Adversarially Robust Stability Certificates Can Be Sample-Efficient

Thomas T.C.K. Zhang TTZ2@SEAS.UPENN.EDU
Stephen Tu STEPHENTU@GOOGLE.COM
Nicholas M. Boffi BOFFI@CIMS.NYU.EDU
Jean-Jacques E. Slotine JJS@MIT.EDU
Nikolai Matni NMATNI@SEAS.UPENN.EDU

Editors: R. Firoozi, N. Mehr, E. Yel, R. Antonova, J. Bohg, M. Schwager, M. Kochenderfer

Abstract

Motivated by bridging the simulation to reality gap in the context of safety-critical systems, we consider learning adversarially robust stability certificates for unknown nonlinear dynamical systems. In line with approaches from robust control, we consider additive and Lipschitz bounded adversaries that perturb the system dynamics. We show that under suitable assumptions of incremental stability on the underlying system, the statistical cost of learning an adversarial stability certificate is equivalent, up to constant factors, to that of learning a nominal stability certificate. Our results hinge on novel bounds for the Rademacher complexity of the resulting adversarial loss class, which may be of independent interest. To the best of our knowledge, this is the first characterization of sample-complexity bounds when performing adversarial learning over data generated by a dynamical system. We further provide a practical algorithm for approximating the adversarial training algorithm, and validate our findings on a damped pendulum example.

Keywords: Stability certificates, adversarial robustness, incremental stability, sim-to-real.

1. Introduction

A challenge to the deployment of modern robotic systems to real-world settings is the overall lack of formal safety guarantees. While controller design for complex robotic systems has received much attention, comparatively less effort has been devoted to verifying the safety of the resulting closed-loop system. Without broadly applicable tools for certifying *a-priori* guarantees, it is difficult to justify deploying these methods in applications where safety is paramount, regardless of the impressive performance that they achieve in simulation or controlled laboratory settings.

An important component of ensuring real-world safety is verifying the stability of a closed-loop system from trajectory data. While recent work (Boffi et al., 2020a) proposes and analyzes a learning-based approach to this problem, a fundamental limitation of the prior art is that learning a stability certificate with failure probability of less than e.g., 1% for high-dimensional systems requires on the order of tens of thousands of trajectories. Realistically, such a large amount of trajectory data can only be collected using a simulation environment. Therefore, in order for a learned certificate to be meaningful for real-world hardware, it is essential for it to be *robust* to modeling errors between simulation and reality, i.e., robust to the so-called sim-to-real gap.

While bridging the sim-to-real gap has traditionally been addressed via domain randomization (Tobin et al., 2017), we take inspiration from the robust control literature, and tackle this challenge by developing an approach for *adversarial learning* of stability certificates for dynamical systems.

We show that under suitable conditions on the underlying system, requiring that a learned certificate is robust to adversarial perturbations that *enter the dynamics* carries little additional statistical overhead. Taking inspiration from Boffi et al. (2020a), we prove our results by converting the *robust* stability certification problem into an adversarial learning problem, and subsequently bounding the Rademacher complexity of the resulting adversarial loss class. To the best of our knowledge, this is the first characterization of sample-complexity bounds when performing adversarial learning over data generated by a dynamical system. Our results build upon and extend a line of work which shows that underlying system-theoretic properties translate into the difficulty (or ease) of learning over data generated by dynamical systems (see e.g., Tsiamis et al. (2020); Tsiamis and Pappas (2021); Lee et al. (2021); Tu et al. (2021) and references therein). We further provide a practical algorithm for approximating the adversarial training algorithm, and show that adversarially trained certificates are robust to various types of model mis-specification on a damped pendulum example. All proofs and more details can be found in the full version of the paper (Zhang et al., 2021) [Link].

1.1. Related Work

Our work draws upon and unifies tools from three areas: (i) learning safety certificates from data, (ii) adversarial robustness, and (iii) statistical learning theory.

Learning safety certificates A wide body of work addresses learning Lyapunov (Giesl et al., 2020; Kenanian et al., 2019; Chen et al., 2020; Richards et al., 2018; Manek and Kolter, 2019; Chang et al., 2019; Ravanbakhsh and Sankaranarayanan, 2019) and barrier (Taylor et al., 2019; Robey et al., 2020; Jin et al., 2020) functions, as well as contraction metrics (Singh et al., 2020; Manchester and Slotine, 2017; Singh et al., 2017) and contracting vector fields (Sindhwani et al., 2018; Khadir et al., 2019) from data. While the generality and strength of guarantees provided vary (see the literature review of Boffi et al. (2020a) for a detailed exposition), all of the aforementioned works consider nominally specified systems without uncertainty, whereas our approach explicitly considers perturbations that can capture model uncertainty and process noise.

Adversarial robustness Traditional approaches (Szegedy et al., 2013; Madry et al., 2018; Zhang et al., 2019; Kurakin et al., 2016) to adversarial learning consider worst-case perturbations to the data during training, i.e., the data is perturbed *after it has been generated*. While such a perturbation model is meaningful in the image classification setting for which adversarial robust training methods were originally developed, it does not immediately translate to the dynamic setting that we consider, where the adversary may be used to capture model uncertainty or process noise. In particular, our adversarial model perturbs the dynamical system which generates the data, a perspective that is more in line with traditional robust control methods. We further show that under suitable stability assumptions on the underlying dynamical system, there is no additional statistical cost to adversarial training, in contrast to results showing that in the traditional setting, adversarial learning algorithms require more data than their nominal counterparts (Schmidt et al., 2018).

Most directly relevant to our work are adversarial deep reinforcement learning methods which learn policies that are robust to various classes of disturbances, such as adversarial observations (Torabi et al., 2019; Gleave et al., 2020), rewards (Fu et al., 2018; Ho and Ermon, 2016), direct disturbances to the system (Pinto et al., 2017), or combinations thereof (Lutter et al., 2021). Nevertheless, there remains a paucity of theoretical guarantees on the generalization error, and thus sample-efficiency, of such learned policies.

Statistical learning theory While such statistical guarantees, to the best of our knowledge, do not exist for adversarial reinforcement learning, the generalization error of an adversarially trained classifier has been studied using uniform convergence (Yin et al., 2019; Attias et al., 2019; Montasser et al., 2019). While our results also rely on uniform convergence, our analysis departs from this existing line of work by allowing adversaries to influence dynamical systems.

