Adapting Value Iteration with Adversarial Risk Analysis for Secure Reinforcement Learning

Jurij Ružejnikov[1,2]

[1] *Department of Adaptive Systems, Institute of Information Theory and Automation, CAS, Pod Vodárenskou věží 1143/4,182 00, Prague, Czech Republic*
[2] *Department of Information Engineering, FEM, Czech University of Life Sciences Prague, Kamýcká 129, 165 00, Prague - Suchdol, Czech Republic*

e-mail: ruzejnikov@utia.cas.cz

Key words: Adversarial Machine Learning, Adversarial Risk Analysis, Reinforcement Learning, Dynamic Programming

The increasing automation of processes via machine learning (ML) algorithms necessitates robust systems, particularly in reinforcement learning (RL) where agents learn through interaction. However, RL systems are highly vulnerable to adversarial attacks that manipulate rewards or observations, compromising their integrity. Traditional adversarial machine learning (AML) often focuses on supervised learning or relies on common knowledge assumption, which is unsuitable for scenarios where adversaries actively conceal information. Moreover, standard single-agent RL methods often falter in dynamic setting with adapting opponents due to the emergent non-stationarity. This work employs framework for enhancing the security of RL agents by leveraging Adversarial Risk Analysis (ARA), by modifying the standard Markov Decision Process (MDP) framework to explicitly account for the presence of an adversary affecting state transition and reward dynamics. The modification incorporates the decision maker's (DM's) subjective beliefs about potential adversarial actions. This enables the DM to use Bayesian principles for predicting adversarial actions in dynamic interactions. The framework adapts the value iteration process by forecasting and averaging over the predicted actions based on DM's beliefs, thereby aiming to enhance policy robustness.To model the DM's uncertainty regarding the adversary's strategy, this approach focuses on methods for belief elicitation and updating. Specifically, the DM maintains and updates a probability distribution over the adversary's likely actions based on past observations. For instance, Bayesian methods with Dirichlet priors can be used to estimate action frequencies in discrete settings, allowing the DM to continuously refine its understanding and adapt its strategy even when the adversary's true policy is unknown and potentially evolving. The primary contribution of this work is a methodology demonstrating how ARA facilitates secure RL by enabling the explicit modelling, prediction, and adaptation to interfering adversarial actions within the augmented MDP formulation. The efficacy and robustness of this approach is tested through a Coin Game experiment. This strategic environment serves as a testbed for adversarial interaction. By operating under the modified MDP, which explicitly models and adapts to the opponent's behaviour, agent aims to manage adversarial interactions more effectively than standard