Adversarially Robust Stability Certificates can be Sample-Efficient

Thomas T.C.K. Zhang[1], Stephen Tu[2], Nicholas M. Boffi[3],
Jean-Jacques E. Slotine[4,2], and Nikolai Matni[1,2]

[1]Department of Electrical and Systems Engineering, University of Pennsylvania
[2]Google Brain Robotics
[3]Courant Institute of Mathematical Sciences, New York University
[4]Nonlinear Systems Laboratory, Massachusetts Institute of Technology

December 21, 2021

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.

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 [4] 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 [39], 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. [4], 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. [42], Tsiamis and Pappas [41], Lee et al. [17], Tu et al. [43] 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. Our results are presented in continuous time; however, they readily admit discrete time analogues, which are detailed in Appendix B.

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 [9, 13, 7, 28, 24, 6, 27] and barrier [37, 29, 12] functions, as well as contraction metrics [33, 23, 32] and contracting vector fields [31, 14] from data. While the generality and strength of guarantees provided vary (see the literature review of Boffi et al. [4] 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 [36, 22, 46, 16] 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 [30].

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 [40, 10], rewards [8, 11], direct disturbances to the system [26], or combinations thereof [21]. 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 [45, 2, 25]. 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. [4]. 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 [4] 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

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

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 [4], 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]:

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

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

Intuitively, $\Delta^u_\varepsilon(\xi)$ 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, $\Delta^x_\varepsilon(\xi)$ 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), misnomer notwithstanding, as *Lipschitz*. Indeed, given $\delta(0) = 0$, $\delta(x)$ being ε-Lipschitz is implied by 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:

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

Here, the δ^u_t 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* [18], 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 $t \geqslant 0$ [34, Prop. C.3.5]. Similarly, the set (6) is well-defined if we additionally assume that $f(x)$ is globally Lipschitz in x [34, 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 $\Delta(\cdot)$. 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 $\tilde{\varphi}_t(\cdot) \in \Delta(\xi)$.*

For norm-bounded adversaries (5), this assumption is satisfied if the underlying nominal dynamics are input-to-state stable [18]. 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 [4], 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 [4], which is in turn adapted from Theorem 5 of [35], 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

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

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. [4, 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 [44, 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 [3, Thm. 12, Property 5], along with the trivial identity $\tilde{h}(\cdot, V) = h(\cdot, V) + \Big(\tilde{h}(\cdot, V) - h(\cdot, V)\Big)$ 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 [1, 5].

Definition 1 ((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-*δ*ISS) 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 [20], as illustrated by the following lemma.

Lemma 3 (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 3 is the analogous result of Proposition 5.3 of [5] for continuous-time systems. We note that this result originally appeared in [20, Section 3.7, Remark (vii)] without explicit 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 1 (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*

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

- *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*

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

In particular, Theorem 1 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 1, 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 1 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. [4] that $\mathcal{R}_n(\mathcal{H}) = O(\sqrt{k/n})$. Combining this with Theorem 1, 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. [4], 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 500 epochs with Adam [15] 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 $m = 5$, 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 $c = \varepsilon \frac{\|x\|_2}{\|\nabla V(x)\|_2}$. Through Algorithm 1, we get an adversarially trained Lyapunov function V_{adv},

which can only be less robust than the true adversarially robust function $\tilde{V}_n$ due to the suboptimal adversary computation. Nevertheless, our approximate robust Lyapunov function V_{adv} is seen to perform well in the face of practically relevant perturbations to the system.

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 for
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 $m = \ell = 1$.

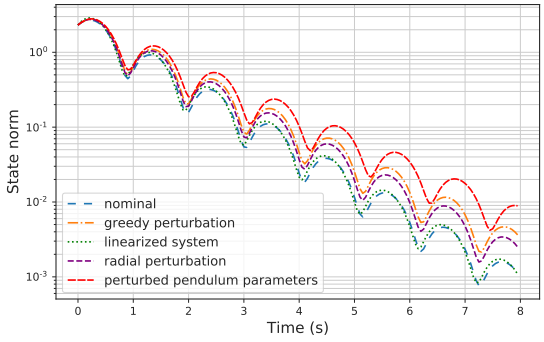

Figure 1: Norm of pendulum state over time starting at a fixed initial condition, for nominal and perturbed trajectories perturbed as described in Section 4.

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 any trajectories when $\eta = 0.4$. In contrast, the robust certificate V_{adv} certifies all trajectories for decrease rates $\eta = 0.4$. 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 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.

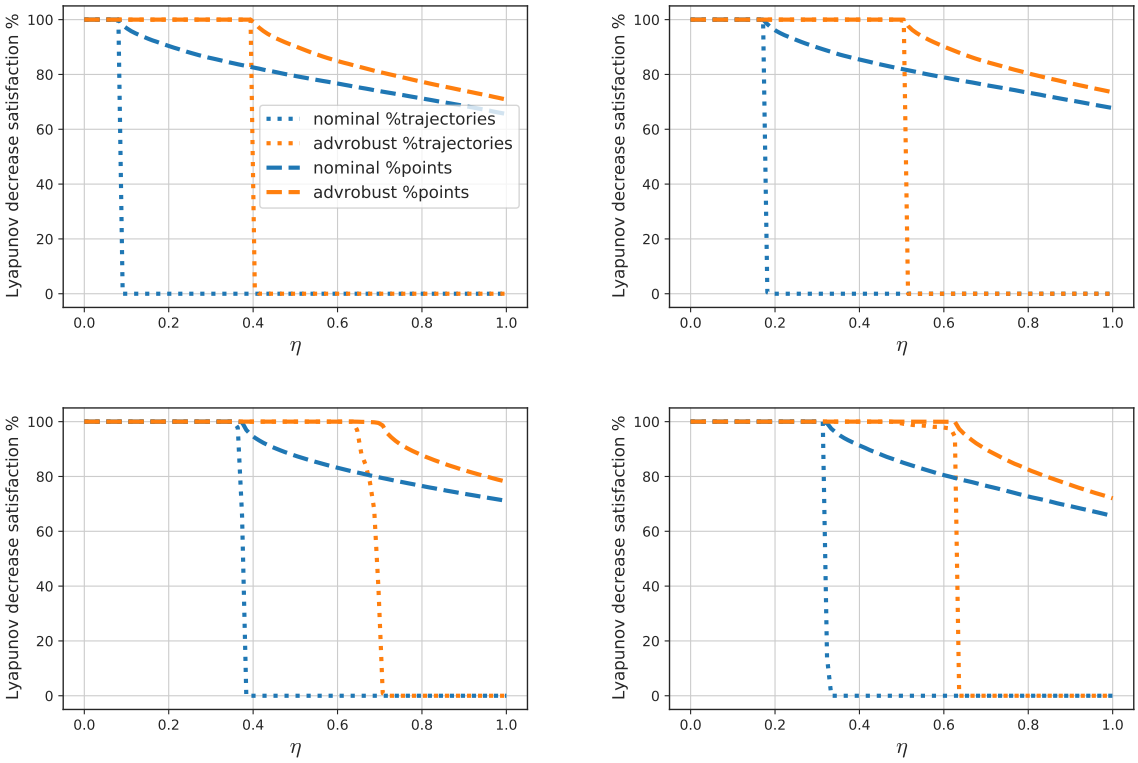

Figure 2: Satisfaction rate of the Lyapunov decrease condition versus the exponential rate parameter η of nominal and adversarially trained certificates V_{nom} and V_{adv} for four classes of perturbed trajectories. The percentage of trajectories and of total points satisfying the Lyapunov decrease 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 $\mathrm{Unif}\left([-2,2]^2\right)$. **Upper left**: dynamics generated from gradient ascent on the adversarial certificate V_{adv}, $\dot{x} = f(x) + \varepsilon \frac{\|x\|}{\|\nabla V_{\mathrm{adv}}\|} \nabla V_{\mathrm{adv}}$. **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, $\tilde{m} = 1.1$, $\tilde{\ell} = 1.1$.

