Scalable Risk Sensitive Reinforcement Learning

Research proposal submitted for Dissertation Year Fellowship of the University of New Hampshire

by

Jia Lin Hau

Advisor

Dr. Marek Petrik

College of Engineering and Physical Sciences

Department of Computer Science

University of New Hampshire

Abstract

Reinforcement learning (RL) is a core component of artificial intelligence that enables decision-making in complex domains. Most existing RL algorithms ignore the risk associated with making decisions, which limits their application to high-stakes decision-making, which can be found in domains such as healthcare, finance, criminal justice, autonomous driving, and others. My goal in this project is to develop scalable risk-sensitive reinforcement learning models and algorithms that can make decisions that balance a decision's expected return with the risk involved. Numerous algorithms for risk-averse decision-making have been studied in the literature, but using them in the context of RL remains challenging. Many attempts to introduce risk-aversion to existing RL methods have led to unsound algorithms that choose actions that fail to optimize the specified risk metric. In the project, I propose taking a different route from the existing work. In particular, I propose building risk-averse RL algorithms by first creating a rigorous understanding of risk-averse algorithms in simple domains and then extending these algorithms to the full RL setting in which the agent has to learn to act in a complex and unknown environment.

1 Introduction

Automatic algorithms for decision-making are essential for many artificial intelligence tasks. Markov Decision Process (MDP) serves as the mathematical framework for modeling sequential decision-making with stochastic outcomes. Reinforcement learning (RL) is a research area that studies algorithms that can learn to act in large, complex MDPs just by interacting with the environment. MDPs and RL have been studied since the late 1950s [4]. Most work in RL has focused on optimizing the expected sum of rewards attained in the interaction with the environment. This goal of maximizing the expected sum of rewards is known as the *risk-neutral* objective.

As RL enters a broader set of application domains, it is increasingly becoming apparent that risk-neutral objectives are often insufficient [11]. In high-stakes environments such as autonomous driving, medical treatment, and fraud detection, the agent needs to consider also the *risk* of catastrophic failures associated with any decision. Over the past 20 years, *monetary risk measures* have become the most popular approach for accounting for both the expected value and possibility of catastrophic risk when evaluating uncertain outcomes. Standard RL methods focus only on risk-neutral objectives and can be very difficult to generalize to objectives that involve risk measures. The one-size-fits-all risk-neutral model is insufficient in model RL, especially when dealing with the risk preferences of diverse decision-makers in a multitude of domains.

Risk-averse RL has become an active research topic in recent years driven by the desire to bring automated, data-driven decision-making to high-stakes domains. Much of the research has focused on developing algorithms for specific application domains, such as motion planning [1, 5, 10], autonomous systems [14, 24], healthcare [16, 23, 25], investment [20] and others. Another stream of research in reinforcement learning has focused on specific risk measures or methods for evaluating the quality of an uncertain decision. Some of the most popular risk measures include value at risk (VaR) [11, 17, 19], conditional value at risk (CVaR) [2, 6], entropic risk measure (ERM) [12], entropic value at risk (EVaR) [21], and others. The risk-sensitive objective utilizes risk measures that map a distribution of possible outcomes to a real number to quantify the risk associated with each action given the current state.

The *main challenge* in risk-averse RL research is that the field has bifurcated into two distinct research streams: one focusing on the theoretical aspects of risk-averse decision-making and another one solely focusing on large-scale implementations without studying the fundamental soundness of the proposed techniques. This bifurcation has led to numerous elegant risk-averse algorithms that do not scale to large problems [12] and scalable algorithms that

are fundamentally flawed [8]. Little effort has been dedicated to bridging the gap between the two extremes.

2 Proposed Research Activities

My aim in this project is to develop risk-averse RL algorithms for intricate real-world high-stakes domains. These risk-averse RL algorithms must be fundamentally sound, mainly when applied to applications where failures can lead to significant financial losses, injury, or loss of life. My research in this project will help bridge the gap between theoretically sound and practically scalable algorithms.

I am well prepared to tackle the proposed research questions. My previous research laid rigorous foundations for risk-averse decision making [12] and identified important gaps and errors in existing risk-averse RL algorithms [11]. In addition, I am also proficient in designing scalable RL algorithms using deep learning.

