修士論文のアウトライン

## Introduction

より良い意思決定をするには、過去の経験を通じて学習しなくてはならない．しかし実世界での意思決定は、不確実で高次元の環境において行われる．そのような環境では、多様な次元での多様な特徴の組み合わせを試すなどの方法で、情報収集して意思決定の質を向上させることができる．しかし、幾らかの次元にのみ限定されるのは実験環境に限られており、実世界では関連する次元の数が多く、情報収集する量が莫大で現実的ではない．

（不確実性について）

また、実世界では、より良い意思決定をするために参照するべき次元が、時間や状況の変化に応じて変化する．

（RLについての研究、探索に焦点を当てて）

探索活用のトレードオフなど

（多次元環境での意思決定についての研究．探索行動として捉えるなら）

Wisconsin card sorting taskとかだと、仮説検証として扱っている．

(Wilson & Niv, 2012; Niv et al., 2015; Leong et al., 2017)

特徴量抽出のような”representation learning”の考え方や

Attentionについての言及

これらの研究はWCSTの方法に当たり確率を25%, 75%にして学習の様子を調べた

（）

（不確実性（変化する環境）での意思決定についての研究）

(Behrens et al., 2007; Daw et al., 2006)

学習において、当たり確率が大きく変動し、より良い選択肢が変化するような研究．

単純に探索行動を調べるためのもの、

Volatitilyが探索行動を高めることを調べたもの．

（本研究の目的）

次元間の探索行動を調べる．

実世界での意思決定は、ものを比較して購入するような場面でも、価格や性能、デザイン、ブランド他多様なことを考慮しつつ意思決定を下す必要がある．次元数が増えるといわゆる「次元の呪い」と呼ばれるように、情報量が膨大に大きくなるため、効率的に処理を行う必要がある．実世界の多次元の特徴をそのまま扱うには処理する情報量が果てしなく多くなってしまうため、判断に重要な次元を選び、その次元を基に判断を下す必要がある．

そのまた、価格や性能などの特定の次元も、刻一刻と変化する環境では、

次元間の探索

## Method

### Subjects

15名の実験参加者（うち２名が女性、13名が男性．年齢は17-28歳の範囲で、平均23.8歳）．Informed consent, reward///

うち２名については、ルールの教示に問題が確認されたので分析から除外した．

### Task

多次元空間での学習行動を調べるために、多次元学習課題を行った．

課題は、Wisconsin Card Sorting課題（Milner, 1958）をベースとした、

色（薄い緑、濃い緑）、形（三角、星）、数（一つ、二つ）の３つの次元の組み合わせで構成される選択を3つ呈示し、それを選び報酬（点数）を得ることを通して学習を進める課題を行なった．色、形、数の次元がそれぞれあたり確率に影響を与え、その影響は課題を通して変化し、実験参加者は選択と報酬を通してその変化に対応して学習を進める、点数を最大化することが求められた．

選択肢を3つ提示され、一つを選ぶ．選ぶと「当たり」か「ハズレ」が表示される．報酬は当たり10点、ハズレ0点であり、実験を通して固定した．全部で300トライアル行い、その選択系列を分析した．

色、形、数の各次元につき要素は2種類で、選択肢は8()通りの選択肢が存在する（（（（（図A）））））．各次元において要素の当たりやすさの優劣は決まっており、実験参加者にも教示した．優劣は、色は濃い方が当たりやすく、形は星が三角より当たりやすく、数は2個が1個よりも当たりやすい（（（（（図B）））））．この優劣は伝えたが、色、形、数のどの特徴が当たり確率にどれだけ寄与するかはわからず、３つの選択肢の当たり確率は独立で、全てがハズレの場合も、全てが当たりの場合もある．優劣が直感的に理解しやすくするために、濃淡、大小、形状を上記のように選択した．

