





Article

Advancing Beam Steering Control: A Comparative Study of Reinforcement Learning and Model Predictive Techniques on CERN AWAKE

Olga Mironova ^{1,†}  0009-0004-3402-8914, Thomas Gallien ^{2,‡}  0000-0003-3331-5917, Lorenz Fischl ^{3,‡}  0000-0002-7893-0641, Simon Hirlaender ^{4,†,‡}  0000-0002-2634-3437

¹ Affiliation 1; e-mail@e-mail.com

² Affiliation 2; e-mail@e-mail.com

* Correspondence: e-mail@e-mail.com; Tel.: (optional; include country code; if there are multiple corresponding authors, add author initials) +xx-xxxx-xxx-xxxx (F.L.)

† Current address: Affiliation.

‡ These authors contributed equally to this work.

Abstract: This paper investigates advanced control strategies for beam steering in the electron line of the AWAKE experiment at CERN. We employ a highly accurate physics-simulation and conduct an in-depth comparison of various control approaches. Model Predictive Control (MPC) utilizes a priori knowledge of the system in the form of a model and is effective with accurate models. Classical analytical inverse control methods use the inverted control matrices for computing control actions, offering straightforward implementation but a limited adaptability to changes. Deep reinforcement learning (RL) algorithms, specifically Proximal Policy Optimization (PPO), do not require explicit system models and can adapt to non-linearities and uncertainties. Finally, model-based RL using Gaussian Processes combined with MPC (GP-MPC) integrates GP regression for learning the system's dynamics with MPC for control. This approach accounts for model uncertainties and non-linearities, providing a probabilistic framework that enhances robustness and adaptability. Our study examines the sensitivities of these control strategies within linear continuous Markov Decision Processes. Although the underlying MDP is linear, the problem introduces slight nonlinearities due to limited actions and the termination criterion. Our analysis involves measurement noise, deviations toward nonlinear dynamics, and nonstationary. Through extensive simulations, we evaluate each method's performance under these challenging conditions. Our findings highlight the potential of RL techniques, particularly those incorporating probabilistic modelling and planning, for real-world accelerator control. This work offers valuable insights into the application of non-linear control methods and reinforcement learning to complex, high-dimensional systems.

Keywords: Reinforcement Learning, Control theory, Gaussian Data-driven Model Predictive Control (GP-MPC)

Citation: Lastname, F.; Lastname, F.; Lastname, F. Title. *Journal Not Specified* **2024**, *1*, 0. <https://doi.org/>

Received:

Revised:

Accepted:

Published:

Copyright: © 2025 by the authors. Submitted to *Journal Not Specified* for possible open access publication under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

0. How to Use this Template

- @Olga: Understand the tutorial in detail, especially the code!
- @Olga: Formulate possible case studies
- @Simon: Guide to the overall process

The template details the sections that can be used in a manuscript. Note that the order and names of article sections may differ from the requirements of the journal (e.g., the positioning of the Materials and Methods section). Please check the instructions on the authors' page of the journal to verify the correct order and names. For any questions, please contact the editorial office of the journal or support@mdpi.com. For LaTeX-related questions please contact latex@mdpi.com.

1. Introduction

The CERN accelerator complex encompasses a diverse array of normal-conducting and superconducting linear and circular accelerators, employing both conventional and advanced acceleration techniques (see Fig. 1)[1?]. The AWAKE (Advanced Proton Driven Plasma Wakefield Acceleration Experiment) aims to utilize high-energy protons from CERN's Super Proton Synchrotron (SPS) to generate plasma wakefields, which serve as a medium to accelerate injected electron bunches to high energies. Effective trajectory control is critical to aligning the electron beam precisely with the plasma channel for optimal acceleration performance. Recent advancements in numerical optimization methods, often augmented with machine learning, have driven progress in tasks such as automated device alignment and parameter optimization in free-electron lasers (FELs) [?]. Reinforcement learning (RL) has emerged as a promising approach to further enhance efficiency in such optimization tasks by minimizing the exploration time needed for solution convergence. Traditional control strategies like Model Predictive Control (MPC) rely on predefined system models, effectively solving sequential decision-making problems when the model is accurate. Analytical inverse control methods, which compute control actions using inverted control matrices, offer simplicity but are less adaptable to dynamic system changes or model inaccuracies. Reinforcement learning, particularly model-free deep RL algorithms such as Proximal Policy Optimization (PPO) and Trust Region Policy Optimization (TRPO), does not require explicit system models, making it well-suited for environments with non-linearities and uncertainties. Meanwhile, model-based RL methods, such as those employing Gaussian Process-based MPC (GP-MPC), combine Gaussian Process regression for dynamic system learning with MPC for control. This hybrid approach integrates probabilistic modeling to account for uncertainties and non-linear behavior, enhancing robustness and adaptability. However, RL's effectiveness is often constrained by challenges such as the availability of sufficient instrumentation for meaningful state observations and the need for high sample efficiency, as RL algorithms can require substantial interaction data for training. These constraints limit the applicability of RL in certain tasks and its deployment in accelerator control rooms. Our study focuses on investigating linear continuous Markov Decision Processes (MDPs) in the context of beam steering control. We address the challenges of applying various control strategies and explore their implementation in dynamic accelerator environments. Through extensive simulations, we compare the performance of RL agents with MPC and analytical control approaches, evaluating metrics such as reward accumulation, state deviations, and action efficiency. This paper is structured as follows: we begin with an overview of reinforcement learning in beam steering control, outlining key concepts, challenges, and algorithmic details. The testing section evaluates the sensitivities of different control strategies under scenarios involving measurement noise, deviations towards non-linear dynamics, and non-stationary behavior. Extensive simulations assess the performance of each method under these challenging conditions. In the discussion, we analyze the results in the context of the encountered challenges and lessons learned. Finally, the outlook highlights potential future directions and broader applications of the approaches explored in this study.