Before describing my proposed research in detail, it is necessary to briefly summarize the limitations of the existing approaches to risk-averse RL. The majority of risk-averse RL literature falls into one of the following three specific categories.

The *first* line of work focuses on the theoretical foundations of risk-averse decision-making in small MDPs in which the distributions of rewards associated with individual actions are known [2, 17, 19]. My earlier work also falls in this category [11, 12]. This line of work develops the fundamental theoretical understanding with concrete mathematical proofs and is an essential stepping stone to developing true RL algorithms. However, proper RL algorithms must be able to *learn* to act even when the model of the environment is unknown and complex.

The *second* line of work focuses on a specific class of risk measures known as dynamic or Markov. When integrated with RL algorithms, these risk measures are especially convenient because they respect the fundamental dynamic decision-making structure [7, 13, 22]. The advantage of this approach is that most RL algorithms can be used with these dynamic risk measures with little modification. Such algorithms let the agents learn from interacting with the environment. Unfortunately, while being computationally convenient, the dynamic risk measures are virtually impossible to interpret and can lead to inexplicable and irrational behavior of agents. Given these limitations, dynamic risk measures have not seen much practical use.

The *third* line of work modifies large-scale RL algorithms to account for risk aversion. These algorithms employ neural nets as universal value function approximators to directly estimate the distribution of potential future outcomes linked to a specific state in the environment [3, 8, 9, 15]. These algorithms are seemingly practical because they can apply to large problems and do not require knowing the models accommodate an infinite state space, encompassing elements like images, videos, and continuous variables. Unfortunately, most of these algorithms are fundamentally flawed as shown by us [11] and others [18]. They compute solutions that do not reliably optimize the desired risk trade-off and can fail unpredictably, even in small domains.

The particular research objective during my dissertation year is to address the limitations of the existing risk-averse RL algorithms. I will leverage the rigorous formulations introduced in my earlier papers to build scalable and sound risk-averse RL algorithms.

1. Identifying dynamic program decompositions for a variety of risk measures. Dynamic programming is the principal component of RL algorithms and can be difficult to establish when the objective is not risk-neutral. My previous work has established dynamic programming equations for VaR, EVaR and ERM, and I will generalize this decomposition to other risk measures, too.

2. Most current risk-sensitive RL methods rely on the knowledge of the system dynamics. Identify and derive appropriate update rules for stochastic approximation to extend risk-sensitive RL methods to model-free environments. I will combine existing risk-averse dynamic programming algorithms with model-free RL algorithms to build new methods that can learn to make risk-averse decisions from interactions with the environment.

3. Design deep neural networks (NNs) that can learn in risk-averse RL settings. If successful, these algorithms will be able to make risk-averse decisions from raw image data without extensive prepossessing.

3 Conclusion

In this proposal, I describe my research on risk-sensitive reinforcement learning and identify three important but distinct and separate groups of work in the field. I explained the limitations of previous works and provided concrete steps to overcome them: (1) identify sufficient statistics, (2) derive and proof stochastic update rule, and (3) design neural network structure to incorporate the optimal policy for risk-sensitive objectives.

This research enables autonomous systems to have tailor-made objectives that correspond with the decision-maker's specific interests for their application. This investigation into risk-sensitive objectives in reinforcement learning not only aids in developing autonomous systems but also enhances our understanding of human decision-making processes. This research will allow decision scientists to gain insights from the theoretically proven optimal decision and learning rule.

References

[1] Mohamadreza Ahmadi, Xiaobin Xiong, and Aaron D Ames. Risk-averse control via CVaR barrier functions: Application to bipedal robot locomotion. *IEEE Control Systems Letters*, 6:878–883, 2021.

[2] Nicole Bäuerle and Jonathan Ott. Markov decision processes with average-value-at-risk criteria. *Mathematical Methods of Operations Research*, 74:361–379, 2011.

[3] Marc G Bellemare, Will Dabney, and Rémi Munos. A distributional perspective on reinforcement learning. In *International Conference on Machine Learning*, pages 449–458. PMLR, 2017.

[4] Richard Bellman. A Markovian decision process. *Journal of Mathematics and Mechanics*, pages 679–684, 1957.

