学 位 論 文 内 容 の 要 旨

博士の専攻分野の名称　博士（情報科学）　氏名　渡邊　僚

学 位 論 文 題 名

Analysis of Policies for Budgeted Multi-Armed Bandit Problem and Matching-Selection Bandit Problem

（予算制約付き多腕バンディット問題およびマッチング選択バンディット問題に対する方策の解析）

多腕バンディット問題とは、与えられた行動の選択肢から一つを選び、対応した報酬を得るというラウンドを繰り返す設定において、報酬の合計 (累積報酬) を最大化する問題である。ただし、選んだ行動に対する報酬しか情報はフィードバックされず、他の行動を選んでいたら報酬はいくらであったかは明らかにされない。そのため、情報収集のための探索とそれまでに得られている情報を使った知識利用をバランス良く行う必要がある問題として知られている。この問題において、ある回数ラウンドを繰り返す場合に、累積報酬を最大化する選択ルール (方策) を設計する研究がなされてきた。

方策の設計は、報酬がどのように生成されるかに大きく影響される。大きく分けて確率的に生成されるという設定と、方策を知っている敵が選ぶという設定がある。本研究ではより応用範囲の広い確率的な設定のみ扱う。また、累積報酬そのものは問題依存であり、方策の評価指標として適切でないため、最適な方策 (期待累積報酬最大の方策) との期待累積報酬の差であるリグレットで方策を評価する。

多腕バンディット問題の方策として、今までに様々な方策が提案され、そのリグレット解析がなされてきた。主なものとしては、UCB (Upper Confidence Bound) 方策や KL-UCB (Kullback-Leibler UCB) 方策 のような期待報酬の信頼上界に基づくもの、最適ではない行動の誤選択率を直接制御する DMED (Deterministic Minimum Empirical Divergence) 方策、事後確率分布に基づいてサンプリングした値で行動を評価する Thompson Sampling が知られており、それぞれリグレットの上界が証明されている。

一方、バンディット問題における理論限界、つまり厳密なリグレット下界も証明されている。実際、上で主なものとして挙げた 4 つの方策のうち UCB 方策を除く 3 つの方策は、それらのリグレット上界の支配項はラウンド数の関数としてみた場合、下界の支配項と漸近的に係数まで一致する。リグレット上界が下界と漸近的に一致する方策は、漸近最適方策 (Asymptotically Optimal Policy) と呼ばれており、KL-UCB 方策、DMED 方策及び Thompson Sampling は漸近最適方策であると言える。

本論文では、多腕バンディット問題の一般化である「予算制約付き多腕バンディット問題」及び「マッチング選択バンディット問題」に関して、主にリグレット解析を行う。予算制約付き多腕バンディット問題では、2 つの方策を提案し、それらの有効性をリグレット解析及びシミュレーション実験により示す。また、マッチング選択バンディット問題に関しては、既存の方策である LLR (Learning with Linear Rewards) 方策のリグレット上界を改善する。

予算制約つき多腕バンディット問題とは、行動の選択に際しコストが発生し、コスト総和が決められた予算を超えるまでのあいだ、ラウンドを繰り返すことのできるバンディット問題である。本

論文では、行動毎にコストが異なるのみでなく、それが未知の確率分布に従って確率的に定まる場合を扱う。このバンディット問題に対するリグレット下界を示すと共に、KL-UCB 方策と UCB 方策をこの問題設定に合うようにそれぞれ拡張した KL-UCB-SC (KL-UCB for Stochastic Costs) 方策及び UCB-SC (UCB for Stochastic Costs) 方策を提案し、リグレット上界を求め、シミュレーション実験により有効性を示す。

KL-UCB 方策が予算制約のない普通のバンディット問題の漸近最適方策であるように、KL-UCB-SC 方策もまた、予算制約付き多腕バンディット問題の漸近最適方策であることを、リグレット上界を分析することにより示す。KL-UCB-SC 方策はリグレット評価においては高性能であるが、各々の行動に対する選択指標値を計算するのに最適化問題を解かなければならず、計算効率のよい実装が難しいという問題点がある。一方、UCB-SC 方策は、漸近最適方策であることを証明できないものの、各々の行動に対する選択指標値の計算式が閉形式で書けるため、計算効率の良い実装が可能である。シミュレーション実験では、これらの理論評価を裏付ける結果が得られたほか、KL-UCB-SC の変種である KL-UCB-SC+ は、既存手法の中で最も性能が良いとされている BTS (Budgeted Thompson Sampling) と同等の性能を示した。また、UCB-SC はこれら 2 つの方策よりかは実験によるリグレット性能も劣るものの、計算効率はこれらより良く、変種の UCB-SC+ は計算効率の良い他の方策 (PD-BwK, UCB-BV1) よりも良いリグレット性能を示した。

マッチング選択バンディット問題は、各ラウンドにおいて完全 2 部グラフにおける最大マッチング (共通の端点を持たない最大辺集合) の 1 つを選択し、選択した辺集合に対応する報酬を得る設定の問題である。この問題は、全ての辺の集合から、条件を満たす辺集合を選択する組み合わせバンディット問題の一種である。報酬は辺毎に発生し、選択した辺集合の辺各々から報酬が得られるという半バンディット (semi-bandit) 設定を扱う。本論文では、$M \leq N$ を満たす自然数 M, N に対し、完全 2 部グラフ $K_{M,N}$ 上のマッチング選択バンディット問題において、既存方策である LLR 方策のリグレット上界を $\Theta(M^{2/3})$ 倍改善する。