On the Sample Complexity of Stability Constrained Imitation Learning

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

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

June 6, 2021

Abstract

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

1 Introduction

Imitation Learning (IL) techniques [1, 2] 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 [3, 4], and can lead to policies that are more computationally efficient to evaluate online [5, 6] than optimization-based experts. Indeed, there is a rich body of work demonstrating the advantages of IL-based methods in a range of applications including video-game playing [4, 7], humanoid robotics [8], and self-driving cars [9]. Safe IL further seeks to provide guarantees on the stability or safety properties of policies produced by IL. Methods drawing on tools from Bayesian deep learning [10–12], PAC-Bayes [13], stability regularization [14], or robust control [5, 6], are able to provide varying levels of guarantees in the context of IL.

However, when applied to continuous control problems, little to no insight is given into how the underlying stability properties of the expert policy affect the sample-complexity of the resulting IL task. In this paper, we address this gap and answer the question: what makes an expert policy easy to learn? Our main insight is that when an expert policy satisfies a suitable quantitative notion of

robust *incremental* stability, i.e., when pairs of system trajectories under the expert policy robustly converge towards each other, and when learned policies are also constrained to satisfy this property, then IL can be made provably efficient. We formalize this insight through the notion of *incremental gain stability* constrained IL algorithms, and in doing so, quantify and generalize previous observations of efficient and robust learning subject to contraction based stability constraints.

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

Nonlinear stability and learning from demonstrations: Our work applies tools from nonlinear stability theory to analyze the sample-complexity of IL algorithms. Concepts from nonlinear stability theory, such as Lyapunov stability or contraction theory [15], have also been successfully applied to learn autonomous nonlinear dynamical systems satisfying desirable properties such as (incremental) stability or controllability. As demonstrated empirically in [14, 16–18], using such stability-based regularizers to trim the hypothesis space results in more data-efficient and robust learning algorithms – however, no quantitative sample-complexity bounds are provided. Most similar in spirit to our work are the results of Boffi et al. [19, 20], which leverage Lyapunov stability and contraction theory to provide finite sample-complexity guarantees for adaptive nonlinear control [19] and learning stability certificates [20] under assumptions of *exponential* incremental-input-to-state-stability.

IL under covariate shift: Vanilla IL (e.g., Behavior Cloning) is known to be sensitive to covariate shift: as soon as the learned policy deviates from the expert policy, errors begin to compound, leading the system to drift to new and possibly dangerous states [4, 21]. Representative IL algorithms that address this issue include DAgger [4] and DART [22]. 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 loss functions that are strongly convex in the policy parameters, DAgger enjoys $\tilde{O}(T)$ sample-complexity in the task horizon T – this bound degrades to $\tilde{O}(T^2)$ when loss functions are only convex. 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. IL algorithms more explicitly focused on stability and safety include methods leveraging tools from Bayesian deep learning [11, 12], robust model-predictive-control [5], robust control [6], and PAC-Bayes theory [13]. While the approach, generality, and strength of guarantees provided by the aforementioned works vary, none provide insight as to how the underlying stability properties of an expert policy affect the sample-complexity of the corresponding IL task.

Contributions.   To provide fine-grained insights into the relationship between system stability and sample-complexity, we first define and analyze the notion of incremental gain stability (IGS) for a nonlinear dynamical system. IGS provides a quantitative measure of robust convergence between system trajectories, that in our context strictly expands on the guarantees provided by contraction theory [15] by allowing for a graceful degradation away from exponential convergence rates.

