審 査 の 結 果 の 要 旨

論 文 提 出 者 氏 名　　小 宮 山　純 平

多腕バンディット問題とは、多数の選択肢の中から1つを選び、選んだ選択肢に関してだけ報酬が分かるという環境で、最適な選択肢をできるだけ少ない回数の選択の繰り返しで推定する問題である。具体的応用としては、オンライン広告におけるクリックスルー率を最適化するために、どのような広告をどの場所に掲示するべきかを推定する問題が挙げられる。一般的に、ABテストのように現実世界における試験的行為によって複数の選択肢から最適なものを選ぶタイプの問題は数多く存在し、数理情報学分野において最適化に関する重要な問題と認識されている。本論文は従来から知られている多腕バンディット問題を拡張し、選択肢がある期間ロックアップされる場合、1回の選択で複数の選択肢を選べる場合、選択肢の直接的な報酬ではなく、2つの選択肢の結果のうちどちらがよいかという情報だけを得られる場合のアルゴリズムを提案している。これらに対して、多腕バンディット問題の評価指標である最適解への漸近性に関して、最適値に比較した場合の提案手法の損失すなわちリグレット下界、上界などの数理モデルを明らかにし、理論的および実験的評価を行っている。

本論文は「Asymptotically Optimal Multi-armed Bandit Algorithms Aimed at Online Contents Selection 」（オンラインコンテンツ選択のための漸近最適な多腕バンディットアルゴリズム）と題し、7章からなる。

第１章「Introduction」（序論）では、多数の選択肢からの選択と報酬という枠組みにおいて、探索と活用という形式で多腕バンディット問題の概念説明を行い、次に本論文で扱うアルゴリズムの導入説明を行っている。最後に本論文の構成を述べている。

第２章「Framework of Multi-armed Bandit Problem」（多腕バンディット問題の枠組み）では、多腕バンディット問題の3つのアプローチ、すなわちベイズ的、確率的、および敵対的アプローチを導入し、各々に関してリグレットを与えて比較している。また、確率的アプローチのリグレット下界を導出している。なお、3章以降で扱っているのは全て確率的アプローチである。

第3章「Algorithms for Multi-armed Bandit Problem」（多腕バンディット問題のためのアルゴリズム）では、多腕バンディット問題に対して提案されている4つのアルゴリズムを紹介している。 ε -greedyは一様なサンプリングによって選択する簡単な方法であるが、漸近最適性が保証されない。Upper Confidence Bound(UCB)は確実なインターバルにおける最適戦略であり、Thompson Sampling (TS) は事後確率を用いたサンプリング方式、Deterministic Minimum Empirical Divergence(DEMD)は尤度を基礎にする探索方式である。これら３つの方式に関して十分長い時間Tを経た後のTをパラメターとした場合の漸近的なリグレットの上界の評価式を求め、いずれもリグレットに関して漸近的最適性があることを示している。ここで示された評価式は従来知られているものより厳密な解になっている。

第4章「Multi-armed Bandit Problem with Lock-up Periods」（ロックアップ期間付き多腕バンディッ

ト問題)では、確率的多腕バンディット問題の拡張として、ひとつの選択肢を選ぶと予め決められた期間はその選択肢しか選ぶことができないロックアップ期間付きの場合を検討している。ロックアップ期間をLmaxとするとリグレットは全選択期間Tに対してlog(T)+Lmaxのオーダーになることを示している。さらにある時点までの実験的に最適な選択肢を長いロックアップ期間に選ぶ手法を提案し、リグレットの評価式を求め、シミュレーションによって有効性を示している。

第5章「Asymptotically Optimal Exploration and Exploitation in Multiple-play Multi-armed Bandit Problem」(複数選択型多腕バンディット問題における漸近最適な探索と活用)では、Thompson Sampling において1回の選択で複数の選択肢を同時に選べるアルゴリズムを提案し、そのリグレットを分析している。その結果、報酬が２値の場合、本論文で求めたリグレットの上界が、既に知られているリグレットの漸近的下界に一致することを示している。また、シミュレーションによって提案したアルゴリズムが他の手法に比べてリグレットを改善していることを示している。

第6章「Regret Lower Bound and Asymptotically Optimal Algorithm in Dueling Bandit Problem」(一対比較型バンディット問題におけるリグレット下界と漸近最適アルゴリズム)では、選んだ選択肢の直接の報酬は分からないが、一対の選択肢を比較して長短を決めることができる問題を扱っている。このような問題設定は、情報検索結果の表示や推薦商品の好感度など数値化しにくい報酬の場合に対応する。提案するアルゴリズムにおいて、他の全ての選択肢に勝る選択肢が存在する場合、情報ダイバージェンスに基づく漸近的リグレット下界を求め、これがリグレット上界と一致することを初めて示した。また、シミュレーションにより、提案したアルゴリズムが同じ問題設定に対する既存のアルゴリズムよりも格段に性能を改善したことも示している。

最後に第7章「Conclusions and Discussions」(結論と議論)では、本論文の成果を簡潔に纏めると共に、今後の研究課題を提示している。

以上を要するに、本論文は広告などのオンラインコンテンツの選択において現れる確率的多腕バンディット問題を対象にして、３種類の問題に対して新規なアルゴリズムを提案して、さらにリグレットの評価式を求め、提案手法の実用性を示している。この結果は、インターネットの普及によって重要性が増しているコンテンツ選択の問題に対する有効な解決策を示すことによって、数理情報学分野の技術発展に寄与した。

よって本論文は博士（情報理工学）の学位請求論文として合格と認められる。