Multi-Armed Bandit Algorithms: A Comprehensive Investigation of Theory, Applications, and Future Directions

Xingjian Qu*

The Department of Computer Science, Georgetown University, Washington D.C, 20057, United States

*Corresponding author's e-mail address: xq59@georgetown.edu

Abstract. This review comprehensively elucidates the dynamics of Multi-Armed Bandit (MAB) algorithms, highlighting their progression, applications, and potential for future research. This paper conducts a meticulous examination of key MAB algorithms, including Explore-Then-Commit (ETC), Upper Confidence Bound (UCB), Thompson Sampling (TS), and a noteworthy variant, was undertaken. Their intrinsic concepts, formulas, and workflows were dissected, anchoring the discussion in a foundation of theoretical understanding. The exploration extended to real-world applications, providing insights into how these algorithms have been actualized across sectors. Real-world deployments, from personalized content recommendations in online platforms to optimizing clinical trial outcomes, were brought to the fore, evaluating both their strides and constraints. MAB algorithms, notably the UCB and TS approaches, have exerted a profound influence across diverse domains, thereby engendering heightened efficiency and facilitating judicious decisional processes characterized by optimality. Nevertheless, persisting challenges manifest, particularly in their capacity to flexibly accommodate dynamic real-world contexts, alongside the ethical considerations arising from their applications. While MAB algorithms have manifestly affected transformative outcomes within environments beset by decisional ambiguity, the scope for further advancement remains conspicuous. Prospective scholarly inquiry might pivot towards the nuanced enhancement of real-time adaptiveness mechanisms and the seamless incorporation of prolonged temporal reward indicators, thereby amplifying their overall effectiveness. Given the poised trajectory of MABs towards an augmented amalgamation with technological frameworks, their indispensably formative role in configuring data-steered decisional paradigms becomes an incontrovertible proposition.

Keywords: MAB algorithm, ETC, UCB, TS.

1. Introduction

Multi-Arm Bandit (MAB) problems constitute an intriguing category of decision-making challenges, where a decision-maker (referred to as the agent) is tasked with the recurrent selection between different alternatives (referred to as arms), in an attempt to optimize a given reward over a prolonged duration [1]. The concept of MAB problems is deeply rooted in the theories of probability and statistics. The origin of the term "Multi-Arm Bandit" lies in the analogy of a gambler, engaged in the task of playing multiple slot machines (also known as one-armed bandits), while simultaneously striving to optimize the total winnings.

MAB problems have demonstrated to be of critical importance due to their extensive applicability in various domains. They are employed in fields as varied as clinical trials, online advertising, recommendation systems, network routing, and even reinforcement learning [2] These problems pose a unique challenge in that they necessitate a balance between exploration, which involves sampling lesser-known arms to gather information, and exploitation, the process of employing the currently known best-performing arm. This balance is crucial for the maximization of the cumulative reward.

Over the course of history, a multitude of MAB algorithms have been proposed to address these problems. One such example is the Upper Confidence Bound (UCB) algorithm, proposed by Auer et al. [3], which utilizes an optimism-in-the-face-of-uncertainty principle to strike a balance between exploration and exploitation. This algorithm has found widespread application in domains ranging from web service recommendation to the selection of news articles. In more recent times, the Thompson Sampling algorithm has surged in popularity, due to its empirical successes and robustness

across a wide array of applications [4]. This Bayesian approach to the MAB problem has demonstrated remarkable efficacy in scenarios like large-scale A/B testing and the placement of online advertisements [5].

Despite the significant advances in the field of Multi-Arm Bandit (MAB) algorithms, there remains an untapped potential in exploring the intricacies of various MAB algorithms and their broad applications. This field, central to reinforcement learning and optimization, has seen substantial growth and diversification, with applications spanning from online advertising to healthcare. However, the complexity and richness of different algorithms, such as Upper Confidence Bound (UCB), Thompson Sampling, and others, require a detailed and nuanced examination. Understanding their underlying principles, the contexts in which they flourish, and the challenges they may face in various applications can provide critical insights for both academia and industry. In response to this, there is a compelling need for a dedicated review that not only introduces these algorithms but also delves into their functionalities, comparisons, and applications across diverse domains, providing a holistic perspective on the current state and future possibilities of MAB algorithms.

