1M3-2

位置による割引を考慮した複数選択バンディット問題

On the position-based model for multi-play multi-armed bandit models

小宮山純平 *1　本多 淳也 *1
Junpei Komiyama　Junya Honda

*1東京大学
The University of Tokyo

We study a version of the multi-armed bandit problem in which several arms are drawn at each round. We propose an algorithm and discuss related optimization problems.

1. 概要

複数のクリック率の不明なオンライン広告をウェブサイトに配置する問題を考える。このとき、クリック率の高い広告から順に配置したいが、ユーザのフィードバックを見ながらクリック率を推定しオンライン的に配置を最適化する必要がある。下の位置にある広告は通常上位より見られないが、この割引効果を考慮した広告配置のオンライン最適化を多腕バンディット問題として定式化し、有効なアルゴリズムを提案する。

2. 背景

近年、多くのウェブサイトに広告が掲載されている。広告をクリックすると、広告主のウェブページに遷移するシステムになっている。検索エンジンに連動した広告（sponsored search）は通常 pay-per-click モデルであり、ユーザが広告をクリックすることによってウェブページの持主は収入を得る [Qin 14]。また、通常のウェブページの画像広告（display advertising）は pay-per-impression モデルであり、広告を表示するたびにウェブページの持ち主は収入を得る [Yuan 13]。いずれのモデルにせよ、ユーザが広告をクリックするということはその広告への興味の顕れであり、高いクリック率の広告を表示することは、ウェブサイトの広告収入の増加につながる。そのため、1 つの広告枠に表示できる関連広告が複数ある場合、それらの中からクリック率の高い広告を選んで表示したい。

多くのウェブサイトは、そのページ配置にもよるが複数の広告枠がある。通常、ユーザはページを上から下へブラウジングするため、上にある広告枠のほうがクリック率が高い。表示可能な K 個の広告のうち L 個の枠に配置する組合せは $K!/(K-L)!$ 通りあるため、この全パターンにクリック率を与えるモデルパラメータが付随すると考えると、K, L がある程度以上大きい場合はモデルの推定・最適化ともに困難である。したがって、より少ないパラメータで現実に沿ったモデルを考えたい。以下 3 つのモデルが考えられる。

1 位置依存なしモデル: それぞれの広告 i にはクリック率パラメータ θ_i が付随し、表示位置にかかわらず確率 θ_i でクリックされる。このモデルでは、同じ広告でも上位の枠に表示したほうがクリック率が高いということをモデル化できない。

2 カスケーディングモデル [Kempe 08]：このモデルでは、ユーザが上位の広告から順番に閲覧を行い、広告をクリックもしくは広告への興味がなくなった時点で離脱すると考える。このモデルでは、ユーザは複数の広告が同時にクリックされることはない。しかし、広告がクリックされる確率 *1 が表示する広告の順序によらないという不自然がある。

3 位置ベースモデル。広告のクリック率パラメータ $\{\theta_i\}$ だけでなく位置ごとのパラメータ $\{\kappa_l\}$ が存在し、広告 i が位置 l に表示された場合、確率 $\theta_i\kappa_l$ でそれぞれの広告が表示される。このモデルは、より良い（パラメータの大きい）広告が上に表示されたほうが全体のクリック率が大きくなる直観を表現できるメリットがある。逆に欠点としては、ある広告をクリックするかどうかが他の広告がクリックされるかどうかに影響を及ぼさず、多数の広告が同時にクリックされることがある可能性が残る点である。

広告のクリック率に関しては、バッチ（オフライン）で推定する手法もあるし、動的（オンライン）に推定するモデルも考えられる。オンラインで推定する方法は、以下の点で望ましい。(i) 新しいウェブサイトに対して、データがほとんどないため広告のクリック率の予測精度が低いような問題に関しても順次適用可能な点。また、(ii) 同じウェブサイトに関しても、モデルの定期的な再学習が容易な点がある。というのは、ウェブ広告のクリック率は時間的な変化が激しい [Agarwal 09] ため、ある時点で作成したモデルが後の時刻で使えないということが往々にして起こりうる。また、広告に対する流行などもあり、一般に同じ広告のクリック率は時間的に落ち込む傾向がある。これらの点を考慮すると、広告の最適化をオンラインで行うことができるアルゴリズムは非常に扱いやすい。