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. [43]), 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 [19, 38].

Acknowledgements

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

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

[2] I. Attias, A. Kontorovich, and Y. Mansour. Improved generalization bounds for robust learning. In *Algorithmic Learning Theory*, 2019.

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

[4] N. M. Boffi, S. Tu, N. Matni, J.-J. E. Slotine, and V. Sindhwani. Learning stability certificates from data. In *Conference on Robot Learning*, 2020.

[5] N. M. Boffi, S. Tu, and J.-J. E. Slotine. Regret bounds for adaptive nonlinear control. In *Learning for Dynamics and Control*, 2020.

[6] Y.-C. Chang, N. Roohi, and S. Gao. Neural lyapunov control. In *Neural Information Processing Systems*, 2019.

[7] S. Chen, M. Fazlyab, M. Morari, G. J. Pappas, and V. M. Preciado. Learning lyapunov functions for piecewise affine systems with neural network controllers. *arXiv preprint arXiv:2008.06546*, 2020.

[8] J. Fu, K. Luo, and S. Levine. Learning robust rewards with adversarial inverse reinforcement learning. In *International Conference on Learning Representations*, 2018.

[9] P. Giesl, B. Hamzi, M. Rasmussen, and K. Webster. Approximation of lyapunov functions from noisy data. *Journal of Computational Dynamics*, 7(1):57–81, 2020.

[10] A. Gleave, M. Dennis, C. Wild, N. Kant, S. Levine, and S. Russell. Adversarial policies: Attacking deep reinforcement learning. In *International Conference on Learning Representations*, 2020.

[11] J. Ho and S. Ermon. Generative adversarial imitation learning. In *Neural Information Processing Systems*, 2016.

[12] W. Jin, Z. Wang, Z. Yang, and S. Mou. Neural certificates for safe control policies. *arXiv preprint arXiv:2006.08465*, 2020.

[13] J. Kenanian, A. Balkan, R. M. Jungers, and P. Tabuada. Data driven stability analysis of black-box switched linear systems. *Automatica*, 109:108533, 2019.

[14] B. E. Khadir, J. Varley, and V. Sindhwani. Teleoperator imitation with continuous-time safety. In *Robotics: Science and Systems*, 2019.

[15] D. P. Kingma and J. Ba. Adam: A method for stochastic optimization. In *International Conference on Learning Representations*, 2015.

[16] A. Kurakin, I. Goodfellow, and S. Bengio. Adversarial examples in the physical world. *arXiv preprint arXiv:1607.02533*, 2016.

[17] B. D. Lee, T. T. C. K. Zhang, H. Hassani, and N. Matni. Adversarial tradeoffs in linear inverse problems and robust state estimation. *arXiv preprint arXiv:2111.08864*, 2021.

[18] Y. Lin, E. Sontag, and Y. Wang. Various results concerning set input-to-state stability. In *1995 34th IEEE Conference on Decision and Control*, 1995.

[19] L. Lindemann, A. Robey, L. Jiang, S. Tu, and N. Matni. Learning robust output control barrier functions from safe expert demonstrations. *arXiv preprint arXiv:2111.09971*, 2021.

[20] W. Lohmiller and J.-J. E. Slotine. On contraction analysis for non-linear systems. *Automatica*, 34(6): 683–696, 1998.

[21] M. Lutter, S. Mannor, J. Peters, D. Fox, and A. Garg. Robust value iteration for continuous control tasks. *arXiv preprint arXiv:2105.12189*, 2021.

[22] A. Madry, A. Makelov, L. Schmidt, D. Tsipras, and A. Vladu. Towards deep learning models resistant to adversarial attacks. In *International Conference on Learning Representations*, 2018.

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

[24] G. Manek and J. Z. Kolter. Learning stable deep dynamics models. In *Neural Information Processing Systems*, 2019.

[25] O. Montasser, S. Hanneke, and N. Srebro. Vc classes are adversarially robustly learnable, but only improperly. In *Conference on Learning Theory*, 2019.

[26] L. Pinto, J. Davidson, R. Sukthankar, and A. Gupta. Robust adversarial reinforcement learning. In *International Conference on Machine Learning*, 2017.

[27] H. Ravanbakhsh and S. Sankaranarayanan. Learning control lyapunov functions from counterexamples and demonstrations. *Autonomous Robots*, 43:275–307, 2019.

[28] S. M. Richards, F. Berkenkamp, and A. Krause. The lyapunov neural network: Adaptive stability certification for safe learning of dynamical systems. In *Conference on Robot Learning*, 2018.

[29] A. Robey, H. Hu, L. Lindemann, H. Zhang, D. V. Dimarogonas, S. Tu, and N. Matni. Learning control barrier functions from expert demonstrations. In *2020 59th IEEE Conference on Decision and Control*, 2020.

[30] L. Schmidt, S. Santurkar, D. Tsipras, K. Talwar, and A. Madry. Adversarially robust generalization requires more data. In *Neural Information Processing Systems*, 2018.

[31] V. Sindhwani, S. Tu, and S. M. Khansari-Zadeh. Learning contracting vector fields for stable imitation learning. *arXiv preprint arXiv:1804.04878*, 2018.

[32] S. Singh, A. Majumdar, J.-J. E. Slotine, and M. Pavone. Robust online motion planning via contraction theory and convex optimization. In *2017 IEEE International Conference on Robotics and Automation*, 2017.

[33] S. Singh, S. M. Richards, J.-J. E. Slotine, V. Sindhwani, and M. Pavone. Learning stabilizable nonlinear dynamics with contraction-based regularization. *International Journal of Robotics Research*, 40(10–11):1123–1150, 2020.

[34] E. Sontag. *Mathematical control theory: deterministic finite dimensional systems*, volume 6. Springer, 2013.

[35] N. Srebro, K. Sridharan, and A. Tewari. Smoothness, low-noise and fast rates. In *Neural Information Processing Systems*, 2010.

[36] C. Szegedy, W. Zaremba, I. Sutskever, J. Bruna, D. Erhan, I. Goodfellow, and R. Fergus. Intriguing properties of neural networks. *arXiv preprint arXiv:1312.6199*, 2013.

[37] A. J. Taylor, A. Singletary, Y. Yue, and A. D. Ames. Learning for safety-critical control with control barrier functions. *arXiv preprint arXiv:1912.10099*, 2019.

[38] A. J. Taylor, V. D. Dorobantu, S. Dean, B. Recht, Y. Yue, and A. D. Ames. Towards robust data-driven control synthesis for nonlinear systems with actuation uncertainty. *arXiv preprint arXiv:2011.10730*, 2021.

[39] J. Tobin, R. Fong, A. Ray, J. Schneider, W. Zaremba, and P. 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.

[40] F. Torabi, G. Warnell, and P. Stone. Generative adversarial imitation from observation. *arXiv preprint arXiv:1807.06158*, 2019.

[41] A. Tsiamis and G. J. Pappas. Linear systems can be hard to learn. *arXiv preprint arXiv:2104.01120*, 2021.

[42] A. Tsiamis, N. Matni, and G. J. Pappas. Sample complexity of kalman filtering for unknown systems. In *Learning for Dynamics and Control*, 2020.

[43] S. Tu, A. Robey, T. Zhang, and N. Matni. On the sample complexity of stability constrained imitation learning. *arXiv preprint arXiv:2102.09161*, 2021.

[44] M. J. Wainwright. *High-Dimensional Statistics: A Non-Asymptotic Viewpoint*. Cambridge University Press, 2019.

