Adaptive Decision-Making in Non-Stationary Markov Decision Processes

Baiting Luo
Vanderbilt University
Nashville, USA
baiting.luo@Vanderbilt.Edu

ABSTRACT

This research addresses a critical and largely unresolved challenge in the field of sequential decision-making: operating effectively in non-stationary environments. These environments are characterized by exogenously-driven changes over time, introducing significant uncertainties in decision-making processes. The urgency lies in devising strategies for optimal decision-making and planning amidst these unpredictable conditions. Central to my research is the concept of 'anytime' decision-making. This approach involves leveraging dynamically learned models that not only mirror the current environmental state but also anticipate its potential evolution. The focus is on how an agent adapts its decision-making process in an ever-changing environment. A key contribution of my work is the exploration of adaptive decision-making strategies employed by an agent whose objectives fluctuate between performance optimization and safety prioritization. This is particularly challenging in dynamic environments where traditional static decision-making models fall short. The paper concludes by presenting future research directions. These aims are to enhance the understanding of adaptive decision-making in non-stationary environments, thereby advancing the field in this complex and constantly evolving area.

KEYWORDS

Sequential Decision-Making; Non-Stationary Environments; Online Planning; Monte Carlo Tree Search

ACM Reference Format:
Baiting Luo. 2024. Adaptive Decision-Making in Non-Stationary Markov Decision Processes. In *Proc. of the 23rd International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2024), Auckland, New Zealand, May 6 – 10, 2024*, IFAAMAS, 3 pages.

1 INTRODUCTION

In the realm of artificial intelligence, incorporating planning and decision-making skills into intelligent agents is a crucial area of research. This endeavor is especially significant in applications such as vehicle routing, emergency management, and autonomous driving [5, 6, 11, 17]. However, a key challenge arises from the limitations in these agents' capabilities, often constrained by their assumptions about the environment's stationarity. This is particularly problematic in real-world scenarios, which are inherently non-stationary and often violate the basic assumption of stationarity that underpins many reinforcement learning algorithms [16].

For these intelligent agents to achieve true autonomy and reliability, they must be capable of not only perceiving and adapting to changes in environmental dynamics but also proactively planning and making decisions in response to these changes. Recent advancements have enabled agents to reactively adapt to new tasks [3, 9] and proactively seek effective policies for anticipated future environmental dynamics [2]. Nevertheless, the pursuit of robust and efficient decision-making that maintains performance objectives during environmental transitions remains a significant and unresolved challenge in non-stationary settings.

One of the potential directions to address this challenge is the concept of adaptive decision-making. This approach necessitates not just a single policy, but an array of diverse policies, each characterized by its unique attributes and designed for specific environmental contingencies. By leveraging this multifaceted policy framework, intelligent agents can navigate the complexities of non-stationary settings more effectively. They can switch between different strategies based on real-time environmental dynamics, thereby maintaining optimal performance despite the ever-evolving nature of their operational contexts. This paradigm shift towards adaptive decision-making, which prioritizes flexibility and responsivenes.

2 OUR WORK TO DATE

Toward reaching our goal of making adaptive robust decision-making for autonomous agents, our works have furthered the state-of-the-art in runtime assurance architecture [12] and robust decision-making for non-stationary MDP [4] to allow the agent to choose among the controllers specializing in different objectives to control the system with the most up to date environment model [7, 13, 14], and to make robust decisions regardless of how much progress has been done towards environment model update [8, 10].

2.1 Dynamic Simplex for Adaptive Decision-Making

In the context of agents operating in non-stationary environments, the challenge of maintaining effective decision-making without immediate access to updated environmental models is crucial. To tackle the challenge, we have taken significant steps by augmenting the traditional simplex architecture [15] with the inclusion of a planner, leading to the development of what we refer to as the dynamic simplex framework [7]. This enhancement allows for a dual-mode operation: a highly conservative safety-oriented controller is activated in response to environmental changes or anomalies, and a goal-focused controller is employed when the objective function's

Proc. of the 23rd International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2024), N. Alechina, V. Dignum, M. Dastani, J.S. Sichman (eds.), May 6 – 10, 2024, Auckland, New Zealand. © 2024 International Foundation for Autonomous Agents and Multiagent Systems (www.ifaamas.org).

constraints are deemed to be met. This approach ensures the system's safety against potential environmental risks, even though it may temporarily set aside other objectives. Simultaneously, model updates could be done while the system operates under the safety controller, whether these updates are executed remotely or directly on-site. This ensures that the model continuously reflects the latest environmental changes without interrupting the system's safety-focused operations.

Additionally, to accelerate the collection of valuable data for model updating, we proposed a simulation framework and methods for optimizing experimentation [13, 14]. Considering the ongoing nature of model updates, the agent is required to engage in online planning that can adapt to changes in environmental parameters in real-time, while also keeping within the domain-specific limits on computation time. The Monte Carlo Tree Search (MCTS) algorithm is an ideal fit for our framework, enabling essential decision-making through an anytime online planning approach.

When compared with various controllers and extensions of the simplex architecture, our approach shows greater resilience to a range of adverse environmental impacts and better performance, all the while maintaining adherence to the objective function's constraints.

