MAGICS: Adversarial RL with Minimax Actors Guided by Implicit Critic Stackelberg for Convergent Neural Synthesis of Robot Safety

Justin Wang[1⋆], Haimin Hu[1⋆], Duy P. Nguyen[1], and Jaime Fernández Fisac[1]

Department of Electrical and Computer Engineering, Princeton University, USA
{jw4406,haiminh,duyn,jfisac}@princeton.edu

Abstract. While robust optimal control theory provides a rigorous framework to compute robot control policies that are provably safe, it struggles to scale to high-dimensional problems, leading to increased use of deep learning for tractable synthesis of robot safety. Unfortunately, existing neural safety synthesis methods often lack convergence guarantees and solution interpretability. In this paper, we present Minimax Actors Guided by Implicit Critic Stackelberg (MAGICS), a novel adversarial reinforcement learning (RL) algorithm that guarantees local convergence to a minimax equilibrium solution. We then build on this approach to provide local convergence guarantees for a general deep RL-based robot safety synthesis algorithm. Through both simulation studies on OpenAI Gym environments and hardware experiments with a 36-dimensional quadruped robot, we show that MAGICS can yield robust control policies outperforming the state-of-the-art neural safety synthesis methods.

Keywords: adversarial reinforcement learning · robot safety · game theory

1 Introduction

The widespread deployment of autonomous robots calls for robust control methods to ensure their reliable operation in diverse environments. Safety filters [21] have emerged as an effective approach to ensure safety for a broad spectrum of robotic systems, such as autonomous driving [48,45,25,24], legged locomotion [23,1,20,38], and aerial robots [14,43,49,8]. Model-based numerical safety synthesis methods [2,6] offer verifiable safety guarantees, but can only scale up to 5-6 state variables. In order to develop safety filters at scale, recent research efforts have been dedicated towards using neural representations of robot policies [15,41,3,22], showing potential for safety filtering with tens [38,20] even hundreds of state variables [26].

While neural safety filters have shown significant progress in scalability, they are often challenging to analyze due to their black-box nature. Moreover, in adversarial settings, naïve training of agent policies can lead to severe oscillatory behaviors, preventing the algorithm from converging. Recently developed game-theoretic machine learning algorithms guarantee convergence to mathematically meaningful equilibrium solutions (*e.g.*, Nash or Stackelberg), even for training of black-box models such as deep

⋆ J. Wang and H. Hu contributed equally.

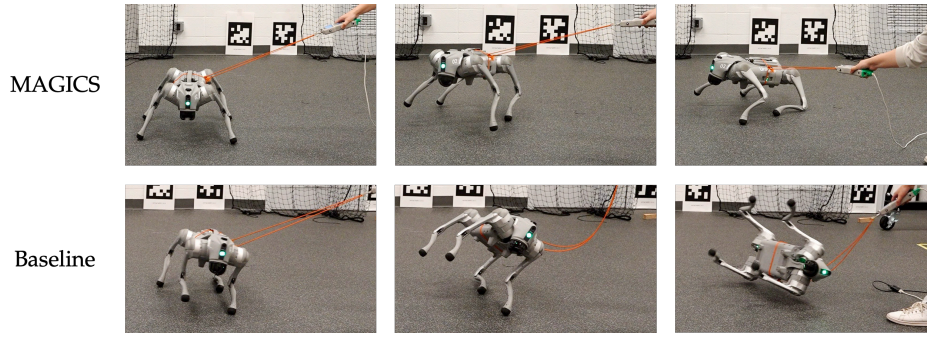


Fig. 1: Comparing to a non-game baseline robust RL, our proposed game-theoretic adversarial RL algorithm yields a control policy consistently more robust when applied to safe quadrupedal locomotion and stress-tested with varying tugging forces.

neural networks [11,29,51,50]. However, providing *convergence guarantees* for training robust *neural* control policies with zero-sum (adversarial) formulations remains an outstanding challenge.

Statement of Contributions. This paper introduces Minimax Actors Guided by Implicit Critic Stackelberg (MAGICS), a novel game-theoretic deep adversarial RL algorithm that guarantees local convergence to a feedback Stackelberg equilibrium strategy. Building on this framework, we propose, for the first time, a provably convergent neural synthesis algorithm for approximate safe analysis in high dimensions. In addition to our theoretical contributions, we demonstrate empirically that MAGICS yields robust control policies consistently outperforming the prior state-of-the-art neural robust controller in both simulations and hardware tests.

2 Related Work

Game-Theoretic Learning Algorithms. Recent years have seen significant progress in machine learning problems formulated as games, such as generative adversarial networks (GANs) [17], adversarial RL [39], and hyperparameter optimization [35]. Those problems involve interacting agents with coupled and potentially competing objectives, which calls for careful design of the training algorithm. Fiez et al. [11] leverages the implicit function theorem to derive a provably convergent gradient-based algorithm for machine learning problems formulated as Stackelberg games. In their follow-up work [13], they show that the simple gradient descent-ascent with finite timescale separation (τ-GDA) algorithm guarantees local convergence to a minimax solution in zero-sum settings, a result also reported by concurrent work [29]. More recent work [36] shows that convergence can be achieved in Stackelberg learning only with first-order gradient information. A closely related topic to Stackelberg learning is the exploitability of suboptimal strategies. Lei et al. [33] use unsupervised learning to estimate the exploited level of each trajectory in a dataset, which is then used in an offline learning process to maximize the influence of the dominant strategy.

Our work builds on insights from the game-theoretic learning community and provides a novel, yet rigorous convergence analysis for adversarial RL.

Adversarial Reinforcement Learning. Multi-agent robust policy synthesis is increasingly solved by adversarial RL methods due to their scalability. Pinto et al. [39] pioneered the idea of robust adversarial reinforcement learning (RARL), which improves robustness by jointly training a stabilizing policy and a destabilizing adversary. However, this approach ignores the coupling between the agents' objectives. Learning with Opponent-Learning Awareness (LOLA) [16] explicitly captures the coupled interests among agents in multi-agent RL by accounting for the influence of one agent's policy on the predicted parameter of the other agents. Huang et al. [27] model policy-gradient-based adversarial RL as a Stackelberg game and apply the Stackelberg learning algorithm from Fiez et al. [11] to solve the RL problem. Although these game-inspired adversarial RL algorithms have shown promising improvement over non-game baselines, they typically lack provable convergence guarantees. Recent work by Zheng et al. [50] models the hierarchical interaction between the actor and critic in single-agent actor-critic-based RL [30,18] as a two-player general-sum game, and provides a gradient-based algorithm that is locally convergent to a Stackelberg equilibrium.

Our work combines the best of both worlds by building on the state-of-the-art game-theoretic RL [50] and adversarial learning [13] approaches to address the missing piece of provably convergent *multi-agent* adversarial RL in continuous state–action spaces.

Data-Driven Safety Filters for Robotics. Deep learning has enabled scalable synthesis of robust control policies. Rapid Motor Adaptation (RMA) [31] uses deep RL to learn a base policy, and supervised learning to train an online adaptation module, allowing legged robots to quickly and robustly adapt to novel environments. Lai et al. [32] propose to train a transformer model for quadrupedal locomotion across diverse terrains. Specially focused on safety, data-driven safety filters [21,47] aim at providing scalable safety assurances. Robey et al. [41] use control barrier function (CBF) constraints as self-supervision signals to learn a CBF from expert demonstrations. A similar approach [10] learns a CBF parameterized as a deep neural network from labeled safe and unsafe state samples. Another line of work focuses on neural approximation of safety Bellman equations. Safety RL [15] proposes to modify Hamilton–Jacobi equation [2] with contraction mapping, rendering RL suitable for approximate safety analysis. Hsu et al. [22] extend the single-agent safety RL to the two-player zero-sum setting by offline co-training a best-effort safety controller and a worst-case disturbance policy. They use model-based rollout to verify safety at runtime. However, existing multi-agent neural safety synthesis approaches predominantly lack convergence guarantees, which can make the training notoriously difficult and time-consuming.

To the best of our knowledge, our proposed game-theoretic deep adversarial RL algorithm is the first *provably-convergent* multi-agent neural safety synthesis approach.

3 Preliminaries and Problem Formulation

Notation. Given a function f, we denote $D_\theta f$ as the total derivative of f with respect to θ, $\nabla_\theta f$ as the partial derivative of f with respect to θ, $\nabla_{\theta\psi} f := \partial^2 f / \partial\theta\partial\psi$, and $\nabla^2_\theta f := \partial^2 f / \partial\theta^2$. We denote $\|\cdot\|$ as the 2-norm of vectors and the spectral norm of matrices. We indicate matrix A is positive and negative definite with $A \succ 0$ and $A \prec 0$. We assume throughout that all functions f are smooth, *i.e.*, $f \in C^q$ for some $q \geq 2$.