2. Problem Setting and Preliminaries

2.1. Problem set-up

Here you describe the real problem. Use the paper [?] and try to understand all details of the problem. What is the basic physics? How do we simulate the problem? What are the limitations? Read about beam steering in general. What is a dipole magnet, what is a quadrupole magnet? What is a BPM? Add citations. Try to think about different interesting scenarios you want to probe.

3. Beam Steering Problem in the AWAKE Electron Line

The practical application of RL in accelerator control presents several challenges alongside its promising potential. These challenges include low sampling efficiency, safety constraints, system non-stationarity, and state space observability. Recent researches have demonstrated the successful application of various RL approaches to address some of these issues [1–4], highlighting the feasibility of implementing RL for accelerator control tasks. The benchmark (see Fig. 1) for these studies was the electron line of the AWAKE experiment, where accurate simulations of the electron beam were used to test optimization and control algorithms. The objective in this episodic task is to minimize the distance between the initial beam trajectory and a target trajectory as fast as possible.

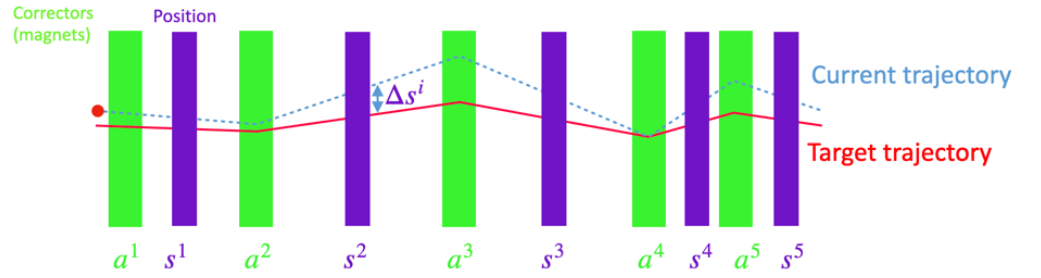


Figure 1. Illustration of a beam steering problem as in the AWAKE electron line

In this environment, ten dipole magnets are used to steer the beam, defining the control actions a , while ten beam position monitors measure the trajectory, forming the state s . The reward r is defined as the negative root mean square (RMS) value of the distance to the target trajectory. The episode is deemed successful if the RMS distance falls below a threshold of -1 mm. Conversely, the episode ends unsuccessfully if the beam hits the wall, defined as any state where the position is ≤ -1 cm or ≥ 1 cm.

Each episode is initialized such that the RMS of the distance to the target trajectory lies between 0.7 cm and 0.8 cm. All states and actions are normalized to the range $[-1, 1]$, ensuring that the task is sufficiently challenging and operates close to safety boundaries to test the robustness of safety constraints.

The electron beamline of the AWAKE experiment provides an appropriate environment for the application of optimization algorithms to address the control problem of steering the beam along a specified target trajectory. This setting offers a unique opportunity to leverage advanced optimization techniques due to several constraints which make this problem complex: actions are bounded to satisfy physical limitations for the safety, and the successful termination below threshold.

3.1. Mathematical description

Mathematical framework Markov Decision Process Describe all details here. What is an MDP, how is our MDP designed? What are critical aspects. How does the dynamics work....