We then propose and analyze the sample-complexity properties of IGS-constrained imitation learning algorithms, and show that the graceful degradation in stability translates into a corre-

sponding degradation of generalization bounds by linking nonlinear stability and statistical learning theory. In particular, we show that when imitating an IGS expert policy, IGS-constrained behavior cloning requires $m \gtrsim qT^{2a_1(1-1/a_1^2)}/\varepsilon^{2a_1}$ trajectories to achieve imitation loss bounded by ε, where T is the task horizon, q is the effective number of parameters of the function class for the learned policy, and $a_1 \in [1, \infty)$ is an IGS parameter determined by the expert policy. We show that $a_1 = 1$ for contracting systems, leading to *task-horizon independent* bounds scaling as $m \gtrsim q/\varepsilon^2$. Furthermore, we construct a simple family of systems that gracefully degrade away from these exponential rates such that $a_1 = 1 + p$ for $p \in (0, \infty)$, yielding sample-complexity that scales as $m \gtrsim qT^{2p(2+p)/1+p}\varepsilon^{2(1+p)}$, which makes clear that an increase/decrease in p yields a corresponding increase/decrease in sample-complexity. Motivated by the empirical success and widespread adoption of DAgger and DAgger-like algorithms, we extend our analysis to an IGS-constrained DAgger-like algorithm. We show that this algorithm enjoys comparable stability dependent sample-complexity guarantees, requiring $m \gtrsim qT^{a_1^2(1-1/a_1)(4+1/a_1)}/\varepsilon^{2a_1^2}$ trajectories to learn an ε-approximate policy, again recovering time-independent bounds for contracting systems that gracefully degrade when applied to our family of systems satisfying $a_1 = 1 + p$. In particular, we provide the first results delineating a class of systems for which sample-complexity bounds scale sublinearly in the task-horizon T, and do so without requiring (strong) convexity of the loss function in the policy parameters.

Finally, we demonstrate the validity of our results on (a) our simple family of nonlinear systems for which the underlying IGS properties can be quantitatively tuned, and (b) a high-dimensional nonlinear quadrupedal robotic system. We show empirically that the sample-complexity scaling predicted by the underlying stability properties of the expert policy are indeed observed in practice.

2 Problem Statement

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

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

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

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

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

The imitation loss function $\ell_{\pi'}(\xi; \pi_1, \pi_2)$ keeps a running tally of the difference $\Delta_{\pi_1,\pi_2}(\varphi_t^{\pi'}(\xi))$ of how actions taken by policies π_1 and π_2 enter the system (2.1) when the system is evolving under policy π' starting from initial condition $x_0 = \xi$.

We now formally state the problem considered in this paper. Fix a known system (2.1), and pick a tolerance $\varepsilon > 0$ and failure probability $\delta \in (0, 1)$. Our goal is to design and analyze imitation learning algorithms that produce a policy $\hat{\pi}$ using $m = m(\varepsilon, \delta)$ trajectories of length T seeded from initial conditions $\{\xi_i\}_{i=1}^m \sim \mathcal{D}^m$, such that with probability at least $1 - \delta$, the learned policy $\hat{\pi}$

induces a state/input trajectory distribution that satisfies $\mathbb{E}_{\xi \sim \mathcal{D}} \ell_{\hat{\pi}}(\xi; \hat{\pi}, \pi_\star) \leqslant \varepsilon$. Crucially, we seek to understand how the underlying stability properties of the expert policy $\pi_\star$ manifest themselves in the number of required trajectories $m(\varepsilon, \delta)$.

We note that bounding the imitation loss has immediate implications on the e.g., safety, stability, and performance of the learned policy $\hat{\pi}$. Concretely, let $h : \mathbb{R}^{n \times (T+1)} \to \mathbb{R}^s$ denote an L_h-Lipschitz observable function of a trajectory: examples of valid observable functions include Lyapunov and barrier inequalities and semi-algebraic constraints on the state or state-feedback policy. Then,

$$\mathbb{E}_{\xi \sim \mathcal{D}} \|h_{\pi_\star}(\xi) - h_{\hat{\pi}}(\xi)\|_2 \leqslant L_h \mathbb{E}_{\xi \sim \mathcal{D}} \sum_{t=0}^{T} \|\varphi_t^{\pi_\star}(\xi) - \varphi_t^{\hat{\pi}}(\xi)\|_2, \tag{2.3}$$

where $h_\pi(\xi) := h(\{\varphi_t^\pi(\xi)\}_{t=0}^T)$. We will subsequently see how the discrepancy term $\sum_{t=0}^{T} \|\varphi_t^{\pi_\star}(\xi) - \varphi_t^{\hat{\pi}}(\xi)\|_2$ can be upper bounded by the imitation loss $\ell_{\hat{\pi}}(\xi; \hat{\pi}, \pi_\star)$, and thus bounds on the imitation loss imply bounds on the deviations between the observables $h_{\pi_\star}$ and $h_{\hat{\pi}}$.

