On the Sample Complexity of Stability Constrained Imitation Learning

Stephen Tu STEPHENTU@GOOGLE.COM
Alexander Robey AROBEY1@SEAS.UPENN.EDU
Tingnan Zhang TINGNAN@GOOGLE.COM
Nikolai Matni NMATNI@SEAS.UPENN.EDU

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

Abstract

We study the following question in the context of imitation learning for continuous control: how are the underlying stability properties of an expert policy reflected in the sample complexity of an imitation learning task? We provide the first results showing that a granular connection can be made between the expert system's *incremental gain stability*, a novel measure of robust convergence between pairs of system trajectories, and the dependency on the task horizon T of the resulting generalization bounds. As a special case, we delineate a class of systems for which the number of trajectories needed to achieve ε-suboptimality is *sublinear* in the task horizon T, and do so without requiring (strong) convexity of the loss function in the policy parameters. Finally, we conduct numerical experiments demonstrating the validity of our insights on both a simple nonlinear system with tunable stability properties, and on a high-dimensional quadrupedal robotic simulation.

Keywords: Imitation learning, incremental stability, behavior cloning, statistical learning.

1. Introduction

Imitation Learning (IL) uses demonstrations of desired behavior, provided by an expert, to train a policy (Hussein et al., 2017; Osa et al., 2018). IL offers many appealing advantages: it is often more sample-efficient than reinforcement learning (Ross et al., 2011; Sun et al., 2017), and can lead to policies that are more computationally efficient to evaluate online (Hertneck et al., 2018; Yin et al., 2020). Indeed, there is a rich body of work demonstrating the advantages of IL-based methods in a range of applications including video-game playing (Ross and Bagnell, 2010; Ross et al., 2011), humanoid robotics (Schaal, 1999), and self-driving cars (Codevilla et al., 2018).

However, when applied to continuous control problems, little to no insight is given into how the underlying stability properties of the expert policy affect the sample complexity of the resulting IL task. In this paper, we address this gap and answer the question: what makes an expert policy easy to learn? Our main insight is that when an expert policy satisfies a suitable quantitative notion of robust *incremental* stability, i.e., when pairs of system trajectories under the expert policy robustly converge towards each other, and when learned policies are also constrained to satisfy this property, then IL can be made provably efficient. We formalize this insight through the notion of *incremental gain stability* constrained IL algorithms, and in doing so, quantify and generalize previous observations of efficient and robust learning subject to contraction based stability constraints.

Related Work. There exist a rich body of work examining the interplay between stability theory and learning dynamical systems/policies satisfying stability/safety properties from demonstrations.

Nonlinear stability and learning from demonstrations: Our work applies tools from nonlinear stability theory to analyze the sample complexity of IL algorithms. Concepts from nonlinear stability theory, such as Lyapunov stability or contraction theory (Lohmiller and Slotine, 1998), have also been successfully applied to learn autonomous nonlinear dynamical systems satisfying desirable properties such as (incremental) stability or controllability. As demonstrated empirically in (Lemme et al., 2014; Ravichandar et al., 2017; Sindhwani et al., 2018; Singh et al., 2020), using such stability-based regularizers to trim the hypothesis space results in more data-efficient and robust learning algorithms – however, no quantitative sample-complexity bounds are provided.

IL under covariate shift: Vanilla IL (e.g., Behavior Cloning) is known to be sensitive to covariate shift: as soon as the learned policy deviates from the expert policy, errors begin to compound, leading the system to drift to new and possibly dangerous states (Pomerleau, 1989; Ross et al., 2011). Representative IL algorithms that address this issue include DAgger (Ross et al., 2011) (on-policy approach) and DART (Laskey et al., 2017) (off policy approach). Both approaches seek to mitigate the effects of system drift at test-time by augmenting the data-set created by the expert: DAgger iteratively augments its data-set of trajectories with appropriately labeled and/or corrected data of the previous policy, whereas DART injects noise into the supervisor demonstrations and allows the supervisor to provide corrections as needed. For loss functions that are strongly convex in the policy parameters, DAgger enjoys $\tilde{O}(T)$ sample-complexity in the task horizon T, and this bound degrades to $\tilde{O}(T^2)$ when loss functions are only convex, whereas we are not aware of finite-data guarantees for DART. IL algorithms more explicitly focused on stability/safety leverage tools from Bayesian deep learning (Menda et al., 2017, 2019), model-predictive-control (Hertneck et al., 2018), robust control (Yin et al., 2020), and PAC-Bayes (Ren et al., 2020). While the approach, generality, and strength of guarantees provided by the aforementioned works vary, none provide insight as to how the stability properties of the expert affect the sample complexity of the corresponding IL task.

Contributions. To provide fine-grained insights into the relationship between system stability and sample complexity, we first define and analyze the notion of incremental gain stability (IGS) for a nonlinear dynamical system. IGS provides a quantitative measure of convergence between system trajectories, that strictly expands on the guarantees provided by contraction theory (Lohmiller and Slotine, 1998) by allowing for a graceful degradation away from exponential convergence rates.

With the aim of understanding what experts of easy to learn, we analyze the sample-complexity properties of IGS-constrained imitation learning algorithms. By linking nonlinear stability and statistical learning theory, we show that a degradation in the stability of the expert policy leads to a corresponding degradation of generalization bounds. In particular, we show that when imitating an IGS expert policy, IGS-constrained behavior cloning requires $m \gtrsim qT^{2a_1(1-1/a_1^2)}/\varepsilon^{2a_1}$ trajectories to achieve imitation loss bounded by ε, where T is the task horizon, q is the effective number of parameters of the function class for the learned policy, and $a_1 \in [1, \infty)$ is an IGS parameter determined by the expert policy. We show that $a_1 = 1$ for contracting systems, leading to *task-horizon independent* bounds scaling as $m \gtrsim q/\varepsilon^2$. Furthermore, we construct a simple family of systems that gracefully degrade away from these exponential rates such that $a_1 = 1 + p$ for $p \in (0, \infty)$, yielding sample-complexity that scales as $m \gtrsim qT^{2p(2+p)/1+p}\varepsilon^{2(1+p)}$, which makes clear that an increase/decrease in p yields a corresponding increase/decrease in sample-complexity. We further extend our analysis to an IGS-constrained DAgger-like algorithm. We show that this algorithm enjoys comparable stability dependent sample-complexity guarantees, requiring $m \gtrsim qT^{a_1^2(1-1/a_1)(4+1/a_1)}/\varepsilon^{2a_1^2}$ trajectories to learn an ε-approximate policy, again recovering time-independent bounds for contracting sys-

tems that gracefully degrade when applied to our family of systems satisfying $a_1 = 1 + p$. As a corollary, we provide the first results delineating a class of systems for which sample-complexity bounds scale sublinearly in the task-horizon T. Our final theoretical contribution is a lower-bound showing that stability-constrained IL is indeed necessary to ensure a graceful sample-complexity dependence on the task-time horizon T. We then empirically validate our results on (a) our simple family of nonlinear systems for which the underlying IGS properties can be quantitatively tuned, and (b) a high-dimensional nonlinear quadrupedal robotic system. We show that the sample-complexity scaling predicted by the underlying stability properties of the expert policy are indeed observed in practice. Omitted proofs and details can be found in the full version of the paper (Tu et al., 2021).

