

Improved Robustness of Deep Reinforcement Learning for Control of Time-Varying Systems by Bounded Extremum Seeking

Shaifalee Saxena^{1,2}, Alan Williams², Rafael Fierro¹, Alexander Scheinker²

Abstract—In this paper, we study the use of robust model-independent bounded extremum seeking (ES) feedback control to improve the robustness of deep reinforcement learning (DRL) controllers for a class of nonlinear time-varying systems. DRL has the potential to learn from large datasets to quickly control or optimize the outputs of many-parameter systems, but its performance degrades catastrophically when the system model changes rapidly over time. Bounded ES can handle time-varying systems with unknown control directions, but its convergence speed slows down as the number of tuned parameters increases and, like all local adaptive methods, it can get stuck in local minima. We demonstrate that together, DRL and bounded ES result in a hybrid controller whose performance exceeds the sum of its parts with DRL taking advantage of historical data to learn how to quickly control a many-parameter system to a desired setpoint while bounded ES ensures its robustness to time variations. We present a numerical study of a general time-varying system and a combined ES-DRL controller for automatic tuning of the Low Energy Beam Transport section at the Los Alamos Neutron Science Center linear particle accelerator.

I. INTRODUCTION

Deep reinforcement learning (DRL) combines aspects of optimal control theory with deep learning techniques. In DRL, deep neural networks parameterize the policy (controller), the value function, or both. DRL is based on dynamic programming, introduced by Bellman to solve sequential decision problems through the Bellman optimality principle [1]. Classical DP assumes known analytic models of the system dynamics and the reward function, which enables the synthesis of controllers u that maximize cumulative return. Reinforcement learning (RL) relaxes this modeling requirement by learning from data collected as state, action, and reward transitions when the dynamics are unknown [2], [3]. DRL extends RL to high-dimensional state and action spaces by using deep neural networks to approximate value functions and/or policies. This enables the application of RL to complex systems such as image-based perception and continuous control, where tabular methods become intractable [4]. A landmark example is deep Q-learning, which approximates the optimal action value function that maximizes cumulative reward [5]. DRL is being actively investigated for a broad set of applications, including robotic

control [6], particle accelerator tuning [7], and the fine-tuning of deep neural networks [8].

Handling time-varying systems remains a central challenge in DRL and is an active area of research [9]. When the system dynamics or reward function change substantially, the learned neural networks require retraining. Ongoing efforts aim to speed up learning from a few observations [10], [11] and to model partially observable RL problems for increased robustness [12]. Contextual RL assumes the system can be parameterized by a set of contexts $\mathcal{C}$ and learns a function that maps each $c \in \mathcal{C}$ to an associated Markov decision process (MDP) [13]. Recent variants exploit Bayesian methods for more sample-efficient training [14]. Despite these advances, such methods still require retraining for any context not represented in $\mathcal{C}$. An emerging approach to improve robustness integrates RL with a stochastic model predictive control framework for nonlinear systems subject to unbounded process noise with closed-loop guarantees [15].

In contrast to DRL, the field of control theory has a long history of developing model-independent feedback control algorithms for unknown time-varying systems [16], [17], [18], [19]. For systems of the form

$$\dot{x} = f(x, \theta(t), t) + g(x, t)u, \quad (1)$$

adaptive controllers have been designed which can handle systems with time-varying parameters $\theta(t)$ and even systems in which the control gain $g(x, t)$ is time-varying, but has a known unchanging sign such that $|g(x, t)| > 0, \forall t$ and the sign of g is known. The challenging problem of stabilizing a system with unknown control direction, in which the sign of $g(x, t)$ is unknown, was solved by Nussbaum [20], but maintained the requirement that the unknown sign cannot change with time i.e., $|g(x, t)| > 0, \forall t$. This limitation was overcome by a recently developed approach of extremum seeking (ES) for the stabilization of unknown systems with unknown and time-varying control direction $g(x, t)$ which could pass through zero and change sign repeatedly [21].

Bounded ES is a new form of ES which has a major advantage of guaranteed bounds on control efforts and parameter update rates despite acting on noisy and analytically unknown and time-varying systems [22]. It has been studied for a wide range of dynamic systems [23], and a general weak-limit averaging analysis has extended its use with non-differentiable dithering functions as well as non-periodic time-varying systems [24]. The ability to stabilize time-varying systems with guaranteed bounds on control effort has made bounded ES useful for safe hardware implementation on high energy systems, where abrupt parameter changes can

This work was supported by the Los Alamos National Laboratory, Los Alamos, NM, USA.

¹Shaifalee Saxena and Rafael Fierro are with Department of Electrical and Computer Engineering, University of New Mexico, Albuquerque, NM 87106, USA, {shaifalisaxena, rfierro}@unm.edu

²Shaifalee Saxena, Alan Williams and Alexander Scheinker are with Los Alamos National Lab, Los Alamos, NM 87547, USA {shaifalees, awilliams, ascheink}@lanl.gov

easily result in damage, such as high energy charged particle accelerators [25], [26]. The bounded ES method has also been applied for GPS-denied source localization [27], [28], optimized path tracking in robotics [29], for biology-inspired 3D source seeking [30], and for tokamak stabilization [31].

Recent research has combined RL with classical controllers to inject robustness, and safety into learning-based control. In [32], a differentiable MPC is integrated with actor-critic policy, providing better real-time performance on agile robotics. In another work [33], MRAC-RL deploys a fast model-reference adaptive controller in the inner loop together with an outer-loop RL policy to mitigate simulation-to-reality mismatch. In [34], robust control-barrier-function (CBF) layers project RL actions onto a certified safe set, improving safe exploration. These directions motivate a promising path toward hybrid controllers that integrate RL with traditional controllers.

While DRL is a powerful method for data-based learning of controllers for dynamic systems, one of the fundamental challenges is handling unknown and rapidly time-varying systems. On the other hand, while adaptive methods such as bounded ES are robust for time-varying systems, they are inherently local feedback-based schemes that do not exploit trajectory histories, are subject to suboptimal solutions, and can converge slowly in high-dimensional parameter spaces. This paper proposes a hybrid framework that combines DRL with bounded ES to leverage the strengths of both. We train a DRL policy using large datasets to find solutions in very few steps when the test dynamics are close to the training distribution. Crucially, the ES layer is warm-started from the DRL policy seeded with DRL-recommended control parameters, which reduces transients and accelerates adaptation when conditions drift. If the system starts to quickly change with time, and the learned DRL policy is no longer valid, robust bounded ES takes over and prevents the system's performance from severe degradation. We demonstrate our results with numerical studies of general time-varying dynamic systems as well as detailed simulation studies of a particle accelerator application with a time-varying magnetic lattice, mimicking natural accelerator behavior.

