Closing the Closed-Loop Distribution Shift in Safe Imitation Learning

Stephen Tu[1], Alexander Robey[2], and Nikolai Matni[2]

[1]Google Brain Robotics
[2]Department of Electrical and Systems Engineering, University of Pennsylvania

February 19, 2021

Abstract

Commonly used optimization-based control strategies such as model-predictive and control Lyapunov/barrier function based controllers often enjoy provable stability, robustness, and safety properties. However, implementing such approaches requires solving optimization problems online at high-frequencies, which may not be possible on resource-constrained commodity hardware. Furthermore, how to extend the safety guarantees of such approaches to systems that use rich perceptual sensing modalities, such as cameras, remains unclear. In this paper, we address this gap by treating safe optimization-based control strategies as experts in an imitation learning problem, and train a learned policy that can be cheaply evaluated at run-time and that *provably satisfies the same safety guarantees as the expert*. In particular, we propose Constrained Mixing Iterative Learning (CMILe), a novel on-policy robust imitation learning algorithm that integrates ideas from stochastic mixing iterative learning, constrained policy optimization, and nonlinear robust control. Our approach allows us to control errors introduced by both the learning task of imitating an expert and by the distribution shift inherent to deviating from the original expert policy. The value of using tools from nonlinear robust control to impose stability constraints on learned policies is shown through sample-complexity bounds that are independent of the task time-horizon. We demonstrate the usefulness of CMILe through extensive experiments, including training a provably safe perception-based controller using a state-feedback-based expert.

1 Introduction

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

However, a key obstacle to deploying policies obtained through IL in real-world settings is guaranteeing safety. While the field of safe RL has seen an explosion of recent results (see e.g., Chow et al. (2018); Fulton and Platzer (2018); Alshiekh et al. (2018); Berkenkamp et al. (2017); Recht (2019); Matni et al. (2019) and references therein), there are relatively fewer results regarding safe IL. Methods based on post-hoc validation (Hertneck et al., 2018), quadratic constraints (Yin et al., 2020), and Bayesian networks (Lee et al., 2018; Menda et al., 2017, 2019) have been used to quantify varying degrees of safety guarantees in limited settings; however, to the best of our knowledge, no provably data-efficient and general method exists.

In this paper we address this gap, and revisit the imitation learning problem with an eye towards providing sample efficient algorithms that provide strong safety guarantees. In particular, we show that by combining ideas from stochastic mixing iterative learning (SMILe) (Ross et al., 2011; Ross and Bagnell, 2010), constrained policy optimization (Schulman et al., 2015), and robust nonlinear control (Angeli, 2002; Lohmiller and Slotine, 1998), such a goal can be met for a broad class of nonlinear dynamical systems being controlled by a general class of expert policies. Our analysis shows that through a careful integration of these tools, we can simultaneously control errors introduced by both the learning task of imitating an expert and by the distribution shift inherent to deviating from the original expert policy.

First, we present and analyze Constrained Mixing Learning (CMILe) (Algorithm 1), a broadly applicable IL algorithm for known nonlinear control affine systems. We prove that under suitable assumptions on the stability of the underlying expert policy, it suffices to collect $\tilde{O}(k/\gamma^2)$ trajectories to ensure that the safety guarantees of the expert policy are enjoyed by the learned policy, for γ the minimum safety-margin of the expert policy and k the effective number of parameters of the function class for the learned policy. Importantly, this bound is *independent* of the task time-horizon T. We then show how CMILe can be applied to a perception-based control IL problem where a perception-based policy is trained using a state-feedback expert policy. We conclude with experiments showcasing our method on: (1) a linear-time-invariant system controlled by a model-predictive-controller, (2) a nonlinear system controlled by a control Lyapunov/Barrier function based controller, and (3) a perception-based control IL task.

1.1 Related Work

Imitation Learning 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 (Ross et al., 2011; Pomerleau, 1989). Representative IL algorithms that address this issue include DAgger (Ross et al., 2011) and DART (Laskey et al., 2017). DAgger proposes an *on-policy* approach, in which the system iteratively augments its data-set of trajectories with those of the previous policy, appropriately labeled and/or corrected by a supervisor. For strongly convex loss functions, DAgger enjoys $\tilde{O}(T)$ sample-complexity in the task horizon T. DART, in contrast is an *off-policy* approach, which injects noise into the supervisor demonstrations and allows the supervisor to provide corrections as needed, thus focusing on states the system may need to recover from at deployment which may lie at the boundary of the ideal policy. To the best of our knowledge, no finite-data guarantees exist for DART. We address the same challenge of covariate shift for systems operating in continuous state and action spaces, and seek to provide finite-sample guarantees under which stability and safety-constraints are maintained. However, in contrast to the aforementioned works, we assume that the system dynamics are known, which allows for stronger guarantees to be made. In particular, we show that by imposing stability constraints on the deployed policies, we can obtain task horizon *independent* sample-complexity guarantees.

Safe Imitation Learning. In Lee et al. (2018); Menda et al. (2017, 2019), variants of Bayesian networks (ensembles, dropout) are used to provide an estimate of uncertainty associated with the learned intermediate (novice) policies, which in turn are used to modify the weight put on the novice policies during RL (Lee et al., 2018) or IL (Menda et al., 2017, 2019) training. These methods are unable to guarantee safety of the resulting learned policies, nor provide sample-complexity guarantees. More closely related to our work, in Hertneck et al. (2018), a supervised learning framework is proposed to approximate a model-predictive-controller (MPC) by combining tools from robust MPC design with statistical learning bounds. While similar in spirit to our proposed method, their statistical guarantees require *post-hoc* sampling of the learned controller to estimate failure rates, which may require an iterative loop of resampling and retraining, and is only applicable to MPC based experts. Our method is applicable to a broader class of expert policies, and does not require post-hoc validation: we apply our method to an MPC expert policy in §6. In Yin et al. (2020), an IL method is presented to synthesize a neural network controller with stability and safety guarantees. Guarantees are ensured through the use of Lyapunov and local quadratic constraints, but only apply to linear time-invariant systems, and no sample-complexity guarantees are given.

Constrained Policy Optimization. Our approach borrows ideas from constrained policy optimization of a surrogate lower-bound cost function. Examples include Trust Region Policy Optimization (TRPO) (Schulman et al., 2015) (and the algorithms it generalizes, e.g., natural policy gradient (Kakade, 2002), and relative entropy policy search (Peters et al., 2010)), and Stochastic Lower Bounds Optimization (SLBO) (Luo et al., 2019) (which is closely related to Farahmand et al. (2017)). At iteration $k+1$, both TRPO and SLBO derive lower-bounds for the reward with respect to the distribution induced by the previous policy π_k by enforcing that the next policy π_{k+1} remain within a neighborhood of π_k, and maximize this lower-bound. To the best of our knowledge, this idea has not been explicitly[1] applied in the context of IL to limit distribution shift and ensure that the safety-guarantees of the expert are preserved.

Robust regression and domain adaptation. In traditional machine learning, designing predictors that perform well on test data that is drawn from a distribution different than that from which the training data is drawn is known as the domain adaptation problem (Daumé III, 2009; Pan et al., 2010). Deep learning is especially brittle to such domain shifts, as highlighted in (Madry et al., 2018; Goodfellow et al., 2015), but domain adaptation (Glorot et al., 2011), data augmentation (Ganin and Lempitsky, 2015), and robust regression methods (Zheng et al., 2016), for example, offer remedies.

In the context of learning models of dynamical systems, in (Liu et al., 2020) and their follow up work (Nakka et al., 2020), the authors consider the problem of safe learning and exploration in sequential control problems. Their goal is to safely collect data samples by operating in an environment and subsequently to learn to solve a challenging control problem. To do so, they propose a deep robust regression method to directly predict the uncertainty bounds needed for safe exploration. Further, they propose a control policy that depends on the learned uncertainty, and

[1] We note that SMILe and DAgger (Ross et al., 2011; Ross and Bagnell, 2010) can be viewed as implicitly imposing such constraints by blending learned and expert policies during data collection.

assume a bound on the true density ratio between the target (test) and source (training) trajectory distributions, allowing for guarantees on safe learning to be derived. This density ratio is assumed to be estimated empirically (Nakka et al., 2020), or obtained by leveraging a priori knowledge about the dynamics of the system and designed training and test trajectories (Liu et al., 2020). The problem we consider is different in nature, but similar in spirit, to those of (Liu et al., 2020; Nakka et al., 2020), and we believe that our approach may suggest appealing alternatives that do not require *a priori bounding* of the density ratio of target and source trajectory distributions – we leave investigating the implications of our approach on safe exploration to future work.

2 Problem Formulation

Consider a nonlinear control-affine dynamical system

$$x_{t+1} = f(x_t) + g(x_t)u_t, \tag{2.1}$$

where $u_t \in \mathbb{R}^q$ is the control input, $x_t \in \mathbb{R}^p$ is the state and is fully observed, and f and g are known functions. We use the notation $\varphi_t^\pi(\xi)$ to denote the value of x_t produced by system (2.1) when starting from initial condition $x_0 = \xi$ and the control input is set to the state-feedback control policy π, i.e., $u_t = \pi(x_t)$. Let $X \subseteq \mathbb{R}^p$ be a compact set and let $T \geqslant 1$ be the time-horizon over which trajectories $\varphi_t^\pi(\xi)$ are considered. We assume that trajectories are generated by drawing random initial conditions. In particular, let $\mathcal{D}$ be a distribution over X, and let $\{\xi_i\}_{i=1}^n \sim \mathcal{D}^n$: then we are give access to $\{\varphi_0^\pi(\xi_i), \varphi_1^\pi(\xi_i), \ldots, \varphi_{T-1}^\pi(\xi_i)\}_{i=1}^n$.

Furthermore, we assume that we are provided with an expert policy $\pi_\star : \mathbb{R}^p \to \mathbb{R}^q$ that ensures that safety constraints, encoded through inequalities:

$$h_j(x_0, x_1, \ldots, x_{T-1}) \leqslant 0,\ \ j = 1, \ldots, J \tag{2.2}$$

are satisfied by all state/input trajectories $(\varphi_t^{\pi_\star}(\xi), \pi_\star(\varphi_t^{\pi_\star}(\xi)))$ for all initial conditions $\xi \in X$. Performance, stability, and safety constraints, as captured by Lyapunov and Barrier functions or polytopic constraints, are naturally encoded through constraints of the form (2.2). Furthermore, under the assumption of state-feedback control policies, input constraints can also be encoded through constraints (2.2) by writing $u_t = \pi(x_t)$.

We may now formally state the problem considered in this paper. Pick a failure probability $\delta \in (0, 1)$ and for any π_1, π_2, π', define the loss function:

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

Our goal is to learn a policy $\hat\pi$ using n trajectories of length T seeded from initial conditions $\{\xi_i\}_{i=1}^n \sim \mathcal{D}^n$ such that with probability at least $1 - \delta$, the learned policy $\hat\pi$ induces a state/input trajectory $\{\varphi_t^{\hat\pi}(\xi), \hat\pi(\varphi_t^{\hat\pi}(\xi))\}_{t=0}^{T-1}$ that

(i) imitates the expert policy $\pi_\star$, as measured by the loss function $\mathbb{E}_{\xi\sim\mathcal{D}} \ell_{\hat\pi}(\xi; \hat\pi, \pi_\star)$,

(ii) satisfies the safety constraints (2.2).

In §3 we introduce and analyze Constrained Mixing Iterative Learning (CMILe), an episodic on-policy robust imitation learning algorithm, and prove generalization bounds for policies learned using CMILe. In §4, we specialize these bounds to various families of policy function classes, and show that the imitation loss $\mathbb{E}_{\xi\sim\mathcal{D}} \ell_{\hat\pi}(\xi; \hat\pi, \pi_\star)$ decays as $\tilde{O}(\sqrt{k/n})$ for k the effective number of parameters of the function class for the learned policy. Before doing so we introduce necessary definitions and regularity assumptions.

2.1 Regularity Assumptions

We now introduce definitions of system, safety constraint, and policy function class regularity.

Definition 2.1. *System (2.1) is (L_f, L_g, B_f, B_g)-regular if: (i) f and g are L_f-Lipschitz and L_g-Lipschitz, respectively, (ii) $\sup_x \|f(x)\| \leqslant B_f$ and $\sup_x \|g(x)\| \leqslant B_g$, and (iii) $f(0) = 0$.*

Definition 2.2. *A policy class Π is (L_Π, B_Π)-regular if (i) Π is convex,[2] (ii) every $\pi \in \Pi$ is L_Π-Lipschitz, (iii) every $\pi \in \Pi$ satisfies $\sup_x \|\pi(x)\| \leqslant B_\Pi$, and (iv) every $\pi \in \Pi$, satisfies $\pi(0) = 0$.*

Definition 2.3. *A safety constraint (2.2) is (γ_h, L_h)-regular over X under a policy π if for $\xi \sim \mathcal{D}$ and all $t \geqslant 0$ (i) $h(\varphi_0^\pi(\xi), \varphi_1^\pi(\xi), \ldots, \varphi_{T-1}^\pi(\xi)) \leqslant -\gamma_h$ a.s. and (ii) $h(\Phi)$ is L_h-Lipschitz in $\Phi = (\varphi_0, \varphi_1, \ldots, \varphi_{T-1})$.*

Assumption 2.4. *We assume that the tuple $(f, g, \Pi, \pi_\star, h_x, h_u, \mathcal{D})$ satisfies:*

(a) The policy class Π is (L_Π, B_Π)-regular.
(b) The dynamics (f, g) are (L_f, L_g, B_f, B_g)-regular.
(c) The distribution $\mathcal{D}$ over initial conditions satisfies $\|\xi\| \leqslant B_0$ a.s. and $\xi \in X$ a.s. for $\xi \sim \mathcal{D}$.
(d) The safety constraints $\{h_j\}$ are (γ_h, L_h)-regular over X under $\pi_\star$.
(e) The expert policy $\pi_\star \in \Pi$.
(f) For $\pi_1, \pi_2 \in \Pi$, we assume Δ_{π_1, π_2} is L_Δ-Lipschitz.

While Assumptions 2.4(a)-(d) are commonly made regularity assumptions, Assumptions 2.4(e)-(f) warrant further justification. Assumption 2.4(e) is made to ensure feasibility of the first and last steps of our imitation learning algorithm, while Assumption 2.4(f) is guaranteed to hold under the regularity assumptions 2.4(a)-(d), with L_Δ being at most $2(B_\Pi L_g + 2B_g L_\Pi)$, but often much smaller. As we will show below, exact knowledge of L_Δ is not required.

2.2 Nonlinear Robust Stability

To bound deviations of state trajectories induced by the learned policy $\hat{\pi}$ and the expert policy $\pi_\star$ on system (2.1), we use the following generalization of input-to-state stability. This definition allows for direct comparison between two system trajectories in terms of the norm of previous inputs.

Definition 2.5 (cf. Angeli (2002); Boffi et al. (2020a))**.** *Let constants β, γ be positive and $\rho \in (0, 1)$. The discrete-time dynamical system $x_{t+1} = f(x_t)$ is (β, ρ, γ)-*exponentially-incrementally-input-to-state-stable *(E-δISS) for a pair of initial conditions (x_0, y_0) and signal u_t (which is possibly adapted to the history $\{x_s\}_{s \leqslant t}$) if the trajectories $x_{t+1} = f(x_t) + u_t$ and $y_{t+1} = f(y_t)$ satisfy for all $t \geqslant 0$:*

$$\|x_t - y_t\| \leqslant \beta \rho^t \|x_0 - y_0\| + \gamma \sum_{k=0}^{t-1} \rho^{t-1-k} \|u_k\| . \tag{2.3}$$

A system is (β, ρ, γ)-E-δISS if it is (β, ρ, γ)-E-δISS for all initial conditions (x_0, y_0) and signals u_t.

Let $\mathcal{S}_\delta(\beta, \rho, \gamma)$ denote the set of policies π such that the resulting closed-loop system $f_{\text{cl}}^\pi(x) := f(x) + g(x)\pi(x)$ is (β, ρ, γ)-E-δISS, and define $\Pi_\delta(\beta, \rho, \gamma) := \mathcal{S}_\delta(\beta, \rho, \gamma) \cap \Pi$. We now add a final assumption on the expert policy $\pi_\star$.

Assumption 2.6. *The expert policy $\pi_\star \in \Pi_\delta(\beta, \rho, \gamma)$.*

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

Algorithm 1 Constrained Mixing Iterative Learning (CMILe)

Input: Total trajectory budget n, number of epochs E that divides n, mixing rate $\alpha \in (0, 1]$, initial conditions $\left\{\{\xi_i^k\}_{i=1}^{n/E}\right\}_{k=0}^{E-1} \sim \mathcal{D}^n$, expert policy $\pi_\star$, stability parameters (β, ρ, γ), and non-negative scalars $\{c_i\}_{i=1}^{E-2}$.

1: $\pi_0 \leftarrow \pi_\star$, $c_0 \leftarrow 0$.
2: **for** $k = 0, \dots, E-2$ **do**
3: Collect trajectories $\left\{\{\varphi_t^{\pi_k}(\xi_i)\}_{t=0}^{T-1}\right\}_{i=1}^{n/E}$.
4: $\pi_k \leftarrow \mathrm{cERM}\left(\left\{\{\varphi_t^{\pi_k}(\xi_i)\}_{t=0}^{T-1}\right\}_{i=1}^{n/E}, \pi_k, c_k, 0\right)$.
5: $\pi_{k+1} \leftarrow (1-\alpha)\pi_k + \alpha\hat{\pi}_k$.
6: **end for**
7: Collect trajectories $\left\{\{\varphi_t^{\pi_{E-1}}(\xi_i)\}_{t=0}^{T-1}\right\}_{i=1}^{n/E}$.
8: $c_{E-1} \leftarrow \frac{(1-\alpha)^E}{\alpha} \frac{1}{n/E} \sum_{i=1}^{n/E} \ell_{\pi_{E-1}}(\xi_i^{E-1}; \pi_{E-1}, \pi_\star)$.
9: $\hat{\pi}_{E-1} \leftarrow \mathrm{cERM}\left(\left\{\{\varphi_t^{\pi_{E-1}}(\xi_i)\}_{t=0}^{T-1}\right\}_{i=1}^{n/E}, \pi_{E-1}, c_{E-1}, (1-\alpha)^E\right)$.
10: $\pi_E \leftarrow \frac{1}{1-(1-\alpha)^E}[(1-\alpha)\pi_{E-1} + \alpha\hat{\pi}_{E-1} - (1-\alpha)^E\pi_\star]$.
11: **return** π_E.

Algorithm 2 Constrained Emprical Risk Minimization $\mathrm{cERM}\left(\left\{\{\varphi_t^{\pi_{\mathrm{roll}}}(\xi_i)\}_{t=0}^{T-1}\right\}_{i=1}^{m}, \pi_{\mathrm{roll}}, c, w\right)$

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

1: **return** the solution to:

$$\mathrm{minimize}_{\bar{\pi}\in\Pi} \ \frac{1}{m}\sum_{i=1}^{m} \ell_{\pi_{\mathrm{roll}}}(\xi_i; \bar{\pi}, \pi_\star) \tag{3.1a}$$
$$\text{subject to } \ \frac{1}{m}\sum_{i=1}^{m} \ell_{\pi_{\mathrm{roll}}}(\xi_i; \bar{\pi}, \pi_{\mathrm{roll}}) \leqslant c, \tag{3.1b}$$
$$(1-w)^{-1}[(1-\alpha)\pi_{\mathrm{roll}} + \alpha\bar{\pi} - w\pi_\star] \in \Pi_\delta(\beta, \rho, \gamma). \tag{3.1c}$$

3 Constrained Mixing Iterative Learning

We propose Constrained Mixing Iterative Learning (CMILe) in Algorithm 1, which blends a deterministic variant of SMILe (Ross and Bagnell, 2010) (policy update step on Line 5) with constrained policy optimization approaches (subroutine Algorithm 2) such as TRPO (Schulman et al., 2015) and SLBO (Luo et al., 2019).

By further restricting ourselves to optimizing over policies that induce data-generation policies π_k that lead to E-δISS closed-loop systems, we are able to prove strong finite-sample guarantees that are *independent of the trajectory length T*. We describe how to search over stable policies in §4 by leveraging tools from (differential) Lyapunov stability theory (Angeli, 2002; Lohmiller and Slotine, 1998).

In order to simplify the statement of our results, we define $\bar{c}_k := \sum_{i=1}^{k-1} c_i$, and $\Gamma_{n,E,\delta}$ as the following expression:

$$\Gamma_{n,E,\delta} := 2\mathcal{R}_{n/E}(\Pi) + \tfrac{\beta B_0 L_\Delta}{1-\rho}\sqrt{\tfrac{E\log(4E/\delta)}{n}}\,, \tag{3.2}$$

where $\mathcal{R}_n(\Pi)$ is the Rademacher complexity of Π:

$$\mathcal{R}_n(\Pi) := \sup_{\pi_d \in \Pi_\delta(\beta,\rho,\gamma)} \sup_{\pi_g \in \Pi} \mathbb{E}_{\{\xi_i\},\{\varepsilon_i\}}\left[\sup_{\pi \in \Pi} \frac{1}{n}\sum_{i=1}^n \varepsilon_i \ell_{\pi_d}(\xi_i;\pi,\pi_g)\right]. \tag{3.3}$$

Above, the $\varepsilon_i \in \{\pm 1\}$ are independent Rademacher random variables, i.e., $\mathbb{P}(\varepsilon_i = 1) = \mathbb{P}(\varepsilon_i = -1) = 1/2$.

Theorem 3.1. *Let Assumptions 2.4 and 2.6 hold. Further assume that* $\alpha E \geqslant \log(2)$, $\frac{(1-\alpha)^E}{\alpha} \leqslant 1$, $T \geqslant \frac{\beta B_0 L_\Delta}{2(1-\rho)B_g B_\Pi}$, *and without loss of generality, let* $\min\{B_g, B_\Pi, \gamma\} \geqslant 1$ *and* $\gamma L_\Delta \geqslant 2(1-\rho)$. *Then with probability at least* $1-\delta$ *over* $\left\{\{\xi_i^k\}_{i=1}^{n/E}\right\}_{k=0}^{E-1} \sim \mathcal{D}^n$, *we have that Algorithm 1 is feasible across all* E *epochs, and that the following are true:*

- *We have that for* $k = 1, \ldots, E-1$

$$\mathbb{E}_{\xi\sim\mathcal{D}}\ell_{\pi_k}(\xi;\pi_k,\pi_\star) \leqslant 2\alpha\left(\tfrac{\gamma L_\Delta}{1-\rho}\right)\left(k\Gamma_{n,E,\delta} + \bar{c}_k\right). \tag{3.4}$$

- *The final policy* π_E *output by Algorithm 1 does not depend on the expert* $\pi_\star$ *and satisfies*

$$\mathbb{E}_{\xi\sim\mathcal{D}}\ell_{\pi_E}(\xi;\pi_E,\pi_\star) \leqslant 18\alpha\left(\tfrac{\gamma L_\Delta}{1-\rho}\right)^2\left(E\Gamma_{n,E,\delta} + \bar{c}_{E-1}\right). \tag{3.5}$$

If we further assume that

$$\alpha E \geqslant \log\left(2 + \tfrac{2L_\Delta\gamma}{1-\rho}\right), \tag{3.6}$$

then π_E *also satisfies*

$$\mathbb{E}_{\xi\sim\mathcal{D}}\ell_{\pi_E}(\xi;\pi_E,\pi_\star) \leqslant 12\alpha\left(\tfrac{\gamma L_\Delta}{1-\rho}\right)\left(E\Gamma_{n,E,\delta} + \bar{c}_{E-1}\right). \tag{3.7}$$

The benefit of multiple epochs. To highlight the benefits of using multiple epochs (i.e., setting $\alpha < 1$ and $E > 1$), we compare the scaling of bound (3.5), which is valid when a single epoch is used, i.e., when $E\alpha = 1 \geqslant \log(2)$, to that of bound (3.7), which may not be valid when a single epoch is used (as we assume $L_\Delta\gamma/(1-\rho) \geqslant 2$). To do so, we assume[3] that $\mathcal{R}_{n/E}(\Pi) \leqslant O(\sqrt{E/n})$ and focus on scaling with respect to the mixing parameter α, the number of epochs E, and the stability parameter ρ. Thus we have that bound (3.5) scales as $O\left(\frac{E^{3/2}\alpha}{(1-\rho)^3}\right)$ and that bound (3.7) scales as $O\left(\frac{E^{3/2}\alpha}{(1-\rho)^2}\right)$. When a single epoch is used ($E = \alpha = 1$), bound (3.5) scales as $O((1-\rho)^{-3})$. In contrast, condition (3.6) is not satisfied by $E = \alpha = 1$; rather, if $\alpha = O(1)$ then we may choose $E = \Theta\left(\log(\frac{1}{1-\rho})\right)$. With such a choice of parameters, bound (3.7) scales as $\tilde{O}((1-\rho)^{-2})$, where we suppress polylogarithimic factors of $1/(1-\rho)$. Thus, up to polylogarithmic factors, multiple epochs can improve the dependency of our bounds by a factor of $1/(1-\rho)$.

