

A Geometric Approach to Soft Actor-Critic with Zonotopes for Locomotion Learning

P. Roditis^{1,2,3}, P. P. Filntsis^{1,2}, and P. Maragos^{1,2,3}

Abstract—Off-policy actor-critic methods control overestimation bias by taking the minimum of two critics. This uses the same aggregation rule everywhere, regardless of how the critics disagree. We propose GeZo-SAC, which uses auxiliary geometric representations to adapt critic pessimism to the state and action. Alongside its scalar value, each critic predicts a set of generators defining a zonotope. Probing this zonotope along sampled directions provides a geometric width, "subtracted from each critic value as a pessimistic offset, and a measure of disagreement between the two critics, aggregated with log-sum-exp. This disagreement controls how the critics are combined, moving from a width-weighted average toward the usual minimum as disagreement increases. At inference, the deployed policy is an unmodified SAC actor, since the generators are used only on the critic side during training.

Across four MuJoCo-v5 locomotion benchmarks and six off-policy baselines, GeZo-SAC achieves the highest mean return on Ant-v5 and Hopper-v5 and remains competitive with other methods on the remaining tasks. Our analysis further shows that GeZo-SAC achieves the lowest average actuator work and action effort per metre among the evaluated methods, while maintaining near-zero measured overestimation frequency across all four environments.

I. INTRODUCTION

Reinforcement learning (RL) has emerged as a powerful framework for learning control policies for complex robotic systems, enabling agents to acquire locomotion and manipulation skills directly from interaction with their environment. In these settings, reliable value estimation is crucial, as inaccurate predictions of future return can lead to unstable policies, inefficient motions, or unsafe behavior.

Modern off-policy actor-critic algorithms such as SAC achieve strong performance in continuous-control tasks due to their sample efficiency, but their effectiveness depends critically on how the learned Q -function guides policy improvement. In these methods, the critic defines both temporal-difference targets and the gradients that shape actor updates; even modest approximation errors can therefore be amplified, as the actor may preferentially select actions that exploit spurious positive Q errors [1], [2]. This phenomenon can lead to systematic *overestimation bias*, which can destabilize training and degrade final policy quality [3], [4], [5].

Twin-critic methods mitigate this bias by taking the minimum of two separately parameterized Q predictions, a strategy adopted by both TD3 [4] and SAC [5]. While effective, this uses a fixed aggregation rule based only on the two scalar value estimates. It therefore does not exploit

richer structure in how the critics disagree. This motivates using structured critic representations to guide pessimism and aggregation.

We propose a geometry-guided extension of SAC in which each critic predicts, alongside its scalar value, a set of generators defining a state-action-dependent zonotope. Directional support probes provide two signals: a geometric band subtracted from each critic value as a pessimistic offset, and a measure of disagreement between the two critics. This disagreement controls how their pessimistic value estimates are combined, shifting smoothly from a weighted average toward the minimum as disagreement increases.

Our design draws on connections between tropical (max-plus) geometry and piecewise-linear function approximators [6], [7], [8], [9]. Prior work has related ReLU networks to polyhedral and zonotopic representations, and approximation error to distances between the associated geometric objects [10]. Motivated by these connections, we use support functions to probe and compare the learned zonotopes, together with temperature-controlled LSE (log-sum-exp) aggregation as a smooth approximation of the largest directional discrepancy.

Our main contributions are as follows:

- We introduce **GeZo-SAC**, a geometry-guided extension of SAC in which each critic predicts an auxiliary zonotope representation. Directional support probes provide geometric width and disagreement signals that adapt pessimistic value estimates and twin-critic aggregation.
- We evaluate GeZo-SAC against six off-policy baselines on four MuJoCo-v5 locomotion benchmarks. GeZo-SAC achieves the highest mean return on Ant-v5 and Hopper-v5 while remaining competitive with other methods on the remaining tasks.
- We complement return-based evaluation with value-estimation and control-quality diagnostics, including overestimation frequency, critic-return discrepancy, actuator work, action variation, and action effort. GeZo-SAC exhibits near-zero measured overestimation frequency and achieves the lowest average actuator work and action effort per metre among evaluated methods.

II. RELATED WORK

Overestimation bias and pessimistic value estimation.

Overestimation bias arises when function approximation and bootstrapping produce optimistic value errors that the policy can exploit. This phenomenon is closely related to the classic *maximization bias* identified in value-based methods, where using the same estimator to both select and evaluate actions

¹Robotics Institute, Athena Research Center, Marousi, Greece

²HERON - Hellenic Robotics Center of Excellence, Athens, Greece

³School of Electrical & Computer Engineering, NTUA, Greece

leads to overestimated targets. Van Hasselt et al. [11] show that DQN suffers from substantial overestimation on Atari and propose Double DQN, which reduces bias by decoupling action selection (online network) from action evaluation (target network). In continuous control, Kuznetsov et al. [2] show that even with common heuristics, inaccurate value estimates can be amplified by policy improvement, degrading stability and performance; they propose Truncated Quantile Critics (TQC), which models a distribution over returns and truncates optimistic quantiles (via an ensemble mixture), providing finer control over overestimation than a simple “take-the-minimum” rule. Generalized Pessimism Learning (GPL) [12] introduces an uncertainty-dependent penalty on critic targets and learns its weight through dual TD-learning, allowing the degree of pessimism to adapt during training. More recently, TDC- λ [13] uses twin distributional critics and selects Bellman targets according to the lower-confidence score $\mu - \lambda\sigma$, providing explicit control over target conservatism. Conservative Q-Learning (CQL) [14] addresses overestimation in offline reinforcement learning by augmenting the Bellman-error objective with a Q-value regularizer that encourages conservative estimates. Under stated assumptions, the expected policy value under the learned critic lower-bounds the true policy value. This regularization discourages inflated values for actions outside the dataset distribution, reducing the incentive to exploit extrapolation errors. Our method uses support-derived bands and directional disagreement between critic zonotopes for adaptive pessimism. The bands determine value offsets and sample-dependent aggregation weights, while support discrepancies determine a minibatch-level mixing coefficient.

