DEIM2020 I2-3 (day1 p59)

# 強化学習を用いた健康行動促進における介入戦略の学習

高橋　公海† 　幸島　匡宏† 　倉島　健† 　戸田　浩之†

† 日本電信電話株式会社 NTT サービスエボリューション研究所　〒 239–0847 神奈川県横須賀市光の丘 1–1
E-mail: †{masami.takahashi.xh,masahiro.kohjima.ev,takeshi.kurashima.uf,hiroyuki.toda.xb}@hco.ntt.co.jp

**あらまし**　生活習慣病の増加は社会的な課題となっており，生活リズムを整えるなど健康行動を促進することが予防に効果的であると知られている．規則正しい食生活や，十分な睡眠を確保することは心身の健康維持に寄与するため，我々は生活リズムを整えるための支援技術について検討を行っている．本研究では，ユーザの理想とする目標 (ユーザが設定した時間と行動) を達成できるように，いつどの行動をすると良いかという方策を学習する手法を提案する．モデルベース強化学習を用いて，ユーザの生活パターンと介入が受容される可能性を考慮した上で介入戦略を学習する手法を構築した．実験では 34 人の参加者から 2ヶ月間行動と受容に関するログを収集し，それらの実データを用いて手法の有効性を検証した．

**キーワード**　強化学習，ヘルスケア，行動変容，健康行動促進，Human-agent Interaction

## 1 はじめに

生活習慣病の増加は社会的な課題であり，その多くは不健全な生活の積み重ねが要因となっている [1]．病気になる前にライフスタイルを見直し，十分な睡眠・適度な運動・規則正しい食生活など健康的な習慣を身に付けることが予防に効果的であると知られている [2]．最近では健康行動を促すアプリケーションやデバイスも普及つつあり，活動量や心拍などを測定可能なセンサを搭載したウェアラブルデバイスは身近なものとなっている．これにより，健康に関するデータの観測・収集や，情報を通知することが容易になった．

例えば Fitbit [1]では 1 時間の歩数が 250 歩以下の場合に運動を促すリマインダや設定された時刻に睡眠を促すリマインダ，Apple watch [2]にはリラックスするよう深呼吸を促す呼吸リマインダなどがある．しかしながら，ユーザに行動を促す (介入する) これらのリマインダは比較的シンプルなルールに基づいており，図 1 に示すように上手くいかない例もある．例えば，ユーザが十分な睡眠時間を確保するために，いつもより 1 時間早く就寝したい場合について考える (図 1 (a))．「夜 11 時に寝る」ことをユーザの目標とした場合，目標とする時間の少し前にリマインダを設定し，寝る時間を知らせるといった方法が考えられる．しかし，普段の生活パターンのまま寝る時間だけを早くしようと試みても，通知を受けた時点でいつも寝る前に行っている一連の行動が終わっていない等の理由で，通知に従って行動することが難しく上手くいかない (図 1 (b))．効果的な介入を行うためには，ユーザが目標としている行動だけを変えようと介入するのではなく，目標達成に影響する他の行動についても考慮する必要がある．

そこで本研究では，強化学習 [3] を用いてユーザの目標達成を支援するための介入戦略を学習する手法を提案する．これは，ユーザの生活パターンと介入が受け入れられる可能性を考慮し，目標 (ユーザが設定した時間と行動) を達成するためにいつどの行動を促すよう介入すると良いか，という方策を学習するものである．我々が強化学習を用いる理由は，現在の決定が将来に及ぼす影響を考慮した上でプランニングを行い，今何をすべきか判断することが出来るためである．例えば囲碁のようなボードゲームでは，勝利することが目標となるがそこに至る過程は自明ではなく，現在の盤面でどこに石を置くと最終的に勝利する可能性が高くなるか判断する必要がある．強化学習はそういった判断に適したアプローチであり，AlphaGo [4] が与えられた目標 (勝利) のみから人間に勝利する手筋を見出すように，ユーザが最終的に目標を達成出来るような最適な介入戦略を見つけることが出来ると考えられる．本問題では，強化学習における状態・行動・報酬を，それぞれユーザの行動・アプリケーションによる介入・ユーザの目標に対応させる (図 2)．図 1 (c) は提案手法の適用例を示したものであり，ユーザに 18 時頃夕食を食べるよう促すことでユーザの将来の状態を変え (いつもより早い時間に入浴)，ユーザが目標とする午後 11 時に寝るという目標を達成出来るように支援する．また，ユーザが介入に従うことが出来なかった場合でも，強化学習では逐次的に判断し，次の時間ステップでユーザが目標に追いつくことが出来るよう介入することが可能である．

我々は強化学習をベースとしたアプローチの効果を検証するため，収集した実データに基づいてユーザのシミュレーションモデルを構築し，評価を行った．ユーザに直接介入するのではなくシミュレーションを行った理由は，強化学習は一般に膨大な試行錯誤の回数を要することが知られており，十分な試行回数を行う前にユーザが離脱してしまい，強化学習の効果を検証することが難しくなることが予測されるためである．また，事前に収集した少量のユーザデータを手掛かりにシミュレータを構築しエージェントを訓練することは，今後ユーザへ介入する際にも同様の手順が想定されるため，今回このような評価方法とした．本実験は次の手順で行った．(i) 2ヶ月間 34 名の参加

1 : https://www.fitbit.com/
2 : https://www.apple.com/watch/

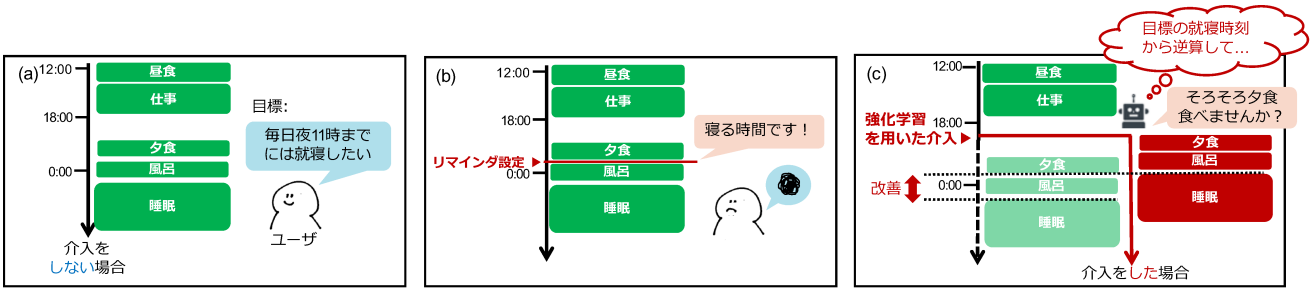



図 1: ユーザがいつもより早く就寝することを目標とした例．(a) ユーザが「夜 11 時，就寝」と目標を設定する．(b) 23 時にアラームを設定するようなシンプルな介入の場合，普段寝る前に行う行動を終えていなければ，介入に従って行動することが難しい．(c) は目標とする時間に就寝できるように逆算して，寝る前に行う一連の行動を早めるようユーザに促して目標達成を支援する．