[3]We derive explicit expressions for $\mathcal{R}_{n/E}(\Pi)$ in §3.2 – when these expressions also depend on $1/(1-\rho)$, they do not change the relative advantages of using multiple epochs described here.

Optimal choice of $\{c_k\}$. The above results and their proofs make clear that the optimal choice of $\{c_k\}$ is $c_k \equiv 0$, i.e., at iteration $k+1$ of the algorithm, it is optimal for $\hat{\pi}_{k+1}$ to *interpolate* the previously played policy π_k over the data $\left\{\{\varphi_t^{\pi_k}(\xi_i^k)\}_{i=1}^{n/E}\right\}_{t=0}^{T-1}$. However, in practice exactly interpolating the data may not be possible. Our results show that using any sequence of constants $\{c_k\}$ such that $\sum_{i=1}^{E-2} c_i$ does not alter the leading order of convergence will not affect the order-wise scaling of our bounds. For example, if $c_i = c^i\sqrt{1/n}$, for some $c \in (0,1)$ then $\bar{c}_k \leqslant \frac{c}{1-c}\sqrt{1/n}$.

3.1 Guaranteeing Safety

We now show how Theorem 3.1 allows us to bound the probability that the final deployed policy $\bar{\pi}_E$, as well as all intermediate policies $\pi_1, \ldots, \pi_E$, satisfy the safety constraints (2.2).

Proposition 3.2. *Fix $\pi \in \Pi$ and let Assumption 2.6 hold. Then with probability at least $1-\delta$, for all safety constraints h that are (γ_h, L_h)-regular over X under $\pi_\star$:*

$$h(\{\varphi_t^{\pi}(\xi)\}_{t=0}^{T-1}) \leqslant -\gamma_h + \tfrac{L_h\gamma}{\delta(1-\rho)}\mathbb{E}_{\xi\sim\mathcal{D}}\ell_\pi(\xi;\pi,\pi_\star)\,. \tag{3.8}$$

Therefore if additionally the assumptions of Theorem 3.1 hold and $aE \geqslant \log\left(2+\frac{2L_\Delta\gamma}{1-\rho}\right)$, then with probability at least $1-\delta$ over $\left(\left\{\{\xi_i^k\}_{i=1}^{n/E}\right\}_{k=0}^{E-1}, \xi\right) \sim \mathcal{D}^{n+1}$, for all $j = 1, ..., J$,

$$h_j(\{\varphi_t^{\pi_E}(\xi)\}_{t=0}^{T-1}) \leqslant -\gamma_h + \tfrac{24\alpha\gamma^2 L_h L_\Delta}{\delta(1-\rho)^2}\left(E\Gamma_{n,E,\delta} + \textstyle\sum_{i=1}^{E-2} c_i\right). \tag{3.9}$$

Proof. For a policy $\tilde{\pi} \in \Pi$, let $\Phi^{\tilde{\pi}}(\xi) := (\varphi_0^{\tilde{\pi}}(\xi), \varphi_1^{\tilde{\pi}}(\xi), ..., \varphi_{T-1}^{\tilde{\pi}}(\xi))$ denote a trajectory starting from ξ and rolled out under $\tilde{\pi}$. By the assumption that h is (γ_h, L_h)-regular:

$$h(\Phi^{\pi}(\xi)) \leqslant h(\Phi^{\pi_\star}(\xi)) + L_h\|\Phi^{\pi}(\xi) - \Phi^{\pi_\star}(\xi)\| \leqslant -\gamma_h + L_h\sum_{t=0}^{T-1}\|\varphi_t^{\pi}(\xi) - \varphi_t^{\pi_\star}(\xi)\|\,.$$

Write $f_{\mathrm{cl}}^{\pi}(x) = f_{\mathrm{cl}}^{\pi_\star}(x) + \Delta_{\pi,\pi_\star}(x)$. Because $f_{\mathrm{cl}}^{\pi_\star}$ is (β,ρ,γ)-E-δISS, for any ξ, t we have:

$$\|\varphi_t^{\pi}(\xi) - \varphi_t^{\pi_\star}(\xi)\| \leqslant \gamma\sum_{s=0}^{t-1}\rho^{t-1-s}\|\Delta_{\pi,\pi_\star}(\varphi_s^{\pi}(\xi))\|\,.$$

Summing the left-hand-side from $t=0$ to $t=T-1$,

$$\sum_{t=0}^{T-1}\|\varphi_t^{\pi}(\xi) - \varphi_t^{\pi_\star}(\xi)\| \leqslant \frac{\gamma}{1-\rho}\ell_\pi(\xi;\pi,\pi_\star)\,.$$

By Markov's inequality, with probability at least $1-\delta$,

$$\ell_\pi(\xi;\pi,\pi_\star) \leqslant \frac{1}{\delta}\mathbb{E}_{\xi\sim\mathcal{D}}\ell_\pi(\xi;\pi,\pi_\star)\,.$$

Equation (3.8) follows from chaining the previous inequalities together. Equation (3.9) follows from (3.8), Theorem 3.1, and a union bound. $\square$

We note that safety of all intermediate policies $\pi_0, \ldots, \pi_{E-1}$ can be shown using a similar approach.

3.2 The Importance of Stability

We show the importance of imposing E-δISS stability constraints on the data generating policies $\pi_0, \ldots, \pi_E$. In particular, this leads to Rademacher complexity bounds that are independent of the horizon T for general parametric function classes with norm-bounded parameters. In contrast, without stability constraints, the corresponding complexity bounds would scale *exponentially* in the time horizon T.

Proposition 3.3. *Let $\Pi = \{\pi(x,\theta) : \theta \in \mathbb{R}^k, \|\theta\| \leqslant B_\theta\}$ for a fixed, twice continuously differentiable map π. Define*

$$L_{\partial^2\pi} := \sup_{\|x\| \leqslant \beta B_0, \|\theta\| \leqslant B_\theta} \left\| \tfrac{\partial^2 \pi}{\partial\theta \partial x} \right\|_{\ell^2(\mathbb{R}^k) \to M(\mathbb{R}^{q \times p})}$$

where $M(\mathbb{R}^{q\times p})$ denotes the Banach space of $q \times p$ real matrices equipped with the operator norm. Then:

$$\mathcal{R}_n(\Pi) \leqslant 32.5 B_g B_\theta L_{\partial^2\pi} \tfrac{B_0 \beta}{1-\rho} \sqrt{\tfrac{k}{n}}\,.$$

Proposition 3.3 shows that stability of the expert and learned policies is key to controlling the Rademacher complexity $\mathcal{R}_n(\Pi)$ of the policy class Π, and when combined with Theorem 3.1 and Proposition 3.2 yields end-to-end sample-complexity bounds on the imitation loss and degradation of the safety margin scaling as $\tilde{O}(\alpha E^{\frac{3}{2}} \sqrt{\frac{k}{n}})$. Full expressions for these bounds, as well as analogous bounds for vector-valued Reproducing Kernel Hilbert Spaces that that scale as $\tilde{O}(\alpha E^{\frac{3}{2}} \sqrt{T \frac{\min\{p,q\}}{n}})$, can be found in Appendix B.

4 Enforcing Exponential δISS

We propose two approaches to imposing the stability constraint (3.1c) in Algorithm 2. We note that all other aspects of Algorithm 1 can be implemented in a straightforward manner. Recall that $f_{\mathrm{cl}}^{\pi}(x) = f(x) + g(x)\pi(x)$, and define a region of attraction $X_{\mathrm{ROA}} \subseteq \mathbb{R}^p$ where stability constraints are enforced, for $X_{\mathrm{ROA}} \supseteq X$ a compact set.

4.1 Contraction Metrics and Exponential δISS

Our first approach is to utilize a key result that states that enforcing $f_{\mathrm{cl}}^{\pi_k}$ to be *contracting* implies exponential δISS.

Definition 4.1 (cf. Lohmiller and Slotine (1998))**.** *The discrete-time dynamical system $f(x)$ is contracting with rate $\gamma \in (0,1)$ in the metric $M(x)$ if for all x,*

$$\tfrac{\partial f}{\partial x}(x)^{\mathsf{T}} M(f(x)) \tfrac{\partial f}{\partial x}(x) \preccurlyeq \gamma M(x)\,.$$

Proposition 4.2 (Proposition 5.3 of Boffi et al. (2020a))**.** *Let $f(x)$ be contracting with rate γ in the metric $M(x)$. Assume that for all x we have $\mu I \preccurlyeq M(x) \preccurlyeq LI$. Then $f(x)$ is $(\sqrt{L/\mu}, \sqrt{\gamma}, \sqrt{L/\mu})$-E-$\delta$ISS.*

Thus, we can replace constraint (3.1c) in optimization problem (3.1) with a search for a metric M satisfying Definition 4.1 and the assumptions of Proposition 4.2. In particular, let $\pi_{k+1} := (1-\alpha)\pi_k + \alpha\bar{\pi}$, and add the constraint

$$\frac{\partial f_{\mathrm{cl}}^{\pi_{k+1}}(x)}{\partial x}^{\mathsf{T}} M\left(f_{\mathrm{cl}}^{\pi_{k+1}}(x)\right) \frac{\partial f_{\mathrm{cl}}^{\pi_{k+1}}(x)}{\partial x} \preccurlyeq \rho^2 M(x), \tag{4.1}$$

for all $x \in X_{\mathrm{ROA}} \subseteq \mathbb{R}^p$. The resulting metric M then certifies (β,ρ,γ)-E-δISS of the closed loop system $f_{\mathrm{cl}}^{\pi_{k+1}}$ over the region of attraction X_{ROA}. Further, if the expert policy $f_{\mathrm{cl}}^{\pi_\star}$ is contracting with rate ρ under some metric M, then following a similar argument as that of Theorem 3.1, we can show recursive feasibility of Algorithm 1 with the contraction constraint.

4.2 Lyapunov Exponential ISS

We next show that if policies satisfy a less restrictive notion of stability that can be certified via a Lyapunov inequality, then we can guarantee that all played policies result in closed systems that are E-δISS at the expense of poorer dependencies on system stability parameters.

Definition 4.3 (cf. Boffi et al. (2020a))**.** *Let constants β, γ be positive and $\rho \in (0,1)$. The discrete-time dynamical system $x_{t+1} = f(x_t)$ is (β,ρ,γ)*-exponentially-input-to-state-stable *(E-ISS) for an initial condition x_0 and a signal u_t (which is possibly adapted to the history $\{x_s\}_{s\leqslant t}$) if the trajectory $x_{t+1} = f(x_t) + u_t$ satisfies for all $t \geqslant 0$:*

$$\|x_t\| \leqslant \beta\rho^t\|x_0\| + \gamma\sum_{k=0}^{t-1}\rho^{t-1-k}\|u_k\|\,. \tag{4.2}$$

A system is (β,ρ,γ)-E-ISS if it is (β,ρ,γ)-E-ISS for all initial conditions x_0 and signals u_t.

The following lemma shows that (β,ρ,γ)-E-ISS can be enforced through Lyapunov certificates.

Lemma 4.4 (Proposition G.2 of Boffi et al. (2020a))**.** *Consider a dynamical system $f(x)$ with $f(0) = 0$. Suppose that $V(x)$ is a Lyapunov function satisfying $k_1\|x\|^2 \leqslant V(x) \leqslant k_2\|x\|^2$ for some positive k_1, k_2 for all x, that $V(x(t+1)) \leqslant \rho^2 V(x(t))$ for all x, t and some $\rho \in (0,1)$, and that $\nabla V(x)$ is L_V-Lipschitz. Then the system defined by $f(x)$ is $(\sqrt{k_2/k_1}, \rho, L_V/(2k_1))$-E-ISS.*

Bootstrapping this weaker notion of stability into the desired E-δISS requires a notion of discrete-time stability for linear systems, which can be certified for the linearization of f about the origin via the same Lyapunov function.

Definition 4.5. *A matrix $A \in \mathbb{R}^{p\times p}$ is (β,ρ) discrete-time stable for some $\beta \geqslant 1$ and $\rho \in (0,1)$ if $\|A^t\| \leqslant \beta\rho^t$ for all $t \geqslant 0$.*

Lemma 4.6. *Let the system $f(x)$ be (β,ρ,γ)-E-ISS, and assume that f is continuously differentiable with respect to x over an open-set containing the origin. Then the matrix $A_0 := \frac{\partial f}{\partial x}(0)$ is a (β,ρ) discrete-time stable matrix.*

Proposition 4.7. *Suppose that Assumption 2.4 holds. Assume that the expert closed-loop system $f_{\mathrm{cl}}^{\pi_\star}$ is (β,ρ,γ)-E-ISS and that $\frac{\partial f_{\mathrm{cl}}^{\pi}}{\partial x}$ is L-Lipschitz for all $\pi \in \Pi$. Suppose we run Algorithm 1 with the constraint in (3.1c) replaced with a (β,ρ,γ)-E-ISS constraint, and with the additional restrictions*

$$\alpha \leqslant \tfrac{(1-\rho)^2}{16B_gB_\Pi\beta L\gamma}, \quad \alpha E \geqslant \log\left(2 + \tfrac{16B_gB_\Pi\beta L\gamma}{(1-\rho)^2}\right). \tag{4.3}$$

Then the conclusions of Theorem 3.1 hold with $(\beta, \rho, \gamma) \leftarrow (\beta', \rho', \gamma')$ *where*

$$\beta' = \beta \exp\left(\tfrac{B_0 \beta^2 L}{1-\rho}\right) \,,\ \rho' = e^{-(1-\rho)/2} \,, \tag{4.4}$$

$$\gamma' = \beta \exp\left(\tfrac{L\beta^2}{1-\rho}\left(\beta B_0 + \tfrac{\gamma D}{1-\rho}\right)\right) \,. \tag{4.5}$$

Thus, we suggest replacing constraint (3.1c) in optimization problem (3.1) to search for a Lyapunov function $V \in \mathcal{V}$ satisfying the assumptions of Lemma 4.4. As before, let $\pi_{k+1} := (1-\alpha)\pi_k + \alpha\bar{\pi}$, and add the constraint

$$V(f_{\text{cl}}^{\pi_{k+1}}(x)) \leqslant \rho^2 V(x) \ \forall x \in X_{\text{ROA}} \subseteq \mathbb{R}^p. \tag{4.6}$$

that certifies (β, ρ, γ)-E-ISS of the next policy to be played π_{k+1} over the region of attraction X_{ROA}. If the expert closed-loop $f_{\text{cl}}^{\pi_\star}$ is (β, ρ, γ)-E-ISS stable, for suitably specified (β, γ), then following a similar argument as that of Theorem 3.1, we can show recursive feasibility Algorithm 1.

Altogether, this analysis shows that an E-ISS system is also E-δISS on a compact set of initial conditions, but only for sufficiently small inputs. This insight allows us to derive analogous results to those in Section 4.1 (see Appendix 4 for details); however, the resulting E-δISS parameters (β, ρ, γ) have a hidden exponential dependence on the original problem constants. We leave resolving this issue to future work, but comment that in practice, such exponential dependencies are not observed (see §6).

4.3 Enforcing Stability in Practice

We note that in practice, enforcing the (differential) Lyapunov inequality constraints uniformly over X_{ROA} is difficult. To address this, we introduce the only approximation between our analysis and practical implementation, and restrict enforcing the (differential) Lyapunov inequalities (4.1) or (4.6) to the sample trajectories $\{\{\varphi_t^{\pi_k}(\xi_i^k)\}_{t=0}^{T-1}\}_{i=1}^{n/E}$. We remark that by combining the results of Boffi et al. (2020b) with those of this paper, constant probability guarantees of validity can be provided for a (differential) Lyapunov certificate that only holds over sample trajectories. However, to bootstrap such guarantees across multiple epochs, high-probability guarantees are needed. We leave closing this gap to future work.

5 Perception-Based Control

We introduce a simple model for training a perception-based policy. In particular, we follow the formalism adopted in Dean et al. (2020a); Dean and Recht (2020); Dean et al. (2020b) and assume a deterministic generative model of the form

$$z(x_t) = I(x_t),\ I(0) = 0, \tag{5.1}$$

for $I : \mathbb{R}^p \to \mathcal{I} := I(\mathbb{R}^p) \subseteq \mathbb{R}^D$, where I is an unknown but L_I-Lipschitz map, and typically $D \gg p$. We make the standard assumption (see Watter et al. (2015); Banijamali et al. (2018)) that the perceptual-measurement z is Markovian such that x can be uniquely recovered from z, e.g., when considering image-based control, z may correspond to a buffer of sequential image frames allowing for the full state x to be recovered. This in particular implies that I^{-1} exists, and we further assume that I^{-1} is $L_{I^{-1}}$-Lipschitz (i.e., we assume that I is invertible and bi-Lipschitz).

The generative model (5.1) is admittedly quite restrictive, in that it assumes a smooth one-to-one mapping between system states x and perceptual-measurements z. However we argue that it is nevertheless applicable in the controlled repetitive environments under which IL is often applied, and we further show first steps towards generalizing model (5.1) to accommodate non-determinism in Appendix D.

As before, we consider a state-feedback-based expert policy $\pi_\star(x)$. However, our goal now is to train a perception-based policy $\pi_z(z)$ through imitation learning. We assume that during training, both state and perceptual measurements are collected, i.e., Line 3 of Algorithm 2 now collects state/perceptual-measurement pair trajectories $\left\{\{\varphi_t^{\pi_k}(\xi_i^k), z(\varphi_t^{\pi_k}(\xi_i^k))\}_{t=0}^{T-1}\right\}_{i=1}^{n/E}$.

We propose two alternative approaches: end-to-end (E2E) and modular. We present results for the E2E architecture here, and defer the modular approach to Appendix D.

E2E Perception-Based IL. Our goal is to train a policy that directly maps the perceptual-measurement z to an action via a perception-based policy $\pi_z(z)$, i.e., $\pi_z : \mathcal{I} \to \mathbb{R}^q$. To that end, we define a class of perception-based policies $\pi_z \in \Pi_z$, for Π_z a (L_z, B_z)-regular policy class with respect to the perceptual-measurements z. Due to the generative model (5.1), Π_z implicitly defines a class of state-based policies by setting $z = I(x)$. Conversely, state-based policies also define perception-based policies via $x = I^{-1}(z)$, where we recall that $I^{-1}(\cdot)$ exists by the Markovian assumption on $z \in \mathcal{I}$. We also assume that $\pi_\star \circ I^{-1} \in \Pi_z$.

We consider optimization over perception-based policies $\pi_z \in \Pi_z$, and in particular, optimization problem (3.1) of Algorithm 2 now becomes:

$$\begin{array}{rl} \text{minimize}_{\bar{\pi}_z\in\Pi_z} & \frac{1}{m}\sum_{i=1}^m \ell_{\pi_{\text{roll}}}(\xi_i; \bar{\pi}_z, \pi_\star \circ I^{-1}) \\ \text{subject to} & \frac{1}{m}\sum_{i=1}^m \ell_{\pi_{\text{roll}}}(\xi_i; \bar{\pi}_z, \pi_{\text{roll}}) \leqslant c, \\ & \pi_{w,I,\star,\alpha} \in \mathcal{S}_\delta(\beta,\rho,\gamma). \end{array} \tag{5.2}$$

Here we define

$$\pi_{w,I,\star,\alpha} := (1-w)^{-1}[(1-\alpha)\pi_{\text{roll}} \circ I + \alpha\bar{\pi}_z \circ I - w\pi_\star],$$

and for policies $\pi_d, \pi_1, \pi_2 \in \Pi_z$, we overload the notation $\ell_{\pi_d}(\xi; \pi_1, \pi_2) = \sum_{t=0}^{T-1}\|\Delta_{\pi_1\circ I,\pi_2\circ I}(\varphi_t^{\pi_d\circ I}(\xi))\|$. Furthermore, if we set $\pi_0 = \pi_\star \circ I^{-1}$, then using the update rule for π_{k+1} defined in Line 5 of Algorithm 1 leads to recursive feasibility of the algorithm, i.e., $\{\pi_i\}_{i=0}^E \in \Pi_z$, and outputs a purely perception-based policy $\pi_E(z)$ that does not rely on the system state x (Lemma D.1).

Finally, under the generative model (5.1), we have that Π_z is $(L_z L_I, B_z)$-regular with respect to x.[4] One can similarly check that $\Delta_{\pi_1\circ I,\pi_2\circ I}(x)$ is L_Δ-Lipschitz with respect to x for some L_Δ under the assumed generative model (5.1) for $\pi_1, \pi_2 \in \Pi_z$.

It follows that Theorem 3.1 and Proposition 3.2 are applicable to learning a perception-based controller using a state-feedback expert $\pi_\star(x)$ under the generative model (5.1), and all that remains is to characterize the Rademacher complexity of Π_z:

$$\mathcal{R}_n(\Pi_z, I) := \sup_{\substack{\pi_d\in\Pi_z,\ \pi_g\in\Pi_z \\ \pi_d\circ I\in\mathcal{S}_\delta(\beta,\rho,\gamma)}} \mathbb{E}_{\{\xi_i\},\{\varepsilon_i\}}\left[\sup_{\pi_z\in\Pi_z}\frac{1}{n}\sum_{i=1}^n \varepsilon_i \ell_{\pi_d}(\xi_i; \pi_z, \pi_g)\right]. \tag{5.3}$$

[4]Note that as $\pi_\star \circ I^{-1} \in \Pi_z$ by assumption, we have that $L_z L_I \geqslant L_\star$, and $B_z \geqslant B_\star$ if $\pi_\star$ is $L_\star$-Lipschitz and $B_\star := \sup_x\|\pi_\star(x)\|$.