Zero-Sum Dynamic Games. We consider discrete-time nonlinear dynamics that describe the robot motion:

$$x_{t+1} = f(x_t, u_t, d_t), \tag{1}$$

where $x_t \in \mathcal{X} \subseteq \mathbb{R}^n$ is the state, $u_t \in \mathcal{U} \subset \mathbb{R}^{n_u}$ is the controller input (belongs to the ego robot), $d_t \in \mathcal{D} \subset \mathbb{R}^{n_d}$ is the disturbance input, and $f : \mathcal{X} \times \mathcal{U} \times \mathcal{D} \to \mathcal{X}$ is a continuous nonlinear function that describes the physical system dynamics. We model the *adversarial* interaction between the controller and disturbance as a *zero-sum dynamic game* with objective $J^{\pi^u,\pi^d}(\mathbf{x})$, where $\pi^u, \pi^d : \mathcal{X} \to \mathcal{U}$ are control and disturbance policies, respectively, $\mathbf{x} := \mathbf{x}_{x,T}^{\pi^u,\pi^d} = (x_0, x_1, \ldots, x_{T-1})$ denotes the state trajectory starting from $x_0 = x$ under dynamics (1), policies π^u, and π^d. In this paper, we seek to compute a control policy π^u and disturbance policy π^d that constitute a *feedback Stackelberg equilibrium (FSE)* of a zero-sum dynamic game, where π^u maximizes $J^{\pi^u,\pi^d}(\mathbf{x})$ and π^d minimizes $J^{\pi^u,\pi^d}(\mathbf{x})$. This is formalized in the following definition.

Definition 1 (Feedback Stackelberg Equilibrium). *The strategy pair $(\pi^{u,*}, \pi^{d,*})$ is a feedback Stackelberg equilibrium if*

$$\inf_{\pi^d \in r(\pi^{u,*})} J^{\pi^{u,*},\pi^d}(\mathbf{x}) \geq \inf_{\pi^d \in r(\tilde{\pi}^u)} J^{\tilde{\pi}^u,\pi^d}(\mathbf{x}), \quad \forall \tilde{\pi}^u \in \Pi^u, \tag{2}$$

and $\pi^{b,*} \in r(\pi^{u,*})$, where $r(\pi^u) := \{\pi^d \in \Pi^d \mid J^{\pi^u,\tilde{\pi}^d} \geq J^{\pi^u,\pi^d}, \forall \tilde{\pi}^d \in \Pi^d\}$ is the optimal response map of the follower. Here, Π^u and Π^d are the sets of control and disturbance policies, respectively.

Definition 1 is aligned with our interest in *worst-case* robot safety analysis: by assigning the controller as the leader and disturbance the follower, we give the disturbance the *instantaneous information advantage* [28]. Note that such worst-case analysis leads to a *least-restrictive* safety filter [21, Sec. 2.3] in that it maintains system safety for all possible disturbance realizations, only overriding nominal control actions when they are incapable of preventing an imminent safety failure.

An FSE can be characterized using the first- and second-order optimality conditions (sufficient conditions of those in Definition 1), leading to the notion of a differential Stackelberg equilibrium (DSE) [11], which can be verified more easily. To this end, we adopt a finite-dimensional parameterization of players' policies: $\pi^u = \pi^u_\theta$ and $\pi^d = \pi^d_\theta$, where $\theta \in \mathbb{R}^{n_\theta}$ and $\psi \in \mathbb{R}^{n_\psi}$. In Section 4, we consider the neural representation of policies, *i.e.*, π^u_θ and π^d_θ are deep neural networks. We denote the resulting game objective value as $J^{\pi^u_\theta,\pi^d_\psi}(\theta, \psi, \mathbf{x})$. We recall that, in zero-sum games, a DSE is equivalent to a local minimax equilibrium.

Definition 2 (Strict Local Minimax Equilibrium [11,12,29]). *Strategy pair $(\pi^u_{\theta^*}, \pi^d_{\psi^*})$ is a differential Stackelberg equilibrium of a zero-sum dynamic game if conditions $\nabla_\theta J^{\pi^u_{\theta^*},\pi^d_{\psi^*}}(\theta^*, \psi^*, \mathbf{x}) = 0$, $\nabla_\psi J^{\pi^u_{\theta^*},\pi^d_{\psi^*}}(\theta^*, \psi^*, \mathbf{x}) = 0$, $\mathtt{S}_1(\mathbf{J}_J(\theta^*, \psi^*, \mathbf{x})) \prec 0$, and $\nabla^2_\psi(\mathbf{J}_J(\theta^*, \psi^*, \mathbf{x})) \succ 0$ hold, where $\mathbf{J}_J(\theta, \psi, \mathbf{x})$ denotes the Jacobian of the* individual *gradient vector:*

$$\mathbf{J}_J(\theta, \psi, \mathbf{x}) = \begin{bmatrix} -\nabla^2_\theta(\theta, \psi, \mathbf{x}) & -\nabla_{\theta\psi}(\theta, \psi, \mathbf{x}) \\ \nabla^\top_{\theta\psi}(\theta, \psi, \mathbf{x}) & \nabla^2_\psi(\theta, \psi, \mathbf{x}) \end{bmatrix}, \tag{3}$$

and $\mathtt{S}_1(\mathbf{J}_J)$ denotes the Schur complement of $\mathbf{J}_J$ with respect to its lower-right block.

Reach–Avoid Robot Safety Analysis. We consider the *worst-case* reach–avoid safety analysis for robot dynamics (1). We define the ego robot's target and failure sets as $\mathcal{T} := \{x \mid \ell(x) \geq 0\} \subseteq \mathbb{R}^n$ and $\mathcal{F} := \{x \mid g(x) < 0\} \subseteq \mathbb{R}^n$, where $\ell(\cdot)$ and $g(\cdot)$ are Lipschitz continuous *margin functions*, encoding problem-specific safety and liveness specifications. We use Hamilton–Jacobi–Isaacs (HJI) reachability analysis to capture the interplay between the best-effort controller policy π^u, *i.e.*, one that attempts to reach the target set $\mathcal{T}$ without entering the failure set $\mathcal{F}$, and worst-case disturbance policy π^d, *i.e.*, one that prevents the controller from succeeding. To this end, we formulate an infinite-horizon *zero-sum* dynamic game with the following objective functional:

$$J_k^{\pi^u,\pi^d}(x) := \max_{\tau \geq k} \min \left\{ \ell(x_\tau), \min_{s \in [k,\tau]} g(x_s) \right\}. \tag{4}$$

The game's minimax solution satisfies the fixed-point Isaacs equation [28]:

$$V(x) = \min \left\{ g(x), \max \left\{ \ell(x), \max_{u \in \mathcal{U}} \min_{d \in \mathcal{D}} V\big(f(x,u,d)\big) \right\} \right\}. \tag{5}$$

Value function $V(\cdot)$ encodes the *reach–avoid set* $\mathcal{RA}(\mathcal{T},\mathcal{F}) := \{x \mid V(x) \geq 0\}$, from which the ego agent is guaranteed a policy to safely reach the target set without entering the failure set. Note that (5) implies that the disturbance has the instantaneous informational advantage and $V(x) = \max_{\pi^u} \min_{\pi^d} J^{\pi^u,\pi^d}(x)$ is known as the *lower* value of the zero-sum game (4). It corresponds to the minimal safety margin g that the controller is able to maintain *at all times* under the worst-case disturbance.

Unfortunately, the time and memory complexity required for numerically solving Isaacs equation (5) is exponential with respect to the dimension of the state space, rendering grid-based dynamic programming approaches impractical for realistic robot control problems. The recently developed—prior state-of-the-art neural safety analysis algorithm—Iterative Soft Adversarial Actor–Critic for Safety (ISAACS) [22] uses adversarial RL to approximately solve the Isaacs equation (5). It alternates between updating the controller and disturbance policy parameters based on *individual* gradients of the players' objectives, which may lead to non-convergent training behaviors or stuck at a poor saddle point. In Section 5, we modify ISAACS with our proposed game-theoretic RL formalism in Section 4 so that it is provably convergent to a minimax equilibrium.

4 Approach: Stackelberg–Minimax Adversarial RL

Soft Adversarial Actor–Critic as a Three-Player Game. We consider a discrete-time adversarial Markov game governed by system (1). The initial state is determined by a given prior distribution $x_0 \sim p_0(x)$. At time t, the controller and disturbance take actions according to their stochastic policies, *i.e.*, $u_t \sim \pi_\theta^u(\cdot|x_t)$, $d_t \sim \pi_\psi^d(\cdot|x_t)$, the controller receives a bounded reward $r_t = r(x_t,u_t,d_t)$ emitted from the environment, and the disturbance receives reward $-r_t$. The single-player version of this Markov game (*i.e.*, with $d_t \equiv 0$) can be solved at scale using deep RL approaches. We will focus on both *on- and off-policy actor-critic* methods; specifically Advantage Actor-Critic (A2C) [37] and Soft Actor–Critic (SAC) [18], respectively. In the following, we present

adversarial variants of the A2C and SAC algorithms. We assume that the critic and actors are deep neural networks parameterized by (ω, θ, ψ). The critic aims to minimize the mean-square Bellman error loss

$$L(\omega, \theta, \psi) = \mathop{\mathbb{E}}_{x \sim \mathbf{x}} \left[(V_\omega(x) - V^\pi(x))^2 \right], \tag{6}$$

or

$$L(\omega, \theta, \psi) = \mathop{\mathbb{E}}_{\xi \sim \mathcal{B}} \left[(Q_{\omega_1}(x, u, d) - r - \gamma Q_{\omega_2}(x', u', d'))^2 \right], \tag{7}$$

