ISBN: 978-3-95450-247-9 ISSN: 2673-5490 doi: 10.18429/JACoW-IPAC2024-TUPS60

TOWARDS FEW-SHOT REINFORCEMENT LEARNING IN PARTICLE ACCELERATOR CONTROL

S. Hirlaender *, L. Lamminger , Paris Lodron Universität Salzburg, Salzburg, Austria
S. Pochaba, Salzburg Research Forschungsgesellschaft mbH, Salzburg Austria
J. Kaiser, Deutsches Elektronen-Synchrotron DESY, Hamburg, Germany
C. Xu, A. Santamaria Garcia, L. Scomparin
Karlsruhe Institute of Technology (KIT), Karlsruhe, Germany
V. Kain, CERN, Geneva, Switzerland

Abstract

This paper addresses the automation of particle accelerator control through Reinforcement Learning (RL). It highlights the potential to increase reliable performance, especially in light of new diagnostic tools and the increasingly complex variable schedules of certain accelerators. We focus on the physics simulation of the AWAKE electron line, an ideal platform for performing in-depth studies that allow a clear distinction between the problem and the performance of different algorithmic approaches for accurate analysis. The main challenges are the lack of realistic simulations and partially observable environments. We show how effective results can be achieved through meta-reinforcement learning, where an agent is trained to quickly adapt to specific real-world scenarios based on prior training in a simulated environment with variable unknowns. When suitable simulations are lacking or too costly, a model-based method using Gaussian processes is used for direct training in a few shots only. This work opens new avenues for the implementation of control automation in particle accelerators, significantly increasing their efficiency and adaptability.

INTRODUCTION

Reinforcement learning presents significant potential for addressing control issues that surpass the capabilities of classical control theory. As a data-driven methodology, RL acquires knowledge through direct interaction with the systems it regulates. Despite its impressive real-world achievements, such as piloting drones with superior skill compared to human operators [1], RL faces several challenges that complicate its application in real-world scenarios. First, these algorithms typically require substantial amounts of data to achieve reliable performance. Secondly, there is an inherent trade-off between training stability and data efficiency, making it difficult to optimise both simultaneously. For particle accelerator control, leveraging the potential to enhance reliable performance is crucial, particularly with the advent of new diagnostic tools and increasingly complex variable schedules of some accelerators. Standard off-the-shelf algorithms may not suffice, necessitating the development of new strategies. We explore two innovative approaches to address some of these challenges: Meta-Reinforcement Learning (Meta-RL) and Model-based Reinforcement Learning (MBRL). These methods are evaluated using the AWAKE electron steering environment, which serves as an excellent benchmark due to its simplicity and non-trivial control task, yet still corresponds to a real, measurable system. All discussed approaches have been successfully implemented and tested in experiments on the actual machine.

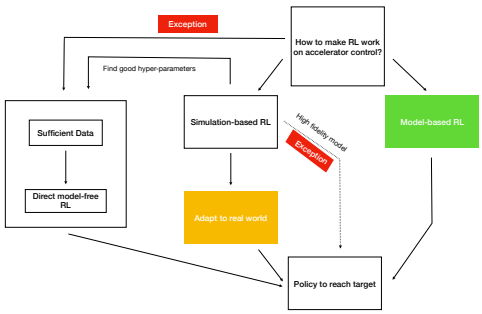

Figure 1: Overview of different approaches to train an RL algorithm in accelerator controls.

METHODICAL APPROACHES

Figure 1 illustrates various applications of reinforcement learning. It highlights that, when the system is accessible and adequate data is obtainable from the real system, direct Model-free RL (MFRL) can be employed using off-the-shelf algorithms such as on-policy trust region policy optimization [2] or sample-efficient off-policy algorithms like soft actor critic [3]. This is only possible in rare cases [4]. Simulations can be useful for determining optimal hyperparameters for the RL algorithms and for making decisions regarding the design of the function approximator before applying direct MFRL on the machine as done in [5, 6]. In some instances, simulations are both fast and precise enough to train the agent entirely in a simulated environment before real-world application [7]. However, often simulations do not perfectly model the real-world scenarios, presenting challenges in directly applying or retraining the agent on the actual machine. In such situations, Meta-RL is beneficial as it integrates prior knowledge from the simulations, ensuring stable adaptation to the real machine in just a few steps. In scenarios lacking even a simulation, MBRL can be advantageous. MBRL is noted for its extreme sample efficiency and the potential to solve tasks in just a few iterations. Nonetheless, this approach places significant computational demands on mak-

* simon.hirlaender@plus.ac.at

ISBN: 978-3-95450-247-9 ISSN: 2673-5490 doi: 10.18429/JACoW-IPAC2024-TUPS60

ing accurate inferences and performing online optimization, which may be a limiting factor.