Proposition 5.1. *Let $\Pi_z = \{\pi(z,\theta) : \theta \in \mathbb{R}^K, \|\theta\| \leqslant B_\theta\}$ for a fixed, twice continuously differentiable map π. Define*

$$L_{\partial^2\pi_z} := \sup_{\|z\| \leqslant \beta B_0 L_I, \|\theta\| \leqslant B_\theta} \left\| \frac{\partial^2 \pi}{\partial\theta\partial z} \right\|_{\ell^2(\mathbb{R}^K) \to M(\mathbb{R}^{q\times D})} .$$

Then if the perceptual-measurements $z(t)$ follow the generative model (5.1), we have that

$$R_n(\Pi_z, I) \leqslant 32.5 B_\theta B_g L_{\partial^2\pi_z} L_I \tfrac{\beta B_0}{1-\rho} \sqrt{\tfrac{K}{n}}. \tag{5.4}$$

As in §3.2, we can combine Proposition 5.1 with Theorem 3.1 and Proposition 3.2 to obtain end-to-end sample-complexity bounds on the imitation loss and safety-margin degradation for E2E perception-based IL, where we now instead consider a $(L_z L_I, B_z)$-regular policy class.

Remark 5.2. *Modern convolutional neural networks are typically* overparameterized *such that $K \gg n$, rendering bound (5.4) vacuous. In order to circumvent this limitation, we propose using data-driven* mid-level representations *trained using supervised learning as described in Sax et al. (2018); Chen et al. (2020). In particular, we learn an embedding from perceptual-measurements (e.g., pixels) to feature space $z \mapsto \Psi(z)$, where $\Psi(z) \in \mathbb{R}^d$ and typically $d \ll D$. We use state/perceptual-measurement pairs generated under the expert policy to learn Ψ, and then fix these features for use by the learned E2E perception-based policy.*

6 Experiments

6.1 The impact of stability constraints

We first study the effects of explicitly constraining the played policies $\{\pi_i\}_{i=1}^E$ to be stabilizing through the use of Lyapunov stability constraints (Appendix C). To do so in a transparent manner, we consider linear quadratic (LQ) control of a linear system $x_{t+1} = Ax_t + Bu_t$, where $A \in \mathbb{R}^{p\times p}$ and $B \in \mathbb{R}^{p\times q}$, so that the resulting stability parameters can be easily quantified.

The LQ control problem can be expressed as

$$\begin{array}{ll} \text{minimize}_{\{x_t\},\{u_t\}} & \sum_{t=0}^{\infty} x_t^\mathsf{T} Q x_t + u_t^\mathsf{T} R u_t \\ \text{subject to} & x_{t+1} = Ax_t + Bu_t,\ x_0 = \xi. \end{array} \tag{6.1}$$

where $Q \in \mathbb{R}^{p\times p}$ and $R \in \mathbb{R}^{q\times q}$ are fixed cost matrices. The optimal policy is linear in the state x_t, i.e., $u_t = K_{\text{lqr}} x_t$, where $K_{\text{lqr}} \in \mathbb{R}^{q\times p}$ can be computed by solving a discrete-time algebraic Riccati equation (Zhou et al., 1996). For our study, we set the task horizon $T = 25$, the state $x_t \in \mathbb{R}^{10}$, the control input $u_t \in \mathbb{R}^4$, and we fix a randomly generated but unstable set of dynamics (A, B). We also set $R = I_4$, and $Q = \nu I_{10}$, and vary ν across three orders of magnitude: $\nu \in \{0.0001, 0.001, 0.01\}$. In the limit of $\nu \to 0$, the optimal LQ controller is the minimum energy stabilizing controller, whereas for larger ν, the optimal LQ controller balances between state-deviations and control effort.

In Fig. 1, we plot the median goal error $\|x_{T-1}\|$ achieved by policies learned via CMILe for different values of ν, both with and without Lyapunov stability constraints (4.6), over 100 i.i.d. test rollouts. The error bars represent the 20th/80th percentiles of the median across ten independent trials. The policy class Π consists of two hidden-layer MLPs with ReLU activations and hidden width 64. We construct a valid robust Lyapunov certificate from the solution to the Riccati Equation

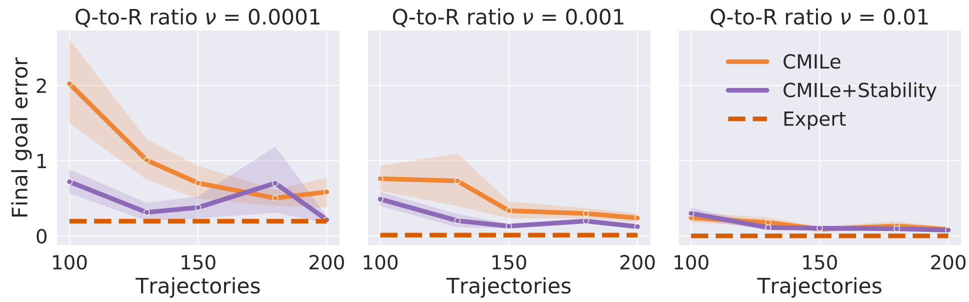

Figure 1: For fixed system matrices (A, B) and cost matrices $R = I_q$ and $Q = \nu I_q$ for $\nu \in \{0.0001, 0.001, 0.01\}$, we show the median goal error obtained by CMILe with and without stability constraints over 100 i.i.d. test trajectories.

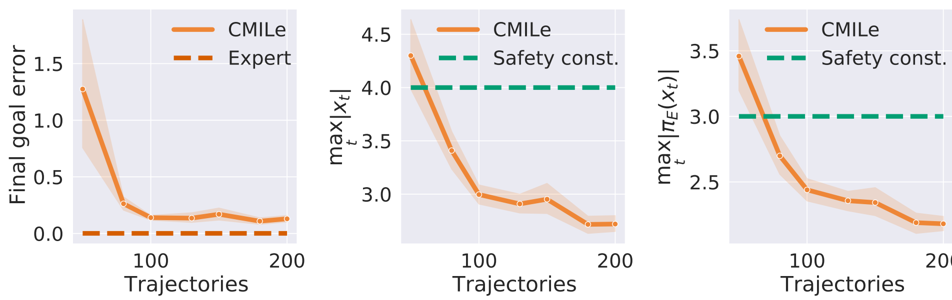

Figure 2: For the MPC problem described in (6.2), on the left we show the median final goal error obtained by CMILe over 100 i.i.d. test trajectories. Middle and right plots show the 90th percentile of $\max_t |x_t|$ and $\max_t |\pi_E(x_t)|$.

(see Appendix E for details), and use it to enforce (4.6) on the resulting closed loop dynamics $f_{cl}^{\pi_k}(x) = Ax + B\pi_k(x)$. Two important trends can be observed in Fig. 1. First, the smaller the weight ν, the more dramatic the effect of the stability constraint, i.e., when the underlying expert is itself fragile (stability radius $\rho \approx 1$), stability constraints are important. Second, the effect of stability constraints are more dramatic in low-data regimes, and by restricting $\pi_k \in \Pi_\delta(\beta, \rho, \gamma)$, we reduce over-fitting. We also evaluated the performance of standard behavior cloning (BC), but even with 200 training trajectories and $\nu = 0.01$, the median final goal error was 504.

6.2 Model-Predictive-Control

We consider a referencing-tracking MPC task in which we explicitly impose safety constraints on the state x_t and control input u_t. We again randomly generate a fixed pair of unstable linear-time-invariant dynamics with state $x_t \in \mathbb{R}^{10}$ and control input $u_t \in \mathbb{R}^4$. MPC is implemented by re-solving a constrained finite-horizon optimal-control problem at each time-step, initialized from the system's true state x_t, and applying the first control input $u_0^\star$ of the solution; through appropriate choices of terminal costs and constraints, MPC controllers can be shown to have strong guarantees of safety and stability, see e.g., Borrelli et al. (2017). For our problem, the resulting sub-problem solved at each time instance is of the form:

$$\begin{aligned} &\text{minimize}_{\{x_k\},\{u_k\}} \textstyle\sum_{k=0}^{H-1} \tilde{x}_k^\top Q \tilde{x}_k + u_k^\top R u_k + \tilde{x}_H^\top Q_H \tilde{x}_H \\ &\text{subject to} \quad\quad \tilde{x}_k = x_k - x_{\text{ref}} \\ &\quad\quad\quad\quad\quad\quad\quad x_{k+1} = A x_k + B u_k,\ x_k = x_t \\ &\quad\quad\quad\quad\quad\quad\quad x_{\min} \leqslant x_k \leqslant x_{\max},\ u_{\min} \leqslant u_k \leqslant u_{\max}. \end{aligned} \tag{6.2}$$

In our study, we set $Q = Q_H = 10 I_{10}$ and $R = 0.01 I_4$. We use a task horizon of $T = 15$, an MPC look-ahead horizon of $H = 10$, and use the state and input constraints $-4 \leqslant x_t \leqslant 4$ and $-3 \leqslant u_t \leqslant 3$.

In the left panel of Fig. 2, we plot the median final goal error $\|x_{T-1} - x_{\text{ref}}\|$ achieved by policies learned by CMILe over 100 i.i.d. test rollouts, where the error bars are the 20th/80th percentiles of the median goal error across ten independent trials. In the middle and rightmost panels of Fig. 2, we show the 90th percentile value of $\max_{0 \leqslant t \leqslant T-1} |x_t|$ and $\max_{0 \leqslant t \leqslant T-1} |\pi_E(x_t)|$ over 100 i.i.d. test rollouts. For these plots, the error bars again represent the 20th/80th percentile across ten trials. Observe that when the total number of trajectories is larger than 50, the safety constraints are satisfied with high-probability. We also evaluated the performance of BC, but even with 200 training trajectories, the median final goal error, state constraint satisfaction, and input constraint satisfaction achieved were given by 14.6, 10.3, and 6.5, respectively, well above the axes of the displayed results.

6.3 Segway dynamics

We next consider a Segway constrained to planar motion as described by Dean et al. (2020a). The resulting degrees of freedom are the horizontal position x, horizontal velocity $\dot{x}$, pitch angle θ, and pitch rate $\dot{\theta}$, which we collect as the state vector $z \in \mathbb{R}^4$. The expert controller is a PD controller tuned to drive the segway to its equilibrium position z_{goal} at the origin, filtered by a Control Barrier Function controller (CBF-QP, see §5 of Dean et al. (2020a)), which ensures that θ does not exceed 0.4 radians in either direction. We discretize the continuous time dynamics using RK4 with a time-step of 1ms; an inner-loop at this resolution is used for simulation, and an outer-loop updated every 0.02s is used for control (see Appendix E). We set the horizon to 5s, resulting in 250 data points per trajectory.

We first perform imitation learning assuming perfect state estimation. We compare the sample-complexity of BC with that of CMILe. For CMILe, we consider values of the mixing parameter $\alpha \in \{0.2, 0.3\}$, choosing the number of epochs E such that $(1-\alpha)^E \approx 0.0138$ is small enough. The policy class Π consists of two hidden-layer MLPs with ReLU activations and hidden width 64. The results are shown in Figs. 3 and 4. Each configuration of algorithm and number of trajectories is repeated for 10 trials.

In Fig. 3, we plot the median final goal error $\|z_{T-1} - z_{\text{goal}}\|$ over an evaluation set of 500 rollouts for each algorithm as a function of the number of training trajectories collected – the error bars represent the 20th/80th percentiles across 10 independent trials. The median final goal error of the expert is shown in the dashed green line. For both the BC algorithm and CMILe, the error decreases as more trajectories are collected. However, we see that CMILe enjoys dramatically better sample complexity. For instance, its performance with 120 trajectories at $\alpha = 0.2$ is comparable to the performance of BC with 200 trajectories. Furthermore, we see that even for a fixed amount of trajectories, the value of α (and hence the number of epochs) affects the final performance.

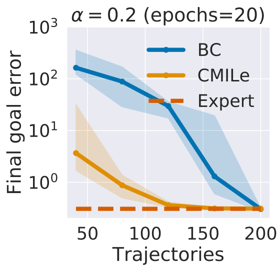

Figure 3: A comparison of the median final goal error performance of BC and CMILe on the Segway system for state-feedback policies. The median final goal error of the expert policy is shown in dashed red.

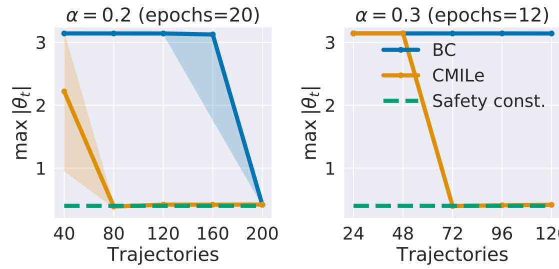

Figure 4: Safety constraint satisfaction for the Segway system using state-feedback policies. Each line is shaded according to the 90th percentile of $\max_t |\theta_t|$.

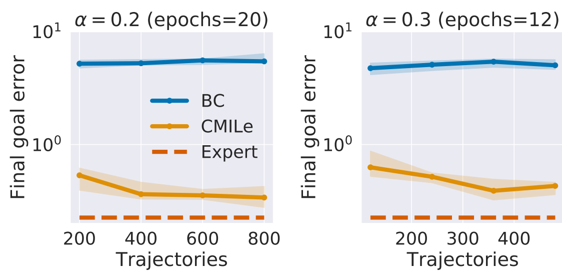

Figure 5: A comparison of BC and CMILe for learning perception-based policies on the Segway system. The median final goal error is shown for each method.

In Fig. 4, we plot the 90th percentile of $\max_{0 \leqslant t \leqslant T-1} |\theta_t|$ over 500 test rollouts. Again, the error bars represent the 20th/80th percentiles across 10 independent trials. Recall that the expert CBF-QP guarantees this quantity remains less than 0.4: we see that CMILe requires much fewer samples than BC to ensure the expert safety-constraint is (approximately) met.

We next perform imitation learning from pixels, where we learn a policy which maps a sequence of images to control actions. Here, the images given to the controller are a side view of a cartoon-representation of the Segway (see Fig. 6 in Appendix E). We train this policy in two steps. First, we collect 500 labelled expert trajectories (using an initial distribution $\widetilde{\mathcal{D}}$ with wider support for more diverse behaviors), and we use supervised learning to fit a simple convolutional neural network to predict both the position and velocity of the base of the segway and its neck from a sequence of 5 consecutive images in time. We then freeze the weights of the learned CNN and use them as *mid-level-representations*, i.e., we append a shallow MLP to the backend of the CNN as our policy. We then use both BC and CMILe to learn the weights of the backend MLP. The results are shown in Figure 5. We see that CMILe is able to successfully learn a perception-based policy with non-trivial performance relative to the expert, whereas BC struggles to generalize with the same amount of data. We believe the gap between the perception-based policy and the expert can be closed with more trajectories and more careful architecture selection.

7 Conclusion and Future Work

We introduced Constrained Mixing Iterative Learning (CMILe), a novel on-policy robust imitation learning algorithm that draws upon tools from stochastic mixing iterative learning, constrained policy optimization, and nonlinear robust control to provide end-to-end sample-complexity guarantees on both the imitation-loss and the stability-margin degradation of learned policies. We demonstrated the applicability of our approach, and its significant advantages over vanilla BC in terms of sample-complexity and safety-guarantees, in the context of a MPC task, a nonlinear control task, and a perception-based control task. For future work, we believe that our method can naturally be extended to accommodate both unknown/uncertain dynamics and driving noise (by using robust Lyapunov functions to enforce stability, for example), which will allow us to consider richer generative models in the context of perception-based IL. We also believe that the the results and proof techniques of this paper will provide a lens through which to study of the role of pre-training of mid-level representations and their effects on the sample-complexity of IL.

Acknowledgements

The authors would like to thank Vikas Sindhwani, Sumeet Singh, Andy Zeng, and Lisa Zhao for their valuable comments and suggestions. NM is generously supported by NSF awards CPS-2038873 and CAREER-2045834.

References

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.

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.

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.

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.

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.

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

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

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

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.

Yinlam Chow, Ofir Nachum, Edgar Duenez-Guzman, and Mohammad Ghavamzadeh. A lyapunov-based approach to safe reinforcement learning. In *Neural Information Processing Systems*, 2018.

Nathan Fulton and André Platzer. Safe reinforcement learning via formal methods. In *AAAI Conference on Artificial Intelligence*, 2018.

Mohammed Alshiekh, Roderick Bloem, Rüdiger Ehlers, Bettina Könighofer, Scott Niekum, and Ufuk Topcu. Safe reinforcement learning via shielding. In *AAAI Conference on Artificial Intelligence*, 2018.

Felix Berkenkamp, Matteo Turchetta, Angela Schoellig, and Andreas Krause. Safe model-based reinforcement learning with stability guarantees. In *Neural Information Processing Systems*, 2017.

Benjamin Recht. A tour of reinforcement learning: The view from continuous control. *Annual Review of Control, Robotics, and Autonomous Systems*, 2:253–279, 2019.

Nikolai Matni, Alexandre Proutiere, Anders Rantzer, and Stephen Tu. From self-tuning regulators to reinforcement learning and back again. In *2019 IEEE 58th Conference on Decision and Control (CDC)*, 2019.

Keuntaek Lee, Kamil Saigol, and Evangelos A. Theodorou. Safe end-to-end imitation learning for model predictive control. *arXiv preprint arXiv:1803.10231*, 2018.

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.

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

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

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

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

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

Sham M. Kakade. A natural policy gradient. In *Neural Information Processing Systems*, 2002.

Jan Peters, Katharina Mulling, and Yasemin Altun. Relative entropy policy search. In *AAAI Conference on Artificial Intelligence*, 2010.

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.

Amir-Massoud Farahmand, Andre Barreto, and Daniel Nikovski. Value-aware loss function for model-based reinforcement learning. In *International Conference on Artificial Intelligence and Statistics*, 2017.

Hal Daumé III. Frustratingly easy domain adaptation. *arXiv preprint arXiv:0907.1815*, 2009.

Sinno J. Pan, Ivor W. Tsang, James T. Kwok, and Qiang Yang. Domain adaptation via transfer component analysis. *IEEE Transactions on Neural Networks*, 22(2):199–210, 2010.

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

Ian J. Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples. In *International Conference on Learning Representations*, 2015.

Xavier Glorot, Antoine Bordes, and Yoshua Bengio. Domain adaptation for large-scale sentiment classification: A deep learning approach. In *International Conference on Machine Learning*, 2011.

Yaroslav Ganin and Victor Lempitsky. Unsupervised domain adaptation by backpropagation. In *International Conference on Machine Learning*, 2015.

Stephan Zheng, Yang Song, Thomas Leung, and Ian J. Goodfellow. Improving the robustness of deep neural networks via stability training. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

Anqi Liu, Guanya Shi, Soon-Jo Chung, Anima Anandkumar, and Yisong Yue. Robust regression for safe exploration in control. In *Learning for Dynamics and Control*, 2020.

Yashwanth K. Nakka, Anqi Liu, Guanya Shi, Anima Anandkumar, Yisong Yue, and Soon-Jo Chung. Chance-constrained trajectory optimization for safe exploration and learning of nonlinear systems. *arXiv preprint arXiv:2005.04374*, 2020.

Nicholas M. Boffi, Stephen Tu, and Jean-Jacques E. Slotine. Regret bounds for adaptive nonlinear control. *arXiv preprint arXiv:2011.13101*, 2020a.

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

Sarah Dean, Andrew J. Taylor, Ryan K. Cosner, Benjamin Recht, and Aaron D. Ames. Guaranteeing safety of learned perception modules via measurement-robust control barrier functions. In *Conference on Robot Learning*, 2020a.

Sarah Dean and Benjamin Recht. Certainty equivalent perception-based control. *arXiv preprint arXiv:2008.12332*, 2020.

Sarah Dean, Nikolai Matni, Benjamin Recht, and Vickie Ye. Robust guarantees for perception-based control. In *Learning for Dynamics and Control*, 2020b.

Manuel Watter, Jost Springenberg, Joschka Boedecker, and Martin Riedmiller. Embed to control: A locally linear latent dynamics model for control from raw images. In *Neural Information Processing Systems*, 2015.

Ershad Banijamali, Rui Shu, Mohammad Ghavamzadeh, Hung Bui, and Ali Ghodsi. Robust locally-linear controllable embedding. In *International Conference on Artificial Intelligence and Statistics*, 2018.

Alexander Sax, Bradley Emi, Amir R. Zamir, Leonidas Guibas, Silvio Savarese, and Jitendra Malik. Mid-level visual representations improve generalization and sample efficiency for learning visuomotor policies. *arXiv preprint arXiv:1812.11971*, 2018.

Bryan Chen, Alexander Sax, Gene Lewis, Iro Armeni, Silvio Savarese, Amir R. Zamir, Jitendra Malik, and Lerrel Pinto. Robust policies via mid-level visual representations: An experimental study in manipulation and navigation. In *Conference on Robot Learning*, 2020.

Kemin Zhou, John C. Doyle, and Keith Glover. *Robust and optimal control.* Prentice Hall, 1996.

Francesco Borrelli, Alberto Bemporad, and Manfred Morari. *Predictive control for linear and hybrid systems.* Cambridge University Press, 2017.

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

Charles A. Micchelli and Massimiliano Pontil. On learning vector-valued functions. *Neural Computation*, 17(1):177–204, 2005.

Andreas Maurer. A vector-contraction inequality for rademacher complexities. In *International Conference on Algorithmic Learning Theory*, 2016.

Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *International Conference on Learning Representations*, 2014.

Geir E. Dullerud and Fernando Paganini. *A course in robust control theory: a convex approach*, volume 36. Springer Science & Business Media, 2013.

Contents

A Proof of Theorem 3.1

A.1 Preliminary Results

We begin with an application of standard uniform convergence bounds to control deviations between empirical and population risks.

