F-040

バンディット問題への保守的な推定の導入に向けた一考察
A Study for Introducing Conservative Estimate to the Bandit Problem

木村 凌大[1)]　菊地 真人[1)]　大囿 忠親[1)]
Ryota Kimura　Masato Kikuchi　Tadachika Ozono

1　はじめに

多腕バンディット問題は，オンライン広告の最適化や，臨床実験の最適化など様々な問題解決に用いる．本研究では，T 回の試行から推定した期待報酬 $\hat{\mu}$ により，真の期待報酬 μ が最大となるアームを識別する “最適腕識別” の問題を扱う．特に，T が，μ の推定に不十分な大きさである場合を考える．このとき，報酬が大きくかつ当たる確率の低いアームの $\hat{\mu}$ を高めに見積もってしまうという課題がある．本研究では，このような場合に保守的な推定 [1] が有効であると考えている．保守的な推定とは，試行回数の少なさに応じて条件付き確率の推定値をあえて低めに見積もる推定法である．

本稿では，本研究における問題設定，ϵ-*greedy* への保守的な推定の導入および実験結果を示す．結果として，保守的な推定を導入することで，報酬が大きく確率が小さいアームに関する予測性能の向上が観測されたことを報告する．

2　問題設定

本研究では 2 腕バンディット問題を考える．プレイヤは，1 回の試行において 2 本のアーム $\{1,2\}$ から 1 本のアーム $i \in \{1,2\}$ を選択する．プレイヤが，アーム i を選択すると，確率 p_i で報酬 r_i が得られる．プレイヤにとって，p_i は未知であり，r_i は既知であるとする．すなわち，プレイヤは p_i の推定値 $\hat{p}_i$ を見積もることができれば，報酬の期待値 μ の推定値 $\hat{\mu}$ を見積もることができる．ここで，

$$\hat{\mu} = \hat{p} \cdot r$$

であり，最適なアーム i^* は，

$$i^* = \underset{i \in \{1,2\}}{arg\,max}\, \hat{\mu}_i$$

である．

プレイヤは，T 回の試行により p_i の推定値 $\hat{p}_i$ を求める必要があるとする．試行回数 T が少ないと，$\hat{p}_i$ が高めに推定される可能性がある．$\hat{p}(x|n)$ は，引いた回数 n のうち当たった回数 x の比 x/n で推定されるため，確率が低いアームが偶然にも最初の数回で当たった場合が当てはまる．例えば，$(p_1, r_1) = (10^{-3}, 10^3)$ の場合を考える．この時，$\mu_1 = 1$ である．また，$T = 10$ とする．この時，1 回目の試行でアーム 1 を引き，報酬が得られたとき，$\hat{\mu}_1 = 10^3 \gg \mu$ となる．以降，アーム 1 から報酬が得られなかったとしても，$\hat{\mu}_1 \gg \mu_1$ は変わらない．すなわち，$\hat{\mu}_1$ を過大に見積もっている．このような場合，保守的な推定では，試行回数が少ない場合に $\hat{\mu}_1$ を低めに見積もった方がよいと考える立場をとる．

1）名古屋工業大学大学院工学研究科工学専攻情報工学系プログラム，Computer Science Program, Department of Engineering, Graduate School of Engineering, Nagoya Institute of Technology

2）http://www.ss.cs.tut.ac.jp/CI-Laplace/

3　保守的な推定の導入

保守的な推定を説明し，その後 ϵ-*greedy* への保守的な想定法の導入について述べる．

まずは，保守的な推定 [1] について説明する．事象の観測頻度から確率の真値を推定するとき，一般的に最尤推定値や期待値が推定値として用いられる．保守的な推定では，観測頻度が小さい場合に，この推定値を小さめに見積もる方が安全と考える．保守的な推定を実現するための一手法として，確率の推定値として事後分布の信頼区間の下限値[2)]を用いることができる．すなわち，事前分布として何らかの分布を仮定し，結果を利用するときの適合率に応じて 2 事象間に成立する関係の強さを保守的に推定する．この推定値は，事後分布の分散を考慮した値となり，最尤推定では扱いにくい低頻度の事象に対しても適切に対処できる．今回，片側 95%信頼区間の下限値を用いることとした．例として，以下の 2 種類のくじに保守的な推定を用いる場合を考える．

- くじ 1：引いた回数 2 回，当たった回数 1 回
- くじ 2：引いた回数 100 回，当たった回数 50 回

この時，最尤推定ではどちらの推定値も 0.5 となる．しかし，くじ 1 では偶然 1 回当たった可能性があり，くじ 2 に比べて信用できない．これらのくじに保守的な推定を用いると，くじ 1 およびくじ 2 の推定値は，それぞれ 0.13 および 0.41 となる．したがって，引いた回数が少ないくじ 1 は推定値が低く見積もられ，引いた回数が多いくじ 2 は最尤推定による推定値に近い値となる．