3.2. Mathematical Description

The beam steering problem can be presented as a Markov Decision Process (MDP), a framework defined by the tuple $(\mathcal{S}, \mathcal{A}, \mathcal{P}, \mathcal{R}, \gamma, \rho)$, where \mathcal{S} is the state space, \mathcal{A} is the action space, $\mathcal{P}(s_{t+1} | s_t, a_t)$ is the state transition probability, $\mathcal{R}(s_t, a_t)$ is the reward function, and the discount factor $\gamma \in [0, 1]$ is set to 1 in our case, indicating that future rewards are valued equally to immediate rewards, which can always be done in episodic scenarios. In this problem, the state $\mathbf{s}_t \in \mathbb{R}^N$ represents the deviations of the electron beam's trajectory from the desired target at time t , where N denotes the number of degrees of freedom in the beamline. The action $\mathbf{a}_t \in \mathbb{R}^N$ corresponds to the adjustments applied to the beamline magnets, constrained to $a_{i,t} \in [a_{\min}, a_{\max}]$. The reward function is defined

as $\mathcal{R}(\mathbf{s}_t) = -\sqrt{\frac{1}{N} \sum_{i=1}^N s_{i,t}^2}$, encouraging minimization of the trajectory deviations. The system dynamics, which are linear time-invariant, are characterized by $\mathbf{s}_{t+1} = \mathbf{B}\mathbf{a}_t + \mathbf{I}\mathbf{s}_t +$ noise term, where \mathbf{B} is the response matrix mapping actions to state changes, and \mathbf{I} is the identity matrix. The environment operates in an episodic manner, where each episode consists of a sequence of interactions (changes in \mathbf{a}_t) until termination criteria are met. Episodes terminate under one of the following criteria: (1) truncation after a maximal number of interactions, (2) successful termination when the root mean square (RMS) of the states falls below the measurement uncertainty, or (3) unsuccessful termination if any state exceeds the beam pipe boundaries. Actions $\mathbf{a}_t \in \mathcal{A}$ are bounded to satisfy physical constraints, ensuring safety. The successful termination condition, involving measurements below a threshold, introduces non-linearities and renders the problem non-trivial. This formulation, incorporating high-dimensional state and action spaces, physical constraints, and stochastic elements, provides a comprehensive yet tractable representation of the beam steering problem, supporting advanced control strategies such as Model Predictive Control (MPC) and Reinforcement Learning (RL). The components are defined as follows:

State Vector (s): Represents the beam positions measured at beam position monitors (BPMs):

$$\mathbf{s}_t = \begin{bmatrix} s_1 \\ s_2 \\ s_3 \\ \vdots \\ s_N \end{bmatrix}.$$

Action Vector (a): Represents the adjustments applied to the dipole magnets:

$$\mathbf{a}_t = \begin{bmatrix} a_1 \\ a_2 \\ a_3 \\ \vdots \\ a_N \end{bmatrix}.$$

Response Matrix (B): Describes the influence of actions on states:

$$\mathbf{B} = \begin{bmatrix} B_{11} & 0 & \cdots & 0 \\ B_{21} & B_{22} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ B_{N1} & B_{N2} & \cdots & B_{NN} \end{bmatrix}.$$

Identity Matrix (I): Represents the persistence of the current state:

$$\mathbf{I} = \begin{bmatrix} 1 & 0 & \cdots & 0 \\ 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & 1 \end{bmatrix}.$$

4. Methods

Describe all approaches in detail with references and some overview of the advantages and disadvantages wrt our problem. Describe the experiments we want to conduct and why these test are done!

4.1. Analytical Approach

There is the SVD method in beam steering which is golden standart. Look for references and try to understand our special case. The analytical approach is a control strategy

that employs a mathematically rigorous framework based on classical control theory to solve beam steering problems. The foundation of this method lies in the response matrix \mathbf{B} , which encapsulates the linear dynamics of the system. This matrix represents how control actions, \mathbf{a}_t , influence the state transitions of the system, \mathbf{s}_{t+1} . For linear time-invariant systems, the dynamics can be expressed as:

$$\mathbf{s}_{t+1} = \mathbf{B}\mathbf{a}_t + \mathbf{I}\mathbf{s}_t + \varepsilon_t \quad (1)$$

where \mathbf{I} is the identity matrix and the noise term ε_t accounts for system uncertainty and is a random variable, which is usually normal distributed with a mean $\mu = 0$ and a variance σ . Using this relationship, the Analytical Approach determines control actions through the inverse of the response matrix:

