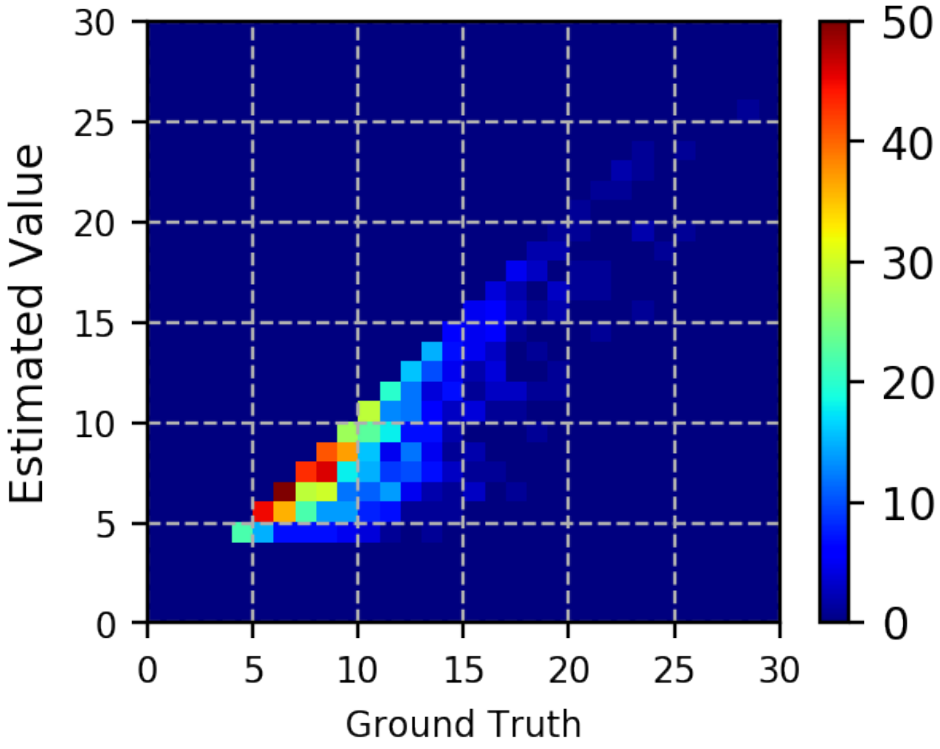
# 3-5. 数値実験

・$N = 1000$
・計算する$\tau$は30ステップまで（データの長さも全て30ステップ分）
・負担率の初期値 $z_{n,30} = 1$

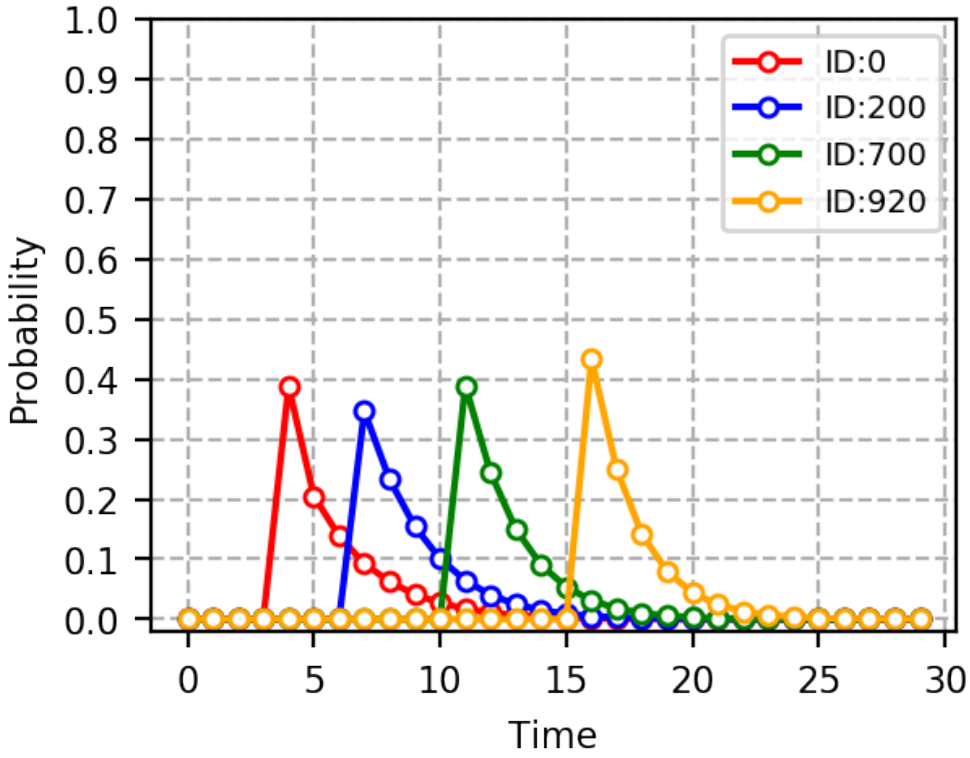


# 3-5. 数値実験




到着時間制約の比較
（推定値は$\tau_{est} = \mathrm{argmax}_{\tau}\, \gamma(z_{n\tau})$）