Uncertainty-aware exploration and critic aggregation. Optimistic Actor-Critic (OAC) [15] addresses the issue of premature convergence in Soft Actor-Critic by distinguishing between the values used for stable updates and those used for exploration. While it maintains an approximate lower confidence bound (LCB) to provide a conservative estimate for policy improvement, mirroring the overestimation control found in TD3, it introduces an approximate upper confidence bound (UCB) to guide exploration. By shifting the exploration policy toward actions with high epistemic uncertainty, OAC enables more purposeful searching of the state-action space compared to standard Gaussian noise. Complementarily, recent advances in distributional actor-critic methods model the full return distribution in the critic, enabling the use of distributional statistics (e.g., variance) to improve stability and mitigate overestimation bias. For example, Distributional Soft Actor-Critic with Three Refinements introduces twin value distribution learning and variance-aware refinements for more robust critic updates [16]. SUNRISE [17] uses a Q-function ensemble to estimate uncertainty, reweight Bellman backups, and select exploratory actions using upper-confidence estimates. Our approach augments scalar critics with zonotope representations. Their directional supports provide geometric bands and a measure of disagreement between these representations. These quantities control pessimistic value offsets and critic

aggregation.

Tropical geometry and piecewise-linear networks. Charisopoulos and Maragos [7], [8] and Zhang et al. [9] were the first to apply tropical geometry to the study of deep neural networks, establishing a connection between feedforward neural networks with ReLU activations and tropical geometry, proving that such networks are equivalent to families of tropical rational mappings. Furthermore, they demonstrated that the hidden layers of feedforward ReLU networks can be represented using zonotopes, which serve as fundamental building blocks for deeper architectures, and that the decision boundaries of these networks correspond to tropical hypersurfaces. Maragos and Theodosis [6] extended tropical geometrical constructs by utilizing weighted lattices and derived optimal solutions to max- $\star$ equations (where $\star$ denotes a binary operation distributive over max) through morphological adjunctions, which serve as projections on weighted lattices. By fitting max- $\star$ tropical curves and surfaces to data approximating polygonal or polyhedral shapes, they established a link between tropical geometry and piecewise-linear regression optimization [6]. Calafiore et al. [18] introduced a novel class of neural networks called log-sum-exp (LSE) networks, which serve as universal approximators of convex functions. They further showed that difference-LSE networks act as smooth universal approximators of continuous functions over compact convex domains [19]. Both LSE and difference-LSE networks exhibit a close connection to tropical mathematics through the so-called *dequantization* process, and provide a smooth counterpart to max-type operators that arise naturally in tropical settings. Misiakos et al. [10] and Fotopoulos et al. [20] further showed that, for ReLU networks, approximation error can be related to Hausdorff-type distances between the associated polyhedral objects. Our work draws on these connections in practice: we use learned zonotope generators as geometric representations, directional support functions to compare these representations, and temperature-controlled LSE aggregation as a smooth approximation of the maximum discrepancy over sampled directions. These components provide practical mechanisms for adaptive pessimism in reinforcement learning.

III. METHOD

GeZo-SAC augments each SAC critic with a learned geometric representation and uses disagreement between the two representations to adapt pessimism. Let $s \in \mathbb{R}^D$ and $a \in \mathbb{R}^A$. Critic $i \in \{1, 2\}$ outputs a scalar $Q_i(s, a)$ and K generators collected as

$$G_i(s, a) = [g_{i,1}(s, a)^\top \quad \cdots \quad g_{i,K}(s, a)^\top]^\top \in \mathbb{R}^{K \times (D+1)}.$$

The generators lie in a fixed auxiliary space $\mathbb{R}^{D+1}$, whose dimension sets the capacity of the geometric representation. Conditioning them on both state and action allows the bands and aggregation weights to vary across candidate actions at the same state, providing action-dependent geometric modulation in both the Bellman target and the actor objective.

Directional support and band. For $N = 32$ random unit

directions $U = \{u_n\}_{n=1}^N \subset \mathbb{R}^{D+1}$, cached and periodically resampled during training, we define

$$h(u; G_i(s, a)) = \sum_{j=1}^K |\langle g_{i,j}(s, a), u \rangle|, \quad (1)$$

$$b_i(s, a) = \frac{1}{N} \sum_{n=1}^N h(u_n; G_i(s, a)). \quad (2)$$

Thus, b_i approximates the zonotope’s average directional support radius.

Generator training signal. Generator-head parameters are trained through a squared-band penalty, a decorrelation penalty on batch-mean generator rows, and a cross-critic band-symmetry penalty, each added to the critic loss with weight 10^{-3} :

$$\mathcal{L}_{\text{band}} = \mathbb{E}_{\mathcal{B}}[b_i(s, a)^2], \quad (3)$$

$$\mathcal{L}_{\text{div}} = \frac{1}{K^2} \|C(\bar{G}_i) - I\|_F^2, \quad \bar{G}_i = \mathbb{E}_{\mathcal{B}}[G_i(s, a)], \quad (4)$$

$$\mathcal{L}_{\text{symm}} = \mathbb{E}_{\mathcal{B}}[(\bar{b}_1 - \bar{b}_2)^2], \quad (5)$$

where $C(\cdot)$ is the row-wise cosine-similarity matrix. The generator heads are trained by geometric regularization, with TD learning acting indirectly through the shared trunk.

Band normalization and pessimistic values. Each critic maintains an EMA $\bar{b}_i$ of its batch-mean band. We define

$$\tilde{b}_i(s, a) = \frac{b_i(s, a)}{\bar{b}_i + \varepsilon}, \quad V_i(s, a) = Q_i(s, a) - \kappa_i \tilde{b}_i(s, a),$$

treating the EMA normalizers as constants during differentiation. The effective coefficient κ_i is shared between critics and takes the target- or actor-side form defined below.

Band-based actor modulation. For actor updates, we apply a second EMA normalization and compute

$$\hat{b}_i = \frac{\tilde{b}_i}{\bar{b}_i + \varepsilon}, \quad \hat{b}_{\text{avg}} = \frac{1}{2}(\hat{b}_1 + \hat{b}_2), \quad (6)$$

$$\hat{b}_{\text{diff}} = |\hat{b}_1 - \hat{b}_2|, \quad u = \frac{\hat{b}_{\text{avg}} + \eta \hat{b}_{\text{diff}}}{\bar{s}(t) + \varepsilon}, \quad (7)$$

$$c(t) = \frac{1}{B} \sum_{m=1}^B \exp[-\tau_c u(s_m, a_m^\pi)], \quad a_m^\pi \sim \pi_\theta(\cdot | s_m), \quad (8)$$

where $\bar{s}(t)$ is an actor-side EMA of the minibatch mean of $\hat{b}_{\text{avg}}$, $\eta = 0.5$, and $\tau_c = 1$. The minibatch scalar $c(t) \in [0, 1]$ is detached during the actor update.