3 Incremental Gain Stability

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

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

We already saw this quantity appear naturally in (2.3). Furthermore, we will reduce analyzing the performance of behavior cloning and our DAgger-like algorithm to bounding the discrepancy (3.1) between trajectories induced by the expert policy $\pi_\star$ and a learned policy $\hat{\pi}$.

The simplest way to bound (3.1) is to use a discrete-time version of Grönwall's inequality: if the maps f and g defining system (2.1) as well as policies π_1 and π_2 are B-bounded and L-Lipschitz, then (assuming $L(1 + 2B) > 1$) we can upper bound the discrepancy (3.1) by:

$$\mathsf{disc}_T(\xi; \pi_1, \pi_2) \leqslant \frac{(L(1 + 2B))^T - 1}{L(1 + 2B) - 1} \ell_{\pi_1}(\xi; \pi_1, \pi_2). \tag{3.2}$$

This bound formalizes the intuition that the discrepancy (3.1) scales in proportion to the deviation between policies π_1 and π_2, summed along the trajectory. Unfortunately, this bound is undesirable due to its exponential dependence on the horizon T. In order to improve the dependence on T, we need to assume some stability properties on the dynamics (f, g). We start by drawing inspiration from the definition of *incremental input-to-state stability* [23].[1] This definition relies on standard comparison class definitions, which we briefly review. A class $\mathcal{K}$ function $\sigma : \mathbb{R}_{\geqslant 0} \to \mathbb{R}_{\geqslant 0}$ is continuous, increasing, and satisfies $\sigma(0) = 0$. A class $\mathcal{K}_\infty$ function is class $\mathcal{K}$ and unbounded. Finally, a class $\mathcal{K}\mathcal{L}$ function $\beta : \mathbb{R}_{\geqslant 0} \times \mathbb{R}_{\geqslant 0} \to \mathbb{R}_{\geqslant 0}$ satisfies (i) $\beta(\cdot, t)$ is class $\mathcal{K}$ for every t and (ii) $\beta(s, \cdot)$ is continuous, decreasing, and tends to zero for every s.

---

[1] Definition 3.1 is more general than Definition 6 of [23] in that we only require a bound with respect to an input perturbation of one of the trajectories, not both.

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

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

Definition 3.1 improves the Grönwall-type estimate (3.2) in the following way. Suppose the closed-loop system defined by $\tilde{f}(x, u) = f(x) + g(x)\pi_2(x) + u$ is δISS. Then the algebraic identity

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

allows us to treat $g(x_t)(\pi_1(x_t) - \pi_2(x_t)) = \Delta_{\pi_1, \pi_2}(x_t)$ as an input signal, yielding

$$\mathsf{disc}_T(\xi; \pi_1, \pi_2) \leqslant T\gamma\left(\ell_{\pi_1}(\xi; \pi_1, \pi_2)\right).$$

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

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

$$\sum_{t=0}^{T} \min\{\|\Delta_t\|_X^{a \wedge a_0}, \|\Delta_t\|_X^{a \vee a_1}\} \leqslant \zeta\|\xi_1 - \xi_2\|_X^a + \gamma \sum_{t=0}^{T-1} \max\{\|u_t\|_U^{b_0}, \|u_t\|_U^{b_1}\}. \tag{3.3}$$