$\gamma(z_{n\tau})$の例

# 3-5. 数値実験

## 実験による検証②

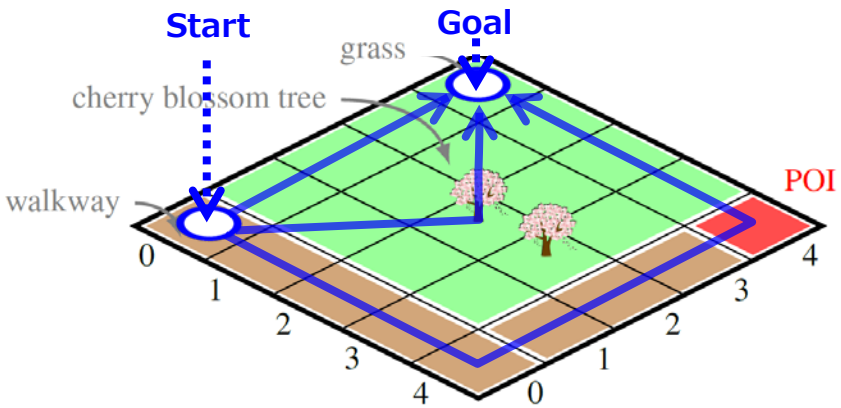
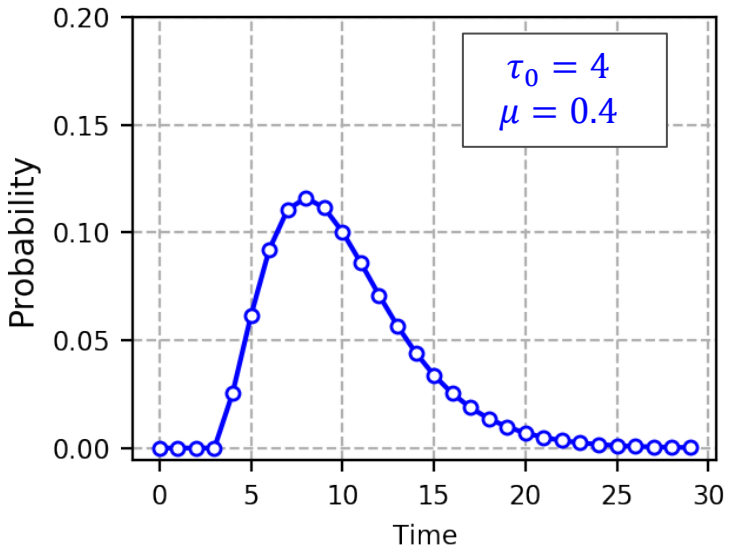



Figure 1: Problem settings




設定した到着時間制約分布

$$[\theta_{\text{walkway}}, \theta_{\text{cherry}}, \theta_{\text{PoI}}] = [2,2,4]$$