2. Problem Framework

2.1. Nominal Stability Certificates

We begin by reviewing the problem setting and results from Boffi et al. (2020a). We assume that the underlying dynamical system is a continuous-time, autonomous system of the form $\dot{x} = f(x)$, where f is continuous and unknown, and that the state $x \in \mathbb{R}^p$ is fully observed. Let $\mathcal{X} \subset \mathbb{R}^p$ be a compact set and $\mathcal{T} \subseteq \mathbb{R}^+$ be the maximum interval such that a unique solution $\varphi_t(\xi)$ exists for all times $t \in \mathcal{T}$ and initial conditions $\xi \in \mathcal{X}$, where $\varphi_t(\xi)$ is the map to the state at time t given initial condition ξ. We assume that we have access to n trajectories initialized from randomly sampled initial conditions. That is, we are given $\{\varphi_t(\xi_i)\}_{i\in[n],\, t\in\mathcal{T}}$, where $\xi_1, \dots, \xi_n \overset{\text{i.i.d.}}{\sim} \mathcal{D}$ and $\mathcal{D}$ is a distribution over $\mathcal{X}$. For simplicity, we assume that we can precisely differentiate $\varphi_t(\xi)$ with respect to time (in practice, we can estimate $\dot{\varphi}_t(\xi)$ numerically).

Let $\mathcal{V}$ be a class of continuously differentiable candidate Lyapunov functions $V : \mathbb{R}^p \to \mathbb{R}_{\geqslant 0}$ satisfying $V(0) = 0$. Fixing a constant $\eta > 0$, we define a scalar violation function $h : \mathcal{X} \times \mathcal{V}$ as:

$$h(\xi, V) := \sup_{t\in\mathcal{T}} \langle \nabla V(\varphi_t(\xi)), f(\varphi_t(\xi))\rangle + \eta V(\varphi_t(\xi)). \tag{1}$$

The violation function $h(\xi, V)$ scans the Lyapunov decrease condition for exponential stability with rate η over the trajectory initialized at ξ, and returns the maximal value. Observe that if $h(\xi, V) \leqslant 0$, then V certifies exponential stability along the trajectory $\varphi_t(\xi)$, $t \in \mathcal{T}$. The nominal stability certification problem is therefore equivalent to the following feasibility problem:

$$\text{Find}_{V\in\mathcal{V}} \text{ s.t. } h(\xi, V) \leqslant 0 \quad \forall \xi \in \mathcal{X}. \tag{2}$$

In general, various choices of $\mathcal{V}$ and $h(\xi, V)$ can encode different notions of stability and accompanying certificates (see Boffi et al. (2020a) for more details). To search for a V that satisfies the above optimization problem given finite data, we solve the following feasibility problem:

$$\text{Find}_{V\in\mathcal{V}} \text{ s.t. } h(\xi_i, V) \leqslant -\tau, \quad i = 1, \dots, n, \tag{3}$$

where $\tau > 0$ is a margin that ensures generalization of the learned stability certificate V on unseen trajectories. Let $\hat{V}_n$ denote a solution to (3) and define the nominal generalization error of $\hat{V}_n$ as

$$\text{err}(\hat{V}_n) := \mathbb{P}_{\xi\sim\mathcal{D}}\left[h\left(\xi, \hat{V}_n\right) > 0\right]. \tag{4}$$

The nominal error (4) characterizes the probability that $\hat{V}_n$ fails to certify stability along a new trajectory with initial condition sampled from $\mathcal{D}$. In Boffi et al. (2020a), it is shown that for general classes of $\mathcal{V}$, $\text{err}(\hat{V}_n)$ decays at a rate $\tilde{O}(k/n)$, where k captures the effective degrees of freedom of the stability function class $\mathcal{V}$ and $\tilde{O}$ suppresses polylog dependence on n and fixed problem parameters.

2.2. Adversarially Robust Stability Certificates

We now consider the stability certification problem under the presence of adversarial perturbations. Consider the following two tubes of perturbed trajectories[1]:

$$\Delta_\varepsilon^u(\xi) := \left\{\tilde{\varphi} : \dot{\tilde{\varphi}}_t = f(\tilde{\varphi}_t) + \delta_t,\ \tilde{\varphi}_0 = \xi,\ \|\delta_t\|_2 \leqslant \varepsilon,\ t \mapsto \delta_t \text{ is locally integrable}\right\}, \tag{5}$$

$$\Delta_\varepsilon^x(\xi) := \left\{\tilde{\varphi} : \dot{\tilde{\varphi}}_t = f(\tilde{\varphi}_t) + \delta(\tilde{\varphi}_t),\ \tilde{\varphi}_0 = \xi,\ \|\delta(\tilde{\varphi}_t)\|_2 \leqslant \varepsilon \|\tilde{\varphi}_t\|_2\right\}. \tag{6}$$

Intuitively, $\Delta_\varepsilon^u(\xi)$ is the tube of perturbed trajectories initialized at ξ for which an additive adversary has an instantaneous norm budget of ε to perturb the dynamics. Analogously, $\Delta_\varepsilon^x(\xi)$ is the tube of perturbed trajectories initialized at ξ for which the adversary satisfies ε-linear growth. We refer adversaries of the form (5) as *norm-bounded*, and adversaries of the form (6) as *Lipschitz* (noting the slight misnomer). Indeed, given $\delta(0) = 0$, $\delta(x)$ being ε-Lipschitz implies ε-linear growth. The norm-bounded adversary can be used to capture small disturbances to the dynamics, such as process noise, while the Lipschitz adversary can be used to capture *model* error between the training and test trajectories. We also define an adversary that is the linear combination of the norm-bounded and Lipschitz adversaries, which leads to the following tube of perturbed trajectories:

$$\Delta_{\varepsilon_x,\varepsilon_u}^{x,u}(\xi) := \{\tilde{\varphi} : \tilde{\varphi}_t = f(\tilde{\varphi}_t) + \delta^x(\tilde{\varphi}_t) + \delta_t^u,\ \tilde{\varphi}_0 = \xi,\ \|\delta^x(\tilde{\varphi}_t)\|_2 \leqslant \varepsilon_x \|\tilde{\varphi}_t\|_2,\ \|\delta_t^u\|_2 \leqslant \varepsilon_u\}. \tag{7}$$

Here, the δ_t^u are additionally assumed to be locally integrable with respect to t. The tube (7) of perturbed trajectories defines a natural way of capturing the sim-to-real gap through the effects of both unmodeled dynamics (δ^x) and process noise (δ^u).