for A2C and SAC, respectively, in which V is the value function, Q is the state–action value function, $\omega = (\omega_1, \omega_2) \in \mathbb{R}^{n_\omega}$ is the critic's parameter, $\xi = (x, u, d, r, x')$ is the transition data, $\mathcal{B}$ is a replay buffer, $\gamma \in (0, 1]$ is a discount factor, $u' \sim \pi^u_\theta(\cdot|x')$, $d' \sim \pi^d_\psi(\cdot|x')$, and the function $V^\pi(x)$ is approximated through a bootstrapped estimator (here we use generalized advantage estimation [42]). For A2C, the controller seeks to maximize the objective

$$J(\omega, \theta, \psi) = \mathop{\mathbb{E}}_{\mathbf{x} \sim (\pi^u, \pi^d)} \left[r(x_0, u_0, d_0) + V_\omega(x_1) \right], \tag{8}$$

while for SAC, the controller seeks to maximize entropy-regularized control objective

$$J(\omega, \theta, \psi) = \\ \mathop{\mathbb{E}}_{x \sim \mathcal{B}} \left[\min_{i \in \{1,2\}} Q_{\omega_i}(x, \tilde{u}, \tilde{d}) - \eta^u \left[\log \pi^u_\theta(\tilde{u}|x) - H^u_0 \right] + \eta^d \left[\log \pi^d_\psi(\tilde{d}|x) - H^d_0 \right] \right], \tag{9}$$

where $\tilde{u} \sim \pi^u_\theta(\cdot|x)$, $\tilde{d} \sim \pi^d_\psi(\cdot|x)$, $\eta^u, \eta^d > 0$ are the entropy regularization coefficients, and H^u_0 and H^d_0 are the minimum entropy (heuristically set to the dimension of the player action space, following [19]) of controller and disturbance policies, respectively. Finally, the disturbance minimizes $J(\omega, \theta, \psi)$.

We cast MAGICS training procedure defined by (7)-(9) as a *three-player* general-sum (*i.e.*, non-cooperative) Stackelberg game, which is modeled by the following *trilevel* optimization problem:

$$\min_{\omega} \quad L(\omega, \theta, \psi) \tag{10a}$$

$$\text{s.t.} \quad \theta \in \arg\max_{\tilde{\theta}} J(\omega, \tilde{\theta}, \bar{\psi}), \tag{10b}$$

$$\text{s.t.} \quad \bar{\psi} \in \arg\min_{\bar{\psi}} J(\omega, \theta, \bar{\psi}), \tag{10c}$$

where the critic is the leader, followed by the controller, and the disturbance plays last, consistent with its information advantage encoded in (5). The most natural way of viewing the three-player game is as follows: in line with [12], the two primary players are the controller u and disturbance d, while the critic sits in judgment of their gameplay.

Solving the Zero-Sum Actor Game with τ-GDA. In this section, we provide a subroutine that solves the inner zero-sum game between the two actors (10b)-(10c) using τ-GDA [12,13] for a fixed critic parameter ω. In essence, this approach scales up the learning rate of the follower so that it can adapt faster, which leads to guaranteed local convergence to a minimax solution of the zero-sum game (10b)-(10c). This subroutine is summarized in Algorithm 1.

Algorithm 1: τ-GDA for Zero-Sum Actor Game

Input: actor parameters (θ_0, ψ_0), critic parameter ω, controller learning rate schedule $\{\alpha_i^u\}_{i=0,1,\ldots}$, learning rate ratio τ_a

$i \leftarrow 0$
while *Not converged* **do**
  $\theta_{i+1} \leftarrow \theta_i + \alpha_i^u \nabla_\theta J(\omega, \theta_i, \psi_i)$ // Updates controller parameters
  $\psi_{i+1} \leftarrow \psi_i - \alpha_i^u \tau_a \nabla_\psi J(\omega, \theta_i, \psi_i)$ // Updates disturbance parameters
  $i \leftarrow i + 1$

Remark 1. While we focus on the local convergence analysis for game-theoretic adversarial RL involving optimization of neural network parameters, recent work [12] shows that, in a zero-sum game, if the maximizing player has a Polyak-Łojasiewicz (PŁ) or strongly-concave (SC) objective, then τ-GDA can converge globally to a strict local minimax equilibrium.

MAGICS: Minimax Actors Guided by Implicit Critic Stackelberg. We now return to the full game (10). Since the controller and the disturbance update their policy parameters using τ-GDA, the critic sees them as two *simultaneous* (*i.e.*, Nash) followers: the controller performs gradient ascent on return $J(\omega, \theta, \psi)$, and the disturbance performs gradient descent on $\tau_a J(\omega, \theta, \psi)$. In the following, we derive the update rule for the MAGICS. We start by providing the critic's Stackelberg learning dynamics [11]. Invoking the implicit function theorem, the total derivative of the critic's cost $D_\omega L(\omega, \theta^*, \psi^*)$ at a minimax equilibrium (θ^*, ψ^*) is:

$$D_\omega L(\omega, \theta^*, \psi^*) = \nabla_\omega L(\omega) + \nabla_\omega \theta^*(\omega) \nabla_\theta L(\omega, \theta^*, \psi^*) + \nabla_\omega \psi^*(\omega) \nabla_\psi L(\omega, \theta^*, \psi^*),$$

where $\theta^*(\omega) = r_\theta(\omega)$ and $\psi^*(\omega) = r_\psi(\omega)$ are the controller's and disturbance's rational response to the critic, respectively. The implicit differentiation terms $\nabla_\omega \theta^*(\omega)$ and $\nabla_\omega \psi^*(\omega)$ can be computed by solving the following linear system of equations:

$$\begin{cases} 0 = \nabla_{\theta\omega} J(\omega, \theta^*, \psi^*) + \nabla_\theta^2 J(\omega, \theta^*, \psi^*) \nabla_\omega \theta^*(\omega) + \nabla_{\theta\psi} J(\omega, \theta^*, \psi^*) \nabla_\omega \psi^*(\omega), \\ 0 = \nabla_{\psi\omega} J(\omega, \theta^*, \psi^*) + \nabla_{\psi\theta} J(\omega, \theta^*, \psi^*) \nabla_\omega \theta^*(\omega) + \nabla_\psi^2 J(\omega, \theta^*, \psi^*) \nabla_\omega \psi^*(\omega). \end{cases}$$

Compactly, the total derivative of the critic can be written as:

$$D_\omega L(\omega, \theta, \psi) = \nabla_\omega L(\omega, \theta, \psi) - h_1(\omega, \theta, \psi)^\top H(\omega, \theta, \psi)^{-1} h_2(\omega, \theta, \psi), \tag{11}$$

where

$$h_1(\omega, \theta, \psi) = \begin{bmatrix} \nabla_{\omega\theta} J(\omega, \theta, \psi) \\ \nabla_{\omega\psi} J(\omega, \theta, \psi) \end{bmatrix}, \quad h_2(\omega, \theta, \psi) = \begin{bmatrix} \nabla_\theta L(\omega, \theta, \psi) \\ \nabla_\psi L(\omega, \theta, \psi) \end{bmatrix},$$

and

$$H(\omega, \theta, \psi) = \begin{bmatrix} \nabla_\theta^2 J(\omega, \theta, \psi) & \nabla_{\theta\psi} J(\omega, \theta, \psi) \\ \nabla_{\psi\theta} J(\omega, \theta, \psi) & \nabla_\psi^2 J(\omega, \theta, \psi) \end{bmatrix}.$$

Each term in the Stackelberg gradient update rule can be computed directly and estimated by samples.

For MAGICS-A2C, the objective functions are defined in expectation over the distribution induced by the players' current policies. Hence, custom gradient estimators are required, as the Stackelberg gradient of (6) is not straightforward. The estimator is given below, and the proof merely requires expanding the value and state–action value function definitions recursively.

Theorem 1. *Given a Markov game with actor parameters* (θ, ψ) *and shared critic parameters* ω*, if critic has objective function* $L(\omega, \theta, \psi)$ *defined in* (6)*, then* $\nabla_\theta L(\omega, \theta, \psi)$ *is given by*

$$\nabla_\theta L(\omega, \theta, \psi) = \mathop{\mathbb{E}}_{\mathbf{x}\sim(\theta,\psi)} \left[2 \sum_{t=0}^{T} \gamma^t \nabla_\theta \log \pi_\theta^u(u_t|x_t) \left(V_\omega(x_0) - V^\pi(x_0)\right) Q^\pi(x_t, u_t, d_t) \right]. \tag{12}$$

MAGICS-SAC, on the other hand, is an off-policy RL scheme, with critic's and actors' objective defined as an expectation over an arbitrary distribution from a replay buffer. An unbiased estimator for each term in the Stackelberg gradient can be computed directly from samples using automatic differentiation, *e.g.*,

$$\begin{aligned} \nabla_\omega L(\omega, \theta, \psi) &= \mathop{\mathbb{E}}_{\xi\sim\mathcal{B}} \left[\nabla_\omega \left(\left(Q_{\omega_1}(x, u, d) - r - \gamma Q_{\omega_2}(x', u', d')\right)^2 \right) \right] \\ &\approx \frac{1}{N} \sum_{n=1}^{N} \nabla_\omega \left(\left(Q_{\omega_1}(x_n, u_n, d_n) - r_n - \gamma Q_{\omega_2}(x'_n, u'_n, d'_n)\right)^2 \right). \end{aligned} \tag{13}$$

For the second term in (11), we can estimate each term $h_1(\cdot)$, $H(\cdot)$, and $h_2(\cdot)$ individually using samples from the replay buffer and resetting the simulator [44, Chapter 11]. In order to numerically determine convergence, we may check the magnitude of the gradient norms, *i.e.*, $\|D_\omega L(\cdot)\| \le \epsilon^c$, $\|\nabla_\theta J(\cdot)\| \le \epsilon^u$, $\|\nabla_\psi J(\cdot)\| \le \epsilon^d$, where ϵ^c, ϵ^u, ϵ^d are small thresholds. The overall procedure of MAGICS is summarized in Algorithm 2.