[45] D. Yin, R. Kannan, and P. Bartlett. Rademacher complexity for adversarially robust generalization. In *International Conference on Machine Learning*, 2019.

[46] H. Zhang, Y. Yu, J. Jiao, E. P. Xing, L. E. Ghaoui, and M. I. Jordan. Theoretically principled trade-off between robustness and accuracy. In *International Conference on Machine Learning*, 2019.

A Proofs for Section 3

A.1 Proof of Lemma 2

The result follows by observing

$$\begin{aligned}
\tilde{h}(\xi, V_1) - \tilde{h}(\xi, V_2) &:= \sup_{\tilde{\varphi}\in\Delta_\varepsilon} \sup_{t\in\mathcal{T}} \langle \nabla V_1(\tilde{\varphi}_t(\xi)), f(\tilde{\varphi}_t(\xi)) + \delta_t\rangle + \eta V_1(\tilde{\varphi}_t(\xi)) - \nu \\
&\quad - \sup_{\tilde{\varphi}\in\Delta_\varepsilon} \left(\langle \nabla V_2(\tilde{\varphi}_t(\xi)), f(\tilde{\varphi}_t(\xi)) + \delta_t\rangle + \eta V_2(\tilde{\varphi}_t(\xi)) - \nu\right) \\
&\leqslant \sup_{\tilde{\varphi}\in\Delta_\varepsilon} \sup_{t\in\mathcal{T}} \langle \nabla V_1(\tilde{\varphi}_t(\xi)), f(\tilde{\varphi}_t(\xi)) + \delta_t\rangle + \eta V_1(\tilde{\varphi}_t(\xi)) \\
&\quad - \left(\langle \nabla V_2(\tilde{\varphi}_t(\xi)), f(\tilde{\varphi}_t(\xi)) + \delta_t\rangle + \eta V_2(\tilde{\varphi}_t(\xi))\right) \\
&\leqslant L_h \left\|V_1 - V_2\right\|_{\mathcal{V}} + \langle \nabla V_1(\tilde{\varphi}_t(\xi)) - \nabla V_2(\tilde{\varphi}_t(\xi)), \delta_t\rangle \\
&\leqslant L_h \left\|V_1 - V_2\right\|_{\mathcal{V}} + \left\|\nabla V_1(\tilde{\varphi}_t(\xi)) - \nabla V_2(\tilde{\varphi}_t(\xi))\right\|_2 \left\|\delta_t\right\|_2 \\
&\leqslant (L_h + B_\delta) \left\|V_1 - V_2\right\|_{\mathcal{V}}.
\end{aligned}$$

Swapping the roles of V_1 and V_2 completes the proof.

A.2 Proof of Lemma 3

We first state a few definitions. Let $\mathsf{Sym}^{n\times n}_{\succcurlyeq 0}$ denote the space of $n \times n$ real-valued positive semi-definite matrices. Given a Riemannian metric $M : \mathbb{R}^n \to \mathsf{Sym}^{n\times n}_{\succcurlyeq 0}$, the geodesic distance associated with M is:

$$d_M(x, y) := \inf_{\gamma\in\Gamma(x,y)} \int_0^1 \sqrt{\gamma'(s)^\top M(\gamma(s))\gamma'(s)} ds,$$

where $\Gamma(x, y)$ denotes the set of smooth curves γ with endpoints at $\gamma(0) = x$ and $\gamma(1) = y$.

Now, given a time-varying metric $M : \mathbb{R}^n \times \mathbb{R} \to \mathsf{Sym}^{n\times n}_{\succcurlyeq 0}$, a function $f(x, t)$ is said to be contracting in the metric $M(x, t)$ at rate λ if for all x and t:

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

Proposition 1. *Consider two nonlinear systems*

$$\begin{aligned}
\dot{x}_p &= f(x_p, t) + d(x_p, t), \\
\dot{x} &= f(x, t),
\end{aligned}$$

where $f(x, t)$ is contracting in the metric $M(x, t)$. Then the geodesic distance $d_{M(\cdot,t)}(x_p(t), x(t))$ satisfies the differential inequality

$$\frac{d}{dt} d_{M(\cdot,t)}(x_p(t), x(t)) \leqslant -\lambda d_{M(\cdot,t)}(x_p(t), x(t)) + \left\|\Theta(x_p(t), t)d(x_p(t), t)\right\|_2,$$

where $M(x, t) = \Theta(x, t)^\top \Theta(x, t)$.

Proof. Consider a geodesic $\gamma_t(s) : [0, 1] \to \mathbb{R}^n$ from $x_p(t)$ to $x(t)$ so that $\gamma_t(0) = x_p(t)$ and $\gamma_t(1) = x(t)$, and let γ_t' denote the derivative of γ_t with respect to its argument. Observe that the Riemannian energy is $E(\gamma_t) = d_M(x_p(t), x(t))^2$, and hence $\dot{E}(\gamma_t) = 2d_M(x_p(t), x(t))\left(\frac{d}{dt} d_M(x_p(t), x(t))\right)$. From the formula

for the first variation of the Riemannian energy for a minimizing geodesic,

$$\dot{E}(\gamma_t) = 2\left\langle \gamma_t'(s), \dot{\gamma}_t(s) \right\rangle |_{s=0}^{s=1} + 2\frac{\partial E}{\partial t},$$

where $\dot{\gamma}_t$ denotes the time derivative of γ_t along the flow of $x_p(t)$, and $\langle \cdot, \cdot \rangle$ denotes the Riemannian inner product. Note that while the functional form of $\gamma_t(s)$ is unknown, the values of its time derivative at the endpoints are fixed to be $\dot{x}_p$ and $\dot{x}$ for $s = 0$ and $s = 1$, respectively. From this, we find that

$$\begin{aligned}\dot{E}(\gamma_t) &= 2\left\langle \gamma_t'(1), f(x) \right\rangle - 2\left\langle \gamma_t'(0), f(x_p) + d(x_p, t) \right\rangle + 2\frac{\partial E}{\partial t} \\ &\leqslant -2\lambda E(\gamma_t) - 2\left\langle \gamma_t'(0), d(x_p, t) \right\rangle,\end{aligned}$$

where the inequality stems from contraction of the nominal dynamics $f(x, t)$. This relation then implies the decrease condition

$$\begin{aligned}\frac{d}{dt} d_{M(\cdot,t)}(x_p(t), x(t)) &\leqslant -\lambda d_{M(\cdot,t)}(x_p(t), x(t)) - \frac{1}{d_{M(\cdot,t)}(x_p(t), x(t))} \gamma_t'(0)^\top \Theta(x_p)^\top \Theta(x_p) d(x_p, t), \\ &\leqslant -\lambda d_{M(\cdot,t)}(x_p(t), x(t)) + \frac{\|\Theta(x_p, t)\gamma_t'(0)\|_2}{d_{M(\cdot,t)}(x_p(t), x(t))} \|\Theta(x_p, t) d(x_p, t)\|_2.\end{aligned}$$

To complete the proof, observe that geodesics have constant energy, so that $\|\Theta(x_p, t)\gamma_t'(0)\|_2 = d_{M(\cdot,t)}(x_p(t), x(t))$. □

We can now prove Lemma 3. By Proposition 1 and the assumption that $M(x, t) \preccurlyeq LI$,

$$\begin{aligned}\frac{d}{dt} d_{M(\cdot,t)}(x_p(t), x(t)) &\leqslant -\lambda d_{M(\cdot,t)}(x_p(t), x(t)) + \|\Theta(x(t), t) d(x(t), t)\|_2 \\ &\leqslant -\lambda d_{M(\cdot,t)}(x_p(t), x(t)) + \sqrt{L} \|d(x_p(t), t)\|_2.\end{aligned}$$

By the comparison lemma,

$$d_{M(\cdot,t)}(x_p(t), x(t)) \leqslant d_{M(\cdot,0)}(x_p(0), x(0)) e^{-\lambda t} + \sqrt{L} \int_0^t e^{-\lambda(t-s)} \|d(x_p(s), s)\|_2 \, ds.$$

Next, by the assumption that $\mu I \preccurlyeq M(x, t) \preccurlyeq LI$, it is not hard to see (see e.g. Proposition D.2 of Boffi et al. [5]) that for all x, y, t,

$$\sqrt{\mu} \|x - y\|_2 \leqslant d_{M(\cdot,t)}(x, y) \leqslant \sqrt{L} \|x - y\|_2$$

Combining these inequalities, we have:

$$\sqrt{\mu} \|x_p(t) - x(t)\|_2 \leqslant \sqrt{L} \|x_p(0) - x(0)\|_2 e^{-\lambda t} + \sqrt{L} \int_0^t e^{-\lambda(t-s)} \|d(x_p(s), s)\|_2 \, ds.$$

The claim now follows by dividing both sides by $\sqrt{\mu}$.