In order to accommodate additive disturbances in our stability analysis, we modify the violation function (1) to certify *practical stability* (Lin et al., 1995), i.e., convergence to a ball about the origin. To that end, for $\nu \geqslant 0$, define the adversarial violation function:

$$\tilde{h}_\nu(\xi, V) := \sup_{\tilde{\varphi}\in\Delta_\varepsilon} \sup_{t\in\mathcal{T}} \left\langle \nabla V(\tilde{\varphi}_t(\xi)), \dot{\tilde{\varphi}}_t(\xi)\right\rangle + \eta V(\tilde{\varphi}_t(\xi)) - \nu. \tag{8}$$

With this definition, finding an adversarially robust certificate of practical stability from data can be posed as solving the following feasibility problem analogous to (3):

$$\text{Find}_{V\in\mathcal{V}} \text{ s.t. } \tilde{h}_\nu(\xi_i, V) \leqslant -\tau, \quad i = 1, \ldots, n. \tag{9}$$

Letting $\tilde{V}_n$ be the solution to (9), we consider the analogous generalization error to (4):

$$\text{err}(\tilde{V}_n) := \mathbb{P}_{\xi\sim\mathcal{D}}\left[\tilde{h}_\nu\left(\xi, \tilde{V}_n\right) > 0\right]. \tag{10}$$

Our goal is to show that the fast rates $\tilde{O}(k/n)$ enjoyed in the nominal setting are *preserved* in the adversarial setting when the underlying system satisfies certain incremental stability conditions.

1. Existence, uniqueness, and completeness of the perturbed trajectories over the interval $[0, T]$ can be guaranteed under various assumptions. As an example, the set (5) is well-defined if $f(x)$ is assumed to be continuous in x and input-to-state stable such that $\tilde{\varphi}_t \in S$ for all $t \geqslant 0$ (Sontag, 2013, Prop. C.3.5). Similarly, the set (6) is well-defined if we additionally assume that $f(x)$ is globally Lipschitz in x (Sontag, 2013, Prop. C.3.8). We note that alternative assumptions on $f(x) + \delta(x)$ can be used to ensure completenes, e.g., that $f(x) + \delta(x)$ is stable in the sense of Lyapunov for all admissible δ.

3. Sample Complexity of Learning Adversarially Robust Stability Certificates

We first introduce our main stability assumption on the system dynamics.

Assumption 1 (Stability in the sense of Lyapunov) *Fix a perturbation set $\Delta_\varepsilon(\cdot)$. There exists a compact set $S \subseteq \mathbb{R}^p$ such that $\tilde{\varphi}_t(\xi) \in S$ for all $\xi \in \mathcal{X}$, $t \in \mathcal{T}$, and $\tilde{\varphi}_t(\cdot) \in \Delta_\varepsilon(\xi)$.*

For norm-bounded adversaries (5), this assumption is satisfied if the underlying nominal dynamics are input-to-state stable (Lin et al., 1995). For Lipschitz (6) and combined (7) adversaries, additional care must be taken to ensure that $f(x)+\delta^x(x)$ remains input-to-state stable for all admissible $\delta^x(x)$.

We further make the following regularity assumptions on the certificate function class $\mathcal{V}$.

Assumption 2 (Regularity of $\mathcal{V}$) *There exists constants L_V, $L_{\nabla V}$ such that for every $V \in \mathcal{V}$, the maps $x \mapsto V(x)$ and $x \mapsto \langle \nabla V(x), f(x) \rangle$ over $x \in S$ are L_V and $L_{\nabla V}$-Lipschitz, respectively.*

Under Assumptions 1 and 2 and the continuity of the nominal dynamics $f(x)$, there exist constants B_V, $B_{\nabla V}$, and $B_{\tilde{h}}$ such that

$$\sup_{V\in\mathcal{V}} \sup_{x\in S} |V(x)| \leqslant B_V, \ \sup_{V\in\mathcal{V}} \sup_{x\in S} \|\nabla V(x)\|_2 \leqslant B_{\nabla V}, \ \sup_{V\in\mathcal{V}} \sup_{\xi\in\mathcal{X}} \left|\tilde{h}(\xi, V)\right| \leqslant B_{\tilde{h}}.$$

Finally let $\|V\|_{\mathcal{V}} := \sup_{x\in S} \left\| \begin{bmatrix} V(x) \\ \nabla V(x) \end{bmatrix} \right\|_2$ denote the supremum norm on the space $\mathcal{V}$.

Borrowing from the key insight in Boffi et al. (2020a), we observe that any feasible solution $\tilde{V}_n$ to (9) achieves zero empirical risk on the loss $\tilde{\ell}_n(V) := \frac{1}{n}\sum_{i=1}^n \mathbf{1}\left\{\tilde{h}(\xi_i, V) > -\tau\right\}$. Therefore, results from statistical learning theory regarding zero empirical risk minimizers can be applied to get fast rates for the generalization error. To do so, we define the adversarial loss class $\tilde{\mathcal{H}} := \left\{\tilde{h}(\cdot, V),\ V \in \mathcal{V}\right\}$. Lemma 4.1 from Boffi et al. (2020a), which is in turn adapted from Theorem 5 of Srebro et al. (2010), immediately gives the following bound on the generalization error.

Lemma 1 (Generalization error bound) *Fix a $\delta \in (0,1)$. Let us assume Assumptions 1 and 2. Suppose that the optimization problem (9) is feasible and $\tilde{V}_n$ is a solution. Then the following holds with probability at least $1-\delta$ over $\xi_1, \ldots, \xi_n$ drawn i.i.d. from $\mathcal{D}$:*

$$\operatorname{err}\left(\tilde{V}_n\right) \leqslant K\left(\frac{\log^3(n)}{\tau^2}\mathcal{R}_n^2(\tilde{\mathcal{H}}) + \frac{\log\left(\log\left(B_{\tilde{h}}/\tau\right)/\delta\right)}{n}\right), \tag{11}$$

where $K > 0$ is a universal constant and

$$\mathcal{R}_n(\tilde{\mathcal{H}}) := \sup_{\xi_1,\ldots,\xi_n\in\mathcal{X}} \mathbb{E}_{\sigma\sim\mathrm{Unif}\{\pm1\}^n}\left[\sup_{\tilde{h}(\cdot,V)\in\tilde{\mathcal{H}}} \frac{1}{n}\left|\sum_{i=1}^n \sigma_i \tilde{h}(\xi_i, V)\right|\right]$$

is the Rademacher complexity of the adversarial loss class $\tilde{\mathcal{H}}$.

Lemma 1 reduces bounding the generalization error of an adversarially robust stability certificate to bounding the Rademacher complexity of the adversarial loss class $\tilde{\mathcal{H}}$. We note that the nominal results of Boffi et al. (2020a, Lemma 4.1) are recovered by setting the perturbation budget $\varepsilon = 0$.

