Research proposal

Pierre Thodoroff

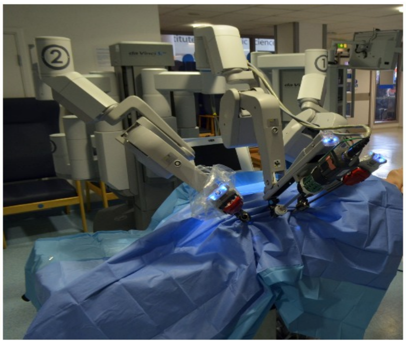
Abstract

This proposal considers the problem of robust sequential decision making in non-linear environments. Reinforcement learning has demonstrated high potential for solving complex problems in non-linear environments but has lacked efficiency and robustness. The objective of this proposal is to leverage the control and probabilistic reasoning literature to improve reinforcement learning agents.

In recent years, machine learning has shown tremendous success in a wide variety of tasks such as computer vision (Krizhevsky et al., 2012), healthcare (Miotto et al., 2017; Faust et al., 2018), and natural language processing (Collobert and Weston, 2008). However, deploying those algorithms in real-world conditions has remained challenging, in particular in critical applications such as healthcare. The main challenge is that deployment conditions are often different than the training ones leading to a problem known as distributional shift (Quionero-Candela et al., 2009). In this context, adaptive algorithms are interesting to consider (Lee and Markus, 1967; Landau et al., 1998; Lattimore and Szepesvári, 2018) as they continuously adapt to the incoming data during deployment. The two leading frameworks for sequential decision-making environment are reinforcement learning (Sutton and Barto, 2018) and control theory (Lee and Markus, 1967). Both attempts to model sequential decision making, however, control theory has a strong emphasis on models, sample efficiency, and theoretical guarantee. In contrast, reinforcement learning is more focused on learning from experiences and data.

Reinforcement learning has had significant success using deep learning on complex games such as Go and Atari (Mnih et al., 2013; Silver et al., 2016; Schrittwieser et al., 2019). Deep learning models possess a remarkable capacity to model complex non-linear phenomena, however, they often lack sample efficiency, reliability, and interpretability. When deployed in the real world, this can lead to unexpected and unstable behaviour (Yuan et al., 2019). In contrast, control theory and probabilistic models are widely used on simple(linear, low dimensional) real-world problems such as control of physical processes (Edwards and Spurgeon, 1998). However, scaling these methods to non-linear high-dimensional data has remained challenging.

In this proposal, I argue that in order to deploy reinforcement learning agents in the real world, it is essential to develop similar efficiency and robustness properties that have been developed in control theory. I propose to leverage the extensive control and probabilistic reasoning literature to improve RL algorithms. First, I present the existing relationships between model-based RL and control theory; then, I discuss two exciting directions for future research. The first one considers using Sequential Monte-Carlo methods to improve planning for non-linear environments. The second direction focuses on designing robust controllers by exploring the connections between adversarial learning, robust control theory, and uncertainty modelling.



1 Background

One of the most popular streams of reinforcement learning, namely model-free RL, attempts to find the solution of a control problem without building a model of the environment. The canonical algorithm in this framework is Q-learning (Watkins and Dayan, 1992). Model-free RL has achieved impressive results in simulated environments such as video games. However, model-free methods are often sample inefficient (Mnih et al., 2013; Recht, 2019). In contrast, model-based methods first build a model of the environment $\mathcal{P}$ and then exploit this model to derive a policy (Deisenroth et al., 2013). This framework is identical to some of the methods developed in control theory. This is why many of the terms used in model-based RL come from control theory. A summary of the basic equivalencies between terms can be found in table 1.

When using model-based RL, the process can be decomposed in two phases. The first phase consists of learning the model (nominal model) of the environment. The second phase considers leveraging the learned model to derive a policy (controller). Both aspects of model-based RL are of interest to me, however, in this proposal, we focus on how to derive a controller from a learned model. In the discrete setting, assuming a perfect learned model, the optimal controller can easily be derived using linear algebra and the Bellman equation (Bellman et al., 1954). If the state-space and action-space are continuous, but the dynamics are linear and the cost quadratic, this problem is called the linear quadratic regulator problem. It is solvable analytically due to the convexity of the optimization problem (Bemporad et al., 2002). However, when such assumptions are not satisfied (non-linearity of the dynamics or cost function), it is necessary to use approximations.