$$\mathbf{a}_t = \mathbf{B}^{-1}\mathbf{s}_t. \quad (2)$$

This direct computation provides an explicit solution for the actions required to bring the system to a desired state. As we can see in Equation (1) the solution is simple. The key advantage of the Analytical Approach is its computational efficiency. By leveraging the inverse of the response matrix, this method avoids iterative optimization, making it significantly faster than other approaches such as reinforcement learning or nonlinear optimization. Additionally, its deterministic nature ensures that the solution is stable and reproducible under the assumption of accurate system modeling. However, the method has notable limitations. First, the assumption of linearity restricts its applicability to systems where nonlinear effects are minimal. In beam steering, for example, deviations due to noise, beam misalignments, or unmodeled disturbances may introduce nonlinearities, reducing the accuracy of the Analytical Approach. Second, the calculation of the inverse matrix \mathbf{B}^{-1} requires that \mathbf{B} be non-singular and well-conditioned, which may not always hold in practice. Poor conditioning of \mathbf{B} can lead to numerical instabilities and unreliable control actions. In summary, the Analytical Approach is well-suited for systems where the response matrix accurately models dynamics, offering computational speed and simplicity. However, its reliance on linearity and the invertibility of \mathbf{B} can limit its performance in more complex or highly perturbed environments. Addressing these limitations may require combining this method with adaptive or robust control techniques to enhance reliability under real-world conditions.

4.2. Model Predictive Control (MPC) Approach

Please bring this into context of the previous notation and mathematical definitions. Model Predictive Control (MPC) is a robust and adaptive control method that calculates control actions by solving an optimization problem over a finite prediction horizon. This approach relies on an internal model of the system dynamics to predict future states and determine the optimal sequence of control actions. The MPC process can be broken into three steps: Predict, Optimize, and Implement.

First, MPC uses the system's model to predict the future states of the beam based on current measurements and anticipated control actions. This prediction is performed over a specified prediction horizon, which is a finite number of steps into the future. Second, it formulates an optimization problem to minimize a predefined cost function, such as the difference between the predicted beam position and the desired trajectory (reference trajectory), while satisfying system constraints. The optimization problem is typically expressed as:

$$\max_{\{\mathbf{a}_t\}} \mathbb{E}_{W_t} \left[\sum_{t=0}^{H-1} R(S_t, A_t, W_t) + V(S_H) \right],$$

where $R(S_t, A_t, W_t)$ is the reward function at time t , $V(S_H)$ represents the terminal cost at the end of the horizon, and H is the prediction horizon. The constraints ensure that the control actions \mathbf{a}_t remain within physical safety limits and system dynamics,

$S_{t+1} = f(S_t, A_t, W_t)$, are respected. Finally, MPC implements only the first control action from the optimized sequence, then repeats the process with updated measurements and predictions in a receding horizon manner. MPC provides significant benefits for beam steering applications. Firstly, it enables real-time adjustments by continuously updating predictions and control actions, making it highly responsive to environmental disturbances or dynamic changes in the system. Secondly, MPC effectively handles constraints, such as the physical limits on actuator movements or safety thresholds for beam angles, ensuring system integrity. Thirdly, its predictive nature improves precision by proactively correcting deviations from the desired trajectory before they become significant. This capability is particularly advantageous in maintaining accurate beam alignment in scenarios requiring high precision. Despite its strengths, MPC presents certain challenges. A major limitation is its computational complexity, as real-time optimization can be demanding for fast-moving systems like beam steering. High computational requirements may necessitate advanced hardware or specialized algorithms to ensure timely execution. Another drawback is the heavy reliance on an accurate system model. Any discrepancies between the model and the actual system dynamics can degrade the control performance. Furthermore, tuning the MPC parameters, such as the prediction horizon or cost weights, is non-trivial and requires expertise in both control theory and the specifics of the beam steering system. MPC is a powerful control strategy for beam steering, balancing adaptability, precision, and constraint handling. It is especially effective in dynamic environments where real-time control and predictive capabilities are crucial. While its computational demands and dependency on accurate modeling are notable challenges, advances in computational power and modeling techniques continue to make MPC increasingly feasible for real-world applications. The figure below illustrates the core concepts of MPC, highlighting the prediction horizon and the iterative optimization process:

5. Experiments

- 5.1. Baseline
- 5.2. Results and Discussions
- 5.3. Ablation Studies

6. Discussion

Authors should discuss the results and how they can be interpreted from the perspective of previous studies and of the working hypotheses. The findings and their implications should be discussed in the broadest context possible. Future research directions may also be highlighted.

7. Conclusions

This section is not mandatory, but can be added to the manuscript if the discussion is unusually long or complex.

