ISBN: 978-3-95450-248-6 ISSN: 2673-5490 doi: 10.18429/JACoW-IPAC2025-THPM113

TRAJECTORY STEERING FOR DC BEAMS AT THE CERN SPS USING REINFORCEMENT LEARNING BASED ON INTENSITY MEASUREMENTS

A. Menor de Oñate*, N. Bruchon, V. Kain, M. Schenk, B. Rodriguez Mateos
CERN, Geneva, Switzerland

Abstract

The slow extracted beams at the CERN Super Proton Synchrotron (SPS) are transported over several 100 m long transfer lines to three targets in the North Area Experimental Hall. The experiments require eliminating intensity fluctuations over the roughly 5 s particle spill and hence to debunch the extracted beams. In this environment, secondary emission monitors (SEMs) have to replace the conventional beam position monitoring systems that rely on RF structure. The intensity difference on split secondary emission foils can be used to indicate beam positions. Traditional trajectory correction algorithms fail however when the beam ends up on one foil only. This paper summarises successful first tests with reinforcement learning (RL) to learn to correct the trajectory based on foil intensity measurements. The RL agents were trained in simulation and then successfully transferred to the real environment. Results of the application of the trained RL agents for target steering will be presented.

INTRODUCTION

The third-order resonant extraction from long straight section 2 (LSS2) at the Super Proton Synchrotron (SPS) at CERN serves the fixed target experiments in the SPS North Area with protons or heavy ions. Figure 1 shows an overview of the transfer lines and three primary beam targets. The particles are extracted for several seconds at an ideally constant extracted intensity rate without intensity fluctuation. The 200 MHz from the main SPS RF system is removed by debunching the beam. The direct current (DC) spills do not allow for conventional beam position monitor systems in the transfer lines as input to trajectory correction algorithms. The intensity readings from secondary emission monitors (SEMs) in the form of split foils are used as a proxy for beam position by measuring the intensity difference between the left and right foil for the horizontal position monitoring or upper and lower foil in case of vertical foils [1]. Standard emittance values and nominal optics are assumed. With these types of monitors, the lines can be successfully corrected with classical trajectory correction algorithms as long as the beams are already relatively well aligned. The conversion to beam position breaks down, however, if the beam ends up on one foil only at a given SEM location. Steering then becomes guesswork. This is the reason why the operations crew spends a significant amount of time (up to hours per line) every year to establish well-steered lines and beams that are well-aligned on the target.

* adrian.menor.de.onate@cern.ch

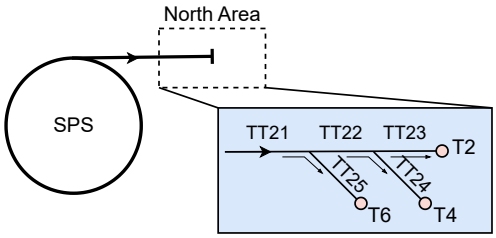

Figure 1: SPS North Area transfer lines.

Data-driven optimization and control of particle accelerators has made steady progress in recent years, with many successful applications deployed at CERN [2, 3]. This paper summarises recent results with reinforcement learning (RL) for the steering problem described above. RL is a machine learning technique that learns a control policy by interacting with an environment (the dipole corrector settings of the transfer lines in this case) and receiving a feedback signal (the reward) based on a performance metric of interest. The final goal of this project is to fully automate beam commissioning in the North Area transfer lines.

THE SIMULATION ENVIRONMENT

One of the limitations of applying RL in real-world scenarios is its limited sample efficiency during training. It can require thousands of iterations to reach convergence. In our case sufficiently accurate transfer line models are available and hence the approach of offline training and simulation-to-real transfer was selected instead of online training.

