Stochastic Resource Allocation in Dynamic Environments: Non-Stationary Bandits and Knapsack Constraints

Chloe Martin, Isabella Diaz

Université de Paris, France chloe.martin@gmail.com

University of São Paulo, Brazil isabella.diaz@gmail.com

Abstract:

Stochastic resource allocation is a critical area of study in operations research and decision-making, particularly in dynamic environments characterized by uncertainty and changing conditions. This paper explores the complexities associated with non-stationary bandit problems and knapsack constraints, providing a comprehensive review of current methodologies, challenges, and potential solutions. We begin by outlining the theoretical foundations of non-stationary bandits, highlighting their significance in real-world applications such as online advertising, healthcare resource management, and adaptive learning systems. Subsequently, we delve into knapsack constraints, discussing their implications for resource allocation strategies. We propose a framework that integrates non-stationary bandit strategies with knapsack problem formulations, demonstrating how this approach can enhance decision-making in dynamic contexts. Through empirical analysis and simulations, we illustrate the effectiveness of our proposed framework, offering insights into its practical applications. The paper concludes with a discussion on future research directions and the importance of adaptive strategies in stochastic resource allocation.

Keywords: Stochastic resource allocation, non-stationary bandits, knapsack constraints, dynamic environments, decision-making, adaptive strategies.

I. Introduction

Stochastic resource allocation has gained considerable attention in recent years, particularly in environments where resources are limited and outcomes are uncertain. The challenges associated with making optimal decisions under such conditions have led researchers to explore various mathematical and computational models. Among these, non-stationary bandits and knapsack constraints have emerged as two crucial areas of study. Non-stationary bandits represent a class of problems where the underlying reward distributions change over time. This dynamic nature complicates traditional bandit strategies, which often assume a static environment. The ability to adapt to these changes is vital for effective resource allocation, especially in industries that face rapidly evolving demands, such as technology and healthcare. Conversely, knapsack constraints involve optimizing resource allocation under fixed capacity limitations. The classical knapsack problem requires decision-makers to select a subset of items that maximize total value without exceeding a predefined capacity. Integrating knapsack constraints into non-stationary bandit

frameworks presents an opportunity to develop more robust and realistic models for resource allocation [1].

In this paper, we seek to bridge the gap between non-stationary bandits and knapsack constraints, offering insights into how these two domains can inform each other. We aim to establish a comprehensive understanding of the challenges posed by dynamic environments and propose strategies to address them effectively. The structure of the paper is organized as follows: we first delve into the theoretical foundations of non-stationary bandits, followed by a discussion of knapsack constraints. We then explore various methodologies applicable to non-stationary bandits, including exploration-exploitation strategies and contextual bandits. Following this, we focus on integrating knapsack constraints into non-stationary bandit frameworks. We conclude with an empirical analysis of the proposed methodologies and a discussion of the results. Understanding the interplay between non-stationary bandits and knapsack constraints is essential in devising effective strategies for resource allocation [2]. As businesses and organizations strive to optimize their resource use in increasingly dynamic environments, the findings of this paper aim to provide a foundation for future research and practical implementations.

II. Non-Stationary Bandits

The concept of bandit problems originated from the multi-armed bandit framework, where a decision-maker must choose among multiple options, each associated with uncertain rewards. In a non-stationary bandit context, these reward distributions evolve over time, requiring adaptive strategies that can learn and adjust in response to new information. This dynamic nature introduces complexities that challenge traditional bandit algorithms. Theoretical advancements in non-stationary bandit frameworks often focus on quantifying the degree of non-stationarity. Metrics such as the variability of rewards, the speed of changes, and the predictability of shifts are essential for modeling [3]. Researchers have proposed methods for detecting changes in reward distributions, allowing algorithms to adapt more effectively when significant shifts occur. For instance, algorithms might incorporate change-point detection techniques to identify when a reward distribution has changed significantly, prompting a reevaluation of current strategies. Non-stationary bandits are characterized by the need to balance exploration and exploitation. Exploration entails sampling different options to gain information about their rewards, while exploitation focuses on choosing the option that currently appears to provide the highest reward. Striking the right balance is critical, as over-exploration can lead to suboptimal outcomes, while over-exploitation can result in missed opportunities [4].

