Decision Making in Non-Stationary Environments with Policy-Augmented Search

Extended Abstract

Ava Pettet[†][*]
Vanderbilt University
Nashville, USA

Yunuo Zhang[†][*]
Vanderbilt University
Nashville, USA

Baiting Luo
Vanderbilt University
Nashville, USA

Kyle Wray
Stanford University
Stanford, USA

Hendrik Baier
Eindhoven University of Technology
Eindhoven, Netherlands

Aron Laszka
Pennsylvania State University
University Park, USA

Abhishek Dubey
Vanderbilt University
Nashville, USA

Ayan Mukhopadhyay
Vanderbilt University
Nashville, USA

ABSTRACT

Sequential decision-making is challenging in non-stationary environments, where the environment in which an agent operates can change over time. Policies learned before execution become stale when the environment changes, and relearning takes time and computational effort. Online search, on the other hand, can return sub-optimal actions when there are limitations on allowed runtime. In this paper, we introduce *Policy-Augmented Monte Carlo tree search* (PA-MCTS), which combines action-value estimates from an out-of-date policy with an online search using an up-to-date model of the environment. We prove several theoretical results about PA-MCTS. We also compare and contrast our approach with AlphaZero, another hybrid planning approach, and Deep Q Learning on several OpenAI Gym environments and show that PA-MCTS outperforms these baselines.

KEYWORDS

Sequential Decision-Making; Non-Stationary Environments; MCTS

ACM Reference Format:
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: Extended Abstract. 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

Sequential decision-making is present in many important problem domains, such as autonomous driving [2], emergency response [7], and medical diagnosis [1]. An open challenge in such settings is non-stationary environments, where the dynamics of the environment can change over time. A decision agent must adapt to these changes to avoid taking sub-optimal actions. Reinforcement learning (RL), especially with deep neural networks, struggles in such settings due to quickly outdated policies and the high cost of retraining [4, 8], while Monte Carlo tree search (MCTS) offers quicker adaptation but faces challenges with slow convergence in complex situations. This can be especially problematic in time-sensitive contexts [9], leading to potentially delayed responses [5]. In this work, we present a novel hybrid decision-making approach called Policy-Augmented Monte Carlo tree search (PA-MCTS), which combines a policy's action-value estimates with the returns generated by MCTS without changing either of the two approaches, i.e., the combination occurs entirely outside the online search tree. We argue that a hybrid decision-making approach that integrates RL and online planning can combine their strengths while mitigating their weaknesses in non-stationary environments. The intuition is that if the environment has not changed too much between when an optimal policy was learned and when a decision needs to be made, the policy can still provide useful information for decision-making. We also show how existing hybrid approaches, e.g., AlphaZero, can also be used for decision-making in non-stationary environments.

2 MARKOV DECISION PROCESSES IN NON-STATIONARY SETTINGS

Our focus is on scenarios where it's impractical to immediately learn a new policy after changes in the environment are detected, aiming to optimize decision-making during the transition to learning a new, nearly optimal policy. We explore this through the lens of non-stationary Markov decision processes (NSMDP), which introduce a temporal dimension to stationary MDPs and assume changes in the transition function are smoothly bounded [6]. Recognizing that some changes can be abrupt and significant, we propose a variant called transition-bounded non-stationary Markov decision processes (T-NSMDP), which limits the overall shift in

[†] These authors contributed equally to this work
[*] ava.pettet@vanderbilt.edu
[*] yunuo.zhang@vanderbilt.edu

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).