2.2 Act As You Learn

We have further advanced our methodology by developing adaptive Monte Carlo Tree Search (MCTS), detailed in our forthcoming full paper at AAMAS'24 [8]. This approach revises the assumptions associated with the always-given safe policy and the separation of model updating from decision-making within the Dynamic Simplex framework. Our methodology initiates with a robust policy derived from risk-averse MCTS (RA-MCTS) when environmental dynamics shift, ensuring safe exploration. In this phase, epistemic uncertainties play a crucial role in determining whether a given state-action pair has been sufficiently explored. This assessment is vital for understanding the extent of our knowledge about the environment and guiding the exploration process.

In the next phase, we implement a hybrid sampling methodology that carefully balances safe exploration with goal-directed actions. The role of aleatoric uncertainties becomes pivotal in this stage. Contrary to standard practice, in our approach, high aleatoric uncertainty, which signals a more non-deterministic environment, leads to the continued use of RA-MCTS to maintain a risk-averse stance. Conversely, when aleatoric uncertainty is lower, indicating a more predictable and deterministic environment, our system may transition to utilizing standard MCTS.

A key aspect of our methodology is the independent updating of model parameters, separate from the parameters describing the environmental dynamics. This distinction allows for more efficient knowledge transfer and helps in preventing the carry-over of uncertainty estimations from the previous model to the new one. Such independent updating ensures that our approach remains agile and accurate, even as environmental conditions evolve.

By benchmarking this method against the state-of-the-art [4] and various MCTS models, our results highlight the adaptability of our approach. It maintains robust performance in highly unpredictable environments due to its reliance on RA-MCTS under high aleatoric uncertainty and exhibits effectiveness similar to traditional MCTS in more stable environments where aleatoric uncertainty is lower.

3 FUTURE DIRECTIONS

The progression from Alphazero to MuZero has been a landmark in the advancement of artificial intelligence, particularly in the sphere of decision-making within nonstationary Markov Decision Processes (MDPs). MuZero's model-based strategy is notably promising for bridging the gap between theoretical reinforcement learning and practical applications, especially in domains where managing risk and adhering to constraints are paramount. This has inspired an in-depth investigation into the unique challenges posed by applying model-based methods to real-world scenarios, with a particular focus on nonstationary MDPs.

One of the primary challenges in these environments, demonstrated by methods like stochastic MuZero [1], is efficient learning for single-task objectives in high-dimensional, stochastic, and continuously evolving spaces. Notably, the continuous action spaces present in many real-world applications, such as robotics, add a layer of complexity to the decision-making process. In continuous action spaces, the agent must choose from an infinite set of possible actions, which significantly increases the complexity of finding optimal strategies, particularly under the changing dynamics of nonstationary MDPs. Furthermore, while these online learning approaches are adept at adapting to changing environments, questions remain regarding their capability to uphold domain-specific constraints in new dynamics and the amount of data required for effectively adapting the model to new tasks.

To address these challenges, a meta-learning approach could be considered, where a generalized model is trained and fine-tuned in response to environmental changes. However, this tends to be a reactive solution, updating models only after changes have occurred. To enhance this, one potential direction could be anticipating future shifts in environmental dynamics and optimizing the model in advance. So it can significantly improve the model's adaptability and decision-making capabilities. Furthermore, in nonstationary MDPs with continuous action spaces, incorporating safe strategies during model updates can help ensure that decision-making remains within constraint thresholds, even during transitional phases.

In conclusion, while the advancements represented by model-based approaches like MuZero are significant, their application to the dynamic, uncertain, and complex realm of nonstationary MDPs with continuous action spaces presents unique challenges. Addressing these challenges is crucial for the successful deployment of AI models in practical, constraint-sensitive environments.

REFERENCES

[1] Ioannis Antonoglou, Julian Schrittwieser, Sherjil Ozair, Thomas K Hubert, and David Silver. 2022. Planning in Stochastic Environments with a Learned Model. In *International Conference on Learning Representations*. https://openreview.net/forum?id=X6D9bAHhBQ1

[2] Yash Chandak, Georgios Theocharous, Shiv Shankar, Martha White, Sridhar Mahadevan, and Philip S. Thomas. 2020. Optimizing for the Future in Non-Stationary MDPs. In *Proceedings of the 37th International Conference on Machine Learning, ICML 2020, 13-18 July 2020, Virtual Event (Proceedings of Machine Learning Research, Vol. 119)*. PMLR, 1414–1425. http://proceedings.mlr.press/v119/chandak20a.html

[3] Taylor W. Killian, George Dimitri Konidaris, and Finale Doshi-Velez. 2017. Robust and Efficient Transfer Learning with Hidden Parameter Markov Decision

Processes. In *Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence, February 4-9, 2017, San Francisco, California, USA*, Satinder Singh and Shaul Markovitch (Eds.). AAAI Press, 4949–4950. https://doi.org/10.1609/aaai.v31i1.11065