オンライン推定に関しては、バンディット問題 [Thompson 33, Robbins 52] による定式化が有効である。バンディット問題ではいくつかのアームと呼ばれる選択肢を考える。それぞれのアームは未知のパラメータを持つ確率分布に対応している。予測者は各ラウンドに、いずれかのアームを選び、アームに対応した確率分布からのサンプルを報酬として受け取る。賢いアルゴリズムを利用し、パラメータを動的に推定しながら期待平均報酬を最大化するのが予測者の目標である。つまり、広告のクリック率最大化問題では、

連絡先: 小宮山純平, 東京大学生産技術研究所, 〒 153-8505 東京都目黒区駒場 4-6-1, junpei@komiyama.info

*1 $\prod_{i\in I(t)}(1-\theta_i)$、ここで $I(t)$ は選択した L 個の広告の集合である。

各広告がアームに 1 対 1 対応しており、各ラウンドがユーザの来訪に対応している。報酬 $\{1,0\}$ をユーザがクリックしたかどうかに対応させると、クリック数の最大化はバンディット問題の報酬最大化に置き換えられる。

バンディット問題が提案されたのは強化学習 [Sutton 98] と呼ばれるクラスのアルゴリズムが現れるより前であるが、現代ではバンディット問題は報酬最大化を目指す強化学習の部分クラスとして位置づけられている。強化学習における普遍的なテーマは、探索と活用である。つまり、より多くの情報の取得を目的として選択を行うことと、現在の情報を利用して高い報酬を得ようとすることである。バンディット問題では、探索は一見クリック率が低そうな広告配置を試し、パラメータの推定精度を向上させること、活用はクリック率が現段階の情報で高そうな広告配置を試すことに対応している。

上記の設定では、各ラウンドに予測者は単一のアームを選択するため、広告枠が 1 つのウェブサイトを扱うことに対応している。しかし、各ラウンドに複数のアームを選択するモデルに拡張することによって、複数の広告枠があるウェブサイトのクリック率最適化に適用可能になる。過去、(i) 位置依存なしモデル [Anantharam 87, Komiyama 15a], (ii) カスケーディングモデル [Kveton 15]、(iii) 位置ベースモデル [Combes 15, Lagrée 16] においての広告クリック率の最適化の研究がある。

これらの既存研究は、いずれも枠の良さがユーザに既知であるということを仮定している。しかし、ウェブページの枠配置を変更した場合、このようなモデルでは再学習が必要になってしまう問題がある。近年の広告配信システムは非常に複雑であり、ある広告出稿者が非常に数多くの枠への配信を行うことが多く、それらの枠ごとに固定の枠の良さを記憶しておくと、そのページの状況の更新についていくのが困難である。本研究は広告の良さだけではなく、枠の良さもともに未知である場合にどのようにオンライン最適化を行えば良いかについて研究を行う。

3. 問題設定

アーム数を K とする。それぞれのアーム $i \in [K] = \{1,2,\dots,K\}$ にはパラメータ $\theta_i \in (0,1)$ が付随している。また、それぞれの広告枠 $l \in [L]$ にはパラメータ $\kappa_l \in [0,1]$ が付随している。このパラメータは $1 = \kappa_1 \geq \kappa_2 \geq \dots \geq \kappa_L$ とする。各ラウンド $t = 1,2,\dots,T$ に、予測者は $L \leq K$ 個のアーム $I(t)$ を選択し、対応する報酬 $\{0,1\}$ を受け取る。選択した $l \in [L]$ 番目のアームを $I_l(t)$ と書くと、このアームの報酬は各ラウンド独立に $\mathrm{Ber}(\theta_i \kappa_l)$ からサンプルされる。つまり、このモデルに出てくるパラメータは $\{\theta_i\}_{i\in[K]}$、$\{\kappa_l\}_{l\in[L]}$、および最終ラウンド数 T となる。これらのパラメータは予測者（アルゴリズム）には与えられない。$\kappa_1 = 1$ であること及び κ が枠位置に対して単調減少であることはアルゴリズムにとって既知とする。$N_{i,l}(t)$ をラウンド t 開始時までにアーム i が位置 l で選ばれた回数とする (つまり, $N_{i,l}(t) = \sum_{t'=1}^{t-1} \mathbf{1}\{i = I_l(t')\}$, ここで $\mathbf{1}\{\mathcal{A}\}$ はイベント $\mathcal{A}$ が成り立つなら 1、そうでなければ 0 と定義する)。また、$\hat{\theta}_{i,l}(t)$ をアーム i を位置 l で選択したときの経験平均とする。