3.1. A Simple Adversary-Agnostic Rademacher Complexity Bound

A standard technique for controlling the Rademacher complexity $\mathcal{R}(\tilde{\mathcal{H}})$ is appealing to Dudley's entropy integral (Wainwright, 2019, Ch 5.). Specifically, if we show that for some $L_{\tilde{h}}$,

$$\left|\tilde{h}(\xi, V_1) - \tilde{h}(\xi, V_2)\right| \leqslant L_{\tilde{h}} \left\|V_1 - V_2\right\|_{\mathcal{V}} \;\; \forall \xi \in \mathcal{X}, \;\; V_1, V_2 \in \mathcal{V},$$

then Dudley's inequality implies the bound $\mathcal{R}_n(\tilde{\mathcal{H}}) \leqslant \frac{24 L_{\tilde{h}}}{\sqrt{n}} \int_0^\infty \sqrt{\log N(\varepsilon; \mathcal{V}, \|\cdot\|_{\mathcal{V}})}\, d\varepsilon$. Our first result shows that our main assumptions are sufficient to ensure that $L_{\tilde{h}}$ can be controlled with a uniform boundedness assumption on the adversary.

Lemma 2 (Uniformly bounded adversaries are sufficient) *Suppose that (i) Assumptions 1 and 2 hold, (ii) $B_\delta := \sup_{x \in S} \sup_{t \in \mathcal{T}} \|\delta(t, x)\|_2$ is finite, and (iii) the flow $\tilde{\varphi}_t(\xi)$ is unique and complete over $\mathcal{T}$ for all $\xi \in \mathcal{X}$ and all admissible $\delta(x, t)$. Let L_h denote any constant such that $|h(\xi, V_1) - h(\xi, V_2)| \leqslant L_h \|V_1 - V_2\|_{\mathcal{V}}$ for all $\xi \in X$ and $V_1, V_2 \in \mathcal{V}$. Then, $L_{\tilde{h}} \leqslant L_h + B_\delta$.*

Lemma 2 shows that if the nominal system is input-to-state stable and if the adversary is uniformly bounded over the set S from Assumption 1, then by Dudley's inequality, the Rademacher complexity $\mathcal{R}_n(\tilde{\mathcal{H}})$ is on the same order as the nominal complexity $\mathcal{R}_n(\mathcal{H})$. Consequently by Lemma 1, the adversarial generalization bound $\mathrm{err}(\tilde{V}_n)$ is on the same order as the nominal bound $\mathrm{err}(\hat{V}_n)$. We show next that with stronger assumptions on the stability of the dynamics, we can obtain bounds on $\mathcal{R}_n(\tilde{\mathcal{H}})$ that are additive, rather than muliplicative, with respect to the nominal complexity $\mathcal{R}_n(\mathcal{H})$. Furthermore, these bounds are also robust to Lipschitz adversarial perturbations.

3.2. Improving the Adversarial Rademacher Complexity via Stability

To improve the bound from Lemma 2, we first adapt a fundamental fact from the calculus of Rademacher complexities (Bartlett and Mendelson, 2002, Thm. 12, Property 5), along with the trivial observation that $\tilde{h}(\cdot, V) = h(\cdot, V) + \left(\tilde{h}(\cdot, V) - h(\cdot, V)\right)$ to conclude that:

$$\mathcal{R}_n(\tilde{\mathcal{H}}) \leqslant \mathcal{R}_n(\mathcal{H}) + \sup_{\xi \in X} \sup_{V \in \mathcal{V}} \frac{1}{\sqrt{n}} \left|\tilde{h}(\xi, V) - h(\xi, V)\right|. \tag{12}$$

Therefore, in order to bound $\mathcal{R}_n(\tilde{\mathcal{H}})$, it suffices to uniformly bound $\tilde{h}(\xi, V) - h(\xi, V)$ over $\xi \in X$ and $V \in \mathcal{V}$. To do so, we introduce the notion of (β, ρ, γ)-exponential-incrementally-input-to-state stability (Angeli, 2002; Boffi et al., 2020b).

Definition 3 ((β, ρ, γ)-E-δISS) *Let $\beta, \rho, \gamma > 0$ be positive constants. A continuous-time dynamical system $\dot{x} = f(x, t)$ is (β, ρ, γ)-exponential-incrementally-input-to-state stable ((β, ρ, γ)-E-δISS) if, for any pair of initial conditions (x_0, y_0) and signal $u(t)$ – which can depend causally on x, y – the trajectories $\dot{x}(t) = f(x(t))$ and $\dot{y}(t) = f(y(t)) + u(t)$ satisfy for all $t \geqslant 0$:*

$$\|x_t - y_t\|_2 \leqslant \beta \|x_0 - y_0\|_2 e^{-\rho t} + \gamma \int_0^t e^{-\rho(t-s)} \|u_s\|_2 \, ds.$$

In short, the dependence of the distance between two trajectories on the initial conditions shrinks exponentially with time (incremental stability), and is input-to-state stable with respect to the inputs entering y_t. This notion of stability is strongly related to notion of contraction (Lohmiller and Slotine, 1998), as illustrated by the following lemma.

Lemma 4 (Contraction implies E-δISS) *Let $M(x,t)$ denote a positive definite Riemannian metric and $f(x,t)$ denote a continuous-time dynamical system. Suppose both M and f are continuously differentable, and that there are constants $0 < \mu \leqslant L < \infty$ and $\lambda > 0$ such that for all $x \in \mathbb{R}^n$ and $t \in \mathbb{R}_{\geqslant 0}$, the metric $M(x,t)$ satisfies $\mu I \preccurlyeq M(x,t) \preccurlyeq LI$, and the function $f(x,t)$ satisfies:*

$$\frac{\partial f}{\partial x}(x,t)^\top M(x,t) + M(x,t)\frac{\partial f}{\partial x}(x,t) + \dot{M}(x,t) \preccurlyeq -2\lambda M(x,t).$$

Then, the dynamical system $\dot{x} = f(x,t)$ is $(\sqrt{L/\mu}, \lambda, \sqrt{L/\mu})$-E-$\delta$ISS.

Lemma 4 is the analogous result of Proposition 5.3 of Boffi et al. (2020b) for continuous-time systems. We note that this result originally appeared in (Lohmiller and Slotine, 1998, Section 3.7, Remark (vii)) without proof. Leveraging (β, ρ, γ)-E-δISS, we can derive a uniform bound on $|\tilde{h}(\xi, V) - h(\xi, V)|$ that scales with the stability parameters of the underlying system, which combined with inequality (12) yields the following bounds on $\mathcal{R}_n(\tilde{\mathcal{H}})$ for the tubes (5)-(7).