WIth this purpose, this review aims to conduct an exhaustive analysis of various MAB algorithms, including but not limited to UCB and Thompson Sampling, along with their applications across diverse domains. Drawing upon a vast body of work, the review offers insights into the progression of MAB algorithms over time and delineates the contexts in which certain algorithms may prove more beneficial than others. It is intended that this work will guide future research in the MAB field and contribute to a deeper understanding and broader application of these algorithms.

2. Method

2.1. Overview of MAB

Multi-Arm Bandit (MAB) algorithms represent a class of optimization and decision-making problems, where an agent must choose among multiple options (or "arms") to maximize a cumulative reward over time. Drawing an analogy to a gambler selecting from a row of slot machines (bandits), the challenge lies in balancing the exploration of unknown or less-tried options with the exploitation of the ones that have provided good returns in the past. The MAB framework is highly adaptable and has found applications in diverse domains such as online advertising, clinical trials, recommendation systems, and more. Its variations and extensions have led to the development of specific algorithms tailored to different scenarios and constraints.

2.2. Explore-then-commit (ETC) algorithm

The Explore-Then-Commit (ETC) algorithm is one of the foundational strategies in the MAB domain, structured around a two-phase approach: exploration and exploitation. It initially explores all available arms for a defined period, gathering empirical data without committing to any specific option. After this exploration phase, it commits to the arm that has performed best, exploiting it for the remaining period. The purpose of exploration is to identify the optimal arm without making premature judgments, while the commitment phase aims to maximize the rewards based on the insights gained.

2.2.1 Formulas and workflow

1) Exploration Phase: For a given number of rounds, m, pull each arm an equal number of times.
2) Calculate Average Reward: Compute the average reward for each arm based on the data collected.
3) Commitment Phase: Select the arm with the highest average reward and exploit it for the remaining rounds. Mathematically, the average reward for arm i is given by:

$$\underline{r_i} = \frac{1}{m}\sum_{j=1}^{m} r_{ij} \quad (1)$$

where r_{ij} is the reward from the j-th pull of arm i.

ETC is simple and robust but may not always be efficient, especially when the exploration phase is not carefully tuned. It serves as a starting point and offers insights for more sophisticated MAB algorithms.

2.3. Upper confidence bound (UCB) algorithm

The UCB algorithm is a popular MAB algorithm that efficiently balances the trade-off between exploration and exploitation. Unlike ETC, UCB doesn't have distinct phases. It continuously adjusts its choices based on the performance of the arms pulled so far. UCB selects arms based on an upper bound on the potential value of each arm. This bound is calculated to be optimistic, favoring arms with less information, thus encouraging exploration.

2.3.1 Formulas and workflow

Initialize: Pull each arm once. **2) Calculate UCB:** For each arm, calculate the UCB index as:

$$UCBi(t) = \underline{r_i}(t) + \sqrt{\frac{2\,log\,t}{n_i(t)}} \tag{2}$$

where $\underline{r_i}(t)$ is the average reward of arm i after t rounds, and $n_i(t)$ is the number of times arm i has been pulled.

3) Select Arm: Choose the arm with the highest UCB index. **4) Update:** Update the average reward for the selected arm.

UCB is particularly effective when the arms' reward distributions are sub-Gaussian. Its logarithmic regret ensures good performance over time.

2.4. Thompson sampling (TS) algorithm

TS is a probabilistic MAB algorithm that has gained attention for its empirical success and theoretical grounding. It utilizes Bayesian inference to model the uncertainty about the true mean reward of each arm. Rather than relying solely on point estimates, TS samples from the posterior distribution of each arm's mean reward. This sampling naturally balances exploration and exploitation.

2.4.1 Formulas and workflow

1) Initialize Prior: Set a prior distribution for each arm's mean reward. **2) Sample Rewards:** Draw a sample from the posterior distribution of each arm's mean reward. **3) Select Arm:** Choose the arm with the highest sampled value. **4) Update Posterior:** Update the chosen arm's posterior distribution using the observed reward.