Proposition A.1. *Under Assumption 2.4, we have that:*

$$\sup_{\pi_d \in \Pi_\delta(\beta,\rho,\gamma)} \sup_{\pi_1,\pi_2 \in \Pi} \operatorname*{ess\,sup}_{\xi \sim \mathcal{D}} \ell_{\pi_d}(\xi; \pi_1, \pi_2) \leqslant B_\ell := \min\left\{ 2TB_g B_\Pi, \frac{\beta B_0 L_\Delta}{1-\rho} \right\} , \tag{A.1}$$

Now fix a data generating policy $\pi_d \in \Pi_\delta(\beta, \rho, \gamma)$ and goal policy $\pi_g \in \Pi$. Furthermore, let $\xi_1, ..., \xi_n$ be drawn i.i.d. from $\mathcal{D}$. With probability at least $1 - \delta$ (over $\xi_1, ..., \xi_n$), we have:

$$\sup_{\pi \in \Pi} \left| \mathbb{E}_{\xi\sim\mathcal{D}} \ell_{\pi_d}(\xi; \pi, \pi_g) - \frac{1}{n} \sum_{i=1}^n \ell_{\pi_d}(\xi_i; \pi, \pi_g) \right| \leqslant 2\mathcal{R}_n(\Pi) + B_\ell \sqrt{\frac{\log(2/\delta)}{n}} , \tag{A.2}$$

where we recall that $\mathcal{R}_n(\Pi)$ is defined as

$$\mathcal{R}_n(\Pi) = \sup_{\pi_d \in \Pi_\delta(\beta,\rho,\gamma)} \sup_{\pi_g \in \Pi} \mathbb{E}_{\{\xi_i\}} \mathbb{E}_{\{\varepsilon_i\}} \left[\sup_{\pi \in \Pi} \frac{1}{n} \sum_{i=1}^n \varepsilon_i \ell_{\pi_d}(\xi_i; \pi, \pi_g) \right] . \tag{A.3}$$

Proof. Since $\pi_1, \pi_2 \in \Pi$, we have $\Delta_{\pi_1,\pi_2}(0) = 0$. Therefore:

$$\ell_{\pi'}(\xi; \pi_1, \pi_2) = \sum_{t=0}^{T-1} \|\Delta_{\pi_1,\pi_2}(\varphi_t^{\pi'}(\xi))\| \leqslant L_\Delta \sum_{t=0}^{T-1} \|\varphi_t^{\pi'}(\xi)\| .$$

Because $f_{\mathrm{cl}}^{\pi'}$ is (β, ρ, γ)-E-δISS, for all $t \geqslant 0$:

$$\|\varphi_t^{\pi'}(\xi)\| \leqslant \beta\rho^t \|\xi\| \leqslant \beta B_0 \rho^t .$$

Hence,

$$\ell_{\pi'}(\xi; \pi_1, \pi_2) \leqslant L_\Delta \sum_{t=0}^{T-1} \beta B_0 \rho^t \leqslant \frac{\beta B_0 L_\Delta}{1-\rho} .$$

Furthermore, by our regularity assumptions we have that $\|\Delta_{\pi_1,\pi_2}(x)\| \leqslant 2B_g B_\Pi$ for all x. This yields (A.1). (A.2) now follows by a standard uniform convergence bound (cf. Chapter 4 of Wainwright (2019)). □

A.2 Proof of Theorem 3.1

Base case: $k = 0$ We first argue the base case. As $\pi_\star \in \Pi$ and $f_{\mathrm{cl}}^{\pi_\star}$ is (β, ρ, γ)-E-δISS by Assumption 2.6, we have that the optimization problem defining $\hat{\pi}_0$ is feasible. By Proposition A.1, there exists an event $\mathcal{E}_0$ such that $\mathbb{P}(\mathcal{E}_0) \geqslant 1 - \delta/E$ and on $\mathcal{E}_0$,

$$\begin{aligned}
\mathbb{E}_{\xi\sim\mathcal{D}}\ell_{\pi_0}(\xi;\hat{\pi}_0,\pi_\star) &\overset{(a)}{\leqslant} \frac{1}{n/E}\sum_{i=1}^{n/E}\ell_{\pi_0}(\xi_i^0;\hat{\pi}_0,\pi_\star) + \Gamma_{n,E,2\delta} \\
&\overset{(b)}{\leqslant} \frac{1}{n/E}\sum_{i=1}^{n/E}\ell_{\pi_0}(\xi_i^0;\pi_\star,\pi_\star) + \Gamma_{n,E,2\delta} \\
&\overset{(c)}{=} \Gamma_{n,E,2\delta} \,,
\end{aligned}$$

where the first inequality (a) follows from Proposition A.1 on event $\mathcal{E}_0$, the second inequality (b) by feasibility of $\pi_\star = \pi_0$ and optimality of $\hat{\pi}_0$ to optimization problem (3.1), and the final equality (c) from $\ell_{\pi'}(\xi;\pi,\pi) = 0$ for all π',π,ξ. This sequence of arguments will be used repeatedly in the sequel.

Our goal is to bound $\mathbb{E}_{\xi\sim\mathcal{D}}\ell_{\pi_1}(\xi;\pi_1,\pi_\star)$. We observe:

$$\begin{aligned}
\Delta_{\pi_1,\pi_\star}(x) &= g(x)(\pi_1(x) - \pi_\star(x)) \\
&= g(x)((1-\alpha)\pi_\star(x) + \alpha\hat{\pi}_0(x) - \pi_\star(x)) \\
&= \alpha g(x)(\hat{\pi}_0(x) - \pi_\star(x)) \\
&= \alpha\Delta_{\hat{\pi}_0,\pi_\star}(x) \,.
\end{aligned}$$

Therefore, by our assumption that $\Delta_{\hat{\pi}_0,\pi_\star}$ is L_Δ-Lipschitz:

$$\begin{aligned}
\mathbb{E}_{\xi\sim\mathcal{D}}\ell_{\pi_1}(\xi;\pi_1,\pi_\star) &= \mathbb{E}_{\xi\sim\mathcal{D}}\sum_{t=0}^{T-1}\|\Delta_{\pi_1,\pi_\star}(\varphi_t^{\pi_1}(\xi))\| = \alpha\mathbb{E}_{\xi\sim\mathcal{D}}\sum_{t=0}^{T-1}\|\Delta_{\hat{\pi}_0,\pi_\star}(\varphi_t^{\pi_1}(\xi))\| \\
&\leqslant \alpha\mathbb{E}_{\xi\sim\mathcal{D}}\sum_{t=0}^{T-1}\|\Delta_{\hat{\pi}_0,\pi_\star}(\varphi_t^{\pi_0}(\xi))\| + \alpha L_\Delta\mathbb{E}_{\xi\sim\mathcal{D}}\sum_{t=0}^{T-1}\|\varphi_t^{\pi_1}(\xi) - \varphi_t^{\pi_0}(\xi)\| \\
&= \alpha\mathbb{E}_{\xi\sim\mathcal{D}}\ell_{\pi_0}(\xi;\hat{\pi}_0,\pi_\star) + \alpha L_\Delta\mathbb{E}_{\xi\sim\mathcal{D}}\sum_{t=0}^{T-1}\|\varphi_t^{\pi_1}(\xi) - \varphi_t^{\pi_0}(\xi)\| \,.
\end{aligned}$$

Now we write:

$$\begin{aligned}
f_{\mathrm{cl}}^{\pi_1}(x) &= f(x) + g(x)((1-\alpha)\pi_\star(x) + \alpha\hat{\pi}_0(x)) \\
&= f_{\mathrm{cl}}^{\pi_\star}(x) + \alpha\Delta_{\hat{\pi}_0,\pi_\star}(x) \,,
\end{aligned}$$

Re-arranging, this shows that $f_{\mathrm{cl}}^{\pi_\star}(x) = f_{\mathrm{cl}}^{\pi_1}(x) - \alpha\Delta_{\hat{\pi}_0,\pi_\star}(x)$. Since $f_{\mathrm{cl}}^{\pi_1}$ is (β,ρ,γ)-E-δISS as a result of constraint (3.1c), we have that for all ξ

$$\|\varphi_t^{\pi_1}(\xi) - \varphi_t^{\pi_0}(\xi)\| \leqslant \gamma\sum_{s=0}^{t-1}\rho^{t-1-s}\|\alpha\Delta_{\hat{\pi}_0,\pi_\star}(\varphi_s^{\pi_0}(\xi))\| \,.$$

Summing the left-hand-side from $t = 0$ to $t = T-1$ yields

$$\sum_{t=0}^{T-1}\|\varphi_t^{\pi_1}(\xi) - \varphi_t^{\pi_0}(\xi)\| \leqslant \frac{\alpha\gamma}{1-\rho}\sum_{t=0}^{T-2}\|\Delta_{\hat{\pi}_0,\pi_\star}(\varphi_t^{\pi_0}(\xi))\| \leqslant \frac{\alpha\gamma}{1-\rho}\ell_{\pi_0}(\xi;\hat{\pi}_0,\pi_\star) \,.$$

Therefore, on the event $\mathcal{E}_0$,

$$\begin{aligned}
\mathbb{E}_{\xi\sim\mathcal{D}}\ell_{\pi_1}(\xi;\pi_1,\pi_\star) &\leqslant \alpha\left(1+\frac{\alpha L_\Delta\gamma}{1-\rho}\right)\mathbb{E}_{\xi\sim\mathcal{D}}\ell_{\pi_0}(\xi;\hat{\pi}_0,\pi_\star) \\
&\leqslant \alpha\left(1+\frac{\alpha L_\Delta\gamma}{1-\rho}\right)\Gamma_{n,E,2\delta} \\
&\leqslant \alpha\left(1+\frac{L_\Delta\gamma}{1-\rho}\right)\Gamma_{n,E,\delta} \\
&=: \alpha\beta_1(n,E,\delta)\,.
\end{aligned}$$

Induction step: We now assume that $1 \leqslant k \leqslant E-2$ and that the event $\mathcal{E}_{0:k-1} := \bigcap_{j=0}^{k-1}\mathcal{E}_j$ holds. Note that the optimization defining $\hat{\pi}_k$ is feasible on $\mathcal{E}_{0:k}$, since π_k trivially satisfies $\ell_{\pi_k}(\xi_i^k;\pi_k,\pi_k) = 0$ for $i = 1,...,n/E$, and $f_{\mathrm{cl}}^{\pi_k}$ is (β,ρ,γ)-E-δISS by constraint (3.1c). Observe that on $\mathcal{E}_{0:k-1}$, we have that

$$\mathbb{E}_{\xi\sim\mathcal{D}}\ell_{\pi_k}(\xi;\pi_k,\pi_\star) \leqslant \alpha\beta_k(n,E,\delta)\,. \tag{A.4}$$

By Proposition A.1 and taking a union bound over $\pi_k,\pi_\star$, there exists an event $\mathcal{E}_k$ with $\mathbb{P}(\mathcal{E}_k) \geqslant 1-\delta/E$ such that on $\mathcal{E}_k$, the following statement holds:

$$\max_{\pi_t\in\{\pi_k,\pi_\star\}}\sup_{\pi\in\Pi}\left|\mathbb{E}_{\xi\sim\mathcal{D}}\ell_{\pi_k}(\xi;\pi,\pi_t) - \frac{1}{n/E}\sum_{i=1}^{n/E}\ell_{\pi_k}(\xi_i^k;\pi,\pi_t)\right| \leqslant \Gamma_{n,E,\delta}\,. \tag{A.5}$$

Furthermore we note that on $\mathcal{E}_k$, it holds that:

$$\begin{aligned}
\frac{1}{n/E}\sum_{i=1}^{n/E}\ell_{\pi_k}(\xi_i^k;\pi_k,\pi_\star) &\overset{(a)}{\leqslant} \mathbb{E}_{\xi\sim\mathcal{D}}\ell_{\pi_k}(\xi;\pi_k,\pi_\star) + \Gamma_{n,E,\delta} \\
&\overset{(b)}{\leqslant} \alpha\beta_k(n,E,\delta) + \Gamma_{n,E,\delta}\,.
\end{aligned} \tag{A.6}$$

Above, (a) follows from (A.5) and (b) follows from (A.4).

Our remaining task is to define $\beta_{k+1}(n,E,\delta)$ such that on $\mathcal{E}_{0:k}$ we have:

$$\mathbb{E}_{\xi\sim\mathcal{D}}\ell_{\pi_{k+1}}(\xi;\pi_{k+1},\pi_\star) \leqslant \alpha\beta_{k+1}(n,E,\delta)\,.$$

We proceed with a similar argument as in the base case. We first write:

$$\begin{aligned}
\Delta_{\pi_{k+1},\pi_\star}(x) &= g(x)(\pi_{k+1}(x)-\pi_\star(x)) \\
&= (1-\alpha)g(x)(\pi_k(x)-\pi_\star(x)) + \alpha g(x)(\hat{\pi}_k(x)-\pi_\star(x)) \\
&= (1-\alpha)\Delta_{\pi_k,\pi_\star}(x) + \alpha\Delta_{\hat{\pi}_k,\pi_\star}(x)\,.
\end{aligned}$$

Therefore since by assumption $\Delta_{\pi_{k+1},\pi_\star}$ is L_Δ-Lipschitz:

$$
\begin{aligned}
\ell_{\pi_{k+1}}(\xi;\pi_{k+1},\pi_\star) &= \sum_{t=0}^{T-1} \|\Delta_{\pi_{k+1},\pi_\star}(\varphi_t^{\pi_{k+1}}(\xi))\| \\
&\leqslant \sum_{t=0}^{T-1} \|\Delta_{\pi_{k+1},\pi_\star}(\varphi_t^{\pi_k}(\xi))\| + L_\Delta \sum_{t=0}^{T-1} \|\varphi_t^{\pi_{k+1}}(\xi) - \varphi_t^{\pi_k}(\xi)\| \\
&\leqslant (1-\alpha)\ell_{\pi_k}(\xi;\pi_k,\pi_\star) + \alpha \ell_{\pi_k}(\xi;\hat{\pi}_k,\pi_\star) + L_\Delta \sum_{t=0}^{T-1} \|\varphi_t^{\pi_{k+1}}(\xi) - \varphi_t^{\pi_k}(\xi)\| \,.
\end{aligned}
$$

Now we write:

$$
\begin{aligned}
f_{\mathrm{cl}}^{\pi_{k+1}}(x) &= f(x) + g(x)((1-\alpha)\pi_k(x) + \alpha\hat{\pi}_k(x)) \\
&= f_{\mathrm{cl}}^{\pi_k}(x) + \alpha \Delta_{\hat{\pi}_k,\pi_k}(x) \,,
\end{aligned}
$$

and therefore given that $f_{\mathrm{cl}}^{\pi_{k+1}}$ is (β,ρ,γ)-E-δISS by constraint (3.1c), we have that for all ξ

$$
\|\varphi_t^{\pi_{k+1}}(\xi) - \varphi_t^{\pi_k}(\xi)\| \leqslant \alpha\gamma \sum_{s=0}^{t-1} \rho^{t-s-1} \|\Delta_{\hat{\pi}_k,\pi_k}(\varphi_s^{\pi_k}(\xi))\| \,.
$$

Summing the left-hand-side above from $t=0$ to $t=T-1$ yields

$$
\sum_{t=0}^{T-1} \|\varphi_t^{\pi_{k+1}}(\xi) - \varphi_t^{\pi_k}(\xi)\| \leqslant \frac{\alpha\gamma}{1-\rho} \ell_{\pi_k}(\xi;\hat{\pi}_k,\pi_k) \,.
$$

Combining this inequality with the inequality above,

$$
\ell_{\pi_{k+1}}(\xi;\pi_{k+1},\pi_\star) \leqslant (1-\alpha)\ell_{\pi_k}(\xi;\pi_k,\pi_\star) + \alpha\ell_{\pi_k}(\xi;\hat{\pi}_k,\pi_\star) + \frac{\alpha\gamma L_\Delta}{1-\rho} \ell_{\pi_k}(\xi;\hat{\pi}_k,\pi_k) \,.
$$

Taking expectations of both sides and applying (A.4), we obtain

$$
\mathbb{E}_{\xi\sim\mathcal{D}} \ell_{\pi_{k+1}}(\xi;\pi_{k+1},\pi_\star) \leqslant (1-\alpha)\alpha\beta_k(n,E,\delta) + \alpha\mathbb{E}_{\xi\sim\mathcal{D}}\ell_{\pi_k}(\xi;\hat{\pi}_k,\pi_\star) + \frac{\alpha\gamma L_\Delta}{1-\rho}\mathbb{E}_{\xi\sim\mathcal{D}}\ell_{\pi_k}(\xi;\hat{\pi}_k,\pi_k) \,.
$$

Now on $\mathcal{E}_k$ we have:

$$
\begin{aligned}
\mathbb{E}_{\xi\sim\mathcal{D}}\ell_{\pi_k}(\xi;\hat{\pi}_k,\pi_k) &\overset{(a)}{\leqslant} \frac{1}{n/E} \sum_{i=1}^{n/E} \ell_{\pi_k}(\xi_i^k;\hat{\pi}_k,\pi_k) + \Gamma_{n,E,\delta} \\
&\overset{(b)}{\leqslant} c_k + \Gamma_{n,E,\delta} \,,
\end{aligned}
$$

where the first inequality (a) follows from (A.5), and the second inequality (b) from $\hat{\pi}_k$ being feasible to the constrained optimization problem (3.1).

Similarly, we have that

$$\begin{aligned}
\mathbb{E}_{\xi\sim\mathcal{D}}\ell_{\pi_k}(\xi;\hat{\pi}_k,\pi_\star) &\overset{(a)}{\leqslant} \frac{1}{n/E}\sum_{i=1}^{n/E}\ell_{\pi_k}(\xi_i^k;\hat{\pi}_k,\pi_\star)+\Gamma_{n,E,\delta} \\
&\overset{(b)}{\leqslant} \frac{1}{n/E}\sum_{i=1}^{n/E}\ell_{\pi_k}(\xi_i^k;\pi_k,\pi_\star)+\Gamma_{n,E,\delta} \\
&\overset{(c)}{\leqslant} \mathbb{E}_{\xi\sim\mathcal{D}}\ell_{\pi_k}(\xi;\pi_k,\pi_\star)+2\Gamma_{n,E,\delta} \\
&\overset{(d)}{\leqslant} \alpha\beta_k(n,E,\delta)+2\Gamma_{n,E,\delta}\ ,
\end{aligned}$$

where (a) follows from (A.5), (b) from using π_k as a feasible point for optimization problem (3.1) and optimality of $\hat{\pi}_k$, (c) from another application (A.5), and (d) follows from (A.4).

Hence, we have:

$$\begin{aligned}
\mathbb{E}_{\xi\sim\mathcal{D}}\ell_{\pi_{k+1}}(\xi;\pi_{k+1},\pi_\star) &\leqslant \alpha\beta_k(n,E,\delta)+\alpha\left(2+\frac{\gamma L_\Delta}{1-\rho}\right)\Gamma_{n,E,\delta}+\frac{\alpha\gamma L_\Delta}{1-\rho}c_k \\
&= \alpha\left[\beta_k(n,E,\delta)+\left(2+\frac{\gamma L_\Delta}{1-\rho}\right)\Gamma_{n,E,\delta}+\frac{\gamma L_\Delta}{1-\rho}c_k\right] \\
&=: \alpha\beta_{k+1}(n,E,\delta)\ .
\end{aligned}$$

This finishes the inductive step. By unrolling the recurrence for $\beta_k(n,E,\delta)$, we conclude that for all $k=1,...,E-1$,

$$\beta_k(n,E,\delta)\leqslant k\left(2+\frac{L_\Delta\gamma}{1-\rho}\right)\Gamma_{n,E,\delta}+\frac{\gamma L_\Delta}{1-\rho}\sum_{i=1}^{k-1}c_i\ .$$

Thus we conclude that on the event $\mathcal{E}_{0:E-2}$, which occurs with probability at least $1-\frac{E-1}{E}\delta$, we have that for $k=1,...,E-1$,

$$\mathbb{E}_{\xi\sim\mathcal{D}}\ell_{\pi_k}(\xi;\pi_k,\pi_\star)\leqslant \alpha k\left(2+\frac{\gamma L_\Delta}{1-\rho}\right)\Gamma_{n,E,\delta}+\alpha\left(\sum_{i=1}^{k-1}c_i\right)\frac{\gamma L_\Delta}{1-\rho}\ .$$

This inequality yields (3.4), after applying the assumption that $2\leqslant\frac{\gamma L_\Delta}{1-\rho}$.

Final bound: We now assume $k=E-1$ and that the event $\mathcal{E}_{0:E-2}$ holds. On this event, we have:

$$\mathbb{E}_{\xi\sim\mathcal{D}}\ell_{\pi_{E-1}}(\xi;\pi_{E-1},\pi_\star)\leqslant\alpha\beta_{E-1}(n,E,\delta)\ . \tag{A.7}$$

We first check feasiblity of the optimization defining $\hat{\pi}_{E-1}$. Define $\tilde{\pi}_{E-1}$ as:

$$\tilde{\pi}_{E-1}:=\frac{(1-\alpha)^E}{\alpha}\pi_\star+\left(1-\frac{(1-\alpha)^E}{\alpha}\right)\pi_{E-1}\ .$$

By our assumption that $\frac{(1-\alpha)^E}{\alpha} \leqslant 1$, we have that $\tilde{\pi}_{E-1} \in \Pi$ by convexity of Π. Next, we have:

$$\begin{aligned}
\Delta_{\tilde{\pi}_{E-1},\pi_{E-1}}(x) &= g(x)\left[\frac{(1-\alpha)^E}{\alpha}\pi_\star(x) + \left(1 - \frac{(1-\alpha)^E}{\alpha}\right)\pi_{E-1}(x) - \pi_{E-1}(x)\right] \\
&= \frac{(1-\alpha)^E}{\alpha} g(x)(\pi_\star(x) - \pi_{E-1}(x)) \\
&= \frac{(1-\alpha)^E}{\alpha} \Delta_{\pi_\star,\pi_{E-1}}(x) \, .
\end{aligned}$$

Therefore:

$$\frac{1}{n/E}\sum_{i=1}^{n/E} \ell_{\pi_{E-1}}(\xi_i^{E-1}; \tilde{\pi}_{E-1}, \pi_{E-1}) = \frac{(1-\alpha)^E}{\alpha}\frac{1}{n/E}\sum_{i=1}^{n/E} \ell_{\pi_{E-1}}(\xi_i^{E-1}; \pi_\star, \pi_{E-1}) = c_{E-1} \, ,$$

which shows that $\tilde{\pi}_{E-1}$ satisfies constraint (3.1b) with equality. Next, we observe that:

$$\frac{1}{1-(1-\alpha)^E}\left[(1-\alpha)\pi_{E-1} + \alpha\tilde{\pi}_{E-1} - (1-\alpha)^E \pi_\star\right] = \pi_{E-1} \, ,$$

and hence $\tilde{\pi}_{E-1}$ satisfies constraint (3.1c) since $f_{\mathrm{cl}}^{\pi_{E-1}}$ is (β, ρ, γ)-E-δISS by constraint (3.1c). This shows the optimization problem defining $\hat{\pi}_{E-1}$ is feasible.

Now as in the inductive step, by Proposition A.1 and taking a union bound over $\pi_{E-1}, \pi_\star$, there exists an event $\mathcal{E}_{E-1}$ with $\mathbb{P}(\mathcal{E}_{E-1}) \geqslant 1 - \delta/E$ such that on $\mathcal{E}_{E-1}$, the following statement holds:

$$\max_{\pi_t \in \{\pi_{E-1},\pi_\star\}} \sup_{\pi\in\Pi} \left| \mathbb{E}_{\xi\sim\mathcal{D}} \ell_{\pi_{E-1}}(\xi; \pi, \pi_t) - \frac{1}{n/E}\sum_{i=1}^{n/E} \ell_{\pi_{E-1}}(\xi_i^{E-1}; \pi, \pi_t) \right| \leqslant \Gamma_{n,E,\delta} \, . \tag{A.8}$$

Furthermore we note that on $\mathcal{E}_{E-1}$, it holds that:

$$\frac{1}{n/E}\sum_{i=1}^{n/E} \ell_{\pi_{E-1}}(\xi_i^{E-1}; \pi_{E-1}, \pi_\star) \leqslant \alpha\beta_{E-1}(n, E, \delta) + \Gamma_{n,E,\delta} \, . \tag{A.9}$$

Therefore we can bound c_{E-1} on $\mathcal{E}_{E-1}$ by:

$$c_{E-1} \leqslant \frac{(1-\alpha)^E}{\alpha}(\alpha\beta_{E-1}(n, E, \delta) + \Gamma_{n,E,\delta}) \, . \tag{A.10}$$

We will use this bound in the sequel.