Width scheduling. Let t denote the number of critic updates. $\lambda_{\text{base}}(t)$ is zero for the first 10^4 updates and increases linearly to λ_{zono} over the next 2×10^5 updates. We define

$$\lambda_i^\pi = \text{clip}(\lambda_{\text{base}}(t)c(t), 0, \lambda_{\text{zono}}), \quad (9)$$

$$\kappa_{i,m}^\pi = \lambda_i^\pi \kappa \beta_{i,m}, \quad \kappa_i^{\text{tar}} = \lambda_{\text{base}}(t) \kappa, \quad (10)$$

$$\beta_{i,m} \sim \text{Beta}(1 + 4c(t), 5 - 4c(t)). \quad (11)$$

During warm-up, the band offsets are zero and the generator penalties are inactive, although geometric aggregation remains active. Actor and target computations use $\kappa_{i,m}^\pi$ and κ_i^{tar} , respectively.

Algorithm 1 GeZo-SAC training loop (high-level): pessimistic targets with adaptive width.

Require: Actor π_θ , twin critics $\{Q_{\phi_i}, G_{\phi_i}\}_{i=1,2}$, targets $\{\bar{Q}_{\phi_i}, \bar{G}_{\phi_i}\}_{i=1,2}$, buffer $\mathcal{B}$

- 1: **for** each update step **do**
- 2: Sample minibatch; sample $a' \sim \pi_\theta(\cdot | s')$
- 3: $V_i \leftarrow \bar{Q}_i(s', a') - \kappa_i^{\text{tar}} \bar{b}_i(s', a')$ // target values
- 4: $\rho^{\text{tar}} \leftarrow \exp(-\beta_{\text{geo}} \bar{d}_{\text{support}}^{\text{tar}})$ // incl. corner directions
- 5: $V_{\text{geo}}^{\text{tar}} \leftarrow (1 - \rho^{\text{tar}}) \min_i V_i + \rho^{\text{tar}} \sum_i w_i V_i$
- 6: $y \leftarrow r + \gamma(1 - d) (V_{\text{geo}}^{\text{tar}} - \alpha(0.5) \log \pi_\theta(a' | s'))$
- 7: Update critics; adjust κ via coverage control
- 8: **if** policy update step **then**
- 9: Compute actor-side ρ^π and $V_{\text{geo}}^\pi(s_m, a_{\pi,m})$ using $\kappa_{i,m}^\pi$
- 10: Update actor and entropy coefficient $\alpha(\cdot)$
- 11: **end if**
- 12: Polyak update targets
- 13: **end for**

Batch-adaptive corner directions. In addition to U , we construct detached directions from the batch-mean differences between corresponding generator rows:

$$\bar{\Delta}g_j = \frac{1}{|\mathcal{B}|} \sum_{(s,a) \in \mathcal{B}} (g_{1,j}(s, a) - g_{2,j}(s, a)), \quad (12)$$

$$u_j^{\text{corner}} = \frac{\bar{\Delta}g_j}{\|\bar{\Delta}g_j\|_2}, \quad \|\bar{\Delta}g_j\|_2 > \varepsilon. \quad (13)$$

This produces at most K additional directions, recomputed independently for target and actor evaluations. They are concatenated with the cached probes as

$$U_\rho = \text{concat}(U, U^{\text{corner}}),$$

with repeated directions retained.

Geometry-aware support disagreement. This construction is applied independently at the two points in training where critic disagreement is needed: at (s', a') for target computation, and at (s, a_π) for the actor update. This yields direction collections U_ρ^{tar} and U_ρ^π built from the corresponding generators $\bar{G}_i(s', a')$ and $G_i(s, a_\pi)$, respectively. For each sample and direction $u \in U_\rho^{(\cdot)}$, where $(\cdot) \in \{\text{tar}, \pi\}$, define

$$\Delta^{(\cdot)}(u) = |h(u; G_1^{(\cdot)}) - h(u; G_2^{(\cdot)})|, \quad (14)$$

$$s_{\text{scale}}^{(\cdot)} = \frac{1}{2}(\tilde{b}_1^{(\cdot)} + \tilde{b}_2^{(\cdot)}) + \varepsilon. \quad (15)$$

We aggregate disagreement using temperature $\tau_d > 0$:

$$d_{\text{support}}^{(\cdot)} = \tau_d \log \sum_{u \in U_\rho^{(\cdot)}} \exp \left(\frac{\Delta^{(\cdot)}(u)}{\tau_d s_{\text{scale}}^{(\cdot)}} \right).$$

The sum includes every entry in the concatenated direction collection, including repetitions. These per-sample quantities are averaged over the minibatch, yielding $\bar{d}_{\text{support}}^{\text{tar}}$ and $\bar{d}_{\text{support}}^\pi$. We define

$$\rho^{(\cdot)} = \exp(-\beta_{\text{geo}} \bar{d}_{\text{support}}^{(\cdot)}), \quad \beta_{\text{geo}} > 0.$$

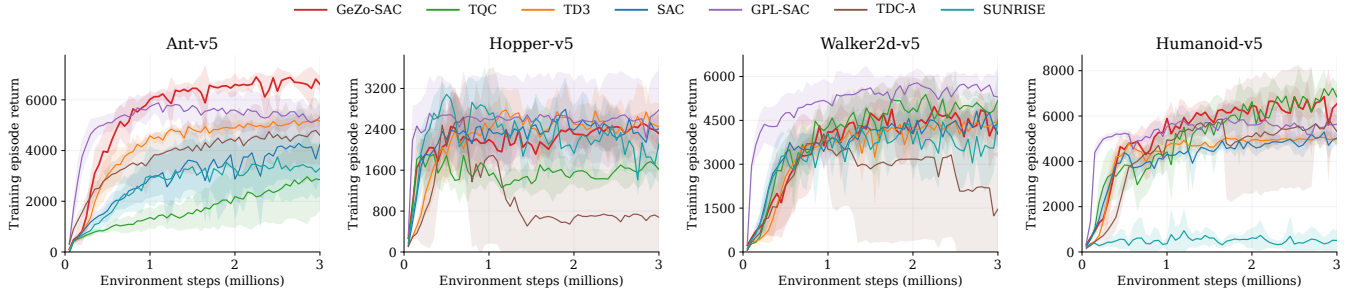


Fig. 1: **Training performance on four MuJoCo-v5 locomotion environments.** Learning curves show episodic training returns for Ant-v5, Hopper-v5, Walker2d-v5, and Humanoid-v5. Per-run returns are smoothed and interpolated onto a common environment-step grid. Curves represent the mean over five independent runs ($n = 5$), and shaded regions denote $\pm$ one standard deviation across runs. GeZo-SAC leads on Ant-v5 and remains competitive on the other tasks.

