2C-07

バンディット問題における満足化基準について

仲里　慎司†　　下川　哲矢‡

東京理科大学　経営学研究科†　　東京理科大学　経営学部ビジネスエコノミクス学科‡

1，本研究の位置づけ

日々の経済活動において，例えば購買行動や求職活動，取引先の選定など，人々は不確実性を伴う多くの意思決定問題に直面する．一般的にこのような意思決定問題においては，逐次的な情報獲得のプロセスを経て，直面する問題に対する信念の更新を伴うものであるが，一方で，環境に対する学習のみに注力することは機会損失の発生に繋がる．このような情報の獲得行動と活用行動のバランスに注視した問題は， the exploration-exploitation dilemma として広く知られる． またこの問題を扱うフレームワークとしては，多腕バンディット(MAB)問題による定式化が広く普及しており，制御工学やロボティックス，生態学など，広範な科学分野で扱われている．

MAB 問題による情報獲得行動の記述は， 人間の意思決定モデリングにもその応用の幅を広げている．このような潮流の中で，現在，ある一つのモデルが人間行動に対する記述精度の高さから注目されている． そのモデルは SGU モデルと呼ばれ， (1)SoftMax 型関数による確率的な選択，(2)Gaussian Process Regression (GPR)による期待更新， (3)Upper Confidence Bound (UCB)型関数による評価の 3 つの要素で構成される． 同モデルは，特に，それまで数個程度の選択肢しか含まれない MAB 問題を扱っていた認知科学の研究領域において，より広大な選択肢空間での人間の意思決定問題を扱えるようにしたという点で大きく評価される．この拡張は，人々が相対するより現実的な問題への応用を可能とするものであり，既に同フレームワークを用いた購買行動データへの応用研究も行われている．

しかしながら，SGUモデルによる人間行動モデリングへの適用は，議論が不十分な点が存在する．第一に，SGU モデルで扱われる UCB 型関数/戦略は，選択肢のもつ標準偏差(リスク)に正の評価を与える性質をもつという点である．経済学等では，一般に人間はリスクに対して回避的な性向をもつことが知られており， この点で矛盾が生じる．第二に，SGU モデルによるモデリングでは，その意思決定に時間割引が影響を及ぼさない点である．しかしながら，プロスペクト理論に代表されるように，時間選好は人間の意思決定に影響を与える重要な要素であり，最近では，神経科学分野からもこの事実に関する実証的な報告がなされている．

以上のように，SGU モデルは，現時点では実データに対する実証的な精度の良さによって評価されているのみであることに注意しなければならない． 本研究では， これらの問題点について，動学的最適化の観点から MAB 問題における行動モデルを記述し直すことによりその解決を図り，また，提案した行動モデルを用いて，人間の満足化行動を説明することを試みる．

2，問題設定/モデリング

MAB 問題に対する動学的最適化からアプローチを行った研究としては AverBeck (2015)などが挙げられる．しかしながら，彼らの研究は選択肢が 2, 3 個の MAB 問題での設定にとどまっており，また状態の遷移にある種の線形性の仮定を置いている。そこで， 本研究では状態空間表現に Gaussian Process (GP)を利用することにより，これらの問題に対する解決を試みる。

今， 選択肢集合Dの中から， 毎期任意の選択肢$x \in D$を選び，対応するリターン$r = f(x) + \varepsilon$を得る動作を計T回行う MAB 問題を考える． ただし$f: D \to R$は未知の報酬関数であり，εは$N(0, \rho^2)$に従うガウスノイズである．また，意思決定主体の期待形成は GP に従うことを仮定し， このことを$f \sim GP(\mu, K)$と表現する. ただしμは平均，Kは共分散行列であり，共分散行列の各要素はカーネル関数$k(x, x')$から与えられる. この時， 任意の時刻tにおいて，それまでの選択行動と報酬の履歴データをそれぞれ$x_{1:t-1} = [x_1, x_2, \cdots, x_{t-1}]^{\mathrm{T}}$， $r_{1:t-1} = [r_1, r_2, \cdots, r_{t-1}]^{\mathrm{T}}$とすれば， 報酬関数の事後分布は次のガウス分布に従う.

$$f(x) \sim N(\mu_{t-1}, \sigma_{t-1}^2), \quad where$$
$$\mu_{t-1}(x) = k(x, x_{1:t-1})[k(x_{1:t-1}, x_{1:t-1}) + I]^{-1} r_{1:t-1},$$
$$\sigma_{t-1}^2(x) = k(x, x) - k(x, x_{1:t-1})[k(x_{1:t-1}, x_{1:t-1}) + \rho^2 I]^{-1} k(x, x_{1:t-1})^{\mathrm{T}}.$$

$k(x, x_{1:t-1})$はxと$x_{1:t-1}$間のカーネル値を計算した列ベクトルであり，同様に$k(x_{1:t-1}, x_{1:t-1})$は$x_{1:t-1}$の各要素同士のカーネル値の行列である．ここでは一般化を失わず$\mu_0(x) = 0, \sigma_0^2(x) = k(x, x) > 0$

Satisficing Criteria in the Bandit Problem

†Shinji Nakazato, Graduates School of Management, Tokyo University of Science

‡Tetsuya Shimokawa, School of Management, Tokyo University of Science

とする．また，本研究では状態空間がGPによって形成されるものとし，毎期意思決定主体が直面する状態を$s_{t-1}=\{\mu_{t-1}(x),\sigma^2_{t-1}(x)|x\in D\}$とする．方策関数にはTompson Samplingを仮定し，$\pi(x|s)$と表記する．