II. BACKGROUND

A. Bounded Extremum Seeking for Time-Varying Systems

We briefly recall the results that summarize bounded ES as developed in [22], [23], [24]. For $x \in \mathbb{R}^n$, consider a system of the form

$$\dot{x} = f(x, t) + g(x, t)u(x, t), \quad (2)$$

where $f: \mathbb{R}^n \times \mathbb{R} \rightarrow \mathbb{R}^n$ and $g: \mathbb{R}^n \times \mathbb{R} \rightarrow \mathbb{R}^{n \times m}$ are unknown, and $u: \mathbb{R}^n \times \mathbb{R} \rightarrow \mathbb{R}^m$ is the control. We focus on two special cases that are most relevant to this work. For stabilization, take $g: \mathbb{R}^n \times \mathbb{R} \rightarrow \mathbb{R}^n$ and a scalar controller $u: \mathbb{R}^n \times \mathbb{R} \rightarrow \mathbb{R}$. Choose a Lyapunov candidate $V(x) = x^T x$, and define the feedback controller as

$$u = \sqrt{\alpha\omega} \cos(\omega t + kV(x)), \quad (3)$$

then for large ω , we can approximate $x(t)$ by the weak limit-averaged dynamics of $\bar{x}(t)$ given by

$$\dot{\bar{x}} = f(\bar{x}, t) - \frac{k\alpha}{2} g(\bar{x}, t) g^T(\bar{x}, t) \nabla_{\bar{x}} V(\bar{x}). \quad (4)$$

Crucially, the averaged system's control direction is now a positive semidefinite matrix $g(\bar{x}, t) g^T(\bar{x}, t) \geq 0$, so we no longer have a control direction sign ambiguity, and we can stabilize the origin relative to $\bar{x}(t)$ by choosing sufficiently large gain $k\alpha > 0$. Another special case of (2), which is most relevant for optimization, is when $f(x, t) = 0$, $g(x, t)$ is a diagonal matrix with diagonal entries $g_i(x, t)$, $i \in \{1, \dots, n\}$, u is a vector, and we have access to measurements y of a noise-corrupted and analytically unknown time-varying cost function $J(x, t)$ which we aim to minimize, so that our system takes on the form

$$\dot{x}_i = g_i(x, t) u_i(x, t), \quad y = J(x, t) + n(t). \quad (5)$$

For this setup, which is typical of optimization problems, we design our feedback controller as

$$u_i = \sqrt{\alpha\omega_i} \cos(\omega_i t + k y(x, t)), \quad \omega_i = r_i \omega, \quad r_i \neq r_j, \quad (6)$$

resulting in averaged system dynamics

$$\begin{aligned} \dot{\bar{x}} &= -\frac{k\alpha}{2} g(\bar{x}, t) g^T(\bar{x}, t) \nabla_{\bar{x}} J(\bar{x}, t) \\ \implies \dot{\bar{x}}_i &= -\frac{k\alpha}{2} g_i^2(\bar{x}, t) \frac{\partial J(\bar{x}, t)}{\partial \bar{x}_i}, \end{aligned} \quad (7)$$

a gradient descent of the unknown $J(x, t)$. In both cases above, the proof in [24] shows that for $x \in K$ for any compact set $K \subset \mathbb{R}^n$, for any $T > 0$, and any desired $\varepsilon > 0$, there exists ω^* such that for all $\omega > \omega^*$ we can guarantee that

$$\|x(t) - \bar{x}(t)\| < \varepsilon \quad \forall t \in [0, T], \quad (8)$$

and this T can be extended to infinity if $\bar{x}(t)$ converges to a stable equilibrium. Therefore, bounded ES is a powerful model-free tool for optimizing a time-varying analytically unknown noise-corrupted output function of a dynamic system or for stabilizing unknown time-varying systems. In what follows, we will use these properties of ES to bring robustness to DRL-based feedback controllers and high-dimensional optimizers.

B. Deep Reinforcement Learning for Feedback Control

We model the DRL as a discounted Markov decision process (MDP) $M = (S, A, p, r, \gamma)$ with $\gamma \in [0, 1)$. At time t , the agent observes $s_t \in S$, applies a continuous control $a_t \in A$, receives $r_t = r(s_t, a_t)$, and transitions to $s_{t+1} \sim p(\cdot | s_t, a_t)$. The objective is

$$J(\pi) = \mathbb{E}_\pi \left[\sum_{t=0}^{\infty} \gamma^t r_t \right]. \quad (9)$$

For high-dimensional continuous actions, deterministic actor-critic methods reduce variance by differentiating through a deterministic policy. The deterministic policy gradient (DPG) theorem [35] states that for a differentiable deterministic policy $\mu_\theta: S \rightarrow A$,

$$\nabla_\theta J(\mu_\theta) = \mathbb{E}_{s \sim \rho^{\mu_\theta}} \left[\nabla_\theta \mu_\theta(s) \nabla_a Q^{\mu_\theta}(s, a) \Big|_{a=\mu_\theta(s)} \right], \quad (10)$$

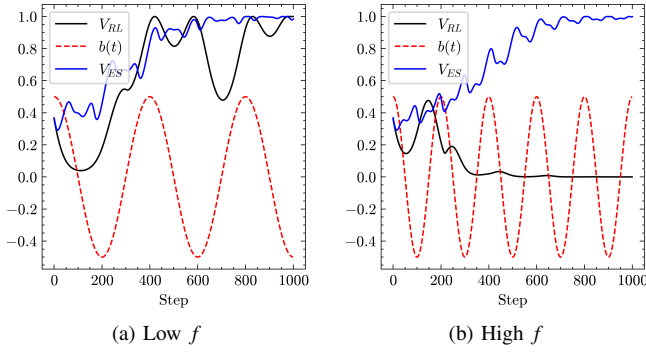


Fig. 1: Maximizing $V(x) = \exp(-x^2)$ under a sinusoidally varying control direction $b(t) = b_0 \cos(2\pi f t)$. (a) Low f : DRL reaches high V temporarily, but diverges during large swings of $b(t)$; ES reaches and maintains high V . (b) High f : DRL diverges; ES maintains high V after convergence.

where ρ^{μ_θ} is the discounted state visitation distribution and $Q^{\mu_\theta}(s, a)$ is the action-value function.