Both mixing coefficients are computed without gradient tracking.

Inverse-band critic weighting and geometric aggregation. Given pessimistic value estimates V_1, V_2 , we define inverse-band weights

$$w_i(s, a) = \frac{(\tilde{b}_i + \varepsilon_w)^{-1}}{(\tilde{b}_1 + \varepsilon_w)^{-1} + (\tilde{b}_2 + \varepsilon_w)^{-1}}, \quad \varepsilon_w = 10^{-6}.$$

We then form

$$V_w(s, a) = w_1 V_1 + w_2 V_2, \quad (16)$$

$$V_{\min}(s, a) = \min\{V_1, V_2\}, \quad (17)$$

$$V_{\text{geo}}^{(\cdot)}(s, a) = (1 - \rho^{(\cdot)})V_{\min} + \rho^{(\cdot)}V_w, \quad (18)$$

where $(\cdot) \in \{\text{tar}, \pi\}$. The coefficient ρ^{tar} combines the target critics' pessimistic value estimates at (s', a') , while ρ^π combines the online critics' pessimistic value estimates at (s, a_π) . In each case, the weights use the corresponding normalized bands.

Intuitive example. In a two-dimensional example, each generator defines one direction of a centrally symmetric zonotope, and $h(u; G_i)$ measures its radius along probe direction u . Similar supports from the two critics yield small disagreement and favor their weighted average. A large discrepancy in any direction is emphasized by the LSE aggregation, reducing ρ and moving the estimate toward the conservative minimum.

Critic targets. Target critics compute $V_{\text{geo}}^{\text{tar}}(s', a')$ using κ_t^{tar} . For $a' \sim \pi_\theta(\cdot | s')$, the Bellman target is

$$y = r + \gamma(1 - d) [V_{\text{geo}}^{\text{tar}}(s', a') - \alpha(0.5) \log \pi_\theta(a' | s')].$$

Here, $d = 1$ for true termination and $d = 0$ otherwise; truncation alone does not stop bootstrapping. The target is computed without gradient tracking.

The base coefficient κ is adapted through empirical target-inclusion control. A center-only target y_{ctr} uses the same backup with $\min_i \bar{Q}_{\tilde{\theta}_i}(s', a')$ as its value term. When $\lambda_{\text{base}}(t) > 0$, the controller measures the fraction of these targets contained in

$$\frac{Q_1(s, a) + Q_2(s, a)}{2} \pm \frac{\lambda_{\text{base}}(t)\kappa}{2} (\tilde{b}_1(s, a) + \tilde{b}_2(s, a)).$$

An EMA of this fraction guides the clipped log-space update of κ toward a target inclusion rate of 0.8. This target-inclusion rate is used only to adapt κ and should not be interpreted as statistical coverage of the true policy value.

Additional critic regularization. Each critic loss includes squared penalties on value magnitudes exceeding an adaptive reward-based threshold and on positive TD residuals, weighted by 10^{-4} and 10^{-3} , respectively.

Actor objective. For each minibatch sample m , the online critics compute pessimistic value estimates at $(s_m, a_{\pi, m})$ using $\kappa_{t, m}^\pi$. Geometric aggregation with ρ^π produces $V_{\text{geo}}^\pi(s_m, a_{\pi, m})$. The minibatch actor loss is

$$J_\pi = \frac{1}{B} \sum_{m=1}^B [\alpha_t^\pi(\beta_{t, m}) \log \pi_\theta(a_{\pi, m} | s_m) - V_{\text{geo}}^\pi(s_m, a_{\pi, m})],$$

where $a_{\pi, m} \sim \pi_\theta(\cdot | s_m)$. The actor uses $\alpha_t^\pi(\beta) = e(t)\alpha(\beta)$, with $e(t) = 1.5$ for $t < 5 \times 10^3$ critic updates and $e(t) = 1$ thereafter. The $\beta_{t, m}$ are sampled conditionally on the modulation scalar $c(t)$ as defined above. Higher $c(t)$ increases their conditional mean.

Update schedule. At each learner update, we update the critics and apply Polyak averaging to the target networks. The coverage controller is updated when active. The actor and entropy-coefficient network are updated every second critic update; t indexes critic updates throughout. The actor-side quantities $c(t)$, λ_t^π , $\beta_{t, m}$ and ρ^π are recomputed only at policy updates.

Entropy coefficient. A small network f_ψ defines $\alpha(\beta) = \exp(\text{clip}(f_\psi(\beta), -3, 2))$ and is optimized using

$$J_\alpha = -\frac{1}{B} \sum_{m=1}^B \log \alpha(\beta_{t, m})(\hat{\ell}_m - A), \quad \hat{\ell}_m = \log \pi_\theta(a_{\pi, m} | s_m).$$

where the target entropy is $-A$. The sampled $\beta_{t, m}$ and policy log-probabilities are treated as constants during this update. Critic targets use $\alpha(0.5)$, whereas actor updates use $\alpha_t^\pi(\beta_{t, m})$ with the early entropy multiplier defined above.

Algorithm 1 summarizes the training procedure.

IV. EXPERIMENTS

A. Experimental Setup

Baselines. We compare GeZo-SAC against six off-policy continuous-control methods: *Soft Actor-Critic* (SAC) [5],

Environment	GeZo-SAC (ours)	TQC	TD3	SAC	GPL-SAC	TDC- λ	SUNRISE
Ant-v5	7.17(0.21)	4.45(2.05)	5.77(0.23)	4.43(1.24)	6.09(0.52)	5.39(1.53)	5.32(0.77)
Hopper-v5	3.77(0.11)	3.44(0.38)	3.52(0.09)	3.51(0.17)	3.38(0.21)	3.59(0.15)	3.63(0.08)
Humanoid-v5	7.89(0.58)	8.61(0.25)	5.23(0.20)	5.60(0.09)	6.13(0.66)	6.85(0.54)	0.62(0.08)
Walker2d-v5	<u>5.74(0.30)</u>	6.22(0.41)	4.61(0.52)	5.25(0.47)	<u>6.06(0.47)</u>	4.99(0.85)	5.79(0.44)

TABLE I: Mean evaluation return (in thousands) across five training runs, using the best checkpoint from each run and 300 deterministic evaluation episodes. Parentheses show the sample standard deviation. Bold and underline denote the best and second-best means. GeZo-SAC uses $K = 15$ in all environments.