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

IGS quantitatively bounds the amplification of an input signal $\{u_t\}_{t \geqslant 0}$ (and differences in initial conditions ξ_1, ξ_2) on the corresponding system trajectory discrepancies $\{\Delta_t\}_{t \geqslant 0}$. Note that a system that is incrementally gain stable is automatically δISS. IGS also captures the phase transition that occurs in non-contracting systems about the unit circle. For example, when $\|\Delta_t\| \leqslant 1$ and $\|u_t\| \leqslant 1$ for all $t \geqslant 0$, inequality (3.3) reduces to $\sum_{t=0}^{T} \|\Delta_t\|_X^{a \vee a_1} \leqslant \zeta\|\xi_1 - \xi_2\|_X^a + \gamma \sum_{t=0}^{T-1} \|u_t\|_U^{b_0}$. Finally, as IGS measures signal-to-signal ($\{u_t\}_{t \geqslant 0} \to \{\Delta_t\}_{t \geqslant 0}$) amplification, it is well suited to analyzing learning algorithms operating on system trajectories.

For a pair of functions $f(x), g(x)$, we say that (f, g) is Ψ-IGS if the system $f(x) + g(x)u$ is Ψ-IGS. We show that (cf. Proposition C.4) if $(f + g\pi_2, \mathsf{Id})$ is Ψ-IGS, then

$$\mathsf{disc}_T(\xi; \pi_1, \pi_2) \leqslant 4(\gamma \vee 1)^{\frac{1}{a \wedge a_0}} T^{1 - \frac{1}{a \vee a_1}} \max\left\{\ell_{\pi_1}(\xi; \pi_1, \pi_2)^{\frac{b_0}{a \vee a_1}}, \ell_{\pi_1}(\xi; \pi_1, \pi_2)^{\frac{b_1}{a \wedge a_0}}\right\}. \tag{3.4}$$

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

---

[2] For $a, b \in \mathbb{R}$, we let $a \vee b := \max\{a, b\}$ and $a \wedge b := \min\{a, b\}$.

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

$$V(f(x, u), f(y, 0)) - V(x, y) \leqslant -\mathfrak{a} \min\{\|x - y\|_X^{a_0}, \|x - y\|_X^{a_1}\} + \mathfrak{b} \max\{\|u\|_U^{b_0}, \|u\|_U^{b_1}\}. \qquad (3.5)$$

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

3.1 Examples of Incremental Gain Stability

Our first example of incremental gain stability is a contracting system [15].

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

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

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

Three concrete examples of autonomously contracting systems include (1) piecewise linear systems $f(x, u) = \left(\sum_{i=1}^K A_i \mathbf{1}\{x \in \mathcal{C}_i\}\right) x + Bu$ where the A_i's are stable, $\{\mathcal{C}_i\}$ partitions $\mathbb{R}^n$, and there exists a common quadratic Lyapunov function $V(x) = x^\mathsf{T} P x$ which yields the metric $M(x) = P$, (2) $f(x, u) = \log(1 + x^2) + u$ with the metric $M(x) = 2[1 + \exp(-|x|)]^{-1}$, and (3) $f(x, u) = x - \eta [\nabla V(x) + u]$ where $V(x)$ is a twice differentiable potential function satisfying $\mu I \preccurlyeq \nabla^2 V(x) \preccurlyeq LI$, and $0 < \eta \leqslant 1/L$ [24]. Importantly, we note that contracting systems enjoy time-independent discrepancy measures such that bound (3.4) reduces to $O(\ell_{\pi_1}(\xi; \pi_1, \pi_2))$, i.e., contracting nonlinear systems behave like stable linear systems, up to contraction metric dependent constants.