Deep Deterministic Policy Gradient (DDPG) instantiates (10) with deep function approximators and two stabilizing mechanisms from deep Q-learning: (i) an experience replay buffer for off-policy updates and (ii) slowly updated target networks for both actor and critic [36]. We maintain an actor μ_{θ^μ} and a critic Q_{θ^Q} , with target networks $\mu_{\theta^{\mu'}}$ and $Q_{\theta^{Q'}}$ updated by Polyak averaging $\theta^{\mu'} \leftarrow \tau \theta^\mu + (1 - \tau) \theta^{\mu'}$ and $\theta^{Q'} \leftarrow \tau \theta^Q + (1 - \tau) \theta^{Q'}$, $0 < \tau \ll 1$. During data collection we use exploratory actions $a_t = \mu_{\theta^\mu}(s(t)) + \varepsilon(t)$.

a) *Critic update*: From a minibatch $\{(s_i, a_i, r_i, s'_i, d_i)\}_{i=1}^N$ sampled from replay $\mathcal{D}$, where s'_i is the next state and $d_i \in \{0, 1\}$ is the terminal indicator, the TD target and loss are

$$y_i = r_i + \gamma(1 - d_i) Q_{\theta^{Q'}}(s'_i, \mu_{\theta^{\mu'}}(s'_i)), \quad (11)$$

$$\mathcal{L}(\theta^Q) = \frac{1}{N} \sum_{i=1}^N (Q_{\theta^Q}(s_i, a_i) - y_i)^2. \quad (12)$$

b) *Actor update*: Approximating the DPG with samples from replay $\mathcal{D}$,

$$\nabla_{\theta^\mu} J \approx \frac{1}{N} \sum_{i=1}^N \left[\nabla_a Q_{\theta^Q}(s, a) \Big|_{a=\mu_{\theta^\mu}(s_i)} \right] \nabla_{\theta^\mu} \mu_{\theta^\mu}(s_i). \quad (13)$$

C. Motivating Example of a Time Varying System

We consider a simple 1D open-loop unstable linear time-varying system with unknown time-varying control direction

$$\dot{x} = ax + b(t)u, \quad b(t) = b_0 \cos(2\pi f t), \quad V(x) = e^{-x^2}, \quad (14)$$

where $a, b_0 > 0$ and f are unknown and our goal is to maximize the analytically unknown objective function $V(x)$ based only on samples. Bounded ES makes $x = 0$ a practically stable equilibrium of this system by

$$u = \sqrt{\alpha \omega} \cos(\omega t - kV), \quad (15)$$

which for large ω gives averaged dynamics

$$\dot{\bar{x}} = a\bar{x} + \frac{k\alpha}{2} b^2(t) \nabla V(\bar{x}). \quad (16)$$

The averaged dynamics control direction is no longer unknown because $b^2 \geq 0$ and therefore sufficiently large $k\alpha > 0$ ensures a gradient ascent of $V(\bar{x})$. As detailed in the bounded ES references, ω must be chosen sufficiently large relative to f . While trivial for bounded ES, this problem is extremely difficult for DRL as shown in Fig 1.

When $b(t)$ varies slowly (low f ; Fig. 1a), ES successfully maximizes the objective $V(x)$ by driving $x \rightarrow 0$. In contrast, the RL policy can reach the maximizer but fails to remain there as $b(t)$ continues to drift. As the drift rate increases (Fig. 1b), the plant moves out of the RL policy's training distribution and the attained V degrades. In contrast, ES continues to ascend V due to its averaging-based gradient property even as the control direction flips.

III. PARTICLE ACCELERATOR PROBLEM FORMULATION

Large particle accelerators are intrinsically time-varying; they are composed of thousands of coupled electromagnetic components whose characteristics drift with temperature, usage, and maintenance; diagnostics are limited and noisy, and they require continual manual retuning. The amplitude and phase of the resonant radio frequency (RF) accelerating cavities drift with time due to temperature, amplifier/distribution effects, and magnet power-supplies ripple and hysteresis break magnetic-field repeatability. Such time-varying behavior is documented for the kilometer-long Los Alamos Neutron Science Center (LANSCE) linear accelerator, for which adaptive ES methods have been developed for beam loss minimization [37]. Because ES is a robust local search-based technique, as the number of parameters increases, convergence can take longer and can get stuck in local minima. Therefore, we are studying the application of a combined DRL-ES controller that can learn and quickly approximately optimize such systems with large numbers of parameters, and then robust ES can keep things stable even as they drift with time.

A. Simulator

We model the initial 12-meter long section of LANSCE, the low energy beam transport (LEBT) using the Kapchinskij-Vladimirskij (KV) envelope model. The KV equations were initially introduced in 1959 [38]; they model space-charge effects in a uniform elliptical beam moving through linear focusing fields. They capture the dominant transverse beam physics while remaining computationally efficient for learning-in-the-loop experiments [39].

The KV equations are two second-order ODEs, where for a **fixed time** t , the ODEs are solved over the independent variable z , where z represents the distance along the beamline in the direction of the beam travel.

Let $X(z, t)$ and $Y(z, t)$ denote the rms-like transverse envelope radii in the horizontal and vertical planes. The quadrupole magnet lattice enters through a signed focusing profile $G(z; u(t))$ (units of T/m) that flips sign with magnet