Method	Work/m $\downarrow$	Dist. $\uparrow$	TV/m $\downarrow$	Effort/m $\downarrow$	Return $\uparrow$
SAC	1.028	0.817	1.266	1.100	4.70 (0.32)
TD3	1.275	0.786	1.951	1.936	4.77 (0.16)
TQC	1.137	1.283	0.941	0.937	<u>5.68</u> (0.51)
GPL-SAC	0.934	0.988	<u>0.901</u>	<u>0.926</u>	5.42 (0.23)
TDC- λ	1.068	1.002	1.463	1.441	5.21 (0.44)
SUNRISE	1.176	0.751	1.890	2.551	3.84 (0.23)
GeZo-SAC (ours)	0.878	<u>1.242</u>	0.837	0.763	6.14 (0.17)

TABLE II: **Mechanical metrics and return across the four locomotion tasks.** Values are averaged across environments; mechanical metrics are normalized within each environment by its median. Lower is better for Work/m, TV/m, and Effort/m; higher is better for Distance and return. Bold and underline denote the best and second-best values.

TD3 [4], *Truncated Quantile Critics (TQC)* [2], *Generalized Pessimism Learning (GPL-SAC)* [12], *Twin Distributional Critics with λ -LCB (TDC- λ)* [13], and *SUNRISE* [17]. We use the same environment and evaluation protocol across methods, while retaining method-specific critic structures and aggregation mechanisms.

Environments. We evaluate Ant-v5, Hopper-v5, Walker2d-v5, and Humanoid-v5, four locomotion benchmarks from Gymnasium [21].

Training. All experiments are conducted for $T = 3 \times 10^6$ interaction steps per seed, with a maximum episode length of 1000. To ensure a fair and balanced evaluation, we use the same hyperparameters: discount factor $\gamma = 0.99$, Polyak averaging coefficient $\tau = 0.005$, Adam optimizer with learning rate 3×10^{-4} , batch size 256 and buffer size 5×10^5 .

Evaluation Setup. For each method, we perform five independent training runs. During training, policies are evaluated every 50 training episodes using 10 deterministic evaluation episodes. For each run, we select the checkpoint achieving the highest mean return in these evaluations. After training, the selected policy is evaluated over 300 deterministic episodes, using reset seeds different from those used for checkpoint selection.

Metrics. Beyond episodic return, we evaluate overestimation frequency, critic-return discrepancy, and mechanical efficiency.

Value diagnostics. Overestimation frequency is measured as $P(\hat{Q}(s, a_\pi) > MC(s))$, where $MC(s)$ is the discounted Monte-Carlo return obtained from the restored simulator state under the deterministic evaluation policy. Lower values indicate less frequent optimistic estimates, but do not by

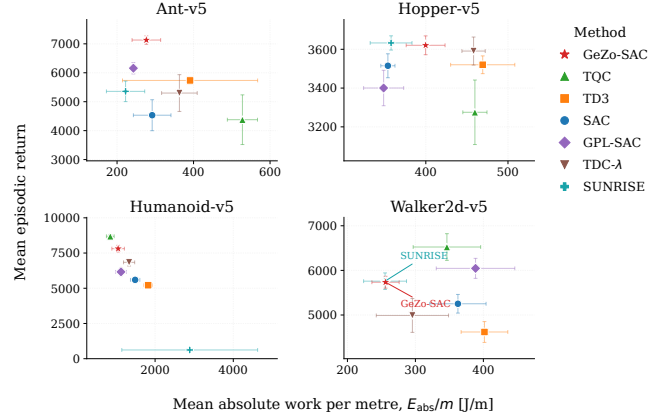


Fig. 2: **Performance–energy trade-off.** Mean episodic return versus mean absolute mechanical work per metre, E_{abs}/m (J/m), across four locomotion tasks. Each marker represents one method; points toward the upper left combine higher return with lower work per metre.

themselves establish greater estimation accuracy. We therefore also report the mean absolute critic–return discrepancy $\mathbb{E}[|\hat{Q}(s, a_\pi) - MC(s)|]$ over the same subset of states.

Mechanical and control metrics. We evaluate locomotion efficiency using absolute actuator work per metre, forward distance, action total variation per metre, and action effort per metre. Let τ_t and $\dot{q}_t$ denote the generalized actuator force and velocity vectors, and $\mathbf{a}_t$ the policy action at step t . For an episode with forward displacement Δx , we compute

$$\frac{E_{\text{abs}}}{m} = \frac{\sum_{t=0}^{T-1} \sum_{d=1}^{n_v} |\tau_t^{(d)} \dot{q}_t^{(d)}| \Delta t}{\max(|\Delta x|, \epsilon)}, \quad (19)$$

$$\text{TV}/m = \frac{\sum_{t=1}^{T-1} \|\mathbf{a}_t - \mathbf{a}_{t-1}\|_1}{\max(|\Delta x|, \epsilon)}, \quad (20)$$

$$\text{Effort}/m = \frac{\sum_{t=0}^{T-1} \|\mathbf{a}_t\|_2^2 \Delta t}{\max(|\Delta x|, \epsilon)}. \quad (21)$$

Lower values indicate lower mechanical cost, smoother control, and lower action effort, respectively. Forward distance is reported separately. For the cross-environment summary, mechanical metrics are normalized within each environment before averaging across tasks.