Algorithm 2: Stackelberg–Minimax Adversarial RL

Input: parameters $(\omega_0, \theta_0, \psi_0)$, critic learning rate schedule $\{\alpha_i^c\}_{i=0,1,\ldots}$, controller learning rate schedule $\{\alpha_i^u\}_{i=0,1,\ldots}$, actor learning rate ratio τ_a

$i \leftarrow 0$
while *Not converged* **do**
  $\omega_{i+1} \leftarrow \omega_i - \alpha_i^c D_\omega L(\omega_i, \theta_i, \psi_i)$ // Updates critic parameters
  $(\theta_{i+1}, \psi_{i+1}) \leftarrow \tau$-GDA$(\omega_i, \theta_i, \psi_i, \alpha_i^u, \tau_a)$ // Updates actor parameters
  $i \leftarrow i + 1$

Complexity and Gradient Computation Details. The critic's Stackelberg gradient requires an inverse-Hessian-vector product (iHvp) and a Jacobian-vector product (jvp). The latter can be computed directly through a call to `torch`'s automatic differentiation engine `autograd.grad`. The Hessian H in (11) is a 2×2 block matrix composed of differentiating the loss functions of the critic and actor(s) against their respective

parameters; accordingly, implicit representations using Hessian-vector products cannot be used. If h_1's differentiation against the player parameters is done first, assembling H requires three additional `autograd.grad` calls—one for each of the block-diagonal elements as well as the $(1,2)$-entry due to symmetry. h_2 can be assembled straightforwardly with two calls to `autograd.grad`, and the iHvp can be computed using `torch`'s built-in linear systems solver `linalg.solve`. The jvp between h_1 and the iHvp can be computed in one final `autograd.grad`for a total of eight calls. For MAGICS-A2C, the sequence of eight `autograd` calls is not prohibitively slow because of the custom gradient estimators—the closed-form estimator with the direct log-likelihood structure ensures that the differentiation does not need to traverse a large computational graph. In contrast, the eight `autograd` calls require massive computational expenditure for MAGICS-SAC. To enhance computational efficiency, in our implementation of MAGICS-SAC, we use an empirical Fisher approximation to the Hessian [7,34] for computing (11). The direct log-likelihood structure and Einstein summation of rank-one matrices can be computed in time of a similar order of magnitude to that of MAGICS-A2C, which ameliorates the computational burden.

For MAGICS-A2C, each Stackelberg gradient step can be computed in ~ 5.5 times the cost of a baseline gradient step. For MAGICS-SAC, the full Stackelberg gradient incurs ~ 10.5 times the computational cost of the baseline gradient, whereas the empirical Fisher implementation significantly reduces this, requiring only about ~ 1.8 times the baseline cost. We find empirically the increase in computation by incorporating the Stackelberg gradient in MAGICS-A2C and MAGICS-SAC (with empirical Fisher approximation) to be bearable even for high-dimensional systems.

Convergence Analysis of MAGICS. Due to sample-based approximation of gradients, the MAGICS procedure can be modeled by the discrete-time dynamical system:

$$\omega_{t+1} = \omega_t - \alpha_t^c \left(D_\omega L(\omega_t, \theta_t, \psi_t) + v_{\omega,t} \right), \tag{14a}$$

$$\theta_{t+1} = \theta_t + \alpha_t^u \left(\nabla_\theta J(\omega_t, \theta_t, \psi_t) + v_{\theta,t} \right), \tag{14b}$$

$$\psi_{t+1} = \psi_t - \tau_a \alpha_t^u \left(\nabla_\psi J(\omega_t, \theta_t, \psi_t) + v_{\psi,t} \right). \tag{14c}$$

In the next, we show that the MAGICS procedure, modeled by system (14), locally converges to a game-theoretically meaningful equilibrium solution. We start by showing that τ-GDA can *robustly* converge to a minimax equilibrium when the critic parameter ω varies within a local region, *i.e.*, $\omega \in U_{\bar{\omega}} := \{\tilde{\omega} \mid \|\tilde{\omega} - \bar{\omega}\| < \epsilon_\omega\}$. This is formalized in the following Lemma, which extends the two-player τ-GDA local convergence result [13, Theorem 1] to additionally account for a "meta" leader (*i.e.*, the critic).

Assumption 1 *We assume that the following hold.*

(a) The maps $D_\omega L(\cdot)$, $\nabla_\theta J(\cdot)$, $\nabla_\psi J(\cdot)$ are L_1, L_2, L_3 Lipschitz, and $\|D_\omega L\| < \infty$.
(b) The critic learning rates are square summable but not summable, i.e., *$\sum_k \alpha_k^i = \infty$, $\sum_k (\alpha_k^i)^2 < \infty$ for $i \in \{c, u\}$.*
(c) The noise processes $\{v_{\omega,t}\}$, $\{v_{\theta,t}\}$, and $\{v_{\psi,t}\}$ are zero-mean, martingale difference sequences (c.f. [50, Assumption 1]).

Lemma 1 (Robust Stability of DSE under τ-GDA). *Consider the zero-sum game between the controller and disturbance $(-J, J)$ parameterized by $\omega \in U_{\bar{\omega}} = \{\tilde{\omega} \mid \|\tilde{\omega} -$*

$\bar{\omega}\| < \epsilon_\omega\}$. *Let Assumption 1 hold. If* $y^* := (\theta^*, \psi^*)$ *is a DSE (minimax equilibrium), then there exists a* $\tau^* \in (0, \infty)$ *such that, for all* $\tau \in (\tau^*, \infty)$ *and for all* $\omega \in U_{\bar{\omega}}$, $\{y_k\}$ *almost surely converges locally asymptotically to* y^*

Proof. Given critic parameter $\bar{\omega}$, the continuous-time limiting system of the noise-free τ-GDA updates (14b)-(14c) is $\dot{y} = (\nabla_\theta J(\bar{\omega}, y), -\tau_a \nabla_\psi J(\bar{\omega}, y))$, with Jacobian denoted as $\mathbf{J}_{\tau_a}(y; \bar{\omega})$. By [13, Theorem 1], it is possible to explicitly construct $\tau_{\bar{\omega}} = \lambda^+_{\max}(Q_{\bar{\omega}}(\mathbf{J}_{\tau_a}(y^*; \bar{\omega})))$, such that $\mathbf{J}_{\tau_a}(y^*; \bar{\omega})$ is Hurwitz, *i.e.*, y^* is locally exponentially stable, for all $\tau \in (\tau_{\bar{\omega}}, \infty)$. Here, matrix $Q_{\bar{\omega}}$ is computed based on blocks of Jacobian $\mathbf{J}_{\tau_a}(y^*; \bar{\omega})$, and is $L_{\bar{\omega}}$ Lipschitz continuous in ω. Next, we extend this result to construct a finite learning rate ratio such that $\mathbf{J}_{\tau_a}(y^*; \omega)$ is Hurwitz for all $\omega \in U_{\bar{\omega}}$. For any $\omega \in U_{\bar{\omega}}$, define $\Delta Q_\omega = Q_\omega - Q_{\bar{\omega}}$. By Weyl's inequality, we have $\lambda^+_{\max}(Q_\omega) \leq \lambda^+_{\max}(Q_{\bar{\omega}}) + \|\Delta Q_\omega\| \leq \lambda^+_{\max}(Q_{\bar{\omega}}) + L_{\bar{\omega}}\|\omega - \bar{\omega}\| < \lambda^+_{\max}(Q_{\bar{\omega}}) + L_{\bar{\omega}}\epsilon_\omega$. Therefore, y^* is locally exponentially stable for all $\tau \in (\tau^*, \infty)$ and all $\omega \in U_{\bar{\omega}}$, where $\tau^* = \lambda^+_{\max}(Q_{\bar{\omega}}) + L_{\bar{\omega}}\epsilon_\omega$. The remainder of the proof follows that of [13, Theorem H.1] and classical results in stochastic approximation theory [4]. That is, there exists a neighborhood U_{y^*} around y^* such that, from an initial point $y_0 \in U_{y^*}$, sequence $\{y_k\}$ converges to an internally chain transitive invariant set contained in U_{y^*} almost surely for all $\omega \in U_{\bar{\omega}}$, and the only such invariant set contained in U_{y^*} is y^*. □

With τ-GDA, we can not only construct a learning rate ratio τ^* to ensure local convergence near a DSE/minimax equilibrium, but also find a τ_0 such that a non-equilibrium critical point is unstable, and thereby can be avoided with arbitrarily small perturbation. This is formalized in the following lemma, of which the proof resembles that of Lemma 1 and [13, Theorem 2].