ここで各選択肢を三つの次元の組み合わせとしての形式で記述した．薄い場合は0、濃い場合に1．三角で０、星で１．一つの時に０、二つの時に１である．

各選択肢の当たり確率の算出は、とする時、以下の式によって行なった．

例えば選択肢(1, 1, 1) つまり濃い星二つの時の確率は．また、選択肢(0, 1, 0) つまり薄い星一つの時の確率はとした．

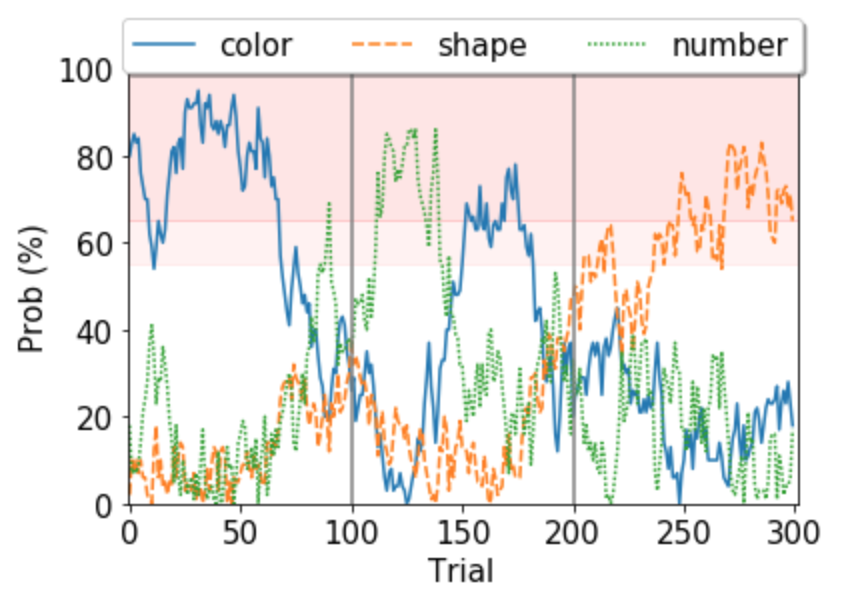
は、確率の変化を含むBandit Taskでの確率推移(Daw et al., 2006)を参考にしてtrialごとに変化させた（（（（（図C））））））．この確率推移は、和が1になるようにしたまま、初期値はランダムで取得し、変化量は３次元の変化量がそれぞれmean=0.2, sd=0.055となるようにgaussian randomで生成し、値は0.01の精度（%では１の精度）で計算した．