[4] Erwan Lecarpentier and Emmanuel Rachelson. 2019. Non-Stationary Markov Decision Processes, a Worst-Case Approach using Model-Based Reinforcement Learning. In *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada*, Hanna M. Wallach, Hugo Larochelle, Alina Beygelzimer, Florence d'Alché-Buc, Emily B. Fox, and Roman Garnett (Eds.). 7214–7223. https://proceedings.neurips.cc/paper/2019/hash/859b00aec8885efc83d1541b52a1220d-Abstract.html

[5] Xiangguo Liu, Chao Huang, Yixuan Wang, Bowen Zheng, and Qi Zhu. 2022. Physics-Aware Safety-Assured Design of Hierarchical Neural Network based Planner. In *2022 ACM/IEEE 13th International Conference on Cyber-Physical Systems (ICCPS)*. 137–146. https://doi.org/10.1109/ICCPS54341.2022.00019

[6] Xiangguo Liu, Ruochen Jiao, Yixuan Wang, Yimin Han, Bowen Zheng, and Qi Zhu. 2023. Safety-Assured Speculative Planning with Adaptive Prediction. arXiv:2307.11876 [cs.RO]

[7] Baiting Luo, Shreyas Ramakrishna, Ava Pettet, Christopher Kuhn, Gabor Karsai, and Ayan Mukhopadhyay. 2023. Dynamic Simplex: Balancing Safety and Performance in Autonomous Cyber Physical Systems. In *Proceedings of the ACM/IEEE 14th International Conference on Cyber-Physical Systems (with CPS-IoT Week 2023)* (San Antonio, TX, USA) *(ICCPS '23)*. Association for Computing Machinery, New York, NY, USA, 177–186. https://doi.org/10.1145/3576841.3585934

[8] Baiting Luo, Yunuo Zhang, Abhishek Dubey, and Ayan Mukhopadhyay. 2024. Act as You Learn: Adaptive Decision-Making in Non-Stationary Markov Decision Processes. arXiv:2401.01841 [cs.AI]

[9] Anusha Nagabandi, Chelsea Finn, and Sergey Levine. 2019. Deep Online Learning via Meta-Learning: Continual Adaptation for Model-Based RL. arXiv:1812.07671 [cs.LG]

[10] Ava Pettet, Yunuo Zhang, Baiting Luo, Kyle Wray, Hendrik Baier, Aron Laszka, Abhishek Dubey, and Ayan Mukhopadhyay. 2024. Decision Making in Non-Stationary Environments with Policy-Augmented Search. arXiv:2401.03197 [cs.AI]

[11] Geoffrey Pettet, Ayan Mukhopadhyay, Mykel J. Kochenderfer, and Abhishek Dubey. 2022. Hierarchical Planning for Dynamic Resource Allocation in Smart and Connected Communities. 6, 4, Article 32 (nov 2022), 26 pages. https://doi.org/10.1145/3502869

[12] Dung T. Phan, Radu Grosu, Nils Jansen, Nicola Paoletti, Scott A. Smolka, and Scott D. Stoller. 2020. Neural Simplex Architecture. In *NASA Formal Methods - 12th International Symposium, NFM 2020, Moffett Field, CA, USA, May 11-15, 2020, Proceedings (Lecture Notes in Computer Science, Vol. 12229)*, Ritchie Lee, Susmit Jha, and Anastasia Mavridou (Eds.). Springer, 97–114. https://doi.org/10.1007/978-3-030-55754-6_6

[13] Shreyas Ramakrishna, Baiting Luo, Yogesh Barve, Gabor Karsai, and Abhishek Dubey. 2022. Risk-Aware Scene Sampling for Dynamic Assurance of Autonomous Systems. In *2022 IEEE International Conference on Assured Autonomy (ICAA)*. 107–116. https://doi.org/10.1109/ICAA52185.2022.00022

[14] Shreyas Ramakrishna, Baiting Luo, Christopher B. Kuhn, Gabor Karsai, and Abhishek Dubey. 2022. ANTI-CARLA: An Adversarial Testing Framework for Autonomous Vehicles in CARLA. In *2022 IEEE 25th International Conference on Intelligent Transportation Systems (ITSC)*. 2620–2627. https://doi.org/10.1109/ITSC55140.2022.9921776

[15] D. Seto, B. Krogh, L. Sha, and A. Chutinan. 1998. The Simplex architecture for safe online control system upgrades. In *Proceedings of the 1998 American Control Conference. ACC (IEEE Cat. No.98CH36207)*, Vol. 6. 3504–3508 vol.6. https://doi.org/10.1109/ACC.1998.703255

[16] Richard S. Sutton and Andrew G. Barto. 1998. *Reinforcement learning - an introduction*. MIT Press. https://www.worldcat.org/oclc/37293240

[17] Michael Wilbur, Salah Uddin Kadir, Youngseo Kim, Geoffrey Pettet, Ayan Mukhopadhyay, Philip Pugliese, Samitha Samaranayake, Aron Laszka, and Abhishek Dubey. 2022. An Online Approach to Solve the Dynamic Vehicle Routing Problem with Stochastic Trip Requests for Paratransit Services. In *2022 ACM/IEEE 13th International Conference on Cyber-Physical Systems (ICCPS)*. 147–158. https://doi.org/10.1109/ICCPS54341.2022.00020