


Analysis of Upper Confidence Boundary Algorithms for the Multi-Armed Bandit Problem

Yitong Song 

Department of Statistics and Actuarial Science, The University of Hong Kong, Hong Kong, China

Keywords: UCB, Machine Learning, Algorithm.

Abstract: The Multi-Armed Bandit (MAB) problem encapsulates the critical exploration and exploitation dilemma inherent in sequential decision-making processes under uncertainty. Central to this problem is the balance between gaining new knowledge (exploration) and leveraging existing knowledge to maximize immediate performance (exploitation). This paper delves into the MAB problem's core, where the Upper Confidence Bound (UCB) strategy emerges as a robust solution that does not necessitate an advanced knowledge of sub-optimality gaps. The methodological contribution is the systematic characterization and comparison of various UCB variants, including the classic UCB, Asymptotically Optimal UCB, KL-UCB, and MOSS. Each variant assigns a UCB index to arms in a bandit setup, by selecting the arm that has the largest index-value in every round, aiming to balance the exploration/exploitation trade-off dynamically. Notably, these algorithms are designed to operate without the abrupt transition from exploration to exploitation, fostering a more seamless and adaptive decision-making process. The paper's conclusion underscores the efficacy of UCB algorithms in optimizing long-term rewards in uncertain environments, highlighting their practical relevance in fields where machine learning algorithms must operate with minimal prior knowledge.

1 INTRODUCTION


In the realm of machine learning and decision making, MAB question serves as a foundational framework for exploring the challenges of exploration and exploitation. At its core, the MAB problem encapsulates a situation where the user must repeatedly choose among multiple options (or arms), each with a reward with unknown distribution, with the objective of maximizing cumulative gain over time. The quintessential dilemma in MAB lies in choosing whether to exploit the arm that has historically given the best rewards (exploitation) or to explore other less-known arms for potentially better rewards (exploration).

In 1985 Lai and Robbins finished the foundational study, establishing the theoretical framework for addressing the MAB problem using the concept of regret. Regret is defined as the difference in expectation of reward between a chosen strategy and the ideal strategy. Auer et al. (2002) introduced the idea of UCB algorithms, a series of strategies that

intelligently balance from exploration to exploitation trade-off by constructing confidence interval on the rewards of diffident choose of arm. These bounds are derived from the concentration inequalities and are used to make an optimistic estimation of the potential of each arm.

Building upon this foundation, Auer et al. (2002) further advanced the UCB methodology, resulting in the development of the UCB algorithm. It simplifies the computation of the upper confidence bounds, making it more practical for real-world applications. It operates by adding a bonus to the estimated rewards that increases with the uncertainty or the lack of knowledge on the correct rewards of different arms. This bonus term, which is influenced by both the number of play times of each arm and all arms, ensures that arm selection is proportional to their respective uncertainties.

The landscape of UCB algorithms has since expanded, with multiple variants being proposed, each tailored to different aspects of the MAB problem. These variants reflect the diverse thinking

 <https://orcid.org/0009-0001-4255-7566>

of different authors on how to best address the exploration-exploitation trade-off. Some, like the KL-UCB algorithm, use the Kullback-Leibler divergence to tighten the confidence bounds for distributions with known parametric forms (Cappé, Garivier, Maillard, Munos, & Stoltz, 2013). Others, such as UCB-V, incorporate variance estimates to adjust the exploration term dynamically, catering to environments with varying noise levels (Audibert, Munos, & Szepesvári, 2009).

The study of MAB and specifically UCB algorithms is of profound importance as it provides insights into optimal decision-making under uncertainty—a frequent occurrence in multiple fields including finance, healthcare, and online recommendation systems. By understanding and comparing different UCB approaches, strategies can be refined to suit specific situations and distributional assumptions, leading to more efficient learning and better performance.

This paper seeks to delve into the topic of UCB algorithms within the MAB framework. The method involves a comparative analysis of the various types of UCB algorithms, assessing their theoretical foundations, performance guarantees, and empirical results. The target of this study is to elucidate the nuanced differences among these algorithms, providing guidance for practitioners on selecting the most appropriate UCB variant for their specific use case. Through this exploration, the goal is to enhance the comprehensive understanding of strategic decision making in uncertain environments.