Lemma 2 (Instability of Spurious Critical Points under τ-GDA). *Consider zero-sum game* $(-J, J)$ *parameterized by* $\omega \in U_{\bar{\omega}}$. *If* $y^* := (\theta^*, \psi^*)$ *is a critical point,* i.e., $\nabla_\theta J(\bar{\omega}, y^*) = 0$, $\nabla_\psi J(\bar{\omega}, y^*) = 0$, $\det \nabla^2_\psi J(\bar{\omega}, y^*) \neq 0$, *but not a DSE, then there exists a* $\tau_0 \in (0, \infty)$ *such that, for all* $\tau \in (\tau_0, \infty)$ *and for all* $\omega \in U_{\bar{\omega}}$, y^* *is an unstable equilibrium of the limiting system* $\dot{y} = (\nabla_\theta J(\omega, y), -\tau_a \nabla_\psi J(\omega, y))$.

Now, we are ready to state our main result, which shows that the MAGICS procedure converges locally to a DSE of game (10).

Theorem 2 (Convergence of MAGICS). *Consider the general-sum game* $(L, -J, J)$ *defined in* (10). *Let Assumption 1 hold. If* $(\omega^*, \theta^*, \psi^*)$ *is a DSE and* $\alpha_i^c = o(\alpha_i^u)$, *there exists a* $\tau_a^* \in (0, \infty)$ *and a neighbourhood* U *around* $(\omega^*, \theta^*, \psi^*)$ *such that, for all* $(\omega_0, \theta_0, \psi_0) \in U$ *and all* $\tau_a \in (\tau_a^*, \infty)$, *iterates* $(\omega_t, \theta_t, \psi_t)$, i.e., *state of system* (14), *converge asymptotically almost surely to* $(\omega^*, \theta^*, \psi^*)$.

Proof. Since $(\omega^*, \theta^*, \psi^*)$ is a DSE, by the implicit function theorem, there exists a neighbourhood $U^1_{\omega^*}$ around ω^* and unique functions r_θ, r_ψ such that $r_\theta(\omega^*) = \theta^*$, $r_\psi(\omega^*) = \psi^*$, $\nabla_\theta J(\omega, r_\theta(\omega), r_\psi(\omega)) = 0$, and $\nabla_\psi J(\omega, r_\theta(\omega), r_\psi(\omega)) = 0$ for all $\omega \in U^1_{\omega^*}$. Moreover, there exists a neighbourhood $U^2_{\omega^*}$ around ω^* on which Hessians $D^2_\omega L(\omega, r_\theta(\omega), r_\psi(\omega)) \succ 0$, $\nabla^2_\theta J(\omega, r_\theta(\omega), r_\psi(\omega)) \prec 0$, $\nabla^2_\psi J(\omega, r_\theta(\omega), r_\psi(\omega)) \succ 0$. Since $\nabla^2_\theta J(\omega^*, \theta^*, \psi^*) \prec 0$ and $\nabla^2_\psi J(\omega^*, \theta^*, \psi^*) \succ 0$, there exists a neighbourhood $U_{\theta^*} \times U_{\psi^*}$ around (θ^*, ψ^*) such that $\nabla^2_\theta J(\omega, \theta, \psi) \prec 0$ and $\nabla^2_\psi J(\omega, \theta, \psi) \succ 0$ for all $(\omega, \theta, \psi) \in U_{\omega^*} \times U_{\theta^*} \times U_{\psi^*}$, where $U_{\omega^*} \subseteq U^1_{\omega^*} \cap U^2_{\omega^*}$ is a non-empty open set.

Next, we show that any $(\omega_0, \theta_0, \psi_0) \in U \subseteq U_{\omega^*} \times U_{\theta^*} \times U_{\psi^*}$ will converge asymptotically almost surely to DSE $(\omega^*, \theta^*, \psi^*)$. The continuous-time limiting system of (14) is $(\dot{\omega}, \dot{\theta}, \dot{\psi}) =$

$(-D_\omega L(\omega,\theta,\psi), \nabla_\theta J(\omega,\theta,\psi), -\tau_a \nabla_\psi J(\omega,\theta,\psi))$. Note that (14a) can be written as $\omega_{t+1} = \omega_t - \alpha_t^u \zeta_t$, where $\zeta_t = \frac{\alpha_t^c}{\alpha_t^u}(D_\omega L(\omega_t,\theta_t,\psi_t) + v_{\omega,t})$. Since $\zeta_t = o(1)$ for all $t = 0,1,\ldots$, the term is asymptotically negligible and discrete-time state $(\omega_t,\theta_t,\psi_t)$ tracks $(\dot\omega = 0, \dot\theta = \nabla_\theta J(\omega,\theta,\psi), \dot\psi = -\tau_a \nabla_\psi J(\omega,\theta,\psi))$. By Lemma 1, this system is locally exponentially stable in a region $\bar U$ around the DSE $(\omega^*,\theta^*,\psi^*)$ for all $\tau_a \in (\tau_a^*, \infty)$. Therefore, there exists a local Lyapunov function on $U = \bar U \cap U_{\omega^*} \times U_{\theta^*} \times U_{\psi^*}$, which shows that the discrete-time state trajectory and continuous-time flow, both starting from any $(\omega_0,\theta_0,\psi_0) \in U$, asymptotically contract onto each other. The remainder of the proof follows [11, Lemma G.2], which shows that $\lim_{k\to\infty} \|(\omega_k,\theta_k,\psi_k) - (\omega^*,\theta^*,\psi^*)\| \to 0$ almost surely for all $(\omega_0,\theta_0,\psi_0) \in U$. □

While Theorem 2 guarantees that MAGICS can converge to a DSE when the network parameters are initialized in a local region around it, we can leverage Lemma 2 to escape a non-DSE critical point.

Proposition 1. *Consider the general-sum game $(L,-J,J)$ defined in (10). If parameters $(\omega^*,\theta^*,\psi^*)$ is a critical point of the game but not a DSE, there exists a $\tau_0 \in (0,\infty)$ such that, for all $\tau_a \in (\tau_0,\infty)$, $(\omega^*,\theta^*,\psi^*)$ is an unstable equilibrium of the limiting system $(\dot\omega,\dot\theta,\dot\psi) = (-D_\omega L(\omega,\theta,\psi), \nabla_\theta J(\omega,\theta,\psi), -\tau_a\nabla_\psi J(\omega,\theta,\psi))$.*

5 Convergent Neural Synthesis of Robot Safety

In this section, we apply MAGICS to high-dimensional robot safety analysis, and propose the MAGICS-Safety algorithm for convergent neural synthesis of safe robot policies. Following prior work on RL-based approximate reachability analysis [15,22], we consider a time-discounted version of Isaacs equation (5):

$$V^{\blacktriangledown}(x) = \min\left\{(1-\gamma)g(x), \gamma \max\left\{\ell(x), \max_{u\in\mathcal{U}} \min_{d\in\mathcal{D}} V^{\blacktriangledown}\big(f(x,u,d)\big)\right\}\right\}. \tag{15}$$

The discount factor $\gamma \in (0,1)$ leads to a probabilistic interpretation, *i.e.*, there is $(1-\gamma)$ probability that the episode terminates immediately due to loss of safety. Note that as $\gamma \to 1$, we recover the undiscounted Isaacs equation (5). Then, we apply SAC to approximately solve (15). The critic minimizes loss

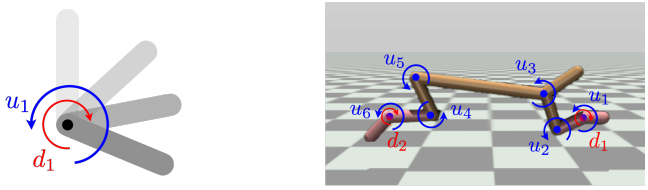
$$L^{\blacktriangledown}(\omega,\theta,\psi) = \mathop{\mathbb{E}}_{\xi\sim\mathcal{B}}\left[\left(Q_{\omega_1}(x,u,d) - (1-\gamma)g' - \gamma \min\left\{g', Q_{\omega_2}(x',u',d')\right\}\right)^2\right], \tag{16}$$

where $g' := g(x')$, $\xi = (x,u,d,g',x')$, $u' \sim \pi_\theta^u(\cdot|x')$, and $d' \sim \pi_\psi^d(\cdot|x')$. The controller maximizes objective

$$J^{\blacktriangledown}(\omega,\theta,\psi) = \mathop{\mathbb{E}}_{x\sim\mathcal{B}}\left[Q_{\omega_1}(x,\tilde u,\tilde d) - \eta^u \log \pi_\theta^u(\tilde u|x) + \eta^d \log \pi_\psi^d(\tilde d|x)\right], \tag{17}$$

where $\tilde u \sim \pi_\theta^u(\cdot|x)$, $\tilde d \sim \pi_\psi^d(\cdot|x)$. The disturbance minimizes $J^{\blacktriangledown}(\omega,\theta,\psi)$.

The MAGICS-Safety training objectives (16)-(17) differ from [22] in that they explicitly capture the coupling among the critic and actors (hence their non-cooperative interactions). This game-theoretic formulation (c.f. (10)) of adversarial RL facilitates developing convergence guarantees using the MAGICS paradigm in Section 4. In the following theorem, we show that MAGICS-Safety locally converges to a DSE, which is a direct consequence of MAGICS convergence in Theorem 2.