The most popular framework to solve this issue originates from control theory and is called Model Predictive Control (MPC). It was initially developed for chemical applications (Qin and Badgwell, 2003). MPC is a finite-horizon iterative method. At each time step t, a sampling method is used to predict different trajectories and find the one that minimizes the cost function. In RL, the process of sampling from a model of the environment to maximize the reward is called planning. The main problem in planning is to balance the computational complexity of the method and its performance.

One of the most successful techniques for planning exploits the differentiability of the cost function to optimize the sequence of states and actions (Deisenroth and Rasmussen, 2011). If the cost function is not differentiable, it is necessary to use sampling-based methods to approximate the cost function at specific points (Bagnell and Schneider, 2001; Ko et al., 2007). Once

	Reinforcement Learning	Control Theory
Nomenclature	Action	Control rule
	Environment	Plant
	Reward	Cost
	Learned model	Nominal model
Methods	Model-based	System identification+Model Predictive Control
	Model-free	Data-driven adaptive control

Table 1: Taxonomy of reinforcement learning versus control theory.

a sampling strategy is defined, there exist many algorithms (gradient-based or gradient-free) to derive the controller(Deisenroth et al., 2013). In this proposal, we focus on the problem of sampling. To plan and sample in discrete environments, one can use a Monte-Carlo tree search (Kearns et al., 2002) algorithm to search through all potential paths. However, in continuous-settings, enumerating all potential paths is intractable. The most popular solutions in continuous setting are cross-entropy (De Boer et al., 2005) and random shooting (Chua et al., 2018) methods. However, these either assume that the future trajectory is Gaussian (unimodal) or do not exploit the learned cost function, leading to inefficient planning. This leads to the first question of this proposal:

How to efficiently plan with non-linear dynamics?

The idea developed later in this proposal is to view planning as a density estimation problem over the future optimal trajectory and to use flexible sampling-based methods such as Sequential Monte-Carlo (SMC) to solve it (Doucet et al., 2001). In particular, SMC has the capacity to model non-linear multimodal distributions.

In all the methods discussed in the previous paragraph, there was the underlying assumption that the learned model was a good representation of the environment. To circumvent this problem, many papers use the certainty assumption, which assumes that the optimal policy for the learned model is the same one as the optimal policy for the true dynamics. However, this is highly unlikely to be satisfied in practice, especially when a few samples are available. This motivated the development of robust model predictive control (Bemporad and Morari, 1999). The goal of robust control is to guarantee performance under model uncertainty and noise. The majority of the results developed in this literature focused on linear dynamics which leads to the second question:

How to guarantee robust control in a non-linear environment when the nominal model is incorrect?

In RL literature, the problem of robust control is tackled using the probabilistic reasoning paradigm by considering a probability distribution over the dynamics of the environment and modelling the potential epistemic uncertainty. Historically, the most successful methods are based on Locally Weighted Bayesian Regression and Gaussian Process (Deisenroth and Rasmussen, 2011). In this proposal, we consider discussing how the probabilistic approach relates to the one developed in robust control theory and how we could develop new theoretical and practical algorithms based on these relationships.

2 Proposal

We discuss two interesting research directions based on the questions developed in the background section.

Sequential Monte-Carlo for planning: There exist many connections between inference in probabilistic models and control theory, one of the oldest ones being between Kalman filters and linear-quadratic-gaussian problems (Todorov, 2008). The relationship between planning and inference has initially been discussed in control theory (Andrieu et al., 2004). More recently, (Piché et al., 2018) casts planning in continuous environments as an instance of a filtering problem in sequential probabilistic models and uses Sequential Monte Carlo(SMC) methods to solve it. The central idea is to view planning as finding states that maximize the probability of observing high-return trajectories. To do so, SMC uses a learned value function to select the most promising particles. As the value function is parametrized by deep neural networks, an interesting extension would be to directly differentiate through the learned value function to optimize the sampled states and compare the performance of this method against SMC. One issue that might arise from differentiating through the value function is obtaining degenerative solutions and falling into local optima. However, this issue also happens in SMC (particle degeneracy), and several solutions have been proposed, such as backward simulation (Lindsten et al., 2013). Another interesting direction considers the problem of planning long term trajectories in physical systems (Saemundsson et al., 2019). One could consider integrating the fundamental laws of physics, such as conservation of energy, in the selection step of the particles. This would enforce predictions that satisfy basic physical laws.