2. Problem Statement

We consider the following discrete-time control-affine dynamical system:

$$x_{t+1} = f(x_t) + g(x_t)u_t, \;\; x_t \in \mathbb{R}^n, \;\; u_t \in \mathbb{R}^d. \tag{2.1}$$

Let $\varphi_t(\xi, \{u_t\}_{t\geqslant 0})$ denote the state x_t of the dynamics (2.1) with input signal $\{u_t\}_{t\geqslant 0}$ and initial condition $x_0 = \xi$. For a policy $\pi : \mathbb{R}^n \to \mathbb{R}^d$, let $\varphi_t^\pi(\xi)$ denote x_t when $u_t = \pi(x_t)$. Let $X \subseteq \mathbb{R}^n$ be a compact set and let $T \in \mathbb{N}_+$ be the time-horizon over which we consider the behavior of (2.1). We generate trajectories by drawing random initial conditions from a distribution $\mathcal{D}$ over X.

We fix an expert policy $\pi_\star : \mathbb{R}^n \to \mathbb{R}^d$ which we wish to imitate. The quality of our imitation is measured through the following *imitation loss*:

$$\ell_{\pi'}(\xi; \pi_1, \pi_2) := \textstyle\sum_{t=0}^{T-1} \|\Delta_{\pi_1,\pi_2}(\varphi_t^{\pi'}(\xi))\|_2, \;\; \Delta_{\pi_1,\pi_2}(x) := g(x)(\pi_1(x) - \pi_2(x)). \tag{2.2}$$

Our goal is to design and analyze imitation learning algorithms that produce a policy $\hat{\pi}$ using $m = m(\varepsilon, \delta)$ trajectories of length T seeded from initial conditions $\{\xi_i\}_{i=1}^m \sim \mathcal{D}^m$, such that with probability at least $1 - \delta$, the learned policy $\hat{\pi}$ induces a state/input trajectory distribution that satisfies $\mathbb{E}_{\xi\sim\mathcal{D}}\ell_{\hat{\pi}}(\xi; \hat{\pi}, \pi_\star) \leqslant \varepsilon$. Crucially, we seek to understand how the underlying stability properties of the expert policy $\pi_\star$ manifest themselves in the number of required trajectories $m(\varepsilon, \delta)$.

Bounding the imitation loss has immediate implications on the safety, stability, and performance of the learned policy $\hat{\pi}$. Concretely, let $h : \mathbb{R}^{n\times(T+1)} \to \mathbb{R}^s$ denote an L_h-Lipschitz observable function of a trajectory: examples of valid observable functions include Lyapunov/barrier inequalities and constraints on the state or policy output. Then $\mathbb{E}_{\xi\sim\mathcal{D}}\|h_{\pi_\star}(\xi) - h_{\hat{\pi}}(\xi)\|_2 \leqslant L_h\mathbb{E}_{\xi\sim\mathcal{D}}\sum_{t=0}^T\|\varphi_t^{\pi_\star}(\xi) - \varphi_t^{\hat{\pi}}(\xi)\|_2$, where $h_\pi(\xi) := h(\{\varphi_t^\pi(\xi)\}_{t=0}^T)$. We subsequently show the discrepancy term $\sum_{t=0}^T\|\varphi_t^{\pi_\star}(\xi) - \varphi_t^{\hat{\pi}}(\xi)\|_2$ can be upper bounded by the imitation loss $\ell_{\hat{\pi}}(\xi; \hat{\pi}, \pi_\star)$, and thus bounds on the imitation loss imply bounds on the deviations between $h_{\pi_\star}$ & $h_{\hat{\pi}}$.

3. Incremental Gain Stability

The crux of our analysis relies on a property which we call *incremental gain stability (IGS)*. Before formally defining IGS, we motivate the need for a quantitative characterization of convergence rates between system trajectories. A key quantity that repeatedly appears in our analysis is the following sum of trajectory discrepancy induced by policies π_1 and π_2:

$$\mathsf{disc}_T(\xi; \pi_1, \pi_2) := \textstyle\sum_{t=0}^T\|\varphi_t^{\pi_1}(\xi) - \varphi_t^{\pi_2}(\xi)\|_2. \tag{3.1}$$

We already saw this quantity appear naturally in the deviation of $h_{\pi_\star}$ and $h_{\hat{\pi}}$. Furthermore, we will reduce analyzing the performance of behavior cloning and our DAgger-like algorithm to bounding the discrepancy (3.1) between trajectories induced by the expert policy $\pi_\star$ and a learned policy $\hat{\pi}$.

The simplest way to control (3.1) is to use a discrete-time version of Grönwall's inequality: if the maps f and g defining system (2.1) as well as policies π_1 and π_2 are B-bounded and L-Lipschitz, then we can upper bound $\mathsf{disc}_T(\xi;\pi_1,\pi_2) \leqslant O((L(1+2B))^T)\ell_{\pi_1}(\xi;\pi_1,\pi_2)$ whenever $L(1+2B)>1$. This bound formalizes the intuition that the discrepancy (3.1) scales in proportion to the deviation between policies π_1 and π_2, summed along the trajectory. Unfortunately, it is undesirable due to its exponential dependence on the horizon T. In order to improve the dependence on T, we need to assume some stability properties on the dynamics (f,g). We start by drawing inspiration from the definition of *incremental input-to-state stability* (Tran et al., 2016).[1]

Definition 3.1 (Incremental input-to-state-stability (δISS)) *Consider the discrete-time dynamics $x_{t+1}=f(x_t,u_t)$, and let $\varphi_t(\xi,\{u_t\}_{t\geqslant 0})$ denote the state x_t initialized from $x_0=\xi$ with input signal $\{u_t\}_{t\geqslant 0}$. The dynamics f is said to be* incremental input-to-state-stable *if there exists a class $\mathcal{KL}$ function ζ and class $\mathcal{K}_\infty$ function γ such that for every $\xi_1,\xi_2\in X$, $\{u_t\}_{t\geqslant 0}\subseteq U$, and $t\in\mathbb{N}$,*

$$\|\varphi_t(\xi_1,\{u_t\}_{t\geqslant 0})-\varphi_t(\xi_2,\{0\}_{t\geqslant 0})\|_X \leqslant \zeta(\|\xi_1-\xi_2\|_X,t)+\gamma\left(\max_{0\leqslant k\leqslant t-1}\|u_k\|_U\right).$$