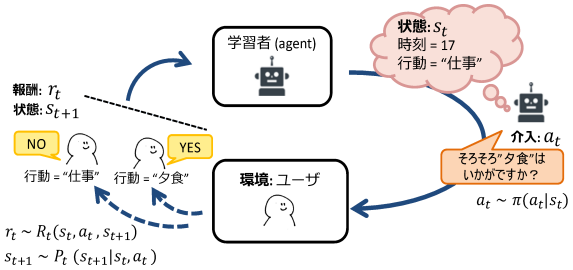



図 2: エージェントと環境 (ユーザ) とのインタラクション．

者から，いつ何を行っていたかを記録した「行動ログ」と呼ぶライフログデータを収集した．同時に，行動ログだけではユーザが介入を受けた場合の反応が分からないため，「許容ログ」と呼ぶユーザが行動を行う時間を変えても良いと考える時間範囲を記録したデータも収集した．(ii) 行動ログと許容ログに基づいて，ユーザの日常行動を模擬したシミュレーションモデルを推定した．このモデルはユーザの現在の行動・時間・介入が与えられた時に，次にユーザがとる行動を出力するものである．(iii) 強化学習により，シミュレーションされたユーザに対する最適な介入方策を求めた．ここで我々が用いる強化学習は一般にモデルベースの強化学習 [3] と呼ばれ，環境モデル (遷移確率) を推定した後に，価値反復によって最適な方策を計算するアプローチである．

実験ではユーザへの介入効果について，実データから各行動の開始時刻の最頻値を算出し，報酬関数の目標時刻をユーザ毎の就寝時刻の (a) 最頻値，(b) 最頻値 −1 時間，(c) 最頻値 −2 時間として検証した結果を示す．提案手法とベースライン (ランダムな介入およびアラーム設定) の平均報酬和を比較したところ，いずれの目標設定においても提案手法の方が高い報酬和を得ており，単純なアラーム設定よりも提案手法のように前もって介入を行うことが効果的であることを確認した．

本研究の貢献は，以下の通りである．

（1） 生活習慣の改善に向けて，強化学習を用いて自動的に介入する手法を提案した．

（2） 2ヶ月間 34 名の実データを収集し，ユーザの行動をシミュレーションするモデルを構築した．

（3） 介入の効果についてシミュレーションによる検証を行い，提案手法はランダムな介入や単純なアラーム設定よりも目標達成に効果的な介入を実現出来ていることを確認した．

本論文の構成は以下の通りである．2 章で関連研究，3 章で収集したデータについて述べる．4 章では提案手法について説明する．5 章でシミュレーションによる検証結果，6 章ではまとめと今後の課題を示す．

## 2 関 連 研 究

健康行動を促すアプリケーションやデバイスは様々なものがあり，活動量・心拍・体温・呼吸数などを測定可能なセンサを搭載したウェアラブルデバイスは身近なものとなっている．これらはユーザの健康に関するデータを収集し，情報を通知することを容易にした．健康行動を促すための介入方法については様々な研究が行われており，Daskalova ら [5] は，ウェアラブルデバイスで蓄積した睡眠データの分析やコホート研究に基づく睡眠習慣を改善するための情報推薦を行っている．実験から得られた知見によると，推薦に従うことが難しかったユーザに共通する理由として，スケジュールに合わなかったことが挙げられている．我々はこの点に着目して研究を進めているが，我々の過去の研究 [6] では初歩的な分析に留まっており手法が確立されていなかった．本研究は [6] のコンセプトをモデルベース強化学習を用いて実現するものである．

強化学習はゲームや自動運転，ロボットの制御，自然言語処理，コンピュータビジョン，信号制御など様々な分野で活用され注目を集めており [7] [8] [9] [10] [11]，本研究でも提案手法に強化学習を用いている．強化学習にはモデルベース学習とモデルフリー学習という 2 つのアプローチがあり，医療・健康などヘルスケアの分野でもその両方が用いられている．例えば，敗血症や統合失調症，糖尿病など様々な病気において長期的な視点で患者の結果が良くなるような意思決定や，健康行動を促進する介入戦略の学習に活用されている [12] [13] [14]．基本的にシミュレータや大量のデータなどを利用できる場合にはモデルフリー学習，それ以外の場合はモデルベース学習を採用する傾向がある．[15] や [16] では，モデルフリー学習でパーソナライズされた介入を行う手法を提案し，3 種類のユーザを仮定してシミュレーションを行い評価している．この研究ではユーザのクラスタリングを行うことで学習を早めることに成功しているが，ユーザにランダムに介入したデータが必要となるため，現実的にはそのような介入を受けている間にユーザの離脱を招く可能性が高い．我々のアプローチでは，ユーザが行動を変えることが可能だったかもしれないと考えた許容ログを用いて遷移確率を推定する手法を構築したため，ランダムに介入した履歴は不要である．

## 3　データ収集と基礎分析

### 3.1　参　加　者

医師から生活習慣病を含め健康上の指導を受けていない 20 代以上の参加者を 40 名募集した．募集及び事前説明会においては，日常行動のモデル化を目的としたライフログデータ収集であり，カレンダーアプリを利用して普段通りの生活を記録するように伝えている．ただし，記録のない日が 5 日以上連続する場合や，2ヶ月のうち記録した日が 50 日未満の場合には辞退したと見なしている．以降では，辞退者を除いた 34 名分のデータで分析を行う [3]．

### 3.2　収 集 方 法

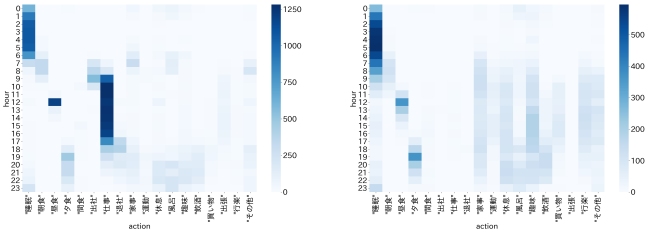
入力データは行動ログおよび許容ログから構成され，参加者は表 1 に示す 15 種類の行動を対象に記録を行う．図 6 (a) は入力例を示している．記録内容は「行動の種類」「開始時刻」「終了時刻」であり，カレンダーアプリケーションに予定として投入するよう依頼した．今回収集対象とした行動は次の 15 種類である．睡眠，朝食，昼食，夕食，間食，出社，仕事，退社，家事，運動，休息，風呂，趣味，飲酒，買い物．時間の粒度は 15 分程度を目安とした．記録の際には，仕事をしながら間食をとる，のように 1 つの時間帯に複数の行動が存在しても良いものとした．また，対象としている行動以外を行っている時間帯は空白として良いと伝えた．