Now we write:

$$\begin{aligned}
\Delta_{\pi_E,\pi_\star}(x) &= g(x)\left(\frac{1}{1-(1-\alpha)^E}\left[(1-\alpha)\pi_{E-1}(x) + \alpha\hat{\pi}_{E-1}(x) - (1-\alpha)^E\pi_\star(x)\right] - \pi_\star(x)\right) \\
&= \frac{1-\alpha}{1-(1-\alpha)^E}\Delta_{\pi_{E-1},\pi_\star}(x) + \frac{\alpha}{1-(1-\alpha)^E}\Delta_{\hat{\pi}_{E-1},\pi_\star}(x) \, .
\end{aligned}$$

Therefore since $\Delta_{\pi_E,\pi_\star}$ is L_Δ-Lipschitz by assumption,

$$\begin{aligned}
\ell_{\pi_E}(\xi;\pi_E,\pi_\star) &= \sum_{t=0}^{T-1}\|\Delta_{\pi_E,\pi_\star}(\varphi_t^{\pi_E}(\xi))\| \\
&\leqslant \sum_{t=0}^{T-1}\|\Delta_{\pi_E,\pi_\star}(\varphi_t^{\pi_{E-1}}(\xi))\| + L_\Delta \sum_{t=0}^{T-1}\|\varphi_t^{\pi_E}(\xi) - \varphi_t^{\pi_{E-1}}(\xi)\| \\
&\leqslant \frac{1-\alpha}{1-(1-\alpha)^E}\ell_{\pi_{E-1}}(\xi;\pi_{E-1},\pi_\star) + \frac{\alpha}{1-(1-\alpha)^E}\ell_{\pi_{E-1}}(\xi;\hat{\pi}_{E-1},\pi_\star) \\
&\qquad + L_\Delta \sum_{t=0}^{T-1}\|\varphi_t^{\pi_E}(\xi) - \varphi_t^{\pi_{E-1}}(\xi)\| \,.
\end{aligned}$$

Now we write

$$\begin{aligned}
f_{\mathrm{cl}}^{\pi_E}(x) &= f(x) + g(x)\left(\frac{1}{1-(1-\alpha)^E}\left[(1-\alpha)\pi_{E-1}(x) + \alpha\hat{\pi}_{E-1}(x) - (1-\alpha)^E\pi_\star(x)\right]\right) \\
&= f_{\mathrm{cl}}^{\pi_{E-1}}(x) + \frac{\alpha}{1-(1-\alpha)^E}\Delta_{\hat{\pi}_{E-1},\pi_{E-1}}(x) + \frac{(1-\alpha)^E}{1-(1-\alpha)^E}\Delta_{\pi_{E-1},\pi_\star}(x) \,.
\end{aligned}$$

From constraint (3.1c), we have that $f_{\mathrm{cl}}^{\pi_E}$ is (β,ρ,γ)-E-δISS, and therefore we have for all ξ

$$\begin{aligned}
&\|\varphi_t^{\pi_E}(\xi) - \varphi_t^{\pi_{E-1}}(\xi)\| \\
&\leqslant \gamma\sum_{s=0}^{t-1}\rho^{t-s-1}\left\|\frac{\alpha}{1-(1-\alpha)^E}\Delta_{\hat{\pi}_{E-1},\pi_{E-1}}(\varphi_s^{\pi_{E-1}}(\xi)) + \frac{(1-\alpha)^E}{1-(1-\alpha)^E}\Delta_{\pi_{E-1},\pi_\star}(\varphi_s^{\pi_{E-1}}(\xi))\right\| \\
&\leqslant \frac{\alpha\gamma}{1-(1-\alpha)^E}\sum_{s=0}^{t-1}\rho^{t-s-1}\|\Delta_{\hat{\pi}_{E-1},\pi_{E-1}}(\varphi_s^{\pi_{E-1}}(\xi))\| + \frac{(1-\alpha)^E\gamma}{1-(1-\alpha)^E}\|\Delta_{\pi_{E-1},\pi_\star}(\varphi_s^{\pi_{E-1}}(\xi))\| \,.
\end{aligned}$$

Summing the left-hand-side above from $t=0$ to $t=T-1$ yields

$$\begin{aligned}
\sum_{t=0}^{T-1}\|\varphi_t^{\pi_E}(\xi) - \varphi_t^{\pi_{E-1}}(\xi)\| &\leqslant \frac{\alpha\gamma}{1-\rho}\frac{1}{1-(1-\alpha)^E}\ell_{\pi_{E-1}}(\xi;\hat{\pi}_{E-1},\pi_{E-1}) \\
&\quad + \frac{(1-\alpha)^E\gamma}{1-\rho}\frac{1}{1-(1-\alpha)^E}\ell_{\pi_{E-1}}(\xi;\pi_{E-1},\pi_\star) \,.
\end{aligned}$$

Combining with the inequality above, we have:

$$\begin{aligned}
(1-(1-\alpha)^E)\ell_{\pi_E}(\xi;\pi_E,\pi_\star) &\leqslant (1-\alpha)\ell_{\pi_{E-1}}(\xi;\pi_{E-1},\pi_\star) + \alpha\ell_{\pi_{E-1}}(\xi;\hat{\pi}_{E-1},\pi_\star) \\
&\quad + \frac{\alpha\gamma L_\Delta}{1-\rho}\ell_{\pi_{E-1}}(\xi;\hat{\pi}_{E-1},\pi_{E-1}) + \frac{(1-\alpha)^E\gamma L_\Delta}{1-\rho}\ell_{\pi_{E-1}}(\xi;\pi_{E-1},\pi_\star) \,.
\end{aligned}$$

Taking expectations of both sides, we obtain:

$$\begin{aligned}
&(1-(1-\alpha)^E)\mathbb{E}_{\xi\sim\mathcal{D}}\ell_{\pi_E}(\xi;\pi_E,\pi_\star) \\
&\leqslant (1-\alpha)\mathbb{E}_{\xi\sim\mathcal{D}}\ell_{\pi_{E-1}}(\xi;\pi_{E-1},\pi_\star) + \alpha\mathbb{E}_{\xi\sim\mathcal{D}}\ell_{\pi_{E-1}}(\xi;\hat{\pi}_{E-1},\pi_\star) \\
&\qquad + \frac{\alpha\gamma L_\Delta}{1-\rho}\mathbb{E}_{\xi\sim\mathcal{D}}\ell_{\pi_{E-1}}(\xi;\hat{\pi}_{E-1},\pi_{E-1}) + \frac{(1-\alpha)^E\gamma L_\Delta}{1-\rho}\mathbb{E}_{\xi\sim\mathcal{D}}\ell_{\pi_{E-1}}(\xi;\pi_{E-1},\pi_\star) \\
&= \left((1-\alpha) + \frac{(1-\alpha)^E\gamma L_\Delta}{1-\rho}\right)\mathbb{E}_{\xi\sim\mathcal{D}}\ell_{\pi_{E-1}}(\xi;\pi_{E-1},\pi_\star) + \alpha\mathbb{E}_{\xi\sim\mathcal{D}}\ell_{\pi_{E-1}}(\xi;\hat{\pi}_{E-1},\pi_\star) \\
&\qquad + \frac{\alpha\gamma L_\Delta}{1-\rho}\mathbb{E}_{\xi\sim\mathcal{D}}\ell_{\pi_{E-1}}(\xi;\hat{\pi}_{E-1},\pi_{E-1})\,.
\end{aligned}$$

Now on $\mathcal{E}_{E-1}$ we have:

$$\begin{aligned}
\mathbb{E}_{\xi\sim\mathcal{D}}\ell_{\pi_{E-1}}(\xi;\hat{\pi}_{E-1},\pi_{E-1}) &\overset{(a)}{\leqslant} \frac{1}{n/E}\sum_{i=1}^{n/E}\ell_{\pi_{E-1}}(\xi_i^{E-1};\hat{\pi}_{E-1},\pi_{E-1}) + \Gamma_{n,E,\delta} \\
&\overset{(b)}{\leqslant} c_{E-1} + \Gamma_{n,E,\delta} \\
&\overset{(c)}{\leqslant} \frac{(1-\alpha)^E}{\alpha}(\alpha\beta_{E-1}(n,E,\delta) + \Gamma_{n,E,\delta}) + \Gamma_{n,E,\delta} \\
&= (1-\alpha)^E\beta_{E-1}(n,E,\delta) + \left(1 + \frac{(1-\alpha)^E}{\alpha}\right)\Gamma_{n,E,\delta}\,.
\end{aligned}$$

Here, (a) follows by (A.8), (b) follows from constraint (3.1b), and (c) follows from (A.10). Next, observe that:

$$\begin{aligned}
\Delta_{\tilde{\pi}_{E-1},\pi_\star}(x) &= g(x)\left[\frac{(1-\alpha)^E}{\alpha}\pi_\star(x) + \left(1 - \frac{(1-\alpha)^E}{\alpha}\right)\pi_{E-1}(x) - \pi_\star(x)\right] \\
&= \left(1 - \frac{(1-\alpha)^E}{\alpha}\right)g(x)(\pi_{E-1}(x) - \pi_\star(x)) \\
&= \left(1 - \frac{(1-\alpha)^E}{\alpha}\right)\Delta_{\pi_{E-1},\pi_\star}(x)\,.
\end{aligned}$$

Therefore, for any π, ξ, we have:

$$\ell_\pi(\xi;\tilde{\pi}_{E-1},\pi_\star) = \left(1 - \frac{(1-\alpha)^E}{\alpha}\right)\ell_\pi(\xi;\pi_{E-1},\pi_\star)\,. \tag{A.11}$$

Hence on $\mathcal{E}_{E-1}$, we have:

$$\begin{aligned}
\mathbb{E}_{\xi\sim\mathcal{D}}\ell_{\pi_{E-1}}(\xi;\hat{\pi}_{E-1},\pi_\star) &\overset{(a)}{\leqslant} \frac{1}{n/E}\sum_{i=1}^{n/E}\ell_{\pi_{E-1}}(\xi_i^{E-1};\hat{\pi}_{E-1},\pi_\star)+\Gamma_{n,E,\delta} \\
&\overset{(b)}{\leqslant} \frac{1}{n/E}\sum_{i=1}^{n/E}\ell_{\pi_{E-1}}(\xi_i^{E-1};\tilde{\pi}_{E-1},\pi_\star)+\Gamma_{n,E,\delta} \\
&\overset{(c)}{=} \left(1-\frac{(1-\alpha)^E}{\alpha}\right)\frac{1}{n/E}\sum_{i=1}^{n/E}\ell_{\pi_{E-1}}(\xi_i^{E-1};\pi_{E-1},\pi_\star)+\Gamma_{n,E,\delta} \\
&\overset{(d)}{\leqslant} \left(1-\frac{(1-\alpha)^E}{\alpha}\right)(\mathbb{E}_{\xi\sim\mathcal{D}}\ell_{\pi_{E-1}}(\xi;\pi_{E-1},\pi_\star)+\Gamma_{n,E,\delta})+\Gamma_{n,E,\delta} \\
&\overset{(e)}{\leqslant} \alpha\beta_{E-1}(n,E,\delta)+2\Gamma_{n,E,\delta}\,.
\end{aligned}$$

Here, (a) follows by (A.8), (b) follows from the optimality of $\hat{\pi}_{E-1}$ and feasibility of $\tilde{\pi}_{E-1}$, (c) follows from (A.11), (d) follows from another application of (A.8), and (e) follows from (A.7).

Hence, we have:

$$\begin{aligned}
&(1-(1-\alpha)^E)\mathbb{E}_{\xi\sim\mathcal{D}}\ell_{\pi_E}(\xi;\pi_E,\pi_\star) \\
&\leqslant \left((1-\alpha)+\frac{(1-\alpha)^E\gamma L_\Delta}{1-\rho}\right)\alpha\beta_{E-1}(n,E,\delta)+\alpha(\alpha\beta_{E-1}(n,E,\delta)+2\Gamma_{n,E,\delta}) \\
&\qquad+\frac{\alpha\gamma L_\Delta}{1-\rho}\left((1-\alpha)^E\beta_{E-1}(n,E,\delta)+\left(1+\frac{(1-\alpha)^E}{\alpha}\right)\Gamma_{n,E,\delta}\right) \\
&=\alpha\left[\left(1+2\frac{(1-\alpha)^E\gamma L_\Delta}{1-\rho}\right)\beta_{E-1}(n,E,\delta)+2\left(1+\frac{\alpha\gamma L_\Delta}{1-\rho}\right)\Gamma_{n,E,\delta}\right]\,.
\end{aligned}$$

This gives the final bound:

$$\mathbb{E}_{\xi\sim\mathcal{D}}\ell_{\pi_E}(\xi;\pi_E,\pi_\star)\leqslant\frac{\alpha}{1-(1-\alpha)^E}\left[\left(1+2\frac{(1-\alpha)^E\gamma L_\Delta}{1-\rho}\right)\beta_{E-1}(n,E,\delta)+2\left(1+\frac{\alpha\gamma L_\Delta}{1-\rho}\right)\Gamma_{n,E,\delta}\right]\,. \tag{A.12}$$

Now, under the Theorem assumptions we have that (i) $\alpha E \geqslant \log(2)$: consequently that $\frac{1}{1-(1-\alpha)^E}\leqslant 2$ and $\frac{(1-\alpha)^E}{1-(1-\alpha)^E}\leqslant 1$, and (ii) $\frac{(1-\alpha)^E}{\alpha}\leqslant 1$. Further, one can check that $q_E(\alpha):=\frac{\alpha}{1-(1-\alpha)^E}\leqslant 1$ for all $\alpha\in(0,1]$ and $E\geqslant 1$ (follows from $q_E'(\alpha)\geqslant 0$ for all $\alpha\in(0,1]$ and $E\geqslant 1$, and that $q_E(1)=1$ for all $E\geqslant 1$).

Combining bound (A.12) with the expression for $\beta_{E-1}(n,E,\delta)$, we then write

$$
\begin{aligned}
&\mathbb{E}_{\xi\sim\mathcal{D}}\ell_{\pi_E}(\xi;\pi_E,\pi_\star) \\
&\leqslant \frac{\alpha}{1-(1-\alpha)^E}\left[\left(1+2\frac{(1-\alpha)^E\gamma L_\Delta}{1-\rho}\right)\beta_{E-1}(n,E,\delta)+2\left(1+\frac{\alpha\gamma L_\Delta}{1-\rho}\right)\Gamma_{n,E,\delta}\right] \\
&\leqslant \frac{\alpha(E-1)}{1-(1-\alpha)^E}\left[\left(1+2\frac{(1-\alpha)^E\gamma L_\Delta}{1-\rho}\right)\left(2+\frac{\gamma L_\Delta}{1-\rho}\right)\Gamma_{n,E,\delta}\right] \\
&\quad+\frac{\alpha}{1-(1-\alpha)^E}\left(2\left(1+\frac{\alpha\gamma L_\Delta}{1-\rho}\right)\Gamma_{n,E,\delta}\right)+\frac{\alpha}{1-(1-\alpha)^E}\left(1+2\frac{(1-\alpha)^E\gamma L_\Delta}{1-\rho}\right)\left(\frac{\gamma L_\Delta}{1-\rho}\right)\sum_{i=1}^{E-2}c_i \\
&\overset{(a)}{\leqslant} \alpha(E-1)\left(2+\frac{2\gamma L_\Delta}{1-\rho}\right)\left(2+\frac{\gamma L_\Delta}{1-\rho}\right)\Gamma_{n,E,\delta}+2\alpha\left(2+\frac{\gamma L_\Delta}{1-\rho}\right)\Gamma_{n,E,\delta}+\alpha\left(2+\frac{2\gamma L_\Delta}{1-\rho}\right)\left(\frac{\gamma L_\Delta}{1-\rho}\right)\sum_{i=1}^{E-2}c_i \\
&\overset{(b)}{\leqslant} 2\alpha E\left(2+\frac{2\gamma L_\Delta}{1-\rho}\right)^2\Gamma_{n,E,\delta}+\alpha\left(2+\frac{2\gamma L_\Delta}{1-\rho}\right)^2\sum_{i=1}^{E-2}c_i \\
&\leqslant 2\alpha\left(2+\frac{2\gamma L_\Delta}{1-\rho}\right)^2\left(E\Gamma_{n,E,\delta}+\sum_{i=1}^{E-2}c_i\right) \\
&\leqslant 18\alpha\left(\frac{\gamma L_\Delta}{1-\rho}\right)^2\left(E\Gamma_{n,E,\delta}+\sum_{i=1}^{E-2}c_i\right).
\end{aligned}
$$

where (a) follows from distributing the $\frac{1}{1-(1-\alpha)^E}$ denominator across terms and applying the bounds discussed before the derivation, and (b) follows by collecting like terms at the expense of worse constants, and by bounding $\left(2+\frac{2\gamma L_\Delta}{1-\rho}\right)\leqslant\left(2+\frac{2\gamma L_\Delta}{1-\rho}\right)^2$. This yields (3.5).

If we further assume that (3.6) holds, we then have that $(1-(1-\alpha)^E)^{-1}\leqslant 2$ and $(1-\alpha)^E\leqslant e^{-\alpha E}\leqslant 1/(2+\frac{2L_\Delta\gamma}{1-\rho})$, and thus inequalities (a) and (b) above instead become

$$
\begin{aligned}
\mathbb{E}_{\xi\sim\mathcal{D}}\ell_{\pi_E}(\xi;\pi_E,\pi_\star) &\overset{(a')}{\leqslant} 4\alpha(E-1)\left(2+\frac{\gamma L_\Delta}{1-\rho}\right)\Gamma_{n,E,\delta}+2\alpha\left(2+\frac{2\gamma L_\Delta}{1-\rho}\right)\Gamma_{n,E,\delta}+4\alpha\left(\frac{\gamma L_\Delta}{1-\rho}\right)\sum_{i=1}^{E-2}c_i \\
&\overset{(b')}{\leqslant} 4\alpha\left(2+\frac{2\gamma L_\Delta}{1-\rho}\right)\left(E\Gamma_{n,E,\delta}+\sum_{i=1}^{E-2}c_i\right) \\
&\leqslant 12\alpha\frac{\gamma L_\Delta}{1-\rho}\left(E\Gamma_{n,E,\delta}+\sum_{i=1}^{E-2}c_i\right).
\end{aligned}
$$

where (a') now follows from the small-shift condition (3.6), and (b') again follows from bounding and collecting like terms at the expense of slightly worse constants. This yields (3.7), and concludes the proof of Theorem 3.1.

B Rademacher Complexity Bounds

We restate Proposition 3.3 here with completely described constants.

Proposition B.1. *Let $\Pi = \{\pi(x,\theta) : \theta \in \mathbb{R}^k, \|\theta\| \leqslant B_\theta\}$ for a fixed map π, and suppose that the map π is twice continuously differentiable. Recall that $L_{\partial^2\pi}$ is defined to be:*

$$L_{\partial^2\pi} = \sup_{\|x\|\leqslant \beta B_0, \|\theta\|\leqslant B_\theta} \left\| \frac{\partial^2 \pi}{\partial\theta\partial x} \right\|_{\ell^2(\mathbb{R}^k)\to M(\mathbb{R}^{q\times p})} .$$

We have that:

$$\mathcal{R}_n(\Pi) \leqslant 32.5 B_0 B_g B_\theta L_{\partial^2\pi} \frac{\beta}{1-\rho}\sqrt{\frac{k}{n}} . \tag{B.1}$$

Therefore, under the assumptions of Theorem 3.1, if $\alpha E \geqslant \log\left(2 + \frac{2L_\Delta \gamma}{1-\rho}\right)$, with probability at least $1-\delta$ over $\left\{ \{\xi_i^k\}_{i=1}^{n/E} \right\}_{k=0}^{E-1} \sim \mathcal{D}^n$:

$$\mathbb{E}_{\xi\sim\mathcal{D}} \ell_{\pi_E}(\xi; \pi_E, \pi_\star) \leqslant 12\alpha \left(\frac{\gamma L_\Delta}{1-\rho} \right) \left[65 B_0 B_g B_\theta L_{\partial^2\pi} \frac{\beta E}{1-\rho} \sqrt{\frac{Ek}{n}} + \frac{\beta B_0 L_\Delta E}{1-\rho} \sqrt{\frac{E}{n}\log(4E/\delta)} + \sum_{i=1}^{E-2} c_i \right] , \tag{B.2}$$

and with probability at least $1-\delta$ over $\left(\left\{ \{\xi_i^k\}_{i=1}^{n/E} \right\}_{k=0}^{E-1}, \xi \right) \sim \mathcal{D}^{n+1}$:

$$h_j(\{\varphi_t^{\pi_E}(\xi)\}_{t=0}^{T-1}) \leqslant -\gamma_h + \frac{24\alpha\gamma^2 L_\Delta L_h}{\delta(1-\rho)^2} \left[65 B_0 B_g B_\theta L_{\partial^2\pi} \frac{\beta E}{1-\rho} \sqrt{\frac{Ek}{n}} + \frac{\beta B_0 L_\Delta E}{1-\rho} \sqrt{\frac{E}{n}\log(8E/\delta)} + \sum_{i=1}^{E-2} c_i \right] . \tag{B.3}$$

Proof. Fix an x and θ_1, θ_2. Because $\pi(0,\theta) = 0$ for all θ, we have by repeated applications of Taylor's theorem:

$$\begin{aligned}
\pi(x,\theta_1) - \pi(x,\theta_2) &= \left(\int_0^1 \left[\frac{\partial \pi}{\partial x}(\alpha x, \theta_1) - \frac{\partial \pi}{\partial x}(\alpha x, \theta_2) \right] d\alpha \right) x \\
&= \left(\int_0^1 \left[\int_0^1 \frac{\partial^2 \pi}{\partial\theta\partial x}(\alpha x, \beta\theta_1 + (1-\beta)\theta_2)(\theta_1 - \theta_2)\, d\beta \right] d\alpha \right) x .
\end{aligned}$$

Now supposing $\|x\| \leqslant \beta B_0$ and $\|\theta_i\| \leqslant B_\theta$ for $i = 1, 2$, then

$$\|\pi(x;\theta_1) - \pi(x;\theta_2)\| \leqslant L_{\partial^2\pi} \|x\| \|\theta_1 - \theta_2\| .$$

Therefore for any goal policy π_g,

$$\begin{aligned}
|\ell_{\pi_d}(\xi; \pi_1, \pi_g) - \ell_{\pi_d}(\xi; \pi_2, \pi_g)| &\leqslant \sum_{t=0}^{T-1} \| g(\varphi_t^{\pi_d}(\xi))(\pi_1(\varphi_t^{\pi_d}(\xi), \theta_1) - \pi_2(\varphi_t^{\pi_d}(\xi), \theta_2)) \| \\
&\leqslant B_g \sum_{t=0}^{T-1} \| \pi_1(\varphi_t^{\pi_d}(\xi), \theta_1) - \pi_2(\varphi_t^{\pi_d}(\xi), \theta_2)) \| \\
&\leqslant B_g L_{\partial^2\pi} \left(\sum_{t=0}^{T-1} \|\varphi_t^{\pi_d}(\xi)\| \right) \|\theta_1 - \theta_2\| \\
&\leqslant B_0 B_g L_{\partial^2\pi} \frac{\beta}{1-\rho} \|\theta_1 - \theta_2\| .
\end{aligned}$$

This shows that the loss function ℓ_{π_d} is uniformly Lipschitz w.r.t. the parameters θ. Now for a fixed $\xi_1, ..., \xi_n$, π_d, and π_g, define the $\|\cdot\|_{\mathbb{P}_n}$ metric over Π as:

$$\|\pi_1 - \pi_2\|_{\mathbb{P}_n}^2 := \frac{1}{n}\sum_{i=1}^{n}(\ell_{\pi_d}(\xi_i; \pi_1, \pi_g) - \ell_{\pi_d}(\xi_i; \pi_2, \pi_g))^2 .$$

The calculation above shows that for all $\pi_1, \pi_2 \in \Pi$,

$$\|\pi_1 - \pi_2\|_{\mathbb{P}_n} \leqslant B_0 B_g L_{\partial^2\pi}\frac{\beta}{1-\rho}\|\theta_1 - \theta_2\|$$