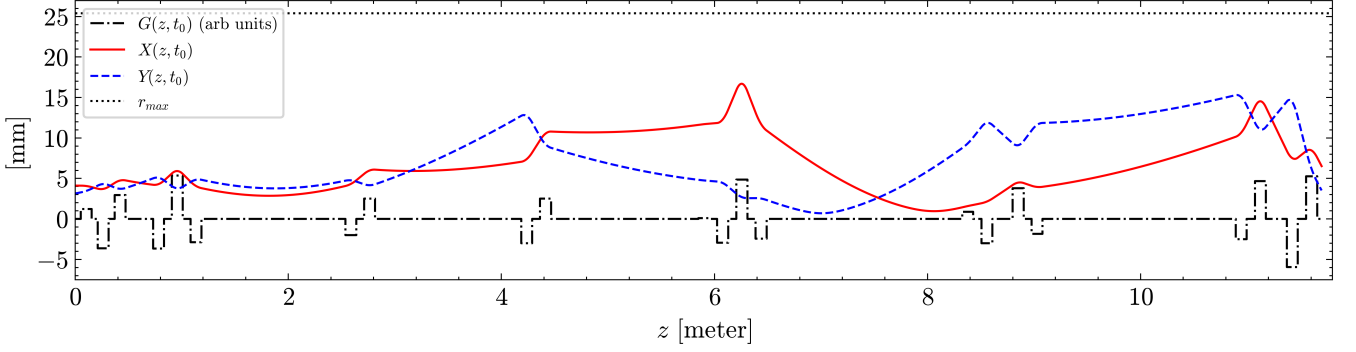


Fig. 2: Solutions of the KV equations based on the 22 quadrupole magnet strengths ($G(z, t_0)$) at initial time t_0 .

polarity. With normalized emittances ϵ_x, ϵ_y and generalized perveance K (proportional to beam current and inversely to $\gamma^3 \beta^3$), the KV envelope dynamics are written as

$$X''(z, t) = -\kappa(z, t)X(z, t) + \frac{\epsilon_x^2}{X^3(z, t)} + \frac{K}{X(z, t) + Y(z, t)}, \quad (17)$$

$$Y''(z, t) = \kappa(z, t)Y(z, t) + \frac{\epsilon_y^2}{Y^3(z, t)} + \frac{K}{X(z, t) + Y(z, t)}, \quad (18)$$

where $\kappa(z, t) := G(z; u(t)) / (\beta \rho)$ and X'', Y'' represent $\frac{d^2 X}{dz^2}, \frac{d^2 Y}{dz^2}$. The κ terms encode linear focusing/defocusing and the K term models space-charge defocusing. An example solution of the KV equations for the LANSCE LEBT at time $t = t_0$ is shown in Fig. 2.

The LANSCE LEBT contains $N = 22$ independently driven quadrupole magnets. We map the controller's action vector $u \in \mathbb{R}^{22}$ (one element per magnet) to a piecewise-constant profile as

$$G(z; u(t)) = \sum_{i=1}^{22} u_i(t) b_i(z), \quad b_i(z) = \sigma_i \mathbf{1}_{[z_i, z_i + \ell_i]}(z), \quad (19)$$

where z_i and ℓ_i are the longitudinal location and effective length of magnet i , where $i \in (1, \dots, 22)$, $\sigma_i \in \{+1, -1\}$ encodes polarity, and $\mathbf{1}$ is an indicator function for the magnet gap. Drift sections (where the function $G(z; u(t)) = 0$) contain no magnets and in these sections the beam is not subject to external fields. See [40] for additional simulation details on implementing the KV equations.

We integrate (17)–(18) forward in z with fixed-step fourth-order Runge-Kutta. Initial conditions ($X(0, t), Y(0, t), X'(0, t), Y'(0, t)$) represent the incoming beam at the LEBT entrance at time t . For initial experiments these values are kept fixed, and are randomized later to improve robustness of the training process.

The four-dimensional observation, which is provided as the state to the RL controller and used to calculate the cost in the ES controller at plane z at time t is defined as:

$$o(t) = [X(z, t), Y(z, t), X'(z, t), Y'(z, t)], \quad (20)$$

where $o(t)$ contains the full transverse profile of the beam for all $z \in [0, z_{\max}]$.

TABLE I: DDPG training hyperparameters

Algorithm	DDPG (off-policy, deterministic)
State (s) dimension	4×4000
Action (a) dimension	22
Discount γ	0.99
Actor learning rate	1×10^{-5} (Adam)
Critic learning rate	1×10^{-4} (Adam)
Replay buffer $ \mathcal{D} $	10^6 transitions (uniform)
Batch size B	128
Target update τ	0.005 (soft, every step)
Exploration noise	$\mathcal{N}(0, 0.1)$ per action

The KV model balances fidelity and speed while accurately describing the physics of a space-charge dominated, nonrelativistic, and high-current beam.

B. DRL Training Procedure

The state stacks KV beamline envelopes $s = [X, Y, X', Y']$ over $N=4000$ longitudinal locations, yielding $s \in \mathbb{R}^{16000}$. Actions are continuous setpoints for 22 quadrupole magnets, $a \in \mathbb{R}^{22}$, applied within machine-valid limits.

Reward Structure: We adopt a measurement-aligned reward that increases when the beam stays compact along the line, is well aligned at the end station, and varies smoothly. Let $z \in [0, z_{\max}]$ denote the longitudinal coordinate, with transverse envelopes $X(z), Y(z)$ and slopes $X'(z), Y'(z)$. Define the hinge $\langle a \rangle_+ \triangleq \max(0, a)$. We then form the averaged path envelopes

$$\bar{X}(t) = \frac{1}{N} \sum_{k \in Z} X(k, t), \quad \bar{Y}(t) = \frac{1}{N} \sum_{k \in Z} Y(k, t), \quad (21)$$

where $Z = \{0, \Delta z, 2\Delta z, \dots, z_{\max}\}$, $\Delta z = 2.92$ mm, and $N = 4000$ is the number of grid points.

We take $r_{\max} = 25.4$ mm as an operational radius bound (Fig. 2) and set the envelope band and terminal target as

$$r_{\text{band}} = \frac{1}{2} r_{\max}, \quad r_{\text{tt}}^2 = \frac{1}{2} r_{\max}^2. \quad (22)$$

Penalty composition: The cumulative penalty is the sum of envelope, smoothness, and terminal terms:

$$P = P_{\text{env}} + P_{\text{smooth}} + P_{\text{term}}, \quad (23)$$

$$P_{\text{env}} = w_e \left(\langle \bar{X} - r_{\text{band}} \rangle_+ + \langle \bar{Y} - r_{\text{band}} \rangle_+ \right), \quad (24)$$