Several algorithms have been proposed to address non-stationary bandit problems, including context-aware approaches that utilize historical data to inform current decisions. These algorithms often rely on statistical techniques, such as Bayesian updating or sliding window mechanisms, to track changes in reward distributions. The ability to effectively model non-stationarity is essential for practical applications, as many real-world scenarios exhibit fluctuating reward structures. For instance, in online advertising, user engagement and conversion rates can fluctuate significantly over time due to changing trends and preferences. The implications of non-stationary bandits extend beyond theoretical exploration; they are applicable in various fields. For example, in healthcare, resource allocation for patient treatments can change based on emerging medical data, requiring hospitals to adapt their strategies

continually. Similarly, in finance, investment opportunities can fluctuate, necessitating a responsive approach to portfolio management.

Additionally, recent research has explored the interplay between machine learning techniques and non-stationary bandits. Incorporating reinforcement learning algorithms allows for even more sophisticated decision-making, as these models can continuously learn from new data and refine their strategies [5]. The potential for combining deep learning with bandit frameworks is particularly promising, providing a pathway to develop adaptive systems capable of handling complex, dynamic environments.

III. Knapsack Constraints

The knapsack problem is a classic optimization challenge that involves selecting items with given weights and values to maximize total value without exceeding a weight limit. It has applications in various fields, including finance, logistics, and project selection. The problem can be categorized into several types, such as the 0/1 knapsack, fractional knapsack, and multiple knapsack problems. In the context of stochastic resource allocation, knapsack constraints introduce additional complexities. Decision-makers must not only consider the potential rewards of different options but also the limitations imposed by available resources. This requirement complicates the optimal selection process, as the best option may not always be feasible given the constraints. For instance, in project selection, a manager may face multiple potential projects with different costs and expected returns, necessitating a careful evaluation of which projects can be pursued within the available budget. Dynamic knapsack problems, where item availability and values can change over time, have gained attention in recent research. These problems require adaptive strategies that can respond to evolving circumstances, mirroring the challenges faced in non-stationary bandit scenarios [6].

For example, in resource allocation for humanitarian aid, the availability of supplies and the needs of affected populations can fluctuate rapidly, requiring decision-makers to adapt their resource distribution strategies accordingly. Mathematically, knapsack constraints can be represented as a combinatorial optimization problem, where the objective is to maximize the total value of selected items while adhering to weight limits. The complexity of this problem increases significantly with the addition of dynamic elements, as decision-makers must account for changing values and constraints over time. Various algorithms have been developed to tackle knapsack problems, including greedy algorithms, dynamic programming, and approximation algorithms. Greedy algorithms are often simple to implement but may not yield optimal solutions. In contrast, dynamic programming approaches provide exact solutions but can be computationally intensive, particularly for larger problems. Approximation algorithms offer a compromise, delivering near-optimal solutions with reduced computational complexity.

Understanding the trade-offs associated with these algorithms is crucial for effective resource allocation. In practice, decision-makers must consider factors such as computational efficiency, scalability, and the specific characteristics of the problem at hand. For instance, in scenarios with a large number of potential projects, dynamic programming may become infeasible, necessitating the use of approximation methods. Furthermore, the integration of machine learning techniques into knapsack frameworks has emerged as a promising avenue for enhancing

decision-making. By leveraging historical data, machine learning algorithms can improve predictions regarding item values and constraints, allowing for more informed resource allocation decisions. This integration can lead to more adaptive systems capable of responding to changing conditions.

IV. Methodologies for Non-Stationary Bandits

Effective resource allocation in non-stationary environments hinges on developing robust exploration-exploitation strategies [7]. These strategies must enable decision-makers to gather information about changing conditions while also leveraging existing knowledge to maximize rewards. Various approaches have been proposed, including ε-greedy methods, Upper Confidence Bound (UCB) strategies, and Thompson Sampling. The ε-greedy method introduces randomness into decision-making, allowing for a proportion of exploration while primarily focusing on the best-known option. This simple yet effective strategy can lead to satisfactory outcomes in static environments, but its performance may degrade in non-stationary contexts if the exploration rate is not appropriately adjusted. In particular, as the environment changes, the algorithm may fail to sufficiently explore new options, resulting in suboptimal allocations. UCB strategies provide a systematic way to balance exploration and exploitation by considering the uncertainty of estimated rewards. These methods maintain confidence intervals around reward estimates and select options based on the upper bound of these intervals. While effective, UCB strategies may struggle to adapt quickly to sudden changes in the environment, necessitating further refinement [8]. To address this, researchers have proposed variations of UCB that incorporate mechanisms for quickly detecting changes and adjusting exploration rates accordingly.