Funding: Please add: “This research received no external funding” or “This research was funded by NAME OF FUNDER grant number XXX.” and and “The APC was funded by XXX”. Check carefully that the details given are accurate and use the standard spelling of funding agency names at <https://search.crossref.org/funding>, any errors may affect your future funding.

Institutional Review Board Statement: In this section, you should add the Institutional Review Board Statement and approval number, if relevant to your study. You might choose to exclude this statement if the study did not require ethical approval. Please note that the Editorial Office might ask you for further information. Please add “The study was conducted in accordance with the Declaration of Helsinki, and approved by the Institutional Review Board (or Ethics Committee) of NAME OF INSTITUTE (protocol code XXX and date of approval).” for studies involving humans. OR “The animal study protocol was approved by the Institutional Review Board (or Ethics Committee) of NAME OF INSTITUTE (protocol code XXX and date of approval).” for studies involving animals. OR “Ethical review and approval were waived for this study due to REASON (please provide a detailed justification).” OR “Not applicable” for studies not involving humans or animals.

Informed Consent Statement: Any research article describing a study involving humans should contain this statement. Please add “Informed consent was obtained from all subjects involved in the study.” OR “Patient consent was waived due to REASON (please provide a detailed justification).” OR “Not applicable” for studies not involving humans. You might also choose to exclude this statement if the study did not involve humans.

Written informed consent for publication must be obtained from participating patients who can be identified (including by the patients themselves). Please state “Written informed consent has been obtained from the patient(s) to publish this paper” if applicable.

Data Availability Statement: We encourage all authors of articles published in MDPI journals to share their research data. In this section, please provide details regarding where data supporting reported results can be found, including links to publicly archived datasets analyzed or generated during the study. Where no new data were created, or where data is unavailable due to privacy or ethical restrictions, a statement is still required. Suggested Data Availability Statements are available in section “MDPI Research Data Policies” at <https://www.mdpi.com/ethics>.

Acknowledgments: In this section you can acknowledge any support given which is not covered by the author contribution or funding sections. This may include administrative and technical support, or donations in kind (e.g., materials used for experiments).

Abbreviations

The following abbreviations are used in this manuscript:

- MDPI Multidisciplinary Digital Publishing Institute
- DOAJ Directory of open access journals
- TLA Three letter acronym
- LD Linear dichroism

Appendix A

Appendix A.1

The appendix is an optional section that can contain details and data supplemental to the main text—for example, explanations of experimental details that would disrupt the flow of the main text but nonetheless remain crucial to understanding and reproducing the research shown; figures of replicates for experiments of which representative data are shown in the main text can be added here if brief, or as Supplementary Data. Mathematical proofs of results not central to the paper can be added as an appendix.

Table A1. This is a table caption.

| Title 1 | Title 2 | Title 3 |
|---------|---------|---------|
| Entry 1 | Data | Data |
| Entry 2 | Data | Data |

Appendix B

All appendix sections must be cited in the main text. In the appendices, Figures, Tables, etc. should be labeled, starting with “A”—e.g., Figure A1, Figure A2, etc.

1. Scheinker, A.; Hirilaender, S.; Velotti, F.M.; Gessner, S.; Della Porta, G.Z.; Kain, V.; Goddard, B.; Ramjiawan, R. Online multi-objective particle accelerator optimization of the AWAKE electron beam line for simultaneous emittance and orbit control. *AIP Advances* **2020**, *10*, 055320, [https://pubs.aip.org/aip/adv/article-pdf/doi/10.1063/5.0003423/12846285/055320_1_online.pdf].
<https://doi.org/10.1063/5.0003423>.

2. Hirilaender, S.; Lamminger, L.; Zevi Della Porta, G.; Kain, V. Ultra fast reinforcement learning demonstrated at CERN AWAKE. *JACoW IPAC 2023*, 2023, THPL038. <https://doi.org/10.18429/JACoW-IPAC2023-THPL038>.

3. Kain, V.; Hirilaender, S.; Goddard, B.; Velotti, F.M.; Della Porta, G.Z.; Bruchon, N.; Valentino, G. Sample-efficient reinforcement learning for CERN accelerator control. *Phys. Rev. Accel. Beams* **2020**, *23*, 124801. <https://doi.org/10.1103/PhysRevAccelBeams.23.124801>.

4. Kain, V.; Bruchon, N.; Hirlander, S.; Madysa, N.; Vojskovic, I.; Skowronski, P.K.; Valentino, G. TEST OF MACHINE LEARNING AT THE CERN LINAC4. 2022.

288

289

Disclaimer/Publisher’s Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.

290

291

292