Probabilistic learning for robust control: In RL literature, the problem of robust control is tackled by designing a probabilistic model capturing the uncertainty of the model (Deisenroth and Rasmussen, 2011) and planning according to this uncertainty. In contrast, the most popular framework in control theory, namely, H_∞ theory, attempts to guarantee the performance of a controller under some bounded disturbance. This inspired further work in RL that considers modelling the noise and uncertainty as induced by an adversary (Morimoto and Doya, 2001). In practice, this yields a min-max optimization problem that can be solved similarly to Generative Adversarial Networks (Goodfellow et al., 2014). I am interested in further investigating the relationship between uncertainty estimation in probabilistic reasoning and adversarial optimization in the context of model-based RL. The relationship between both phenomena has started to be investigated in supervised learning (Smith and Gal, 2018). Another interesting direction would be to extend the work of (Vinogradska et al., 2016) that attempts to provide stability guarantee of the controller when the model is learned using a Gaussian Process. In particular, a first step could be to consider whether similar guarantee can be obtained using deep probabilistic models.

Conclusion: To summarize, the overall theme of this proposal is to design robust sequential decision-making algorithms that are capable of dealing with complex non-linear environments. To do so, I propose to design algorithms at the intersection of deep RL, control theory, and probabilistic reasoning.

References

Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*, pages 1097–1105, 2012.

Riccardo Miotto, Fei Wang, Shuang Wang, Xiaoqian Jiang, and Joel T Dudley. Deep learning for healthcare: review, opportunities and challenges. *Briefings in bioinformatics*, 19(6):1236–1246, 2017.

Oliver Faust, Yuki Hagiwara, Tan Jen Hong, Oh Shu Lih, and U Rajendra Acharya. Deep learning for healthcare applications based on physiological signals: A review. *Computer methods and programs in biomedicine*, 161:1–13, 2018.

Ronan Collobert and Jason Weston. A unified architecture for natural language processing: Deep neural networks with multitask learning. In *Proceedings of the 25th international conference on Machine learning*, pages 160–167. ACM, 2008.

Joaquin Quionero-Candela, Masashi Sugiyama, Anton Schwaighofer, and Neil D Lawrence. *Dataset shift in machine learning*. The MIT Press, 2009.

Ernest Bruce Lee and Lawrence Markus. Foundations of optimal control theory. Technical report, Minnesota Univ Minneapolis Center For Control Sciences, 1967.

Ioan Doré Landau, Rogelio Lozano, Mohammed M'Saad, et al. *Adaptive control*, volume 51. Springer New York, 1998.

Tor Lattimore and Csaba Szepesvári. Bandit algorithms. 2018.

Richard S Sutton and Andrew G Barto. *Reinforcement learning: An introduction*. 2018.

Volodymyr Mnih, Koray Kavukcuoglu, David Silver, Alex Graves, Ioannis Antonoglou, Daan Wierstra, and Martin Riedmiller. Playing atari with deep reinforcement learning. *arXiv preprint arXiv:1312.5602*, 2013.

David Silver, Aja Huang, Chris J Maddison, Arthur Guez, Laurent Sifre, George Van Den Driessche, Julian Schrittwieser, Ioannis Antonoglou, Veda Panneershelvam, Marc Lanctot, et al. Mastering the game of go with deep neural networks and tree search. *nature*, 529 (7587):484, 2016.

Julian Schrittwieser, Ioannis Antonoglou, Thomas Hubert, Karen Simonyan, Laurent Sifre, Simon Schmitt, Arthur Guez, Edward Lockhart, Demis Hassabis, Thore Graepel, Timothy Lillicrap, and David Silver. Mastering atari, go, chess and shogi by planning with a learned model, 2019.

Xiaoyong Yuan, Pan He, Qile Zhu, and Xiaolin Li. Adversarial examples: Attacks and defenses for deep learning. *IEEE transactions on neural networks and learning systems*, 2019.

Christopher Edwards and Sarah Spurgeon. *Sliding mode control: theory and applications*. Crc Press, 1998.

Christopher JCH Watkins and Peter Dayan. Q-learning. *Machine learning*, 8(3-4):279–292, 1992.