Therefore by Dudley's entropy integral (cf. Wainwright (2019)):

$$\begin{aligned}
\mathcal{R}_n(\Pi) &\leqslant 24 \sup_{\pi_d \in \Pi_\delta(\beta,\rho,\gamma)} \sup_{\pi_g \in \Pi} \mathbb{E}_{\{\xi_i\}} \frac{1}{\sqrt{n}} \int_0^\infty \sqrt{\log N(\varepsilon; \Pi, \|\cdot\|_{\mathbb{P}_n})}\, d\varepsilon \\
&\leqslant 24 B_0 B_g L_{\partial^2\pi}\frac{\beta}{1-\rho}\frac{1}{\sqrt{n}}\int_0^\infty \sqrt{\log N(\varepsilon; \Theta, \|\cdot\|)}\, d\varepsilon \\
&\leqslant 24 B_0 B_g B_\theta L_{\partial^2\pi}\frac{\beta}{1-\rho}\frac{1}{\sqrt{n}}\int_0^\infty \sqrt{\log N(\varepsilon; \mathbb{B}_2(1), \|\cdot\|)}\, d\varepsilon \\
&\leqslant 24 B_0 B_g B_\theta L_{\partial^2\pi}\frac{\beta}{1-\rho}\sqrt{\frac{k}{n}}\int_0^1 \sqrt{\log(1+2/\varepsilon)}\, d\varepsilon \\
&\leqslant 32.5 B_0 B_g B_\theta L_{\partial^2\pi}\frac{\beta}{1-\rho}\sqrt{\frac{k}{n}} .
\end{aligned}$$

This yields (B.1). Both (B.2) and (B.3) follow from Theorem 3.1 and Proposition 3.2. □

One drawback of Proposition B.1 is that it only applies to function classes which are parameterized by a finite number of parameters. In the next proposition, we bound the Rademacher complexity for a family of infinite-dimensional functions that live in a vector-valued reproducing kernel Hilbert space (see e.g., Micchelli and Pontil (2005) for background on vector-valued RKHS). We note that our next result is pessimistic by a factor of $\sqrt{T}$ in the bound– we leave completely removing the dependence on T in the RKHS setting to future work. Below, we let $\mathrm{Sym}_{\geqslant 0}^{q\times q}$ denote the space of $q \times q$ symmetric PSD matrices. We also define the following function $\phi_\pi^{\pi_d} : \mathbb{R}^p \to \mathbb{R}^{T\times p}$ for $\pi, \pi_d \in \Pi$:

$$\phi_\pi^{\pi_d}(\xi) = \begin{bmatrix} g(\varphi_0^{\pi_d}(\xi))\pi(\varphi_0^{\pi_d}(\xi)) \\ \vdots \\ g(\varphi_{T-1}^{\pi_d}(\xi))\pi(\varphi_{T-1}^{\pi_d}(\xi)) \end{bmatrix} . \tag{B.4}$$

Proposition B.2. *Let $(\mathcal{H}, \|\cdot\|_{\mathcal{H}})$ be a vector-valued RKHS of functions mapping $\mathbb{R}^p \to \mathbb{R}^q$ and let $K : \mathbb{R}^p \times \mathbb{R}^p \to \mathrm{Sym}_{\geqslant 0}^{q\times q}$ be its associated reproducing kernel. Let $\Pi = \{\pi \in \mathcal{H} : \|\pi\|_{\mathcal{H}} \leqslant B_{\mathcal{H}}\}$. We have that:*

$$\mathcal{R}_n(\Pi) \leqslant \sup_{\pi_d \in \Pi_\delta(\beta,\rho,\gamma)} B_{\mathcal{H}}\sqrt{\frac{2T}{n}}\sqrt{\mathbb{E}_\xi\, \mathbf{Tr}(g(\varphi_t^{\pi_d}(\xi))K(\varphi_t^{\pi_d}(\xi), \varphi_t^{\pi_d}(\xi))g(\varphi_t^{\pi_d}(\xi))^\mathsf{T})} .$$

Therefore, if for some $c > 0$,

$$\|g(x)K(x,x)g(x)^\mathsf{T}\| \leqslant G\|x\|^c\,,$$

we have

$$\mathcal{R}_n(\Pi) \leqslant B_{\mathcal{H}}\sqrt{\frac{2TG\min\{p,q\}}{n}}\sqrt{\frac{\beta^c}{1-\rho^c}\mathbb{E}\|\xi\|^c}\,.$$

Proof. For a $x \in \mathbb{R}^p$ and $w \in \mathbb{R}^q$, let $K_x w = K(\cdot,x)w \in \mathcal{H}$. By the reproducing property, $\langle \pi(x), w\rangle = \langle \pi, K_x w\rangle$. Let the shorthand $x_{i,t} = \varphi_t^{\pi_d}(\xi_i)$. We have:

$$\begin{aligned}
\mathcal{R}_n(\Pi) &\overset{(a)}{\leqslant} \sup_{\pi_d}\sqrt{2T}\mathbb{E}_{\{\xi_i\}}\mathbb{E}_{\{\varepsilon_i\}}\left[\sup_{\pi\in\Pi}\frac{1}{n}\sum_{i=1}^n \langle \varepsilon_i, \phi_\pi^{\pi_d}(\xi_i)\rangle\right] \\
&= \sup_{\pi_d}\frac{\sqrt{2T}}{n}\mathbb{E}_{\{\xi_i\}}\mathbb{E}_{\{\varepsilon_i\}}\left[\sup_{\|\pi\|_{\mathcal{H}}\leqslant B_{\mathcal{H}}}\left\langle \pi, \sum_{i=1}^n\sum_{t=0}^{T-1} K_{x_{i,t}} g(x_{i,t})^\mathsf{T}\varepsilon_i\right\rangle\right] \\
&= \sup_{\pi_d} B_{\mathcal{H}}\frac{\sqrt{2T}}{n}\mathbb{E}_{\{\xi_i\}}\mathbb{E}_{\{\varepsilon_i\}}\left\|\sum_{i=1}^n\sum_{t=0}^{T-1} K_{x_{i,t}} g(x_{i,t})^\mathsf{T}\varepsilon_i\right\|_{\mathcal{H}} \\
&\overset{(b)}{\leqslant} \sup_{\pi_d} B_{\mathcal{H}}\frac{\sqrt{2T}}{n}\mathbb{E}_{\{\xi_i\}}\sqrt{\mathbb{E}_{\{\varepsilon_i\}}\sum_{i_1,t_1,i_2,t_2}\langle K_{x_{i_1,t_1}} g(x_{i_1,t_1})^\mathsf{T}\varepsilon_{i_1}, K_{x_{i_2,t_2}} g(x_{i_2,t_2})^\mathsf{T}\varepsilon_{i_2}\rangle_{\mathcal{H}}} \\
&= \sup_{\pi_d} B_{\mathcal{H}}\frac{\sqrt{2T}}{n}\mathbb{E}_{\{\xi_i\}}\sqrt{\mathbb{E}_{\{\varepsilon_i\}}\sum_{i_1,t_1,i_2,t_2}\langle K(x_{i_2,t_2}, x_{i_1,t_2}) g(x_{i_1,t_1})^\mathsf{T}\varepsilon_{i_1}, g(x_{i_2,t_2})^\mathsf{T}\varepsilon_{i_2}\rangle} \\
&= \sup_{\pi_d} B_{\mathcal{H}}\frac{\sqrt{2T}}{n}\sqrt{\mathbb{E}_{\{\xi_i\}}\sum_{i,t}\mathbf{Tr}(g(x_{i,t})K(x_{i,t},x_{i,t})g(x_{i,t})^\mathsf{T})} \\
&= \sup_{\pi_d} B_{\mathcal{H}}\sqrt{\frac{2T}{n}}\sqrt{\mathbb{E}_{\xi_1}\sum_{t}\mathbf{Tr}(g(x_{1,t})K(x_{1,t},x_{1,t})g(x_{1,t})^\mathsf{T})}\,.
\end{aligned}$$

Above, the supremum is over $\pi_d \in \Pi_\delta(\beta,\rho,\gamma)$, (a) follows from Proposition B.3, and (b) follows from Jensen's inequality.

To finish the proof, note that for all $t \geqslant 0$ we have $\|\varphi_t^{\pi_d}(\xi)\| \leqslant \beta\rho^t\|\xi\|$. Therefore:

$$\begin{aligned}
&\sum_t \mathbf{Tr}(g(x_{1,t})K(x_{1,t},x_{1,t})g(x_{1,t})^\mathsf{T}) \leqslant \sum_t \min\{p,q\}\|g(x_{1,t})K(x_{1,t},x_{1,t})g(x_{1,t})^\mathsf{T}\| \\
&\leqslant \sum_t \min\{p,q\}\|x_{1,t}\|^c \leqslant \sum_t \min\{p,q\}\beta^c\rho^{ct}\|\xi\|^c \leqslant \min\{p,q\}\frac{\beta^c}{1-\rho^c}\|\xi\|^c\,.
\end{aligned}$$

□

The following contraction inequality was used in Proposition B.2.

Proposition B.3. *We have that:*

$$\mathcal{R}_n(\Pi) \leqslant \sup_{\pi_d\in\Pi_\delta(\beta,\rho,\gamma)}\sqrt{2T}\mathbb{E}_{\{\xi_i\}}\mathbb{E}_{\{\varepsilon_i\}}\sup_{\pi\in\Pi}\frac{1}{n}\sum_{i=1}^n\langle\varepsilon_i, \phi_\pi^{\pi_d}(\xi_i)\rangle\,.$$

Here, each $\varepsilon_i \in \{\pm 1\}^{T\times q}$ is a vector of independent Rademacher random variables.

Proof. The proof is a simple consequence of a standard vector-valued Rademacher contraction result (Maurer, 2016).

Let $h : \mathbb{R}^{T\times q} \to \mathbb{R}$ be defined as $h(z_0, ..., z_{T-1}) = \sum_{t=0}^{T-1} \|z_t\|$. We claim that h is $\sqrt{T}$-Lipschitz w.r.t. the ℓ_2-norm, since For any $z, \tilde{z} \in \mathbb{R}^{T\times q}$,

$$|h(z) - h(\tilde{z})| = \left|\sum_{t=0}^{T-1}(\|z_t\| - \|\tilde{z}_t\|)\right| \leqslant \sum_{t=0}^{T-1}\|z_t - \tilde{z}_t\| \leqslant \sqrt{T}\sqrt{\sum_{t=0}^{T-1}\|z_t - \tilde{z}_t\|^2} = \sqrt{T}\|z - \tilde{z}\| \,.$$

Now we write:

$$\begin{aligned}
\mathcal{R}_n(\Pi) &= \sup_{\pi_d}\sup_{\pi_g}\mathbb{E}_{\{\xi_i\}}\mathbb{E}_{\{\varepsilon_i\}}\left[\sup_{\pi\in\Pi}\frac{1}{n}\sum_{i=1}^n \varepsilon_i \ell_{\pi_d}(\xi_i;\pi,\pi_g)\right] \\
&= \sup_{\pi_d}\sup_{\pi_g}\mathbb{E}_{\{\xi_i\}}\mathbb{E}_{\{\varepsilon_i\}}\left[\sup_{\pi\in\Pi}\frac{1}{n}\sum_{i=1}^n \varepsilon_i h(\phi_\pi^{\pi_d}(\xi_i) - \phi_{\pi_g}^{\pi_d}(\xi_i))\right] \\
&\leqslant \sup_{\pi_d}\sup_{\pi_g}\sqrt{2T}\mathbb{E}_{\{\xi_i\}}\mathbb{E}_{\{\varepsilon_i\}}\left[\sup_{\pi\in\Pi}\frac{1}{n}\sum_{i=1}^n \langle\varepsilon_i, \phi_\pi^{\pi_d}(\xi_i) - \phi_{\pi_g}^{\pi_d}(\xi_i)\rangle\right] \\
&= \sup_{\pi_d}\sqrt{2T}\mathbb{E}_{\{\xi_i\}}\mathbb{E}_{\{\varepsilon_i\}}\left[\sup_{\pi\in\Pi}\frac{1}{n}\sum_{i=1}^n \langle\varepsilon_i, \phi_\pi^{\pi_d}(\xi_i)\rangle\right] \,.
\end{aligned}$$

Above, the inequality follows from Corollary 4 of Maurer (2016). □

C Guaranteeing E-δISS Stability of Played Policies through Lyapunov Certificates

C.1 Proof of Proposition 4.7

We first give a proof of Lemma 4.6.

Proof. (of Lemma 4.6) Wlog we assume the open-set containing the origin contains the ℓ_2-ball of radius β. Let $z_{t+1} = f(z_t)$ with initial condition z_0 and $\|z_0\| \leqslant 1$. Because f is (β, ρ, γ)-E-ISS, we have that $\|z_t\| \leqslant \beta\rho^t\|z_0\| \leqslant \beta$ for all $t \geqslant 0$. Let L_β be smallest constant so that $\frac{\partial f}{\partial x}$ is L_β-Lipschitz for all $\|x\| \leqslant \beta$ (such a constant must exist and be finite). Let $\Delta(x) := f(x) - A_0 x$. By Taylor's theorem, for every $\|x\| \leqslant \beta$ we have $\|\Delta(x)\| \leqslant L_\beta\|x\|^2$. Therefore we can write $z_{t+1} = A_0 z_t + \Delta(z_t)$ with $\|\Delta(z_t)\| \leqslant L_\beta\|z_t\|^2$ for all $t \geqslant 0$. Unrolling this recurrence:

$$z_t = A_0^t z_0 + \sum_{k=0}^{t-1} A_0^{t-1-k}\Delta(z_k) \,.$$

By triangle inequality:

$$\begin{aligned}
\|A_0^t z_0\| &\leqslant \|z_t\| + \left\| \sum_{k=0}^{t-1} A_0^{t-1-k} \Delta(z_k) \right\| \leqslant \beta\rho^t \|z_0\| + \sum_{k=0}^{t-1} \|A_0^{t-1-k}\| \cdot \|\Delta(z_k)\| \\
&\leqslant \beta\rho^t \|z_0\| + L_\beta \sum_{k=0}^{t-1} \|A_0^{t-1-k}\| \cdot \|z_k\|^2 \\
&\leqslant \beta\rho^t \|z_0\| + L_\beta \beta^2 \|z_0\|^2 \sum_{k=0}^{t-1} \rho^{2k} \|A_0^{t-1-k}\| \,.
\end{aligned}$$

We now prove that $\|A_0^t\| \leqslant \beta\rho^t$ for all $t \geqslant 0$. Clearly the bound holds for $t = 0$. Now fix a $t \geqslant 1$ and let $z_0 \neq 0$ be such that $\|A_0^t z_0\| = \|A_0^t\|\|z_0\|$ and $\|z_0\| \leqslant 1$. We have:

$$\|A_0^t\|\|z_0\| = \|A_0^t z_0\| \leqslant \beta\rho^t \|z_0\| + L_\beta \beta^2 \|z_0\|^2 \sum_{k=0}^{t-1} \rho^{2k} \|A_0^{t-1-k}\| \,.$$

Dividing both sides by $\|z_0\|$,

$$\|A_0^t\| \leqslant \beta\rho^t + L_\beta \beta^2 \|z_0\| \sum_{k=0}^{t-1} \rho^{2k} \|A_0^{t-1-k}\| \,.$$

Taking the limit as $\|z_0\| \searrow 0$ yields $\|A^t\| \leqslant \beta\rho^t$ as desired. □

Proposition C.1 (Proposition G.4 of Boffi et al. (2020a)). *Let A be a (C, ρ) discrete-time stable matrix. Let $\Delta_1, ..., \Delta_t$ be arbitrary perturbations. We have that for all $t \geqslant 1$:*

$$\left\| \prod_{i=1}^{t} (A + \Delta_i) \right\| \leqslant C \prod_{i=1}^{t} (\rho + C\|\Delta_i\|) \,.$$

Lemma C.2 (cf. Lemma G.5 of Boffi et al. (2020a)). *Consider an autonomous system $f(x)$ with $f(0) = 0$. Suppose that $f(x)$ is (β, ρ, γ)-E-ISS, that the linearization $A_0 := \frac{\partial f}{\partial x}(0)$ is a (C, ζ) discrete-time stable matrix, and that $\frac{\partial f}{\partial x}$ is L-Lipschitz. Let $\{d_t\}$ be an arbitrary sequence such that $\|d_t\| \leqslant D$ with:*

$$D \leqslant \frac{(1-\rho)(1-\zeta)}{2CL\gamma} \,. \tag{C.1}$$

Let x_0, y_0 have norm bounded by B_0. Then $f(x)$ is (β', ρ', γ')-E-δISS for the initial conditions (x_0, y_0) and signal $\{d_t\}$ with:

$$\beta' = C \exp\left(\frac{B_0 C L \beta}{1-\rho}\right) \,,\ \rho' = e^{-(1-\zeta)/2} \,,\ \gamma' = C \exp\left(\frac{CL\beta}{1-\rho}\left(\beta B_0 + \frac{\gamma D}{1-\rho}\right)\right) \,. \tag{C.2}$$

Proof. We follow the proof of Lemma G.5 in Boffi et al. (2020a). By E-ISS (4.2), we have that for all $t \geqslant 0$, for the dynamics $x_{t+1} = f(x_t) + d_t$:

$$\|x_t\| \leqslant \beta\rho^t \|x_0\| + \gamma \sum_{s=0}^{t-1} \rho^{t-1-s} \|d_s\| \,.$$

In particular, this implies that for all $t \geqslant 0$:

$$\|x_t\| \leqslant \beta\|x_0\| + \frac{\gamma D}{1-\rho} \,.$$

Define $g_t(x) := f(x) + d_t$ and for $t \geqslant 1$:

$$\Phi_t(x_0, d_0, ..., d_{t-1}) := (g_{t-1} \circ g_{t-2} \circ ... \circ g_0)(x_0) \,.$$

Observe that $\frac{\partial g_t}{\partial x}(x) = \frac{\partial f}{\partial x}(x)$. By the chain rule:

$$\begin{aligned}
\frac{\partial \Phi_t}{\partial x_0}(x_0, d_0, ..., d_{t-1}) &= \frac{\partial g_{t-1}}{\partial x}(x_{t-1})\frac{\partial g_{t-2}}{\partial x}(x_{t-2}) \cdots \frac{\partial g_0}{\partial x}(x_0) \\
&= \frac{\partial f}{\partial x}(x_{t-1})\frac{\partial f}{\partial x}(x_{t-2}) \cdots \frac{\partial f}{\partial x}(x_0) \\
&= \left(A_0 + \frac{\partial f}{\partial x}(x_{t-1}) - A_0\right)\left(A_0 + \frac{\partial f}{\partial x}(x_{t-2}) - A_0\right) \cdots \left(A_0 + \frac{\partial f}{\partial x}(x_0) - A_0\right) \,.
\end{aligned}$$

Define $\Delta_t := \frac{\partial f}{\partial x}(x_t) - A_0$. By the assumption that $\frac{\partial f}{\partial x}$ is L-Lipschitz, we have that $\|\Delta_t\| \leqslant L\|x_t\|$. Therefore by Proposition C.1:

$$\begin{aligned}
\left\|\frac{\partial \Phi_t}{\partial x_0}(x_0, d_0, ..., d_{t-1})\right\| &\leqslant C \exp\left(-t(1-\zeta) + CL\sum_{s=0}^{t-1}\|x_s\|\right) \\
&\leqslant C \exp\left(-t(1-\zeta) + CL\sum_{s=0}^{t-1}\left(\beta\rho^s\|x_0\| + \gamma\sum_{k=0}^{s-1}\rho^{s-1-k}\|d_k\|\right)\right) \\
&\leqslant C \exp\left(-t(1-\zeta) + \frac{CL\beta}{1-\rho}\|x_0\| + \frac{CL\gamma}{1-\rho}\sum_{s=0}^{t-1}\|d_s\|\right) \\
&\leqslant C \exp\left(-t(1-\zeta) + \frac{CL\beta}{1-\rho}\|x_0\| + \frac{CDL\gamma}{1-\rho}t\right) \\
&\leqslant C \exp\left(-t(1-\zeta)/2 + \frac{CL\beta}{1-\rho}\|x_0\|\right) \,.
\end{aligned}$$

Now let us look at $\frac{\partial \Phi_t}{\partial d_k}(x_0, d_0, ..., d_{t-1})$ for some $0 \leqslant k \leqslant t-1$. Again by the chain rule:

$$\begin{aligned}
\frac{\partial \Phi_t}{\partial d_k}(x_0, d_0, ..., d_{t-1}) &= \frac{\partial g_{t-1}}{\partial x}(x_{t-1})\frac{\partial g_{t-2}}{\partial x}(x_{t-2}) \cdots \frac{\partial g_{k+1}}{\partial x}(x_{k+1}) \\
&= \frac{\partial f}{\partial x}(x_{t-1})\frac{\partial f}{\partial x}(x_{t-2}) \cdots \frac{\partial f}{\partial x}(x_{k+1}) \\
&= \left(A_0 + \frac{\partial f}{\partial x}(x_{t-1}) - A_0\right)\left(A_0 + \frac{\partial f}{\partial x}(x_{t-2}) - A_0\right) \cdots \left(A_0 + \frac{\partial f}{\partial x}(x_{k+1}) - A_0\right) \,.
\end{aligned}$$

Using Proposition C.1 again:

$$
\begin{aligned}
\left\| \frac{\partial \Phi_t}{\partial d_k}(x_0, d_0, ..., d_{t-1}) \right\| &\leqslant C \exp\left(-(t-k-1)(1-\zeta) + CL \sum_{s=k+1}^{t-1} \|x_s\| \right) \\
&\leqslant C \exp\left(-(t-k-1)(1-\zeta) + CL \sum_{s=k+1}^{t-1} \left(\beta \rho^{s-(k+1)} \|x_{k+1}\| + \gamma \sum_{\ell=0}^{s-(k+1)-1} \rho^{s-(k+1)-1-\ell} \|d_{k+1+\ell}\| \right) \right) \\
&\leqslant C \exp\left(-(t-k-1)(1-\zeta) + \frac{CL\beta}{1-\rho} \|x_{k+1}\| + \frac{CL\gamma}{1-\rho} \sum_{s=k+1}^{t-1} \|d_s\| \right) \\
&\leqslant C \exp\left(-(t-k-1)(1-\zeta) + \frac{CL\beta}{1-\rho} \left(\beta \|x_0\| + \frac{\gamma D}{1-\rho} \right) + \frac{CL\gamma}{1-\rho} \sum_{s=k+1}^{t-1} \|d_s\| \right) \\
&\leqslant C \exp\left(-(t-k-1)(1-\zeta) + \frac{CL\beta}{1-\rho} \left(\beta \|x_0\| + \frac{\gamma D}{1-\rho} \right) + \frac{CDL\gamma}{1-\rho}(t-k-1) \right) \\
&\leqslant C \exp\left(-(t-k-1)(1-\zeta)/2 + \frac{CL\beta}{1-\rho} \left(\beta \|x_0\| + \frac{\gamma D}{1-\rho} \right) \right) .
\end{aligned}
$$

Now let x_0, y_0 be norm bounded by B_0. Let $(\tilde{z}_0, \tilde{d}_0, ..., \tilde{d}_{t-1})$ be an element along the ray connecting $(x_0, d_0, ..., d_{t-1})$ with $(y_0, 0, ..., 0)$. Observe that $\|\tilde{z}_0\| \leqslant B_0$ and $\|\tilde{d}_t\| \leqslant D$. Let $a_1 = (x_0, d_0, ..., d_{t-1})$, $a_2 = (y_0, 0, ..., 0)$, and $a(t) = ta_1 + (1-t)a_2$. By the mean-value theorem:

$$
\begin{aligned}
\Phi_t(a_1) - \Phi_t(a_2) &= \int_0^1 \frac{\partial \Phi_t}{\partial a}(a(t))(a_1 - a_2) \, dt \\
&= \int_0^1 \left[\frac{\partial \Phi_t}{\partial x_0}(a(t))(x_0 - y_0) + \sum_{s=0}^{t-1} \frac{\partial \Phi_t}{\partial d_s}(a(t)) d_s \right] dt .
\end{aligned}
$$

Therefore by the triangle inequality,

$$
\begin{aligned}
&\|\Phi_t(x_0, d_0, ..., d_{t-1}) - \Phi_t(y_0, 0, ..., 0)\| \\
&\leqslant \left(\int_0^1 \left\| \frac{\partial \Phi_t}{\partial x_0}(a(t)) \right\| dt \right) \|x_0 - y_0\| + \sum_{s=0}^{t-1} \left(\int_0^1 \left\| \frac{\partial \Phi_t}{\partial d_s}(a(t)) \right\| dt \right) \|d_s\| \\
&\leqslant C \exp\left(\frac{B_0 CL\beta}{1-\rho} \right) e^{-t(1-\zeta)/2} \|x_0 - y_0\| \\
&\qquad + C \exp\left(\frac{CL\beta}{1-\rho} \left(\beta B_0 + \frac{\gamma D}{1-\rho} \right) \right) \sum_{s=0}^{t-1} e^{-(t-s-1)(1-\zeta)/2} \|d_s\| .
\end{aligned}
$$

□

C.2 Proof of Proposition 4.7

The proof reduces to showing that every π_k is (β', ρ', γ')-E-δISS. For $k = 0, ..., E-2$, following the proof of Theorem 3.1, we write

$$
\begin{aligned}
f_{\mathrm{cl}}^{\pi_{k+1}}(x) &= f(x) + g(x)((1-\alpha)\pi_k(x) + \alpha \hat{\pi}_k(x)) \\
&= f_{\mathrm{cl}}^{\pi_k}(x) + \alpha \Delta_{\hat{\pi}_k, \pi_k}(x) .
\end{aligned}
$$

Therefore it suffices to pick α sufficiently small so as to ensure that $\|\alpha\Delta_{\hat{\pi}_k,\pi_k}(x)\| \leqslant \frac{(1-\rho)^2}{2\beta L\gamma}$, which then allows us to apply Lemma C.2. By Assumption 2.4, we can bound $\sup_x\|\Delta_{\hat{\pi}_k,\pi_k}(x)\| \leqslant 2B_g B_\Pi$ and therefore, as long as α satisfies $\alpha \leqslant \frac{(1-\rho)^2}{4B_g B_\pi \beta L\gamma}$, then by Lemma C.2, for almost all ξ, we have that $f_{\mathrm{cl}}^{\pi_k}$, for $k = 1, \ldots, E-1$, are (β', ρ', γ')-E-δISS. Notice that this condition on α is implied by (4.3).