行動ログは実際に行った行動を記録していくため比較的理解しやすいが，許容ログについては日頃意識しない概念であり，事前説明会及び実験の初期段階で考え方や入力時の注意事項などフォローを行った．許容ログを入力する際には次の 2 点について考えるよう促した．まず，仮に昼食を 11:30〜12:30 の時間でとっていた時に，その日一日を振り返って「11:15〜12:30 の時間幅の中であれば，昼食をとる時間を動かすことが可能だったかもしれない」と考えられる場合は許容可能な時間を「11:15〜12:30」と記録する (図 6 (a))．これにより，我々は参加者が昼食をとる時間帯の変更を許容できる時間幅を知ることができる．また，その日行っていない行動 (例えば運動) であっても，「19:00〜19:30 の時間帯ならば運動することが可能だったかもしれない」と考えられる場合は「19:00〜19:30」と記録する．許容ログは 1 日の行動の記録を振り返り，出来るだけ当日もしくは翌日に入力するよう依頼した．入力の負荷に関しては終了後にアンケートを行なっており，慣れるまで 1 週間程度かかったが慣れてきたら気にならないと述べる参加者が多かった．

### 3.3　データ分析

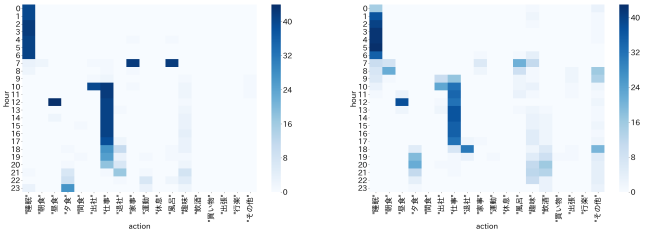
まず，収集した行動ログから各行動がどの時間帯に行われたかを平日と休日に分けカウントしたものを図 3 に示す．例えば 12:00〜12:45 に昼食をとった場合は，12 時に”昼食”が 1 回

3：参加者には事前説明会の際に実験の説明書および同意書を配布し，実験を通じて収集したデータは個人を特定出来ないよう加工した統計情報として扱うことや，途中で辞退の自由が保証されることなど，内容の確認を促し同意を得た．




(a) 平日の行動ログ (全員)　(b) 休日の行動ログ (全員)

図 3: 参加者全員の行動ログを (a) 平日と (b) 休日に分け，どの時間に何回行動が出現していたかカウントしたもの．縦軸は時間，横軸は行動を示す．色が濃いほど出現回数が多い．




(a) 平日の行動ログ (ID:9)　(b) 平日の行動ログ (ID:33)

図 4: 各参加者 (a) ID:9 と (b) ID:33 の平日の行動ログについて，どの時間に何回行動が出現していたかカウントしたもの．

出現したとしてカウントする．図 3 では 1 時間刻みで集計を行い，参加者全員分のログを合計した回数を可視化している．平日 (図 3a) と休日 (図 3b) を比較すると，出社・仕事・退社のように明らかに平日だけに出現する行動があると分かる．食事のように平日・休日問わず行われる行動を見ると，平日よりも休日の方が行われる時間帯にブレがあることが分かる．例えば昼食は平日 12 時に集中しているが，休日は 12 時を中心に前後 11 時〜13 時にも行われている．次に，参加者毎の傾向を図 4 に示す．いずれも平日の行動ログを集計したものだが，ID:9 (図 4a) と ID:33 (図 4b) を比較すると，参加者によって出現する行動や時間帯が異なる．例えば運動は ID:9 のみ，朝食・飲酒は ID:33 のみで出現する行動である．また，同じ行動であっても ID:9 では夕食が 23 時，ID:33 では 20 時が中心となっているなど行われる時間帯は異なっていることが分かる．このように曜日や参加者によっていつどの行動を行うかという傾向が異なるため，5 章で述べる実験では平日のデータだけを利用し，参加者一人一人のデータから個別にシミュレーションモデルを構築し各人に適した方策を求めた．

## 4　提 案 手 法

提案手法はモデルベースの強化学習 [3] を用いて自動的に先を見越した介入を行うもので，環境のモデル (遷移確率) を推定し，後ろ向き帰納法アルゴリズムで最適な方策を求める．概要を図 5 に示す．我々は以下の MDP の定義に基づいて，エージェントが最適な方策を獲得できるよう提案手法を設計した．

### 4.1　有限期間非斉時的マルコフ決定過程

ユーザとエージェント間のインタラクションは有限

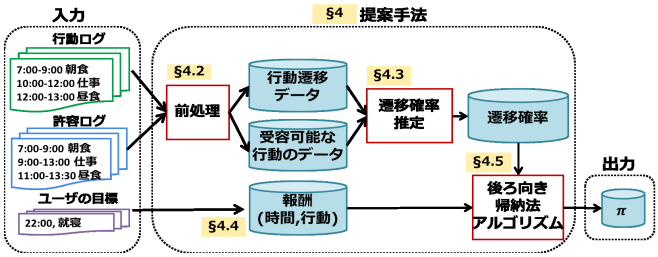



図 5: 提案手法の概要．入力は行動ログ，許容ログ，ユーザの目標達成に応じて定められた報酬である．行動ログと許容ログから遷移確率を推定し，後ろ向き帰納法アルゴリズムで得た最適な方策 $\pi$ に基づいてユーザへの介入を決定する．

表 1: 状態，行動 (介入)，報酬

| | |
|---|---|
| State | 1: 睡眠, 2: 朝食, 3: 昼食, 4: 夕食, 5: 間食,<br>6: 出社, 7: 仕事, 8: 退社, 9: 家事,<br>10: 運動, 11: 休息, 12: 風呂, 13: 趣味, 14: 飲酒,<br>15: 買い物 |
| Action | 1: 睡眠, 2: 朝食, 3: 昼食, 4: 夕食, 5: 間食,<br>6: 出社, 7: 仕事, 8: 退社, 9: 家事,<br>10: 運動, 11: 休息, 12: 風呂, 13: 趣味, 14: 飲酒,<br>15: 買い物，16: 何も介入しない (介入無し) |
| Reward | ユーザの目標 (例：夜 10 時に寝る) |

期間非斉時的マルコフ決定過程 [17] でモデル化し，$\{\mathcal{S}, \mathcal{A}, \{\bar{\mathcal{P}}_t\}_{t=0}^{T-1}, \{\mathcal{R}_t\}_{t=0}^{T-1}\}$ の 4 つ組で定義する．

***State:*** $\mathcal{S} = \{1, 2, \cdots, |\mathcal{S}|\}$ は状態の有限集合である．本研究における状態とはユーザが行っている行動を指し，次の 15 種類を対象とした：睡眠，朝食，昼食，夕食，間食，出社，仕事，退社，家事，運動，休息，風呂，趣味，飲酒，買い物 (表 1)．例えば，「8 時 朝食」「10 時 仕事」「12 時 昼食」が状態となる．

***Action:*** $\mathcal{A} = \{1, 2, \cdots, |\mathcal{A}|\}$ は行動の集合であり，本研究においてはユーザへの介入を行動とする．$|\cdot|$ は集合中の要素の数を示しており，$T$ は行動の長さを示す．行動空間 $\mathcal{A}$ はエージェントがユーザに表 1 中のいずれかの行動を促す介入をする，もしくは介入を行わないという変数で構成される．i.e., $|\mathcal{A}| = |\mathcal{S}| + 1$．エージェントは介入を選択した後，ユーザに「夕食はいかがですか？」「寝る時間です」といったメッセージを提示することを想定している．