つぎに，ϵ-*greedy* への保守的な推定の導入について説明する．ϵ-*greedy* は，バンディット問題の戦略の一種である．ϵ-*greedy* では，アームの条件付き確率の推定値 $\hat{p}(x|n)$ を，引いた回数 n に対する当たった回数 x の比で推定する．このとき，確率 ϵ で $\hat{\mu}$ が最大となるアーム i を選択し，それ以外は，ランダムにアームを選択する．ここで，$\hat{\mu}$ の推定に保守的な推定を用いた ϵ-*greedy* を，*conservative-*ϵ*-greedy* (以降，*consv-*ϵ*-greedy*) と呼ぶ．*consv-*ϵ*-greedy* では，$\hat{\mu} = \theta_{lb} \cdot r$ とする．ここで，θ_{lb} は，保守的な推定により得られた $\hat{p}(x|n)$ の推定値である．

保守的な推定の効果について説明する．例として，以下の 2 種類のアームをどちらも 10 回引いた場合 ($T = 10$) を考える．

- アーム 1：$p_1 = 1/20$，$r_1 = 6$，$\mu_1 = 0.3$
- アーム 2：$p_2 = 1/2$，$r_2 = 1$，$\mu_2 = 0.5$

$\mu_1 < \mu_2$ より，アーム 2 が最適である．p_1 および p_2 の大きさを考えると，$T = 10$ では，アーム 1 からは 1 回も報酬が得られないことが少なくない．一方，アーム 2 からは 5 回程度は報酬が得られると期待できる．ここで，偶然にもアーム 1 が 1 回当たった場合，

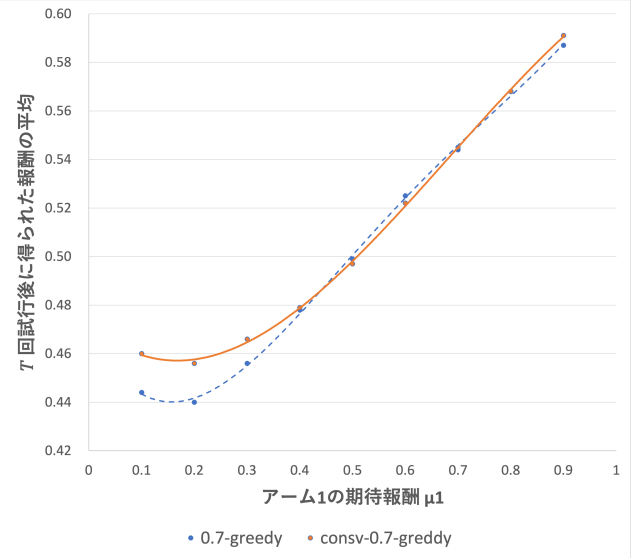

図 1　アーム 1 の期待報酬 μ_1 と平均獲得報酬（$\epsilon = 0.7$）

$\hat{p}_1(1|10) = 1/10$ となり，$\hat{\mu}_1 = 6 \cdot 1/10 = 0.6$ となる．このとき，$\hat{\mu}_1 > \hat{\mu}_2 \approx 0.5$ となる可能性があり，アーム 1 を最適と誤認する．一方，保守的な推定を用いれば，$\theta_{lb1} = 0.033$，$\theta_{lb2} = 0.271$ となる．よって，$\hat{\mu}_1 = 0.198$，$\hat{\mu}_2 = 0.271$ となるから $\hat{\mu}_1 < \hat{\mu}_2$ となる．よって，試行回数が少ない場合でも，保守的な推定によりアーム 2 が最適であると判断できる．

4　実験

最適アームの選択における $\hat{p}_1$ の過大推定による失敗の抑制が保守的な推定により可能になるかを調べたい．本実験の目的は，アームを引く試行回数 T が少なく，$p_1 \gg p_2$ である場合における，保守的な推定方法による挙動の変化を観察することである．具体的には，p_1, p_2, r_2 を固定し，r_1 を変化させたときの挙動の変化を観察する．$\mu_1 < \mu_2$ の場合は，$\hat{\mu}_1 < \hat{\mu}_2$ となることが好ましいが，T が少なく，p が微少な場合には，推定を誤る可能性がある．ここでは，保守的な推定によりその誤りを抑制可能であることを確認したい．

本実験では，次のように 2 本のアーム 1, 2 を設定した．

- アーム 1：$p_1 = 1/20$，$r_1 \in \{2, 4, 6, \ldots 18\}$
- アーム 2：$p_2 = 1/2$，$r_2 = 1$，$\mu_2 = 0.5$