Theorem 5 (E-δISS yields additive bounds) *Put $B_X := \sup_{\xi \in \mathcal{X}} \|\xi\|_2$, let Assumption 2 hold, and assume that the nominal system $f(x)$ is (β, ρ, γ)-E-δISS. Then for*

- *adversarial trajectories drawn from the norm-bounded tube $\Delta_\varepsilon^u(\xi)$ defined in (5), Assumption 1 holds and*

$$\mathcal{R}_n(\tilde{\mathcal{H}}) \leqslant \mathcal{R}_n(\mathcal{H}) + \left[(L_{\nabla V} + \eta L_V)\gamma\varepsilon\rho^{-1} + B_{\nabla V}\varepsilon + \nu\right]\frac{1}{\sqrt{n}}, \tag{13}$$

- *adversarial trajectories drawn from the Lipschitz tube $\Delta_\varepsilon^x(\xi)$ defined in (6), if $\varepsilon > 0$ is small enough such that $\gamma\varepsilon < \rho$, then Assumption 1 holds and*

$$\mathcal{R}_n(\tilde{\mathcal{H}}) \leqslant \mathcal{R}_n(\mathcal{H}) + \Bigg[(L_{\nabla V} + \eta L_V + B_{\nabla V}\varepsilon)\frac{\gamma\varepsilon\rho^{-1}}{1-\gamma\varepsilon\rho^{-1}}e^{-1}B_X\beta\varepsilon + B_{\nabla V}B_X\beta\varepsilon + \nu\Bigg]\frac{1}{\sqrt{n}}, \tag{14}$$

- *adversarial trajectories drawn from the combined tube $\Delta_{\varepsilon_x,\varepsilon_u}^{x,u}$ defined in (7), if $\varepsilon_x > 0$ is small enough such that $\gamma\varepsilon_x < \rho$, then Assumption 1 holds and*

$$\mathcal{R}_n(\tilde{\mathcal{H}}) \leqslant \mathcal{R}_n(\mathcal{H}) + \Bigg[(L_{\nabla V} + \eta L_V + B_{\nabla V}\varepsilon_x)\frac{\gamma\varepsilon_u\rho^{-1} + \gamma\varepsilon_x\rho^{-1}e^{-1}B_X\beta\varepsilon_x}{1-\gamma\varepsilon_x\rho^{-1}} + B_{\nabla V}\beta\varepsilon_x B_X + B_{\nabla V}\varepsilon_u + \nu\Bigg]\frac{1}{\sqrt{n}}. \tag{15}$$

In particular, Theorem 5 shows that $\mathcal{R}_n(\tilde{\mathcal{H}}) \leqslant \mathcal{R}_n(\mathcal{H}) + O(1)\frac{1}{\sqrt{n}}$ for all the aforementioned adversary classes. Here, $O(1)$ suppresses all problem specific constants. This demonstrates that under the assumptions of Theorem 5, the Rademacher complexity of the resulting adversarial loss class is no more than an additive factor of order $O(1/\sqrt{n})$ greater than the Rademacher complexity class of the nominal loss class. Because a typical scaling of $\mathcal{R}_n(\mathcal{H}) \asymp \sqrt{k/n}$ where k is the effective degrees of freedom of $\mathcal{V}$, the $O(1/\sqrt{n})$ term is often negligible compared to $\mathcal{R}_n(\mathcal{H})$.

The bounds in Theorem 5 involving the Lipschitz adversary are only valid when the denominator $1-\gamma\varepsilon\rho^{-1}$ is positive, hence the necessary assumption that $\gamma\varepsilon < \rho$. This is a necessary assumption; when the budget for the Lipschitz adversary is too large, then an adversary can cause the system to diverge exponentially. To illustrate this, consider the scalar system $\dot{x} = -\rho x$, which we can

verify is $(1, \rho, 1)$-E-δISS, perturbed by a ε-Lipschitz adversary that adds εx to the dynamics such that $\dot{y} = -(\rho - \varepsilon)y$. If $\varepsilon > \rho$, then the perturbed trajectory will diverge away from 0 exponentially and we cannot hope to find a uniform bound on $\tilde{h}(\xi, V) - h(\xi, V)$ for all t.

We conclude this section with an important example of a certificate function class and its associated Rademacher complexities. This example further highlights that the additive $O(1/\sqrt{n})$ factor is comparatively negligible for many certificate function classes of interest.

Example 1 (Lipschitz Parametric Function Classes) *Consider the parametric function class*

$$\mathcal{V} = \Big\{ V_\theta(\cdot) = g(\cdot, \theta) : \theta \in \mathbb{R}^k,\ \|\theta\| \leqslant B_\theta \Big\}, \tag{16}$$

where we assume $g : \mathbb{R}^p \times \mathbb{R}^k \to \mathbb{R}$ is twice-continuously differentiable. The description (16) is very general; for example, feed-forward neural networks with differentiable activation functions and sum-of-squares polynomials lie in this function class. It is shown in Boffi et al. (2020a) that $\mathcal{R}_n(\mathcal{H}) = O(\sqrt{k/n})$. Combining this with Theorem 5, we conclude that

$$\mathcal{R}_n(\tilde{\mathcal{H}}) \leqslant \mathcal{R}_n(\mathcal{H}) + O(1/\sqrt{n}) = O(\sqrt{k/n}).$$

4. Learning Adversarially Robust Certificates in Practice

In this section, we illustrate the practicality and effectiveness of learning adversarial certificates. We consider the damped pendulum with dynamics $m\ell^2\ddot{\theta} + b\dot{\theta} + mg\ell\sin(\theta) = 0$, where we set $m = 1$, $\ell = 1$, $b = 2$, and $g = 9.81$. The state space is given by $x = (\theta, \dot{\theta}) \in \mathbb{R}^2$ with stable equilibrium is at the origin and we wrap θ to the interval $(-\pi, \pi]$. Consider the following certificate function class

$$\mathcal{V} = \Big\{ V_\theta(x) = x^\top \Big(L_\theta(x)^\top L_\theta(x) + I \Big) x,\ \theta \in \mathbb{R}^{p \times h \times h \times p \cdot (2p)} \Big\}, \tag{17}$$

where $L_\theta(x) \in \mathbb{R}^{2p \times p}$ is the re-shaped output of a fully-connected neural network with 2 hidden layers of width $h = 20$ and $\tanh$ activations.