意思決定主体は，現時点での状態(環境情報)を下に，将来利得の割引現在価値を最大化することを目的とする．しかしながら，各時刻において将来の選択経路は不明である。そのため本研究では，意思決定主体が将来利得について方策に対する期待値で評価をするものとする．このとき，t期における目的関数は時間割引因子をγとすれば，以下のように与えられる．

$$G(x_t|s_{t-1}) \coloneqq f(x_t)+\sum_{i=t+1}^{T}\gamma^{i-t}\pi(x_i|s_{i-1})f(x_i).$$

本モデルにおいては，意思決定主体は毎期Gで各選択肢に対する評価を行うことになる．ここで，上記を計算するにあたり，将来の状態経路が必要となるが，今回の問題設定では，GPを用いて状態空間の表現をしているため，GPRの更新式より，現時点での状態から，将来の状態経路を推定することが可能となる．結果，$G(x_t|s_{t-1})$は次のガウス分布の形で書き換えられる．

$$G(x_t|s_{t-1}) \sim N\Big(\mu_{t-1}(x_t)+\sum_{i=t+1}^{T}\gamma^{i-t}\sum_{x_i\in D}\pi(x_i|\tilde{s}_{i-1})\tilde{\mu}_{i-1}(x_i|s_{t-1}),\ \sigma^2_{t-1}(x_t)+\sum_{i=t+1}^{T}\gamma^{2(i-t)}\sum_{x_i\in D}\pi^2(x_i|\tilde{s}_{i-1})\tilde{\sigma}^2_{i-1}(x_i|x_{i-1},s_{t-1})\Big).$$

ただし，

$$\tilde{s}_t=\{\tilde{\mu}_t(x|s_{t-1}),\tilde{\sigma}^2_{t-1}(x|s_{t-1})|x\in D\},$$
$$\tilde{\mu}_t(x|s_{t-1})=\mu_{t-1}(x),$$
$$\tilde{\sigma}^2_t(x|x_t,s_{t-1})=\sigma^2_{t-1}(x_t)-\frac{\big(\sigma_{t-1}(x,x_t)\big)^2}{\sigma^2_{t-1}(x_t)+\rho^2}.$$

3，満足化行動

続いて，上記の行動モデルを下に，人間の意思決定でしばしば観測される満足化行動についての説明を試みる，満足化行動/基準とはSimon(1955)によって提唱されたものであり，目的関数を最大化する選択肢ではなく，個人によって異なる閾値以上を観測した別の選択肢で行動選択を収束させることである．ただし，現時点において同基準の数学的な定義は統一されていない．そこで本研究では，満足化行動を次のように定義する．

Definition(満足化行動)

任意の時刻tにおいて，$x^*\coloneqq\operatorname{argmax}_{x\in D}\mu(x)$に対して、(a)$\mu_{t-1}(x^*)\le\mu_{t-1}(x')$を満たすある選択肢$x'\in D\backslash x^*$が選択され，(b)また，その結果$\mu_t(x^*)\le\mu_t(x')$が維持されることを，行動が$x$にスタックすると呼ぶ．また，任意の時刻$t$以降のすべての時刻において行動が$x'$にスタックすることを，$x'$で満足化すると呼ぶ．

この定義は，現時点でグローバルに最適な選択肢よりも期待値高いと推定される選択肢が選択され，その結果，最適な選択肢とその選択肢の推定される期待値の大小関係が維持されることを要求するものである．

以上より，この定義の下で満足化行動が起こることを説明するためには，(a)と(b)の条件が生起する確率が時間経過とともに増加することを示すことになる．このことはGPRにおいて分散が単調に減少する性質と$G(x_t|s_{t-1})$の正規分布での表現から証明される．また，(a)と(b)の生起確率を増加させる要因について，その証明から，(1)$T-t$の減少はx'でスタックする確率を増加させ，(2)γの減少もまたx'でスタックする確率を増加させることが導かれる．この性質はSGUモデルでは考慮されない要素である．このことから，同問題における人間の行動モデルとして，どちらがより適切かを判断するためには，T,t,γの3つの値の変化が，行動選択に影響を及ぼすかどうかを確認すればよい．

4，実験/検証

本研究では，この点について東京理科大学の学生108名を対象にMAB問題を扱った実験を行いその検証を行った．実験には$T=5,8,15,25$の4種類の異なる意思決定期間のMAB問題を採用した．また、各被験者のもつ時間割引因子は実験での行動データと本研究の行動モデルを用いた最尤法から推定した．

実験データによる分析の結果，T,t,γがそれぞれ人間の意思決定に影響を及ぼすことが確認された．具体的には，意思決定期間の長い問題では，初期の時刻帯で，スタックする確率の低い選択肢を積極的に選択する傾向が観測されたが，この傾向は意思決定期間の短い問題では確認されなかった．また時間割引因子が高い被験者群と低い被験者群で探索行動の終了時刻に差異があるかどうかを検証した結果，そこに統計的に有意な差があることが確認され，時間割引因子の低いリスク回避的な被験者群ほど，探索行動を早期に終わらせる傾向をもつことが確認された．これらの結果はSGUモデルでは記述されることがないものである．

［主要参考文献］

[1] Wu, Charley M., et al. “Generalization guides human exploration in vast spaces.” *Nature human behavior*, 2(12):915-924, 2018.
[2] Burno B Averbeck. “Theory of choice in bandit, information sampling and foraging tasks.” *PLoS Comput Biol*, 11(3):e1004164, 2015.
[3] Herbert A Simon. “A behavioral model of rational choice.” *The quarterly journal of economics*, 99-118, 1955.