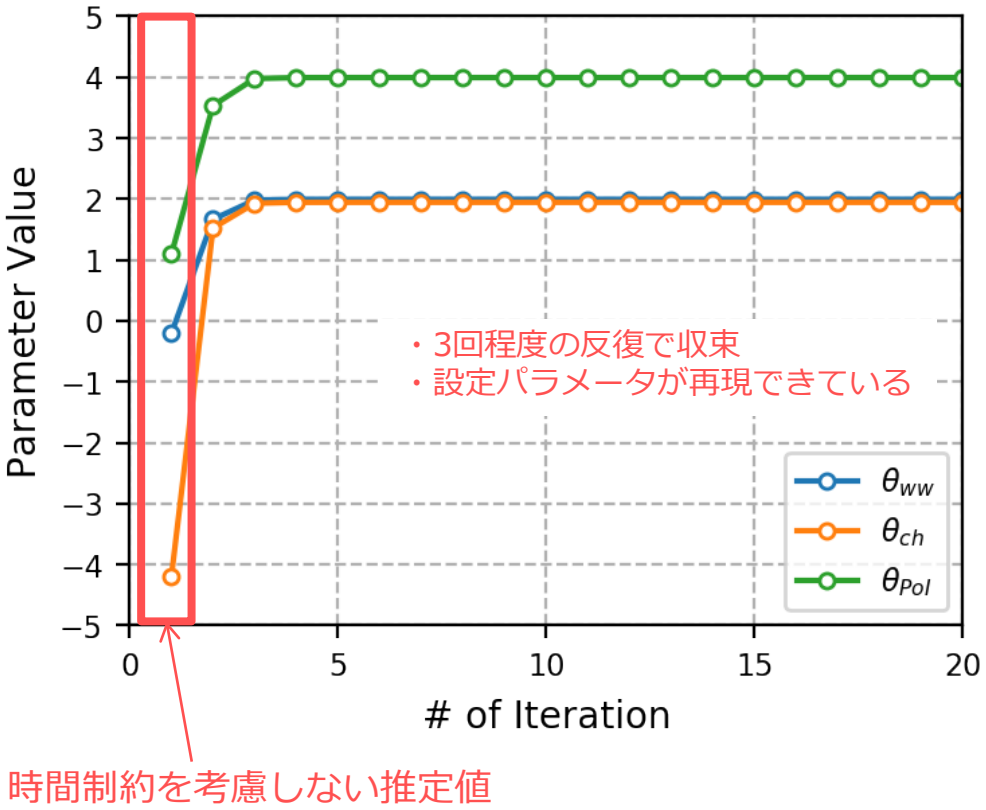
$$\bar{\tau} = \frac{\tau_0}{\mu} = 10$$