The TT20 primary beam transfer lines to the North Area targets are modelled in MAD-X [4]. The RL steering policies were trained by simulating the trajectories for the dipole corrector settings defined by the agent and translating those to intensity differences at the various SEM locations. Gaussian beams, nominal optics, and a normalised emittance of 2 μm rad in both planes were used. In reality, the emittance depends on the beam intensity. Due to the nature of the beam transfer with slow extraction in the horizontal plane and splitting the beams vertically at two locations to balance beam sharing to the targets (see Fig. 1), the beam will not necessarily be Gaussian everywhere. Tests in simulation with different distributions and emittances have, however, shown that RL training with Gaussian beams and very small emittances generalises well enough to all other scenarios.

ISBN: 978-3-95450-248-6 ISSN: 2673-5490 doi: 10.18429/JACoW-IPAC2025-THPM113

Following the layout of the transfer lines in Fig. 1, with splitter magnets that allow three targets to be served simultaneously, 9 policies were trained for the various line segments. Note that this paper only discusses the results for the final parts of the lines, right in front of the targets – TT23, TT24 and TT25. The trained RL agents were used for what is usually called "target steering" in the control room. Testing threading capabilities for all of TT20 will be tested later in the year with the heavy ion run.

IMPLEMENTATION DETAILS

The control problem in RL is described as a Markov Decision Process in the form of a so-called *environment*. The RL algorithm interacts with this environment – its policy sets so-called actions in the environment and receives state information and reward signals from the environment. The metric of interest in the context of SEM foils is symmetry S (see Eq. 1, where I_1 and I_2 are the intensities on foil 1 and 2, respectively) per monitor. The symmetry is 1 if the beam is centered and 0 if on one foil only.

$$S = \sqrt{1 - \frac{|I_1 - I_2|}{I_1 + I_2}}. \tag{1}$$

RL follows an episodic training. For each episode, the agent needs to solve a new problem, starting from random initial settings of the corrector magnets. The episodes have a maximum length (50 iterations in our case) and are terminated if the target is reached, or truncated otherwise.

The required input states for our steering problem are the normalised intensity differences between the foils for each SEM monitor (ΔI) as well as a flag I_{on} indicating whether the SEM saw intensity above a certain threshold. This resolves the ambiguity wherein both a perfectly centered beam and a completely off-centered beam yield identical measurements of $\Delta I = 0$. I_{on} was scaled to ± 0.1 instead of $[0, 1]$ to be in a similar range as ΔI and allow the policy to converge. As explained below the input state also needs a so-called "goal" vector. The policy π takes the state consisting of ΔI and I_{on} at the different monitors as well as the goal G as input and produces the actions corresponding to relative steering angle changes at the different dipole corrector magnets, i.e. $\pi : (\Delta I \times I_{\text{on}} \times G) \rightarrow \Delta\theta$.

The reward function is defined via the resulting ΔI (rather than symmetries directly) and I_{on} flags:

$$\text{Reward} = -\sqrt{\frac{1}{n}\sum_{i=1}^{n}(\Delta I_i)^2} + 1.1 \cdot \frac{\sum_{j=1}^{n} I_{\text{on}_j}}{n}, \tag{2}$$

where n is the total number of monitors. The transfer line is considered to be steered if $S \geq 0.85$. The reward threshold for terminating an episode during training was set to -0.01.

The TD3 [5] RL algorithm was used with Hindsight Experience Replay (HER) [6] to enhance the experience buffer. This combination increases learning robustness and avoids the need to explicitly create a curriculum learning scheme for the different transfer lines, which was otherwise necessary.

Integrating TD3 with HER reformulates the RL problem as a goal-conditioned task. Specifically, the inputs to both the policy and value networks are augmented to include a goal vector g, which, in this context, consists of target values for the observables: ΔI and I_{on}. Ultimately, our goal is that $\Delta I = 0$ and $I_{\text{on}} = 0.1$ on all SEMs.