A.3 Proof of Theorem 1

We observe that for an arbitrary perturbation tube $\Delta(\xi)$

$$\begin{aligned}
\tilde{h}(\xi, V) - h(\xi, V) = & \sup_{\tilde{\varphi} \in \Delta(\xi)} \sup_{t \in \mathcal{T}} \langle \nabla V(\tilde{\varphi}_t(\xi)), f(\tilde{\varphi}_t(\xi)) + \delta_t \rangle + \eta V(\tilde{\varphi}_t(\xi)) - \nu \\
& - \sup_{t \in \mathcal{T}} (\langle \nabla V(\varphi_t(\xi)), f(\varphi_t(\xi)) \rangle + \eta V(\varphi_t(\xi))) \\
\leqslant & \sup_{\tilde{\varphi} \in \Delta(\xi)} \sup_{t \in \mathcal{T}} \langle \nabla V(\tilde{\varphi}_t(\xi)), f(\tilde{\varphi}_t(\xi)) + \delta_t \rangle + \eta V(\tilde{\varphi}_t(\xi)) - \nu \\
& - (\langle \nabla V(\varphi_t(\xi)), f(\varphi_t(\xi)) \rangle + \eta V(\varphi_t(\xi))) \\
\leqslant & \sup_{\tilde{\varphi} \in \Delta(\xi)} \sup_{t \in \mathcal{T}} (L_{\nabla V} + \eta L_V) \left\| \tilde{\varphi}_t(\xi) - \varphi_t(\xi) \right\|_2 + B_{\nabla V} \left\| \delta_t \right\|_2 - \nu,
\end{aligned}$$

and thus

$$\left| \tilde{h}(\xi, V) - h(\xi, V) \right| \leqslant \sup_{\tilde{\varphi} \in \Delta(\xi)} \sup_{t \in \mathcal{T}} (L_{\nabla V} + \eta L_V) \left\| \tilde{\varphi}_t(\xi) - \varphi_t(\xi) \right\|_2 + B_{\nabla V} \left\| \delta_t \right\|_2 + \nu \tag{19}$$

- **Norm-bounded Adversary** Δ_ε^u: from EδISS, we know that the deviation can be bounded by:

$$\begin{aligned}
\left\| \varphi_t - \tilde{\varphi}_t \right\|_2 &\leqslant \beta \left\| \varphi_0 - \tilde{\varphi}_0 \right\| e^{-\rho t} + \gamma \int_0^t e^{-\rho(t-s)} \left\| \delta_s \right\|_2 \, ds \\
&\leqslant \gamma \varepsilon \int_0^t e^{-\rho(t-s)} \, ds \\
&\leqslant \gamma \varepsilon \rho^{-1}.
\end{aligned}$$

  Plugging this back into (19), we get

$$\left| \tilde{h}(\xi, V) - h(\xi, V) \right| \leqslant (L_{\nabla V} + \eta L_V) \gamma \varepsilon \rho^{-1} + B_{\nabla V} \varepsilon + \nu.$$

  Applying (12) yields the desired result.

- **Lipschitz Adversary** Δ_ε^x: from EδISS, we can bound the deviation by:

$$\begin{aligned}
\left\| \varphi_t - \tilde{\varphi}_t \right\|_2 &\leqslant \beta \left\| \varphi_0 - \tilde{\varphi}_0 \right\| e^{-\rho t} + \gamma \int_0^t e^{-\rho(t-s)} \left\| \delta(\tilde{\varphi}_s) \right\|_2 \, ds \\
&\leqslant \gamma \varepsilon \int_0^t e^{-\rho(t-s)} \left\| \tilde{\varphi}_s \right\|_2 \, ds \\
&\leqslant \gamma \varepsilon \int_0^t e^{-\rho(t-s)} (\left\| \varphi_s - \tilde{\varphi}_s \right\|_2 + \left\| \varphi_s \right\|_2) \, ds \\
&\leqslant \gamma \varepsilon \int_0^t e^{-\rho(t-s)} \left\| \varphi_s - \tilde{\varphi}_s \right\|_2 \, ds + \gamma \varepsilon \int_0^t e^{-\rho(t-s)} \beta \left\| \xi \right\|_2 e^{-\rho s} \, ds \\
&= \gamma \varepsilon \int_0^t e^{-\rho(t-s)} \left\| \varphi_s - \tilde{\varphi}_s \right\|_2 \, ds + \gamma \varepsilon \beta \left\| \xi \right\|_2 t e^{-\rho t}.
\end{aligned}$$

Taking the supremum over t on both sides, we have

$$\begin{aligned}\sup_{t\in\mathcal{T}} \|\varphi_t - \tilde{\varphi}_t\|_2 &\leqslant \sup_{t\in\mathcal{T}} \gamma\varepsilon \int_0^t e^{-\rho(t-s)} \|\varphi_s - \tilde{\varphi}_s\|_2 \, ds + \gamma\varepsilon\beta \|\xi\|_2 te^{-\rho t} \\ &\leqslant \gamma\varepsilon \sup_{t\in\mathcal{T}} \int_0^t e^{-\rho(t-s)} \sup_{s\in\mathcal{T}} \|\varphi_s - \tilde{\varphi}_s\|_2 \, ds + \sup_{t\in\mathcal{T}} \gamma\varepsilon\beta \|\xi\|_2 te^{-\rho t} \\ &\leqslant \gamma\varepsilon\rho^{-1} \sup_{s\in\mathcal{T}} \|\varphi_s - \tilde{\varphi}_s\|_2 + \gamma\varepsilon\rho^{-1}\beta \|\xi\|_2 e^{-1},\end{aligned}$$

where the second term in the last line comes from optimizing $\max_t te^{-\rho t}$, which attains its maximum $\frac{1}{\rho e}$ at $t = 1/\rho$. Since by assumption, $\frac{\gamma\varepsilon}{\rho} < 1$, we have

$$\sup_{t\in\mathcal{T}} \|\varphi_t - \tilde{\varphi}_t\|_2 \leqslant \frac{\gamma\varepsilon\rho^{-1}}{1-\gamma\varepsilon\rho^{-1}} \beta e^{-1} \|\xi\|_2 .$$