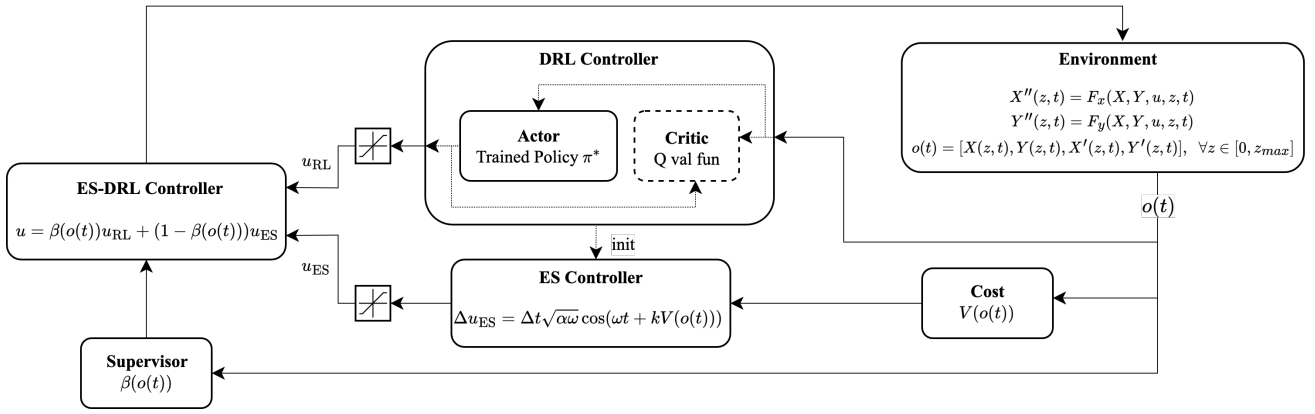


Fig. 3: Architecture of the ES-DRL controller for accelerator tuning. A supervisor selects β based on safety constraints and combines $u = \beta(o(t))u_{\text{RL}} + (1 - \beta(o(t)))u_{\text{ES}}$. ES may be warm-started from DRL (dotted). The critic is not used during evaluation.

$$P_{\text{smooth}} = w_s (\overline{X'^2} + \overline{Y'^2}). \quad (25)$$

The averaged values of $\overline{X'^2}$ and $\overline{Y'^2}$ are computed similarly to those given in Eq. (21).

At $z = z_{\text{max}}$, the terminal cost drives the beam to become circular (equal X and Y), flat (small derivatives), and of a prescribed radius r_{tt} according to:

$$P_{\text{term}} = w_r |X(z_{\text{max}}, \cdot) - Y(z_{\text{max}}, \cdot)| + w_w (|X'(z_{\text{max}}, \cdot)| + |Y'(z_{\text{max}}, \cdot)|) + w_t |X(z_{\text{max}}, \cdot)^2 + Y(z_{\text{max}}, \cdot)^2 - r_{\text{tt}}^2|. \quad (26)$$

w_e, w_s, w_r, w_w, w_t denote the weights.

The instantaneous reward is the bounded inverse

$$R = \frac{1}{1 + P} \in (0, 1], \quad (27)$$

which increases monotonically as envelope excursions, slope, and terminal mismatches decrease.

If the IVP solver fails to return valid envelopes in an episode, we assign a large penalty and terminate the episode. The environment is then reset to the last feasible setting. Directly training a controller over all 22 quadrupole inputs caused frequent failures of the initial-value problem (IVP) solver for certain magnet combinations, which prevented the simulator from returning a solution to the KV equations. To stabilize learning while preserving the intended control objective, we adopt a curricular procedure that progressively increases the control dimension.

Phase I-Group-wise training (stabilization): We partition the 22 quadrupoles into seven longitudinally contiguous groups. One group is trained at a time while the rest are held at nominal settings. For group g , the actor produces incremental updates, the IVP for (17)–(18) is solved, and the reward is evaluated. Groups are visited sequentially, carry forward the actor-critic parameters to preserve lattice knowledge and stabilize learning.

Phase II-Full 22-input training (coordination): After the group-wise pass, all 22 quadrupole magnets are trained

together and training continues with the actor-critic initialized from Phase I. We use fixed initial beam conditions $(X(0, \cdot), Y(0, \cdot), X'(0, \cdot), Y'(0, \cdot))$ to promote coordinated moves and prevent solver failures, aligning the controller with the final deployment objective in the full actuation space.

Phase III-Robust training (Random initial conditions): To improve robustness, we randomize the initial beam state each episode drawing $X(0, \cdot), Y(0, \cdot) \sim \mathcal{U}[1.5 \times 10^{-3}m, 4.5 \times 10^{-3}m]$ and $X'(0, \cdot), Y'(0, \cdot) \sim \mathcal{U}[-10^{-2}, 10^{-2}]$ and continue full 22 input training.

Curriculum transitions (Phase I→II→III) are triggered by the reward saturation test and a minimum-episode budget per phase. Across all phases we use an off-policy deterministic actor-critic (DDPG) with experience replay, slowly updated target networks[35]. We adopt the settings in Table I. The actor μ_θ is a 3-layer MLP with 512 units per layer, LayerNorm and ReLU nonlinearities, followed by a tanh output that scales actions to the permitted range. The critic Q value function is a 3-layer MLP (512 units per layer, LayerNorm+ReLU) applied to the concatenated (s, a) and outputs a scalar value.

Runtime policy: We deploy the policy trained by DDPG. At run time we use only the frozen actor μ_θ and define the RL control command as

$$u_{\text{RL}}(o(t)) = \text{sat}(\mu_\theta(o(t))), \quad (28)$$

where $o(t)$ is the current observation and $\text{sat}(\cdot)$ enforces elementwise actuator limits. Exploration noise is disabled at evaluation. The critic and target networks are used only during training and are not invoked online.