予測者の目的は報酬の最大化である。簡単のため、すべてのアームのパラメータ θ_i は互いに異なるとする。一般性を失わずに、$\theta_1 > \theta_2 > \theta_3 > \dots > \theta_K$ とおける。もちろん、予測者はこのアームの配置順番を知らず、順番に依存したようなアルゴリズムを利用しないものとする。全パラメータの正確な推定ができているとすると、最も良いアームの選択方法は、パラメータの大きい順にアーム $1,2,\dots,L$ を位置 $1,2,\dots,L$ に配置することである。リグレット $\mathrm{Reg}(T)$ を最も良いアームを選択したときとアルゴリズムの選択との期待累計報酬の差として定義する：

$$\mathrm{Reg}(T) = \sum_{t=1}^{T} \left(\sum_{i\in[L]} (\theta_i - \theta_{I_i(t)})\kappa_i \right).$$

期待リグレット $\mathbb{E}[\mathrm{Reg}(T)]$ をアルゴリズムの性能指標として用いる（小さいほどよい）。

4. アルゴリズム

バンディット問題のアルゴリズムでよく知られているものは次の 3 つである。つまり、信頼上界（Upper Confidence Bound, UCB）法 [Lai 85, Auer 02]、事後確率サンプリング（Posterior sampling, PS もしくは Thompson sampling）法 [Thompson 33]、または最小経験ダイバージェンス（Minimum empirical Divergence, MED）法 [Honda 10] の 3 つである。UCB はそのコンセプトの簡単さや信頼区間が導出可能であること、PS は経験的な性能に優れるが、今回の問題には使いにくい。これは、今回の問題では広告をある枠に配置してクリックが得られなかった場合、広告のクリック率の低さと枠のクリック率の低さの両方が考えられ、直接パラメータを観測できないためである。このような問題を扱うには、MED が他の 2 手法とくらべて優れている [Komiyama 15b]。

直接パラメータを観測できない、今回のような問題は部分モニタリング問題（partial monitoring problem)[Piccolboni 01] という広いクラスの問題として知られている。部分モニタリング問題に対する最適な（Regret を漸近的に最小化する）アルゴリズムとしては PM-DMED[Komiyama 15b] が知られているが、PM-DMED を本問題に対して適用するのは以下の理由で非効率である。つまり、今回の広告最適化は $K + L$ 個のパラメータを持つが、これを部分モニタリング問題として表現すると、K, L に指数的な数のパラメータがある問題として冗長な表現をしなければいけない。そのため、計算的に非効率であり、また Regret の意味でも大きくなってしまう可能性がある。

これらを勘案しつつ、MED のあらたな拡張アルゴリズムを提案する。MED の原理は (i) 現在のデータからパラメータに対して一致性のある推定パラメータを作る (ii) 推定パラメータが真のパラメータだと仮定し、その結果最適なアーム配置の尤度を計算し、有意水準 $1/t$ 以下で最適な配置を決定できた場合にのみ活用を行い、それ以外では探索を行う、というものである。これに基づき、パラメータを最尤推定し推定したパラメータから探索と活用を決定するアルゴリズムを提案する（アルゴリズム 1)。

4.1 モデルパラメータの経験推定 $\{\hat{\theta}\}$, $\{\hat{\kappa}\}$

モデルパラメータの経験推定には最尤推定を用いる。本モデルの最尤推定は凸最適化であり、勾配法やヘッシアンを使う方法（Newton 法）などによって最適化可能である。