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

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

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

4 Algorithms and Theoretical Results

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

Assumption 4.1 (Regularity). *We assume that the dynamics (f, g), policy class Π, expert $\pi_\star$, and initial condition distribution $\mathcal{D}$ satisfy:*

(a) *The dynamics (f, g) satisfy (i) $f(0) = 0$ and (ii) $\sup_{x \in \mathbb{R}^n} \|g(x)\|_{\mathrm{op}} \leqslant B_g$.*
(b) *The policy class Π is convex[3] and $\pi(0) = 0$ for all $\pi \in \Pi$.*
(c) *The distribution $\mathcal{D}$ over initial conditions satisfies $\|\xi\|_2 \leqslant B_0$ a.s. for $\xi \sim \mathcal{D}$.*
(d) *The expert policy $\pi_\star \in \Pi$.*
(e) *Δ_{π_1, π_2} is L_Δ-Lipschitz for all $\pi_1, \pi_2 \in \Pi$.*
(f) *The constants $B_g, B_0, L_\Delta \in [1, \infty)$.*

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

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

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

Algorithm 1 Constrained Mixing Iterative Learning (CMILe)

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

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

---

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

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

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

$$\text{minimize}_{\bar{\pi} \in \Pi} \ \frac{1}{m} \sum_{i=1}^{m} \ell_{\pi_{\text{roll}}}(\xi_i; \bar{\pi}, \pi_\star) \tag{4.1a}$$

$$\text{subject to } \frac{1}{m} \sum_{i=1}^{m} \ell_{\pi_{\text{roll}}}(\xi_i; \bar{\pi}, \pi_{\text{roll}}) \leqslant c, \tag{4.1b}$$

$$\frac{1}{1-w}[(1-\alpha)\pi_{\text{roll}} + \alpha\bar{\pi} - w\pi_\star] \in \mathcal{S}_\Psi. \tag{4.1c}$$

With these definitions and assumptions in place, we introduce IGS-Constrained Mixing Iterative Learning (CMILe) in Algorithm 1 and state our main theoretical results. CMILe draws upon and integrates ideas from Stochastic Mixing Iterative Learning (SMILe) [7], constrained policy optimization [25, 26], and the IGS tools developed in Section 3. As in SMILe and DAgger, CMILe proceeds in epochs, beginning with data generated by the expert policy, and iteratively shifts towards a learned policy via updates of the form $\pi_{k+1} = (1-\alpha)\pi_k + \alpha\hat{\pi}_k$, where π_k is the current data-generating policy, $\hat{\pi}_k$ is the policy learned using the most recently generated data, and $\alpha \in (0, 1]$ is a mixing parameter. However, CMILe contains two key departures from traditional IL algorithms: (i) it constrains the learned policy at each epoch to remain appropriately close to the previous epoch's data-generating policy (constraint (4.1b)), and (ii) all data-generating policies $\{\pi_k\}$ are constrained to induce Ψ-IGS closed-loop systems (constraint (4.1c)). The latter constraint allows us to leverage the Ψ-IGS machinery of §3 to analyze Algorithm 1.

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

$$\Pi = \{\pi(x, \theta) : \theta \in \mathbb{R}^q, \|\theta\|_2 \leqslant B_\theta\}, \tag{4.2}$$

with $B_\theta \geqslant 1$, and π a fixed twice continuously differentiable map. As an example, neural networks with q weights and twice continuously differentiable activation functions are captured by the policy class (4.2). We note that our results do not actually require a parameteric representation: as long as a particular policy class Rademacher complexity (defined in the appendix) can be bounded, then our results apply. In what follows, we define the constant $L_{\partial^2 \pi}$ to be:

$$L_{\partial^2 \pi} := 1 \vee \sup_{\|x\| \leqslant \zeta^{1/a} B_0, \|\theta\| \leqslant B_\theta} \left\| \frac{\partial^2 \pi}{\partial\theta\partial x} \right\|_{\ell^2(\mathbb{R}^q) \to M(\mathbb{R}^{d \times n})}.$$

Here, $M(\mathbb{R}^{d \times n})$ is the Banach space of $d \times n$ real-valued matrices equipped with the operator norm.

IGS-Constrained Behavior Cloning. We first analyze a single epoch version of Algorithm 1, which reduces to Behavior Cloning (BC) subject to interpolating the expert policy on the training data (constraint (4.1b)) and inducing a Ψ-IGS closed-loop system (constraint (4.1c)).

