REINFORCEMENT LEARNING BEYOND GREEDY OPTIMISATION FOR DELAYED-CONSEQUENCE ACCELERATOR CONTROL

S. Hirlaender*, K. M. Björkbom†, S. Trausner, O. Mironova,
University of Salzburg, Salzburg, Austria
L. Fischl, EBG MedAustron GmbH, Wiener Neustadt, Austria
P. Auer, R. Ortner, Montanuniversität Leoben, Leoben, Austria
V. Kain, CERN, Geneva, Switzerland

Abstract

Most accelerator control systems assume that the effect of an action can be evaluated locally and immediately. While greedy approaches work in near-linear regimes and Bayesian Optimisation (BO) is now standard for black-box tuning, both are essentially static optimisers and struggle in dynamic tasks with delayed consequences. In these environments, even adaptive BO remains time-myopic and lacks explicit temporal credit assignment for system memory and long-range machine evolution. We investigate three relevant forms of delayed consequences: explicit action latency (field settling delays response), magnetic hysteresis (output depends on change history), and ballistic amplification (small upstream kicks grow through nonlinear optics and apertures, causing downstream loss). Using a high-fidelity XSuite model of the AWAKE electron line, we benchmark a reinforcement learning (RL) controller against an inverse-response greedy optimiser and BO. The learning-based method anticipates delayed effects and avoids failure regions where standard baselines falter, although spatially non-local delays remain an open challenge for all tested methods. These results indicate that delayed-consequence regimes are a key class of accelerator control problems where horizon-aware methods can clearly outperform current practice.

INTRODUCTION

Modern accelerator control relies on greedy correction: measure a deviation, apply the inverse response matrix, reduce the error step-by-step. This works well in near-linear, quasi-static regimes [1], and Bayesian Optimisation (BO) has become standard for black-box tuning [2]. Kaiser *et al.* [10] showed BO is sample-efficient but can become trapped in local optima in dynamic environments. Meanwhile, CERN deployed transformer feed-forward models for SPS hysteresis compensation [11]; Scomparin *et al.* [12] achieved real-time RL on Field-Programmable Gate Arrays (FPGAs) at KARA; and the community now recognises delayed consequences as a primary motivation for RL adoption [13].

However, both Singular Value Decomposition (SVD) controllers and BO share a fundamental limitation: they evaluate actions by their *immediate, local* effect. In practice, a locally optimal correction may amplify downstream β-beating, trigger beam loss through nonlinear optics, or degrade emittance measurable only later. These *delayed consequences* arise naturally from beam transport physics, magnet dynamics, and diagnostic latencies, yet no standard benchmark tests algorithmic robustness against them.

We formalise three categories of delayed consequences, propose DelayedBeamBench built on XSuite [3] / AWAKE [4], and benchmark Proximal Policy Optimisation (PPO) [5] against SVD and Gaussian-Process-based BO (GP-BO), demonstrating that horizon-aware control outperforms greedy methods when delays are present.

DELAYED-CONSEQUENCE TAXONOMY

We identify and isolate three physically motivated categories of delayed consequences in accelerator beam steering that routinely break standard control assumptions. These fall into two classes: *temporal* delayed consequences, where the effect of an action manifests later in time (action latency and magnetic hysteresis), and *spatial* delayed consequences, where a local action causes effects at a distant location in the beam line (ballistic amplification). From a reinforcement learning perspective, each category violates a different aspect of the standard Markov Decision Process (MDP) formulation: action latency transforms the problem into a Partially Observable MDP (POMDP), hysteresis introduces non-Markovian state transitions where the reward depends on action history, and ballistic amplification creates a spatial credit-assignment problem where locally benign actions cause distant failures.

Action Latency

Magnet settling times (1–100 ms [8]) mean the controller often observes the state *before* its previous action has taken effect, transforming the system into a POMDP. We model this as an explicit d-step delay: action a_t alters the state from step $t+d$ onward, as illustrated in Fig. 1.

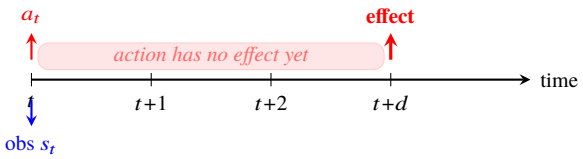

Figure 1: Action latency model: the controller issues action a_t and observes state s_t at time t, but the action only takes effect at step $t+d$.

* Equal contribution. simon.hirlaender@plus.ac.at
† Equal contribution.