B. Results

Performance and Training Dynamics. Table I shows that GeZo-SAC achieves the highest mean evaluation return on

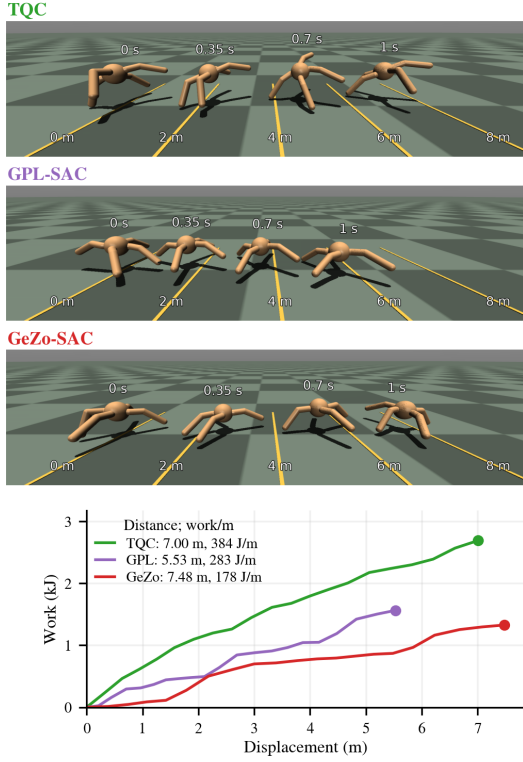


Fig. 3: **Locomotion and mechanical cost on Ant-v5.** **Top:** Overlaid poses of TQC, GPL-SAC, and GeZo-SAC at matched times over a one-second interval, shown at a common spatial scale. **Bottom:** Cumulative actuator work versus forward displacement over the same interval. GeZo-SAC travels farther while requiring less work per metre in these illustrative rollouts.

Ant-v5 and Hopper-v5, while remaining competitive with the strongest baselines on Humanoid-v5 and Walker2d-v5. All GeZo-SAC results use $K = 15$ generators across the four environments.

Figure 1 shows different learning patterns across tasks. On Ant-v5, GPL improves fastest initially, but GeZo-SAC overtakes it and remains among the strongest methods later in training. On Hopper-v5, the leading methods show overlapping learning curves, with GeZo-SAC achieving the highest mean return in the selected-checkpoint evaluation. On Humanoid-v5, TQC achieves the highest returns in later training, with GeZo-SAC generally following ahead of the remaining methods. On Walker2d-v5, GeZo-SAC reaches a comparable return range, although its mean curve generally remains below the leading baselines. SUNRISE remains at low returns on Humanoid-v5, while TDC exhibits a marked decline in return on Hopper-v5 and Walker2d-v5.

Mechanical Cost and Control Quality. GeZo-SAC combines the highest average return with the lowest average actuator work, action variation, and action effort per metre (Table II). It also ranks second in forward distance. These cross-task rankings show that its strong returns are accompanied by lower mechanical and action costs per unit of forward progress.

Figure 2 illustrates the trade-offs within each environment.

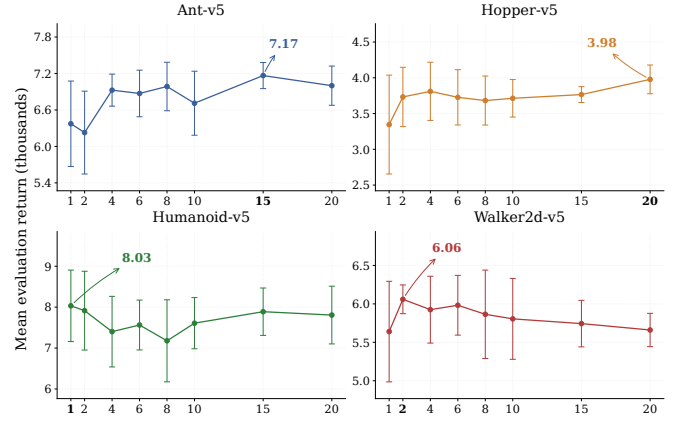


Fig. 4: **Sensitivity to the number of zonotope generators.** Curves show mean evaluation return over five training seeds, with shaded regions and error bars indicating $\pm$ one standard deviation. Bold labels mark the highest observed mean in each environment.

On Ant-v5, GeZo-SAC achieves the highest return while requiring less work per metre than SAC, TD3, TQC, and TDC. On Humanoid-v5, GeZo-SAC and TQC stand out by combining the highest returns with the lowest work per metre. On Walker2d-v5, GeZo-SAC and SUNRISE achieve similar returns at the lowest work per metre, whereas TQC and GPL obtain higher returns at greater mechanical cost. On Hopper-v5, GeZo-SAC reaches returns comparable to the leading methods, using less work per metre than TD3 and TDC but more than SAC and SUNRISE. The aggregate advantage therefore reflects different trade-offs across tasks, with GeZo-SAC offering particularly favorable combinations of return and mechanical cost on Ant-v5 and Humanoid-v5.

Figure 3 also complements these results with a direct comparison of learned locomotion on Ant-v5. Synchronized poses show the agents’ forward progress, while the work–displacement curves track the associated mechanical cost. Over the illustrated one-second interval, GeZo-SAC travels farther than TQC and GPL-SAC while expending less absolute actuator work, achieving the lowest work per metre of the three displayed rollouts.

C. Value Estimation and Policy Performance

Figure 5 compares GeZo-SAC with SAC, TD3, TQC, and GPL-SAC using the mean online critic prediction for each method, without the pessimistic adjustments used during training. Metrics are computed over sampled states from the deterministic evaluation policy, and error bars show variability across training seeds.

GeZo-SAC exhibits near-zero measured overestimation frequency across all four environments. On Ant-v5, it combines the highest mean Monte-Carlo return with lower measured overestimation frequency than SAC and TQC. On Hopper-v5, GeZo-SAC, SAC, TD3, and TQC show near-zero measured overestimation frequency, while GPL-SAC has a higher mean frequency. GeZo-SAC attains the highest