We first demonstrate the robustness properties of an adversarially trained Lyapunov function versus a nominal one. We collect $n = 1000$ trajectories with randomly sampled initial conditions $\xi \sim \mathrm{Unif}\big([-2, 2]^2\big)$. Each trajectory is rolled out using `scipy.integrate.solve_ivp` with horizon $T = 8$ and $dt = 0.05$, such the size of the total dataset is $1000 \times 160 \times 2$. Following Boffi et al. (2020a), the nominal Lyapunov function V_{nom} is learned by minimizing the surrogate loss

$$L(\theta; \eta, \lambda) = \sum_{i=1}^{1000} \sum_{k=1}^{160} \mathrm{ReLU}[\langle \nabla V_\theta(x_i(k)), \dot{x}_i(k) \rangle + \eta V_\theta(x_i(k))] + \lambda \|\theta\|_2^2, \tag{18}$$

where we set the exponential rate $\eta = 0.4$ and regularization parameter $\lambda = 0.1$. The loss is minimized for 1000 epochs with Adam (Kingma and Ba, 2015) with cosine decay, initialized at step size 0.005, and batch size 1000.

Solving for the adversarially robust Lyapunov function is challenging due to the inner maximization problem over perturbations entering through the dynamics. As is standard in the adversarial learning literature, we instead approximate the true adversarially robust loss function via an alternating scheme, summarized in Algorithm 1. We set $m = 10$, and each inner minimization of $L(\theta; \eta, \lambda)$ runs for 100 epochs. The approximate adversarial computation uses a simple greedy heuristic: at any x, the maximal direction to increase the Lyapunov decrease condition $\langle \nabla V(x), f(x) + \delta \rangle + \eta V(x)$ is $\delta = c \nabla V(x)$, where $c > 0$ is a normalizing factor to adjust δ for the

adversarial budget ε. In this experiment, we use the Lipschitz adversary, and thus $c = \varepsilon \frac{\|x\|_2}{\|\nabla V(x)\|_2}$. We note that Algorithm 1 does not necessarily compute optimal adversarial perturbations, and therefore the output V_{adv} can only be less robust than the true adversarially robust function $\tilde{V}_n$. Nevertheless, V_{adv} performs well in the face of practically relevant perturbations to the system, highlighting the benefit of adversarial training.

Algorithm 1: Training adversarially robust Lyapunov function V_{adv} (Lipschitz adversary)

Input: Initial conditions $\{\xi_i\}_{i=1}^n$, rate $\eta > 0$, adversarial budget $\varepsilon > 0$, alternations m.

Compute nominal trajectories $\mathbf{T} = \{x(\xi_i)\}_{i=1}^n$.

for $i = 1, ..., m-1$ **do**

  Minimize $L(\theta; \eta, \lambda)$ with respect to $\mathbf{T}$ to get V.

  Re-compute $\mathbf{T}$ using dynamics $\dot{x}_i(t) = f(x_i(t)) + \varepsilon \frac{\|x_i(t)\|_2}{\|\nabla V(x_i(t))\|_2} \nabla V(x_i(t))$, $x_i(0) = \xi_i$.

end

Minimize $L(\theta; \eta, \lambda)$ with respect to $\mathbf{T}$ to get V.

Output: Adversarially trained Lyapunov function $V_{\text{adv}} = V$.

We assess the robustness of the nominal and robust certificates V_{nom} and V_{adv} by measuring how well they certify stability on various classes of perturbed trajectories. We first draw an additional test set of $n = 1000$ initial conditions from $\text{Unif}\left([-2, 2]^2\right)$. For each class of perturbation, we vary the decrease rate parameter $\eta \in [0, 1]$ (recall that the certificates V_{nom} and V_{adv} were trained with decrease rate $\eta = 0.4$) and measure both the proportion of whole trajectories as well as the total proportion of the 1000×160 states that satisfy the Lyapunov decrease condition with rate η.

We consider the following four classes of perturbed trajectories:

1. $\dot{x} = f(x) + \varepsilon \frac{\|x\|_2}{\|\nabla V_{\text{adv}}(x)\|_2} \nabla V_{\text{adv}}(x)$, analogous to the adversarial training process,
2. $\dot{x} = f(x) + \varepsilon x$, which is a Lipschitz adversary that aims to greedily maximize $\|x(t)\|_2^2$ at any given time t,
3. the dynamics resulting from using the linearization of the damped pendulum at the origin to generate the trajectories, and
4. the dynamics resulting from setting $\tilde{m} = \tilde{\ell} = 1.1$ instead of $m = \ell = 1$.

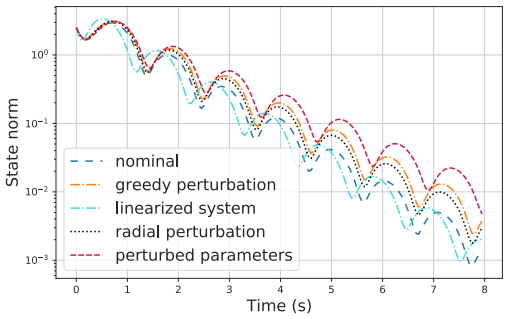

Figure 1: Norm of pendulum state over time starting at a fixed initial condition, for nominal and perturbed trajectories perturbed as described in Section 4.

The perturbation class 1 acts in the direction ∇V_{adv}, and thus the perturbed trajectories are tuned to degrade the performance of V_{adv}. Additionally, the perturbation classes 3 and 4 can be viewed as instances of the sim-to-real gap, where there are model discrepancies between training and test.

Figure 2 plots the resulting Lyapunov decrease satisfaction rates for each type of perturbation. We observe that for each type of perturbation, the nominal certificate V_{nom} fails to certify a significant proportion of trajectories when $\eta = 0.4$. In contrast, the robust certificate V_{adv} certifies all trajectories for decrease rate $\eta = 0.4$. We further observe that the robust certificate is also able to certify *faster* decrease rates as well. Finally, we note that the trajectories resulting from perturbed