4.2 アームの必要探索量 $\tilde{N}_{i,l}$

$(1),\dots,(K)$ を $\hat{\theta}$ の降順で並べた順序とする。経験推定されたパラメータから、必要な探索量は測度変換の議論 [Lai 85, Graves 97] によって与えられる。この議論は以下のようになる。つまり、現在予想するアームの順番が仮に間違っていた場合、真のパラメータ θ' では、 1 つ以上のペア $i < j$ で

$$\theta'_{(i)} < \theta'_{(j)} \tag{1}$$

Algorithm 1 提案アルゴリズム

1: モデルパラメータ: $\alpha > 0$
2: **while** $t \leq T$ **do**
3: モデルパラメータの経験推定 $\{\hat{\theta}\}$, $\{\hat{\kappa}\}$ をこれまでの観測から得る
4: 経験推定されたパラメータが正しいと仮定して、それぞれのアームの必要探索量 $\tilde{N}_{i,l}$ を計算
5: 次の優先度順で枠 1 から昇順に各枠 $l \in [L]$ にアームを割り当てる。複数の同レベルの優先順位のアームがある場合は任意に選ぶ： 1. $N_{i,l} < \alpha\sqrt{\log t}$ であるアーム i 2. $N_{i,l} < \tilde{N}_{i,l}$ であるアーム 3. $\hat{\theta}$ の降順でソートして、これまでに割当されていないもので最も大きいもの
6: **end while**

となっていることになる。このようなパラメータから現在のデータが生成されるというある種の帰無仮説を有意性 $1/t$ で一様に棄却する必要がある。パラメータ $N_{i,s}, \hat{\theta}_i, \hat{\kappa}_s$ で特徴づけられる経験分布が真のパラメータ θ'_i, κ'_s から生成される負の対数尤度が

$$N_{i,s} d_{\mathrm{KL}}(\hat{\theta}_i \hat{\kappa}_s, \theta'_i \kappa'_s)$$

で与えられる。対数尤度を $\log t$ にするために最低限必要な探索量は、以下の最適化問題の解を $q_{i,j}$ とすると、$\tilde{N}_{i,j} = q_{i,j} \log t$ で与えられる:

$$\begin{aligned}
&\text{minimize} \\
&\quad \sum_{i\in[K]} \sum_{s\in[L]} q_{(i),s}(\hat{\theta}_s - \hat{\theta}_{(i)})\hat{\kappa}_s \\
&\text{subject to} \\
&\quad \forall_{1=\kappa'_1 \geq \cdots \geq \kappa'_L > 0} \\
&\quad \forall_{\theta'_{(i)},\ldots,\theta'_{(L)} \in (0,1): \forall_{i\in[L]} \hat{\theta}_{(i)}\hat{\kappa}_i = \theta'_{(i)}\kappa'_i} \\
&\qquad \sum_{i\in[K]} \sum_{s\in[L]} q_{(i),s} d_{\mathrm{KL}}(\hat{\theta}_{(i)}\hat{\kappa}_s, \theta'_{(i)}\kappa'_s) \geq 1 \\
&\quad \forall_{i\in[K], s\in[K]} q_{i,s} \geq 0 \\
&\quad \forall_{i,j} \sum_{s\in[K]} (q_{i,s} - q_{j,s}) = 0 \\
&\quad \forall_{s,t} \sum_{i\in[K]} (q_{i,s} - q_{i,t}) = 0
\end{aligned}$$

上記の最適化は、線形の目的関数に対して連続なパラメータ $\{\theta'_i, \kappa'_s\}$ による無限個の線形制約が入る半無限計画法 (Linear Semi-Infinite Programming, LSIP) である。LSIP を解く方法としては、Cutting-Set 法と呼ばれる動的に有限個の制約を追加･それらの制約に対して線形計画問題 (Linear Programming, LP) を解くことを繰り返す方法があるため [Mutapcic 09]、この方法を用いて最適化を行う。

5. おわりに

本研究では、複数の広告を表示する問題をバンディット問題として扱った。とくに、広告の良さだけではなく広告枠の良さが未知という既存研究の枠組を超えた問題設定を研究した。尤度に基づくアルゴリズムを提案した。このアルゴリズムによって、既存研究では枠の良さをバッチ学習、広告の良さをオンライン学習しているという「半オンライン学習状態」であったものが、完全にオンライン化する方法に目処がたったと言える。今後の目標としては、(i) 実装とアルゴリズムの性能検証 (ii) 最適化の理論的保証 (iii) アルゴリズムの Regret バウンドの証明が挙げられる。これらの点については現在進行形で研究を行っており本稿に載せられなかったが、講演時にはお見せできる予定である。

参考文献

[Agarwal 09] Agarwal, D., Chen, B.-C., and Elango, P.: Spatio-temporal models for estimating click-through rate, in *WWW*, pp. 21–30 (2009)