Meta Reinforcement Learning

Meta-RL advances machine learning by developing algorithms that are adept at quickly adapting to new tasks, essentially embodying the concept of "learning to learn" in RL. Among the notable techniques is Model Agnostic Meta-Learning (MAML), which seeks an optimal initial model setting that can be rapidly adjusted to a diverse array of tasks with minimal modifications, leveraging gradient-based optimization to efficiently identify parameters conducive to quick adaptability [8]. Our focus on MAML stems from its broad utility and effectiveness across various tasks. Within this framework, RL tasks are treated as Markov Decision Processs (MDPs), with variations in tasks reflected through differences in initial states, dynamics, and rewards. The versatility of MAML permits customization to diverse problem types, optimizing learning by fine-tuning initial model parameters for enhanced performance and adaptability over conventional pre-training methods. Our implementation of MAML utilises an action-dependent baseline and a trust region method, which boost the efficiency and stability of the learning process [9].

Model-based Reinforcement Learning

Model-based Reinforcement Learning (MBRL) contrasts with model-free approaches by constructing an internal model of the world, which it uses to simulate interactions. This method enhances sample efficiency by reducing the need for direct system interaction. The process involves gathering data to refine the model and leveraging the model to improve the control policy. However, developing an effective policy can be challenging if the model is under-trained. The GP-MPC algorithm [10] applied in the AWAKE project, as described in [11] and [12], employs Gaussian Processs (GPs) to model system dynamics and quantify epistemic uncertainty—uncertainty due to limited data—thereby enhancing the model's robustness. The Model Predictive Control (MPC) aspect optimizes future actions based on these predictions, adapting to changes in the environment. This integration results in a highly sample-efficient algorithm, beneficial in scenarios where data is expensive or difficult to collect.

PROBLEM DEFINITION

The AWAKE electron line is an excellent environment for testing various algorithms [5, 12–15]. Initial RL agents were developed for trajectory optimization on the AWAKE electron line, aiming to match the efficiency of traditional Singular Value Decomposition (SVD) algorithms [16] used in control rooms. These agents guide the beam along a specified path to achieve critical parameters at the line's end for further processes. The electron production at AWAKE starts with a 5 MV RF gun that boosts electrons to 18 MeV, traveling through a 12 m beam line to the plasma cell. This path includes a vertical shift of 1 meter and a 60-degree bend to meet the proton beam at the plasma cell entrance. Beam trajectory is adjusted using 10 horizontal and 10 vertical steering dipoles, monitored by 10 Beam Position Monitors (BPMs) in each plane.

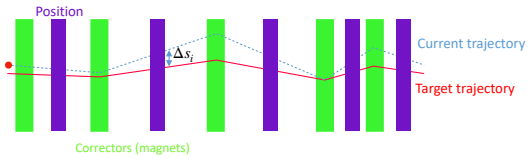

Figure 2: Visualization of a beam steering problem in the AWAKE electron line. Correctors are marked in green and are succeeded by BPMs, depicted in violet. The state vector s, consisting of components Δs_i for each BPM indexed by i, represents the distance to the target. The measured trajectory is shown as a dashed blue line, while the target trajectory is displayed in red.

Defining the Markov Decision Process

The electron transfer line and its various components are modeled using MAD-X [17], which simulates the transfer functions from field to current for different magnets at normalized strengths. Steering dipoles typically adjust trajectories by about 1 mrad per corrector. Using MAD-X, the response matrix, which shows BPM changes in relation to corrector adjustments, is computed. For the RL agent to operate effectively in the simulated environment, the observations s (BPM deviations from a reference trajectory, as depicted in Fig. 2) and actions a (adjustments to the dipole currents) need to align with the units and normalization of the actual equipment settings. The reward metric is the negative root mean square (RMS) of deviations from the target trajectory, defined as $r \propto -||s||$, as shown in Fig. 2. To increase the challenge of the control task, initial trajectories are purposefully set far from desired paths, and action amplitudes are limited. This approach ensures that resolving an episode is not a simple one-step process but requires a nontrivial control strategy. In instances where trajectory deviations become excessively large $||s||_{\max} >= 10$ mm, resulting in contact with the beam pipe, which possesses a diameter of 20 mm, the episode undergoes a reset. Subsequently, the point of impact on the wall and all ensuing measurements are assigned a value of 10 mm. This assignment is justified by the fact that the beam is effectively considered lost beyond the point of impact.

Distribution of MDP in the AWAKE Environment

To assess different scenarios of realisations of the environment we utilize to train agents across a diverse range of MDPs. We uniformly vary the quadrupole settings in the AWAKE setup by $\pm 25\%$ from standard values to capture possible variability and uncertainty of model. Figure 3 displays the diversity in the linear response matrices used in our experiments, with the "original task" serving as a baseline,

ISBN: 978-3-95450-247-9 ISSN: 2673-5490 doi: 10.18429/JACoW-IPAC2024-TUPS60

situated centrally in the distribution and based on actual measurements. We also present five varied MDP realizations to illustrate the scope of the task distribution, which remain fixed to assess the learning progress and the effectiveness of the algorithms under conditions that mimic real-world scenarios.