It remains to consider the case when $k = E-1$. We want to show that $f_{\mathrm{cl}}^{\pi_E}$ is (β', ρ', γ')-E-δISS. We have that

$$f_{\mathrm{cl}}^{\pi_E}(x) = f_{\mathrm{cl}}^{\pi_{E-1}}(x) + \frac{\alpha}{1-(1-\alpha)^E}\Delta_{\hat{\pi}_{E-1},\pi_{E-1}}(x) + \frac{(1-\alpha)^E}{1-(1-\alpha)^E}\Delta_{\pi_{E-1},\pi_\star}(x)\,.$$

By triangle inequality, we bound the perturbation:

$$\begin{aligned}
&\left\|\frac{\alpha}{1-(1-\alpha)^E}\Delta_{\hat{\pi}_{E-1},\pi_{E-1}}(x) + \frac{(1-\alpha)^E}{1-(1-\alpha)^E}\Delta_{\pi_{E-1},\pi_\star}(x)\right\| \\
&\leqslant \frac{\alpha}{1-(1-\alpha)^E}\|\Delta_{\hat{\pi}_{E-1},\pi_{E-1}}(x)\| + \frac{(1-\alpha)^E}{1-(1-\alpha)^E}\|\Delta_{\pi_{E-1},\pi_\star}(x)\| \\
&\leqslant \frac{\alpha}{1-(1-\alpha)^E}2B_g B_\Pi + \frac{(1-\alpha)^E}{1-(1-\alpha)^E}2B_g B_\Pi\,.
\end{aligned}$$

If we ensure that:

$$\max\left\{\frac{\alpha}{1-(1-\alpha)^E}2B_g B_\Pi, \frac{(1-\alpha)^E}{1-(1-\alpha)^E}2B_g B_\Pi\right\} \leqslant \frac{(1-\rho)^2}{4\beta L\gamma}\,,$$

then the perturbation will be bounded by $\frac{(1-\rho)^2}{2\beta L\gamma}$ from which we can apply Lemma C.2 to conclude that for almost all ξ we have that $f_{\mathrm{cl}}^{\bar{\pi}_E}$ is (β', ρ', γ')-E-δISS. First, by assuming that $\alpha E \geqslant \log(2)$, we have that $(1-\alpha)^E \leqslant 1/2$. Thus it suffices to require that:

$$\alpha E \geqslant \log(2)\,,\ \ \max\{\alpha, (1-\alpha)^E\} \leqslant \frac{(1-\rho)^2}{16 B_g B_\Pi \beta L\gamma}\,.$$

These conditions are implied by (4.3).

D Perception-Based Control

D.1 End-to-End Perception-Based Control

Lemma D.1. *Under the Assumptions of Section 5, we have that $\pi_0, \ldots, \pi_E \in \Pi_z$, and that $\pi_E \in \Pi_z$ can be implemented without knowing the state x.*

Proof. Noting that $x = I^{-1}(z)$, we rewrite $\pi_\star(x) = \pi_\star \circ I^{-1}(z) =: \pi_\star^z(z) \in \Pi_z$ by assumption. Thus $\pi_0 = \pi_\star^z \in \Pi_z$. Then for $k = 1, \ldots, E-1$, under the update rule defined in line 5 of Algorithm 1, we have that $\pi_k(z) = (1-\alpha)^k\pi_\star(I^{-1}(z)) + \alpha\sum_{i=0}^{k-1}(1-\alpha)^{k-1-i}\hat{\pi}_i(z)$, and we therefore conclude that $\pi_k(z) = \lambda_k\pi_\star^z(z) + \sum_{i=0}^{k-1}\lambda_i\hat{\pi}_i(z)$ for some $\lambda_i \geqslant 0$, $\sum_i\lambda_i = 1$. Similarly, for $k = E$, $\pi_E = \sum_{i=0}^{E-1}\frac{\alpha(1-\alpha)^{E-1-i}}{1-(1-\alpha)^E}\hat{\pi}_i = \sum_{i=0}^{E-1}\lambda_i\hat{\pi}_i$ for some $\lambda_i \geqslant 0$, $\sum_i\lambda_i = 1$. Therefore $\pi_k(z) \in \Pi_z$ for all $k = 0, \ldots, E$ as Π_z is convex by assumption, and it follows that $\pi_k(z)$ is a feasible solution to the k-th epoch iteration of the end-to-end perception-based instantiation of Algorithm 1 using optimization problem (5.2). Finally, by construction, π_E only depends on $\hat{\pi}_k(z)$, $k = 0, \ldots, E-1$, which do not require knowledge of the state x (or the inverse model I^{-1}) to implement. $\square$

The next proposition considers the generative model

$$z_t = I(x_t) + e_t,\ I(0) = 0, \tag{D.1}$$

where we assume that I is L_I-Lipschitz and that $\|e\|_{\ell_2(T)} \leqslant \eta$, where $\|x\|_{\ell_2(T)}^2 := \sum_{t=0}^{T-1} \|x_t\|^2$. Note that we recover the generative model (5.1) and the results of Section 5 by setting $\eta = 0$ and $e_t \equiv 0$.

Define the Rademacher complexity

$$\mathcal{R}_n(\Pi_z, I, e) := \sup_{\substack{\pi_d \in \Pi_z \\ \pi_d \circ I \in S_\delta(\beta,\rho,\gamma)}} \sup_{\pi_g \in \Pi_z} \mathbb{E}_{\{\xi_i\}} \mathbb{E}_{\{\varepsilon_i\}} \left[\sup_{\pi_z \in \Pi_z} \frac{1}{n} \sum_{i=1}^n \varepsilon_i \ell_{\pi_d}(\xi_i; \pi_z, \pi_g) \right], \tag{D.2}$$

noting that $z(x_t) = I(x_t) + e_t$ when evaluating the expression on the right-hand-side.

Proposition D.2. *Fix a data generating policy $\pi_d \in \Pi_z$ such that $\pi_d \circ I \in S_\delta(\beta, \rho, \gamma)$ when $e_t \equiv 0$ in the generative model (D.1), and let $\Pi_z = \{\pi(z, \theta) : \theta \in \mathbb{R}^K, \|\theta\| \leqslant B_\theta\}$ for a fixed map π, and suppose that the map π is twice continuously differentiable. Define $L_{\partial^2 \pi}$ to be:*

$$L_{\partial^2 \pi_z} := \sup_{\|z\| \leqslant \beta B_0 L_I + \eta, \|\theta\| \leqslant B_\theta} \left\| \frac{\partial^2 \pi}{\partial \theta \partial z} \right\|_{\ell^2(\mathbb{R}^K) \to M(\mathbb{R}^{q \times D})}.$$

Then if the perceptual-measurements z_t follow the generative model (D.1):

$$\mathcal{R}_n(\Pi_z, I, e) \leqslant \frac{32.5 B_\theta B_g L_{\partial^2 \pi_z} L_I}{1-\rho} \left(\beta B_0 + 2\gamma B_g L_\Pi \eta \sqrt{T} \right) \sqrt{\frac{K}{n}}. \tag{D.3}$$

If we further assume that e_t is S-sparse, i.e., has at most S-nonzero elements over $t = 0, \ldots, T-1$, and that $\|e_t\| \leqslant \eta_s$, then the Rademacher complexity (D.2) is given by

$$\mathcal{R}_n(\Pi_z, I, e) \leqslant \frac{32.5 B_\theta B_g L_{\partial^2 \pi_z} L_I}{1-\rho} \left(\beta B_0 + 2\gamma B_g L_\Pi S \eta_s \right) \sqrt{\frac{K}{n}}. \tag{D.4}$$

Proof. Fix a z and θ_1, θ_2. Because $\pi(0, \theta) = 0$ for all θ, we have by repeated applications of Taylor's theorem:

$$\begin{aligned} \pi(z, \theta_1) - \pi(z, \theta_2) &= \left(\int_0^1 \left[\frac{\partial \pi}{\partial z}(\alpha z, \theta_1) - \frac{\partial \pi}{\partial z}(\alpha z, \theta_2) \right] d\alpha \right) z \\ &= \left(\int_0^1 \left[\int_0^1 \frac{\partial^2 \pi}{\partial \theta \partial z}(\alpha z, \beta \theta_1 + (1-\beta)\theta_2)(\theta_1 - \theta_2)\, d\beta \right] d\alpha \right) z\,. \end{aligned}$$

Now supposing $\|z\| \leqslant \beta B_0 L_I + \eta$ and $\|\theta_i\| \leqslant B_\theta$ for $i = 1, 2$, then

$$\|\pi(z; \theta_1) - \pi(z; \theta_2)\| \leqslant L_{\partial^2 \pi_z} \|z\| \|\theta_1 - \theta_2\|\,.$$

Now, let $\{\tilde{\varphi}_t^{\pi_d}(\xi)\}_{t=0}^{T-1}$ the state trajectory of system (2.1) under the policy $\pi_d \in \Pi_z$ under the *noiseless* generative model $z_t = I(x_t)$. Observe that under the loss function introduced in Section 5,

$$
\begin{aligned}
|\ell_{\pi_d}(\xi;\pi_1,\pi_t) - \ell_{\pi_d}(\xi;\pi_2,\pi_t)| &\leqslant \sum_{t=0}^{T-1} \|g(\tilde{\varphi}_t^{\pi_d}(\xi))(\pi_1(I(\tilde{\varphi}_t^{\pi_d}(\xi)),\theta_1) - \pi_2(I(\tilde{\varphi}_t^{\pi_d}(\xi)),\theta_2)\| \\
&\leqslant B_g \sum_{t=0}^{T-1} \|\pi_1(I(\tilde{\varphi}_t^{\pi_d}(\xi)),\theta_1) - \pi_2(I(\tilde{\varphi}_t^{\pi_d}(\xi)),\theta_2)\| \\
&\overset{(a)}{\leqslant} B_g L_{\partial^2 \pi_z} \left(\sum_{t=0}^{T-1} \|I(\tilde{\varphi}_t^{\pi_d}(\xi))\| \right) \|\theta_1 - \theta_2\| \\
&\overset{(b)}{\leqslant} B_g L_{\partial^2 \pi_z} L_I \left(\sum_{t=0}^{T-1} \|\tilde{\varphi}_t^{\pi_d}(\xi)\| \right) \|\theta_1 - \theta_2\|,
\end{aligned}
$$

where (a) follows because $\|I(\tilde{\varphi}_t^{\pi_d}(\xi))\| \leqslant L_I \beta \rho^t \|\xi\| + \|e(t)\| \leqslant \beta B_0 L_I + \eta$, and (b) from the properties of the generative model. We now use that $\pi_d \circ I \in S_\delta(\beta,\rho,\gamma)$, i.e., that it leads to a closed-loop system that is (β,ρ,γ)-E-δISS, to couple the trajectory $\{\tilde{\varphi}_t^{\pi_d}(\xi)\}_{t=0}^{T-1}$ induced by the policy π_d acting based on the noiseless generative model $z(x_t) = I(x_t)$ to those acting based on the noisy generative model (D.1). In particular, let $x_{t+1} = f(x_t) + g(x_t)\pi_d(I(x_t)) := f_{\text{cl}}^0$ and $y(t+1) = f(x_t) + g(x_t)\pi_d(I(x_t) + e_t)$. Then we can write $y_{t+1} = f_{\text{cl}}^0 + \delta(y_t)$, where

$$
\delta(y_t) := g(y_t)(\pi_d(I(y_t) + e_t) - \pi_d(I(y_t))).
$$

It follows immediately that $\|\delta(y_t)\| \leqslant B_g L_\Pi \eta$ for all y, t, and given that f_{cl}^0 is (β,ρ,γ)-E-δISS by assumption, we have that

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

for all initial conditions x_0 and y_0 and input signals u_t. In particular, if we set $x_0 = y_0$ and $u_t = \delta(y_t)$, this implies that

$$
\sum_{t=0}^{T-1} \|\varphi_t^{\pi_d}(\xi) - \tilde{\varphi}_t^{\pi_d}(\xi)\| \leqslant \frac{\gamma B_g L_\Pi}{1-\rho} \sum_{t=0}^{T-1} \|e_t\|.
$$

Similarly, setting $x_0 = 0$ and $u_t = \delta(y_t)$ we see that

$$
\sum_{t=0}^{T-1} \|\varphi_t^{\pi_d}(\xi)\| \leqslant \frac{\beta B_0}{1-\rho} + \frac{\gamma B_g L_\Pi}{1-\rho} \sum_{t=0}^{T-1} \|e_t\|.
$$

We can therefore write

$$
\begin{aligned}
|\ell_{\pi_d}(\xi;\pi_1,\pi_t)-\ell_{\pi_d}(\xi;\pi_2,\pi_t)| &\leqslant B_g L_{\partial^2\pi_z} L_I \left(\sum_{t=0}^{T-1}\|\tilde{\varphi}_t^{\pi_d}(\xi)\|\right)\|\theta_1-\theta_2\| \\
&\leqslant B_g L_{\partial^2\pi_z} L_I \left(\sum_{t=0}^{T-1}\|\varphi_t^{\pi_d}(\xi)\|+\sum_{t=0}^{T-1}\|\tilde{\varphi}_t^{\pi_d}(\xi)-\varphi_t^{\pi_d}(\xi)\|\right)\|\theta_1-\theta_2\| \\
&\leqslant \frac{B_g L_{\partial^2\pi_z} L_I}{1-\rho}\left(\beta B_0+2\gamma B_g L_\Pi B_\eta\right)\|\theta_1-\theta_2\|,
\end{aligned}
$$

where $B_\eta=\eta\sqrt{T}$ if $\|e\|_{\ell_2(T)}\leqslant\eta$, $\mathcal{B}_\eta=S\eta_s$ if e_t follows the sparse noise model, and $B_\eta=0$ recovers noiseless generative model of Section 5 and the results of Proposition 5.1.

As in the proof of Proposition B.1, this shows that the loss function ℓ_{π_d} is uniformly Lipschitz with respect to the parameter θ. Therefore by similar arguments,

$$
\mathcal{R}_n(\Pi_z,I,e)\leqslant\frac{32.5B_\theta B_g L_{\partial^2\pi_z}L_I}{1-\rho}\left(\beta B_0+2\gamma B_g L_\Pi B_\eta\right)\sqrt{\frac{k}{n}}.
$$

□

D.2 Modular Perception-Based IL

In the modular approach, our goal is to learn both a perception-map $p(z)$ from perceptual-measurement z to *state-estimate* $p(z)=\hat{x}$, i.e., $p:\mathbb{R}^D\to\mathbb{R}^p$ as well as a state-feedback based policy $\pi_x:\mathbb{R}^p\to\mathbb{R}^q$ acting on the the state-estimate, such that the composition $\pi_x(p(z))$ imitates the expert $\pi_\star(x)$. To that end, we assume that the perception-maps p lie in a perception-map class $\mathcal{P}$ of L_p-Lipschitz maps, from which it follows that the class of perception based policies defined as:

$$
\Pi_{x,p}:=\left\{\pi_{x,p}(z)=\pi_x(p(z))\,|\,p\in\mathcal{P},\pi\in\Pi_x\right\}, \tag{D.5}
$$

for a (L_x,B_x)-regular policy class Π_x is itself a $(L_xL_pL_I,B_x)$-regular policy class with respect to x under the generative model (5.1). We assume that $\pi_\star\in\Pi_x$ and that $I^{-1}\in\mathcal{P}$ such that $\pi_\star\circ I^{-1}\in\Pi_{x,p}$

In order to simultaneously optimize the perceptual and control error, we propose the following modification to optimization problem (3.1) in Algorithm 2, which is now assumed to take as an input a rollout:

$$
\begin{array}{ll}
\underset{\bar{\pi}_x\in\Pi_x,\bar{p}\in\mathcal{P}}{\text{minimize}} & \frac{1}{n/E}\sum_{i=1}^{n/E}\ell_{\pi_{\text{roll}}}(\xi_i^k;\bar{\pi}_x\circ\bar{p},\pi_\star\circ I^{-1}) \\
\text{subject to} & \frac{1}{n/E}\sum_{i=1}^{n/E}\ell_{\pi_{\text{roll}}}(\xi_i^k;\bar{\pi}_x\circ\bar{p},\pi_{\text{roll}})\leqslant c_k, \\
& \frac{1}{n/E}\sum_{i=1,t=0}^{n/E,T-1}\|p(z(\varphi_t^{\pi_k}(\xi_i^k)))-\varphi_t^{\pi_k}(\xi_i^k)\|\leqslant d_k \\
& \quad(1-w)^{-1}\left[(1-\alpha)\pi_{\text{roll}}\circ I+\alpha\bar{\pi}_x\circ\bar{p}\circ I-w\pi_\star\right]\in S_\delta(\beta,\rho,\gamma),
\end{array} \tag{D.6}
$$

where for policies $\pi_d,\pi_1,\pi_2\in\Pi_{x,p}$, we overload the loss notation $\ell_{\pi_d}=\sum_{t=0}^{T-1}\Delta_{\pi_1\circ I,\pi_2\circ I}(\varphi_t^{\pi_d\circ I}(\xi))$. It can be checked that the resulting $\Delta_{\pi_1\circ I,\pi_2\circ I}(\varphi_t^{\pi_d\circ I}(x))$ is L_Δ-Lipschitz with respect to x under the noisy generative model (D.1) (i.e., that $z_t=I(x_t)+e_t$), and as we assume that $\pi_\star\in\Pi_x$, $I^{-1}\in\mathcal{P}$, setting $\pi_0=\pi_\star$ allows for a similar argument to the proof of Lemma D.1 to show that Algorithm

1 using optimization problem (D.6) for subroutine cERM defined in Algorithm 2 is recursively feasible and outputs a purely perception-based policy for any choice of $d_k \geqslant 0$.

It follows that Theorem 3.1 and Proposition 3.2 are applicable to learning a modular perception-based controller in this setting using a state-feedback-based expert $\pi_\star(x)$ under the generative model (5.1), and all that remains is to characterize the Rademacher complexity:

$$\mathcal{R}_n(\Pi_{x,p}, I, e) := \sup_{\substack{\pi_d \in \Pi_{x,p} \\ \pi_d \circ I \in S_\delta(\beta,\rho,\gamma)}} \sup_{\pi_g \in \Pi_{x,p}} \mathbb{E}_{\{\xi_i\},\{\varepsilon_i\}} \left[\sup_{\pi \in \Pi_{x,p}} \frac{1}{n} \sum_{i=1}^n \varepsilon_i \ell_{\pi_d}(\xi_i; \pi, \pi_g) \right], \tag{D.7}$$

for suitable choices of policy classes Π_x, perception-map classes $\mathcal{P}$, and noisy generative models (D.1).