Magnetic Hysteresis

Ferromagnetic yokes make the effective field path-dependent: identical setpoints yield different kicks depending on excitation history. Greedy one-step planning is insufficient since optimal actions depend on previous sequences. We model this via a Bouc–Wen inspired play-backlash wrapper with configurable width w. Fig. 2 illustrates the resulting play model: the effective field B lags behind the requested current H by $\pm w$, creating a dead-band that frustrates greedy correction.

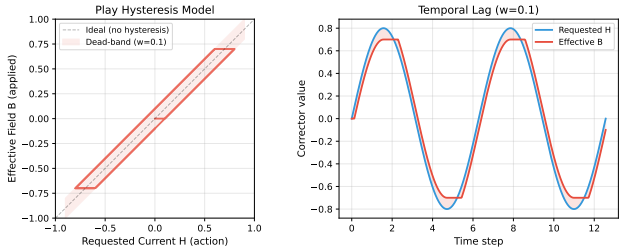

Figure 2: Play hysteresis model (w=0.1). *Left:* B–H loop showing the dead-band region. *Right:* Temporal lag between requested and effective corrector values.

Ballistic Amplification

A small upstream kick angle ($\Delta x'$) can grow through downstream quadrupoles, dictated by R_{12} and phase advance. We model this with two physically motivated mechanisms applied to the real AWAKE optics matrix R: (i) multiplicative β-beating amplification that scales downstream BPM responses by a factor α=5, and (ii) a cubic octupole-like nonlinearity: $\Delta s_{\text{nl}} = \kappa\, R\,(a_{\text{up}})^{\circ 3}$ with κ=10, where the exponent denotes element-wise (Hadamard) power. The linear SVD controller solves the amplified linear part but is blind to the cubic term, which makes large upstream corrections catastrophically unstable. Fig. 3 illustrates this spatial delay.

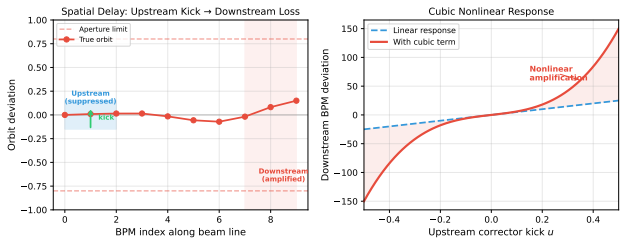

Figure 3: Ballistic amplification. *Left:* A small upstream kick appears benign locally but grows into catastrophic downstream orbit deviations. *Right:* Cubic nonlinearity makes large kicks disproportionately dangerous.

SIMULATION ENVIRONMENT

We use a high-fidelity XSuite [3] model of the AWAKE electron line wrapped as a Gymnasium [7] environment: 10 correctors, 10 dual-plane Beam Position Monitors (BPMs) (Gaussian noise σ=0.01), configurable delay $d \in \{0, 1, 3, 5\}$ and hysteresis $w \in \{0, 0.05, 0.1, 0.2\}$. The reward is $r_t = -\text{RMS}(x_t)$; beam loss ($|x_i| \geq 1$) saturates BPMs to ± 1 and terminates the episode. Success requires $r_t > -0.1$ within T=100 steps. The **success rate** (SR) averages over N=40 randomly seeded episodes.

Benchmark Scenarios

Three core benchmark scenarios (S1–S3) are explicitly defined, each cleanly isolating one specific form of delayed consequence to diagnose algorithmic failure modes: **S1** (Action Latency, d=3), **S2** (Hysteresis, w=0.1), and **S3** (Ballistic Amplification, α=5, κ=10).

CONTROL METHODS

Greedy Baseline (SVD)

The analytical greedy optimiser uses the pseudo-inverse of the empirically derived response matrix $R \in \mathbb{R}^{n_{\text{BPM}} \times n_{\text{corr}}}$ to compute a global correction vector:

$$a_t = -g R^{+} x_t\,, \tag{1}$$

where x_t is the current BPM reading vector, R^{+} is the Moore–Penrose pseudo-inverse computed via Singular Value Decomposition (SVD), and $g = 0.5$ is a conservative feedback gain. This represents standard global orbit correction used across modern accelerator facilities [1].

Bayesian Optimisation Baseline

We employ GP-BO with Expected Improvement via BoTorch [9] (Matérn 5/2 kernel), treating each interaction as a single-step black-box evaluation [2]. GP-BO is *time-myopic*: under delay d>0 it observes the state before its action takes effect, and as a black-box optimiser it is structurally blind to spatial optics.