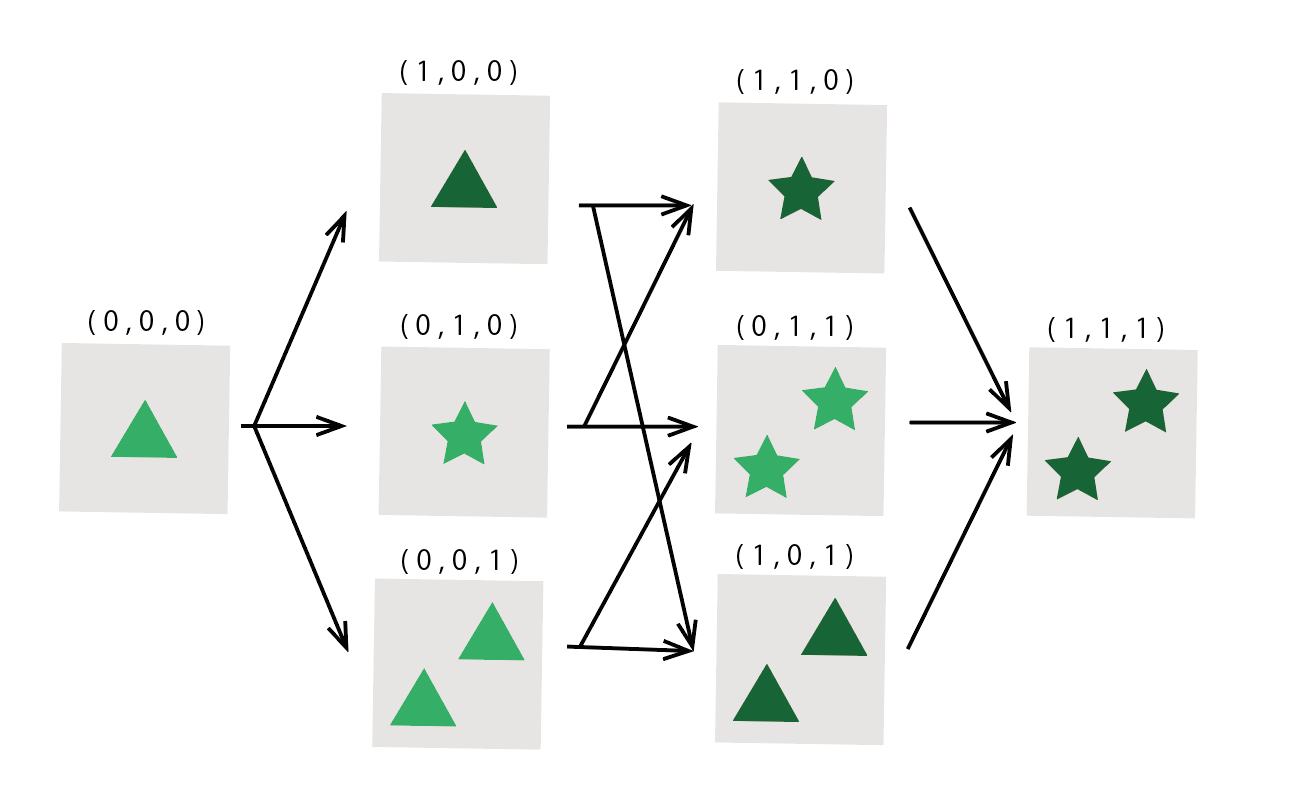
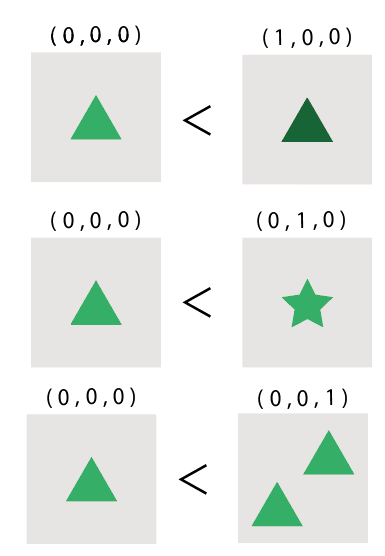
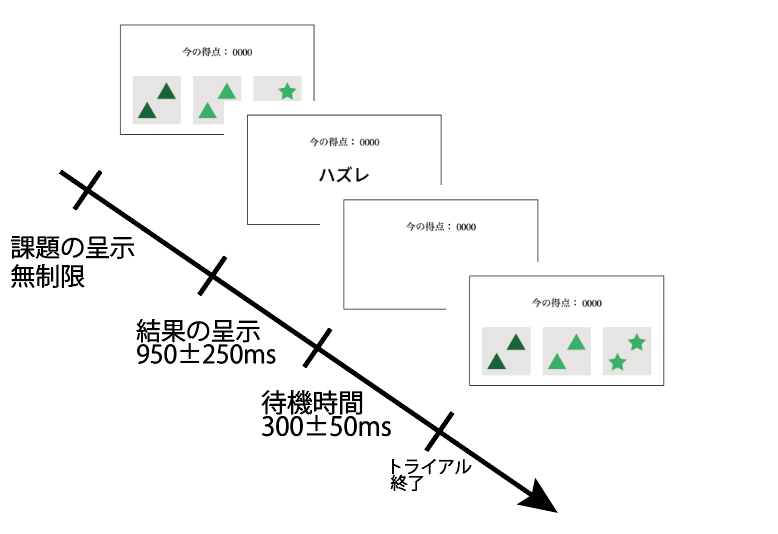
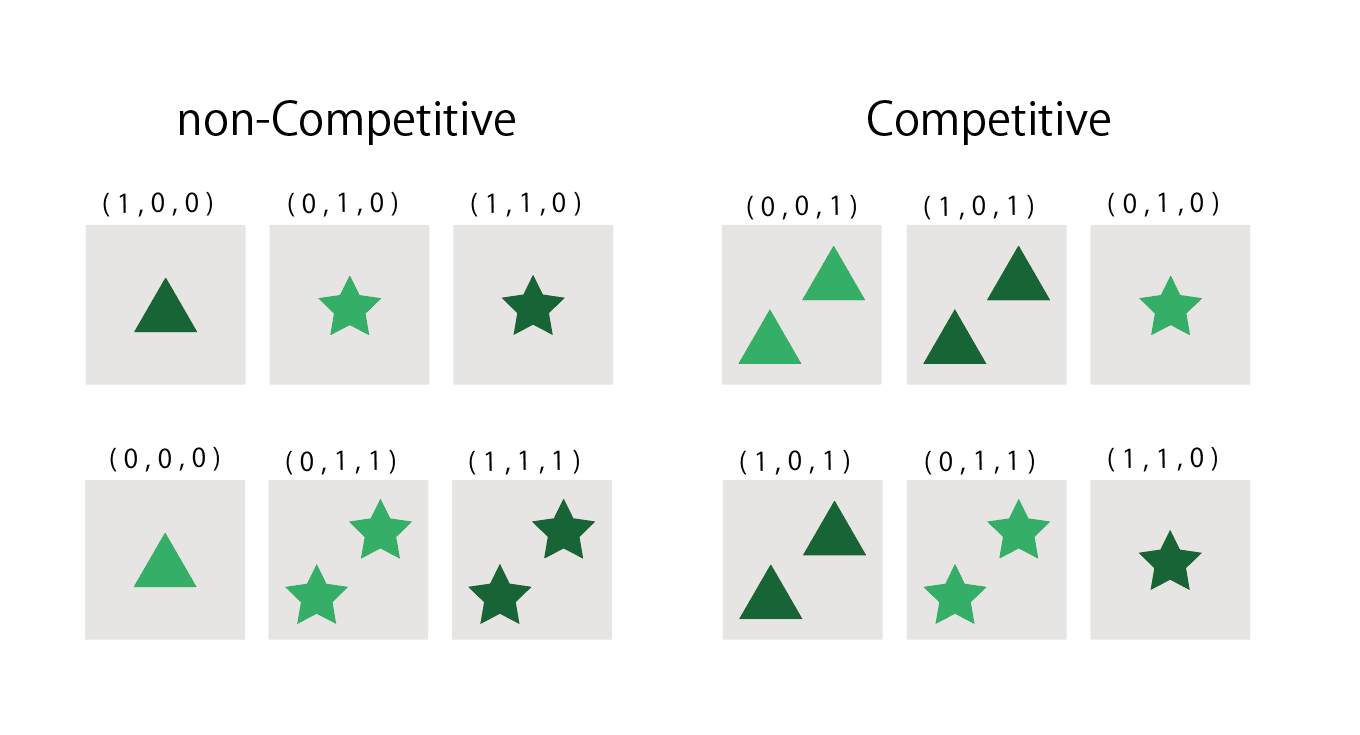
（ここに、想定される学習の流れを説明する）