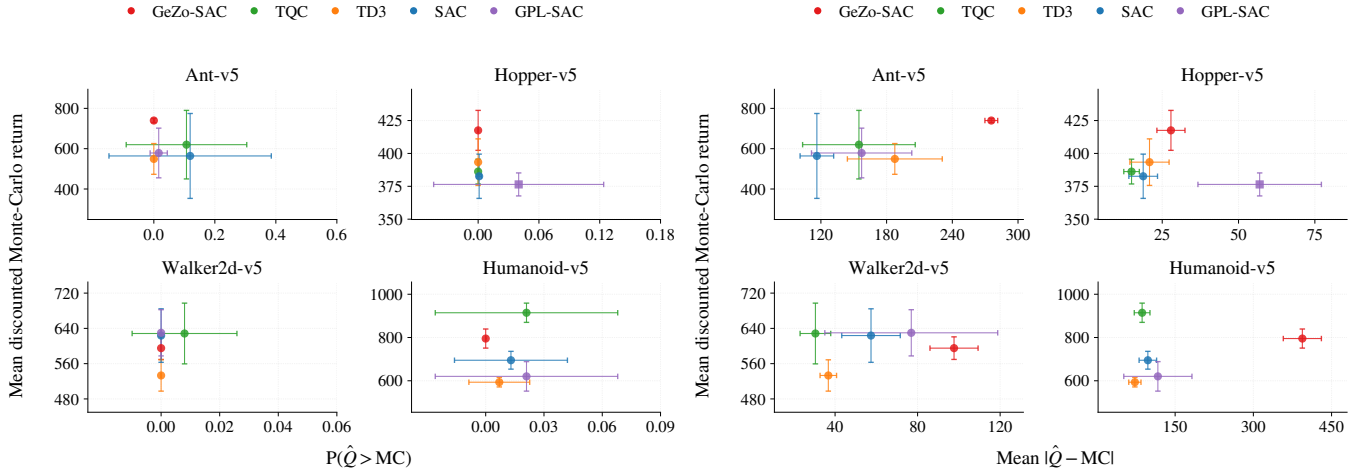


Fig. 5: **Value-estimation diagnostics and policy performance.** Comparison of GeZo-SAC, SAC, TD3, TQC, and GPL-SAC across four locomotion tasks. **Left:** measured overestimation frequency $P(\hat{Q}(s, a_\pi) > MC(s))$. **Right:** mean absolute critic–return discrepancy $\mathbb{E}[|\hat{Q}(s, a_\pi) - MC(s)|]$. The vertical axis shows mean discounted, reward-only Monte-Carlo return under deterministic evaluation. Markers average policy-level metrics across training seeds; horizontal and vertical error bars show $\pm$ one sample standard deviation.

mean Monte-Carlo return. On Humanoid-v5, TQC achieves a higher return despite more frequent overestimation.

The absolute discrepancies reveal a different ordering. GeZo-SAC has the largest mean critic–return discrepancy on Ant-v5, Humanoid-v5, and Walker2d-v5, despite achieving the highest return on Ant-v5 and the second-highest on Humanoid-v5. Absolute agreement with Monte-Carlo returns therefore does not directly predict policy quality. An action-independent offset, for example, changes absolute value errors without changing action preferences. On Walker2d-v5, however, several baselines combine smaller discrepancies with higher returns. These results show that strong policy performance can coexist with substantial critic–return discrepancies.

D. Ablation: Number of Zonotope Generators

We vary the number of generators $K \in \{1, 2, 4, 6, 8, 10, 15, 20\}$ across all four environments, keeping the remaining hyperparameters fixed and using five independent training seeds per configuration.

Figure 4 shows that the effect of K is task-dependent. Ant-v5 and Hopper-v5 attain their highest mean returns with larger generator sets, at $K = 15$ and $K = 20$, respectively. Larger configurations also generally show less variability across seeds on these tasks. In contrast, Humanoid-v5 and Walker2d-v5 achieve their highest mean returns at $K = 1$ and $K = 2$, with no consistent benefit from additional generators. Increasing the generator count therefore does not uniformly improve performance. All main comparisons use a shared setting of $K = 15$ across the four environments.

E. Ablation: Fixed versus Adaptive Pessimism

We compare fixed $\kappa = 1$, near-zero $\kappa = 10^{-6}$, and adaptive κ , using $K = 15$. Each setting is trained for 3×10^6 environment steps with three independent seeds per environment

Method	Act. (M)	Crit. (M)	Train (ms)	Infer. (ms)
SAC	0.101	0.199	9.83	0.181
TD3	0.099	0.199	5.25	0.161
TQC	0.101	0.527	16.90	0.181
GPL-SAC	0.101	1.656	41.62	0.662
TDC- λ	0.099	0.231	6.65	0.153
SUNRISE	0.506	0.993	55.63	0.900
GeZo-SAC (ours)	0.101	1.133	15.71	0.181

TABLE III: **Computational overhead.** Parameter counts include the actor and all online critics. Timings calculated on an NVIDIA A100-SXM-80GB.

(Figs. 6–7). In the displayed runs, adaptive κ achieves higher returns on Ant-v5 and Humanoid-v5. Hopper-v5 shows overlapping training curves but the highest final evaluation return with adaptation, whereas near-zero κ performs best on Walker2d-v5. Figure 7 also illustrates the evolution of the adaptive pessimism coefficient on Hopper-v5. Starting from its initial value, the controller decreases κ substantially during early training and subsequently stabilizes at an intermediate level, in contrast to the fixed coefficients used in the ablations.

F. Computational Cost

Table III compares model size and runtime across methods. GeZo-SAC increases critic capacity relative to SAC, but its training cost remains comparable to TQC and substantially below larger ensemble-based methods such as GPL-SAC and SUNRISE. Importantly, the additional zonotope representation is confined to the critics. The deployed actor therefore has the same architecture and inference latency as SAC, so the extra computational cost is incurred only during training.

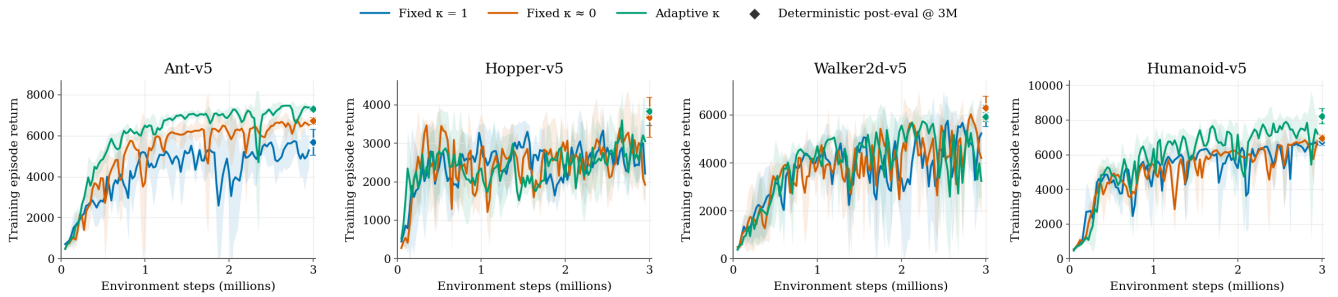


Fig. 6: **Effect of fixed versus adaptive pessimism.** Evaluation returns for GeZo-SAC with fixed $\kappa = 1$, fixed $\kappa \approx 0$, and the adaptive κ controller across the four locomotion tasks. Curves show the mean over three training runs, with shading indicating $\pm$ one sample standard deviation. Diamonds mark the final deterministic evaluation over 300 episodes per policy, with error bars showing the standard deviation across runs.