PPO Agent

We train PPO [5] via Stable-Baselines3 [6] (γ=0.99, GAE parameter $\lambda_{\text{GAE}} = 0.95$, clip 0.2, LR $3{\times}10^{-4}$, $5{\times}10^{4}$ steps for S1/S2, $5{\times}10^{5}$ for S3, 10 seeds). Frame stacking (k=5) provides implicit state history to resolve the POMDP induced by action latency.

CORE RESULTS

The quantitative success rates and mean rewards for the three core scenarios are summarised in Table 1.

Action Latency (S1)

SVD is remarkably robust (92% SR) despite correcting from stale state. GP-BO plateaus at 48% regardless of budget. PPO (k=5) achieves 100% SR, surpassing SVD, but requires $5{\times}10^{4}$ training interactions.

Table 1: Success Rates (SR) and Mean Episode Rewards (Rew.)

Method	S1: Latency		S2: Hyst.		S3: Ballistic	
	Rew.	SR	Rew.	SR	Rew.	SR
SVD (Greedy)	−0.19	92%	−0.17	92%	−0.94	18%
GP-BO (seq.)	−0.26	48%	−0.20	88%	−1.00	0%
PPO (k=5)	**−0.18**	**100%**	**−0.09**	**100%**	**−0.68**	**40%**

Magnetic Hysteresis (S2)

Under mild hysteresis (w=0.10), SVD achieves 92% SR. GP-BO performs well at 88% SR. PPO (k=5) surpasses both with 100% SR and the best reward (−0.09).

Ballistic Amplification (S3)

Spatial non-locality is the most challenging scenario. The cubic nonlinearity makes large upstream corrections catastrophically unstable: SVD achieves only 18% SR by solving the amplified linear part but overcorrecting through the $\kappa(a_{\text{up}})^3$ term. GP-BO collapses to 0% SR. PPO (k=5, $5{\times}10^5$ steps) reaches **40**% SR—more than doubling SVD—by learning cautious, multi-step corrections that avoid exciting the nonlinear regime.

Extended Ablations (E1–E5)

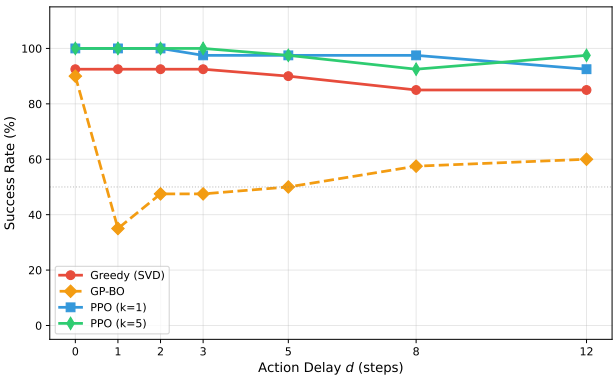

Figure 4: E1: Delay severity sweep. PPO (k=5) and PPO (k=1) maintain ≥92% SR across all delays while SVD degrades to 85% at large d. GP-BO (dashed) drops to 35% at d=1 and remains below 60%.

Fig. 4 shows that both PPO variants maintain high SR up to d=12 while SVD degrades to 85% and GP-BO collapses to 35–60%. Fig. 5 reveals that frame stacking outperforms RecurrentPPO (82% vs 60%) under hysteresis (E2), and that training–test distribution match is critical: a PPO agent trained on full BPMs drops to 45% SR when tested with slow downstream diagnostics, while retraining on the correct configuration recovers 100% (E3). Compound delays (E4), combining latency (d=2), hysteresis (w=0.1), and ballistic effects simultaneously, drop SVD to 20% SR, GP-BO to 0% SR, and PPO to 31% SR (mean over

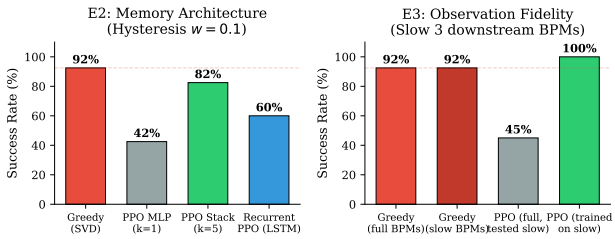

Figure 5: E2–E3 ablations. *Left (E2):* Memory architecture under hysteresis (w=0.1): frame stacking (k=5, 82%) outperforms RecurrentPPO (Long Short-Term Memory (LSTM), 60%) and memoryless PPO (42%). Note: the 82% figure reflects this architecture comparison using shorter training; Table 1 reports the fully trained agent (100%). *Right (E3):* Observation fidelity: PPO trained on the actual slow-BPM configuration reaches 100%, while a mismatched agent drops to 45%.