Benjamin Recht. A tour of reinforcement learning: The view from continuous control. *Annual Review of Control, Robotics, and Autonomous Systems*, 2(1):253–279, May 2019. ISSN 2573-5144. doi: 10.1146/annurev-control-053018-023825. URL `http://dx.doi.org/10.1146/annurev-control-053018-023825`.

Marc Peter Deisenroth, Gerhard Neumann, Jan Peters, et al. A survey on policy search for robotics. *Foundations and Trends® in Robotics*, 2(1–2):1–142, 2013.

Richard Bellman et al. The theory of dynamic programming. *Bulletin of the American Mathematical Society*, 60(6):503–515, 1954.

Alberto Bemporad, Manfred Morari, Vivek Dua, and Efstratios N Pistikopoulos. The explicit linear quadratic regulator for constrained systems. *Automatica*, 38(1):3–20, 2002.

S Joe Qin and Thomas A Badgwell. A survey of industrial model predictive control technology. *Control engineering practice*, 11(7):733–764, 2003.

Marc Deisenroth and Carl E Rasmussen. Pilco: A model-based and data-efficient approach to policy search. In *Proceedings of the 28th International Conference on machine learning (ICML-11)*, pages 465–472, 2011.

J Andrew Bagnell and Jeff G Schneider. Autonomous helicopter control using reinforcement learning policy search methods. In *Proceedings 2001 ICRA. IEEE International Conference on Robotics and Automation (Cat. No. 01CH37164)*, volume 2, pages 1615–1620. IEEE, 2001.

Jonathan Ko, Daniel J Klein, Dieter Fox, and Dirk Haehnel. Gaussian processes and reinforcement learning for identification and control of an autonomous blimp. In *Proceedings 2007 ieee international conference on robotics and automation*, pages 742–747. IEEE, 2007.

Michael Kearns, Yishay Mansour, and Andrew Y Ng. A sparse sampling algorithm for near-optimal planning in large markov decision processes. *Machine learning*, 49(2-3):193–208, 2002.

Pieter-Tjerk De Boer, Dirk P Kroese, Shie Mannor, and Reuven Y Rubinstein. A tutorial on the cross-entropy method. *Annals of operations research*, 134(1):19–67, 2005.

Kurtland Chua, Roberto Calandra, Rowan McAllister, and Sergey Levine. Deep reinforcement learning in a handful of trials using probabilistic dynamics models. In *Advances in Neural Information Processing Systems*, pages 4754–4765, 2018.

Arnaud Doucet, Nando De Freitas, and Neil Gordon. An introduction to sequential monte carlo methods. In *Sequential Monte Carlo methods in practice*, pages 3–14. Springer, 2001.

Alberto Bemporad and Manfred Morari. Robust model predictive control: A survey. In *Robustness in identification and control*, pages 207–226. Springer, 1999.

Emanuel Todorov. General duality between optimal control and estimation. In *2008 47th IEEE Conference on Decision and Control*, pages 4286–4292. IEEE, 2008.

Christophe Andrieu, Arnaud Doucet, Sumeetpal S Singh, and Vladislav B Tadic. Particle methods for change detection, system identification, and control. *Proceedings of the IEEE*, 92(3):423–438, 2004.

Alexandre Piché, Valentin Thomas, Cyril Ibrahim, Yoshua Bengio, and Chris Pal. Probabilistic planning with sequential monte carlo methods. 2018.

Fredrik Lindsten, Thomas B Schön, et al. Backward simulation methods for monte carlo statistical inference. *Foundations and Trends® in Machine Learning*, 6(1):1–143, 2013.

Steindor Saemundsson, Alexander Terenin, Katja Hofmann, and Marc Peter Deisenroth. Variational integrator networks for physically meaningful embeddings, 2019.

Jun Morimoto and Kenji Doya. Robust reinforcement learning. In *Advances in Neural Information Processing Systems*, pages 1061–1067, 2001.

Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *Advances in neural information processing systems*, pages 2672–2680, 2014.

Lewis Smith and Yarin Gal. Understanding measures of uncertainty for adversarial example detection. *arXiv preprint arXiv:1803.08533*, 2018.

Julia Vinogradska, Bastian Bischoff, Duy Nguyen-Tuong, Anne Romer, Henner Schmidt, and Jan Peters. Stability of controllers for gaussian process forward models. In *International Conference on Machine Learning*, pages 545–554, 2016.