**遷移確率:** $\mathcal{P}_t$ は遷移確率，$\mathcal{P}_t(j|i,k)$ は行動 $k$ が $t$ 番目の時間に行われた時に状態 $i$ から状態 $j$ へ遷移する確率を示している．これは，介入有りもしくは無い場合のユーザの行動遷移を定義している．もし $\mathcal{P}_t(j|i,j)$ の値が大きければ，ユーザは $j$ という行動を行うよう促すエージェントの介入に従う確率が高く，次の時間に行動 $j$ に遷移する．もしユーザへの介入が行われない場合には，ユーザの次の行動は $\mathcal{P}_t(j|i, k = 16\text{: none})$ によって決定する．このように非斉時的な MDP を用いることで，朝や夜のように時間帯に依存したユーザの行動遷移をモデル化することができる．

***Reward:*** $\mathcal{R}_t$ は報酬関数であり，$\mathcal{R}_t(i,k,j)$ は $t$ 番目の時間に介入 $k$ を行うことで状態 $i$ から状態 $j$ に遷移する時に得られる報酬である．既に述べたように，我々は時間に依存した遷移確率と報酬関数を用いる．ユーザが目標とする時間に目標の行動を実施出来た場合に，正の報酬を得るものとして設計する．

図 2 は我々の設定における環境とエージェントとのインタラクションを示しており，最終的にユーザへの最適な介入方策を獲得することが目的である．$t$ 番目に状態 $i$ にいる時に介入 $k$ を行う確率を $\pi_t(k|i)$ とし，$\boldsymbol{\pi} = \{\boldsymbol{\pi}_t\}_{t=1}^{T-1}$ を方策とする．方策 $\boldsymbol{\pi}$ の時の $T$ ステップ目の遷移履歴 (エピソード)$h_T$ は，$P(h_T|\mathcal{P}, \boldsymbol{\pi}) = q(s_0)\prod_{t=0}^{T-1}\pi_t(a_t|s_t)\mathcal{P}_t(s_{t+1}|s_t, a_t)$, で与えられる．ただし，$s_t$ と $a_t$ は $t$ 番目の時間の遷移した状態と行われた行動をそれぞれ示す．$q(i)$ は状態 $i$ の初期状態の確率を示す．

### 4.2 前 処 理

図 6 に示すようなユーザの $L$ 日間の行動ログと許容ログを想定する．行動ログから作成した行動遷移データを $\mathcal{D}_{tr} = \{N_{tij}\}_{ij\in\mathcal{X}}$ とし，$N_{tij}$ は行動ログ中の $t$ 番目の期間における行動 $i$ から行動 $j$ への遷移回数とする．同様に許容ログから作成した受容可能な行動データを $\mathcal{D}_{apt} = \{M_{tj}\}_{ij\in\mathcal{X}}$ とし，$M_{tj}$ は許容ログの $t$ 番目の期間において行動 $j$ が記録された回数とする．これらの集合を $\mathcal{D} = \mathcal{D}_{tr} \cup \mathcal{D}_{apt}$ とする．

### 4.3 遷移確率推定

本研究においてモデルベース強化学習を採用する難しさは，$(s_t, a_t, s_{t+1})$ から成るユーザに実際に介入した履歴データの入手が困難な点にあり，通常の方法では遷移確率の推定が出来ない．そのため我々は行動ログと許容ログを構成要素に組み込み，遷移確率を推定する手法を構築した．我々の手法は，(i) 実行しやすい介入 (当該時間の許容ログに出現する行動を促す等) を受けた場合には，行動遷移する確率が上がる．(ii) 従うことが現実的ではない介入 (就寝中に昼食をとるよう促す等) を受けた場合には行動は遷移しないよう，振る舞うことが望ましいと考えて設計している．推定した遷移確率と報酬関数を用いて，最適な方策を獲得することが可能である．

MDP の遷移確率をモデル化するため，パラメータ $\theta = \{\boldsymbol{u}, \boldsymbol{v}\}$ を導入する．パラメータの依存度を強調するため，遷移確率のモデルを $\{\mathcal{P}_t^\theta\}$ とする．今回は下記の対数線形モデルを用いる．

$$\mathcal{P}_t^\theta(s_{t+1} = j|s_t = i, a = k) = \begin{cases} \frac{\exp(v_{tij})}{\sum_{j'}\exp(v_{tij'})} & (\text{if } k = |\mathcal{A}| : \text{介入無し}) \\ \frac{\exp(v_{tij}+u_{tk})}{\exp(v_{tik}+u_{tk})+\sum_{j'\neq k}\exp(v_{tij'})} & (\text{if } j = k, k \neq |\mathcal{A}|) \\ \frac{\exp(v_{tij})}{\exp(v_{tik}+u_{tk})+\sum_{j'\neq k}\exp(v_{tij'})} & (\text{if } j \neq k, k \neq |\mathcal{A}|). \end{cases}$$

直感的には，パラメータ $\boldsymbol{v}$ は介入無しの遷移確率で定義されるが，$\boldsymbol{u}$ は介入の効果を示す．介入は，もし $u_{tj}$ が大きい場合には次の時間に行動 $j$ のトリガになる．以降では，$\mathcal{P}_t^\theta(s_{t+1} = j|s_t = i, a = k)$ を省略して $p_{tijk}^\theta$ と書く．

行動遷移データは介入が行われていない場合のユーザの行動遷移を表しているため，上記のモデルを用いることで尤度関数が次のように与えられる．

$$P(\mathcal{D}_{tr}|\theta) = \prod_{t=1}^{T}\prod_{i,j\in\mathcal{S}}\left(p_{tij|\mathcal{A}|}^\theta\right)^{N_{tij}}.$$

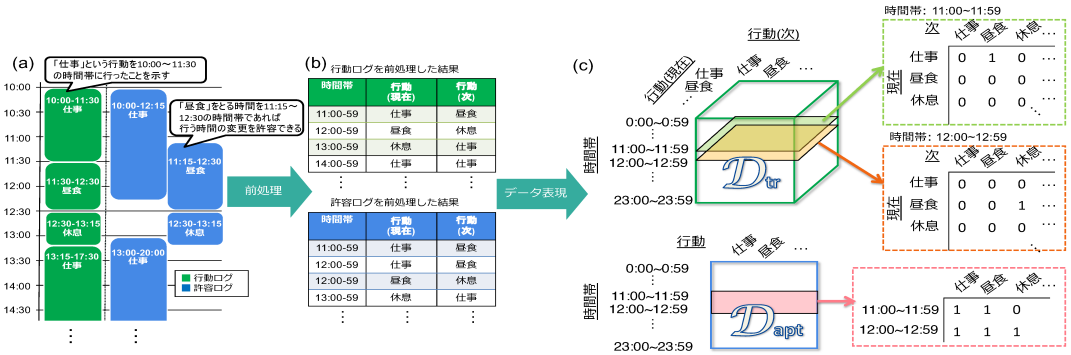



図 6: 行動遷移データ及び受容可能な行動のデータを作成する前処理の概要．(a) 行動ログと許容ログの入力例．(b) 入力データから中間結果を作成する処理．(c) 遷移確率の推定に使用する行動遷移データと受容可能な行動のデータ．