(a) Inverted pendulum. **(b)** Half Cheetah in Mujoco.

Fig. 2: The Pendulum and Half Cheetah environments in OpenAI Gym [5], with control and disturbance actions represented in blue and red arrows, respectively. The Pendulum's control inputs are one-dimensional torques applied to the end of the rod in opposition to each other. The Half Cheetah has a six-dimensional control input on the notated joints, with the disturbance acting to destabilize the cheetah through additional torques on its paws.

Theorem 3 (Convergence of MAGICS-Safety). *Consider the general-sum game defined in* (10) *with game objectives* $(L^{\textcolor{red}{\blacklozenge}}, -J^{\textcolor{red}{\blacklozenge}}, J^{\textcolor{red}{\blacklozenge}})$ *defined in* (16)-(17)*. Let Assumption 1 hold. If* $(\omega^*, \theta^*, \psi^*)$ *is a DSE and* $\alpha_i^c = o(\alpha_i^d)$*, there exists a* $\tau_a^* \in (0, \infty)$ *and a neighbourhood* $U^{\textcolor{red}{\blacklozenge}} = U_{\omega^*}^{\textcolor{red}{\blacklozenge}} \times U_{\theta^*}^{\textcolor{red}{\blacklozenge}} \times U_{\psi^*}^{\textcolor{red}{\blacklozenge}}$ *around* $(\omega^*, \theta^*, \psi^*)$ *such that, for all* $(\omega_0, \theta_0, \psi_0) \in U^{\textcolor{red}{\blacklozenge}}$ *and all* $\tau_a \in (\tau_a^*, \infty)$*, iterates* $(\omega_t, \theta_t, \psi_t)$ *of the MAGICS-Safety training procedure converge asymptotically almost surely to* $(\omega^*, \theta^*, \psi^*)$*.*

6 Experiments

In this section, we illustrate the strength of our approach in two simulated and one hardware examples that differ in task, problem scale, and computation approach. *Our main hypothesis is that the Stackelberg-minimax robust RL is likelier to yield a stronger policy (for both controller and disturbance) than non-game robust RL baselines.*

6.1 Simulated Examples: Robust Control in OpenAI Gym

Our simulation experiments build upon the OpenAI [5] Gym with the Mujoco physics simulator [46] and Stable-Baselines3 [40] platforms. Specifically, we adapt the Pendulum and Half Cheetah environments (Figure 2), originally created for single-agent RL, to account for two inputs–controller and disturbance.

We compare MAGICS-A2C and MAGICS-SAC against an ablation study and a baseline method—the ablation removes the critic's Stackelberg gradient but preserves τ-GDA between the agents, and the baseline enforces all learning rates to be identical. Table 1 and Table 2 display results from a series of round-robin matches between the learned control and disturbance policies for all three methods. For each match, the selected control and disturbance play five sets of 100 games. We report the win rate of each controller for each controller/disturbance strategy pair. We assign a controller failure should any of the following occur: (1) the controller swings the pendulum into the upright position, but the disturbance is able to destabilize it (the pendulum tip moves outside $\pm 10^\circ$ from the vertical), (2) the controller swings the pendulum up, but cannot swing it soon enough for more than 5 seconds to remain, or (3) the controller completely fails to bring the pendulum to the upright position. If none of these three failure modes occur, we consider the controller to have won the game.

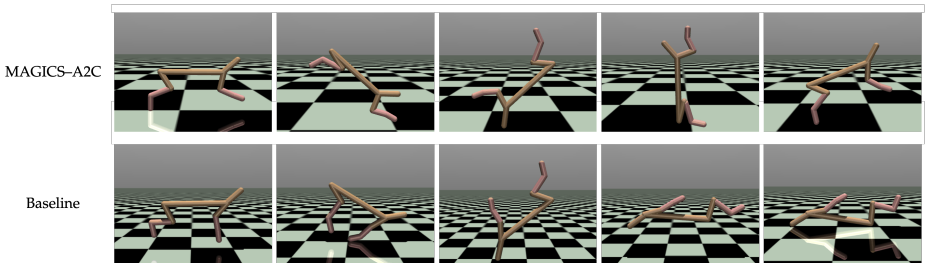


Fig. 3: Snapshots of the Half Cheetah controlled by MAGICS-A2C and baseline-A2C. Despite an excessively large disturbance torque, the MAGICS-A2C policy manage to flip the robot back upright and resumed normal gaits. In contrast, the baseline-A2C policy is unable to recover the robot from the overturn; it moved awkwardly on its face and back, wiggling its feet.

In these tables, for each disturbance strategy, we highlight, in bold-type, the most performant control strategy. All hyperparameters and model architecture (*e.g.*, number of hidden layers, activation functions) among the models are identical with the exception of the learning rates as per τ-GDA.

For the pendulum, our results demonstrate that the MAGICS controller outperforms the ablation and baseline controllers against all three disturbances. In addition, since the two actors evolve together, there is strong evidence to believe that, for the snapshot of performance depicted in the tables, each disturbance relative to its associated control, is also performant—the diagonal entries depict a win rate that is close to even. However, given that MAGICS's disturbance can resoundingly defeat the other individually strong control strategies, there is strong evidence that the critic's Stackelberg gradient encourages the disturbance to learn alternative strategies to attack the control.

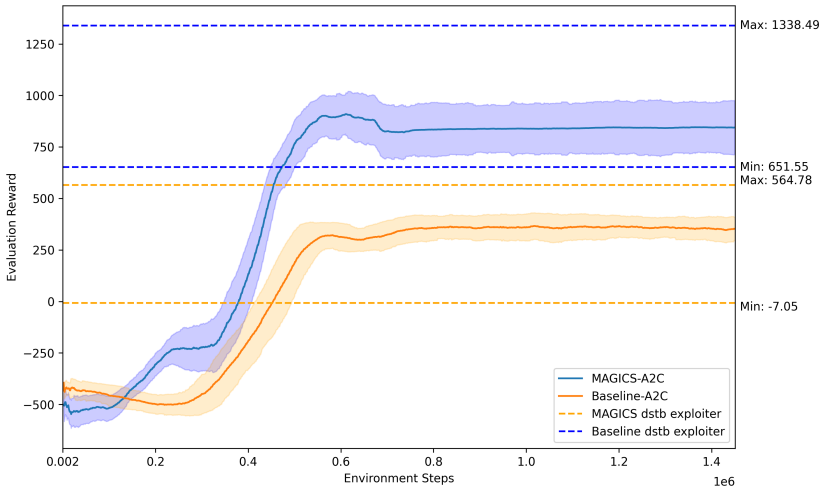


Fig. 4: Cumulative reward curves across five seeds of MAGICS-A2C (blue) and baseline-A2C (orange) for the adversarial Half Cheetah environment. MAGICS-A2C converges to an equilibrium that outperforms the converged baseline equilibrium by ~ 2.7 times. Dashed lines represent exploiter disturbances against the same controller color.

We also perform an analogous tournament in the Half Cheetah example. However, for this environment, success or failure can no longer be straightforwardly defined with a single logical variable. For the Half Cheetah, the objective is to run to the right as quickly and for as long as possible (until the episode is terminated at a maximum step count). To this end, we report the raw cumulative reward in Table 3. MAGICS-A2C achieves significantly higher reward compared to both ablation and baselines. Furthermore, the MAGICS-A2C disturbance is quantitatively stronger than that of the baseline, as the controller reward increases when MAGICS-A2C's controller is pitted against the baseline disturbance. Figure 3 showcases the robustness of the MAGICS controller with a representative example. The baseline disturbance learns that a large torque to flip the cheetah prevents the controller from getting a large reward. Concordantly, the baseline controller fails to learn how to recover the cheetah from turnover. In contrast, MAGICS's controller has learned to leverage the momentum—when the disturbance tips the cheetah too far forward, the controller adapts the fall into a roll—to get the cheetah back to its feet.

ctrl. strategy \ dstb. strategy	MAGICS-A2C	MAGICS-A2C-ablation	Baseline-A2C
MAGICS-A2C ↑	**78.8**%	**67.4**%	**74.6**%
MAGICS-A2C-ablation ↑	36.4%	40.6%	38.2%
Baseline-A2C ↑	59.6%	58.0%	53.0%

Table 1: Win rates across five sets of 100 zero-sum games between the corresponding controller and disturbance from a random initial state in the Pendulum environment. Stackelberg gradient boosts the self-play win rate by approximately 38%, as evidenced by the diagonal entries.

ctrl. strategy \ dstb. strategy	MAGICS-SAC	MAGICS-SAC-ablation	Baseline-SAC
MAGICS-SAC ↑	**66.5**%	**63.25**%	80.75%
MAGICS-SAC-ablation ↑	51.5%	57.5%	**83.0**%
Baseline-SAC ↑	61.25%	60.25%	80.25%

Table 2: Win rates obtained from 100 zero-sum games between the corresponding five controller and disturbance random seeds for the Pendulum. MAGICS-SAC outperforms almost all the other disturbance classes.

ctrl. strategy \ dstb. strategy	MAGICS-A2C	MAGICS-A2C-ablation	Baseline-A2C
MAGICS-A2C ↑	$\mathbf{972.14 \pm 289.19}$	$\mathbf{967.78 \pm 376.77}$	$\mathbf{1121.81 \pm 137.13}$
MAGICS-A2C-ablation ↑	316.46 ± 281.73	285.41 ± 275.00	325.56 ± 277.35
Baseline-A2C ↑	360.85 ± 154.86	335.57 ± 115.96	409.04 ± 118.44

Table 3: Episodic reward obtained from five sets of 100 zero-sum games between the corresponding controller and disturbance from a random initial state for the Half Cheetah. MAGICS-A2C achieves *significantly* higher reward against all other disturbance classes, with qualitatively more innovative strategies found relative to that of the baseline (Figure 3).