この当たりやすさの優劣（以後ルールと呼ぶ）に基づくと、8つの選択肢には順序関係が生じる．順序関係は以下の通りで、濃い星2つのような他のどの選択肢よりも当たりやすいものや、逆にどれよりも当たりにくい選択肢が存在する．

このルールに基づくと、二つの選択肢に明らかな優劣が存在するかどうかが分かる．図のように、<<>>と＜＜＞＞の場合は、順序が存在し、＜＜＞＞が明らかに当たりやすい．一方、＜＜＞＞と＜＜＞＞の場合は順序が存在せず、どちらが当たりやすいかをそれまでの選択と報酬に応じて意思決定する必要がある．

同様に3つの選択肢が提示された時に、ルールに基づいてもどれが一番当たりやすいか判断できない場合はcompetitiveなトライアル、明らかにどれが当たりやすいか判断できる場合はnon-competitiveなトライアルと呼んで分類した．

3つの選択肢を生成する方法は表示順を無視すると56通りあるが、non-competitiveな組み合わせは半分以上の30通り存在する．competitiveすべての選択肢が等確率で出現させた場合、ルールに準拠するだけで点数が上がってしまい、選択行動とその学習を調べられない．本研究では、次元に対する学習を調べる目的より、異なる次元で比較するtrialで分析をするため、一定数のcompetitiveなトライアルが存在する必要がある．したがって実験においては、non-competitiveなトライアルは10%に設定し、ルールを理解できているかの確認と、実験参加者がルールを再認識するトライアルとして利用した．実際の実験では、non-competitiveなトライアルは7-14%の範囲（平均9.8%）であった．



### Procedure

実験は、練習と本番と行った．練習はルールを説明しつつ、選択に慣れるために、10回当たりが出るまで行った．本番は300トライアルで所要時間は15分程度であった．

また表示位置にも関係ないことを伝えた．

### Model fit

モデルのフィッティングには基本的なモデルと、情報量および温度の探索行動への影響を調べるための以下のモデルを利用した．

#### Naïve RL

8()つある刺激に対して独立にTD学習（Rescorla & Wagner, 1972; Sutton & Barto, 1998）を行うようなモデルである．すでに与えられた順序を考慮せず、またある星形の刺激を選んでも、他の星形の刺激のvalueには影響を及ばさない点でNaïveなモデルである．