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

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

8

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

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

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

IGS-CMILe. Next we analyze Algorithm 1 as stated, and show that if the mixing parameter α and number of episodes E are chosen appropriately with respect to the Ψ-IGS parameters of the underlying expert system, sample-complexity guarantees similar to those in the IGS-constrained BC setting can be obtained. As described above, the key to ensuring that guarantees can be bootstrapped across epochs is the combination of a trust-region constraint (4.1b) and IGS-stability constraints (4.1c) on the intermediate data-generating policies.

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

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

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

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

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

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

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

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

5 Experiments

In our experiments, we implement neural network training by combining the `haiku` NN library [27] with `optax` [28] in `jax` [29]. We use the CPU backend of `jax`, and run on our internal cluster.

5.1 Practical Algorithm Implementation

In order to implement Algorithm 1, a constrained ERM subproblem (Algorithm 2) over the policy class must be solved. Two elements make this subproblem practically challenging: (i) the trust-region constraint (4.1b), and (ii) the IGS-stability constraint (4.1c).

Trust-regions constraints. We implement the trust-region constraint (4.1b) by initializing the weights $\hat{\theta}_k$ parameterizing the policy $\hat{\pi}_k$ at those of the previous epoch's parameters $\hat{\theta}_{k-1}$, and using a small learning rate during training. Alternative viable approaches include imposing trust-region constraints on the parameters of the form $\|\hat{\theta}_k - \hat{\theta}_{k-1}\|_2 \leqslant \kappa$, or explicitly enforcing the trust-region constrain (4.1b). These latter options would be implemented through either a suitable Lagrangian relaxation to soft-penalties in the objective, or by drawing on recent results in constrained empirical risk minimization [30].

Enforcing IGS. Enforcing the Ψ-IGS constraint (4.1c) via an incremental Lyapunov function (cf. Proposition 3.3) is more challenging, as it must be enforced for all x within a desired region of attraction. If such a Lyapunov function is known for the expert, then it can be used to only enforce stability constraints on trajectory data, an approximation/heuristic that is common in the constrained policy optimization literature (see for example [25, 26]). However, if a Lyapunov certificate of Ψ-IGS stability for the expert is not known, then options include (a) learning such a certificate for the expert from data, see for example [20, 31], or (b) jointly optimizing over a Ψ-IGS certificate and learned policy. Although this may be computationally challenging, alternating optimization schemes have been proposed and successfully applied in other contexts, see for example [16, 32–34].

Fortunately, we note that empirically, explicit stability constraints seem not to be required. In the next subsection, we study the quantitative effects of enforcing the stability constraint (4.1c) for a linear system, for which a stability certificate is available, and for which the level of stability of the expert system can be quantitatively tuned. We observe that only when (a) the expert is nearly unstable, and (b) we are in a low-data regime, that a small difference in performance between stability-constrained and unconstrained algorithms occurs, suggesting that optimal policies are naturally stabilizing. Therefore, we simply omit constraint (4.1c) from our implementation and take care to ensure that sufficient data is provided to the IL algorithms to yield stabilizing policies. We also note that once stability constraints are dropped, system dynamics are no longer needed, as similar guarantees can be shown to hold for imitation losses defined with respect to the model-free quantity $\tilde{\Delta}_{\pi_1,\pi_2}(x) := \pi_1(x) - \pi_2(x)$.

5.2 Stability Experiments

We study the effects of explicitly constraining the played policies $\{\pi_i\}_{i=1}^{E}$ to be IGS through the use of Lyapunov certificates. The use of Lyapunov constraints to enforce incremental stability was studied in Section G.1 of Boffi et al. [19], with the main takeaway being that systems satisfying suitable exponential Lyapunov conditions, in particular those certifying exponential input-to-state-stability, are also exponentially IGS (i.e., satisfy $a = a_0 = a_1 = b_0 = b_1 = 1$) on a compact set of initial conditions and bounded inputs.