Definition 3.1 improves the Grönwall-type estimate in the following way. Suppose the closed-loop system defined by $\tilde{f}(x,u)=f(x)+g(x)\pi_2(x)+u$ is δISS. Then the algebraic identity

$$x_{t+1}=f(x_t)+g(x_t)\pi_1(x_t)=f(x_t)+g(x_t)\pi_2(x_t)+\Delta_{\pi_1,\pi_2}(x_t),$$

allows us to treat $\Delta_{\pi_1,\pi_2}(x_t)$ as an input signal, yielding $\mathsf{disc}_T(\xi;\pi_1,\pi_2)\leqslant T\gamma\left(\ell_{\pi_1}(\xi;\pi_1,\pi_2)\right)$. This bound certainly improves the dependence on T, but is not sharp: for stable linear systems, it is not hard to show that $\mathsf{disc}_T(\xi;\pi_1,\pi_2)\leqslant O(1)\ell_{\pi_1}(\xi;\pi_1,\pi_2)$. In order to capture sharper rate dependence on T, we need to modify the definition to more explicitly quantify convergence rates.

Definition 3.2 (Incremental gain stability) *Consider the discrete time dynamics $x_{t+1}=f(x_t,u_t)$. Let $a,a_0,a_1,b_0,b_1\in[1,\infty)$ and ζ,γ be positive finite constants satisfying $a_0\leqslant a_1$ and $b_0\leqslant b_1$. Put $\Psi:=(a,a_0,a_1,b_0,b_1,\zeta,\gamma)$. We say that f is* Ψ-incrementally-gain-stable *(Ψ-IGS) if for all horizon lengths $T\in\mathbb{N}$, initial conditions $\xi_1,\xi_2\in X$, and input sequences $\{u_t\}_{t\geqslant 0}\subseteq U$, we have:*[2]

$$\textstyle\sum_{t=0}^{T}\min\{\|\Delta_t\|_X^{a\wedge a_0},\|\Delta_t\|_X^{a\vee a_1}\}\leqslant\zeta\|\xi_1-\xi_2\|_X^a+\gamma\sum_{t=0}^{T-1}\max\{\|u_t\|_U^{b_0},\|u_t\|_U^{b_1}\}. \tag{3.2}$$

Here, $\Delta_t:=\varphi_t(\xi_1,\{u_t\}_{t\geqslant 0})-\varphi_t(\xi_2,\{0\}_{t\geqslant 0})$.

IGS quantitatively bounds the amplification of an input signal $\{u_t\}_{t\geqslant 0}$ (and differences in initial conditions ξ_1,ξ_2) on the corresponding system trajectory discrepancies $\{\Delta_t\}_{t\geqslant 0}$. Note that a system that is incrementally gain stable is automatically δISS. IGS also captures the phase transition that occurs in non-contracting systems about the unit circle. For example, when $\|\Delta_t\|\leqslant 1$ and $\|u_t\|\leqslant 1$ for all $t\geqslant 0$, inequality (3.2) reduces to $\sum_{t=0}^{T}\|\Delta_t\|_X^{a\vee a_1}\leqslant\zeta\|\xi_1-\xi_2\|_X^a+\gamma\sum_{t=0}^{T-1}\|u_t\|_U^{b_0}$. Finally, as IGS measures signal-to-signal ($\{u_t\}_{t\geqslant 0}\to\{\Delta_t\}_{t\geqslant 0}$) amplification, it is well suited to analyzing learning algorithms operating on system trajectories.

For a pair of functions $f(x),g(x)$, we say that (f,g) is Ψ-IGS if the system $f(x)+g(x)u$ is Ψ-IGS. It turns out that if $(f+g\pi_2,\mathsf{Id})$ is Ψ-IGS, then

$$\mathsf{disc}_T(\xi;\pi_1,\pi_2)\leqslant 4\,(\gamma\vee 1)^{\frac{1}{a\wedge a_0}}\,T^{1-\frac{1}{a\vee a_1}}\max\left\{\ell_{\pi_1}(\xi;\pi_1,\pi_2)^{\frac{b_0}{a\vee a_1}},\ell_{\pi_1}(\xi;\pi_1,\pi_2)^{\frac{b_1}{a\wedge a_0}}\right\}. \tag{3.3}$$

1. Definition 3.1 is more general than Definition 6 of (Tran et al., 2016) in that we only require a bound with respect to an input perturbation of one of the trajectories, not both.
2. For $a,b\in\mathbb{R}$, we let $a\vee b:=\max\{a,b\}$ and $a\wedge b:=\min\{a,b\}$.

With this bound, the dependence on T is allowed to interpolate between 1 and T, and the dependence on $\ell_{\pi_1}(\xi;\pi_1,\pi_2)$ is made explicit. Next, we state a Lyapunov based sufficient condition for Definition 3.2 to hold, which is checked pointwise in space rather than over entire trajectories.

Proposition 3.3 (Incremental Lyapunov function implies stability) *Let $a, a_0, a_1, b_0, b_1 \in [1,\infty)$ and $\underline{\alpha}, \overline{\alpha}, \mathfrak{a}, \mathfrak{b}$ be positive finite constants satisfying $a_0 \leqslant a_1$, $b_0 \leqslant b_1$, and $\underline{\alpha} \leqslant \overline{\alpha}$. Suppose there exists a non-negative function $V : \mathbb{R}^n \times \mathbb{R}^n \to \mathbb{R}_+$ satisfying $\underline{\alpha}\|x-y\|_X^a \leqslant V(x,y) \leqslant \overline{\alpha}\|x-y\|_X^a$, such that for all $x, y \in \mathbb{R}^n$ and $u \in U$,*

$$V(f(x,u),f(y,0)) - V(x,y) \leqslant -\mathfrak{a}\min\{\|x-y\|_X^{a_0}, \|x-y\|_X^{a_1}\} + \mathfrak{b}\max\{\|u\|_U^{b_0}, \|u\|_U^{b_1}\}. \quad (3.4)$$

Then, f is Ψ-incrementally-gain-stable with $\Psi = \left(a, a_0, a_1, b_0, b_1, \frac{\overline{\alpha}}{\underline{\alpha}\wedge\mathfrak{a}}, \frac{\mathfrak{b}}{\underline{\alpha}\wedge\mathfrak{a}}\right)$.

An example of an IGS system is a contracting system (Lohmiller and Slotine, 1998).