10 seeds)—confirming that real-world conditions where all imperfections compound remain challenging. Trajectory BO yields 0% SR across all tested budgets (E5).

DISCUSSION AND CONCLUSION

Temporal delays are surprisingly manageable. SVD achieves high SR under both latency (S1) and hysteresis (S2), matching or exceeding GP-BO with zero training cost; the conservative gain (g=0.5) provides inherent stability margins. PPO nonetheless reaches 100% SR in both scenarios, confirming that temporal credit assignment yields measurable gains even where classical methods are robust.

Spatial non-locality remains challenging. The cubic nonlinearity in S3 punishes large corrections disproportionately, reducing SVD to 18% SR despite its correct linear model. Controllers must learn cautious, multi-step strategies—precisely the temporal reasoning that RL's discounted return objective provides.

Practical implications. For latency-dominated facilities, SVD remains highly effective across all tested delays (Fig. 4). RL adds greatest value where history-dependent errors accumulate: frame stacking (k=5) outperforms both memoryless and recurrent architectures (E2).

Limitations. Our benchmark isolates each failure mode independently; real machines compound all three simultaneously (E4).

ACKNOWLEDGEMENTS

The authors gratefully acknowledge the support of the WISS 2025 project 'IDA Lab Salzburg' (Land Salzburg, 20102/F2300464-KZP, 20204-WISS/225/348/3-2023).

REFERENCES

[1] R. Steinhagen, "Beam instrumentation and diagnostics for particle accelerators," in *CERN Accelerator School*, CAS, CERN, 2013, doi:10.5170/CERN-2014-009.303.

[2] A. Scheinker *et al.*, "Demonstration of Model-Independent Control of the Longitudinal Phase Space of Electron Beams in the Linac-Coherent Light Source with Femtosecond Resolution" *Phys. Rev. Lett.*, vol. 121, p. 044801, 2018, doi:10.1103/PhysRevLett.121.044801.

[3] G. Iadarola *et al.*, "Xsuite: An Integrated Beam Physics Simulation Framework", in *Proc. HB'23*, Geneva, Switzerland, Oct. 2023, pp. 73–80. doi:10.18429/JACoW-HB2023-TUA2I1

[4] E. Gschwendtner *et al.*, "AWAKE, the advanced proton driven plasma wakefield acceleration experiment at CERN," *Nucl. Instrum. Methods Phys. Res. A*, vol. 829, pp. 76–82, 2016, doi:10.1016/j.nima.2016.02.026.

[5] J. Schulman *et al.*, "Proximal Policy Optimization Algorithms", 2017. doi:10.48550/arXiv.1707.06347

[6] A. Raffin *et al.*, "Stable-Baselines3: Reliable reinforcement learning implementations," *JMLR*, vol. 22, no. 268, pp. 1–8, 2021.

[7] Farama Foundation, "Gymnasium: A standard API for reinforcement learning," https://gymnasium.farama.org, 2023.

[8] D. Tommasini, "Practical definitions and formulae for normal conducting magnets," CERN Yellow Report CERN-2010-004, 2010, doi:10.5170/CERN-2010-004.1.

[9] M. Balandat *et al.*, "BoTorch: A framework for efficient Monte-Carlo Bayesian optimization," in *Proc. NeurIPS*, 2020.

[10] J. Kaiser *et al.*, "Reinforcement learning-trained optimisers and Bayesian optimisation for online particle accelerator tuning," *Sci. Rep.*, vol. 14, p. 15733, 2024, doi:10.1038/s41598-024-66263-y.

[11] L. Felsberger *et al.*, "Accurate control of accelerator magnet fields using transformer-based models," in *NeurIPS ML4PS Workshop*, 2024.

[12] L. Scomparin *et al.*, "Real-time control with reinforcement learning on hardware at KARA," in *Proc. IPAC'24*, Nashville, TN, USA, 2024, doi:10.18429/JACoW-IPAC2024-THPR050.

[13] A. Santamaria Garcia *et al.*, "Reinforcement learning in particle accelerators", in *Proc. IPAC'25*, Taipei, Taiwan, Jun. 2025, pp. 2481-2486. doi:10.18429/JACoW-IPAC2025-THYD1