[Anantharam 87] Anantharam, V., Varaiya, P., and Walrand, J.: Asymptotically efficient allocation rules for the multiarmed bandit problem with multiple plays-Part I: I.I.D. rewards, *Automatic Control, IEEE Transactions on*, Vol. 32, No. 11, pp. 968–976 (1987)

[Auer 02] Auer, P., Cesa-bianchi, N., and Fischer, P.: Finite-time Analysis of the Multiarmed Bandit Problem, *Machine Learning*, Vol. 47, pp. 235–256 (2002)

[Combes 15] Combes, R., Magureanu, S., Proutière, A., and Laroche, C.: Learning to Rank: Regret Lower Bounds and Efficient Algorithms, in *Proceedings of the 2015 ACM SIGMETRICS International Conference on Measurement and Modeling of Computer Systems, Portland, OR, USA, June 15-19, 2015*, pp. 231–244 (2015)

[Graves 97] Graves, T. L. and Lai, T. L.: Asymptotically Efficient Adaptive Choice of Control Laws in Controlled Markov Chains, *SIAM Journal on Control and Optimization*, Vol. 35, No. 3, pp. 715–743 (1997)

[Honda 10] Honda, J. and Takemura, A.: An Asymptotically Optimal Bandit Algorithm for Bounded Support Models, in *COLT*, pp. 67–79 (2010)

[Kempe 08] Kempe, D. and Mahdian, M.: A Cascade Model for Externalities in Sponsored Search, in *WINE*, pp. 585–596 (2008)

[Komiyama 15a] Komiyama, J., Honda, J., and Nakagawa, H.: Optimal Regret Analysis of Thompson Sampling in Stochastic Multi-armed Bandit Problem with Multiple Plays, in *ICML*, pp. 1152–1161 (2015)

[Komiyama 15b] Komiyama, J., Honda, J., and Nakagawa, H.: Regret Lower Bound and Optimal Algorithm in Finite Stochastic Partial Monitoring, in *NIPS*, pp. 1792–1800 (2015)

[Kveton 15] Kveton, B., Szepesvári, C., Wen, Z., and Ashkan, A.: Cascading Bandits: Learning to Rank in the Cascade Model, in *ICML*, pp. 767–776 (2015)

[Lagrée 16] Lagrée, P., Vernade, C., and Cappé, O.: Multiple-Play Bandits in the Position-Based Model, in *NIPS*, pp. 1597–1605 (2016)

[Lai 85] Lai, T. L. and Robbins, H.: Asymptotically Efficient Adaptive Allocation Rules, *Advances in Applied Mathematics*, Vol. 6, No. 1, pp. 4–22 (1985)

[Mutapcic 09] Mutapcic, A. and Boyd, S. P.: Cutting-set methods for robust convex optimization with pessimizing oracles., *Optimization Methods and Software*, Vol. 24, No. 3, pp. 381–406 (2009)

[Piccolboni 01] Piccolboni, A. and Schindelhauer, C.: Discrete Prediction Games with Arbitrary Feedback and Loss, in *Computational Learning Theory, 14th Annual Conference on Computational Learning Theory, COLT 2001 and 5th European Conference on Computational Learning Theory, EuroCOLT 2001, Amsterdam, The Netherlands, July 16-19, 2001, Proceedings*, pp. 208–223 (2001)

[Qin 14] Qin, T., Chen, W., and Liu, T.: Sponsored Search Auctions: Recent Advances and Future Directions, *ACM TIST*, Vol. 5, No. 4, pp. 60:1–60:34 (2014)

[Robbins 52] Robbins, H.: Some aspects of the sequential design of experiments, *Bulletin of the AMS*, Vol. 58, pp. 527–535 (1952)

[Sutton 98] Sutton, R. S. and Barto, A. G.: *Introduction to Reinforcement Learning*, MIT Press, Cambridge, MA, USA, 1st edition (1998)

[Thompson 33] Thompson, W. R.: On The Likelihood That One Unknown Probability Exceeds Another In View Of The Evidence Of Two Samples, *Biometrika*, Vol. 25, pp. 285–294 (1933)

[Yuan 13] Yuan, S., Wang, J., and Zhao, X.: Real-time Bidding for Online Advertising: Measurement and Analysis, *CoRR*, Vol. abs/1306.6542, (2013)