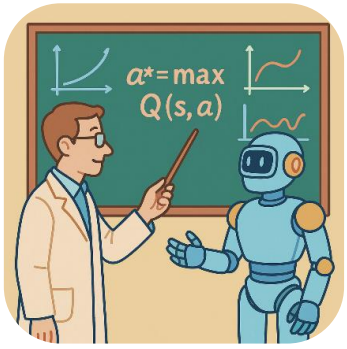
SSM
Scuola Superiore Meridionale

MERC PhD Project Proposal 2024/2025

Smarter, Safer, Faster: Bridging Control Theory and Reinforcement Learning for Efficient and Trustworthy Decision-Making



Keywords

Reinforcement learning; Control theory; Data-efficiency; Formal guarantees

Key research question(s)	• How can we make learning faster in RL? • How can we provide guarantees on learned policies?
Tools and techniques	Reinforcement learning; Control theory; Markov decision processes; Stochastic approximation; Probability
Target application(s)	Robotics; Mobile autonomous agents
Beneficiaries	Industry; Researchers

Supervisors

Supervisor 1: **Marco Coraggio** (marco.coraggio@unina.it)
http://marco-coraggio.com
Expertise: Control theory, Complex networks, Data-driven control

Supervisor 2: **Francesco De Lellis** (francesco.delellis@unina.it)
https://sites.google.com/site/dibernardogroup/group/francesco-de-lellis
Expertise: Control theory, Machine learning, Reinforcement learning

Supervisor 3: **Giovanni Russo** (giovarusso@unisa.it)
www.sites.google.com/view/giovanni-russo
Expertise: Theory of decision-making, Data-driven systems, Control Theory, Complex Cyber-physical systems, Network systems

Supervisor 4: **Mirco Musolesi** (m.musolesi@ucl.ac.uk)
https://www.mircomusolesi.org
Expertise: Machine learning, Reinforcement learning, Computational models

Project description

Introduction

Context and motivation

Reinforcement learning (RL) h as achieved impressive success across a wide range of complex tasks—from predicting protein 3D structures [Jumper, 2021], to controlling plasma in fusion reactors [Degrave, 2022], high-performance drone navigation [Kaufmann, 2023], and mastering collaborative or competitive games such as two-player sports, Go, and advanced video games [Won, 2021]. More recently, RL has even played a key role in fine-tuning large language models like GPT-3.5 and GPT-4. Despite these breakthroughs, current RL methodologies still face critical limitations. Specifically:

1. they often demand prohibitively long training times, especially when dealing with large state and action spaces, limiting the accessibility and scalability of the technology;
2. they typically lack theoretical guarantees on the quality or safety of the learned policies, hindering their adoption in safety-critical or high-performance domains.

This project seeks to address these challenges by integrating insights and tools from control theory, with the goal of developing RL methods that are not only more efficient but also more reliable and grounded in formal guarantees.

State of the art

Recent studies have demonstrated that integrating principles from control theory into reinforcement learning (RL) can lead to more effective and sample-efficient control strategies. For example, [Zanon, 2021] used RL to dynamically tune the parameters of both the model and the objective function in Model Predictive Control (MPC). In a different approach, [Gu, 2016] accelerated learning by iteratively fitting local linear models to data gathered through exploration.

A particularly compelling contribution is found in [De Lellis, 2023a], where the concept of a *control tutor* was introduced. This framework assumes the availability of a (possibly imperfect) feedback control law, which is used to intermittently guide the agent during exploration. Experiments showed that this guidance significantly reduces learning time, both for tabular methods and deep value-based algorithms. Despite these promising results, several open questions remain. Namely, (i) Does the benefit of the control tutor extend consistently to policy-based methods? (ii) In a multi-agent scenario, can multiple control tutors coexist, or will their guidance conflict? (iii) How can the advantage provided by a control tutor be formally quantified and theoretically guaranteed?

A further meaningful contribution is found in [De Lellis, 2023b], which introduced a novel reward shaping technique that brings formal performance guarantees into the realm of reinforcement learning—an area where such assurances are notoriously rare. By evaluating the cumulative reward function, this approach enables principled assessments of policy quality, a critical advancement for deploying RL in safety-sensitive and mission-critical domains. However, this rigor comes at a cost: the reward signal becomes sparse, potentially hindering the agent's ability to efficiently learn optimal policies, especially in deep RL settings.

Objectives

The workplan and objectives are flexible and will be adapted depending on the inclination of the student and the results obtained in the early phases of the project.

O1. Develop and validate formally control-theoretical based methodologies to increase data efficiency in reinforcement learning.
O2. Develop practical control-theoretical based methodologies that provide stability and performance guarantees on the learned policy in reinforcement learning.
O3. Validate the developed strategies on the problem of agent navigation.

Methodology

The project will begin with a comprehensive and structured classification of existing approaches that integrate reinforcement learning (RL) with tools from control theory. This survey will highlight the strengths, limitations, and unresolved challenges of each method, setting a solid foundation for the contributions to follow.

Research focus will initially address Objective O1: to provide a formal, analytical validation of the improved data-efficiency introduced by *control tutors*—a concept already demonstrated numerically in [De Lellis, 2023a] across various settings. Our strategy will begin by analyzing how the tutor's action suggestions alter a uniform exploration policy and yield a new probability function. We will then extend the classical convergence proof of the Q-learning algorithm for discrete Markov decision processes to account for this non-uniform, tutor-influenced exploration. The investigation will explore tools such as stochastic approximation theory and convergence of Markov decision processes. This theoretical development will also allow us to quantify how the quality of a control tutor—measured by the optimality of its suggestions—impacts learning speed. Building on this foundation, we will explore the multi-agent scenario: heterogeneous robotic agents (e.g., land vs. aerial units, fast vs. slow movers) collaborating in an unstructured environment. We aim to determine whether the benefits of a single control tutor naturally generalize to settings with multiple tutors, or whether coordinated schemes—such as a leader tutor dynamically modulating others—are necessary to avoid interference or inefficiency.