2 MAB PROBLEM

In general, number of actions has been given, denote as k ; at each time horizon $n = 1, 2, 3, \dots$, one of the action has been chosen. Once the action i is played, a reward has been gained simultaneously, with the support in $[1, 0]$ from a fixed but unknown distribution. After repeatedly choosing independent and identically distributed arms, a random reward will be obtained for each round, and the selection of each action is independent of others.

The method behind the MAB problem goes to the choice of arm at each horizon n . The choice of selection will base on the rewards for previous $n-1$ round. The unknown expected rewards of an action i is denoted as μ_i . In real-time application the objective is to make sure that the total rewards in horizon is the largest, $E[\sum_{i=1}^k \mu_i \cdot T_i(n)]$, where $i(n)$ is the arm that has been selected at round t and the algorithm is

chosen randomly. An equivalent express for the result is the expected total regret: the reward lost by taking sub-optimal decisions, which denotes the difference between the reward gains from the arm has the potentially largest reward and the actual reward received.

$$R(N) = n \cdot \mu^* - E[\sum_{i=1}^N \mu_i(n)] = \sum_i \Delta_i \cdot E[T_i(n)] \quad (1)$$

where μ^* denotes the largest mean reward in all actions, Δ_i denotes the sub-optimality, $T_i(n)$ denotes the number of times arm i has been selected in $n-1$ round.

2.1 Algorithms Employed

UCB is a better algorithm compare to the most basic method ETC. Comparing with ETC, UCB strategy does not require advanced information of the suboptimality gaps and tends to outperform ETC when there are more than two actions (Auer, Cesa-Bianchi, & Fischer, 2002). The UCB algorithm follows the rules of optimism, operating under the assumption that the environment is as favorable as is plausibly conceivable (Lattimore & Szepesvári, 2020). Consider the sequence of independent random variables $(X_t)_{t=1}^n$, which follows normal distribution with 1 as standard deviation.

$$P(\mu \geq \bar{\mu} + \sqrt{[2 \log(1/c)]/n}) \forall c \text{ belongs to } (0,1) \quad (2)$$

When evaluating the option in round t , the learner bases their decision on the $T_i(t-1)$ observed samples, which have an mean value of $\mu_i(t-1)$. Under these circumstances, a logical estimate for the covered μ of the next action would be

$$UCB_i(t-1, c) = \begin{cases} \text{infinity} & \text{if } T_i(t-1) = 0 \\ \mu_i(t-1) + \sqrt{[2 \log(1/c)]/n} & \text{otherwise} \end{cases} \quad (3)$$

When comparing (2) and (3), a great care should be taken. As for (2) the number of sample is a constant n , while for (3) it comes to the number of selection in $n-1$ round. In the formula, c serves as an approximate upper limit on the probability that the given quantity underestimates the actual mean value. Then it comes to the algorithm of UCB(c), which a input of number of actions and the error probability c is required.

The UCB algorithm, an index-based method (select largest value), where the index is represented by the summation of the $E(\mu_i)$ observed up to that point. The value within the argument of argmax corresponds to the index i of the arm.

For this UCB method on a stochastic k -armed bandit problem which follows a 1-subGaussian distribution. For any horizon n , if $c = 1/n^2$, then

$$R(N) \leq 3 \sum_{i=1}^k \Delta_i + \sum_{i: \Delta_i > 0} 16 \log(n) / \Delta_i \quad (4)$$

The regret of UCB algorithm follows a time complexity with $O(\log n)$, as for the part $3 \sum_{i=1}^k \Delta_i$ stays constant with the horizon n , and is negligible compared to the second term, which is very large, thus the time complexity primarily depends on the latter part. After analysis the researcher found that the worst-case regret of $R_n = O(\sqrt{kn \log(n)})$.