なお，行動 $k$ は「介入なし」に対応し $k = |\mathcal{A}|$ と定義する．

受容可能な行動のデータ (ユーザが行動を変えることが可能だったかもしれないと考えるデータ) をモデル化するために，エージェントの介入に対するユーザの決定を次の 3 つのケースに分類した; (i) 許容ログのデータに現れる行動の介入を受け入れる．(ii) 許容ログのデータに現れていたとしても，その行動の介入は受け入れない．(iii) 許容ログのデータに現れていない行動であり，介入は受け入れない．ここでは，介入がユーザに受容される確率 (i) を $\beta$ で表す．[4] これを用いて，受容可能なデータは介入を受け入れた回数 $\beta M_{tj}$ と，受け入れなかった回数 $(1-\beta)M_{tj}$ であると見なすことが出来る．また，ユーザが介入を $L - M_{tj}$ 回受け入れないと見なすが，これは行動 $j$ における日数であって現在の時間 $t$ におけるものではない．よって，受容可能な行動データの尤度関数は次の通り記述できる．

$$P(\mathcal{D}_{apt}|\theta) = \prod_{t=1}^{T} \prod_{i,j \in \mathcal{S}} \left(p^{\theta}_{tijj}\right)^{\beta M_{tj}} \left(1-p^{\theta}_{tijj}\right)^{\{(1-\beta)M_{tj}+(L-M_{tj})\}}.$$

上記 2 つの尤度関数の負の対数をとり，以下の目的関数を導くことができる．

$$\begin{aligned}\mathcal{L}(\theta) &= -\sum_{t=1}^{T} \sum_{i,j \in \mathcal{S}} N_{tij} \log p^{\theta}_{tij0} - \gamma \sum_{t=1}^{T} \sum_{i,j \in \mathcal{S}} \Big\{ \beta M_{tj} \log p^{\theta}_{tijj} \\ &\quad + (N - \beta M_{tj}) \log(1 - p^{\theta}_{tijj}) \Big\} + \Omega(\theta), \qquad (1)\end{aligned}$$

$\Omega(\theta)$ は過学習を避けるための正則化項であり，$L_2$ ノルム正則化 $\Omega(\theta) = \lambda\|\theta\|^2$ を実験では用いる．$\gamma, \lambda$ は目的関数の各項の寄与度を制御するハイパーパラメータである．この目的関数の設計によって，許容ログに多数現れる行動 $j$ については $u_{tj}$ の値を大きくすることで目的関数を小さくすることとなる．また，同様に許容ログに全く現れない行動 $j'$ については，$u_{tj'}$ を小さくすることで目的関数が小さくなる．よって確かに実行しやすい介入は受け入れやすく，介入ログに現れない従うことが現実でない介入は受け入れない推定結果が得られる．パラメータ $\theta$ は正則化項を用いて目的関数を最適化することで推定できる．

$$\hat{\theta} = \arg\min_{\theta} \mathcal{L}(\theta), \quad \mathcal{L}(\theta) := -\log P(\mathcal{D}|\theta) + \Omega(\theta), \quad (2)$$

式 (2) の目的関数の最小化には勾配降下法やニュートン法などの最適化手法を用いることができ，実験では L-BFGS 法 [19] を用いた．

### 4.4 報酬設計

**Case1. Good sleep reward**: 報酬関数 $\mathcal{R}$ はユーザの理想とする目標 (ユーザが設定した時間と行動) によって定義される．例えば，ユーザが十分な睡眠を確保するために目標とする状態 $s_g$ を「睡眠」，$t_g$ を「22:00」と設定した場合について考える．報酬は基本的には，目標を達成した場合に大きな正の値 $r_g$ $(r_g > 0)$，介入無しの場合に 0，それ以外の場合は負の値 $r_{itv}$ $(r_{itv} < 0)$ を与える [5] $(|r_g| >> |r_{itv}|)$．つまり，次の状態 $s_{t+1}$ が $t_g$ (22:00) に目標とする状態 $s_g$ (睡眠) であれば，報酬 $r_g$ を得る (それ以外は $r_{itv}$ を得る)．このように設計することで，エージェントは目標達成に繋がる時だけ介入を行い，それ以外は介入しないという振る舞いをする．そのため，不必要に何度もユーザに介入することを避けることが出来る．よって，報酬は下記のように設計する．

$$\mathcal{R}_t(s_t, a_t, s_{t+1}) = r_g \mathbb{1}(s_{t+1} = s_g, t+1 = t_g) + r_{itv} \mathbb{1}(a_t \neq \text{"}none\text{"})$$

この Good sleep reward の設計は他の事例，例えばユーザが新しい習慣を身に付けたい場合にも適用可能である．新しい習慣を始めるには，1 日の中で一定の時間を確保する必要がある．仕事の後に運動したい場合，$s_g$ を「退社」として $t_g$ を普段の退社時刻より早い時間を指定すると，退社時刻を早めるような介入戦略が学習され時間的な余裕を作ることが出来る．

**Case2. Good exercise reward**: ユーザが特定の行動を行う時間を増やしたい場合，例えば運動時間を増やしたい場合について考える．このとき，ユーザは $s_g$ を「exercise」と指定すれば良い．下記のように，次の状態 $s_{t+1}$ が $s_g$ であれば報酬 $r_g$ を与えるようにし，介入無しは 0，その他は $r_{itv}$ とする．

$$\mathcal{R}_t(s_t, a_t, s_{t+1}) = r_g I(s_{t+1} = s_g) + r_{itv} I(a_t \neq \text{"}none\text{"})$$

4 : もしユーザがエージェントの介入を受け入れる可能性が高い場合，$\beta$ の値は 1.0 に近付く．もしユーザが介入に従う可能性が低い場合であっても，ユーザがログに介入を受け入れ可能だと記録していたのであれば，$\beta$ は少なくとも $\epsilon > 0$ よりは大きくなる．$\beta$ はラベルのノイズの確率と見なすことも出来るが，これは [18] と類似するものである．なお，今回の実験では $\beta = 0.5$ とした．

5 : 5 章の実験では目標を達成した場合に大きな正の値 $r_g = 100$，何も介入を行わない場合は 0，それ以外の場合は $r_{itv} = -3$ と設計した．

**Algorithm 1** 後ろ向き帰納法アルゴリズム

**Input:** $\mathcal{P}$: transition probability, $\mathcal{R}$: reward function, $\alpha$: hyper-parameter

**Output:** $\{Q_t^*\}_t, \{V_t^*\}_t$: value function, $\{\pi_t^*\}_t$: policy

1: Set $t \leftarrow T$ and $V_T(s) = 0$ for all $s \in \mathcal{S}$.
2: Set $t \leftarrow t-1$
3: Compute $Q_t(s,a)$ following Eq. (3) for all $s \in \mathcal{S}$ and $a \in \mathcal{A}$.
4: Compute $V_t(s)$ for all $s \in \mathcal{S}$ following Eq. (4).
5: Compute $\pi_t(a|s)$ for all $s \in \mathcal{S}$ and $a \in \mathcal{A}$ following Eq. (5).
6: If $t = 0$, stop. Otherwise, return to step 3.