The reward function is correspondingly modified to support HER. In particular, the buffer is relabeled according to alternative goals g', and the reward includes a term $||g - g'||$, where g is the achieved goal at a given timestep in the episode and g' the desired goal, as prescribed by the HER framework. This approach enables the agent to extract informative learning signals from episodes that do not reach the intended steering goal by treating the final achieved state as a substitute goal g' thereby converting unsuccessful episodes into useful training data.

Both the policy and value functions are modeled using multi-layer perceptrons. The algorithm uses the implementation of `stable baselines 3` [7].

RESULTS

The RL policies were tested both in simulation and on the real accelerator during the 2025 commissioning. The simulation results are used to assess the expected performance of the RL policies, while experiments on the real machine demonstrate the viability of this approach in practice. Table 1 summarises the parameters used for training and evaluation.

Table 1: Simulation parameters for the various lines. At the beginning of each episode, the initial corrector settings are sampled in the range indicated in the column of *Corrector initialisation*. Max $\Delta\theta$ corresponds to the maximum corrector step per iteration. In this first implementation, all correctors for a given line have the same maximum allowed setting change.

Line		Corrector initialisation θ_0	Max $\Delta\theta$
TT20	H	$\mathcal{U}(-50, 50)$ µrad	10 µrad
TT21	H	$\mathcal{U}(-100, 100)$ µrad	10 µrad
	V	$\mathcal{U}(-100, 100)$ µrad	10 µrad
TT23	H	$\mathcal{U}(-200, 200)$ µrad	20 µrad
	V	$\mathcal{U}(-200, 200)$ µrad	20 µrad
TT24	H	$\mathcal{U}(-200, 200)$ µrad	20 µrad
	V	$\mathcal{U}(-200, 200)$ µrad	20 µrad
TT25	H	$\mathcal{U}(-150, 150)$ µrad	20 µrad
	V	$\mathcal{U}(-320, 320)$ µrad	20 µrad

Figure 2 shows the number of iterations the trained policies took on 40 evaluation episodes to steer the lines to better than the threshold. Typically less than 30 iterations are necessary, where most took less than 15, suggesting a significant time reduction compared to manual steering; i.e. with beam every ≈ 30 s, steering one particular line would take a maximum of 15 minutes.

ISBN: 978-3-95450-248-6 ISSN: 2673-5490 doi: 10.18429/JACoW-IPAC2025-THPM113

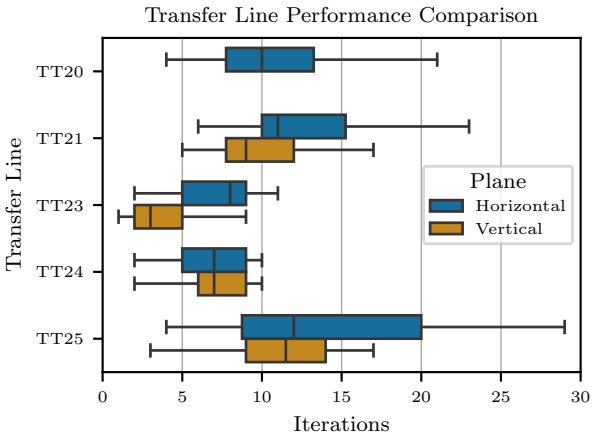

Figure 2: Distribution of the number of iterations to steer to better than the reward threshold for the different policies in simulation. Each policy is evaluated 40 times.

Results from a Real Environment

As part of the 2025 commissioning, time was allocated to test the trained RL agents on the real hardware with beam extracted to the SPS North Area targets. As the control system of the SPS works with high-level parameters (i.e. steering angles for dipole corrector magnets instead of currents), the transfer from simulation to real-world was almost seamless. Only the polarity conventions for the SEM foils needed attention. Figure 3 summarises the successful outcome of target steering for all three targets in the horizontal and vertical plane (only TT24H could not be tested due to lack of time). Note that the initial symmetries were particularly low for some of the cases. The number of iterations needed to reach termination is in a similar range as in simulation.

