Ensuring safe state space exploration of a Markov decision process (MDP) using Bayesian non-parametric (BNP) models for reinforcement learning (RL)

Lancaster University

Researcher: Xavier Hickman
Supervisors: Prof. Dan Prince, Prof. Neeraj Suri

Brief Overview of RL and Safe RL

Reinforcement learning addresses optimisation problems such as optimal control where other machine learning paradigms such as supervised and unsupervised fail [1]. Fundamentally optimal control is a field of mathematics that studies the problem of finding the best control strategy for a system, given a set of constraints and a criterion that defines optimality. Reinforcement learning algorithms have been shown to handle highly complex and uncertain environment dynamics which makes them well suited to a plethora of real-world applications such as autonomous vehicles and robotics. RL is also highly data efficient, in that it can learn from a limited quantity of data, which in many problem spaces is real barrier to entry for conventional ML paradigms.

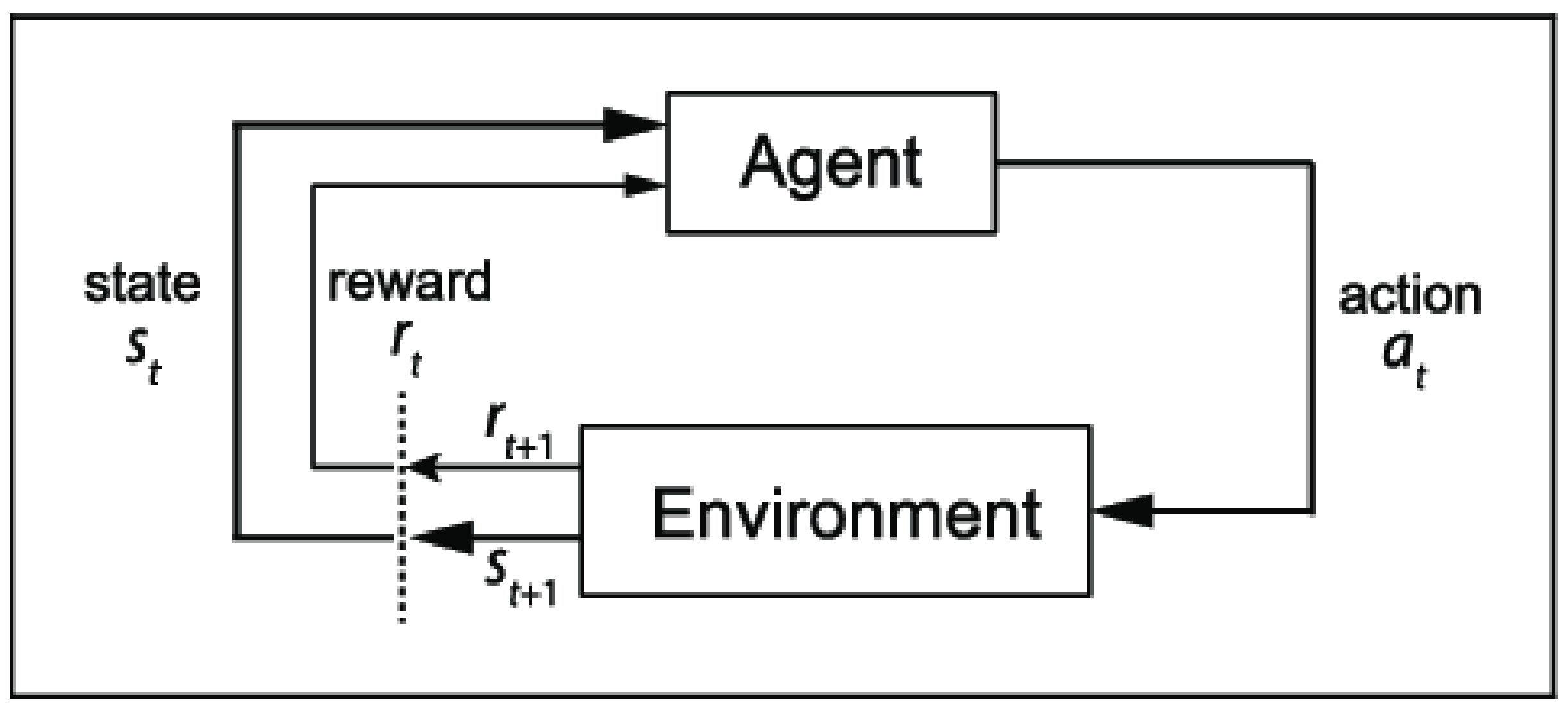

Figure 1: Schematic of Reinforcement learning Paradigm [1].

Safe reinforcement learning is a subfield of RL that focuses on the safety and reliability of reinforcement learning algorithms. The key motivation behind the discipline is the exploration-exploitation mechanism that enables conventional RL algorithms so learn. When training begins there is more weight given to choosing random actions to expand the known state space and nearer the end of training there is more weight given to choosing actions based existing knowledge of state space. This mechanism creates a problem in safety critical environments because the nature of the stochastic policy could potentially cause damage to the agent or the environment which in application contexts such as self-driving cars or military drones would be detrimental.

Safe RL Problem Formulation

Markovian sequential decision-making problems are often formulated as Markov decision process (MDP). MDPs are used across a variety of fields including RL, operations research and control theory. An extension of the MDP is the constrained MDP (CMDP) where the tuple includes a set of constraints which can be used to model properties such as safety. A formulation of a constrained Markov decision process is shown in figure 2.