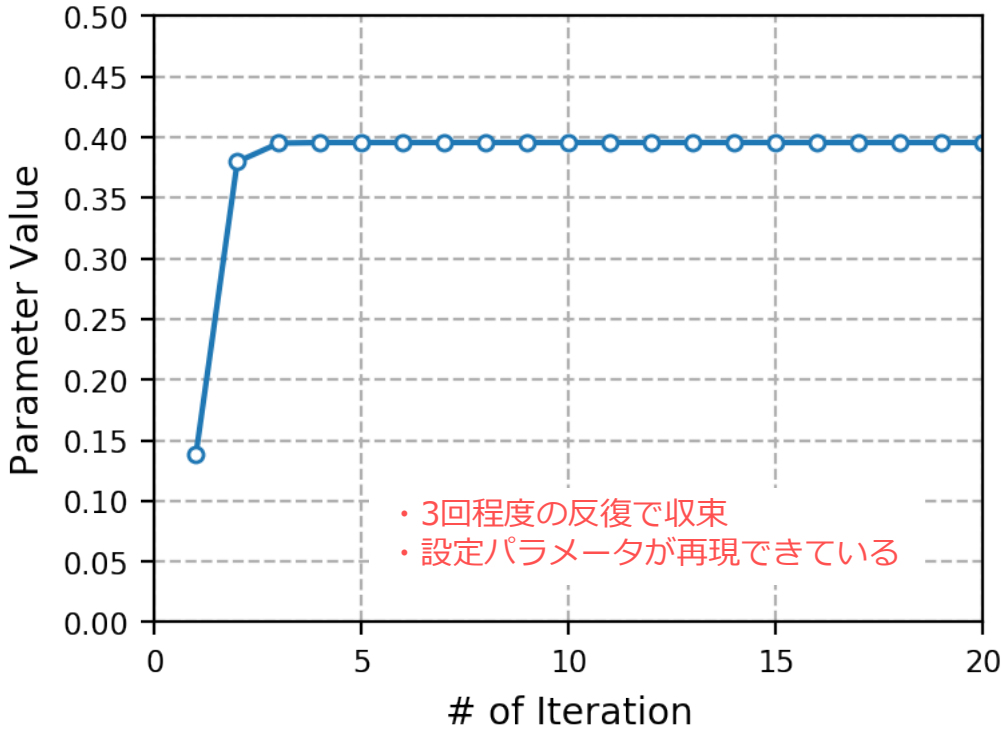
到着時間制約の平均
10ステップ

経路データのみから到着時間制約分布のパラメータμと特徴量の重みθを推定できるか検証

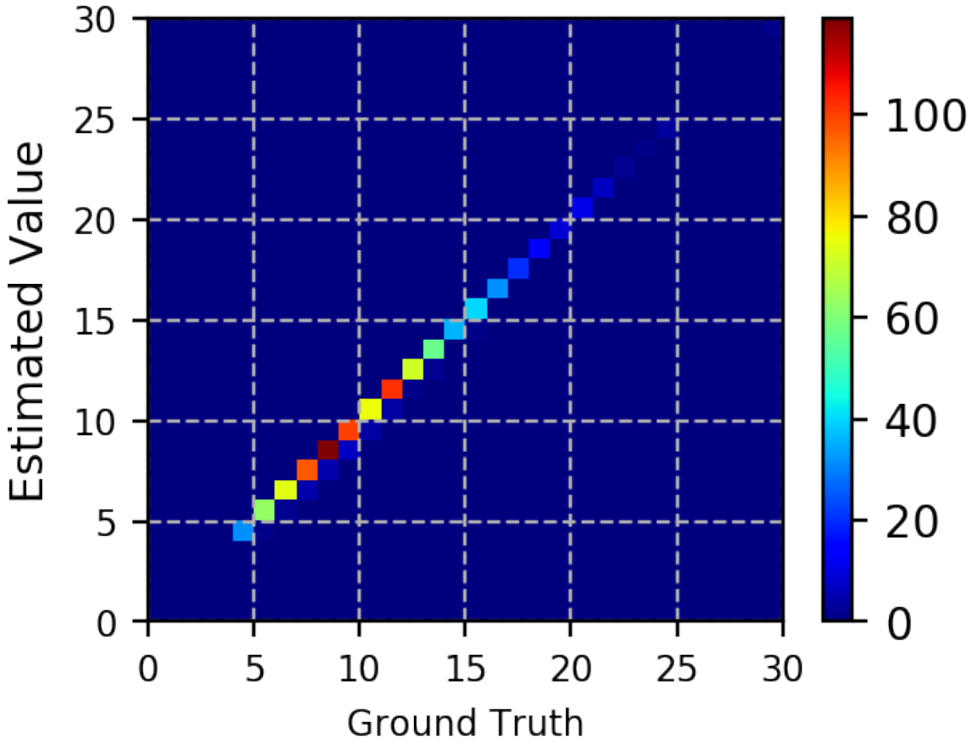
# 3-5. 数値実験

・$N = 1000$
・計算する$\tau$は30ステップまで
・負担率の初期値 $z_{n,30} = 1$

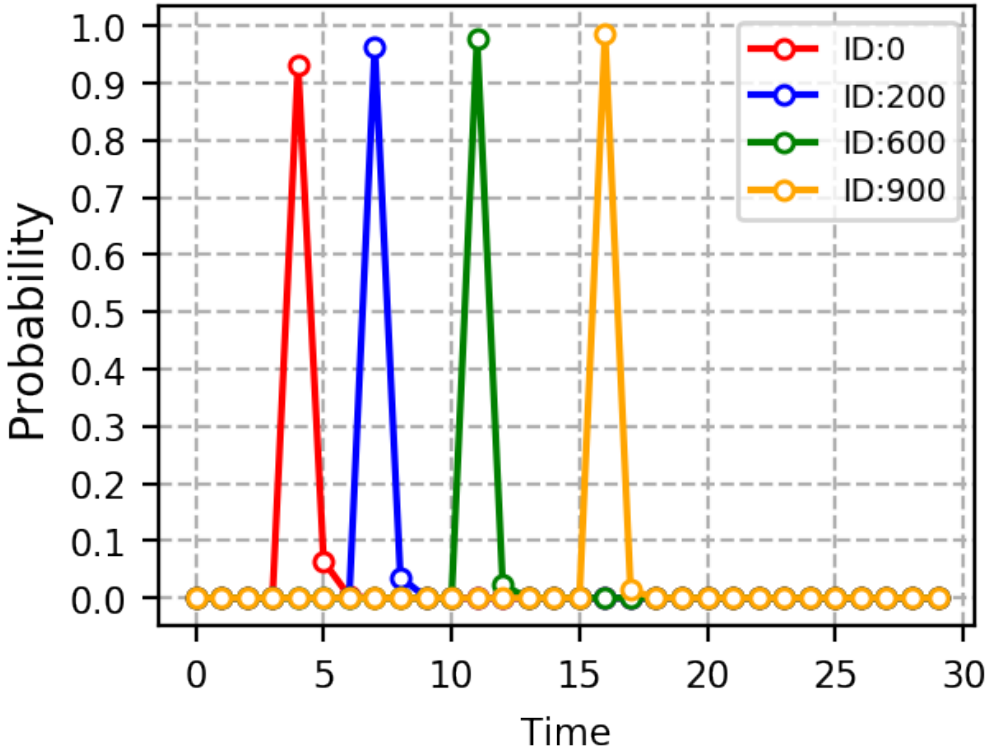


# 3-5. 数値実験

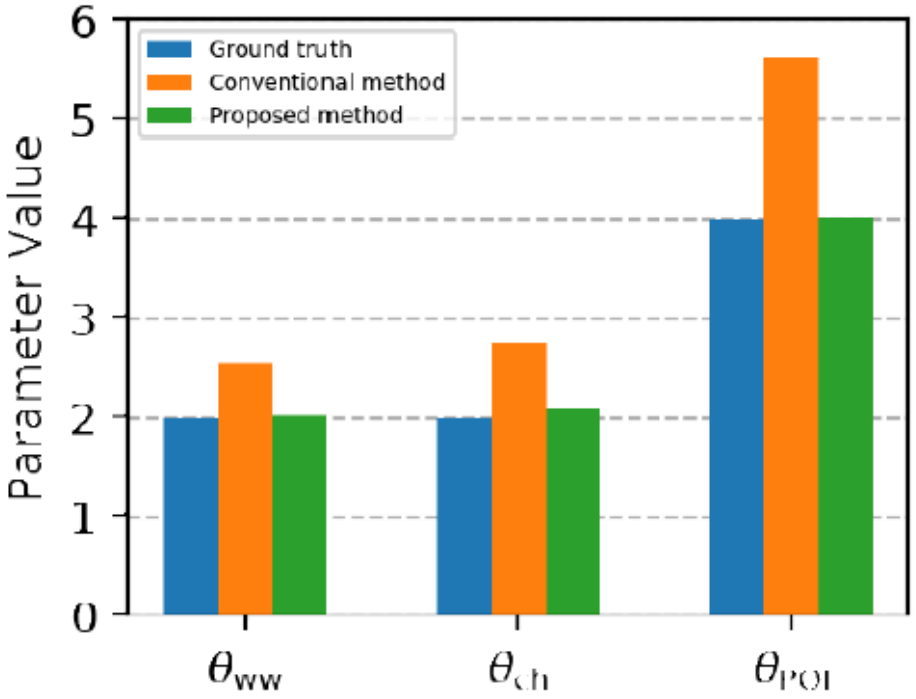



到着時間制約の比較
（推定値は$\tau_{est} = \mathrm{argmax}_{\tau}\, \gamma(z_{n\tau})$）

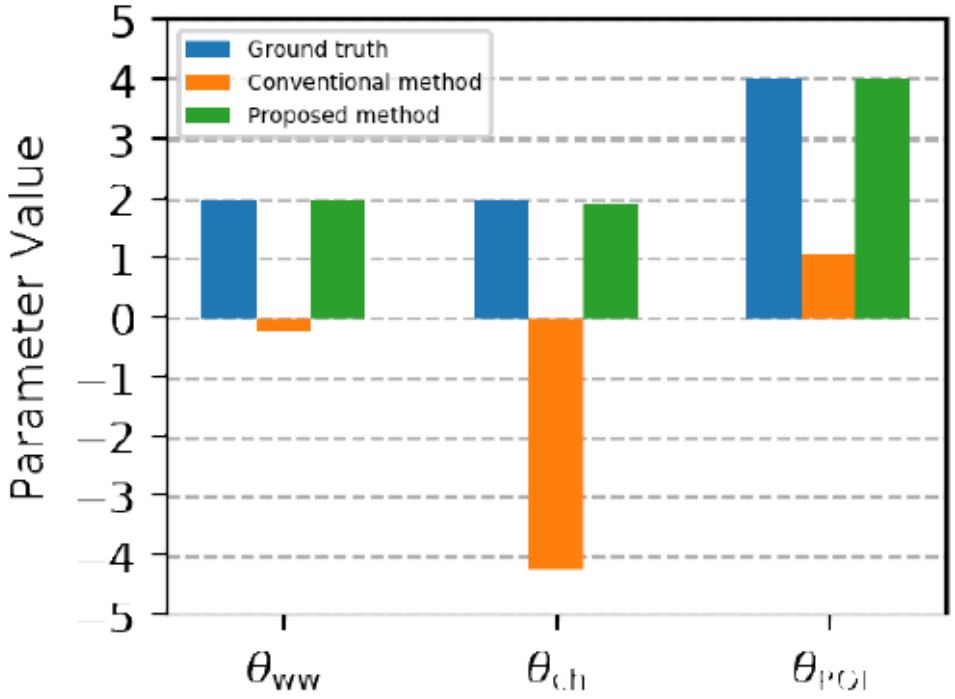



$\gamma(z_{n\tau})$の例

# 3-5. 数値実験




(a) Experiment 1