Proposition 3.4 *Consider the dynamics $x_{t+1} = f(x_t, u_t)$. Suppose that f is* autonomously contracting*, i.e., there exists a positive definite metric $M(x)$ and a scalar $\rho \in (0,1)$ such that:*

$$\frac{\partial f}{\partial x}(x,0)^\mathsf{T} M(f(x,0))\frac{\partial f}{\partial x}(x,0) \preccurlyeq \rho M(x) \;\; \forall x \in \mathbb{R}^n.$$

Suppose also that the metric M satisfies $\underline{\mu} I \preccurlyeq M(x) \preccurlyeq \overline{\mu} I$ for all $x \in \mathbb{R}^n$, and that there exists a finite L_u such that the dynamics satisfies $\|f(x,u) - f(x,0)\|_2 \leqslant L_u\|u\|_2$ for all $x \in \mathbb{R}^n$, $u \in \mathbb{R}^d$. Then we have that f is Ψ-IGS, with $\Psi = \left(1,1,1,1,1, \sqrt{\frac{\overline{\mu}}{\underline{\mu}}}\frac{1}{1-\sqrt{\rho}}, L_u\sqrt{\frac{\overline{\mu}}{\underline{\mu}}}\frac{1}{1-\sqrt{\rho}}\right)$.

Proposition 3.4 shows that for contracting systems, (3.3) is bounded by $O(\ell_{\pi_1}(\xi;\pi_1,\pi_2))$. A concrete example is a piecewise linear system $f(x,u) = \left(\sum_{i=1}^K A_i \mathbf{1}\{x \in \mathcal{C}_i\}\right)x + Bu$ where the A_i's are stable, $\{\mathcal{C}_i\}$ partitions $\mathbb{R}^n$, and there exists a common quadratic Lyapunov function $V(x) = x^\mathsf{T}Px$ which yields the metric $M(x) = P$.

Our next example is a family of systems that gracefully degrade away from exponential rates.

Proposition 3.5 *Consider the scalar dynamics $x_{t+1} = x_t - \eta x_t \frac{|x_t|^p}{1+|x_t|^p} + \eta u_t$ for $p \in (0,\infty)$. Then as long as $0 < \eta < \frac{4}{5+p}$, we have that f is Ψ-IGS, with $\Psi = \left(1,1,1+p,1,1,\frac{2^{2+p}}{\eta}, 2^{2+p}\right)$*

The systems described in Proposition 3.5 behave like stable linear systems when $|x_t| \geqslant 1$ (hence $a_0 = 1$), and like polynomial systems when $|x_t| < 1$ (hence $a_1 = 1+p$). This also highlights the need to be able to capture this phase-transition within our definitions.

4. Algorithms and Theoretical Results

In this section we define and analyze IGS-constrained imitation learning algorithms. We begin by introducing our main assumption of dynamics and policy class regularity.

Assumption 4.1 (Regularity) *We assume that the dynamics (f,g), policy class Π, expert $\pi_\star$, and initial condition distribution $\mathcal{D}$ satisfy, for some $B_g, B_0, L_\Delta \in [1,\infty)$:*

(a) The dynamics (f,g) satisfy (i) $f(0) = 0$ and (ii) $\sup_{x\in\mathbb{R}^n}\|g(x)\|_{\mathrm{op}} \leqslant B_g$.
(b) The policy class Π contains $\pi_\star$, is convex[3], and $\pi(0) = 0$ for all $\pi \in \Pi$.

3. Convexity is a stronger assumption than required and made to streamline the presentation. It can be relaxed to Π being closed under a finite number of convex combinations $(1-\alpha)\pi_1 + \alpha\pi_2$, $\alpha \in [0,1]$, $\pi_1, \pi_2 \in \Pi$.

(c) Δ_{π_1,π_2} *is* L_Δ*-Lipschitz for all* $\pi_1, \pi_2 \in \Pi$.
(d) The distribution $\mathcal{D}$ *over initial conditions satisfies* $\|\xi\|_2 \leqslant B_0$ *a.s. for* $\xi \sim \mathcal{D}$.

We now turn to our main stability assumption. Let $f_{\mathrm{cl}}^{\pi}(x) := f(x) + g(x)\pi(x)$ denote the closed-loop dynamics induced by a policy π. Our main stability assumption is that $(f_{\mathrm{cl}}^{\pi_\star}, \mathsf{Id})$ is Ψ-IGS. For ease of exposition, we restrict the degrees of freedom of the Ψ-IGS parameters, but note that our results extend to the general case at the expense of more cumbersome expressions.

Assumption 4.2 (Incremental Gain Stability) *Let* $\Psi = (a, a_0, a_1, b_0, b_1, \zeta, \gamma)$ *be a tuple satisfying* $a = a_0$, $b := b_0 = b_1$, $\zeta \geqslant 1$, $\gamma \geqslant 1$, *and* $a \geqslant b$. *Let* $\mathcal{S}_\Psi$ *denote the set of policies* π *such that* $(f_{\mathrm{cl}}^{\pi}, \mathsf{Id})$ *is* Ψ*-IGS. We assume that* $\pi_\star \in \mathcal{S}_\Psi$.

Algorithm 1: Constrained Mixing Iterative Learning

Data: Trajectory budget m, number of epochs E that divides m, mixing rate $\alpha \in (0, 1]$, initial conditions $\left\{\{\xi_i^k\}_{i=1}^{m/E}\right\}_{k=0}^{E-1} \sim \mathcal{D}^n$, expert $\pi_\star$, stability parameters Ψ, and scalars $\{c_i\}_{i=1}^{E-2}$.

$\pi_0 \leftarrow \pi_\star$, $c_0 \leftarrow 0$.
for $k = 0, \ldots, E-2$ **do**
  Collect trajectories $\mathcal{T}_k = \left\{\{\varphi_t^{\pi_k}(\xi_i^k)\}_{t=0}^{T-1}\right\}_{i=1}^{m/E}$.
  $\hat{\pi}_k \leftarrow$ cERM $(\mathcal{T}_k, \pi_k, c_k, 0)$.
  $\pi_{k+1} \leftarrow (1-\alpha)\pi_k + \alpha\hat{\pi}_k$.
end
Collect trajectories $\mathcal{T}_{E-1} = \left\{\{\varphi_t^{\pi_{E-1}}(\xi_i)\}_{t=0}^{T-1}\right\}_{i=1}^{m/E}$.
$c_{E-1} \leftarrow \frac{(1-\alpha)^E}{\alpha} \frac{1}{m/E} \sum_{i=1}^{m/E} \ell_{\pi_{E-1}}(\xi_i^{E-1}; \pi_{E-1}, \pi_\star)$.
$\hat{\pi}_{E-1} \leftarrow$ cERM $\left(\mathcal{T}_{E-1}, \pi_{E-1}, c_{E-1}, (1-\alpha)^E\right)$.
$\pi_E \leftarrow \frac{1}{1-(1-\alpha)^E}[(1-\alpha)\pi_{E-1} + \alpha\hat{\pi}_{E-1} - (1-\alpha)^E\pi_\star]$.
return π_E.