Plugging this into (19), we get

$$\begin{aligned}\left|\tilde{h}(\xi, V) - h(\xi, V)\right| &\leqslant \sup_{\tilde{\varphi}\in\Delta_\varepsilon(\xi)} \sup_{t\in\mathcal{T}} (L_{\nabla V} + \eta L_V) \|\tilde{\varphi}_t(\xi) - \varphi_t(\xi)\|_2 + B_{\nabla V} \|\delta(\tilde{\varphi}_t(\xi))\|_2 + \nu \\ &\leqslant \sup_{\tilde{\varphi}\in\Delta_\varepsilon(\xi)} \sup_{t\in\mathcal{T}} (L_{\nabla V} + \eta L_V) \|\tilde{\varphi}_t - \varphi_t\|_2 + B_{\nabla V}\varepsilon(\|\varphi_t - \tilde{\varphi}_t\|_2 + \|\tilde{\varphi}_t\|_2) + \nu \\ &\leqslant (L_{\nabla V} + \eta L_V + B_{\nabla V}\varepsilon) \frac{\gamma\varepsilon\rho^{-1}}{1-\gamma\varepsilon\rho^{-1}} \beta e^{-1} \|\xi\|_2 + \sup_t B_{\nabla V}\beta\varepsilon \|\xi\|_2 e^{-\rho t} + \nu \\ &\leqslant \left[(L_{\nabla V} + \eta L_V + B_{\nabla V}\varepsilon) \frac{\gamma\varepsilon\rho^{-1}}{1-\gamma\varepsilon\rho^{-1}} e^{-1} + B_{\nabla V}\right] \beta\varepsilon \|\xi\|_2 + \nu.\end{aligned}$$

Applying (12) yields the desired result.

- **Combined Adversary** $\Delta^{x,u}_{\varepsilon_x,\varepsilon_u}$: proof follows similarly to the Lipschitz adversary case. Using EδISS, we have

$$\begin{aligned}\|\varphi_t - \tilde{\varphi}_t\|_2 &\leqslant \beta \|\varphi_0 - \tilde{\varphi}_0\| e^{-\rho t} + \gamma \int_0^t e^{-\rho(t-s)} \|\delta(\tilde{\varphi}_s) + \delta_t\|_2 \, ds \\ &\leqslant \gamma\varepsilon_x \int_0^t e^{-\rho(t-s)} \|\tilde{\varphi}_s\|_2 \, ds + \gamma\varepsilon_u\rho^{-1} \\ &= \gamma\varepsilon_x \int_0^t e^{-\rho(t-s)} \|\varphi_s - \tilde{\varphi}_s\|_2 \, ds + \gamma\varepsilon_x\beta \|\xi\|_2 te^{-\rho t} + \gamma\varepsilon_u\rho^{-1}.\end{aligned}$$

Since $\gamma\varepsilon_x < \rho$, we take the supremum of both sides and shift terms around to get

$$\sup_{t\in\mathcal{T}} \|\varphi_t - \tilde{\varphi}_t\|_2 \leqslant \frac{\gamma\varepsilon_u\rho^{-1} + \gamma\varepsilon_x\rho^{-1}\beta e^{-1} \|\xi\|_2}{1-\gamma\varepsilon_x\rho^{-1}},$$

and thus plugging into (19), we get

$$\begin{aligned}\left|\tilde{h}(\xi, V) - h(\xi, V)\right| \leqslant (L_{\nabla V} + \eta L_V + B_{\nabla V}\varepsilon_x) \frac{\gamma\varepsilon_u\rho^{-1} + \gamma\varepsilon_x\rho^{-1}e^{-1}\beta\varepsilon_x \|\xi\|_2}{1-\gamma\varepsilon_x\rho^{-1}} \\ + B_{\nabla V}\beta\varepsilon_x \|\xi\|_2 + B_{\nabla V}\varepsilon_u + \nu.\end{aligned}$$

Applying (12) yields the desired result.

B Adversarially Robust Certificates in Discrete Time

In the discrete time setting, we consider the system $x_{t+1} = f(x_t)$. Like in the continuous time case, f is continuous and unknown, the state $x \in \mathbb{R}^p$ is fully observed, the initial conditions ξ are drawn from some compact set $\mathcal{X}$, and $\varphi_t(\xi)$ denotes the map to the state at time t given initial condition ξ. Given a candidate Lyapunov function V, we define the nominal and adversarial (exponential) Lyapunov decrease conditions as:

$$h(\xi, V) = \max_{t \leqslant T} V(\varphi_{t+1}(\xi)) - \eta^2 V(\varphi_t(\xi))$$
$$\tilde{h}_\nu(\xi, V) = \max_{\tilde{\varphi} \in \Delta_\varepsilon(\xi)} \max_{t \leqslant T} V(\tilde{\varphi}_{t+1}(\xi)) - \eta^2 V(\tilde{\varphi}_t(\xi)) - \nu, \quad \nu \geqslant 0,$$

where $0 < \eta < 1$, as well as the corresponding loss classes $\mathcal{H}$ and $\tilde{\mathcal{H}}$. The stability certification problem can be posed as the following feasibility problem

$$\text{Find}_{V \in \mathcal{V}} \text{ s.t. } \tilde{h}_\nu(\xi, V) \leqslant -\tau, \quad i = 1, \ldots, n. \tag{20}$$

We make the following assumptions.

Assumption 3 (Discrete-time stability in the sense of Lyapunov)**.** *Fix a perturbation set $\Delta(\cdot)$. There exists a compact set $S \subseteq \mathbb{R}^p$ such that $\tilde{\varphi}_t(\xi) \in S$ for all $\xi \in \mathcal{X}$, $t \leqslant T$, and $\tilde{\varphi}_t(\xi) \in \Delta(\xi)$.*

Assumption 4 (Regularity of $\mathcal{V}$)**.** *There exist a constant L_V such that for every $V \in \mathcal{V}$, the map $x \to V(x)$ over $x \in S$ is L_V-Lipschitz.*

Under Assumptions 3 and 4, and the continuity of the dynamics $f(x)$, there exist constants B_V and $B_{\tilde{h}}$ such that

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

Finally, let $\|V\|_{\mathcal{V}} := \sup_{x \in S} |V(x)|$ denote the supremum norm on the space $\mathcal{V}$. Under Assumptions 3 and 4, and the discrete time definition of $\tilde{h}$, Lemma 1 holds in precisely the same form, such that we once again need only to bound the Rademacher complexity of the adversarial loss class. We now prove the discrete-time variant of the adversary-agnostic Rademacher complexity bound.

Lemma 4 (Discrete-time analogue of Lemma 2)**.** *Suppose that Assumptions 3 and 4 hold. 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, $L_{\tilde{h}} \leqslant L_h + 2$.*

Proof. Writing out $\tilde{h}(\xi, V_1) - \tilde{h}(\xi, V_2)$, for arbitrary ξ, V_1, V_2, we have:

$$\begin{aligned}\tilde{h}(\xi, V_1) - \tilde{h}(\xi, V_2) &= \max_{\tilde{\varphi} \in \Delta(\xi)} \max_{t \leqslant T} V_1(\tilde{\varphi}_{t+1}) - \eta^2 V_1(\tilde{\varphi}_t) - \max_{\tilde{\varphi} \in \Delta(\xi)} \max_{t \leqslant T} \big(V_2(\tilde{\varphi}_{t+1}) - \eta^2 V_2(\tilde{\varphi}_t)\big) \\ &\leqslant \max_{\tilde{\varphi} \in \Delta(\xi)} \max_{t \leqslant T} V_1(\tilde{\varphi}_{t+1}) - \eta^2 V_1(\tilde{\varphi}_t) - \big(V_2(\tilde{\varphi}_{t+1}) - \eta^2 V_2(\tilde{\varphi}_t)\big).\end{aligned}$$