Environment	Setting	DDQN	MCTS	AlphaZero	PA-MCTS
Cartpole (varying *g*)	g = 9.8	**2500.0 ± 0.0**	846.456±43.228	2403.261±35.946	**2500.0±0.0**
	g = 20	**2500.0±0.0**	918.022±46.554	2278.90±52.9	**2500.0±0.0**
	g = 50	22.061±0.729	778.511±44.90	1920.261±78.547	**2500.0±0.0**
	g = 500	7.083±0.126	111.578±17.705	626.178±80.091	**954.656±90.150**
Frozen Lake	[1.000, 1.000, 1.000]	**1.0 ± 0.0**	**1.0 ± 0.0**	**1.0 ± 0.0**	**1.0±0.0**
	[0.833, 0.083, 0.083]	**0.830 ± 0.012**	0.806 ± 0.012	0.809±0.013	**0.830±0.012**
	[0.633, 0.183, 0.183]	0.522 ± 0.016	0.56 ± 0.017	0.523±0.017	**0.587±0.016**
	[0.433, 0.283, 0.283]	0.26 ± 0.014	0.764 ± 0.013	0.235±0.014	**0.796±0.013**
	[0.333, 0.333, 0.333]	0.12 ± 0.01	0.866 ± 0.01	0.114±0.011	**0.936±0.009**

Table 1: Results for all four environments with varying levels of non-stationarity. For each environment, the degree of *change* increases from top to bottom. We observe PA-MCTS comprehensively outperforms baseline approaches, including AlphaZero.

transition probabilities between the initial learning phase and execution, to better manage the adaptability of decision-making processes in dynamic environments. Consider the transition probability function $P_t(s' \mid s, a)$, where the subscript t denotes the time step under consideration. Now, consider that the environment undergoes *some* change between time steps 0 and t. We assume that: $\forall s, a : \sum_{s' \in S} |P_t(s' \mid a, s) - P_0(s' \mid a, s)| \leq \eta$ where $t \in \mathcal{T}$ (i.e., some point in time after the original policy was learned), and $\eta \in \mathbb{R}^+$ is a scalar bound. Although our algorithm only tackles discrete changes for now, our problem definition is agnostic to whether the change is continuous or discrete.

Our key hypothesis is that with small changes in the transition function, the Q function under an optimal policy does not change much, i.e., "good" actions remain valuable, and "bad" actions do not suddenly become promising. we show that the change in Q is bounded with respect to the change in P:

Theorem 2.1. *If* $\forall s, a\colon \sum_{s' \in S} |P_t(s' \mid a, s) - P_0(s' \mid a, s)| \leq \eta$, *and*$\forall s, a\colon |r(s, a)| \leq R$, *and the discount factor* $\gamma < 1$, *then* $|Q_0^{\pi_0^*}(s, a) - Q_t^{\pi_t^*}(s, a)| \leq \epsilon$ $\forall s, a$, *where* $\epsilon = \frac{\gamma \cdot \eta \cdot R}{(1-\gamma)^2}$ *(The proof is presented in the arXiv version).*[1]

The objective is to identify optimal actions at time t, maximizing future rewards under certain conditions: an initial optimal action-value function $Q_0^{\pi_0^*}(s, a)$ learned under different conditions, bounded changes in transition probabilities by η; an up-to-date black-box simulator for the current environment, and constrained computational resources preventing the learning of a new optimal policy at execution time.

3 POLICY AUGMENTED MONTE CARO TREE SEARCH

Policy-Augmented Monte Carlo Tree Search (PA-MCTS) integrates an online search with Q-values learned on the environment at an earlier decision epoch, even if the environment has changed. Rather than selecting an action based on the highest expected return estimated by the online search, PA-MCTS instead chooses the action that maximizes a convex combination of the previously learned Q-values and the MCTS estimates $\overline{G}$:

$$\operatorname*{arg\,max}_{a \in \mathcal{A}_s} \quad \alpha Q_0^{\pi_0^*}(s, a) + (1 - \alpha)\overline{G}_t(s, a) \tag{1}$$

where $Q_0^{\pi_0^*}(s, a)$ is the optimal[2] Q-function previously learned by the decision agent. The hyper-parameter α, set between 0 and 1, moderates the balance between learned Q-values and MCTS-derived expected returns, aiming to find a middle ground between low-variance, biased $Q_0^{\pi_0^*}$ estimates and high-variance, unbiased $\overline{G}_t$ estimates. Below, we quantify the total error in the expected return using PA-MCTS compared to an optimal (updated) policy.