$$M = \langle\ S, A, P, r(.,.,.), \gamma, C \rangle$$

Figure 2: Constrained Markov decision process [2].

Figure 3 shows is an alternative formulation of the safe RL problem. We aim to maximize the value function for some policy π of some state *s* at a given timestep *t* subject to the safety function evaluation of that state *s* and time step *t* being at or above some scalar threshold *h* which is problem specific [2].

$$\textit{maximise}: V_N^{\pi}(s_t) = E\left[\sum_{t=1}^{N} \gamma^{t-1} r(s_t, a_t))\right]$$

$$\textit{subject to}: g(s_t) \geq h, \forall t = [1, N].$$

Figure 3: Safe RL problem formulation [2].

Why Bayesian non-parametric models?

Deep neural networks (DNNs) have been a very popular choice for function approximators in classical reinforcement learning in recent years [3][4], however Bayesian non-parametric (BNP) models offer some unique advantages over DNN's when optimizing for safety.

- DNN's have been shown to be very sensitive to distributional shifts in input data which in reinforcement learning problems is very common. Distributional shifts in observation data can occur for several reasons, but the most common is the non-stationarity of environments where the underlying distribution of states and actions change over time. In contrast BNP models are designed to be robust to distributional shifts in data and can learn flexible distributions that capture the underlying structure of the data.

- BNP models can capture and quantify uncertainty which is especially useful in safe RL as many environments where safe RL algorithms are applicable often have high levels of uncertainty. BNP models can use this when making decisions in these uncertain environments to ensure safe actions are taken whereas DNNs have no natural way of modelling uncertainty.

- DNNs are notoriously uninterpretable [5] whereas BNP models provide highly interpretable models as a result of their simpler model structure and transparent uncertainty approximations. The inherent interoperability can help in explaining an agent's behavior and actions which is especially useful in safety critical application contexts.