Sを刺激とする時、Valueの更新式は、

ここでαは学習率のパラメータである．報酬Rは当たりで１、ハズレで0として、valueも(0,1)の範囲である．

Valueに基づく選択確率はsoftmax関数を用い、表示されている３つの選択肢が（）とする時、その式は以下の通りである．

ここでβは意思決定のノイズである逆温度をパラメータで、βが大きくなるほどvalueが高い刺激が選択されるようになり、βが小さくなるとvalueが小さい刺激でも選択される確率が高くなる．

パラメータはここまでに示したα、βの他に、valueの初期値もパラメータとして設定しモデルフィットをした．

このモデルは、ベースラインとして利用するとともに、ルールの教示が正しく行われていたかを確かめるために利用した．

以上より、フリーパラメータは３つ存在し、．

#### Model-based Naïve RL

Naïve RLに、刺激の順序関係を加味したモデル．3つ呈示された刺激の内、順序が明らかに低いもののvalueをそのトライアル中だけ０とするようにした．

他の更新式および、選択確率を求めるsoftmax関数も同一のものを使用した．

このモデルは、順序関係というルールを理解しているものの、次元を跨いで効率的な学習ができているかどうかを調べるために利用した．

以上より、フリーパラメータは３つ存在し、．

#### Dimension RL

刺激自体ではなく、選択した刺激の次元に対して学習を行うモデル．色、形、数の３つのvalueを持たせた．として、刺激を<<<図>>>のように(1,0,1)などと表示するようにとして、またで示す時、価値関数の更新式と、ある刺激のvalueの導出は以下のように行なった．

*ただし、*

価値関数の更新は、選択肢した刺激の特徴のうち、優位な特徴を有する次元でのみ更新されるようになっており、例えば濃い三角一つ（1,0,0）の刺激を選んだ場合は、色にのみ適用した．

したがって、フリーパラメータは３つ存在し、．

#### Dimension RL + Decay

選択されなかった次元についてはvalueが減少するようなモデル．人は過去の選択履歴については忘れてしまうため、学習についても忘れられる．この「忘れる」ことを含めたモデルである（Niv et al., 2015）．忘れることはdecay率dをパラメータとしておき、以下のように行なった：

すなわち、刺激に含まれない次元のvalueに(1 – d)を乗して、減衰させる．

したがって、フリーパラメータは４つ存在し、．

#### Dimension RL + Info

情報収集において、各次元の情報量を考慮したモデル．Valueは期待値であり、報酬は1としているため、valueに対して、以下の式で定義した．

ここで、

は、Information bonusの重みであり、どれだけ情報量が重視されるのかをフリーパラメータとして導入した．

したがって、フリーパラメータは４つ存在し、．βとInformation bonus weightは、いずれも探索行動に影響を与えるパラメータであり、同時にフィットすることでどちらの影響が強く出るかを検証するために用いた．

1. ここに、モデル上での与える影響、βとの違いについて説明する．

#### Dimension RL + Decay + Info (without β)

Dimension RLに対して、Information bonusを加味し、Decayを含むようにした上で、逆温度βを固定したモデル．

βとInformation bonus weightはどちらも探索を促進させるパラメータであるため、探索行動をβで設定したDimension RL + Decayとこのモデルを比較することで、行動系列への逆温度βとInformation bonus weightの影響を分析した．

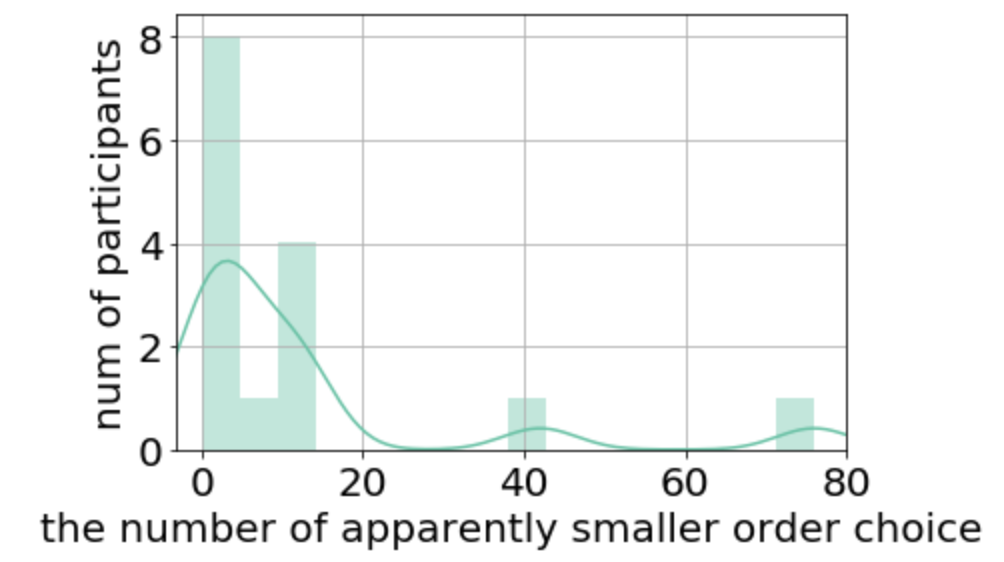
1. Betaがなくなることで、どんなモデル上の影響があるかを、上記の再記して説明する