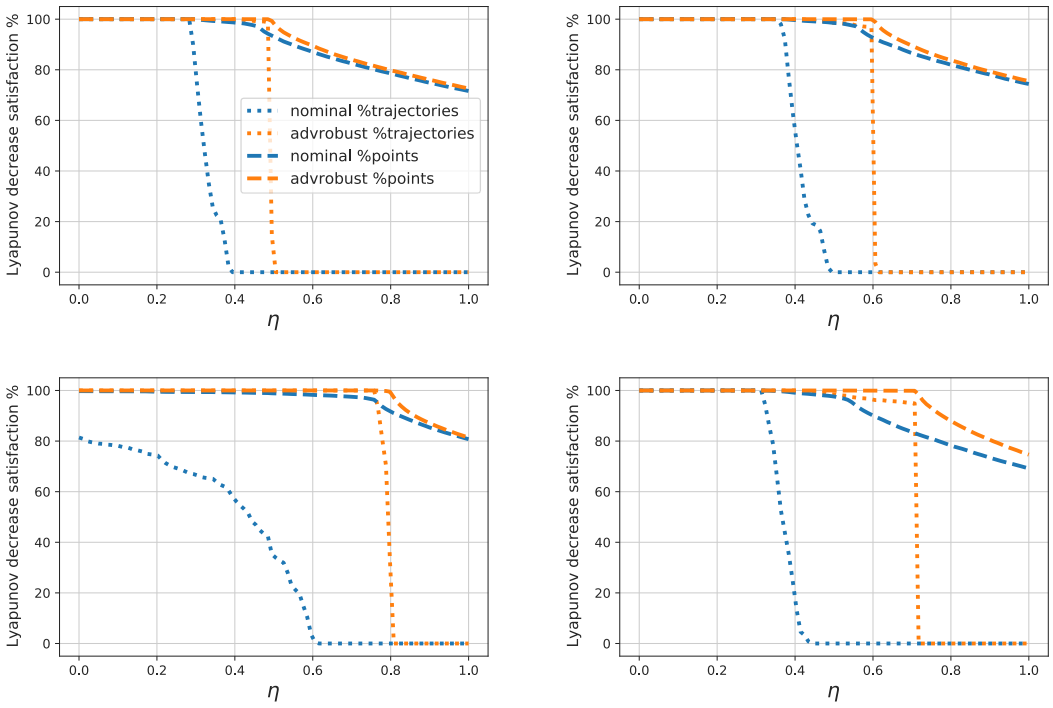

Figure 2: Satisfaction rate of the Lyapunov decrease condition versus the exponential rate parameter η of nominal and adversarially trained certificates V_{nom} and V_{adv} for four classes of perturbed trajectories. The percentage of trajectories and of total points satisfying the Lyapunov decrease condition for V_{nom} and V_{adv} are shown. Both V_{nom} and V_{adv} were trained with $\eta = 0.4$. Trajectories were generated by rolling out a fixed set of 1000 initial conditions sampled from $\text{Unif}\left([-2,2]^2\right)$. **Upper left**: dynamics generated from gradient ascent on the adversarial certificate V_{adv}, $\dot{x} = f(x) + \varepsilon \frac{\|x\|}{\|\nabla V_{\text{adv}}\|} \nabla V_{\text{adv}}$. **Upper right**: dynamics generated from a radial perturbation, $\dot{x} = f(x) + \varepsilon x$. **Lower left**: dynamics generated from system linearized at origin, $\dot{x} = J_{(0,0)} x$. **Lower right**: dynamics generated from perturbing the pendulum parameters, $\tilde{m} = 1.1$, $\tilde{\ell} = 1.1$.

pendulum parameters (perturbation class 4) actually cause the system to be more unstable than the greedy perturbations (perturbation class 1) used during training (see Figure 1). Nevertheless, the robust certificate V_{adv} is able to certify stability for a large range of η.

5. Conclusion

Motivated by bridging the sim-to-real gap, we proposed and analyzed an approach to learning adversarially robust Lyapunov certificates. We showed that for systems that enjoy exponential incremental input-to-state stability, stability certificate functions that are robust to norm-bounded and Lipschitz adversarial perturbations to the system dynamics can be learned with negligible statistical overhead as compared to the nominal case. Future research directions include exploring the statistical tradeoffs occurring from progressively weaker notions of stability (e.g., incremental gain stability as defined in Tu et al. (2021)), providing approximation guarantees for the adversarial training algorithm proposed in Section 4, and extending our results to provide statistical guarantees for policies synthesized from robust certificate functions (Lindemann et al., 2021; Taylor et al., 2021).

Acknowledgments

The authors thank Alexander Robey and Bruce D. Lee for various helpful discussions. Nikolai Matni is funded by NSF awards CPS-2038873, CAREER award ECCS-2045834, and a Google Research Scholar award.

References

David Angeli. A lyapunov approach to incremental stability properties. *IEEE Transactions on Automatic Control*, 47(3):410–421, 2002.

Idan Attias, Aryeh Kontorovich, and Yishay Mansour. Improved generalization bounds for robust learning. In *Algorithmic Learning Theory*, 2019.

Peter L. Bartlett and Shahar Mendelson. Rademacher and gaussian complexities: Risk bounds and structural results. *Journal of Machine Learning Research*, 3(Nov):463–482, 2002.

Nicholas M. Boffi, Stephen Tu, Nikolai Matni, Jean-Jacques E. Slotine, and Vikas Sindhwani. Learning stability certificates from data. In *Conference on Robot Learning*, 2020a.

Nicholas M. Boffi, Stephen Tu, and Jean-Jacques E. Slotine. Regret bounds for adaptive nonlinear control. In *Learning for Dynamics and Control*, 2020b.

Ya-Chien Chang, Nima Roohi, and Sicun Gao. Neural lyapunov control. In *Neural Information Processing Systems*, 2019.

Shaoru Chen, Mahyar Fazlyab, Manfred Morari, George J. Pappas, and Victor M. Preciado. Learning lyapunov functions for piecewise affine systems with neural network controllers. *arXiv preprint arXiv:2008.06546*, 2020.

Justin Fu, Katie Luo, and Sergey Levine. Learning robust rewards with adversarial inverse reinforcement learning. In *International Conference on Learning Representations*, 2018.

Peter Giesl, Boumediene Hamzi, Martin Rasmussen, and Kevin Webster. Approximation of lyapunov functions from noisy data. *Journal of Computational Dynamics*, 7(1):57–81, 2020.

Adam Gleave, Michael Dennis, Cody Wild, Neel Kant, Sergey Levine, and Stuart Russell. Adversarial policies: Attacking deep reinforcement learning. In *International Conference on Learning Representations*, 2020.

Jonathan Ho and Stefano Ermon. Generative adversarial imitation learning. In *Neural Information Processing Systems*, 2016.

Wanxin Jin, Zhaoran Wang, Zhuoran Yang, and Shaoshuai Mou. Neural certificates for safe control policies. *arXiv preprint arXiv:2006.08465*, 2020.

Joris Kenanian, Ayca Balkan, Raphael M. Jungers, and Paulo Tabuada. Data driven stability analysis of black-box switched linear systems. *Automatica*, 109:108533, 2019.