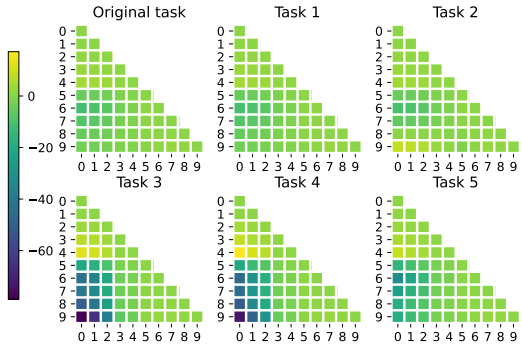

Figure 3: The response matrices of the different test tasks as variations of the real settings. Original task corresponds to measured values of AWAKE.

EXPERIMENTS

To evaluate the adaptability of the MAML and MBRL frameworks to varying MDPs, these approaches were evaluates on the previously mentioned six test tasks. In all plots average values over the test tasks are shown in solid lines with shaded areas indicating standard deviations. The Meta-RL training involved developing a meta-policy that could adapt to a broad range of system changes, followed by an adaptation phase as outlined previously. During testing, a policy gradient method fine-tuned the meta-policy, targeting the specific dynamics. Three scenarios were evaluated to determine how different starting conditions influence agent performance and adaptability, with results averaged across the six test tasks as shown in Fig. 4.

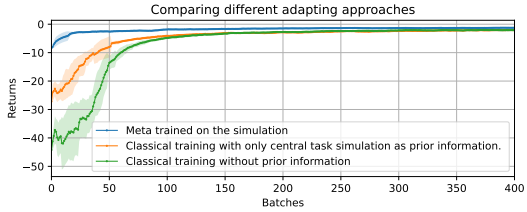

Figure 4: A comparative evaluation of adaptation strategies, underscoring the efficacy and efficiency of the meta-training approach.

1. **"Classical training with only central task as prior information"**: Begins with a pre-trained agent optimized for the central task.
2. **"Classical training without prior information"**: Starts with randomly selected weights, serving as a baseline.
3. **"Meta trained on the simulation"**: Uses weights refined through meta-training for enhanced adaptability.

The results clearly demonstrate the advantages of MAML, as it rapidly and stably adapts to various scenarios in just a few steps. The MBRL approach was tested on the six tasks with no prior training, taking only ten random steps to probe the system. Results show rapid learning of the control problem, as depicted in Fig. 5, which illustrates episode and cumulative lengths (top plot), and total rewards per episode (bottom plot), demonstrating the GP-MPC's ability to optimize its strategy across episodes for enhanced stability and efficiency in trajectory control within 20 steps.

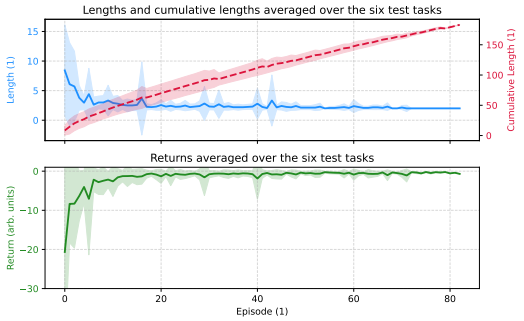

Figure 5: Evaluation of the MBRL approach without any prior training.

CONCLUSION AND OUTLOOK

In conclusion, this study has explored the significant potential of reinforcement learning for controlling particle accelerators, with a particular focus on the AWAKE project. Through our investigations, we have demonstrated that Meta-RL and MBRL approaches, especially those leveraging GP-MPC, provide frameworks for enhancing adaptability and efficiency in control automation. Our experiments confirm that Meta-RL facilitates rapid adaptation to new and varying conditions, markedly improving upon traditional methods that rely on extensive pre-training on a central task. Meanwhile, the GP-MPC algorithm stands out for its extreme sample efficiency, enabling effective control with minimal interaction, which is ideal in environments where data acquisition is challenging. Looking ahead, further improvements could include using residual models, which enhance adaptability by adjusting from simulated environments to real-world scenarios and develop more robustness approaches. Ultimately, the continued development and application of these advanced RL techniques will play a crucial role in the future of autonomous control systems for particle accelerators and similar complex systems.