IV. COMBINED ES-DRL CONTROLLER

We consider time-varying accelerator tuning in which the control input consists of the 22 quadrupole magnet strengths $Q = (Q_1, \dots, Q_{22})$. In Fig. 2, each Q_i corresponds to one positive or negative peak of the gradient profile $G(z)$. We propose the combined ES-DRL iterative controller

$$Q_i(t+1) = u(o(t), t, Q(t)), \quad (29)$$

where

$$u(o(t), t) = \beta(o(t)) u_{\text{RL}}(o(t)) + (1 - \beta(o(t))) u_{\text{ES}}(o(t), t), \quad (30)$$

whose architecture is shown in Fig. 3. We let $o(t) = [X(z, t), Y(z, t), X'(z, t), Y'(z, t)]$ denote the observation vector, where each component is the corresponding trajectory sampled along $z \in [0, z_{\text{max}}]$ on a uniform grid of 4000 points, with $z_{\text{max}} = 11.70$ m. In this setup, the RL-based controller always updates magnet settings Q_i relative to their initial conditions according to:

$$Q_i(t+1) = Q_i(0) + \underbrace{\hat{u}_{\text{RL},i}(o(t))}_{u_{\text{RL},i}}. \quad (31)$$

The ES controller is instead uses a finite difference approximation of Eq. (6):

$$\dot{Q}_i(t) \approx \frac{Q_i(t+1) - Q_i(t)}{\Delta_t} = \sqrt{\alpha \omega_i} \cos(\omega_i \Delta_t t - kV(o(t))), \quad (32)$$

where $\Delta_t \ll 1$ is sufficiently small relative to $\max \omega_i$ for the approximation to hold, which gives

$$Q_i(t+1) = Q_i(t) + \underbrace{\Delta_t \sqrt{\alpha \omega_i} \cos(\omega_i \Delta_t t - kV(o(t)))}_{u_{\text{ES},i}}, \quad (33)$$

DRL controller: We train a Deep Deterministic Policy Gradient (DDPG) agent. At evaluation, only the trained actor π^* is used. It maps observations to actions that are subsequently passed through a saturation block to enforce actuator limits. The critic (action-value) function Q_{val} is used exclusively during training; see Section III-B for details.

ES controller: To reduce transients and accelerate adaptation, the ES controller is warm-started with the RL actor's output. The ES objective is to maximize the same reward $V(o(t))$. We use $\alpha > 0$ and distinct $\omega_i > 0$ as dither parameters, and set feedback gain $k = 15$.

Safety supervisor: The supervisor generates the binary switch $\beta(o(t))$ from envelope measurements. Beam loss increases as the beam envelope approaches the beam-pipe aperture (maximum radius r_{max}), we enforce a safety margin by constraining the beam to remain within 70% of r_{max} . Given the RL action u_{RL} , we integrate the KV equations (17)–(18) to obtain $X(z, t)$ and $Y(z, t)$ along the line (cf. Fig. 2).

Let r_{max} denote the allowable beam-pipe radius (Fig. 2), with $r_{\text{max}} = 25.4$ mm. The switching law is

$$\beta := \begin{cases} 1, & \text{if } \bar{X}(t) \text{ or } \bar{Y}(t) < 0.7 r_{\text{max}}, \\ 0, & \text{otherwise,} \end{cases} \quad (34)$$

so that RL is used when the average envelopes are well within the aperture and control reverts to the robust ES component when either envelope approaches the aperture, which indicates higher beam loss. Other beam-loss models may also be used; see [40]. Applying this rule in (35) switches to ES whenever the RL policy violates the constraint or fails to stabilize the plant. For large ω_i , when $\beta = 0$ and the ES controller is activated, the ES weak-limit averaging

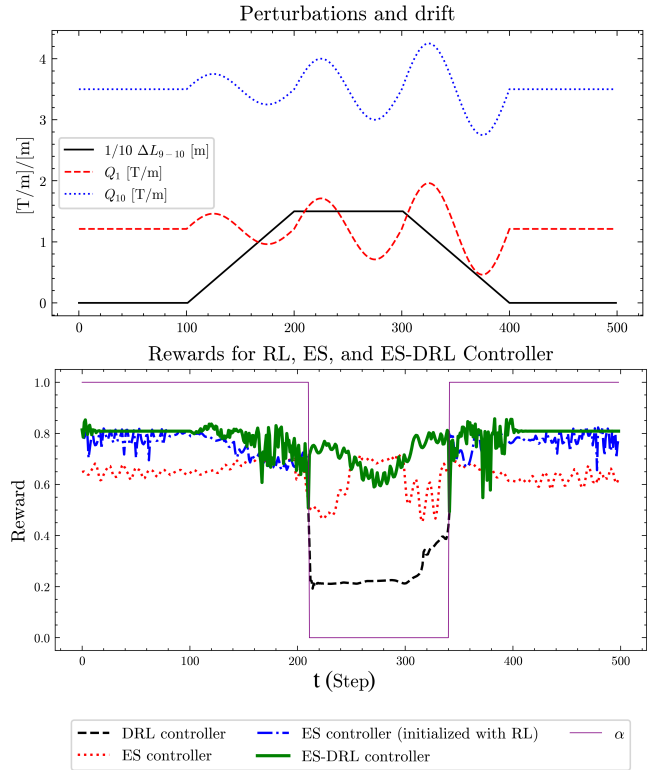


Fig. 4: Perturbations and performance: (top) injected sinusoidal perturbations at Q_1 and Q_{10} and the drift in segment length between Q_9 and Q_{10} over 500 steps. (bottom) Resulting reward trajectories; the hybrid ES-DRL controller achieves the best overall reward.

results hold and the overall ES-DRL magnet field dynamics evolve according to

$$\dot{Q}(t) \approx \frac{k\alpha}{2} \nabla_Q V(o(t)), \quad (35)$$

a local model-independent gradient ascent of the time-varying $V(o(t))$, relative to the controlled subset of quadrupole magnets Q .

V. SIMULATION RESULTS

We evaluate four controllers in the KV-based simulator: (i) a DDPG policy (DRL); (ii) bounded extremum seeking (ES); (iii) ES warm-started with the DRL action at engagement; and (iv) the proposed combined ES-DRL controller.

In this experiment the agent's action space *excludes* Q_1 and Q_{10} . These two magnets are driven exogenously by sinusoids, and Q_{10} is additionally perturbed via a geometric drift. To probe generalization beyond training, we apply a perturbation schedule that pushes the policy far outside its training distribution. Specifically, we excite Q_1 and Q_{10} with discrete-time, amplitude-ramped sinusoids

$$Q_j(t) = Q_j^* + A(t) \sin(\nu t), \quad j \in \{1, 10\}, t = 0, \dots, 500,$$

with setpoints $Q_1^* = 1.21$ T/m and $Q_{10}^* = 3.5$ T/m. The angular frequency is $\nu = \pi/50$ rad/step (period 100 steps). At step 100, the amplitude starts to increase from 0.25 to 0.75

over 400 steps and then drops back to 0, as shown in the top part of Fig. 4. Here Q_1 and Q_{10} correspond to the first and tenth extrema of $G(z)$ in Fig. 2.