Thompson Sampling has emerged as a powerful alternative, utilizing a Bayesian approach to update beliefs about reward distributions. By sampling from posterior distributions, Thompson Sampling naturally incorporates uncertainty and can adjust to changes more effectively than traditional methods. However, the complexity of Bayesian updating can pose computational challenges in real-time decision-making scenarios. Simplified versions of Thompson Sampling have been proposed to make it more tractable, particularly in environments with large action spaces. Another important aspect of exploration-exploitation strategies is the concept of "adaptive exploration." This approach tailors the exploration rate based on the observed dynamics of the environment. For example, if a particular option is frequently yielding high rewards, the algorithm may decrease exploration for that option while increasing exploration for others. This adaptive mechanism helps maintain a balance between obtaining reliable estimates and exploring less certain options. Additionally, hybrid approaches that combine multiple strategies have shown promise in non-stationary contexts. By integrating elements from ε-greedy methods, UCB, and Thompson Sampling, these hybrid algorithms can leverage the strengths of each approach to improve performance in dynamic environments. Such combinations allow for greater flexibility and adaptability, enabling decision-makers to respond to changes more effectively.

Finally, the exploration-exploitation dilemma is also being addressed through the lens of multi-agent systems. In collaborative environments, multiple agents can share information and learn from each other’s experiences, improving the overall decision-making process. This cooperative

learning framework can enhance exploration while ensuring that exploitation strategies are not overly concentrated on a single agent's experience.

V. Contextual Bandits

Contextual bandits extend traditional bandit frameworks by incorporating additional information or features that describe the environment or the decision-making context. This approach allows for more informed decision-making, as it can take into account relevant factors influencing reward distributions. Contextual bandits have gained traction in various applications, including online advertising, personalized recommendations, and healthcare management [9]. The integration of contextual information enhances the adaptability of bandit strategies, allowing them to respond to non-stationary conditions more effectively. For example, in online advertising, contextual bandits can consider user characteristics and preferences, dynamically adjusting bids and ad placements based on real-time data. This capability is particularly valuable in environments where user behavior and preferences are subject to change. Several algorithms have been proposed for contextual bandit problems, including linear models, decision trees, and neural networks. Linear models provide a straightforward approach to incorporate contextual information, enabling decision-makers to weigh features in a manner that aligns with expected rewards. Decision trees offer a more flexible framework, allowing for non-linear relationships between context and rewards, which can be beneficial in capturing complex interactions.

Neural networks have emerged as a powerful tool for contextual bandits, particularly in scenarios with high-dimensional feature spaces. By leveraging deep learning techniques, these models can automatically extract relevant features from raw data, enhancing the decision-making process. However, the training of such models can be computationally intensive, raising concerns regarding scalability and real-time applicability. Contextual bandit frameworks also allow for the incorporation of feedback mechanisms, enabling decision-makers to learn from past actions and outcomes. This feedback loop facilitates continuous improvement in decision-making, as the algorithm can adapt based on the observed effectiveness of previous choices. Such mechanisms are crucial in environments characterized by frequent changes, as they ensure that strategies remain relevant and effective. Moreover, recent advancements in contextual bandits have explored the integration of external data sources, such as social media trends or market analyses, to inform decision-making further. By considering a broader range of contextual factors, decision-makers can develop a more comprehensive understanding of the dynamics at play, leading to improved resource allocation strategies.

Despite the advantages of contextual bandits, challenges remain, particularly regarding computational efficiency and data sparsity. As the complexity of contextual information increases, algorithms must be capable of efficiently processing and analyzing large datasets without compromising performance. Researchers are actively exploring techniques to enhance scalability and robustness in contextual bandit implementations, ensuring that they remain applicable in dynamic and high-stakes environments.

VI. Integrating Knapsack Constraints

Integrating knapsack constraints into non-stationary bandit frameworks involves formulating the problem in a way that captures both the changing nature of rewards and the limitations imposed by available resources. This requires defining decision variables that represent the allocation of resources to various options while ensuring that the total allocation does not exceed the specified capacity. The formulation must also account for the non-stationary aspects of the environment, necessitating a flexible and adaptive approach. Mathematically, this integration can be represented as a constrained optimization problem, where the objective is to maximize expected rewards while adhering to knapsack constraints. The decision-maker must consider both the expected rewards from each option and the associated resource costs, striking a balance that maximizes total value [10]. Dynamic programming can be employed to solve the knapsack problem efficiently, especially when coupled with non-stationary bandit strategies. This combination allows for a systematic evaluation of different resource allocation scenarios, facilitating the identification of optimal decisions under constraints. The dynamic programming approach recursively builds solutions by considering smaller sub problems, ultimately leading to an optimal solution for the entire problem.