We also display reward curves for the Half Cheetah environment in Figure 4. From the same initial conditions, qualitatively, MAGICS converges to more performant equilibria than the baseline does, with a controller reward approximately ~ 2.7 times greater than that of the baseline controller and disturbance pair. We also plot, in dashed lines,

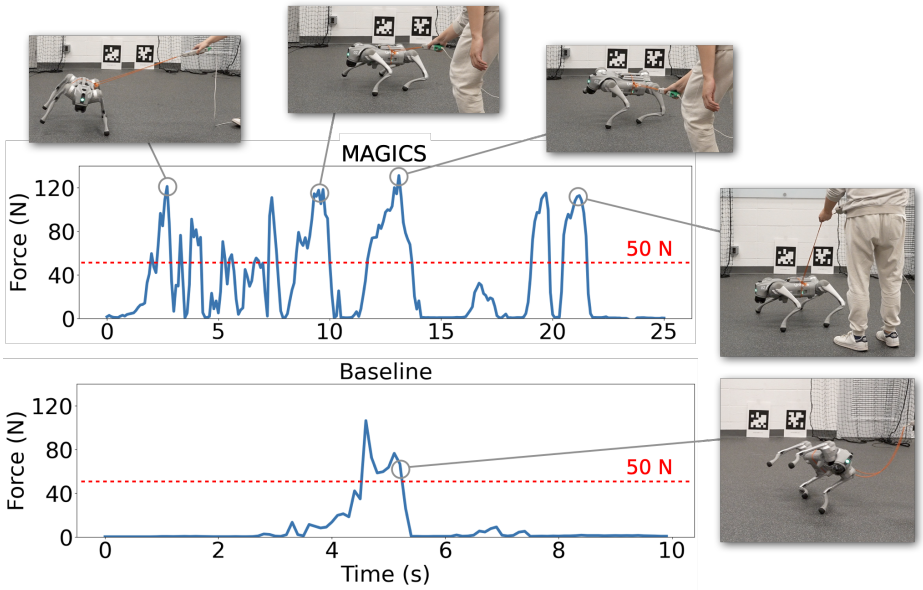


Fig. 5: Time evolution of the human's tugging forces (disturbance) with MAGICS-Safety and the baseline. Both policies are trained in simulation with a maximum of 50 N tugging force disturbance. The MAGICS-Safety policy is robust against the varying tugging forces from different angles, while the baseline failed even with tugging forces of smaller magnitude.

the final exploit curve results: the "best" and "worst" disturbances against the "best" controllers of that color, *e.g.*, dashed-blue lines represent the strongest MAGICS-A2C controller against the strongest and weakest of the baseline-A2C disturbances. Even against the strongest baseline-trained disturbance, MAGICS-A2C's performance only degrades slightly, as opposed to the baseline controller against the MAGICS disturbance, in which all strategies fail as evidenced by the extremely poor reward.

6.2 Hardware Demonstration: Safe Quadrupedal Locomotion

Next, we apply MAGICS-Safety (Section 5) to a safe quadrupedal locomotion task.

Hardware. We utilize the Unitree Go-2 as the test robot platform. The robot is equipped with built-in IMUs for obtaining measurements about body angular velocities and linear acceleration, and internal motor encoders that measure joint positions and velocities. The robot also provides a Boolean contact signal for each foot. No visual perception is used for computing the control policy.

System Dynamics. The quadrupedal robot's state is 36-dimensional, including positions of the body frame, velocities of the robot's torso, rotation angles, axial rotational rates, angle, angular velocity, and commanded angular increment of the robot's joints. The robot's control inputs include independent torques applied on each of its 12 rotational joints provided by an electric motor. Our neural control policy sends a reference signal to each motor, which is tracked by a low-level controller. See [38] for a detailed explanation of the robot's dynamic model and safety specification.

Baseline. We compare to ISAACS [22,38], the former state-of-the-art adversarial neural safety synthesis method, which uses individual gradients and alternating optimization for updating the critic and actor parameters. Both policies are trained with the PyBullet [9] physics engine. We use the same neural network architecture as [38].

Policy Training. We train both policies for 1.5 million steps. During training, we use a maximum of 50 N tugging force disturbance for both policies. Once the policies are trained offline, they are deployed online within a *value-based* safety filter [21, Sec. 3.2] to prevent the robot from falling under adversarial human tugging forces.

Evaluation: The Tugging Experiment. We evaluate the robustness of the trained policies with a human tugging the robot from different angles and using forces with a varying magnitude (Figure 1). The time evolution of the forces applied by the human is plotted in Figure 5. The MAGICS-Safety policy is able to withstand the external force, with a peak value above 120 N, for the entire test horizon of 25 seconds. On the other hand, the baseline fails to resist a force with a peak value less than 120 N. This test empirically demonstrates the superior robustness of the MAGICS-Safety policy compared to the baseline.

7 Limitations and Future Work

Our proposed MAGICS algorithm requires the computation of second-order information over the space of neural network parameters (c.f. (11)), which can be expensive to obtain. Using appropriate estimators in lieu of intensive gradient computations uncovers several open opportunities. In another vein, recent advances in Stackelberg learning using only first-order information [36] offer a promising pathway to ease the computation burden of MAGICS. In this paper, we focus exclusively on developing a game-theoretic variant of the adversarial on- and off-policy actor–critic algorithm. Our algorithmic and theoretical framework may be extended to broader multi-agent RL settings with different algorithms, such as policy gradient, and with general-sum objectives among players.

8 Conclusions

In this paper, we introduced Minimax Actors Guided by Implicit Critic Stackelberg (MAGICS), a novel game-theoretic reinforcement learning algorithm that is provably convergent to an equilibrium solution. Building on MAGICS, we also offered convergence assurances for an RL-based robot safety synthesis method. Our empirical evaluations, conducted through simulations in OpenAI Gym and hardware tests using a 36-dimensional quadruped robot, demonstrated that MAGICS produced robust control policies consistently outperforming the state-of-the-art neural safe control method.

Acknowledgements

The authors thank Chi Jin, Wenzhe Li, Zixu Zhang, Kai-Chieh Hsu, and Kaiqu Liang for helpful discussions throughout the project. Compute time was provided by Princeton Research Computing HPC facilities. The work has been partly supported by the DARPA LINC program and an NSF CAREER Award.

References

1. Agrawal, A., Sreenath, K.: Discrete control barrier functions for safety-critical control of discrete systems with application to bipedal robot navigation. In: Robotics: Science and Systems. vol. 13, pp. 1–10. Cambridge, MA, USA (2017)
2. Bansal, S., Chen, M., Herbert, S., Tomlin, C.J.: Hamilton-jacobi reachability: A brief overview and recent advances. In: 2017 IEEE 56th Annual Conference on Decision and Control (CDC). pp. 2242–2253. IEEE (2017)
3. Bansal, S., Tomlin, C.J.: Deepreach: A deep learning approach to high-dimensional reachability. In: 2021 IEEE International Conference on Robotics and Automation (ICRA). pp. 1817–1824. IEEE (2021)
4. Borkar, V.S.: Stochastic approximation: a dynamical systems viewpoint, vol. 48. Springer (2009)
5. Brockman, G., Cheung, V., Pettersson, L., Schneider, J., Schulman, J., Tang, J., Zaremba, W.: Openai gym. arXiv preprint arXiv:1606.01540 (2016)
6. Bui, M., Lu, M., Hojabr, R., Chen, M., Shriraman, A.: Real-time hamilton-jacobi reachability analysis of autonomous system with an fpga. In: 2021 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS). pp. 1666–1673. IEEE (2021)
7. Chaudhari, P., Choromanska, A., Soatto, S., LeCun, Y., Baldassi, C., Borgs, C., Chayes, J., Sagun, L., Zecchina, R.: Entropy-sgd: Biasing gradient descent into wide valleys (2017), https://arxiv.org/abs/1611.01838
8. Chen, M., Herbert, S.L., Hu, H., Pu, Y., Fisac, J.F., Bansal, S., Han, S., Tomlin, C.J.: Fastrack: a modular framework for real-time motion planning and guaranteed safe tracking. IEEE Transactions on Automatic Control **66**(12), 5861–5876 (2021)
9. Coumans, E., Bai, Y.: Pybullet, a python module for physics simulation for games, robotics and machine learning (2016)
10. Dawson, C., Qin, Z., Gao, S., Fan, C.: Safe nonlinear control using robust neural lyapunov-barrier functions. In: Conference on Robot Learning. pp. 1724–1735. PMLR (2022)
11. Fiez, T., Chasnov, B., Ratliff, L.: Implicit learning dynamics in stackelberg games: Equilibria characterization, convergence analysis, and empirical study. In: International Conference on Machine Learning. pp. 3133–3144. PMLR (2020)
12. Fiez, T., Ratliff, L., Mazumdar, E., Faulkner, E., Narang, A.: Global convergence to local minmax equilibrium in classes of nonconvex zero-sum games. Advances in Neural Information Processing Systems **34**, 29049–29063 (2021)
13. Fiez, T., Ratliff, L.J.: Local convergence analysis of gradient descent ascent with finite timescale separation. In: Proceedings of the International Conference on Learning Representation (2021)
14. Fisac, J.F., Akametalu, A.K., Zeilinger, M.N., Kaynama, S., Gillula, J., Tomlin, C.J.: A general safety framework for learning-based control in uncertain robotic systems. IEEE Transactions on Automatic Control **64**(7), 2737–2752 (2018)
15. Fisac, J.F., Lugovoy, N.F., Rubies-Royo, V., Ghosh, S., Tomlin, C.J.: Bridging hamilton-jacobi safety analysis and reinforcement learning. In: 2019 International Conference on Robotics and Automation (ICRA). pp. 8550–8556. IEEE (2019)
16. Foerster, J., Chen, R., Al-Shedivat, M., Whiteson, S., Abbeel, P., Mordatch, I.: Learning with opponent-learning awareness. Autonomous Agents and Multi-Agent Systems (AAMAS 2018) (2018)
17. Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., Bengio, Y.: Generative adversarial networks. Communications of the ACM **63**(11), 139–144 (2020)