We also introduce a *geometric drift* by shifting the longitudinal location of Q_{10} in (17)–(18):

$$L_{9-10}(t) = L_{9-10}^* + \Delta L_{9-10}(t),$$

where $L_{9-10}^* = 176\text{mm}$ is the distance between Q_9 and Q_{10} . The inter-magnet spacing $\Delta L_{9-10}(t)$ is ramped from 0 to 150mm over steps $t = 100:200$, held at 0.15m for $t = 200:300$, and ramped back to 0 over $t = 300:400$, as shown in the top part of Fig. 4. Aggregate performance under these perturbations is summarized in the bottom part of Fig. 4.

Because Q_1 and Q_{10} are not actuated by the controller, the agent must compensate by coordinating the *remaining* 20 quadrupoles. Note that the DRL policy was trained to command all 22 magnets; in this test it must recover performance without authority over Q_1 and Q_{10} .

Findings. The standalone DRL policy maintains a high reward (≈ 0.8) up to $\sim$ step 160, when $\Delta L_{9-10} \approx 100\text{mm}$ and the sinusoid amplitude is 0.25. As the amplitude increases and the spacing reaches 150mm, the DRL reward degrades (out-of-distribution behavior), the supervisor control hands off to ES, keeping updates bounded. The combined ES-DRL controller maintains rewards well above 0.6 throughout the plateau and the return ramp. When the spacing decreases toward 100mm and the perturbation weakens, the DRL policy recovers and β returns toward 1, restoring fast, coordinated adjustments. Warm-starting ES with DRL actions improves transients relative to standalone ES. Overall, the hybrid ES-DRL controller achieves the highest and most stable reward trajectory over the full 500 steps, consistent with Fig. 4.

VI. CONCLUSIONS

We presented a hybrid ES–DRL control framework that leverages the complementary strengths of deep reinforcement learning and bounded extremum seeking for stabilizing and optimizing nonlinear time-varying systems. Numerical studies and accelerator-inspired simulations demonstrated that while DRL policies can achieve rapid convergence in-distribution regimes, their performance degrades under unmodeled dynamics and distribution shifts. In contrast, bounded ES guarantees robustness to unknown and drifting control directions, though at the cost of slower convergence. By combining these approaches through a safety-aware supervisor and warm-starting ES from DRL actions, the proposed controller maintained high rewards under severe perturbations and outperformed either method alone. These results suggest a principled path toward deploying learning-based controllers in high-dimensional, safety-critical applications such as particle accelerators, where both adaptability and robustness are essential. Future work will extend this framework to hardware-in-the-loop studies at LANSCE and will include a theoretical study of the convergence properties of this approach for a wide range of nonlinear time-varying dynamic systems.

ACKNOWLEDGMENTS

This research has been supported by the Laboratory Directed Research and Development program of Los Alamos National Laboratory under project “Charged Particle Beam Control and Diagnostics using Adaptive Machine Learning,” project number 20220074D. Los Alamos National Laboratory is operated by Triad National Security, LLC, for the National Nuclear Security Administration of U.S. Department of Energy (Contract No. 89233218CNA000001).

REFERENCES

- [1] R. Bellman, *Dynamic Programming*. Princeton University Press, 1957.
- [2] R. S. Sutton, A. G. Barto *et al.*, *Reinforcement learning: An introduction*. MIT press Cambridge, 1998, vol. 1, no. 1.
- [3] D. Bertsekas, *Dynamic programming and optimal control: Volume I*. Athena scientific, 2012, vol. 4.
- [4] K. Arulkumaran, M. P. Deisenroth, M. Brundage, and A. A. Bharath, “Deep reinforcement learning: A brief survey,” *IEEE signal processing magazine*, vol. 34, no. 6, pp. 26–38, 2017. [Online]. Available: <https://ieeexplore.ieee.org/abstract/document/8103164>
- [5] V. Mnih, K. Kavukcuoglu, D. Silver, A. A. Rusu, J. Veness, M. G. Bellemare, A. Graves, M. Riedmiller, A. K. Fidjeland, G. Ostrovski *et al.*, “Human-level control through deep reinforcement learning,” *nature*, vol. 518, no. 7540, pp. 529–533, 2015.
- [6] C. Tang, B. Abbatemateo, J. Hu, R. Chandra, R. Martín-Martín, and P. Stone, “Deep reinforcement learning for robotics: A survey of real-world successes,” *Annual Review of Control, Robotics, and Autonomous Systems*, vol. 8.
- [7] S. Hirlaender and N. Bruchon, “Model-free and bayesian ensembling model-based deep reinforcement learning for particle accelerator control demonstrated on the fermi fel,” *arXiv preprint arXiv:2012.09737*, 2020. [Online]. Available: <https://doi.org/10.48550/arXiv.2012.09737>
- [8] L. Ouyang, J. Wu, X. Jiang, D. Almeida, C. Wainwright, P. Mishkin, C. Zhang, S. Agarwal, K. Slama, A. Ray *et al.*, “Training language models to follow instructions with human feedback,” *Advances in neural information processing systems*, vol. 35, pp. 27 730–27 744, 2022.
- [9] C. Benjamins, T. Eimer, F. Schubert, A. Biedenkapp, B. Rosenhahn, F. Hutter, and M. Lindauer, “Carl: A benchmark for contextual and adaptive reinforcement learning,” *arXiv preprint arXiv:2110.02102*, 2021.
- [10] C. Finn, P. Abbeel, and S. Levine, “Model-agnostic meta-learning for fast adaptation of deep networks,” in *International conference on machine learning*. PMLR, 2017, pp. 1126–1135. [Online]. Available: <https://proceedings.mlr.press/v70/finn17a.html>
- [11] Y. Duan, J. Schulman, X. Chen, P. L. Bartlett, I. Sutskever, and P. Abbeel, “ RL^2 : Fast reinforcement learning via slow reinforcement learning,” *arXiv preprint arXiv:1611.02779*, 2016.
- [12] M. Lauri, D. Hsu, and J. Pajarinen, “Partially observable markov decision processes in robotics: A survey,” *IEEE Transactions on Robotics*, vol. 39, no. 1, pp. 21–40, 2022.
- [13] A. Hallak, D. Di Castro, and S. Mannor, “Contextual markov decision processes,” *arXiv preprint arXiv:1502.02259*, 2015.
- [14] J.-H. Cho, V. Jayawardana, S. Li, and C. Wu, “Model-based transfer learning for contextual reinforcement learning,” *Advances in Neural Information Processing Systems*, vol. 37, pp. 88 279–88 319, 2024.
- [15] J. Köhler and M. N. Zeilinger, “Predictive control for nonlinear stochastic systems: Closed-loop guarantees with unbounded noise,” *IEEE Transactions on Automatic Control*, 2025.
- [16] K. Tsakalis and P. Ioannou, “Adaptive control of linear time-varying plants,” *Automatica*, vol. 23, no. 4, pp. 459–468, 1987.
- [17] R. Marino and P. Tomei, “Adaptive control of linear time-varying systems,” *Automatica*, vol. 39, no. 4, pp. 651–659, 2003.
- [18] K. Chen and A. Astolfi, “Adaptive control for systems with time-varying parameters,” *IEEE Transactions on Automatic Control*, vol. 66, no. 5, pp. 1986–2001, 2020.
- [19] H. K. Khalil, *Nonlinear systems*. Upper Saddle River, N.J.: Prentice Hall, 2002.
- [20] R. D. Nussbaum, “Some remarks on a conjecture in parameter adaptive control,” *Systems & control letters*, vol. 3, no. 5, pp. 243–246, 1983.