To address Objective O2, the research will build on the reward shaping method introduced in [De Lellis, 2023b], which marked a significant step forward by providing formal performance guarantees in reinforcement learning—an area where such assurances are typically elusive. However, a drawback of this method is that it introduces *sparsity* into the reward signal: reward values can vary sharply across the state-action space, making it difficult for learning algorithms—especially those relying on function approximators such as deep neural networks—to generalize effectively. To overcome this limitation, the PhD student will explore alternative shaping strategies that maintain the theoretical guarantees while improving the reward signal's smoothness and continuity. Specifically, we will design shaping functions based on smooth, energy-like potential fields, which are expected to guide the learning process more gently and reliably than the discrete approach used in [De Lellis, 2023b]. In parallel, we will investigate the feasibility of combining *deterministic* guarantees (as provided by the original framework) with *probabilistic* ones, offering a broader and more flexible spectrum of safety and performance assurances. This analysis will possibly be based on

probabilistic convergence methods and a Bayesian approach to the characterization of system trajectories.

Finally, in the context of Objective O3, the combination will be investigated numerically of different techniques developed in the project, to obtain both reduced learning time and guarantees of performance. The best performing algorithms developed in the project will be validated on the task, for a ground mobile robot, of reaching a specified region, while avoiding unsafe regions, mimicking vehicle driving or robot navigation on an extraterrestrial planet.

Relevance to the MERC PhD program

Relevance and beneficiaries

The project is grounded in a strong methodological foundation, aiming to bridge control theory and reinforcement learning in a principled and impactful way. This integration is expected to yield two key advantages:

1. Significantly reduced learning times, and
2. Formal certification of properties of the learned policies.

Faster learning will help make reinforcement learning more accessible and scalable—enabling its application to complex tasks even in environments with limited computational resources, rather than relying solely on high-performance supercomputers. At the same time, the ability to certify policy properties marks a critical step toward the safe and trustworthy deployment of RL in high-stakes, real-world scenarios. This includes domains such as autonomous driving, search-and-rescue operations, and robotic exploration in extraterrestrial environments, where safety and reliability are non-negotiable. Together, these advancements will help push reinforcement learning from powerful yet opaque tools toward rigorously grounded and widely applicable technologies.

Relevance to the MERC PhD program

This project is highly interdisciplinary, sitting at the intersection of dynamical systems, control theory, and machine learning. Its aim is to develop both practical algorithms and theoretical advances that address complex control problems from a fresh, integrative perspective.

Skills

Throughout the project, the student will acquire a robust and versatile set of skills, supported by both hands-on supervision and independent study. These will include:

- Machine learning algorithms and methods, with a strong emphasis on reinforcement learning and its integration with control-theoretic tools,
- Dynamical systems and Markov decision processes, particularly with regard to stability and convergence analysis,
- Advanced computer programming skills, including proficiency with modern languages and cutting-edge machine learning libraries.

In addition to technical expertise, the student will receive targeted guidance to refine their scientific communication skills—including technical writing and oral presentations—and to develop the ability to critically and efficiently navigate the

scientific literature. These competencies will position the student for success in both academic and industry settings

References

Key references

- F. De Lellis, M. Coraggio, G. Russo, M. Musolesi, M. di Bernardo, "CT-DQN: control-tutored deep reinforcement learning," in Proceedings of the 5th Annual Learning for Dynamics and Control Conference, PMLR, pp. 941–953, 2023a,
- F. De Lellis, M. Coraggio, G. Russo, M. Musolesi, M. di Bernardo, "Guaranteeing control requirements via reward shaping in reinforcement learning." arXiv.2311.10026, 2023b.
- S. Gu, T. Lillicrap, I. Sutskever, S. Levine, "Continuous deep Q-learning with model-based acceleration," in International Conference on Machine Learning (ICML'16), pp. 2829–2838, 2016.
- M. Zanon and S. Gros, "Safe reinforcement learning using robust MPC," IEEE Transactions on Automatic Control, 66(8):3638–3652, 2021.

Additional references

- J. Degrave et al., "Magnetic control of tokamak plasmas through deep reinforcement learning," Nature, 602(7897):414–419, 2022.
- J. Jumper et al., "Highly accurate protein structure prediction with AlphaFold," Nature, 596(7873), 2021.
- E. Kaufmann, L. Bauersfeld, A. Loquercio, M. Müller, V. Koltun, and D. Scaramuzza, "Champion-level drone racing using deep reinforcement learning," Nature, 620(7976):982–987, 2023.
- J. Won, D. Gopinath, and J. Hodgins, "Control strategies for physically simulated characters performing two-player competitive sports," ACM Transaction on Graphics, 40(4):146:1–146:11, 2021.

Joint supervision arrangements

The student will meet at least weekly with at least one of the supervisors. The whole team will meet at least once every 1 or 2 months for a progress update.

Location and length of the study period abroad (min 12 months)

The student will be able to spend a research period (or research periods) at the lab of Mirco Musolesi at the University College London, or of a scientist with whom a collaboration is active.