Note that the assumptions of boundedness of g and Lipschitz continuity of Δ_{π_1,π_2} in Assumption 4.1 can be relaxed to continuity of the dynamics (f, g) and boundedness of $\pi \in \Pi$ by our assumption that $\pi_\star \in \mathcal{S}_\Psi$, which ensures boundedness over state trajectories under bounded inputs. However, explicitly defining B_g and L_Δ streamlines the exposition. Next, we assume that the expert policy lies in our policy class, i.e., that $\pi_\star \in \Pi$, to guarantee that zero imitation loss can be achieved in the limit of infinite data; it is straightforward to relax this assumption to $\Pi \cap \mathcal{S}_\Psi \neq \emptyset$ and prove results with respect to the best stabilizing policy in class.

IGS-Constrained Mixing Iterative Learning (CMILe) is presented in Algorithm 1. CMILe integrates ideas from Stochastic Mixing Iterative Learning (SMILe) (Ross and Bagnell, 2010), constrained policy optimization (Luo et al., 2019; Schulman et al., 2015), and the IGS tools developed in Section 3. As in SMILe and DAgger, CMILe proceeds in epochs, beginning with data generated by the expert policy, and iteratively shifts towards a learned policy via updates of the form $\pi_{k+1} = (1-\alpha)\pi_k + \alpha\hat{\pi}_k$, where π_k is the current data-generating policy, $\hat{\pi}_k$ is the policy learned using the most recently generated data, and $\alpha \in (0, 1]$ is a mixing parameter. However, CMILe contains two key differences: it constrains π_k to (i) remain close to π_{k-1} (constraint (4.2b)), and (ii) to induce Ψ-IGS closed-loop systems (constraint (4.2c)). The latter allows us to leverage the Ψ-IGS machinery of §3 to analyze Algorithm 1.

In presenting our results, we specialize the policy class Π to have the particular parametric form:

$$\Pi = \{\pi(x, \theta) : \theta \in \mathbb{R}^q, \|\theta\|_2 \leqslant B_\theta\}, \tag{4.1}$$

with $B_\theta \geqslant 1$, and π a fixed twice continuously differentiable map. As an example, neural networks with q weights and twice continuously differentiable activation functions are captured by the policy

class (4.1). We note that our results do not actually require a parameteric representation: as long as a particular policy class Rademacher complexity can be bounded, then our results apply. In what follows, we define $L_{\partial^2\pi} := 1 \vee \sup_{\|x\|\leqslant\zeta^{1/a}B_0,\|\theta\|\leqslant B_\theta} \left\| \frac{\partial^2\pi}{\partial\theta\partial x} \right\|_{\ell^2(\mathbb{R}^q)\to M(\mathbb{R}^{d\times n})}$, with $M(\mathbb{R}^{d\times n})$ the space of $d \times n$ real-valued matrices equipped with the operator norm.

IGS-Constrained Behavior Cloning. We first analyze a single epoch version of Algorithm 1, which reduces to Behavior Cloning (BC) subject to interpolating the expert policy on the training data (constraint (4.2b)) and inducing a Ψ-IGS closed-loop system (constraint (4.2c)). Our analysis is summarized in the following theorem, which bounds the closed-loop imitation loss.

Algorithm 2: cERM $(\mathcal{T}, \pi_{\text{roll}}, c, w)$

Data: $\mathcal{T} = \left\{ \{\varphi_t^{\pi_{\text{roll}}}(\xi_i)\}_{t=0}^{T-1} \right\}_{i=1}^{m}$, policy $\pi_{\text{roll}} \in \Pi$, constraint $c \geqslant 0$, weight $w \in [0, 1)$

return the solution to:

$$\text{minimize}_{\bar{\pi}\in\Pi} \ \tfrac{1}{m}\textstyle\sum_{i=1}^m \ell_{\pi_{\text{roll}}}(\xi_i; \bar{\pi}, \pi_\star) \tag{4.2a}$$

$$\text{subject to } \tfrac{1}{m}\textstyle\sum_{i=1}^m \ell_{\pi_{\text{roll}}}(\xi_i; \bar{\pi}, \pi_{\text{roll}}) \leqslant c, \tag{4.2b}$$

$$\tfrac{1}{1-w}[(1-\alpha)\pi_{\text{roll}} + \alpha\bar{\pi} - w\pi_\star] \in \mathcal{S}_\Psi. \tag{4.2c}$$

Theorem 4.3 (IGS-constrained BC) *Suppose that Assumption 4.1 and Assumption 4.2 hold. Set $\alpha = E = 1$ in Algorithm 1. Suppose that m satisfies:*

$$m \geqslant \Omega(1) q \zeta^{\frac{2}{a}} B_0^2 T^{2\left(1-\frac{1}{a_1}\right)} \max\{B_g B_\theta L_{\partial^2\pi}, L_\Delta\}^2.$$

With probability at least $1 - e^{-q}$ over the randomness of Algorithm 1, we have that:

$$\mathbb{E}_{\xi\sim\mathcal{D}} \ell_{\pi_1}(\xi; \pi_1, \pi_\star) \leqslant O(1) \gamma^{\frac{1}{a}} \zeta^{\frac{b}{aa_1}} B_0^{\frac{b}{a_1}} T^{\left(1-\frac{1}{a_1}\right)\left(1+\frac{b}{a_1}\right)} L_\Delta \max\{B_g B_\theta L_{\partial^2\pi}, L_\Delta\}^{\frac{b}{a_1}} \left(\frac{q}{m}\right)^{\frac{b}{2a_1}}.$$

Theorem 4.3 shows that the imitation loss for IGS-constrained BC decays as $O\big(T^{2(1-1/a_1)}(\frac{q}{m})^{b/2a_1}\big)$. We discuss implications on sample-complexity after analyzing the general setting.

IGS-CMILe. Next we show that if the mixing parameter α and number of episodes E are chosen appropriately with respect to the Ψ-IGS parameters of the underlying expert system, sample-complexity guarantees similar to those in the IGS-constrained BC setting can be obtained. The key to ensuring that guarantees can be bootstrapped across epochs is the combination of a trust-region constraint (4.2b) and IGS-stability constraints (4.2c) on the intermediate policies.