Theorem 3.1. *When PA-MCTS is used for sequential decision making, the maximum difference between the return from an optimal policy and the return from following PA-MCTS is at most* $\frac{2(\alpha\epsilon - \alpha\delta + \delta)}{1-\gamma}$. *(The proof is presented in the arxiv version).*

4 EXPERIMENTS

Our approach is tested using four OpenAI Gym environments: Cart Pole, Frozen Lake, Cliff Walking, and Lunar Lander [3]. In Cart Pole, we introduce non-stationarity by varying the gravitational constant and the pole's mass. For Frozen Lake and Cliff Walking, non-stationarity is simulated by adding probabilities for unintended movements. In Lunar Lander, wind is introduced as a non-stationary factor, requiring the agent to adapt its actions to wind force.

Results for partial environments and settings we use to induce non-stationarity are presented in Table 1. The complete results are presented in the arXiv version. Our implementation is available at `https://github.com/scope-lab-vu/PAMCTS`. We observe as the change in the environment increases, the performance of the DDQN policy in isolation degrades as hypothesized. Second, we observe that PA-MCTS converges significantly faster than standard MCTS (with appropriate α). Third, in most environment settings, PA-MCTS outperforms Alphazero, Standard MCTS and DDQN.

ACKNOWLEDGMENTS

This material is based upon work sponsored by the National Science Foundation (NSF) under Grant CNS-2238815, the Defense Advanced Research Projects Agency (DARPA), and the Air Force Research Lab (AFRL). Results presented in this paper were obtained using the Chameleon testbed supported by the National Science Foundation. Any opinions, findings, conclusions, or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the NSF, AFRL, or DARPA.

[1] arXiv version: https://arxiv.org/abs/2401.03197

[2] In principle, we do not require the optimal Q-function. As shown in the experiments, an approximation also works well in practice.

REFERENCES

[1] Turgay Ayer, Oguzhan Alagoz, and Natasha K Stout. 2012. OR Forum—A POMDP approach to personalize mammography screening decisions. *Operations Research* 60, 5 (2012), 1019–1034.

[2] Maxime Bouton, Alireza Nakhaei, Kikuo Fujimura, and Mykel J Kochenderfer. 2019. Safe reinforcement learning with scene decomposition for navigating complex urban environments. In *2019 IEEE Intelligent Vehicles Symposium (IV)*. IEEE, 1469–1476.

[3] Greg Brockman, Vicki Cheung, Ludwig Pettersson, Jonas Schneider, John Schulman, Jie Tang, and Wojciech Zaremba. 2016. OpenAI Gym. *arXiv preprint arXiv:1606.01540* (2016).

[4] Wang Chi Cheung, David Simchi-Levi, and Ruihao Zhu. 2019. Non-stationary reinforcement learning: The blessing of (more) optimism. *Available at SSRN 3397818* (2019).

[5] Mykel J Kochenderfer, Tim A Wheeler, and Kyle H Wray. 2022. *Algorithms for decision making*. MIT Press.

[6] Erwan Lecarpentier and Emmanuel Rachelson. 2019. Non-stationary Markov decision processes, a worst-case approach using model-based reinforcement learning. *Advances in Neural Information Processing Systems* 32 (2019), 7216–7225.

[7] Ayan Mukhopadhyay, Geoffrey Pettet, Chinmaya Samal, Abhishek Dubey, and Yevgeniy Vorobeychik. 2019. An online decision-theoretic pipeline for responder dispatch. In *ACM/IEEE International Conference on Cyber-Physical Systems*. 185–196.

[8] Ronald Ortner, Pratik Gajane, and Peter Auer. 2020. Variational regret bounds for reinforcement learning. In *35th Uncertainty in Artificial Intelligence Conference*, Vol. 115. 81–90.

[9] Geoffrey Pettet, Ayan Mukhopadhyay, Mykel J. Kochenderfer, and Abhishek Dubey. 2021. Hierarchical Planning for Dynamic Resource Allocation in Smart and Connected Communities. *ACM Transactions on Cyber-Physical Systems* (2021). arXiv:2107.01292 [cs.MA]