したがって、フリーパラメータは４つ存在し、．βは固定し、固定したβの値は、前述のDimensionRLを含むモデルでフィットされたβの範囲で、全体へのフィットが最も良くなる値を採用した．

## results

### 行動分析

参加者の基本的な行動分析を行なった．



Caption:：図---: 実験を通して順序の小さい刺激を選択した回数と、その人数．

刺激の順序列を考慮したときに、明らかに当たり確率の小さい、順序の小さい刺激を選択した回数別のヒストグラム．全300 trials中40~,70~回順序の小さい刺激を選択した参加者がいた．

この二人の参加者については、教示がうまく行えておらずルールを理解していなかったとして、この後の分析では除外した．他の参加者は全体の5%程度にとどまっており、正しくルールが教示されていたとして、分析の対象とした．

他の実験に関する基本的な事項としては、平均の点数は1973点、当たり確率は197.3回．スクリーンショット が含まれている画像

自動的に生成された説明ルールに順守しつつも他学習をせずランダムに選択した場合の点数は178点程度（呈示される刺激の組み合わせにより変動する）なので、ランダムよりはいい成績を残していると．点数が低く1800点台の参加者もいるが、選択系列を見る限り学習を行なっていたため分析の対象とした．

呈示された刺激の組み合わせについては、順序を考慮すれば当たり確率が最大のものを選べるトライアルは全体の1割程度、2つの刺激で比較をすれば良い（１つだけ順序から明らかに当たり確率が低い刺激がある）トライアルは全体の8割超、３つ全ての刺激を比較するトライアルは1割弱だった．

学習がなされているかを確認するために、「正しい選択」をしたかどうかを調べた．「正しい選択」は呈示された３つの刺激から最も当たり確率の高い刺激を選択したことを指す．課題の確率変化と学習の進捗に応じて変化すると考えられる．100 trial, 150 trial, 200 trialあたりで確率が変化するため、正しい選択の確率はそのときに低下し、そこから学習が進むにつれて上昇すると想定される．実際の結果もそのように変化していた．

（Caption:図---： 20 trialごとの「正しい選択」をした確率の推移．「正しい選択」は、呈示された３つの刺激の中から最も当たり確率の高い刺激を選択したことを指す．）

テキスト, 地図 が含まれている画像

自動的に生成された説明

#### フィットの結果

スクリーンショット が含まれている画像

自動的に生成された説明

テキスト, 地図 が含まれている画像

自動的に生成された説明テキスト, 地図 が含まれている画像

自動的に生成された説明

（全体のモデルフィットを概観）

・どのモデルのLLが良くなっていたか？

Naïve RLが最も低い尤度で、次がDimension RL、Rule-based Naïve RL、DimensionRL+info．より良い尤度を示したのはdimensionRL+Decayのモデルで、探索に関する項目でBeta（逆温度）を使用した場合でもinformation bonus rateを使用した場合でもあまり変わらない尤度となった．

・Decay有りのモデルで大きくLLがいい．

・Dim RLとrb-NRLだと、若干rb-NRLがいい．どちらも順序の情報が入った状態だが、そこまでうまくdimを処理できているのか？などの何かしらの踏み込みすぎない説明を入れる．

・infoを入れてもLLが向上せず、表２のparamを見ても、0にフィットされており、infoがdimRLの性能を向上させないことが分かる．Infoに関わる要素は他のパラメータや式によって説明されていると言える．