(b) Experiment 2

どちらの実験でもバイアスのない推定ができた

# （付録）到着時間制約下の特徴量期待値の求め方

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

時間制約$\tau$の下での特徴量期待値

時間制約$\tau$の下でのSVF

到着時間制約

$$SVF_\tau=\sum_t P(s_t|s_0,s_\tau)$$

$$P(s_t|s_0,s_\tau)=\frac{P(s_\tau|s_0,s_t)P(s_0,s_t)}{P(s_0,s_\tau)}$$

$$=\frac{P(s_\tau|s_t)P(s_t|s_0)P(s_0)}{P(s_0,s_\tau)}$$

$$=\frac{P(s_\tau|s_t)P(s_t|s_0)}{P(s_\tau|s_0)}$$

$$=\frac{\alpha(s_t)\beta(s_t)}{\beta(s_0)}$$

**forward probability**

$$\alpha(s_t)\equiv P(s_t|s_0)$$

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

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

**backward probability**

$$\beta(s_t)\equiv P(s_\tau|s_t)$$

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

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

# （付録）到着時間制約下の特徴量期待値の求め方

まとめると

時間制約$\tau$の下でのSVF

$$SVF_\tau = \sum_t \frac{\alpha(s_t)\beta(s_t)}{\beta(s_0)}$$

時間制約のない一般のSVF

$$SVF = \sum_t \alpha(s_t)$$