Theorem 4.4 (IGS-CMILe) *Suppose that Assumption 4.1 and Assumption 4.2 hold, and that:*

$$m \geqslant \Omega(1) E(q \vee \log E) \zeta^{\frac{2}{a}} B_0^2 T^{2\left(1-\frac{1}{a_1}\right)} \max\{B_g B_\theta L_{\partial^2\pi}, L_\Delta\}^2.$$

Suppose further that for $k \in \{1, ..., E-2\}$, we have:

$$c_k \leqslant O(1) \zeta^{\frac{1}{a}} B_0 T^{1-\frac{1}{a_1}} \max\{B_g B_\theta L_{\partial^2\pi}, L_\Delta\} \sqrt{\frac{E(q \vee \log E)}{m}},$$

that E divides m, $E \geqslant \frac{1}{\alpha}\log\left(\frac{1}{\alpha}\right)$, and $\alpha \leqslant \min\left\{\frac{1}{2}, \frac{1}{L_\Delta \gamma^{1/a} T^{1-1/a_1}}\right\}$. Then with probability at least $1 - e^{-q}$ over the randomness of Algorithm 1, Algorithm 1 is feasible for all epochs, and:

$$\mathbb{E}_{\xi\sim\mathcal{D}} \ell_{\pi_E}(\xi; \pi_E, \pi_\star) \leqslant O(1) \zeta^{\frac{b}{aa_1}} \gamma^{\left(1-\frac{b^2}{a_1^2}\right)\frac{1}{a}} B_0^{\frac{b}{a_1}} T^{1-\frac{1}{a_1}} L_\Delta^{1-\frac{b^2}{a_1^2}} \times \max\{B_g B_\theta L_{\partial^2\pi}, L_\Delta\}^{\frac{b}{a_1}} E^{1+\frac{b}{2a_1}} \left(\frac{q \vee \log E}{m}\right)^{\frac{b^2}{2a_1^2}}.$$

Theorem 4.4 states that if the mixing parameter α and number of episodes E are set according to the underlying IGS-stability parameters of the expert system then the imitation loss of the final policy π_E scales as $\tilde{O}\big(T^{(1-1/a_1)(2+b/2a_1)}(\frac{q}{m})^{b^2/2a_1^2}\big)$.

Sublinear rates. From the above discussion, we can delineate classes of systems for which imitation loss bounds are sublinear in the task horizon T. Using that $b \geqslant 1$, we see that IGS-constrained behavior cloning requires $m \gtrsim T^{2a_1(1-1/a_1^2)}/\varepsilon^{2a_1}$ trajectories to achieve ε-bounded imitation loss: hence, if $a_1 \in [1, \frac{1}{4}(1+\sqrt{17}) \approx 1.28)$, this sample-complexity bound scales sublinearly in T. A similar analysis shows that CMILe achieves sublinear scaling in T whenever $a_1 \in [1, \frac{1}{8}(3+\sqrt{41}) \approx 1.175)$. Finally, when a system is contracting, $a = a_1 = b = 1$ (Prop. 3.4), and hence the imitation-loss bounds for both IGS-constrainted BC and CMILe reduce to $\tilde{O}\big(\sqrt{\frac{q}{m}}\big)$.

Necessity of stability constraints. We illustrate the necessity of imposing stability constraints to derive high probability sub-exponential in T bounds on the imitation error. Consider the LTI system $x_{t+1} = Ax_t + u_t$, with expert $\pi_\star(x) = -Ax_t$, and $A = \mathrm{diag}(2, 2, 0, ..., 0)$. Observe that $\varphi_t^{\pi_\star}(\xi) = 0$ for all $t \geqslant 1$ and $\xi \in \mathbb{R}^n$. Hence, for any m initial conditions $\xi_1, ..., \xi_m$, all the informative data we will see from the expert is $\{(\xi_i, y_i := \pi_\star(\xi_i))\}_{i=1}^m$.

Let m_0 be large enough so that $(1-1/m)^m \geqslant 1/(2e)$ for all $m \geqslant m_0$ (such an m_0 exists since $\lim_{m\to\infty}(1-1/m)^m = 1/e$), and fix any $m \geqslant m_0$. Consider $\mathcal{D}$ defined as $\mathbb{P}_{\xi\sim\mathcal{D}}(\xi = e_1) = 1 - 1/m$ and $\mathbb{P}_{\xi\sim\mathcal{D}}(\xi = e_2) = 1/m$, where $e_i \in \mathbb{R}^n$ is the i-th standard basis vector. Let $\mathcal{E}_m := \bigcap_{i=1}^m \{\xi_i = e_1\}$, and observe that $\mathbb{P}(\mathcal{E}_m) = (1-1/m)^m \geqslant 1/(2e)$. We will consider optimization over the compact convex policy class $\Pi = \{x \mapsto Kx : K \in \mathbb{R}^{n\times n}, \|K\|_F \leqslant 2\sqrt{2}\}$ which contains $\pi_\star$. Note that Assumptions 4.1 and 4.2 hold for the closed-loop expert dynamics and policy class. The ERM problem is $\min_{K\in\mathbb{R}^{n\times n}:\|K\|_F\leqslant 2\sqrt{2}} \frac{1}{m}\sum_{i=1}^m \|K\xi_i - y_i\|_2$. It is easy to check that on $\mathcal{E}_m$ (a constant probability event), $\hat{K} = -2e_1e_1^\mathsf{T}$ is an interpolating solution of this ERM problem. Let $\hat{\pi}(x) = \hat{K}x$. Since $\varphi_t^{\hat{\pi}}(e_2) = 2^t e_2$, a straightforward computation shows that $\mathbb{E}_{\xi\sim\mathcal{D}}\ell_{\hat{\pi}}(\xi; \hat{\pi}, \pi_\star) \geqslant \frac{4(2^T-1)}{m} = \Omega(2^T/m)$. This illustrates that under our assumptions, sub-exponential bounds are impossible without enforcing stability constraints.

5. Experiments

In our experiments, we implement neural network training by combining the `haiku` NN library (Hennigan et al., 2020) with `optax` (Hessel et al., 2020) in `jax` (Bradbury et al., 2018).

Tuneable Ψ-IGS System. We consider the dynamical system in $\mathbb{R}^{10}$:

$$x_{t+1} = x_t - \eta x_t \frac{|x_t|^p}{1+|x_t|^p} + \frac{\eta}{1+|x_t|^p}(h(x_t) + u_t), \quad \eta = 0.3. \tag{5.1}$$

All arithmetic operations in (5.1) are element-wise. We set $h : \mathbb{R}^{10} \to \mathbb{R}^{10}$ to be a randomly initialized two layer MLP with zero biases, hidden width 32, and $\tanh$ activations. The expert policy is set to be $\pi_\star = -h$, so that the expert's closed-loop dynamics are given by $x_{t+1} = x_t - \eta x_t \frac{|x_t|^p}{1+|x_t|^p}$. From Proposition 3.5, the incremental stability of the closed-loop system degrades as p increases.