ここで，$r_1 = 2, 4, 6, \ldots, 18$ と変化させた．すなわち，$\mu_1 = 0.1, 02, \ldots, 0.9$ と変化させた．$r_1 = 10$ のとき，$\mu_1 = \mu_2 = 0.5$ となり，問題の性質が変わる．具体的には，$\mu_1 < \mu_2$ のとき，すなわち，$r_1 < 10$ の場合は，アーム 2 を選択すべきである．また，$\mu_1 = \mu_2 = 0.5$ を境界として，$\mu_1 > \mu_2$ となるため，アーム 1 が最適となる．

プレイヤは，T 回の試行後にアームを選択する．これを 1 セットと呼び，10^6 セット実行した．本実験では，ϵ-*greedy* および *consv*-ϵ-*greedy* の 2 戦略を用いた．$\epsilon \in \{0.1, 0.2, ..., 0.8, 0.9\}$ とし，ϵ-*greedy* および *consv*-ϵ-*greedy* において，同一の ϵ を用いた．

図 1 に実験結果を示す．ここでは，$\epsilon = 0.7$ の ϵ-*greedy* である，0.7-*greedy* および *consv*-0.7-*greedy* の結果を示している．図 1 の横軸は μ_1 を表し，縦軸は T 回の試行後に得られた報酬の平均を表している．*consv*-ϵ-*greedy* は，$\mu_1 < 0.4$ において，ϵ-*greedy* を上回っている．これは，ϵ-*greedy* に比べて，アーム 1 よりもアーム 2 をより多く選択していることを意味している．$\mu_1 \geq 0.4$ では，両手法間に大差がないようにみえる．

予想通り，$\mu_1 < 0.4$ では，*consv*-ϵ-*greedy* が $\hat{p}_1$ の過大推定を抑制できているといえる．特に，$\mu_1 = 0.1$および0.2 の場合の有効性は明らかである．しかし，$\mu_1 = 0.4$ における保守的な推定の効果は確認できなかった．これは，μ_1 と μ_2 との差が小さいことから，妥当であるとも考えられる．

一方，保守的な推定により $\hat{p}_1$ が過小推定され，$\mu_1 > \mu_2$ であるにも関わらず，$\hat{\mu}_1 < \hat{\mu}_2$ と過小評価されると予想していた．すなわち，*consv*-ϵ-*greedy* の結果が，$\mu_1 > 0.5$ において悪化することを予想していた．結果としては，ϵ-*greedy* および *consv*-ϵ-*greedy* は類似した結果となった．この理由について考察する．今回の設定では，$p_1 \ll p_2$ であるから，保守的な推定により $p_1 - \theta_{lb1} > p_2 - \theta_{lb2}$ となる可能性が高い．これは，p_1 が小さい場合，報酬が得られる回数が少ないことから，保守的な推定により $p_1 - \theta_{lb1}$ が大きくなるからである．ただし，p よりも r の定義域が大きいことから，$r_1 \gg r_2$ の場合，$\theta_{lb1} \cdot r_1 > \theta_{lb2} \cdot r_2$，すなわち，$\hat{\mu}_1 > \hat{\mu}_2$ となる可能性が高くなる．今回の事例では，これが *consv*-ϵ-*greedy* にとって良い効果を与えている．言い換えれば，r_1/r_2 によって，$p_1 - \theta_{lb1}$ をよりよく制御する必要性があることを示唆している．例えば，適切に信頼区間を変えることが考えられる．

本研究では，試行回数 T が不十分な場合という特殊な状況を考えている．このような状況において，多腕バンディット問題に関する他の戦略に関しても，$\hat{\mu} = \hat{p} \cdot r$ とする場合は，同様の問題を持つ．多腕バンディット問題において一般的な，累積報酬最大化問題では，報酬の期待値の推定値を高めに見積もることで，真っ当な戦略となるように設計されている．一方，本研究と同様な最適腕識別問題では，期待値の推定値を低めに見積もることも有効である．ただし，T や p が小さな状況においては確率収束が期待できないため，経験的な手法の導入が必要であると考えられる．保守的な推定に基づく分析により，何らかの知見が得られることが期待される．

5　おわりに

2 腕バンディット問題における ϵ-*greedy* への保守的な推定の導入および実験結果を示した．保守的な推定とは，試行回数の少なさに応じて条件付き確率の推定値をあえて低めに見積もる推定法である．ϵ-*greedy* に保守的な推定を導入することで，報酬が大きく確率が小さなアームに関する予測性能の向上が観測された．

謝辞

本研究の一部は JSPS 科研費 JP19K12266, JP22K18006 の助成を受けたものです．

参考文献

[1] 菊地 真人, 山本 英子, 吉田 光男, 岡部 正幸, 梅村 恭司, “条件付き確率の保守的な推定”, 電子情報通信学会論文誌 D, Vol.J100-D, No.4, pp.544-555, 2017.