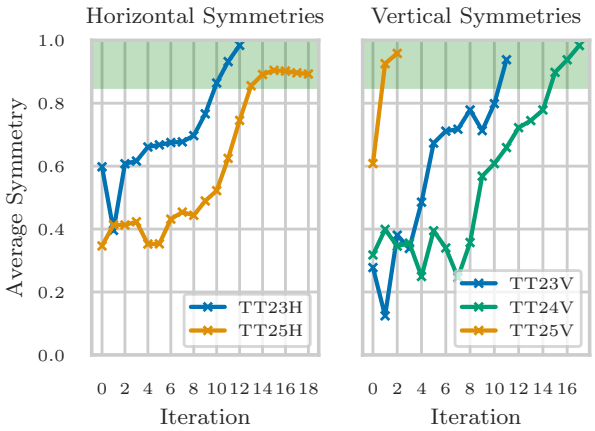

Figure 3: Evolution of the transfer line steering with the trained RL agents on the real hardware for the three different targets. The RL agents successfully steered the beam on target, reaching final symmetries that well surpass $S \geq 0.85$. TT24H could not be tested due to lack of time.

This initial steering at the beginning of the commissioning phase was carried out with relatively low intensity ($\sim 5{\times}10^{11}$ p) and the noise on the SEMs was problematic (i.e. even if the beam was on one foil only, the resulting symmetries were close to $0.2 - 0.4$). But even under these conditions, the agents proved to be robust.

CONCLUSION AND OUTLOOK

The trajectory of the DC beams in the several 100 m long transfer lines between the SPS extraction point and North Area targets is measured with secondary emission monitors arranged as split foils. Inferring beam position from these split foils breaks down when the beam ends up on one foil only and classical trajectory correction algorithms do not readily apply anymore. This paper proposed to train reinforcement learning agents for trajectory correction for the different line segments that only rely on intensity measurements at the various monitors. The agents were fully trained in simulation and then transferred to the real environment. During 2025 commissioning the agents could be tested on the task of target steering whereby the angle and position of the beam on target were optimised. The test of the RL agents was successful and now paves the way for beam threading along the entire transfer line during the ion run in autumn 2025. Eventually, it is planned to fully automate the setting-up of the lines for physics and in this way reduce the commissioning time from several 8-h shifts to potentially less than 1 h if all is run in parallel.

REFERENCES

[1] K. Bernier *et al.*, "Calibration of secondary emission monitors of absolute proton beam intensity in the CERN SPS North Area", CERN, Geneva, Switzerland, 1997.

[2] L. Foldesi *et al.*, "Model-based optimization for automated Multi-Turn Extraction tuning at the CERN Proton Synchrotron", presented at the IPAC'25, Taipei, Taiwan, Jun. 2025, paper THPM009, this conference.

[3] A. Lu *et al.*, "Data-driven hysteresis compensation in the CERN SPS main magnets", presented at the IPAC'25, Taipei, Taiwan, Jun. 2025, paper WEAN2, this conference.

[4] CERN, "MAD-X: Methodical Accelerator Design," [Online]. Available: `https://madx.web.cern.ch`

[5] S. Fujimoto *et al.*, "Addressing Function Approximation Error in Actor-Critic Methods", *CoRR*, vol. abs/1802.09477, 2018. `doi:10.48550/arXiv.1802.09477`

[6] M. Andrychowicz *et al.*, "Hindsight Experience Replay", in *Advances in Neural Information Processing Systems*, vol. 30, Curran Associates, Inc., 2017. Available: `https://proceedings.neurips.cc/paper_files/paper/2017/file/453fadbd8a1a3af50a9df4df899537b5-Paper.pdf`

[7] A. Raffin *et al.*, "Stable-Baselines3: Reliable Reinforcement Learning Implementations," *J. Mach. Learn. Res.*, vol. 22, no. 268, pp. 1–8, 2021. Available: `http://jmlr.org/papers/v22/20-1364.html`