### 4.5 エントロピー正則化強化学習

推定された遷移確率と報酬関数が与えられた時，提案手法では最適な方策を求める．ここではエントロピー正則化強化学習 (RLER) を用いるものとし，最適化の定義は RLER に従う．RLER の目的は，報酬和と方策エントロピーを最大化する最適な方策 $\boldsymbol{\pi}^*$ を求めることである．

$$\boldsymbol{\pi}^* = \arg\max_{\pi} \mathbb{E}^{\pi}_{h_T}\Big[\sum\nolimits_{t=0}^{T-1}\big\{\mathcal{R}(s_t,a_t,s_{t+1}) + \alpha\mathcal{H}(\pi_t(\cdot|s_t))\big\}\Big],$$

$\mathbb{E}^{\pi}_{h_T}$ は $\pi$ に従うエピソード $h_T$ の期待値，$\mathcal{H}(\pi(\cdot|s_k))$ は分布 $\{\pi(k|s_t)\}_k$ のエントロピー，$\alpha$ はエントロピー項の寄与度を制御するハイパーパラメータ [6]である．エントロピー項は一様分布に近付くほど大きくなり，大きなエントロピー項を持つ場合に方策は決定論的ではなく確率的なものになる．報酬和の値が近い複数の介入行動が存在する時に確率的に行動が選択され，ユーザが介入に飽きてしまうことを避けることができる．そのため，最適な方策は大きな報酬を得ることができ，大抵の場合決定論的ではないことが期待できる．なお，この RLER はハイパーパラメータ $\alpha = 0$ の時に一般的な強化学習の枠組みと同一である．我々は RLER の価値関数を下記のように定義した．

$$Q_t^{\pi}(s,a) = \mathbb{E}^{\pi}_{h_T}\Big[\sum_{t'=t}^{T-1}\{\mathcal{R}_t(s_{t'},a_{t'},s_{t'+1}) + \alpha\mathcal{H}(\pi_{t'}(\cdot|s_{t'}))\}\Big|s_0{=}s, a_0{=}a\Big]$$

また，RLER の最適なベルマン方程式は，

$$Q_t^{\pi^*}(s,a) = \mathbb{E}_{s'\sim\mathcal{P}_t(s'|s,a)}[\mathcal{R}_t(s,a,s') + V_{t+1}^{\pi^*}(s')] \quad (3)$$

$$\text{where } V_t^{\pi^*}(s) = \alpha\log\sum\nolimits_{a'}\exp\Big(\alpha^{-1}Q_t^{\pi^*}(s,a')\Big) \quad (4)$$

この最適な価値関数と方策は Alg. 1 に示す後ろ向き帰納法アルゴリズムで求めることが出来る．RLER の最適な方策は次の確率的な方策で与えられる:

$$\pi_t^*(a|s) = \exp\Big(\alpha^{-1}\big\{Q_t^{\pi^*}(s,a) - V_t^{\pi^*}(s)\big\}\Big), \quad (5)$$

## 5 実　　験

本章では, § 4.3 の遷移確率推定結果の妥当性を検証したのち, 提案する強化学習による介入手法の有効性を確認する．

6 : 実験では $\alpha = 0.01$ を用いた．

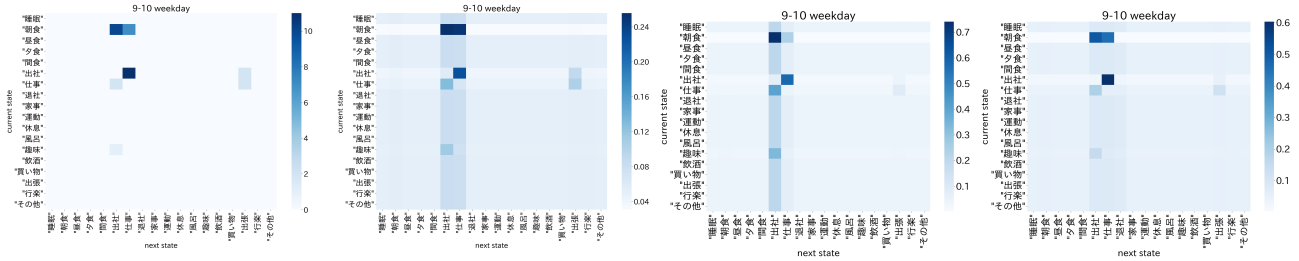
### 5.1 許容ログを用いた遷移確率推定の検証

まず，介入を行う場合と行わない場合の遷移確率を可視化することで，その妥当性を定性的に検証する．介入が無い状態での行動遷移について検証を行うため，ID:39 の行動ログを用いて作成したデータ $D_{tr}$ を可視化したものを図 7a を示す．これは行動ログを用いて行動遷移回数をカウントしたもので，9:00〜10:00 の時間帯にどの行動からどの行動に何回遷移したかを示す．図 7b は，提案手法の推定結果により得た，何も介入を行わない時の遷移確率を可視化したものである．色の濃淡に着目して図 7a と図 7b を比較すると傾向が類似しており，提案手法が MDP のパラメータを正しく推定できていることを示している．さらに，介入を行った場合の推定結果についても検証する．9:00〜10:00 において実行しやすいと考えられる (i) 出社を促す介入 (図 7c) と，従うことが難しいと考えられる (ii) 退社を促す介入 (図 7d) を行なった場合の遷移確率の推定結果を示す. 図 7c では，介入を受けて出社に遷移する列の確率が高くなり，図 7a や図 7b と比較して色が濃くなっている．一方で図 7d の場合は介入を受けても傾向がほぼ変わらない．このように，実行しやすい介入であれば遷移確率が上がるが，そうでない場合は影響しないように推定出来ていることを確認した．

次に，介入無しの遷移確率について定量評価を行う．実際にユーザが介入を受けた時の行動遷移データは入手することが難しいため，代わりに行動ログと許容ログを用いて評価を行う．評価に用いるデータは，参加者が記録したログのうち最後の 5 日間のログからテストデータを作成し，他の日のログから訓練データを作成した．訓練データの行動ログと許容ログはパラメータ推定に利用し，テストデータ中の行動ログ評価に利用した．提案手法の性能は負の対数尤度を評価指標とする．これは値が小さいほど真の確率分布に近い推定結果を表し，確率モデルの評価に広く利用されている．$G = (1/\mathcal{T}_{test})\sum_t\sum_{i,j\in\mathcal{X}} -N_{tij}^{test}\log p_{tij}^{\theta}$. テストデータ中の行動ログにおける $\mathcal{T}_{test}$ は遷移数の合計，$N_{tij}^{test}$ は $t$ 番目の時間に状態 $i$ から状態 $j$ へ遷移した数を示す．我々は提案手法と，許容ログを用いないベースライン手法 ($\gamma = 0.0$) とを比較する．図 8 は負の対数尤度を示しており，値が小さいほど良い．提案手法 ($\beta > 0.0$) はベースライン手法 ($\beta = 0.0$) よりも性能が高く，$\beta = 0.1$ の時にはどちらの参加者においても最も良い性能であった．この結果から，許容ログを用いることで精度が低下することは無いことを確認した．