- [21] A. Scheinker and M. Krstić, "Minimum-seeking for CLFs: Universal semiglobally stabilizing feedback under unknown control directions," *IEEE Transactions on Automatic Control*, vol. 58, no. 5, pp. 1107–1122, 2012.
- [22] A. Scheinker *et al.*, "Model independent beam tuning," in *Proceedings of the 2013 International Particle Accelerator Conference, Shanghai, China, 2013*. [Online]. Available: <https://proceedings.jacow.org/IPAC2013/papers/tupwa068.pdf>
- [23] A. Scheinker and M. Krstić, "Extremum seeking with bounded update rates," *Systems & Control Letters*, vol. 63, pp. 25–31, 2014. [Online]. Available: <https://doi.org/10.1016/j.sysconle.2013.10.004>
- [24] A. Scheinker and D. Scheinker, "Bounded extremum seeking with discontinuous dithers," *Automatica*, vol. 69, pp. 250–257, 2016. [Online]. Available: <https://doi.org/10.1016/j.automatica.2016.02.023>
- [25] A. Scheinker, S. Baily, D. Young, J. S. Kolski, and M. Prokop, "In-hardware demonstration of model-independent adaptive tuning of noisy systems with arbitrary phase drift," *Nuclear Instruments and Methods in Physics Research Section A: Accelerators, Spectrometers, Detectors and Associated Equipment*, vol. 756, pp. 30–38, 2014.
- [26] A. Scheinker, E.-C. Huang, and C. Taylor, "Extremum seeking-based control system for particle accelerator beam loss minimization," *IEEE Transactions on Control Systems Technology*, vol. 30, no. 5, pp. 2261–2268, 2021.
- [27] M. A. Ghadiri-Modarres, M. Mojiri, and H. R. Zangeneh, "New schemes for gps-denied source localization using a nonholonomic unicycle," *IEEE Transactions on control systems technology*, vol. 25, no. 2, pp. 720–727, 2016.
- [28] M. A. Ghadiri Modarres, M. Mojiri, and H. R. Zangeneh, "Non-holonomic source localization in 3-d environments without position measurement," *IEEE Transactions on Automatic Control*, vol. 61, no. 11, pp. 3563–3567, 2016.
- [29] S. Bajpai, "Investigating the performance of different controllers in optimized path tracking in robotics: A lie bracket system and extremum seeking approach," Master's thesis, University of Cincinnati, 2024.
- [30] M. Abdelgalil and H. Taha, "Recursive averaging with application to bio-inspired 3-d source seeking," *IEEE Control Systems Letters*, vol. 6, pp. 2816–2821, 2022.
- [31] G. De Tommasi, S. Dubbioso, A. Mele, and A. Pironti, "Event-driven adaptive vertical stabilization in tokamaks based on a bounded extremum seeking algorithm," in *2022 IEEE Conference on Control Technology and Applications (CCTA)*. IEEE, 2022, pp. 831–836.
- [32] A. Romero, E. Aljalbout, Y. Song, and D. Scaramuzza, "Actor-critic model predictive control: Differentiable optimization meets reinforcement learning," *arXiv preprint arXiv:2306.09852*, 2024.
- [33] A. Guha and A. M. Annaswamy, "Online policies for real-time control using MRAC-RL," in *2021 60th IEEE Conference on Decision and Control (CDC)*. IEEE, 2021, pp. 1808–1813.
- [34] Y. Emam, G. Notomista, P. Glotfelter, Z. Kira, and M. Egerstedt, "Safe reinforcement learning using robust control barrier functions," *IEEE Robotics and Automation Letters*, 2022.
- [35] D. Silver, G. Lever, N. Heess, T. Degris, D. Wierstra, and M. Riedmiller, "Deterministic policy gradient algorithms," in *International conference on machine learning*. Pmlr, 2014, pp. 387–395.
- [36] T. P. Lillicrap, J. J. Hunt, A. Pritzel, N. Heess, T. Erez, Y. Tassa, D. Silver, and D. Wierstra, "Continuous control with deep reinforcement learning," *arXiv preprint arXiv:1509.02971*, 2015.
- [37] A. Scheinker, P. Naffziger, and A. Garcia, "Extremum seeking for minimization of beam loss in the lansce linear accelerator by tuning rf cavities," in *2020 American Control Conference (ACC)*. IEEE, 2020, pp. 5071–5075.
- [38] I. Kapchinskij and V. Vladimirkij, "Limitations of proton beam current in a strong focusing linear accelerator associated with the beam space charge," in *Proceedings of the International Conference on High Energy Accelerators and Instrumentation*, vol. 1957. CERN Scientific Information Service, 1959, pp. 274–288.
- [39] S. M. Lund and B. Bukh, "Stability properties of the transverse envelope equations describing intense ion beam transport," *Physical Review Special Topics—Accelerators and Beams*, vol. 7, no. 2, p. 024801, 2004.
- [40] A. Williams, A. Scheinker, E.-C. Huang, C. Taylor, and M. Krstić, "Safe extremum seeking applications in particle accelerators," in *2024 American Control Conference (ACC)*. IEEE, 2024, pp. 4314–4319.