Betaがあれば十分なのか？

・Decayを入れると大きく向上した．

#### ・dimRLのbetaとinfoについて

DimensionRL+Decayでは、上述したように探索行動にBeta（逆温度）を使用した場合でもinformation bonus rateを使用した場合でも変わらないfitの精度を示した．ここで、betaとinfoが同様の行動を示すパラメータであるかを比較するためにPearsonの相関分析を行なったところ強い相関が見られた（図---; *r*=.929, *p*<.001）．また、他のパラメータを確認しても、初期value・decayについても強い相関が見られた． 学習率についてはあまりいい相関が見られなかった．

フィットされた両モデルでの学習率を見ると、βを利用したモデルでは .22, .29, .33などの大きな値が取られているのに対して、info bonusを利用したモデルでは一つ .11を除き.03~.06付近に集まっている．学習率.11もβを利用したもモデルに比べるとかなり小さい値となっており、info bonusが学習率の分散を小さくしたと考えられる．

βを利用したモデルでの学習率とβを比較すると実際軽い相関関係が見られる（（（図）））．この実験課題とこれらのモデル上では、両パラメータは大きくなると、より当たりやすい次元の選択を促進させ、小さくなると当たりにくい次元の選択をする確率を上げるため、この側面において、モデル上の学習率と逆温度βが同様の背景ないし特徴を有していると考えられる．

一方でinfo bonusを利用したモデルでは学習率の分散がとても小さくなっており、info bonusが単に逆温度βと共有している特徴を表現しているだけでなく、学習率と逆温度βの双方の特徴を表現していると考えられる．これらのことより学習率と逆温度と分けていたパラメータが本実験課題においてはinfo bonusによって説明できることが示唆された．

Log01.csvのfitした時のvalueの推移．

Valueのテーブル

LLの数字．．

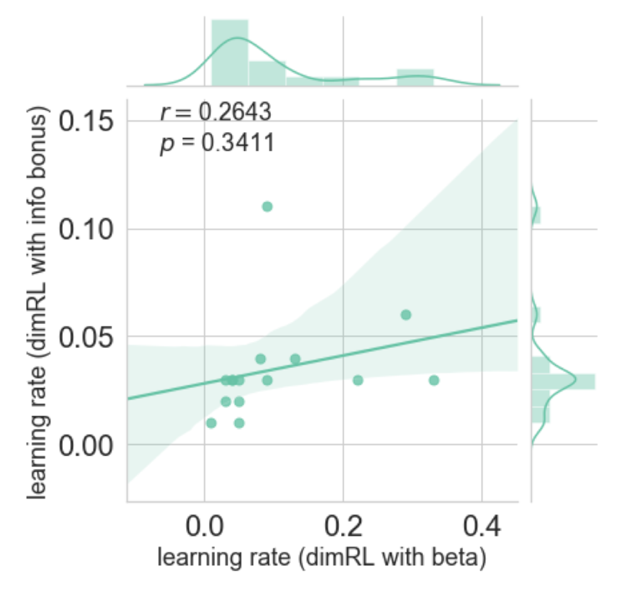
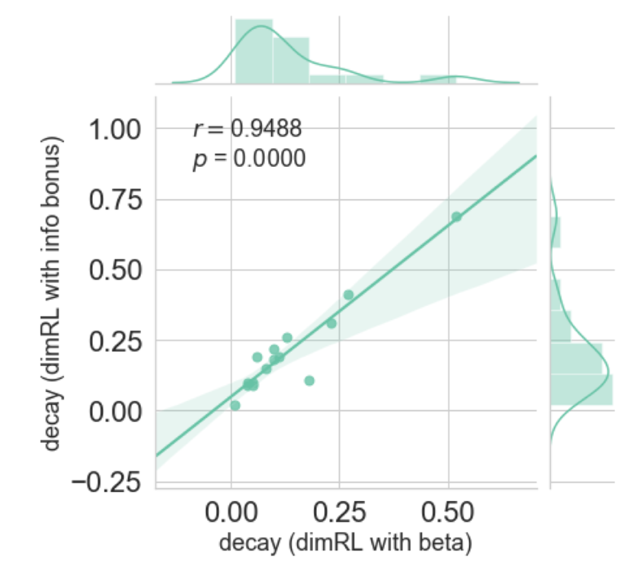
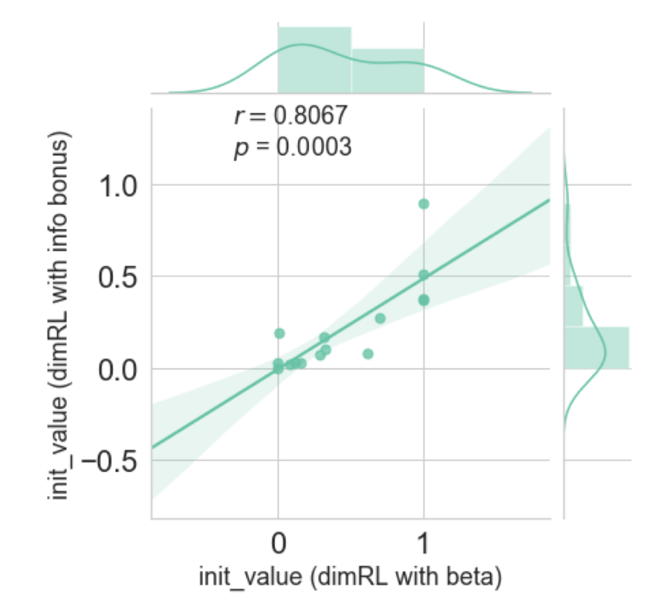
Dimension RL + Decay のBy BetaとBy Infoの比較．