For any time t, we have

$$\begin{aligned} V_1(\tilde{\varphi}_{t+1}) - \eta^2 V_1(\tilde{\varphi}_t) - \big(V_2(\tilde{\varphi}_{t+1}) - \eta^2 V_2(\tilde{\varphi}_t)\big) &= V_1(f(\tilde{\varphi}_t)) - \eta^2 V_1(\tilde{\varphi}_t) - \big(V_2(f(\tilde{\varphi}_t)) - \eta^2 V_2(\tilde{\varphi}_t)\big) \\ &\quad + (V_1(\tilde{\varphi}_{t+1}) - V_2(\tilde{\varphi}_{t+1})) \\ &\quad - (V_1(f(\tilde{\varphi}_t)) - V_2(f(\tilde{\varphi}_t))) \\ &\leqslant |h(\xi, V_1) - h(\xi, V_2)| + 2 \sup_{x \in S} |V_1(x) - V_2(x)| \\ &\leqslant L_h \left\| V_1 - V_2 \right\|_{\mathcal{V}} + 2 \left\| V_1 - V_2 \right\|_{\mathcal{V}}, \end{aligned}$$

where added and subtracted $V_1(f(\tilde{\varphi}_t))$ and $V_2(f(\tilde{\varphi}_t))$, and used Assumption 1 to bound the leftover terms using the definition of $\|\cdot\|_{\mathcal{V}}$. The argument is symmetric for V_1, V_2, so we have

$$\left| \tilde{h}(\xi, V_1) - \tilde{h}(\xi, V_2) \right| \leqslant (L_h + 2) \left\| V_1 - V_2 \right\|_{\mathcal{V}},$$

for all $\xi \in X$ and $V_1, V_2 \in \mathcal{V}$. □

We note this is at first glance a better bound than the continuous-time version. This can be attributed to the fact that Assumption 3 is at face value more restrictive than its continuous-time analogue Assumption 1, since it implicitly enforces a norm-constraint on δ_t such that it cannot push x_t out of the compact set S, which is a property independent of the certificate V. On the other hand, Assumption 1 does not immediately enforce a norm-constraint on δ_t–the implicit constraint depends on the choice of V, where δ_t cannot render optimization problem (9) infeasible. As Assumption 3 is in a sense more restrictive than Assumption 1, the bound we get is stronger.

We now re-introduce the norm-bounded, Lipschitz, and combined adversarial trajectory tubes in discrete time:

$$\Delta_\varepsilon^u(\xi) := \{\tilde{\varphi} : \tilde{\varphi}_{t+1} = f(\tilde{\varphi}_t) + \delta_t,\ \tilde{\varphi}_0 = \xi,\ \|\delta_t\|_2 \leqslant \varepsilon\} \tag{21}$$

$$\Delta_\varepsilon^x(\xi) := \{\tilde{\varphi} : \tilde{\varphi}_{t+1} = f(\tilde{\varphi}_t) + \delta(\tilde{\varphi}_t),\ \tilde{\varphi}_0 = \xi,\ \|\delta(\tilde{\varphi}_t)\|_2 \leqslant \varepsilon \|\tilde{\varphi}_t\|_2\} \tag{22}$$

$$\Delta_{\varepsilon_x, \varepsilon_u}^{x,u}(\xi) := \{\tilde{\varphi} : \tilde{\varphi}_{t+1} = f(\tilde{\varphi}_t) + \delta^x(\tilde{\varphi}_t) + \delta_t^u,\ \tilde{\varphi}_0 = \xi,\ \|\delta^x(\tilde{\varphi}_t)\|_2 \leqslant \varepsilon_x \|\tilde{\varphi}_t\|_2,\ \|\delta_t^u\|_2 \leqslant \varepsilon_u\}. \tag{23}$$

Accordingly, we introduce (β, ρ, γ)-EδISS in discrete time.

Definition 2 (Discrete-time (β, ρ, γ)-EδISS)**.** *Let $\beta, \gamma > 0$ be positive constants and $\rho \in (0, 1)$. A discrete-time dynamical system $f(x, t)$ is (β, ρ, γ)-exponential-incrementally-input-to-state stable if for every pair of initial conditions (x_0, y_0) and signal u_t (which can depend causally on x, y), the trajectories $x_{t+1} = f(x_t, t)$ and $y_{t+1} = f(y_t, t) + u_t$ satisfy for all $t \geqslant 0$:*

$$\|x_t - y_t\|_2 \leqslant \beta \rho^t \|x_0 - y_0\|_2 + \gamma \sum_{k=0}^{t-1} \rho^{t-1-k} \|u_k\|_2. \tag{24}$$

As mentioned earlier in this paper, Proposition 5.3 of [5] gives that a discrete-time system contracting with respect to some metric is (β, ρ, γ)-EδISS. We are now ready to provide the discrete-time analogues to the adversarial Rademacher complexities provided in Theorem 1.

Theorem 2 (Discrete-time analogue to Theorem 1)**.** *Put $B_X := \sup_{\xi \in \mathcal{X}} \|\xi\|_2$, let Assumption 4 hold, and assume that the nominal discrete-time system $f(x)$ is (β, ρ, γ)-EδISS. Then for*

- *adversarial trajectories drawn from the norm-bounded tube $\Delta_\varepsilon^u(\xi)$ defined in (21), Assumption 3 holds and*

$$\mathcal{R}_n(\tilde{\mathcal{H}}) \leqslant \mathcal{R}_n(\mathcal{H}) + \left[(1 + \eta^2) L_V \frac{\gamma \varepsilon}{1 - \rho} + \nu \right] \frac{1}{\sqrt{n}}, \tag{25}$$

- *adversarial trajectories drawn from the Lipschitz tube $\Delta_\varepsilon^x(\xi)$ defined in (22), if $\varepsilon > 0$ is small enough such that $\rho + \gamma\varepsilon < 1$, then Assumption 3 holds and*

$$\mathcal{R}_n(\tilde{\mathcal{H}}) \leqslant \mathcal{R}_n(\mathcal{H}) + \left[L_V \beta B_X(\rho + \gamma\varepsilon + \eta^2) + \nu\right] \frac{1}{\sqrt{n}}, \tag{26}$$

- *adversarial trajectories drawn from the combined tube $\Delta_{\varepsilon_x,\varepsilon_u}^{x,u}$ defined in (23), if $\varepsilon_x > 0$ is small enough such that $\rho + \gamma\varepsilon_x < 1$, then Assumption 3 holds and*

$$\mathcal{R}_n(\tilde{\mathcal{H}}) \leqslant \mathcal{R}_n(\mathcal{H}) + \left[(1+\eta^2)L_V\left(\frac{1-\rho}{1-(\rho+\gamma\varepsilon_x)}\frac{\gamma\beta B_X \varepsilon_x}{e\rho\log(\rho^{-1})} + \frac{\gamma\varepsilon_u}{1-\rho}\right) + \nu\right]\frac{1}{\sqrt{n}}. \tag{27}$$

We note that the necessary condition that $\rho + \gamma\varepsilon < 1$ for the Lipschitz and combined adversaries is nicely analogous to the continuous-time case where we needed $\gamma\varepsilon < \rho$; in both cases, the adversary cannot be powerful enough to de-stabilize the system. One can consider the scalar system $x_{t+1} = \rho x_t$, $0 < \rho < 1$ and the adversary $\delta(x) = \varepsilon x$ to see why this condition cannot be loosened in general. We now provide the proof to Theorem 2.