[5] Daniel A Braun, Arne J Nagengast, and Daniel M Wolpert. Risk-sensitivity in sensorimotor control. *Frontiers in Human Neuroscience*, 5:1, 2011.

[6] Yinlam Chow, Aviv Tamar, Shie Mannor, and Marco Pavone. Risk-sensitive and robust decision-making: a CVaR optimization approach. *Advances in Neural Information Processing Systems*, 28, 2015.

[7] Anthony Coache and Sebastian Jaimungal. Reinforcement learning with dynamic convex risk measures. *Mathematical Finance*, 2023.

[8] Will Dabney, Georg Ostrovski, David Silver, and Rémi Munos. Implicit quantile networks for distributional reinforcement learning. In *International Conference on Machine Learning*, pages 1096–1105. PMLR, 2018.

[9] Will Dabney, Mark Rowland, Marc Bellemare, and Rémi Munos. Distributional reinforcement learning with quantile regression. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32, 2018.

[10] Astghik Hakobyan and Insoon Yang. Wasserstein distributionally robust motion control for collision avoidance using conditional value-at-risk. *IEEE Transactions on Robotics*, 38(2):939–957, 2021.

[11] Jia Lin Hau, Erick Delage, Mohammad Ghavamzadeh, and Marek Petrik. On dynamic programming decompositions of static risk measures in Markov decision processes. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023.

[12] Jia Lin Hau, Marek Petrik, and Mohammad Ghavamzadeh. Entropic risk optimization in discounted MDPs. In *International Conference on Artificial Intelligence and Statistics*, pages 47–76. PMLR, 2023.

[13] Wenjie Huang and William B Haskell. Risk-aware Q-learning for Markov decision processes. In *2017 IEEE 56th Annual Conference on Decision and Control (CDC)*, pages 4928–4933. IEEE, 2017.

[14] I Ge Jin, Bastian Schürmann, Richard M Murray, and Matthias Althoff. Risk-aware motion planning for automated vehicle among human-driven cars. In *2019 American Control Conference (ACC)*, pages 3987–3993. IEEE, 2019.

[15] Ramtin Keramati, Christoph Dann, Alex Tamkin, and Emma Brunskill. Being optimistic to be conservative: Quickly learning a CVaR policy. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 4436–4443, 2020.

[16] Ümit Emre Köse. *Optimal timing of living-donor liver transplantation under risk-aversion*. PhD thesis, Bilkent Universitesi (Turkey), 2016.

[17] Xiaocheng Li, Huaiyang Zhong, and Margaret L Brandeau. Quantile Markov decision processes. *Operations Research*, 70(3):1428–1447, 2022.

[18] Shiau Hong Lim and Ilyas Malik. Distributional reinforcement learning for risk-sensitive policies. *Advances in Neural Information Processing Systems*, 35:30977–30989, 2022.

[19] Yuanlie Lin, Congbin Wu, and Boda Kang. Optimal models with maximizing probability of first achieving target value in the preceding stages. *Science in China Series A: Mathematics*, 46:396–414, 2003.

[20] Seungki Min, Ciamac C Moallemi, and Costis Maglaras. Risk-sensitive optimal execution via a conditional value-at-Risk objective. *arXiv preprint arXiv:2201.11962*, 2022.

[21] Xinyi Ni and Lifeng Lai. EVaR optimization for risk-sensitive reinforcement learning. *IEEE Transactions on Information Theory*, 2022.

[22] Yun Shen, Michael J Tobia, Tobias Sommer, and Klaus Obermayer. Risk-sensitive reinforcement learning. *Neural Computation*, 26(7):1298–1328, 2014.

[23] Ashudeep Singh, Yoni Halpern, Nithum Thain, Konstantina Christakopoulou, E Chi, Jilin Chen, and Alex Beutel. Building healthy recommendation sequences for everyone: A safe reinforcement learning approach. In *FAccTRec Workshop*, 2020.

[24] Yuheng Wang and Margaret P Chapman. Risk-averse autonomous systems: A brief history and recent developments from the perspective of optimal control. *Artificial Intelligence*, 311:103743, 2022.

[25] Huaiyang Zhong. *Decision making for disease treatment: Operations Research and Data Analytic Modeling*. Stanford University, 2020.