Mathematically, if the rewards are Bernoulli, the posterior update with a Beta prior is: Beta ($\boldsymbol{\alpha}$ + successes, $\boldsymbol{\beta}$ + failures). TS's flexibility in modeling the arms' reward distributions allows it to be adapted to various contexts. It has shown empirical effectiveness in a variety of applications, ranging from recommendation systems to clinical trials, underscoring its versatility.

2.5. LinUCB algorithm

The LinUCB algorithm is a contextual MAB algorithm that extends the idea of UCB to situations where contextual information about the arms is available. In many real-world scenarios, the decision on which arm to pull may depend on contextual information. LinUCB leverages this context to make more informed decisions. LinUCB assumes a linear relationship between the context and the expected reward of an arm. It then utilizes ridge regression to estimate the model parameters.

2.5.1 Formulas and workflow

1) Initialize Parameters: Set the parameters for ridge regression, including the regularization term. **2) Calculate Confidence Bound:** For each arm, calculate the upper confidence bound as:

$$UCBi(t) = \theta^T x_i + \alpha\sqrt{x_i^T A_i^{-1} x_i} \tag{3}$$

Where θ is the estimated parameter vector, x_i is the context vector, and A_i is the matrix related to the arm's historical context.

3) Select Arm: Choose the arm with the highest UCB index. **4) Update Parameters:** Update the ridge regression parameters using the observed reward.

LinUCB provides a powerful framework for handling complex environments where arms are associated with contextual information. Its ability to leverage context has found applications in targeted advertising, personalized recommendations, and other domains where context is crucial.

3. Applications and discussion

3.1. Online advertising

The landscape of online advertising platforms has undergone a significant metamorphosis through the integration of MAB algorithms. Among these algorithms, the UCB approach and the LinUCB algorithm have emerged as pivotal contributors to this transformative process. A concrete manifestation of this influence is observed in the strategic adoption of the LinUCB algorithm by Yahoo!. This implementation found application within Yahoo!'s personalized news recommendation system, precisely situated on its front-page interface [6]. This allowed them to make real-time ad bidding decisions based on a user's history, behavior, and other contextually relevant factors. Online advertising is a dynamic domain, marked by rapidly changing user preferences, competitors' actions, and market conditions. Ensuring optimal ad delivery in such a volatile environment is challenging. MABs sometimes make simplistic assumptions which may not hold true in real-world scenarios. There's also the risk of falling into local optima rather than achieving a global understanding of user behavior. Advanced machine learning models are expected to converge with MABs, offering more adaptive strategies. This holds the promise of real-time online advertising strategies that not only focus on immediate user engagement metrics like click-through rates but also take into account longer-term brand building and loyalty generation.

3.2. Healthcare

Healthcare, a domain where decisions can have life-altering consequences, has embraced the MAB approach, especially in clinical trials. Thompson Sampling has found utility in adaptive patient assignment to treatment arms, thus optimizing treatment outcomes [7]. This results in personalized medicine, tailoring treatments to individual patient profiles. Within the domain of healthcare, the utilization of MAB strategies navigates a nuanced equilibrium between exploratory endeavors involving novel treatments and the exploitation of established efficacious interventions. This duality encapsulates the potential hazard inherent in explorative pursuits, which might inadvertently result in suboptimal therapeutic selections. Alongside this concern, the ethical dimensions underpinning the notion of "experimentation" within healthcare are notably consequential. The healthcare landscape, by its very nature, presents an environment conducive to the integration of MABs that can assimilate more granular patient-specific data and domain expertise. Evident on the horizon is the evolutionary trajectory toward adaptive clinical trials, characterized by a dual emphasis on upholding patient well-being and concurrently unearthing innovative avenues for therapeutic intervention. This paradigmatic evolution envisages a harmonious reconciliation of the imperatives of safety and the pursuit of pioneering treatment modalities.