Define

$$\begin{aligned} \Pi_{x,p} &:= \{\pi_x(p(z)) \,|\, p \in \mathcal{P},\ \pi_x \in \Pi_x\}, &\text{(D.8)}\\ \Pi_x &:= \{\pi_x(x;\theta) \,|\, \theta \in \mathbb{R}^k,\ \|\theta\| \leqslant B_\theta\}, &\text{(D.9)}\\ \mathcal{P} &:= \{p(z) \,|\, p(0) = 0,\ p \text{ is } L_p\text{-Lipschitz}\}. &\text{(D.10)} \end{aligned}$$

Proposition D.3. *Fix a data generating policy $\pi_d \circ p \in \Pi_{x,p}$. Suppose that $\pi_d \circ p$ is (β, ρ, γ)-E-δISS when $e_t \equiv 0$ in the generative model (D.1), and let Π_x be as defined in equation (D.9) for a fixed map π_x that is twice continuously differentiable, and let $\mathcal{P}$ be as defined in equation (D.10). Define $L_{\partial^2 \pi_x}$ to be:*

$$L_{\partial^2 \pi_x} := \sup_{\|x\| \leqslant \beta B_0 L_p L_I + L_p \eta, \|\theta\| \leqslant B_\theta} \left\| \frac{\partial^2 \pi_x}{\partial \theta \partial x} \right\|_{\ell^2(\mathbb{R}^k) \to M(\mathbb{R}^{q \times p})}.$$

Then if the perceptual-measurements z_t follow the generative model (D.1):

$$\mathcal{R}_n(\Pi_{x,p}, I, e) \leqslant \frac{32.5 B_\theta B_g L_{\partial^2 \pi_x} L_p L_I}{1-\rho} \left(\beta B_0 + 2\gamma B_g L_\Pi L_p \eta \sqrt{T} \right) \sqrt{\frac{k}{n}}. \tag{D.11}$$

If we further assume that e_t is $S-$sparse, i.e., has at most S-nonzero elements over $t = 0, \ldots, T-1$, and that $\|e_t\| \leqslant \eta_s$, then the Rademacher complexity (D.2) is given by

$$\mathcal{R}_n(\Pi_{x,p}, I, e) \leqslant \frac{32.5 B_\theta B_g L_{\partial^2 \pi_x} L_p L_I}{1-\rho} \left(\beta B_0 + 2\gamma B_g L_\Pi L_p S \eta_s \right) \sqrt{\frac{k}{n}}. \tag{D.12}$$

Proof. Fix a z, p_1, p_2, and θ_1, θ_2. Because $\pi(0, \theta) = 0$ for all θ, we have by repeated applications of the Fundamental Theorem of Calculus that:

$$\pi(p_1(z), \theta_1) - \pi(p_2(z), \theta_2) = \\ \left(\int_0^1 \left[\int_0^1 \frac{\partial^2 \pi}{\partial \theta \partial p} (\alpha p_1(z) + (1-\alpha) p_2(z); \beta\theta_1 + (1-\beta)\theta_2)(\theta_1 - \theta_2)\, d\beta \right] d\alpha \right) (p_1(z) - p_2(z)).$$

Now supposing $\|p_i(z)\| \leqslant \beta B_0 L_I L_p + L_p \eta$ and $\|\theta_i\| \leqslant B_\theta$ for $i = 1, 2$, then

$$\|\pi(p_1(z); \theta_1) - \pi(p_2(z); \theta_2)\| \leqslant L_{\partial^2 \pi_x} \|p_1(z) - p_2(z)\| \|\theta_1 - \theta_2\|.$$

Now, let $\{\tilde{\varphi}_t^{\pi_d}(\xi)\}_{t=0}^{T-1}$ be the state trajectory of system (2.1) under the policy $\pi_d \in \Pi_{x,z}$ under the *noiseless* generative model $z(t) = I(x(t))$. Observe that under the loss function introduced in Section 5,

$$
\begin{aligned}
|\ell_{\pi_d}(\xi;\pi_1,\pi_t) - \ell_{\pi_d}(\xi;\pi_2,\pi_t)| &\leqslant \sum_{t=0}^{T-1} \|g(\tilde{\varphi}_t^{\pi_d}(\xi))(\pi_x(p_1(I(\tilde{\varphi}_t^{\pi_d}(\xi))),\theta_1) - \pi_x(p_2(I(\tilde{\varphi}_t^{\pi_d}(\xi))),\theta_2)\| \\
&\leqslant B_g \sum_{t=0}^{T-1} \|\pi_x(p_1(I(\tilde{\varphi}_t^{\pi_d}(\xi))),\theta_1) - \pi_x(p_2(I(\tilde{\varphi}_t^{\pi_d}(\xi))),\theta_2)\| \\
&\overset{(a)}{\leqslant} B_g L_{\partial^2\pi_x} \left(\sum_{t=0}^{T-1} \|p_1(I(\tilde{\varphi}_t^{\pi_d}(\xi))) - p_2(I(\tilde{\varphi}_t^{\pi_d}(\xi)))\|\right) \|\theta_1 - \theta_2\| \\
&\overset{(b)}{\leqslant} 2B_g L_{\partial^2\pi_x} L_p L_I \left(\sum_{t=0}^{T-1} \|\tilde{\varphi}_t^{\pi_d}(\xi))\|\right) \|\theta_1 - \theta_2\| \,,
\end{aligned}
$$

where (a) follows because $\|p_i(I(\tilde{\varphi}_t^{\pi_d}(\xi)))\| \leqslant L_p L_I \beta\rho^t\|\xi\| + L_p\|e(t)\| \leqslant \beta B_0 L_I L_p + L_p\eta$, and (b) from the properties of the generative model. We now use that π_d leads to a closed-loop system that is (β,ρ,γ)-E-δISS to couple the trajectory $\{\tilde{\varphi}_t^{\pi_d}(\xi)\}_{t=0}^{T-1}$ induced by the policy $\pi_d(p(z))$ acting based on the noiseless generative model $z(x_t) = I(x_t)$ to those acting based on the noisy generative model (D.1). In particular, let $x_{t+1} = f(x_t) + g(x_t)\pi_d(I(x_t) := f_{\mathrm{cl}}^0$ and $y_{t+1} = f(x_t) + g(x_t)\pi_d(I(x_t) + e_t)$. Then we can write $y_{t+1} = f_{\mathrm{cl}}^0 + \delta(y_t)$, where

$$
\delta(y_t) := g(y_t)(\pi_d(I(y_t) + e_t) - \pi_d(I(y_t))).
$$

It follows immediately that $\|\delta(y_t)\| \leqslant B_g L_\Pi L_p \eta$ for all y, t, and given that f_{cl}^0 is (β,ρ,γ)-E-δISS by assumption, we have that

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

for all initial conditions x_0 and y_0 and input signals u_t. In particular, if we set $x_0 = y_0$ and $u_t = \delta(y_t)$, this implies that

$$
\sum_{t=0}^{T-1}\|\varphi_t^{\pi_d}(\xi) - \tilde{\varphi}_t^{\pi_d}(\xi)\| \leqslant \frac{\gamma B_g L_\Pi L_p}{1-\rho}\sum_{t=0}^{T-1}\|e_t\|.
$$

Similarly, setting $x_0 = 0$ and $u_t = \delta(y_t)$ give us that

$$
\sum_{t=0}^{T-1}\|\varphi_t^{\pi_d}(\xi)\| \leqslant \frac{\beta B_0}{1-\rho} + \frac{\gamma B_g L_\Pi L_p}{1-\rho}\sum_{t=0}^{T-1}\|e_t\|.
$$

We can therefore write

$$
\begin{aligned}
|\ell_{\pi_d}(\xi;\pi_1,\pi_t)-\ell_{\pi_d}(\xi;\pi_2,\pi_t)| &\leqslant 2B_g L_{\partial^2\pi_x} L_p L_I \left(\sum_{t=0}^{T-1}\|\tilde{\varphi}_t^{\pi_d}(\xi)\|\right)\|\theta_1-\theta_2\| \\
&\leqslant 2B_g L_{\partial^2\pi_x} L_p L_I \left(\sum_{t=0}^{T-1}\|\varphi_t^{\pi_d}(\xi)\|+\sum_{t=0}^{T-1}\|\tilde{\varphi}_t^{\pi_d}(\xi)-\varphi_t^{\pi_d}(\xi)\|\right)\|\theta_1-\theta_2\| \\
&\leqslant \frac{2B_g L_{\partial^2\pi_x} L_p L_I}{1-\rho}\left(\beta B_0+2\gamma B_g L_\Pi L_p B_\eta\right)\|\theta_1-\theta_2\|,
\end{aligned}
$$

where $B_\eta=\eta\sqrt{T}$ if $\|e\|_{\ell_2(T)}\leqslant\eta$, $B_\eta=S\eta_s$ if e_t follows the sparse noise model, and setting $B_\eta=0$ recovers the noiseless generative model (5.1).

As in the proof of Proposition B.1, this shows that the loss function ℓ_{π_d} is uniformly Lipschitz with respect to the parameter θ. Therefore by similar arguments,

$$
\mathcal{R}_n(\Pi_{x,p},I,e)\leqslant\frac{32.5B_\theta B_g L_{\partial^2\pi_x} L_p L_I}{1-\rho}\left(\beta B_0+2\gamma B_g L_\Pi L_p B_\eta\right)\sqrt{\frac{k}{n}}.
$$

□

Remark D.4. *Note that this approach can also be used to convert a state-feedback based policy $\pi_\star(x)$ to a perception-based policy $\pi_\star(p(z))$ by only learning a perception map at each epoch, and keeping $\bar{\pi}_x=\pi_\star$ fixed.*

D.3 Sample Complexity Bounds and Discussion

Proposition D.5 (End-to-end Perception-Based IL)**.** *Under the assumptions of Theorem 3.1, if $\alpha E\geqslant\log\left(2+\frac{2L_\Delta\gamma}{1-\rho}\right)$, with probability at least $1-\delta$ over $\left\{\{\xi_i^k\}_{i=1}^{n/E}\right\}_{k=0}^{E-1}\sim\mathcal{D}^n$:*

$$
\begin{aligned}
&\mathbb{E}_{\xi\sim\mathcal{D}}\ell_{\pi_E}(\xi;\pi_E,\pi_\star)\leqslant \\
&12\alpha\left(\frac{\gamma L_\Delta}{1-\rho}\right)\left[\frac{65B_\theta B_g L_{\partial^2\pi_z} L_I}{1-\rho}\left(\beta B_0+2\gamma B_g L_\Pi B_\eta\right)\sqrt{\frac{K}{n}}+\frac{\beta B_0 L_\Delta E}{1-\rho}\sqrt{\frac{E}{n}\log(4E/\delta)}+\sum_{i=1}^{E-2}c_i\right],
\end{aligned}
\tag{D.13}
$$

and with probability at least $1-\delta$ over $\left(\left\{\{\xi_i^k\}_{i=1}^{n/E}\right\}_{k=0}^{E-1},\xi\right)\sim\mathcal{D}^{n+1}$:

$$
\begin{aligned}
&h_j(\{\varphi_t^{\pi_E}(\xi)\}_{t=0}^{T-1})\leqslant-\gamma_h \\
&+\frac{24\alpha\gamma^2 L_\Delta L_h}{\delta(1-\rho)^2}\left[\frac{65B_\theta B_g L_{\partial^2\pi_z} L_I}{1-\rho}\left(\beta B_0+2\gamma B_g L_\Pi B_\eta\right)\sqrt{\frac{K}{n}}+\frac{\beta B_0 L_\Delta E}{1-\rho}\sqrt{\frac{E}{n}\log(8E/\delta)}+\sum_{i=1}^{E-2}c_i\right],
\end{aligned}
\tag{D.14}
$$

where $B_\eta=\eta\sqrt{T}$ if $\|e\|_{\ell_2(T)}\leqslant\eta$, $B_\eta=S\eta_s$ if e_t follows the sparse noise model, and setting $B_\eta=0$ recovers the noiseless generative model (5.1).

Proof. Follows immediately from Theorem 3.1 and Proposition 3.2 equipped with a $(L_z L_I, B_z)$-regular policy class, and Proposition 5.1. $\square$

Proposition D.6 (Modular Perception-Based IL)**.** *Under the assumptions of Theorem 3.1, if* $\alpha E \geqslant \log\left(2 + \frac{2L_\Delta \gamma}{1-\rho}\right)$*, with probability at least* $1-\delta$ *over* $\left\{\{\xi_i^k\}_{i=1}^{n/E}\right\}_{k=0}^{E-1} \sim \mathcal{D}^n$*:*

$$\begin{aligned}&\mathbb{E}_{\xi\sim\mathcal{D}}\ell_{\pi_E}(\xi;\pi_E,\pi_\star) \leqslant \\ &12\alpha\left(\frac{\gamma L_\Delta}{1-\rho}\right)\left[\frac{65 B_\theta B_g L_{\partial^2\pi_x} L_p L_I}{1-\rho}\left(\beta B_0 + 2\gamma B_g L_p L_\Pi B_\eta\right)\sqrt{\frac{K}{n}} + \frac{\beta B_0 L_\Delta E}{1-\rho}\sqrt{\frac{E}{n}\log(4E/\delta)} + \sum_{i=1}^{E-2} c_i\right],\end{aligned} \tag{D.15}$$

and with probability at least $1-\delta$ *over* $\left(\left\{\{\xi_i^k\}_{i=1}^{n/E}\right\}_{k=0}^{E-1}, \xi\right) \sim \mathcal{D}^{n+1}$*:*

$$\begin{aligned}&h_j(\{\varphi_t^{\pi_E}(\xi)\}_{t=0}^{T-1}) \leqslant -\gamma_h \\ &+\frac{24\alpha\gamma^2 L_\Delta L_h}{\delta(1-\rho)^2}\left[\frac{65 B_\theta B_g L_{\partial^2\pi_x} L_p L_I}{1-\rho}\left(\beta B_0 + 2\gamma B_g L_p L_\Pi B_\eta\right)\sqrt{\frac{K}{n}} + \frac{\beta B_0 L_\Delta E}{1-\rho}\sqrt{\frac{E}{n}\log(8E/\delta)} + \sum_{i=1}^{E-2} c_i\right],\end{aligned} \tag{D.16}$$

where $B_\eta = \eta\sqrt{T}$ *if* $\|e\|_{\ell_2(T)} \leqslant \eta$*,* $B_\eta = S\eta_s$ *if* e_t *follows the sparse noise model, and setting* $B_\eta = 0$ *recovers the noiseless generative model* (5.1)*.*

Proof. Follows immediately from Theorem 3.1 and Proposition 3.2 equipped with a $(L_x L_p L_I, B_x)$-regular policy class, and Proposition D.3. $\square$

Which architecture is best? Notice that the terms under the radicals in Propositions D.5 and D.6 differ in terms of the function class free parameters K and k, where we typically expect $K > k$ given that in the end-to-end architecture, $\pi_z : \mathbb{R}^d \to \mathbb{R}^q$ (we assume the approach described in Remark 5.2 is taken to reduce D to d), whereas in the modular architecture, $\pi_x : \mathbb{R}^p \to \mathbb{R}^q$, where typically $d > p$. Furthermore, we expect $L_{\partial^2\pi_x} < L_{\partial^2\pi_z}$ for analogous reasons. However, we note that because we assume $I^{-1} \in \mathcal{P}$, then if the generative model (5.1) is not well-conditioned, i.e., if $L_{I^{-1}}$ is large, we may see a considerable degradation in the modular Rademacher bounds (D.11), (D.12) without a corresponding increase in the end-to-end Rademacher bounds (D.3), (D.4). Finally we note that despite the naming convention that we adopt, the modular architecture is trained "end-to-end" as both the state-based policy π_x and the perception-map p are learned simultaneously. We therefore conjecture that the modular architecture is likely to be better performing than the end-to-end architecture given a fixed number of trajectories when the generative model is well-conditioned, that is to say, when $L_{I^{-1}}$ is not too large. We leave empirical investigation of this hypothesis to future work.

E Experiment Details

E.1 Linear System Experiments

For this experiment, we used the linear dynamical system with matrices $A \in \mathbb{R}^{10\times 10}$ and $B \in \mathbb{R}^{10\times 4}$, where $\|A\| = 5.893$ and $\|B\| = 4.964$. The open loop system is unstable, with $\rho(A) =$

3.638. We synthesize the optimal state-feedback LQR controller for the dynamics (A, B) and prescribed cost matrices (Q, R) by solving the Discrete Algebraic Riccati Equation (DARE) using `scipy.linalg.solve_discrete_are`. The resulting closed loop LQR norms of the resulting systems for $\nu = 0.01$, 0.001, and 0.0001 were 10.909, 10.784, and 10.739 respectively.

We drew initial conditions according to the distribution $\mathcal{N}(0, 4)$. We used a policy parameterized by a two-hidden-layer feed-forward neural network with ReLU activations. Each hidden layer in this network had a width of 64 neurons. For both BC and CMILe without stability constraints, we train the policy for 500 epochs; for CMILe with stability constraints, we trained policies for 1000 epochs. All neural networks were optimized with the Adam optimizer (Kingma and Ba, 2014) with a learning rate of 0.01.

Stability experiments. In order to compute a robust Lyapunov function for use in the stability experiments, we use the following approach. Let $A_{\mathrm{lqr}} := A + BK_{\mathrm{lqr}}$ and consider the following Lyapunov equation:

$$A_{\mathrm{lqr}}^{\mathsf{T}} X A_{\mathrm{lqr}} - \gamma^2 X + \varepsilon I = 0,$$

for $\gamma \in (0, 1)$ and $\varepsilon > 0$. Note that by rewriting this equation as

$$\left(\frac{A_{\mathrm{lqr}}}{\gamma}\right)^{\mathsf{T}} X \left(\frac{A_{\mathrm{lqr}}}{\gamma}\right) - X + \frac{\varepsilon}{\gamma^2} I = 0,$$

we see that this equation has a unique positive define solution so long as A_{lqr}/γ is stable, i.e., so long as $\gamma \geqslant \rho(A_{\mathrm{lqr}})$ Zhou et al. (1996).

Let $X_\star$ denote the solution to the above Lyapunov equation. We now make two obervations. First, notice that $X_\star$ can be used to define a quadratic Lyapunov function $V(x) = x^{\mathsf{T}} X_\star x$ satisfying the assumptions of Lemma 4.4:

$$\begin{aligned} V(x_{t+1}) &= x_{t+1}^{\mathsf{T}} X_\star x_{t+1} \\ &= x_t^{\mathsf{T}} A_{\mathrm{lqr}}^{\mathsf{T}} X_\star A_{\mathrm{lqr}} x_t \\ &= x_t^{\mathsf{T}} (\gamma^2 X_\star - \varepsilon I) x_t \\ &\leqslant \gamma^2 x_t^{\mathsf{T}} X_\star x_t^{\mathsf{T}} = \gamma^2 V(x_t) \end{aligned}$$

Then note that we can rewrite the Lyapunov equation as

$$A_{\mathrm{lqr}}^{\mathsf{T}} X_\star A_{\mathrm{lqr}} - X_\star + Q = 0,$$

for $Q = (1 - \gamma^2) X_\star + \varepsilon I$. To that end, we suggest solving the Lyapunov equation

$$A_{\mathrm{lqr}}^{\mathsf{T}} X A_{\mathrm{lqr}} - X + Q = 0,$$

with $Q = (1 - \gamma^2) P_\star + \varepsilon I$, for a small $\varepsilon > 0$ and $P_\star$ the solution to the DARE, so as to obtain a a Lyapunov certificate with similar convergence properties to that of the expert, where ε trades off between how small γ can be and how robust the resulting Lyapunov function is to mismatches between the learned policy and the expert policy. We note that the existence of solutions to this Lyapunov equation are guaranteed by continuity of the solution of the Lyapunov equation and that the solution to the DARE $P_\star$ is the maximizing solution among symmetric solutions (Dullerud and Paganini, 2013).

E.2 MPC Experiments

For the MPC experiment, we used the linear dynamical system with matrices $A \in \mathbb{R}^{10\times 10}$ and $B \in \mathbb{R}^{10\times 4}$, where $\|A\| = 5.406$ and $\|B\| = 3.739$. As in the linear system experiment, the open loop system is unstable, with $\rho(A) = 2.596$.

As described in §6, the constraints where that $|x_t| \leqslant 4$ and $|u_t| \leqslant 3$ for all t. We drew initial conditions according to the distribution $\mathrm{Unif}([-1, 1])$. This conservatism in initial conditions relative to the constraints resulted in robust feasibility during much of the training procedure; however, when the expert policy encountered infeasibility during training, we heuristically doubled the size of the feasible set until feasibility was achieved. We note that more principled approaches to constructing robust forward invariant sets that guarantee robust feasibility of the expert during training are possible, see for example, the methods described in Hertneck et al. (2018). However, given that robust MPC was not the focus of this paper, we opted for this simple yet effective approach.

We used a parameterized policy. We used a policy parameterized by a two-hidden-layer feed-forward neural network with ReLU activations. Each hidden layer in this network had a width of 64 neurons. All neural networks were trained for 500 epochs with the Adam optimizer using a learning rate of 0.01.

E.3 Segway Experiments

Our segway implementation is based heavily on the code accompanying Dean et al. (2020a)[5]. The continuous dynamics are derived using the constrained Euler-Lagrange equations. We integrate the continuous dynamics using a standard RK4 integrator for $T = 5$ seconds with $dt = 0.001$, and each control input is zero-order-held for 20 timesteps; therefore, each trajectory is discretized into 250 time-steps.

The segway state $z \in \mathbb{R}^7$ is a 7-dimensional vector $z = (x, y, \psi, \dot{x}, \dot{\psi}, \theta, \dot{\theta})$. The y and ψ coordinates are held constant throughout, since the segway is constrained to move only in the x-direction by applying equal torque to both wheels. Hence, the relevant degrees of freedom are $(x, \dot{x}, \theta, \dot{\theta})$. The (relevant degrees of) the goal state are $(0, 0, 0.138, 0)$. The CBF-QP is described in Dean et al. (2020a), and is implemented using the `ecos` optimization library. We use the version of the CBF-QP with the robustness parameter ε set to zero. In our experiments, whenever the CBF-QP optimization was infeasible, we simply play the unfiltered input. We note that infeasiblity only occurred when the expert was asked to label a trajectory that was very far from optimal.

We use several different initial distributions $\mathcal{D}$ over the initial state $z(0)$ which we describe as follows. For the state-based experiments (Figure 3 and Figure 4), we draw $x \sim \mathrm{Unif}([-4, 4])$ and $\theta \sim 0.138 + \mathrm{Unif}([-.4, .4])$. For the perception-based experiments (Figure 5), we draw x from mixture model $\frac{1}{2}\mathrm{Unif}([-2, -1]) + \frac{1}{2}\mathrm{Unif}([1, 2])$ and $\theta \sim 0.138 + \mathrm{Unif}([-.2, .2])$. For pretraining our mid-level representations, we use $x \sim \mathrm{Unif}([-3, 3])$ and $\theta \sim 0.138 + \mathrm{Unif}([-1, 1])$. All velocity coordinates (i.e., $\dot{x}$ and $\dot{\theta}$) are initialized at zero.

For the state-based experiments, we use a policy parameterized by a 7-64-64-1 MLP. For both BC and CMILe, we train the imitation loss with 200 epochs of Adam using a step size of 10^{-4} and batch size of 512. For CMILe, after the first epoch, we initialize the weights with the weights found by the previous epoch.

[5]https://github.com/rkcosner/cyberpod_sim_ros

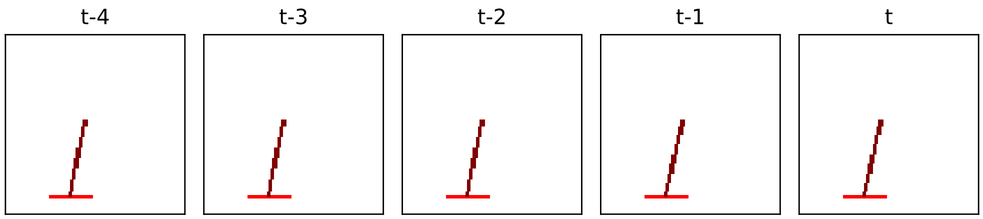

Figure 6: Frames provided as visual input to the perception-based policy.

```
hk.Sequential([
  lambda x: x / 255.,
  hk.Conv3D(32, kernel_shape=3, stride=2,
            padding='VALID', data_format='NDHWC'),
  jax.nn.relu,
  hk.Conv3D(64, kernel_shape=(2, 3, 3), stride=2,
            padding='VALID', data_format='NDHWC'),
  lambda x: jnp.squeeze(x, axis=1),
  jax.nn.relu,
  hk.Conv2D(128, kernel_shape=3, stride=2),
  jax.nn.relu,
  hk.Conv2D(256, kernel_shape=3, stride=2),
  jax.nn.relu,
  hk.Conv2D(256, kernel_shape=3, stride=2),
  jax.nn.relu,
  lambda x: x.reshape((x.shape[0], -1)),
  hk.Linear(64),
  jax.nn.relu,
  hk.Linear(64),
  jax.nn.relu,
  hk.Linear(1)
])
```

Figure 7: CNN policy architecture.

For the perception-based experiments, we render the segway into an 84×84 RGB image using `PIL`[6]. In order to provide enough context to estimate velocities, we append the last four frames. An example of the visual input supplied to the policy is shown in Figure 6.

For pretraining the CNN features, we use 100 epochs of Adam with learning rate 10^{-3}, batch size 512, and weight decay 10^{-3}. As in the state-based case, for both BC and CMILe, we train the imitation loss with 200 epochs of Adam using a step size of 10^{-4} and batch size of 512. Again, for CMILe, after the first epoch, we initialize the weights with the weights found by the previous epoch. Figure 7 describes our CNN architecture using the `haiku` library[7].

[6] https://pillow.readthedocs.io/en/stable/
[7] https://github.com/deepmind/dm-haiku