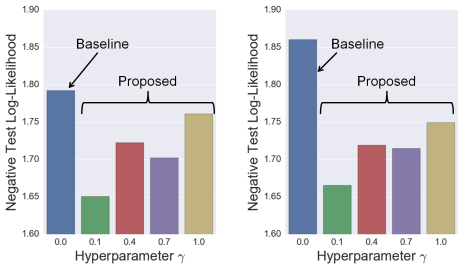
### 5.2 介入効果の検証

ここでは，前節で検証した遷移確率の推定結果をユーザのシミュレータとして用いて，§ 4.5 で述べた強化学習アルゴリズムによる介入効果を検証する．本実験ではベースラインとしてランダムに介入を行う方策 (random) と，目標の時間にのみ介入する方策 (onetime) を用いる．ランダムな介入方策では，各時間ステップ毎に 16 種類の介入の選択肢の中からランダムに 1 つを選択する．目標の時間のみ介入する方策は，例えば目標が「23 時に寝る」と定められた場合，23 時にだけ就寝を促す介入を行い他の時間帯は全て介入なしを選択するもので，寝る時間のリマインダを設定した図 1 (b) に該当する方策である．



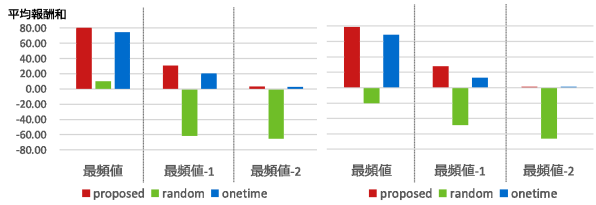

(a) 元データ　(b) 推定結果　(c) 「出社」と介入した時の推定結果　(d) 「退社」と介入した時の推定結果

図 7: 9:00〜10:00 における行動遷移の元データと遷移確率推定結果を，現在の状態 × 次の状態で可視化したもの (ID:39)．色の濃さは各時間帯における遷移の頻度を表し，元データと推定結果の色の濃淡が類似しているほど良い．




(a) ID:39　(b) ID: 14

図 8: (a) ID:39 と (b) ID:14 における負の対数尤度．




(a) ID: 6　(b) ID:12

図 9: 参加者の平均報酬和の比較．値が大きいほど良い．

今回の報酬は § 4.4 で述べた Good sleep reward を用いて，目標行動は睡眠と定めた．目標の時刻はユーザ毎に設定しており，収集した行動ログから睡眠開始時刻の最頻値を算出し，(i) 最頻値の時刻，(ii) 最頻値から 1 時間早い時刻，(iii) 最頻値から 2 時間早い時刻という 3 種類の目標を設定した．例えば，毎日 0 時に就寝しているユーザであれば (i) 0:00，(ii) 23:00，(iii) 22:00 が目標時刻となる．なお，介入効果の評価指標としては，平均報酬 (報酬和) を用いる．これはユーザを目標達成に導くよう適切な介入を実現出来ている度合いを示すものであり，値が大きいほど良い．

定量評価として，平均報酬和を用いて評価を行う．まず，参加者 2 名の報酬和を図 9 に示す．いずれの目標時刻においても提案手法は最も高い報酬和を得ており，ランダムや単純なアラーム設定よりも我々の手法で目標時刻の前段階から介入する方が効果的であることを示している．次に，全ての参加者 34 名の平均報酬和を累積相対度数分布で示す (図 10)．例えば図 10a を見ると，参加者の 60%以上はランダム方策 (random) では報酬和が 0 より小さい値をとることが分かる．参加者全体で見ても提案手法はどの目標時刻においても概ねベースラインよりも高い報酬和を得ており，目標達成に導く介入を実現出来ていることを確認した．目標時刻毎に見ると，就寝時刻を 1 時間早める場合 (図 10b)，報酬和の低い全体の 50%程度の参加者については proposed と onetime とで差が無かったが，報酬和 10〜70 前後の参加者 (全体の 40%程度) に関しては差がついており，提案手法の介入効果が高いことが分かった．

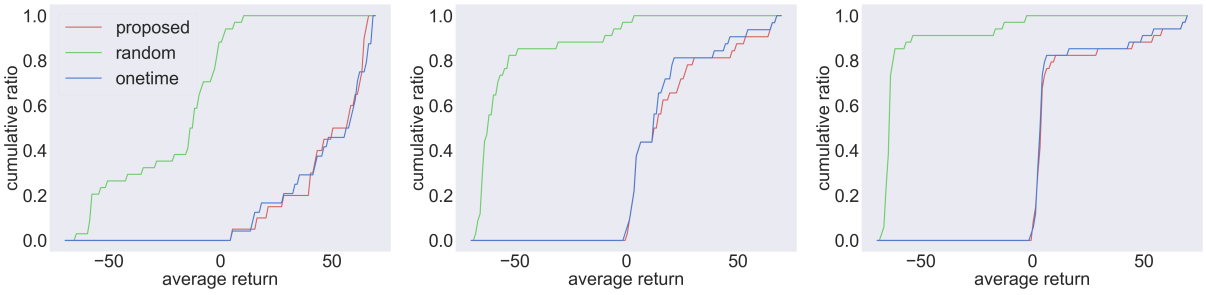
定性評価として，目標を 22 時就寝と設定した時の 19〜22 時の各時間ステップにおける最適方策 $\pi(a_t|s_t)$ を可視化した (図 11，図 12)．図 11 の参加者については，19 時は介入なし，20 時夕食, 21 時風呂，22 時就寝という順番で参加者に介入することが目標達成に最適の方策であると分かる．図 12 では，20 時にもし夕食や家事を行っていた場合には入浴を促すが，それ以外の行動であれば介入しない．その後，21 時に家事，22 時に就寝するように促している．このように，参加者一人一人に合わせて方策が最適化されており，提案手法はユーザが目標を達成出来るような行動を適切に見定めていることが分かる．

## 6 おわりに

本研究では，モデルベース強化学習を用いてユーザの目標達成を支援するための介入戦略を学習する手法を提案した．ユーザの生活パターンと介入が受け入れられる可能性を考慮して環境のモデル (遷移確率) を推定し，後ろ向き帰納法アルゴリズムで目標から逆算して最適な方策を得る方法を構築した．実験では 34 名の参加者から収集した実データを用いて遷移確率の推定を行い妥当性を確認した後，推定結果をユーザのシミュレータとして用いて強化学習による介入手法の有効性を検証した．平均報酬和を指標として介入効果を定量的に検証したところ，提案手法はランダムな介入方策や単純なアラーム設定よりも目標達成のために効果的な介入を実現できていることを確認した．今後の課題としては，$\beta$ の値を変更した場合や他の報酬関数における評価，本手法をアプリケーションに実装した実環境での検証，類似ユーザのデータを活用することによりデータが無いユーザにも適用出来るよう手法を拡張することを検討している．