3.3. E-commerce

E-commerce giants like Amazon and Alibaba harness the power of MAB algorithms for enhancing their product recommendation systems [8, 9]. The Explore-Then-Commit algorithm is of particular interest in dynamic e-commerce settings, where inventory and product ranges are constantly evolving. While MABs are instrumental in driving immediate e-commerce metrics like clicks and purchases, there's a potential blind spot when considering long-term user engagement. Metrics like customer

loyalty, return rates, and long-term value can sometimes be overshadowed by the immediate rewards that MABs focus on. The future beckons e-commerce platforms to incorporate long-term reward signals into their MAB frameworks. This would mean crafting recommendation algorithms that don't just lead to an immediate sale but foster long-term customer relationships.

3.4. Network routing

As the digital world becomes more connected, the importance of efficient data flow across networks cannot be overstated. MABs, especially the UCB algorithms, have been critical in dynamic routing decisions. For instance, in managing congestion in data center networks, MAB-based algorithms have proven to be robust and efficient [10]. The unpredictable nature of real-world networks, marked by unforeseen disruptions and spikes in traffic, demands MAB algorithms that can swiftly adapt. A failure to do so can lead to significant downtimes or suboptimal network performance. The emergence of IoT and the increasing reliance on edge computing forecast a central role for MABs in decentralized network management. This would entail not just managing data flow but ensuring consistent connectivity in an increasingly complex web of devices. With these applications, it becomes evident that while MAB algorithms have significantly impacted diverse sectors, the journey is just beginning. Continuous research will likely see them morph, adapt, and become even more integral in the future.

4. Conclusion

This study provides a comprehensive review for the progress of MAB algorithms. The landscape of decision-making in uncertain environments has been revolutionized by MAB algorithms. These algorithms, with their adaptive nature and foundational principles grounded in balancing exploration and exploitation, have found applications spanning from online advertising to healthcare. As elucidated, while algorithms like UCB and Thompson Sampling have immensely shaped sectors like e-commerce and network routing, the journey of MABs is continually evolving. The achievements documented in this review, from personalized news recommendation systems to enhancing clinical trial outcomes, underscore the transformative power of MABs. However, it's also imperative to recognize the limitations. The dynamic nature of real-world scenarios often demands more swift adaptability than some MAB models can currently provide. Ethical implications, especially in sensitive sectors like healthcare, also need more contemplative consideration. As the horizon of MAB continues to expand, future research endeavors should prioritize refining these algorithms for real-time adaptability. Additionally, the integration of long-term reward signals, particularly in sectors like e-commerce, will be paramount. As technology and data science further intertwine, the prospect of MABs becoming even more integral and refined in decision-making systems seems not just probable, but inevitable. The quest for optimal decision-making under uncertainty marches on, and MABs will undeniably be at the forefront of this journey.

References

[1] Robbins H 1952 Some aspects of the sequential design of experiments Bulletin of the American Mathematical Society

[2] Thall P F & Wooten L H 2007 Bayesian Designs for Phase I–II Clinical Trials CRC Press

[3] Auer P Cesa-Bianchi N & Fischer P 2002 Finite-time Analysis of the Multiarmed Bandit Problem. Machine Learning

[4] Chapelle O & Li L 2011 An empirical evaluation of Thompson sampling Advances in Neural Information Processing Systems

[5] Scott S L 2010 A modern Bayesian look at the multi-armed bandit Applied Stochastic Models in Business and Industry

[6] Li L et al 2010 A contextual-bandit approach to personalized news article recommendation. Proceedings of the 19th international conference on World wide web

[7] Villar S S and Bowden J & Wason J 2015 Multi-armed bandit models for the optimal design of clinical trials: benefits and challenges Statistical Science 30(2) 199-215

[8] Dobson A Bekris K E 2015 Planning representations and algorithms for prehensile multi-arm manipulation 2015 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS) IEEE 6381-6386

[9] Deva A Abhishek K Gujar S A 2021 multi-arm bandit approach to subset selection under constraints arXiv preprint arXiv:2102.04824

[10] Neely M J 2010 Stochastic network optimization with application to communication and queueing systems Synthesis Lectures on Communication Networks 3(1) 1-211