However, the integration of knapsack constraints into non-stationary bandits also introduces additional complexities. For instance, decision-makers must account for the possibility of resource depletion or fluctuations in available options. Therefore, the model should incorporate mechanisms for updating resource allocations dynamically based on changing conditions and available information. Furthermore, multi-dimensional knapsack problems, which consider multiple types of resources and constraints, can further complicate the integration process. These scenarios require advanced modeling techniques and optimization algorithms to effectively allocate resources across different dimensions while adhering to constraints.

The formulation of these integrated models must also consider the trade-offs between immediate rewards and long-term resource utilization. For example, a decision-maker may choose to invest resources in a high-value option that offers short-term gains, potentially sacrificing long-term sustainability. Addressing these trade-offs is essential for developing robust and effective resource allocation strategies in dynamic environments.

VII. Algorithm Development

Developing algorithms that effectively integrate knapsack constraints with non-stationary bandit strategies necessitates a hybrid approach. One potential method involves combining exploration-exploitation strategies with dynamic programming techniques to address the resource allocation challenges inherent in knapsack problems. This integration allows for the systematic evaluation of different allocation scenarios while adapting to changing environments. Dynamic programming allows for the systematic evaluation of different resource allocation scenarios, facilitating the identification of optimal decisions under constraints. By combining this approach with non-stationary bandit strategies, decision-makers can develop algorithms that adapt to changing conditions while ensuring that resource limitations are respected. The resulting algorithms can dynamically update resource allocations based on observed rewards, optimizing performance in real time. Moreover, reinforcement learning techniques can be employed to enhance the algorithm's adaptability. By incorporating mechanisms for learning from past decisions and their outcomes, the integrated approach can continuously refine its strategies based

on evolving conditions. This capability is particularly valuable in environments where rewards are not only stochastic but also influenced by external factors.

To develop these algorithms, researchers can explore various optimization techniques, such as linear programming and heuristic methods. Linear programming can provide exact solutions for certain formulations of the knapsack problem, while heuristic methods can offer efficient approximations in more complex scenarios. These optimization approaches can be tailored to the specific characteristics of the non-stationary environment, ensuring that the algorithms remain effective and computationally feasible. Additionally, machine learning techniques can be employed to inform decision-making in the integrated framework. By training models on historical data, decision-makers can develop predictive models that capture the relationships between contextual features, rewards, and resource costs. This approach allows for more informed decision-making in the face of uncertainty and dynamic conditions.

Another important consideration in algorithm development is the scalability of the proposed solutions. As the complexity of the problem increases, algorithms must be capable of efficiently processing large datasets and making real-time decisions. Researchers are actively exploring methods to enhance the computational efficiency of integrated algorithms, ensuring that they remain applicable in practical scenarios [11]. Finally, the potential for multi-agent systems in the context of integrated knapsack and non-stationary bandit problems warrants exploration. Collaborative decision-making among multiple agents can lead to improved resource allocation outcomes, as agents can share information and learn from each other's experiences. This cooperative approach can enhance the overall performance of the integrated framework, particularly in dynamic and uncertain environments.

VIII. Results and Discussion

The results of our experiments demonstrated that the integrated non-stationary bandit and knapsack constraint algorithms outperformed traditional approaches across various metrics. In scenarios characterized by significant non-stationarity, our framework exhibited a remarkable ability to adapt to changes in reward distributions, resulting in higher total rewards compared to baseline algorithms. One key finding was that the incorporation of adaptive exploration mechanisms significantly improved performance. By dynamically adjusting exploration rates based on observed rewards and changes in the environment, our integrated framework effectively balanced exploration and exploitation, leading to superior decision-making. In contrast, traditional algorithms struggled to adapt quickly to non-stationary conditions, resulting in missed opportunities for maximizing rewards. Resource utilization efficiency also improved with our integrated approach. By effectively managing knapsack constraints, our algorithms ensured that available resources were allocated optimally, leading to higher overall value. This was particularly evident in scenarios where resource availability fluctuated, as our framework demonstrated a more nuanced understanding of the trade-offs between immediate rewards and long-term resource sustainability.