Proof. We observe that for an arbitrary perturbation tube $\Delta(\xi)$

$$\begin{aligned}
\tilde{h}(\xi, V) - h(\xi, V) &= \sup_{\tilde{\varphi}\in\Delta(\xi)} \sup_{t\leqslant T} V(\tilde{\varphi}_{t+1}) - \eta^2 V(\tilde{\varphi}_t) - \nu - \sup_{t\leqslant T}\big(V(\varphi_{t+1}) - \eta^2 V(\varphi_t)\big) \\
&\leqslant \sup_{\tilde{\varphi}\in\Delta(\xi)} \sup_{t\leqslant T} V(\tilde{\varphi}_{t+1}) - V(\varphi_{t+1}) - \eta^2(V(\tilde{\varphi}_t) - V(\varphi_t)) - \nu \\
&\leqslant \sup_{\tilde{\varphi}\in\Delta(\xi)} \sup_{t\leqslant T} (1+\eta^2)L_V \left\|\tilde{\varphi}_t - \varphi_t\right\|_2 - \nu,
\end{aligned}$$

and thus

$$\left|\tilde{h}(\xi, V) - h(\xi, V)\right| \leqslant \sup_{\tilde{\varphi}\in\Delta(\xi)} \sup_{t\leqslant T} (1+\eta^2)L_V \left\|\tilde{\varphi}_t - \varphi_t\right\|_2 + \nu. \tag{28}$$

Thus, it suffices to establish bounds on $\|\tilde{\varphi}_t - \varphi_t\|_2$.

- **Norm-bounded Adversary** Δ_ε^u: from EδISS, we know that the deviation can be bounded by:

$$\begin{aligned}
\left\|\tilde{\varphi}_t - \varphi_t\right\|_2 &\leqslant \beta\rho^t \left\|\tilde{\varphi}_0 - \varphi_0\right\|_2 + \gamma \sum_{k=0}^{t-1} \rho^{t-1-k} \left\|\delta_k\right\|_2 \\
&\leqslant \gamma \sum_{k=0}^{t-1} \rho^{t-1-k}\varepsilon \\
&\leqslant \frac{\gamma\varepsilon}{1-\rho}.
\end{aligned}$$

  Plugging this back into (28), we get

$$\left|\tilde{h}(\xi, V) - h(\xi, V)\right| \leqslant (1+\eta^2)L_V \frac{\gamma\varepsilon}{1-\rho} + \nu.$$

- **Lipschitz Adversary** Δ_ε^x: using a more careful analysis, we can get a finer bound in the discrete-time setting than the continuous-time setting, where we get an explicit time-dependent upper bound on the

deviation $\|\tilde{\varphi}_t - \varphi_t\|_2$. From E$\delta$ISS, observe that for any $k \leqslant T$,

$$\begin{aligned}
\|\tilde{\varphi}_k - \varphi_k\|_2 &\leqslant \gamma \sum_{i=0}^{k-1} \rho^{k-1-i} \|\delta(\tilde{\varphi}_i)\|_2 \\
&\leqslant \gamma\varepsilon \sum_{i=0}^{k-1} \rho^{k-1-i} \|\tilde{\varphi}_i\|_2 \\
&\leqslant \gamma\varepsilon \sum_{i=0}^{k-1} \rho^{k-1-i} (\|\tilde{\varphi}_i - \varphi_i\|_2 + \|\varphi_i\|_2) \\
&\leqslant \gamma\varepsilon \sum_{i=0}^{k-1} \rho^{k-1-i} \|\varphi_i\|_2 + \gamma\varepsilon \sum_{i=0}^{k-1} \rho^{k-1-i} \|\tilde{\varphi}_i - \varphi_i\|_2 \\
&\leqslant \gamma\varepsilon \sum_{i=0}^{k-1} \rho^{k-1} \beta \|\xi\|_2 + \gamma\varepsilon \sum_{i=1}^{k-1} \rho^{k-1-i} \|\tilde{\varphi}_i - \varphi_i\|_2 . \qquad (29)
\end{aligned}$$

We keep the sum in the first term to keep the our later algebraic manipulations clear. We also observe we can move the starting index of the second sum from 0 to 1, since $\tilde{\varphi}_0 - \varphi_0 = 0$. Now fixing any timestep $t \leqslant T$ for $t \geqslant 2$, we apply (29) recursively:

$$\begin{aligned}
\|\tilde{\varphi}_t - \varphi_t\|_2 &\leqslant \gamma\varepsilon \sum_{k_1=0}^{t-1} \rho^{t-1} \beta \|\xi\|_2 + \gamma\varepsilon \sum_{k_1=1}^{t-1} \rho^{t-1-k_1} \|\tilde{\varphi}_{k_1} - \varphi_{k_1}\|_2 \\
&\leqslant \gamma\varepsilon \sum_{k_1=0}^{t-1} \rho^{t-1} \beta \|\xi\|_2 \\
&\quad + \gamma\varepsilon \sum_{k_1=1}^{t-1} \rho^{t-1-k_1} \left(\gamma\varepsilon \sum_{k_2=0}^{k_1-1} \rho^{k_1-1} \beta \|\xi\|_2 + \gamma\varepsilon \sum_{k_2=1}^{k_1-1} \rho^{k_1-k_2-1} \|\tilde{\varphi}_{k_2} - \varphi_{k_2}\|_2 \right) \\
&\leqslant \gamma\varepsilon \sum_{k_1=0}^{t-1} \rho^{t-1} \beta \|\xi\|_2 + (\gamma\varepsilon)^2 \sum_{k_1=1}^{t-1} \sum_{k_2=0}^{k_1-1} \rho^{t-2} \beta \|\xi\|_2 \\
&\quad + (\gamma\varepsilon)^2 \sum_{k_1=1}^{t-1} \sum_{k_2=1}^{k_1-1} \rho^{t-2-k_2} \|\tilde{\varphi}_{k_2} - \varphi_{k_2}\|_2 \\
&\leqslant \gamma\varepsilon \sum_{k_1=0}^{t-1} \rho^{t-1} \beta \|\xi\|_2 + (\gamma\varepsilon)^2 \sum_{k_1=1}^{t-1} \sum_{k_2=0}^{k_1-1} \rho^{t-2} \beta \|\xi\|_2 + \cdots \\
&\quad + (\gamma\varepsilon)^j \sum_{k_1=1}^{t-1} \sum_{k_2=1}^{k_1-1} \cdots \sum_{k_j=0}^{k_{j-1}-1} \rho^{t-j} \beta \|\xi\|_2 \\
&\quad + (\gamma\varepsilon)^j \sum_{k_1=1}^{t-1} \sum_{k_2=1}^{k_1-1} \cdots \sum_{k_j=1}^{k_{j-1}-1} \rho^{t-j-k_j} \left\|\tilde{\varphi}_{k_j} - \varphi_{k_j}\right\|_2 .
\end{aligned}$$

This recursive process terminates when there does not exist an assignment of indices $k_1, \ldots, k_{j-1}$ such that the summand $\sum_{k_j=1}^{k_j-1}$ is non-empty, i.e. $1 = k_j > k_{j-1} - 1$. The largest j such that the aforementioned summand is possibly non-empty is when $k_i = k_{i-1} - 1$ for all $i < j$ and $k_1 = t - 1$, which implies $k_i = t - i$. In order for $k_j \geqslant 1$, we must have $k_{j-1} - 1 = t - j \geqslant 1$, i.e. $j \leqslant t - 1$,

and thus our recursive expansion terminates when $j = t - 1$. Therefore, continuing our earlier series of inequalities, we have