18. Haarnoja, T., Zhou, A., Abbeel, P., Levine, S.: Soft actor-critic: Off-policy maximum entropy deep reinforcement learning with a stochastic actor. In: International conference on machine learning. pp. 1861–1870. PMLR (2018)
19. Haarnoja, T., Zhou, A., Hartikainen, K., Tucker, G., Ha, S., Tan, J., Kumar, V., Zhu, H., Gupta, A., Abbeel, P., Levine, S.: Soft actor-critic algorithms and applications (2019), https://arxiv.org/abs/1812.05905
20. He, T., Zhang, C., Xiao, W., He, G., Liu, C., Shi, G.: Agile but safe: Learning collision-free high-speed legged locomotion. In: Proceedings of Robotics: Science and Systems (RSS) (2024)
21. Hsu, K.C., Hu, H., Fisac, J.F.: Safety filter: A unified view of safety-critical control in autonomous systems. Annual Review of Control, Robotics, and Autonomous Systems (2023, to appear)
22. Hsu, K.C., Nguyen, D.P., Fisac, J.F.: Isaacs: Iterative soft adversarial actor-critic for safety. In: Proceedings of the 5th Conference on Learning for Dynamics and Control (2023)
23. Hsu, S.C., Xu, X., Ames, A.D.: Control barrier function based quadratic programs with application to bipedal robotic walking. In: 2015 American Control Conference (ACC). pp. 4542–4548. IEEE (2015)
24. Hu, H., Isele, D., Bae, S., Fisac, J.F.: Active uncertainty reduction for safe and efficient interaction planning: A shielding-aware dual control approach. The International Journal of Robotics Research **43**(9), 1382–1408 (2024)
25. Hu, H., Nakamura, K., Fisac, J.F.: SHARP: Shielding-aware robust planning for safe and efficient human-robot interaction. IEEE Robotics and Automation Letters (2022)
26. Hu, H., Zhang, Z., Nakamura, K., Bajcsy, A., Fisac, J.F.: Deception game: Closing the safety-learning loop in interactive robot autonomy. In: 7th Annual Conference on Robot Learning (2023)
27. Huang, P., Xu, M., Fang, F., Zhao, D.: Robust reinforcement learning as a stackelberg game via adaptively-regularized adversarial training. In: the 31st International Joint Conference on Artificial Intelligence (IJCAI) (2022)
28. Isaacs, R.: Differential games i: Introduction. Tech. rep., RAND CORP SANTA MONICA CA SANTA MONICA (1954)
29. Jin, C., Netrapalli, P., Jordan, M.: What is local optimality in nonconvex-nonconcave minimax optimization? In: International conference on machine learning. pp. 4880–4889. PMLR (2020)
30. Konda, V., Tsitsiklis, J.: Actor-critic algorithms. Advances in neural information processing systems **12** (1999)
31. Kumar, A., Fu, Z., Pathak, D., Malik, J.: Rma: Rapid motor adaptation for legged robots. Robotics: Science and Systems XVII (2021)
32. Lai, H., Zhang, W., He, X., Yu, C., Tian, Z., Yu, Y., Wang, J.: Sim-to-real transfer for quadrupedal locomotion via terrain transformer. In: 2023 IEEE International Conference on Robotics and Automation (ICRA). pp. 5141–5147. IEEE (2023)
33. Lei, S., Lee, K., Li, L., Park, J., Li, J.: Ela: Exploited level augmentation for offline learning in zero-sum games. arXiv preprint arXiv:2402.18617 (2024)
34. Liao, Z., Drummond, T., Reid, I., Carneiro, G.: Approximate fisher information matrix to characterise the training of deep neural networks (2018), https://arxiv.org/abs/1810.06767
35. Maclaurin, D., Duvenaud, D., Adams, R.: Gradient-based hyperparameter optimization through reversible learning. In: International conference on machine learning. pp. 2113–2122. PMLR (2015)
36. Maheshwari, C., Sasty, S.S., Ratliff, L., Mazumdar, E.: Convergent first-order methods for bi-level optimization and stackelberg games. arXiv preprint arXiv:2302.01421 (2023)

37. Mnih, V., Badia, A.P., Mirza, M., Graves, A., Lillicrap, T.P., Harley, T., Silver, D., Kavukcuoglu, K.: Asynchronous methods for deep reinforcement learning (2016), https://arxiv.org/abs/1602.01783
38. Nguyen, D.P., Hsu, K.C., Yu, W., Tan, J., Fisac, J.F.: Gameplay filters: Safe robot walking through adversarial imagination. arXiv preprint arXiv:2405.00846 (2024)
39. Pinto, L., Davidson, J., Sukthankar, R., Gupta, A.: Robust adversarial reinforcement learning. In: International Conference on Machine Learning. pp. 2817–2826. PMLR (2017)
40. Raffin, A., Hill, A., Gleave, A., Kanervisto, A., Ernestus, M., Dormann, N.: Stable-baselines3: Reliable reinforcement learning implementations. Journal of Machine Learning Research **22**(268), 1–8 (2021), http://jmlr.org/papers/v22/20-1364.html
41. Robey, A., Hu, H., Lindemann, L., Zhang, H., Dimarogonas, D.V., Tu, S., Matni, N.: Learning control barrier functions from expert demonstrations. In: 2020 59th IEEE Conference on Decision and Control (CDC). pp. 3717–3724. IEEE (2020)
42. Schulman, J., Moritz, P., Levine, S., Jordan, M., Abbeel, P.: High-dimensional continuous control using generalized advantage estimation (2018), https://arxiv.org/abs/1506.02438
43. Singletary, A., Swann, A., Chen, Y., Ames, A.D.: Onboard safety guarantees for racing drones: High-speed geofencing with control barrier functions. IEEE Robotics and Automation Letters **7**(2), 2897–2904 (2022)
44. Sutton, R.S., Barto, A.G.: Reinforcement learning: An introduction. MIT press (2018)
45. Tearle, B., Wabersich, K.P., Carron, A., Zeilinger, M.N.: A predictive safety filter for learning-based racing control. IEEE Robotics and Automation Letters **6**(4), 7635–7642 (2021)
46. Todorov, E., Erez, T., Tassa, Y.: Mujoco: A physics engine for model-based control. In: 2012 IEEE/RSJ international conference on intelligent robots and systems. pp. 5026–5033. IEEE (2012)
47. Wabersich, K.P., Taylor, A.J., Choi, J.J., Sreenath, K., Tomlin, C.J., Ames, A.D., Zeilinger, M.N.: Data-driven safety filters: Hamilton-jacobi reachability, control barrier functions, and predictive methods for uncertain systems. IEEE Control Systems Magazine **43**(5), 137–177 (2023)
48. Zeng, J., Zhang, B., Sreenath, K.: Safety-critical model predictive control with discrete-time control barrier function. In: 2021 American Control Conference (ACC). pp. 3882–3889. IEEE (2021)
49. Zhang, S., So, O., Garg, K., Fan, C.: Gcbf+: A neural graph control barrier function framework for distributed safe multi-agent control. arXiv preprint arXiv:2401.14554 (2024)
50. Zheng, L., Fiez, T., Alumbaugh, Z., Chasnov, B., Ratliff, L.J.: Stackelberg actor-critic: Game-theoretic reinforcement learning algorithms. In: Proceedings of the AAAI conference on artificial intelligence. vol. 36, pp. 9217–9224 (2022)
51. Zrnic, T., Mazumdar, E., Sastry, S., Jordan, M.: Who leads and who follows in strategic classification? In: Advances in Neural Information Processing Systems. vol. 34, pp. 15257–15269 (2021), https://proceedings.neurips.cc/paper/2021/file/812214fb8e7066bfa6e32c626c2c688b-Paper.pdf