In this experiment, we vary $p \in \{1, ..., 5\}$ to see the effect of p on the final task goal error and imitation loss. We compare four different algorithms, and investigate the inclusion of IGS constraints (4.2c). **BC** is standard behavior cloning. **CMILe** is Algorithm 1, with hard constraints replaced by soft constraints in the cost. **CMILe+Agg** is **CMILe**, except that at epoch k, the data from previous epochs $j \in \{0, ..., k-1\}$ is also used in training. **DAgger** is the imitation learning

p	**BC+IGS**	**BC**	**CMILe+IGS**	**CMILe**
1	0.149 ± 0.020	0.335 ± 0.073	0.167 ± 0.013	0.199 ± 0.047
2	0.454 ± 0.032	0.782 ± 0.158	0.510 ± 0.018	0.692 ± 0.026
3	0.829 ± 0.131	1.128 ± 0.118	0.852 ± 0.057	1.099 ± 0.046
4	1.220 ± 0.176	1.737 ± 0.126	1.041 ± 0.045	1.412 ± 0.052
5	1.899 ± 0.160	2.067 ± 0.214	1.236 ± 0.035	1.535 ± 0.042

Table 1: Final $\|x_T^{\text{expert}} - x_T^{\text{IL}}\|_2$ of **BC** and **CMILe** with and without IGS constraints on (5.1).

algorithm from Ross et al. (2011). For all algorithms, we fix the number of trajectories m from (5.1) to be $m = 250$. The horizon length is $T = 100$. The distribution $\mathcal{D}$ over initial condition is set as $N(0, I)$. We set the policy class Π to be two layer MLPs with hidden width 64 and $\tanh$ activations. Each algorithm minimizes the imitation loss using 300 epochs of Adam with learning rate 0.01 and batch size 512. For all algorithms except **BC**, we use $E = 25$ epochs with $\alpha = 0.15$ (in DAgger's notation, we set $\beta_k = 0.85^k$), resulting in 10 trajectories per epoch.

In Table 1, we track the difference in norm $\|x_T^{\text{expert}} - x_T^{\text{IL}}\|_2$ between the expert's (x_T^{expert}) and the IL algorithm's final state (x_T^{IL}), both seeded from the same initial conditions. The table entries are computed by rolling out 500 test trajectories and computing the median quantity over the test trajectories. Each algorithm is repeated for 50 trials, and the median quantity $\pm \max(80\text{th percentile} - \text{median}, \text{median} - 20\text{th percentile})$ (over the 50 trials) is shown. In Table 1, we see that as p decreases, the goal deviation error decreases, showing that the task becomes easier, as predicted by our main results.[4] Finally, we note that while the same trends hold for both vanilla and IGS-constrained versions of **BC** and **CMILe**, the quantitative performance degrades, showing that stability constraints reduce sample-complexity by trimming the hypothesis space.

Unitree Laikago. We now study IL on the Unitree Laikago robot, an 18-dof quadruped with 3-dof of actuation per leg. We use PyBullet (Coumans and Bai, 2016–2021) for our simulations. The goal of this experiment is to demonstrate, much like for the previous tuneable family of Ψ-IGS systems, that increasing the stability of the underlying closed-loop expert decreases the sample-complexity of imitation learning. We do this qualitatively by studying a sideways walking task where the robot tracks a constant sideways linear velocity: the larger the desired linear velocity, the more unstable the resulting expert. Our expert controller is a model-based predictive controller using a simplified center-of-mass dynamics as described in Di Carlo et al. (2018). The stance and swing legs are controlled separately, and we restrict our imitation learning to the stance leg controller, as it is significantly more complex than the swing leg controller.[5] Furthermore, instead of randomizing over initial conditions, we inject randomization into the environment by subjecting the Laikago to a sequence of random push forces throughout the entire trajectory. We compare **BC**, **CMILe**, **CMILe+Agg** (omitting **DAgger** for space reasons). We set the horizon length to $T = 1000$. We featurized the robot state into a 14-dimensional feature vector, and the output of the policy is a 12-dimensional vector (x, y, z contact forces for each of the 4 legs). We used a policy class of two layer MLPs of hidden width 64 with ReLU activations. For training, we ran 500 epochs of Adam with a batch size of 512 and step size of 0.001. To assess the effect of the number of samples on

4. This trend is reflected in all the imitation learning algorithms that we evaluated, but in the interest of space, we only show results for **BC** and **CMILe** and IGS-constrained versions thereof – qualitatively similar results are obtained for the other algorithms, and can be found in the full paper (Tu et al., 2021)
5. More details about the expert controller can be found in the full paper.

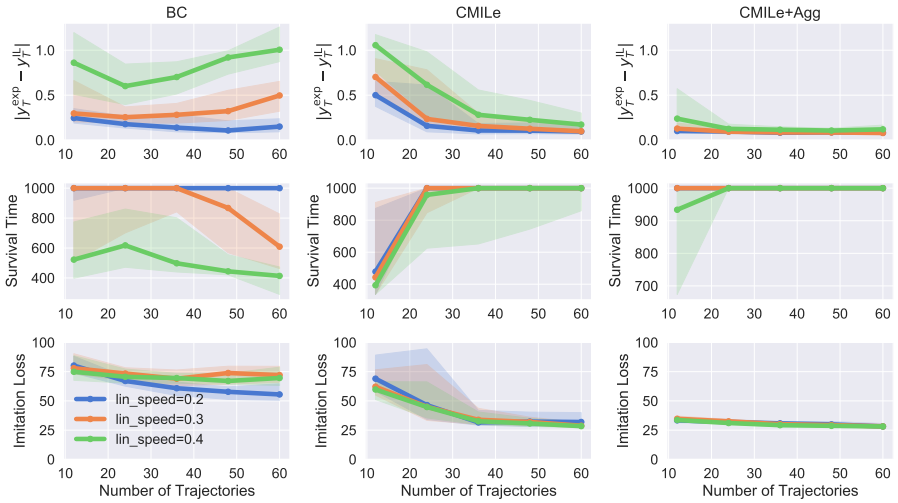

Figure 1: IL on a sideways walking task. The top, middle, and bottom rows show the deviation error $|y_T^{\text{exp}} - y_T^{\text{IL}}|$, the survival times, and the imitation loss $\frac{1}{T}\mathbb{E}_{\xi\sim\mathcal{D}}\ell_{\pi_E}(\xi;\pi_E,\pi_\star)$ of the various algorithms.

IL, we vary the number of rollouts per epoch $S \in \{1, ..., 5\}$. For **CMILe** and **CMILe+Agg**, we fix $\alpha = 0.3$ and $E = 12$. We provide **BC** with $S \times E$ total trajectories.