$$\begin{aligned}
\|\tilde{\varphi}_t - \varphi_t\|_2 &\leqslant \gamma\varepsilon \sum_{k_1=0}^{t-1} \rho^{t-1}\beta \|\xi\|_2 + \cdots + (\gamma\varepsilon)^{t-1} \sum_{k_1=1}^{t-1} \sum_{k_2=1}^{k_1-1} \cdots \sum_{k_{t-1}=0}^{k_{t-2}-1} \rho^{1-k_{t-1}}\beta \|\xi\|_2 \\
&\quad + (\gamma\varepsilon)^{t-1} \sum_{k_1=1}^{t-1} \sum_{k_2=1}^{k_1-1} \cdots \sum_{k_{t-1}=1}^{k_{t-2}-1} \rho^{t-(t-1)-k_{t-1}} \left\|\tilde{\varphi}_{k_{t-1}} - \varphi_{k_{t-1}}\right\|_2 \\
&= \gamma\varepsilon \sum_{k_1=0}^{t-1} \rho^{t-1}\beta \|\xi\|_2 + \cdots \\
&\quad + (\gamma\varepsilon)^{t-1} \sum_{k_1=1}^{t-1} \sum_{k_2=1}^{k_1-1} \cdots \sum_{k_{t-1}=0}^{k_{t-2}-1} \rho^{1-k_{t-1}}\beta \|\xi\|_2 + (\gamma\varepsilon)^{t-1} \|\tilde{\varphi}_1 - \varphi_1\|_2 \\
&\leqslant \gamma\varepsilon \sum_{k_1=0}^{t-1} \rho^{t-1}\beta \|\xi\|_2 + \cdots \\
&\quad + (\gamma\varepsilon)^{t-1} \sum_{k_1=1}^{t-1} \sum_{k_2=1}^{k_1-1} \cdots \sum_{k_{t-1}=0}^{k_{t-2}-1} \rho^{1-k_{t-1}}\beta \|\xi\|_2 + (\gamma\varepsilon)^{t-1}\gamma\varepsilon\rho\beta \|\xi\|_2 \\
&\leqslant \beta \|\xi\|_2 \sum_{j=1}^{t-1} (\gamma\varepsilon)^j \rho^{t-j} \left(\sum_{k_1=1}^{t-1} \cdots \sum_{k_{j-1}=1}^{k_{j-2}-1} \sum_{k_j=0}^{k_{j-1}-1} 1 \right) + \beta \|\xi\|_2 (\gamma\varepsilon)^t \rho \\
&\leqslant \beta \|\xi\|_2 \sum_{j=1}^{t-1} (\gamma\varepsilon)^j \rho^{t-j} \left(\sum_{k_1=1}^{t-1} \cdots \sum_{k_{j-1}=1}^{k_{j-2}-1} \sum_{k_j=0}^{k_{j-1}-1} 1 \right) + \beta \|\xi\|_2 (\gamma\varepsilon)^t + \beta \|\xi\|_2 \rho^t.
\end{aligned}$$

Now it remains to determine the value of $\sum_{k_1=1}^{t-1} \cdots \sum_{k_{j-1}=1}^{k_{j-2}-1} \sum_{k_j=0}^{k_{j-1}-1} 1$. Observe the sum is only non-empty if for each $1 \leqslant i \leqslant j$, $k_{i-1} - 1 - k_i \geqslant 0$, where we define $k_0 = t$. Let us define the variables $c_i = k_{i-1} - k_i \geqslant 1$ for $i = 1, \ldots, j$, and we define $c_{j+1} := k_j - 0 = k_j \geqslant 0$. The tuple $(c_1, \ldots, c_{j+1})$ thus satisfies $\sum_{i=1}^{j} c_i = \sum_{i=1}^{j+1} k_{i-1} - k_i = t - 0 = t$. Therefore, the number of terms in the nested summand is equal to the number of integer tuples $(c_1, \ldots, c_{j+1})$, where $c_1, \ldots, c_j$ are positive and $c_{j+1} \geqslant 0$, that sum up to t, which in turn can be transformed into a balls-and-bins problem where we have t total balls, and $j + 1$ bins, but with the first j bins already containing 1 ball. Thus applying the standard balls-and-bins formula for $t - j$ balls and $j + 1$ bins, we get

$$\sum_{k_1=1}^{t-1} \cdots \sum_{k_{j-1}=1}^{k_{j-2}-1} \sum_{k_j=0}^{k_{j-1}-1} 1 = \binom{(t-j) + (j+1) - 1}{(j+1) - 1} = \binom{t}{j}.$$

Plugging this back into the last line of our series of inequalities, we get

$$\begin{aligned}
\|\tilde{\varphi}_t - \varphi_t\|_2 &\leqslant \beta \|\xi\|_2 \sum_{j=1}^{t-1} (\gamma\varepsilon)^j \rho^{t-j} \left(\sum_{k_1=1}^{t-1} \cdots \sum_{k_{j-1}=1}^{k_{j-2}-1} \sum_{k_j=0}^{k_{j-1}-1} 1 \right) + \beta \|\xi\|_2 (\gamma\varepsilon)^t + \beta \|\xi\|_2 \rho^t \\
&= \beta \|\xi\|_2 \sum_{j=1}^{t-1} (\gamma\varepsilon)^j \rho^{t-j} \binom{t}{j} + \beta \|\xi\|_2 (\gamma\varepsilon)^t + \beta \|\xi\|_2 \rho^t \\
&= \beta \|\xi\|_2 \sum_{j=0}^{t} \binom{t}{j} (\gamma\varepsilon)^j \rho^{t-j} \\
&= \beta \|\xi\|_2 (\rho + \gamma\varepsilon)^t.
\end{aligned}$$

This gives us the bound:

$$\left| \tilde{h}(\xi, V) - h(\xi, V) \right| \leqslant L_V \beta B_X (\rho + \gamma\varepsilon + \eta^2) + \nu.$$

- **Combined Adversary** $\Delta^{x,u}_{\varepsilon_x,\varepsilon_u}$: from (β, ρ, γ)-EδISS, we have

$$\begin{aligned}
\|\tilde{\varphi}_t - \varphi_t\|_2 &\leqslant \gamma \sum_{k=0}^{t-1} \rho^{t-1-k} \|\delta^x(\tilde{\varphi}_k) + \delta^u_t\|_2 \\
&\leqslant \gamma \sum_{k=0}^{t-1} \rho^{t-1-k} \|\delta^x(\tilde{\varphi}_k)\|_2 + \gamma \sum_{k=0}^{t-1} \rho^{t-1-k} \|\delta^u_t\|_2 \\
&\leqslant \gamma \sum_{k=0}^{t-1} \rho^{t-1-k} \|\delta^x(\tilde{\varphi}_k)\|_2 + \gamma \frac{1}{1-\rho} \varepsilon_u \\
&\leqslant \gamma\varepsilon_x \sum_{k=0}^{t-1} \rho^{t-1-k} (\|\tilde{\varphi}_k - \varphi_k\|_2 + \|\varphi_k\|_2) + \gamma \frac{1}{1-\rho} \varepsilon_u \\
&\leqslant \gamma\varepsilon_x \Big(\max_t \|\tilde{\varphi}_t - \varphi_t\|_2 \Big) \frac{1}{1-\rho} + \gamma\varepsilon_x \beta t \rho^{t-1} \|\xi\|_2 + \gamma \frac{1}{1-\rho} \varepsilon_u.
\end{aligned}$$

Similarly taking a maximum with respect to t on both sides, we get

$$\max_t \|\tilde{\varphi}_t - \varphi_t\|_2 \leqslant \gamma\varepsilon_x \Big(\max_t \|\tilde{\varphi}_t - \varphi_t\|_2 \Big) \frac{1}{1-\rho} + \max_t \gamma\varepsilon_x \beta t \rho^{t-1} \|\xi\|_2 + \gamma \frac{1}{1-\rho} \varepsilon_u.$$

It is straightforward to compute $\max_t t\rho^{t-1} = \frac{1}{e\rho \log(\rho^{-1})}$. Thus, rearranging some terms, and assuming that $\rho + \gamma\varepsilon_x < 1$, we get

$$\max_t \|\tilde{\varphi}_t - \varphi_t\|_2 \leqslant \frac{1-\rho}{1-(\rho+\gamma\varepsilon_x)} \gamma \left(\frac{\beta}{e\rho \log(\rho^{-1})} \|\xi\|_2 \varepsilon_x + \frac{1}{1-\rho} \varepsilon_u \right),$$

which yields our desired bound.

□