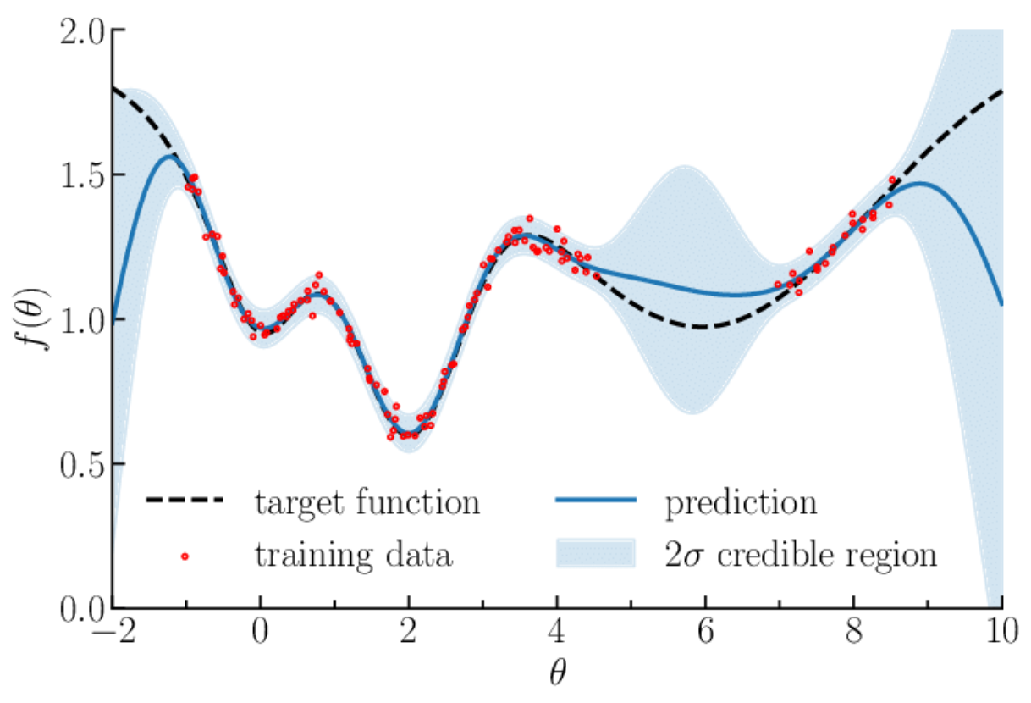

Figure 4: Illustration of gaussian process regression in 1 dimension [6]

Figure 4 shows an illustration of a 1-dimensional gaussian process which is a popular BNP model. The uncertainty is captured as the blue shaded area, so the larger the shaded are at certain points in the function the less certainty there is in the approximation.

Ongoing work

Stability of MARL in the face of perturbed communication:
Some of our ongoing work is looking at the stability and resilience of certain multi-agent reinforcement learning (MARL) algorithms in the face of severe network interruptions. These interruptions could affect the availability of communications mediums, or the integrity of messages sent. We are specifically interested in the multi-agent deep deterministic policy gradient algorithm (MADDPG) and the multi-agent proximal policy optimization (MAPPO) algorithm as they both demonstrate good performance on the multi-particle environments that best simulate groups of cooperating and competing autonomous systems [7][8].