Betaとinfoが相関．

Init\_valueもdecay rateも相関している．

一方learning rateに関しては、

テキスト, 地図 が含まれている画像

自動的に生成された説明

テキスト が含まれている画像

自動的に生成された説明

## Discussion

不確実で多次元の環境においての学習・探索行動を調べるために、本研究では多次元学習課題を行ない、行動の特徴を抽出するためのモデルをフィットした．その結果、人が多次元環境において次元間の学習を行う際に、情報量ボーナスを考慮して意思決定をしていると考えらえる．特に、学習と探索と分けてパラメータを置き、考慮していた部分を情報量ボーナスで説明できる可能性が示唆された．

ここで、

また、Decayのフィットが良くなったことへの指摘―――――．本実験課題での学習においては、常に当たりやすい刺激‘(当たりやすい次元を含む刺激)を選ぶ頻度に対して、当たりにくい刺激を選ぶ頻度が少なくvalueを下げる働きが弱い．そのためDecayを採用したvalueを下げる効果のあるモデルでのフィットが良くなったと考えられる．Valueの上昇と下降について扱いを変えることは、valueの上下によって学習率を変化させるモデルが人で良いフィットをしていること（Farreny et al., 2016; Gershman, 2016）にも共通する部分がある．

## Reference

Behrens, T.E., Woolrich, M.W., Walton, M.E., & Rushworth, M.F. (2007). Learning the value of information in an uncertain world. *Nature Neuroscience*, 10, 1214-1221.

Cohen, J.D., McClure, S., & Yu, A. (2007). Should I stay or should I go? How the human brain manages the trade-off between exploitation and exploration. *Philosophical transactions of the Royal Society of London.* Series B, Biological sciences, 362 1481, 933-42 .

Daw, N.D., O'Doherty, J.P., Dayan, P., Seymour, B., & Dolan, R.J. (2006). Cortical substrates for exploratory decisions in humans. *Nature*, 441, 876-879.

Farreny, A., Rey-Mejías, Á.D., Escartín, G., Usall, J., Tous, N., Haro, J.M., & Ochoa, S. (2016). Study of positive and negative feedback sensitivity in psychosis using the Wisconsin Card Sorting Test. *Comprehensive psychiatry*, 68, 119-28 .

Gershman, S. J. (2016). Empirical priors for reinforcement learning models. *Journal of Mathematical Psychology*, 71, 1–6.

Kaelbling, L. P., & Littman, M. L. (1996). Reinforcement learning: a survey. *Journal of Artificial Intelligence Research*, 4, 237–285.

Leong, Y.C., Radulescu, A., Daniel, R., DeWoskin, V., & Niv, Y. (2017). Dynamic Interaction between Reinforcement Learning and Attention in Multidimensional Environments. Neuron, 93, 451-463.

Niv, Y., Daniel, R., Geana, A., Gershman, S.J., Leong, Y.C., Radulescu, A., & Wilson, R.C. (2015). Reinforcement learning in multidimensional environments relies on attention mechanisms. The Journal of neuroscience : the official journal of the Society for Neuroscience, 35 21, 8145-57 .

Sutton, R.S., & Barto, A.G. (1998). *Reinforcement Learning: An Introduction*. Cambridge: MIT Press.

Wilson, R.C., & Niv, Y. (2011). Inferring Relevance in a Changing World. Front. Hum. Neurosci..