For the previous UCB algorithm that mentioned in this section, where $c=1/n^2$. This requires the knowledge of horizon n , which is not an anytime algorithm (Lattimore & Szepesvári, 2020). Also, the exploration bonus does not grow with t , i.e., there is no built-in mechanism to choose an arm that the number of selection stays constant for many rounds. The algorithm of Asymptotically optimal UCB is similar to the previous one, just using a new version of the UCB index.

The exploration bonus changes from $\sqrt{[2 \log(1/c)]/n}$ to $\sqrt{[2 \log(f(t))]/T_i(t-1)}$, where $f(t) = 1 + t \log^2(t)$. This modification will give a tighter upper bound for the user, sometimes user can change $f(t) = t$, but the performance is slightly worse. Before the modification, the exploration bonus remains the same for the arms that are not selected and the bonus goes down for the selected arm. After the changes, the UCB index is updated at every round for all arms. Exploration bonus increases for arms not selected, and decreases for the selected arm. Also, the latter form of bonus does not require a knowledge of n . So, it is an anytime algorithm.

The regret of the asymptotically optimal UCB follows:

$$\limsup_{n \rightarrow \infty} \frac{R_n}{\log n} \leq \sum_{i: \Delta_i > 0} \frac{2}{\Delta_i} \quad (5)$$

After simplification the regret is $O(\log n)$. The improvement of this method compared to the UCB is that the confidence interval is slightly smaller. The key insight is that users do not need to show that the $\mu_{is} \geq \mu_l$ for all s with high probability, it is sufficient to show that $\mu_{is} \geq \mu_l - c$ for some small c . Katherakis and Robbins (1995), Garivier et al. (2016), remarked on the unusual appearance of the function $f(t) = 1 + t \log^2(t)$. However, with a more complicated calculation user can choose $f(t) = t \log^a(t)$ for any a larger than 0. If the reward follows the normal distribution, then a more thorough analysis of concentration enables the selection of $f(t) = t$ or potentially a function with a slightly slower growth rate. Also, the asymptotic regret typically reflects finite-time performance, yet caution is advised.

Lower-order terms, which are hidden in asymptotic expressions, may dominate in practical applications.

In this part, a modification of UCB and basic ETC algorithm will be introduced. This method is called as Elimination algorithm. This represents a direct extension of the ETC algorithm to accommodate more arms, which also addresses the issue of selecting an appropriate commitment duration, involves the use of an elimination algorithm. This algorithm functions in distinct phases, each maintaining a group of potentially optimal arms known as the active set. During the ℓ -Th phase, the objective is to remove the arm i , which the inequality $\Delta_i \geq 2^{-\ell}$ holds.

2.2 MOSS and KL-UCB

Part 2.1 mentions that in worst-case the UCB regret follows $O(\sqrt{kn \log(n)})$ and in the elimination method the regret follows $O(\sqrt{kn \log(n)})$ (Lattimore & Szepesvári, 2020). It is feasible to entirely remove the logarithmic factor by modifying the confidence levels in the algorithm. The MOSS algorithm builds on the principles of UCB and was the first to implement this adjustment. A detailed presentation of the MOSS algorithm follows.

The MOSS algorithm was introduced as a variant of the UCB algorithm specifically designed to achieve minimax optimal regret in the problems with a number of arms and rewards. The key feature of MOSS is its adjustment of the exploration term in the confidence bound, which becomes more conservative as the number of selections of single arm increases. This adjustment allows MOSS to handle the trade-off more effectively in certain scenarios.

The performance of the MOSS algorithm is noteworthy. It has been proven to be asymptotically optimal, meaning that when the number of rounds increases, it achieves the same behaviors as the best possible strategy. This optimality holds for both finite and infinite action sets (Audibert & Bubeck, 2009). The MOSS algorithm strikes a balance between the trade-off of exploration and exploitation by incorporating an adaptive exploration parameter. This allows it to explore arms sufficiently while still exploiting the arms with the highest estimated rewards.

The regret of the MOSS follows a log function in terms of the time horizon and the number of arms.

$$R_n \leq 39\sqrt{kn} + \sum_{i=1}^k \Delta_i \quad (7)$$

This logarithmic regret bound ensures that the algorithm learns to make near-optimal decisions over time. The MOSS algorithm achieves this by carefully balancing exploration and exploitation, resulting in a

regret that grows logarithmically with the number of rounds.

