RADemics

# Application of Proximal Policy Optimization for Stable Decision Making in Non-Stationary Environments

Paduchuru Viswanath, Abhithya T. P
JNTUA, VELALAR COLLEGE OF ENGINEERING AND TECHNOLOGY

# 9. Application of Proximal Policy Optimization for Stable Decision Making in Non-Stationary Environments

[1]Paduchuru Viswanath, Assistant Professor (A), School of Management Studies, JNTUA, Ananthapuramu - 515002, Andhra Pradesh, India, pvnmbajntua@gmail.com

[2]Abhithya T. P, Assistant Professor, Department of Computer Science and Engineering, Velalar College of Engineering and Technology, Erode, Tamil Nadu, India, abhithya1701@gmail.com

## Abstract

This book chapter delves into the application of Proximal Policy Optimization (PPO) for achieving stable decision-making in non-stationary environments. PPO, a widely recognized reinforcement learning algorithm, was examined through the lens of its stability mechanisms, including clipping, adaptive learning rates, entropy regularization, and discount factors. The chapter explores how these elements contribute to maintaining reliable performance in dynamic settings where the environment was continuously evolving. The integration of experience replay with PPO was discussed, emphasizing its role in enhancing stability and sample efficiency. Challenges such as balancing exploration and exploitation, optimizing hyperparameters, and managing computational costs are highlighted. Potential future directions for improving PPO's adaptability and robustness in non-stationary environments are proposed, including dynamic discounting and novel replay strategies. This chapter provides critical insights into leveraging PPO for real-world applications in uncertain and rapidly changing conditions, making it a valuable resource for researchers and practitioners in reinforcement learning.

**Keywords:**

Proximal Policy Optimization, Non-Stationary Environments, Stability Mechanisms, Adaptive Learning Rates, Experience Replay, Reinforcement Learning.

## Introduction

Proximal Policy Optimization (PPO) was a powerful reinforcement learning algorithm that has gained prominence due to its ability to balance exploration and exploitation while ensuring stability in policy updates [1,2]. Reinforcement learning (RL) has become a cornerstone for decision-making in complex environments, but its application in non-stationary environments presents unique challenges [3-5]. Non-stationary environments are characterized by dynamic changes in state distributions, rewards, or the underlying task structure, making it difficult for agents to maintain consistent performance [6-8]. PPO addresses some of these challenges through key mechanisms designed to stabilize learning, but the effectiveness of these mechanisms in such environments requires deeper investigation [9,10].

A critical aspect of PPO’s functionality was its use of stability mechanisms, which are aimed at preventing large, destabilizing updates to the policy [11,12]. One of the most important of these was the clipping mechanism, which ensures that the new policy does not deviate excessively from the old one [13-15]. This helps avoid the problem of policy overfitting, a common issue in reinforcement learning [16]. The algorithm also relies on adaptive learning rates, entropy regularization, and discount factors to further promote stability and prevent overfitting to outdated states, especially when the environment undergoes continuous changes [17,18]. Each of these components plays a crucial role in enabling PPO to maintain a balance between exploration and exploitation in dynamic environments, ensuring that the agent was able to learn efficiently and make reliable decisions in a fluctuating environment [19].

Non-stationary environments are inherently unpredictable, with changes occurring that can render past experiences irrelevant or even harmful to future decision-making [20,21]. This creates a risk of the agent making poor decisions based on outdated or misleading information [22]. Experience replay, which stores past experiences and allows for their reuse, offers a potential solution by helping to stabilize learning [23]. However, in non-stationary settings, the effectiveness of experience replay becomes limited, as the stored experiencesno longer accurately reflect the current dynamics of the environment [24]. This chapter explores the potential of experience replay in enhancing PPO's stability and offers insights into how to modify and optimize this technique for more effective learning in such environments [25].