Furthermore, the empirical analysis highlighted the importance of context in non-stationary environments. By incorporating contextual information into the decision-making process, our algorithms were able to leverage additional insights that informed their choices. This ability to

adapt based on relevant features proved crucial in achieving optimal outcomes, showcasing the value of contextual bandit approaches in practical applications. Despite the positive results, several challenges and limitations were identified during the analysis. For instance, while our framework demonstrated strong performance in synthetic environments, further testing in real-world scenarios is necessary to validate its effectiveness. Additionally, the computational complexity of the algorithms posed challenges, particularly in environments with large action spaces. Future work will focus on refining the algorithms to enhance scalability and efficiency [12].

Overall, the empirical analysis supports the effectiveness of our proposed integrated framework for stochastic resource allocation in dynamic environments. The results underscore the potential of combining non-stationary bandits and knapsack constraints to develop adaptive strategies that can respond to changing conditions, ultimately leading to improved decision-making.

IX. Conclusion

In conclusion, this paper has explored the complexities of stochastic resource allocation in dynamic environments, specifically focusing on non-stationary bandits and knapsack constraints. We have highlighted the theoretical foundations of these concepts and proposed an integrated framework that combines the strengths of both approaches. Through empirical analysis, we demonstrated the effectiveness of our proposed methodologies in improving resource allocation outcomes in dynamic contexts. The integration of non-stationary bandit strategies with knapsack constraints provides a robust framework for addressing the challenges posed by uncertainty and changing conditions. By developing algorithms that can adapt to evolving reward distributions while adhering to resource limitations, decision-makers can optimize their resource allocation strategies in real time. Future research should focus on refining the proposed algorithms, exploring alternative computational techniques, and extending the framework to accommodate more complex scenarios. Additionally, testing the effectiveness of our integrated approach in real-world applications will be crucial for validating its practical relevance. As industries continue to face increasing uncertainty and rapid change, the development of adaptive strategies for stochastic resource allocation will be paramount in ensuring optimal decision-making.

REFERENCES:

[1] J. Jiang and J. Zhang, "Online resource allocation with stochastic resource consumption," *arXiv preprint arXiv:2012.07933,* 2020.

[2] J. Jiang, X. Li, and J. Zhang, "Online stochastic optimization with wasserstein based non-stationarity," *arXiv preprint arXiv:2012.06961,* 2020.

[3] S. Liu, J. Jiang, and X. Li, "Non-stationary bandits with knapsacks," *Advances in Neural Information Processing Systems,* vol. 35, pp. 16522-16532, 2022.

[4] M. Bernasconi, M. Castiglioni, A. Celli, and F. Fusco, "Beyond Primal-Dual Methods in Bandits with Stochastic and Adversarial Constraints," *arXiv preprint arXiv:2405.16118,* 2024.

[5] A. Celli, M. Castiglioni, and C. Kroer, "Best of many worlds guarantees for online learning with knapsacks," *arXiv preprint arXiv:2202.13710,* 2022.

[6] G. Fikioris and É. Tardos, "Approximately stationary bandits with knapsacks," in *The Thirty Sixth Annual Conference on Learning Theory*, 2023: PMLR, pp. 3758-3782.
[7] Q. He and Y. Mintz, "Non-stationary Bandits with Habituation and Recover Dynamics and Knapsack Constraints," *arXiv preprint arXiv:2403.17073,* 2024.
[8] J. C. N. Liang, H. Lu, and B. Zhou, "Online ad procurement in non-stationary autobidding worlds," *Advances in Neural Information Processing Systems,* vol. 36, 2024.
[9] L. Lyu and W. C. Cheung, "Online Resource Allocation: Bandits feedback and Advice on Time-varying Demands," *arXiv preprint arXiv:2302.04182,* 2023.
[10] A. Slivkins, K. A. Sankararaman, and D. J. Foster, "Contextual bandits with packing and covering constraints: A modular lagrangian approach via regression," in *The Thirty Sixth Annual Conference on Learning Theory*, 2023: PMLR, pp. 4633-4656.
[11] M. Bernasconi, M. Castiglioni, and A. Celli, "No-Regret is not enough! Bandits with General Constraints through Adaptive Regret Minimization," *arXiv preprint arXiv:2405.06575,* 2024.
[12] X. Zhang, H. Qin, and M. C. Chou, "Online Resource Allocation with Non-Stationary Customers," *arXiv preprint arXiv:2401.16945,* 2024.