Furthermore, the K-armed bandit problem serves as a well-known model for studying decision-making in uncertain conditions, involving a player's choice among K different options, each with unknown payout probabilities. The objective is to optimize the total payout accumulated over time. A principal strategy for addressing this challenge is the KL-UCB algorithm, an advanced version of the broader UCB methodology. The KL-UCB algorithm is an advanced method for the MAB problem (stochastic) designed to minimize regret, which is the loss in potential reward due to not picking the best arm at each trial. It achieves this by balancing exploration (trying out less chosen arms to discover their potential) and exploitation (picking the arm that has historically given the best rewards). The KL-UCB algorithm, is particularly well-suited for distributions that can be parameterized by a single parameter, like the Bernoulli or Poisson distributions (Cappé et al. 2013).

The algorithm follows the same rule with UCB algorithm. In the KL-UCB algorithm, the index calculation involves resolving an optimization issue that applies the KL divergence to quantify the discrepancy between the empirical mean of rewards and the possible true mean (Cappé et al., 2011). The KL divergence assesses the deviation of one probability distribution from another, expected probability distribution.

Garivier & Cappé (2011) points out that when there is K arms and rewards is bounded between 0 and 1 independently, and a^* denotes for the optimal arm. Then the regret is:

$$\limsup_{n \rightarrow \infty} \frac{E[R_n]}{\log(n)} \leq \sum_{a: \mu_a < \mu_{a^*}} \frac{\mu_{a^*} - \mu_a}{d(\mu_{a^*}, \mu_a)} \quad (6)$$

However, in 2002 Auer et al., uses the empirical mean plus a term that encourages exploration proportional to the square root of the logarithm of the total number of plays divided by the number of times that particular arm has been played, KL-UCB replaces the exploration term with a KL divergence-based confidence bound. This change ensures a tighter confidence interval and thus a more informed selection strategy, particularly for reward distributions that are not sub-Gaussian (Garivier & Cappé, 2011). Garivier & Cappé (2011) also did more comparison in more difficult situation, with the Bernoulli rewards. In this experiment, the distinction between KL-UCB and UCB was marked, and the

performance of UCB-T, which is another UCB method, was considerably less notable.

2.3 Comparisons

In this section the performances of UCB, Asymptotically Optimal UCB (AO-UCB), MOSS and KL-UCB is going to be compared.

Auer's UCB method (2002) is designed to balance exploration and exploitation by using an upper confidence bound to select actions. The basic UCB algorithm adds a confidence interval to the estimated rewards, which depends on the number of times an arm has been pulled. The term ensures that arms not recently chosen are revisited, thus exploring potentially underestimated options.

Asymptotically Optimal UCB (AO-UCB), on the other hand, refines the confidence bounds to minimize the regret asymptotically. According to Lattimore and Szepesvári (2020), AO-UCB adjusts the exploration term to be more sensitive to the variance in arm rewards, which theoretically reduces the cumulative regret more efficiently than standard UCB in the long run. Empirical studies, such as those by Cowan and Katehakis (2015), have shown that AO-UCB outperforms UCB in environments with high variance in rewards, primarily due to its more nuanced exploration mechanism.

MOSS, introduced by Audibert and Bubeck (2009), aims to minimize the worst-case regret across all sub-optimal arms. Unlike AO-UCB, which adapts based on the variance of rewards, MOSS sets a uniform exploration term that decreases only with the number of times an arm is played, independent of the total number of pulls. This approach can lead to better performance in situations with many arms or non-stationary reward distributions. In comparing AO-UCB and MOSS, Bubeck and Slivkins (2012) found that MOSS tends to perform better in scenarios with many arms, as it does not over-penalize less frequently chosen arms, unlike AO-UCB. However, in settings with fewer arms and clear distinctions in arm quality, AO-UCB's variance-sensitive exploration can achieve lower regret.