Bachir El Khadir, Jake Varley, and Vikas Sindhwani. Teleoperator imitation with continuous-time safety. In *Robotics: Science and Systems*, 2019.

Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *International Conference on Learning Representations*, 2015.

Alexey Kurakin, Ian Goodfellow, and Samy Bengio. Adversarial examples in the physical world. *arXiv preprint arXiv:1607.02533*, 2016.

Bruce D. Lee, Thomas T. C. K. Zhang, Hamed Hassani, and Nikolai Matni. Adversarial tradeoffs in linear inverse problems and robust state estimation. *arXiv preprint arXiv:2111.08864*, 2021.

Yuandan Lin, Eduardo Sontag, and Yuan Wang. Various results concerning set input-to-state stability. In *1995 34th IEEE Conference on Decision and Control*, 1995.

Lars Lindemann, Alexander Robey, Lejun Jiang, Stephen Tu, and Nikolai Matni. Learning robust output control barrier functions from safe expert demonstrations. *arXiv preprint arXiv:2111.09971*, 2021.

Winfried Lohmiller and Jean-Jacques E. Slotine. On contraction analysis for non-linear systems. *Automatica*, 34(6):683–696, 1998.

Michael Lutter, Shie Mannor, Jan Peters, Dieter Fox, and Animesh Garg. Robust value iteration for continuous control tasks. *arXiv preprint arXiv:2105.12189*, 2021.

Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards deep learning models resistant to adversarial attacks. In *International Conference on Learning Representations*, 2018.

Ian R. Manchester and Jean-Jacques E. Slotine. Control contraction metrics: Convex and intrinsic criteria for nonlinear feedback design. *IEEE Transactions on Automatic Control*, 62(6):3046–3053, 2017.

Gaurav Manek and J. Zico Kolter. Learning stable deep dynamics models. In *Neural Information Processing Systems*, 2019.

Omar Montasser, Steve Hanneke, and Nathan Srebro. Vc classes are adversarially robustly learnable, but only improperly. In *Conference on Learning Theory*, 2019.

Lerrel Pinto, James Davidson, Rahul Sukthankar, and Abhinav Gupta. Robust adversarial reinforcement learning. In *International Conference on Machine Learning*, 2017.

Hadi Ravanbakhsh and Sriram Sankaranarayanan. Learning control lyapunov functions from counterexamples and demonstrations. *Autonomous Robots*, 43:275–307, 2019.

Spencer M. Richards, Felix Berkenkamp, and Andreas Krause. The lyapunov neural network: Adaptive stability certification for safe learning of dynamical systems. In *Conference on Robot Learning*, 2018.

Alexander Robey, Haimin Hu, Lars Lindemann, Hanwen Zhang, Dimos V. Dimarogonas, Stephen Tu, and Nikolai Matni. Learning control barrier functions from expert demonstrations. In *2020 59th IEEE Conference on Decision and Control*, 2020.

Ludwig Schmidt, Shibani Santurkar, Dimitris Tsipras, Kunal Talwar, and Aleksander Madry. Adversarially robust generalization requires more data. In *Neural Information Processing Systems*, 2018.

Vikas Sindhwani, Stephen Tu, and Seyed Mohammad Khansari-Zadeh. Learning contracting vector fields for stable imitation learning. *arXiv preprint arXiv:1804.04878*, 2018.

Sumeet Singh, Anirudha Majumdar, Jean-Jacques E. Slotine, and Marco Pavone. Robust online motion planning via contraction theory and convex optimization. In *2017 IEEE International Conference on Robotics and Automation*, 2017.

Sumeet Singh, Spencer M. Richards, Jean-Jacques E. Slotine, Vikas Sindhwani, and Marco Pavone. Learning stabilizable nonlinear dynamics with contraction-based regularization. *International Journal of Robotics Research*, 40(10–11):1123–1150, 2020.

Eduardo Sontag. *Mathematical control theory: deterministic finite dimensional systems*, volume 6. Springer, 2013.

Nathan Srebro, Karthik Sridharan, and Ambuj Tewari. Smoothness, low-noise and fast rates. In *Neural Information Processing Systems*, 2010.

Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian Goodfellow, and Rob Fergus. Intriguing properties of neural networks. *arXiv preprint arXiv:1312.6199*, 2013.

Andrew J. Taylor, Andrew Singletary, Yisong Yue, and Aaron D. Ames. Learning for safety-critical control with control barrier functions. *arXiv preprint arXiv:1912.10099*, 2019.

Andrew J. Taylor, Victor D. Dorobantu, Sarah Dean, Benjamin Recht, Yisong Yue, and Aaron D. Ames. Towards robust data-driven control synthesis for nonlinear systems with actuation uncertainty. *arXiv preprint arXiv:2011.10730*, 2021.

Josh Tobin, Rachel Fong, Alex Ray, Jonas Schneider, Wojciech Zaremba, and Pieter Abbeel. Domain randomization for transferring deep neural networks from simulation to the real world. In *2017 IEEE/RSJ International Conference on Intelligent Robots and Systems*, 2017.

Faraz Torabi, Garrett Warnell, and Peter Stone. Generative adversarial imitation from observation. *arXiv preprint arXiv:1807.06158*, 2019.

Anastasios Tsiamis and George J. Pappas. Linear systems can be hard to learn. *arXiv preprint arXiv:2104.01120*, 2021.

Anastasios Tsiamis, Nikolai Matni, and George J. Pappas. Sample complexity of kalman filtering for unknown systems. In *Learning for Dynamics and Control*, 2020.

Stephen Tu, Alexander Robey, Tingnan Zhang, and Nikolai Matni. On the sample complexity of stability constrained imitation learning. *arXiv preprint arXiv:2102.09161*, 2021.

Martin J. Wainwright. *High-Dimensional Statistics: A Non-Asymptotic Viewpoint*. Cambridge University Press, 2019.

Dong Yin, Ramchandran Kannan, and Peter Bartlett. Rademacher complexity for adversarially robust generalization. In *International Conference on Machine Learning*, 2019.

Hongyang Zhang, Yaodong Yu, Jiantao Jiao, Eric P. Xing, Laurent El Ghaoui, and Michael I. Jordan. Theoretically principled trade-off between robustness and accuracy. In *International Conference on Machine Learning*, 2019.

Thomas TCK Zhang, Stephen Tu, Nicholas M Boffi, Jean-Jacques E Slotine, and Nikolai Matni. Adversarially robust stability certificates can be sample-efficient. *arXiv preprint arXiv:2112.10690*, 2021.