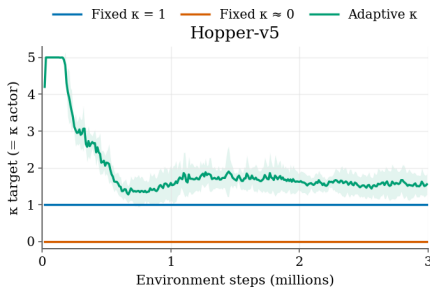


Fig. 7: **Evolution of the adaptive pessimism coefficient on Hopper-v5.** The adaptive controller changes κ throughout training, while the two ablated variants keep it fixed. The curve shows the mean across three training seeds, with shading indicating $\pm$ one standard deviation.

V. CONCLUSIONS

We presented GeZo-SAC, which augments SAC’s twin critics with auxiliary zonotope generators whose directional supports provide heuristic geometric bands and a disagreement signal for adapting pessimism and critic aggregation. Across four MuJoCo-v5 tasks, GeZo-SAC achieves the highest mean return on Ant-v5 and Hopper-v5, outperforms SAC and TD3 on every task, and attains the lowest average work and action effort per metre, with near-zero measured overestimation frequency. In the future we also plan to extend and evaluate GeZo-SAC to robotic manipulation tasks.

REFERENCES

- [1] S. Thrun and A. Schwartz, “Issues in using function approximation for reinforcement learning,” in *Proceedings of the Fourth Connectionist Models Summer School*. Hillsdale, NJ: Lawrence Erlbaum, Dec. 1993.
- [2] A. Kuznetsov, P. Shvechikov, A. Grishin, and D. Vetrov, “Controlling overestimation bias with truncated mixture of continuous distributional quantile critics,” in *Proceedings of the 37th International Conference on Machine Learning*, 2020.
- [3] V. Mnih, K. Kavukcuoglu, D. Silver, A. Graves, I. Antonoglou, D. Wierstra, and M. Riedmiller, “Playing atari with deep reinforcement learning,” *arXiv preprint arXiv:1312.5602*, 2013.
- [4] S. Fujimoto, H. van Hoof, and D. Meger, “Addressing function approximation error in actor-critic methods,” in *Proceedings of the 35th International Conference on Machine Learning*, 2018.
- [5] T. Haarnoja, A. Zhou, P. Abbeel, and S. Levine, “Soft actor-critic: Off-policy maximum entropy deep reinforcement learning with a stochastic actor,” *arXiv preprint arXiv:1801.01290*, 2018.
- [6] P. Maragos and E. Theodosis, “Multivariate tropical regression and piecewise-linear surface fitting,” in *Proc. IEEE ICASSP*, 2020.
- [7] V. Charisopoulos and P. Maragos, “Morphological Perceptrons: Geometry and Training Algorithms,” in *Mathematical Morphology and Its Applications to Signal and Image Processing - ISMM 2017*, ser. LNCS, J. A. et al., Ed., vol. 10225. Springer, 2017, pp. 3–15.
- [8] —, “A Tropical Approach to Neural Networks with Piecewise Linear Activations,” *arXiv:1805.08749*, 2018.
- [9] X. Zhang, G. Naitzat, and L.-H. Lim, “Tropical geometry of deep neural networks,” in *Proceedings of the 35th International Conference on Machine Learning*, 2018.
- [10] P. Misiakos, G. Smyrnis, G. Retsinas, and P. Maragos, “Neural network approximation based on hausdorff distance of tropical zonotopes,” in *International Conference on Learning Representations (ICLR)*, 2022.
- [11] H. van Hasselt, A. Guez, and D. Silver, “Deep reinforcement learning with double q-learning,” *arXiv preprint arXiv:1509.06461*, 2016. [Online]. Available: <https://arxiv.org/abs/1509.06461>
- [12] E. Cetin and O. Celiktutan, “Learning pessimism for reinforcement learning,” in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 37, no. 6, 2023, pp. 6971–6979.
- [13] O. Osman, B. Yalcin Kavus, T. K. Karaca, G. Tulum, and M. A. Karabulut, “Risk sensitive twin distributional critics with a lambda lower confidence bound for continuous control reinforcement learning,” *Scientific Reports*, vol. 16, p. 6699, 2026.
- [14] A. Kumar, A. Zhou, G. Tucker, and S. Levine, “Conservative q-learning for offline reinforcement learning,” in *Advances in Neural Information Processing Systems*, ser. NeurIPS, 2020.
- [15] K. Ciosek, Q. Vuong, R. Loftin, and K. Hofmann, “Better exploration with optimistic actor-critic,” in *Advances in Neural Information Processing Systems (NeurIPS)*, vol. 32, 2019.
- [16] J. Duan, W. Wang, L. Xiao, J. Gao, S. E. Li, C. Liu, Y.-Q. Zhang, B. Cheng, and K. Li, “Distributional soft actor-critic with three refinements,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 47, no. 5, pp. 3935–3946, 2025.
- [17] K. Lee, M. Laskin, A. Srinivas, and P. Abbeel, “SUNRISE: A simple unified framework for ensemble learning in deep reinforcement learning,” in *Proceedings of the 38th International Conference on Machine Learning*, vol. 139, 2021, pp. 6131–6141.
- [18] G. C. Calafiore, S. Gaubert, and C. Possieri, “Log-sum-exp neural networks and posynomial models for convex and log-log-convex data,” *IEEE Transactions on Neural Networks and Learning Systems*, 2019.
- [19] —, “A universal approximation result for difference of log-sum-exp neural networks,” *IEEE Transactions on Neural Networks and Learning Systems*, 2020.
- [20] K. Fotopoulos, P. Maragos, and P. Misiakos, “TropNNC: Structured neural network compression using tropical geometry,” *Transactions on Machine Learning Research*, Jun. 2026. [Online]. Available: <https://openreview.net/forum?id=u7DRq1icmY>
- [21] M. Towers, A. Kwiatkowski, J. Terry, J. U. Balis, G. De Cola, T. Deleu, M. Goulão, A. Kallinteris, M. Krimmel, A. KG et al., “Gymnasium: A standard interface for reinforcement learning environments,” *arXiv preprint arXiv:2407.17032*, 2024.