Cappé et al. (2013) Proposed KL-UCB that uses the KL divergence to tailor the exploration term more closely to the true distribution of rewards. This approach is particularly beneficial in environments where the reward distributions are known to be non-Gaussian, as it can more accurately estimate the upper confidence bounds. When comparing MOSS and KL-UCB, in 2012 Garivier and Cappé noted that KL-UCB often achieves significantly lower regret in practice, especially in problems with skewed or

bounded reward distributions, like Bernoulli or exponential rewards. The tailored exploration term in KL-UCB allows for more efficient exploration by focusing more precisely on the statistical properties of each arm's reward distribution. However, KL-UCB is not without its limitations. Calculating the KL divergence can be computationally more intensive than the simpler calculation required for UCB. This can make KL-UCB less appealing for problems where computational resources are constrained or when very fast decision-making is required. Additionally, KL-UCB's performance guarantee is mainly for single-parameter distributions; for more complex distribution families, its optimality isn't always guaranteed (Maillard, 2018).

In this paper there are four popular multi-armed bandit (MAB) algorithms have been explored: UCB, Asymptotically Optimal UCB, MOSS, and KL-UCB. Each algorithm aims to achieve a balance arise from explore and exploit, addressing the challenges posed by MAB problems. Firstly, UCB algorithm offers a simple and effective approach, providing an effective balance between exploration and exploitation. It achieves sublinear regret and follows the function of $O(k \log t)$, where k means the number of arms and t denotes the number of time steps. Asymptotically Optimal UCB, on the other hand, comes with a more sophisticated exploration strategy. It achieves an even lower regret rate than UCB, specifically a logarithmic regret. However, it comes at the cost of increased time complexity, $O(k \log^2 t)$. Moving on to MOSS, this algorithm introduces a different exploration mechanism by focusing on the arms that have shown promising rewards in the past. It achieves sublinear regret, similar to UCB, but with a slightly higher time complexity of $O(k^2 \log T)$. Lastly, KL-UCB algorithm leverages the Kullback-Leibler divergence to balance exploration and exploitation. It achieves logarithmic regret also obeys the performance of $O(k \log t)$. Although it requires more computations compared to UCB, it can lead to improved performance in certain scenarios. Determining which algorithm is better depends on the specific problem and its requirements. Asymptotically Optimal UCB is preferable in settings with significant reward variance, MOSS excels in environments with a large number of arms, and KL-UCB is ideal for handling non-Gaussian reward distributions. The choice of algorithm should thus be guided by the nature of the reward structure and the specific goals of the exploration-exploitation trade-off.

There are several potential future extensions to explore. Firstly, this algorithm can expand into more diverse field, UCB algorithms have already made

significant impacts in areas such as recommendation advertisement systems, clinical medicine trials, and financial management. Future research could expand these applications into more complex and dynamic environments. For example, in the field of personalized medicine, UCB algorithms could be employed to adaptively select among treatment options for patients based on real-time responses. Similarly, in automated trading systems, these algorithms could dynamically adjust trading strategies to maximize financial returns under volatile market conditions. Furthermore, integrating this algorithm with Emerging Technologies can improve a lot, the integration of UCB algorithms with emerging technologies such as artificial intelligence (AI) and machine learning could open more spaces for smarter, more efficient decision-making systems. For instance, incorporating UCB algorithms into AI-driven IoT (Internet of Things) devices could enhance decision-making processes in smart homes and smart cities by learning and adapting to the preferences and behaviors of users. Thirdly, UCB algorithm can gain Enhancement through Advanced Computational Techniques, the development of more sophisticated computational techniques can further enhance the performance of UCB algorithms. Techniques such as deep learning could be used to approximate the reward distributions more accurately, especially in complex scenarios where traditional statistical methods fall short. This could lead to more refined and effective exploration-exploitation balances in UCB implementations. Also, people should focus on the Ethical Considerations and Bias Mitigation of UCB, as UCB algorithms and their applications grow, it becomes crucial to consider the ethical implications of automated decision-making systems, particularly in terms of fairness and bias. In the future researchers should also focus on developing mechanisms within these algorithms to detect and mitigate biases, ensuring that decisions made by automated systems do not inadvertently disadvantage any group or individual.