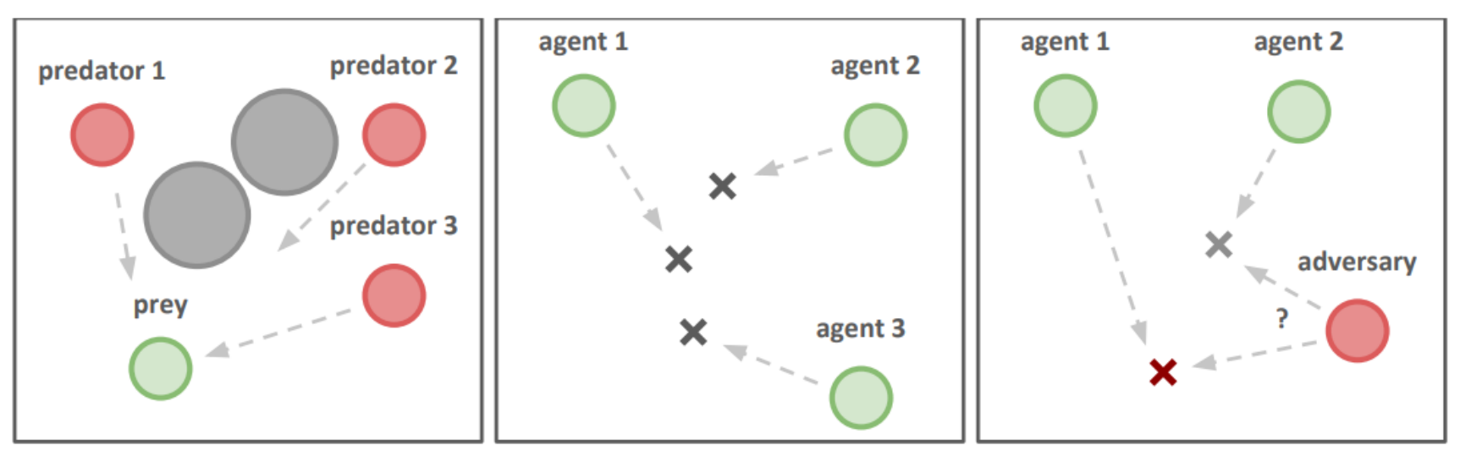

Figure 5: Screenshot of environments in the multi-particle environment (MPE) from the MADDPG paper [7]

Improving the SafeMDP algorithm:
Our ongoing work also includes a paper that is looking to improve the robustness of the SafeMDP algorithm [9] which utilizes a gaussian process to approximate safety in highly uncertain environments. SafeMDP provides desirable theoretical guarantees and demonstrates good empirical performance. We aim to use the SafeMDP algorithm as a base and develop a new algorithm which is based on a similar model to the gaussian process but with improved empirical performance w.r.t outlier observations and to build on the existing theoretical guarantees.

References

[1] Sutton, R. S., & Barto, A. G. (2018). Reinforcement learning: An introduction (2nd ed.). The MIT Press.
[2] Akifumi Wachi, Yanan Sui, Yisong Yue, and Masahiro Ono. Safe exploration and optimization of constrained mdps using gaussian processes. In Proceedings of the AAAI Conference on Artificial Intelligence, volume 32, 2018.
[3] Mnih, V., Kavukcuoglu, K., Silver, D., Graves, A., Antonoglou, I., Wierstra, D. & Riedmiller, M. (2013). Playing Atari with Deep Reinforcement Learning. , .
[4] Schulman, J., Wolski, F., Dhariwal, P., Radford, A. & Klimov, O. (2017). Proximal Policy Optimization Algorithms.. *CoRR*, abs/1707.06347.
[5] S. Chakraborty *et al.*, "Interpretability of deep learning models: A survey of results," *2017 IEEE SmartWorld, Ubiquitous Intelligence & Computing, Advanced & Trusted Computed, Scalable Computing & Communications, Cloud & Big Data Computing, Internet of People and Smart City Innovation*
[6] Leclercq, Florent. (2018). Bayesian optimization for likelihood-free cosmological inference. Physical Review D. 98. 10.1103/PhysRevD.98.063511.
[7] Lowe, Ryan & Wu, Yi & Tamar, Aviv & Harb, Jean & Abbeel, Pieter & Mordatch, Igor. (2017). Multi-Agent Actor-Critic for Mixed Cooperative-Competitive Environments.
[8] Yu, Chao & Velu, Akash & Vinitsky, Eugene & Wang, Yu & Bayen, Alexandre & Wu, Yi. (2021). The Surprising Effectiveness of PPO in Cooperative, Multi-Agent Games.
[9] Turchetta, Matteo & Berkenkamp, Felix & Krause, Andreas. (2016). Safe Exploration in Finite Markov Decision Processes with Gaussian Processes.

This work is supported by the Engineering and Physical Sciences Research Council [grant number: EP/V026763/1]