ACKNOWLDEGEMENTS

The authors acknowledge support from DESY (Hamburg, Germany) and KIT (Karlsruhe, Germany), members of the Helmholtz Association HGF. This work has in part been funded by the IVF project InternLabs-0011 (HIR3X) and the Initiative and Networking Fund by the Helmholtz Association (Autonomous Accelerator, ZT-I-PF-5-6). We also gratefully acknowledge the support of the WISS 2025 project 'IDA-lab Salzburg' (20204-WISS/225/197-2019 and 20102-F1901166-KZP).

ISBN: 978-3-95450-247-9 ISSN: 2673-5490 doi: 10.18429/JACoW-IPAC2024-TUPS60

REFERENCES

[1] Elia Kaufmann, Leonard Bauersfeld, Antonio Loquercio, Matthias Müller, Vladlen Koltun, and Davide Scaramuzza, "Champion-level drone racing using deep reinforcement learning", *Nature*, vol. 620, no. 7976, pp. 982–987, Aug. 2023. doi:10.1038/s41586-023-06419-4

[2] Schulman John, Levine Sergey, Moritz Philipp, Jordan Michael, and Abbeel Pieter, "Trust region policy optimization", in *International Conference on Machine Learning*, 2015, pp. 1889–1897.

[3] Tuomas Haarnoja, Aurick Zhou, Pieter Abbeel, and Sergey Levine, "Soft actor-critic algorithms and applications", *arXiv preprint*, 2018. doi:10.48550/arXiv.1812.05905

[4] L. Scomparin *et al.*, "A low-latency feedback system for the control of horizontal betatron oscillations", in *Proc. IPAC'23*, Venice, Italy, Sep. 2023, pp. 4479–4482. doi:10.18429/JACoW-IPAC2023-THPL027

[5] Verena Kain *et al.*, "Sample-efficient reinforcement learning for CERN accelerator control", *Phys. Rev. Accel. Beams*, vol. 23, no. 12, p. 124801, Dec. 2020. doi:10.1103/PhysRevAccelBeams.23.124801

[6] Simon Hirlaender and Niky Bruchon, "Model-free and Bayesian Ensembling Model-based Deep Reinforcement Learning for Particle Accelerator Control Demonstrated on the FERMI FEL", *arXiv*, 2022. doi:10.48550/arXiv.2012.09737

[7] Jan Kaiser, Oliver Stein, and Annika Eichler, "Learning-based Optimisation of Particle Accelerators Under Partial Observability Without Real-World Training", in *Proceedings of the 39th International Conference on Machine Learning*, Baltimore, Maryland, USA, July 2022, pp. 10575–10585. https://proceedings.mlr.press/v162/kaiser22a.html

[8] Chelsea Finn, Pieter Abbeel, and Sergey Levine, "Model-Agnostic Meta-Learning for Fast Adaptation of Deep Networks", in *Proceedings of the 34th International Conference on Machine Learning*, Sydney, NSW, Australia, Aug. 2017, pp. 1126–1135. https://dl.acm.org/doi/10.5555/3305381.3305498

[9] John Schulman, Philipp Moritz, Sergey Levine, Michael Jordan, and Pieter Abbeel, "High-Dimensional Continuous Control Using Generalized Advantage Estimation", *arXiv*, 2018. doi:10.48550/arXiv.1506.02438

[10] Sanket Kamthe and Marc Peter Deisenroth, "Data-Efficient Reinforcement Learning with Probabilistic Model Predictive Control", *arXiv*, 2018. doi:10.48550/arXiv.1706.06491

[11] S. Hirlaender, L. Lamminger, G. Zevi Della Porta, and V. Kain, "Ultra fast reinforcement learning demonstrated at CERN AWAKE", in *Proc. IPAC'23*, Venice, Italy, Sep. 2023, pp. 4510–4513. doi:10.18429/JACoW-IPAC2023-THPL038

[12] Simon Hirlaender, Jan Kaiser, Chenran Xu, and Andrea Santamaria Garcia, "Tutorial at the RL4AA'24 Workshop", *Zenodo*, Mar. 2024. doi:10.5281/zenodo.10886639

[13] Lukas Lamminger, "Model Based Reinforcement Learning and Meta Reinforcement Learning for Accelerator Control at CERN", MA thesis, PLUS University Salzburg, 2023.

[14] Michael Schenk *et al.*, "Hybrid actor-critic algorithm for quantum reinforcement learning at CERN beam lines", *arXiv*, 2022. doi:10.48550/arXiv.2209.11044

[15] David Michalik, "A Model-based Optimal Control Approach for CERN's AWAKE Electron Line Trajectory Correction Problem", Master thesis, Aalborg University, 2021.

[16] Anthony A. Maciejewski and Charles A. Klein, "The Singular Value Decomposition: Computation and Applications to Robotics", *The International Journal of Robotics Research*, vol. 8, no. 6, pp. 63–79, 1989. doi:10.1177/027836498900800605

[17] *MAD-X Documentation and Source Code*. https://mad.web.cern.ch/mad