Figure 1 shows the result of our experiments. In the top row, we plot the *deviation* $|y_T^{\text{exp}} - y_T^{\text{IL}}|$ between the expert's (y_T^{exp}) and the IL algorithm's final y position (y_T^{IL}). We observe that as the target linear speed decreases, the deviation between the expert and IL algorithms also decreases; this qualitative trend is consistent with Theorem 4.3 and Theorem 4.4.[6] In the middle row, we plot the survival times for each of the algorithms, which is the number of simulation steps (out of 1000) that the robot executes before a termination criterion triggers when the robot is about to fall. We see that for all algorithms, by decreasing the sideways linear velocity, the resulting learned policy is able to avoid falling more. In the bottom row, we plot the average closed-loop imitation loss $\frac{1}{T}\mathbb{E}_{\xi\sim\mathcal{D}}\ell_{\pi_E}(\xi;\pi_E,\pi_\star)$. We see that for **BC**, the imitation loss shows improvement with increased samples for linear speed of 0.2, but none for the larger linear speeds of $0.3, 0.4$. This trend is less apparent for **CMILe** and **CMILe+Agg**, but is reflected in the deviation error $|y_T^{\text{exp}} - y_T^{\text{IL}}|$.

6. Conclusions & Future Work

We showed that IGS-constrained IL algorithms allow for a granular connection between the stability properties of an underlying expert system and the resulting sample-complexity of an IL task. Our future work will focus on theoretically characterizing why CMILe and DAgger significantly outperform BC in our experiments. We will also look to apply the general framework for reasoning about learning over trajectories in continuous state and action spaces that we developed to other settings.

6. Note that the y positions are computed by subjecting the Laikago to the same sequence of random force pushes for both the expert and IL algorithm.

References

James Bradbury, Roy Frostig, Peter Hawkins, Matthew James Johnson, Chris Leary, Dougal Maclaurin, George Necula, Adam Paszke, Jake VanderPlas, Skye Wanderman-Milne, and Qiao Zhang. JAX: composable transformations of Python+NumPy programs, 2018. URL http://github.com/google/jax.

Felipe Codevilla, Matthias Miiller, Antonio López, Vladlen Koltun, and Alexey Dosovitskiy. End-to-end driving via conditional imitation learning. In *2018 IEEE International Conference on Robotics and Automation (ICRA)*, 2018.

Erwin Coumans and Yunfei Bai. Pybullet, a python module for physics simulation for games, robotics and machine learning. http://pybullet.org, 2016–2021.

Jared Di Carlo, Patrick M. Wensing, Benjamin Katz, Gerardo Bledt, and Sangbae Kim. Dynamic locomotion in the mit cheetah 3 through convex model-predictive control. In *2018 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2018.

Tom Hennigan, Trevor Cai, Tamara Norman, and Igor Babuschkin. Haiku: Sonnet for JAX, 2020. URL http://github.com/deepmind/dm-haiku.

Michael Hertneck, Johannes Köhler, Sebastian Trimpe, and Frank Allgöwer. Learning an approximate model predictive controller with guarantees. *IEEE Control Systems Letters*, 2(3):543–548, 2018.

Matteo Hessel, David Budden, Fabio Viola, Mihaela Rosca, Eren Sezener, and Tom Hennigan. Optax: composable gradient transformation and optimisation, in jax!, 2020. URL http://github.com/deepmind/optax.

Ahmed Hussein, Mohamed M. Gaber, Eyad Elyan, and Chrisina Jayne. Imitation learning: A survey of learning methods. *ACM Computing Surveys (CSUR)*, 50(2):1–35, 2017.

Michael Laskey, Jonathan Lee, Roy Fox, Anca Dragan, and Ken Goldberg. Dart: Noise injection for robust imitation learning. In *Conference on Robot Learning*, 2017.

Andre Lemme, Klaus Neumann, R. Felix Reinhart, and Jochen J. Steil. Neural learning of vector fields for encoding stable dynamical systems. *Neurocomputing*, 141:3–14, 2014.

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

Yuping Luo, Huazhe Xu, Yuanzhi Li, Yuandong Tian, Trevor Darrell, and Tengyu Ma. Algorithmic framework for model-based deep reinforcement learning with theoretical guarantees. In *International Conference on Learning Representations*, 2019.

Kunal Menda, Katherine Driggs-Campbell, and Mykel J. Kochenderfer. Dropoutdagger: A bayesian approach to safe imitation learning. *arXiv preprint arXiv:1709.06166*, 2017.

Kunal Menda, Katherine Driggs-Campbell, and Mykel J. Kochenderfer. Ensembledagger: A bayesian approach to safe imitation learning. In *2019 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2019.

Takayuki Osa, Joni Pajarinen, Gerhard Neumann, J. Andrew Bagnell, Pieter Abbeel, and Jan Peters. An algorithmic perspective on imitation learning. *Foundations and Trends® in Robotics*, 7(1–2): 1–179, 2018.

Dean A. Pomerleau. Alvinn: An autonomous land vehicle in a neural network. In *Neural Information Processing Systems*, 1989.

Harish Ravichandar, Iman Salehi, and Ashwin Dani. Learning partially contracting dynamical systems from demonstrations. In *Conference on Robot Learning*, 2017.

Allen Z. Ren, Sushant Veer, and Anirudha Majumdar. Generalization guarantees for imitation learning. In *Conference on Robot Learning*, 2020.

Stéphane Ross and J. Andrew Bagnell. Efficient reductions for imitation learning. In *International Conference on Artificial Intelligence and Statistics*, 2010.

Stéphane Ross, Geoffrey J. Gordon, and J. Andrew Bagnell. A reduction of imitation learning and structured prediction to no-regret online learning. In *International Conference on Artificial Intelligence and Statistics*, 2011.

Stefan Schaal. Is imitation learning the route to humanoid robots? *Trends in cognitive sciences*, 3 (6):233–242, 1999.

John Schulman, Sergey Levine, Pieter Abbeel, Michael Jordan, and Philipp Moritz. Trust region policy optimization. In *International Conference on Machine Learning*, 2015.

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

Sumeet Singh, Spencer M. Richards, Vikas Sindhwani, Jean-Jacques E. Slotine, and Marco Pavone. Learning stabilizable nonlinear dynamics with contraction-based regularization. *The International Journal of Robotics Research*, 2020.

Wen Sun, Arun Venkatraman, Geoffrey J. Gordon, Byron Boots, and J. Andrew Bagnell. Deeply aggrevated: Differentiable imitation learning for sequential prediction. In *International Conference on Machine Learning*, 2017.

Duc N. Tran, Björn S. Rüffer, and Christopher M. Kellett. Incremental stability properties for discrete-time systems. In *2016 IEEE 55th Conference on Decision and Control (CDC)*, 2016.

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

He Yin, Peter Seiler, Ming Jin, and Murat Arcak. Imitation learning with stability and safety guarantees. *arXiv preprint arXiv:2012.09293*, 2020.