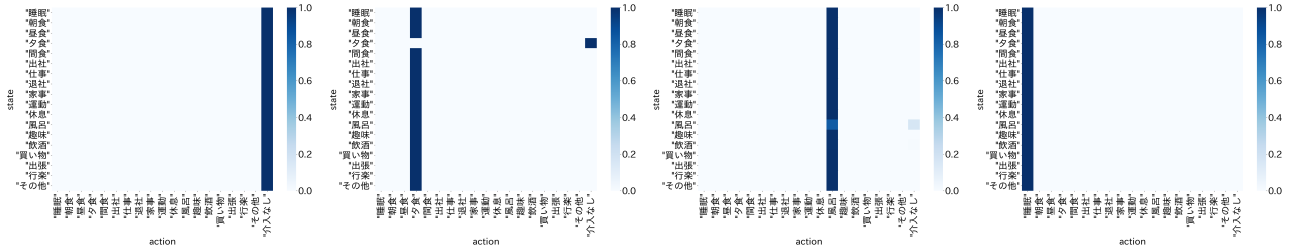
## 文　献

[1] Nursing for the people with lifestyle-related diseases in Japan, https://www.nurse.or.jp/jna/english/pdf/lifestyle–01.pdf (参照日：2020 年 2 月 13 日).
[2] 厚生労働省：生活習慣病予防，https://www.mhlw.go.jp/stf/seisakunitsuite/bunya/kenkou_iryou/kenkou/seikatsu/ seikatusyuukan.html (参照日：2020 年 1 月 9 日).
[3] Sutton, Richard S., and Andrew G. Barto. Reinforcement learning: An introduction. MIT press, 2018.
[4] Silver, David, et al. Mastering the game of go without human knowledge. Nature 550.7676 (2017): 354-359.
[5] Daskalova, N., Lee, B., Huang, J., Ni, C., and Lundin, J.



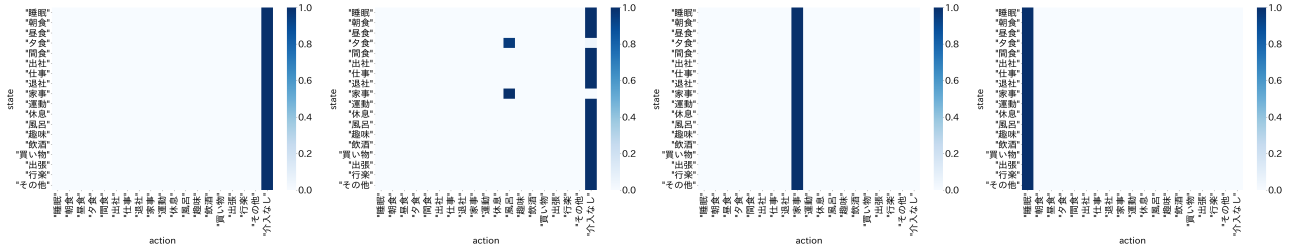

(a) 最頻値　(b) 最頻値-1 時間　(c) 最頻値-2 時間

図 10: 全参加者の平均報酬和分布 (提案手法，ベースライン手法 (ランダム)，1 度だけ介入)．縦軸は参加者の累積比率，横軸は報酬和を示す．報酬関数の目標時刻はユーザ毎の就寝時刻の (i) 最頻値，(ii) 最頻値-1 時間，(iii) 最頻値-2 時間とした結果を示す．




(a) $t = 19$　(b) $t = 20$　(c) $t = 21$　(d) $t = 22$

図 11: 19〜22 時の各時間ステップにおける方策 $\pi(a_t|s_t)$ (ID:39)




(a) $t = 19$　(b) $t = 20$　(c) $t = 21$　(d) $t = 22$

図 12: 19〜22 時の各時間ステップにおける方策 $\pi(a_t|s_t)$ (ID:14)

Investigating the effectiveness of cohort-based sleep recommendations. Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies, 2018.

[6] Takahashi, M., Kohjima, M., Kurashima, T., Matsubayashi, T., and Toda, H. Identifying self-changeable actions toward regulating rhythm of daily life. In Proceedings of the 2019 ACM International Joint Conference on Pervasive and Ubiquitous Computing and Proceedings of the 2019 ACM International Symposium on Wearable Computers (pp. 218-221).

[7] Mnih, V., Kavukcuoglu, K., Silver, D., Rusu, A. A., Veness, J., Bellemare, M. G., and Petersen, S. Human-level control through deep reinforcement learning. Nature, 2015.

[8] Silver, D., Huang, A., Maddison, C. J., Guez, A., Sifre, L., Van Den Driessche, G., and Dieleman, S. Mastering the game of Go with deep neural networks and tree search. nature, 529(7587), 484. 2016.

[9] Shani, G., Heckerman, D., and Brafman, R. I. An MDP-based recommender system. Journal of Machine Learning Research, 6(Sep), 1265-1295. 2005.

[10] Lillicrap, T. P., Hunt, J. J., Pritzel, A., Heess, N., Erez, T., Tassa, Y., and Wierstra, D. Continuous control with deep reinforcement learning. arXiv:1509.02971. 2015.

[11] Li, J., Monroe, W., Ritter, A., Galley, M., Gao, J., and Jurafsky, D. Deep reinforcement learning for dialogue generation. arXiv preprint arXiv:1606.01541. 2016.

[12] Raghu, A., Komorowski, M., Celi, L. A., Szolovits, P., and Ghassemi, M. Continuous state-space models for optimal sepsis treatment-a deep reinforcement learning approach. arXiv preprint arXiv:1705.08422. 2017.

[13] Shortreed, S. M., Laber, E., Lizotte, D. J., Stroup, T. S., Pineau, J., and Murphy, S. A. Informing sequential clinical decision-making through reinforcement learning: an empirical study. Machine learning, 84(1-2), 109-136. 2011.

[14] Yom-Tov, Elad and Feraru, Guy and Kozdoba, Mark and Mannor, Shie and Tennenholtz, Moshe and Hochberg, Irit. Encouraging Physical Activity in Patients With Diabetes: Intervention Using a Reinforcement Learning System, J Med Internet Res, 2017.

[15] el Hassouni, A., Hoogendoorn, M., van Otterlo, M., and Barbaro, E. Personalization of health interventions using cluster-based reinforcement learning. In International Conference on Principles and Practice of Multi-Agent Systems (pp. 467-475). Springer, Cham. 2018.

[16] Tabatabaei, Seyed and Hoogendoorn, Mark and Halteren, Aart. Narrowing Reinforcement Learning: Overcoming the Cold Start Problem for Personalized Health Interventions. PRIMA 2018 Principles and Practice of Multi-Agent Systems, Proceedings. 2018.

[17] Puterman, Martin L. Markov Decision Processes. Discrete Stochastic Dynamic Programming. John Wiley and Sons, 2014.

[18] Natarajan, N., Dhillon, I. S., Ravikumar, P. K., and Tewari, A. Learning with noisy labels. In Advances in neural information processing systems (pp. 1196-1204). 2013.

[19] Liu, D. C., and Nocedal, J. On the limited memory BFGS method for large scale optimization. Mathematical programming, 45(1-3), 503-528. 1989.