3 CONCLUSION

In conclusion, each of the four MAB algorithms that have been discussed has its pros and cons. The choice of algorithm is decided on the specific situation and trade-offs between performance and computational complexity. By considering future extensions and adapting these algorithms to different scenarios, people can continue advancing the field of multi-

armed bandit problems and finding even more effective solutions.

REFERENCES

- Agrawal, R. (1995). Sample mean based index policies with $O(\log n)$ regret for the multi-armed bandit problem. *Advances in Applied Probability*, 27(4), 1054-1078.
- Audibert, J.-Y., & Bubeck, S. (2009). *Minimax policies for adversarial and stochastic bandits*. 22nd Annual Conference on Learning Theory (COLT 2009).
- Audibert, J.-Y., Munos, R., & Szepesvári, C. (2009). *Exploration - exploitation trade-off using variance estimates in multi-armed bandits*. *Theoretical Computer Science*, 410(19), 1876-1902.
- Auer, P., Cesa-Bianchi, N., & Fischer, P. (2002). *Finite-time analysis of the multi-armed bandit problem*. *Machine Learning*, 47(2-3), 235-256.
- Bubeck, S., & Slivkins, A. (2012). The Best of Both Worlds: Stochastic and Adversarial Bandits. *Proceedings of the 25th Annual Conference on Learning Theory*.
- Burnetas, A. N., & Katehakis, M. N. (1996). *Optimal adaptive policies for sequential allocation problems*. *Advances in Applied Mathematics*, 17(2), 122-142.
- Cappé, O., Garivier, A., Maillard, O. A., Munos, R., & Stoltz, G. (2013). *Kullback-Leibler upper confidence bounds for optimal sequential allocation*. *Annals of Statistics*, 41(3), 1516-1541.
- Chapelle, O., & Li, L. (2011). *An empirical evaluation of Thompson Sampling*. In *NIPS 2011*.
- Clopper, C. J., & Pearson, E. S. (1934). The use of confidence of fiducial limits illustration in the case of the binomial. *Biometrika*, 26, 404-413.
- Cowan, W., & Katehakis, M. N. (2015). *Asymptotically Optimal Multi-Armed Bandit Algorithms*. *Mathematics of Operations Research*, 40(3), 576-602.
- Cowan, W., Honda, J., & Katehakis, M. N. (2017). *Normal bandits of unknown means and variances*. *Journal of Machine Learning Research*, 18(1), 5638-5665.
- Filippi, S., Cappé, O., & Garivier, A. (2010). *Optimism in reinforcement learning and Kullback-Leibler divergence*. *Allerton Conference on Communication, Control, and Computing*, Monticello, US.
- Garivier, A., & Cappé, O. (2011). *The KL-UCB algorithm for bounded stochastic bandits and beyond*. In *Conference on Learning Theory (COLT 2011)*.
- Garivier, A., Ménard, P., & Stoltz, G. (2016). *Explore first, exploit next: The true shape of regret in bandit problems*. *Mathematics of Operations Research*, 41(4), 1436-1454.
- Honda, J., & Takemura, A. (2010). An asymptotically optimal bandit algorithm for bounded support models. *Proceedings of COLT 2010*, 67-79.
- Katehakis, M. N., & Robbins, H. (1995). *Sequential choice from several populations*. *Proceedings of the National Academy of Sciences*, 92(19), 8584-8585.
- Lai, T. L., & Robbins, H. (1985). *Asymptotically efficient adaptive allocation rules*. *Advances in Applied Mathematics*, 6, 4-22.
- Lattimore, T. (2016). *Regret analysis of the finite-horizon Gittins index strategy for multi-armed bandits*. *Probability in the Engineering and Informational Sciences*, 30(4), 530-553.
- Lattimore, T., & Szepesvári, C. (2020). *Bandit Algorithms*. Cambridge University Press.
- Maillard, O. (2018). *Boundary crossing probabilities for general exponential families*. *Mathematical Methods of Statistics*, 27(1), 1-31.