In order to easily tune the underlying stability properties of the resulting expert system, we consider linear quadratic (LQ) control of a linear system $x_{t+1} = Ax_t + Bu_t$, where $A \in \mathbb{R}^{n \times n}$ and
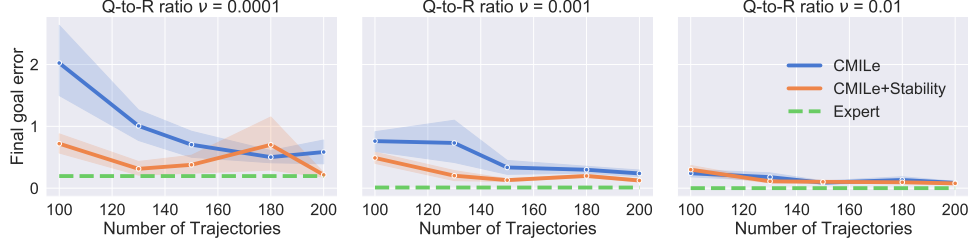
Figure 1: For fixed system matrices (A, B) and cost matrices $R = I_q$ and $Q = \nu I_q$ for $\nu \in \{0.0001, 0.001, 0.01\}$, we show the median goal error obtained by CMILe with and without stability constraints over 100 i.i.d. test trajectories.

$B \in \mathbb{R}^{n \times d}$.

The LQ control problem can be expressed as

$$
\begin{aligned}
\text{minimize}_{\{x_t\}, \{u_t\}} \quad & \sum_{t=0}^{\infty} x_t^{\mathsf{T}} Q x_t + u_t^{\mathsf{T}} R u_t \\
\text{subject to} \quad & x_{t+1} = A x_t + B u_t, \ x_0 = \xi.
\end{aligned}
\tag{5.1}
$$

where $Q \in \mathbb{R}^{n \times n}$ and $R \in \mathbb{R}^{d \times d}$ are fixed cost matrices. The optimal policy is linear in the state x_t, i.e., $u_t = K_{\mathrm{lqr}} x_t$, where $K_{\mathrm{lqr}} \in \mathbb{R}^{d \times n}$ can be computed by solving a discrete-time algebraic Riccati equation [35]. 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). Specifically, our realization satisfied $\|A\|_{\mathrm{op}} = 5.893$, $\|B\|_{\mathrm{op}} = 4.964$, and the open loop system was unstable with spectral radius $\rho(A) = 3.638$. 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. 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 with a learning rate of 0.01.

In Figure 1, we plot the median goal error $\|x_T\|_2$ achieved by policies learned via CMILe for different values of ν, both with and without Lyapunov stability constraints, over 100 test rollouts. The error bars represent the 20th/80th percentiles of the median across ten independent trials. We construct a valid robust Lyapunov certificate from the solution to the DARE (see Appendix A for details), and use it to enforce the IGS constraint (4.1c) on the resulting closed loop dynamics $f_{cl}^{\pi_k}(x) = Ax + B\pi_k(x)$. Two important trends can be observed in Figure 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 (closed-loop spectral radius $\rho(A + BK_{\mathrm{lqr}}) \approx 1$), stability constraints have a measurable effect. Second, the effect of stability constraints is more dramatic in low-data regimes, and by restricting $\pi_k \in \Pi_\Psi$, we reduce over-fitting. We also evaluated the performance of standard

11

behavior cloning (BC), but even with 200 training trajectories and $\nu = 0.01$, the median final goal error was 504.

5.3 Tuneable Ψ-IGS System

We consider the dynamical system in $\mathbb{R}^{10}$:

$$x_{t+1} = x_t - 0.5x_t \frac{|x_t|^p}{1 + 0.5|x_t|^p} + \frac{1}{1 + |x_t|^p}(h(x_t) + u_t). \tag{5.2}$$

All arithmetic operations in (5.2) are element-wise. We set $h : \mathbb{R}^{10} \to \mathbb{R}^{10}$ to be a randomly initialized two layer MLP with zero biases, hidden width 32, and tanh activations. The expert policy is set to be $\pi_\star = -h$, so that the expert's closed-loop dynamics are given by $x_{t+1} = x_t - 0.5x_t \frac{|x_t|^p}{1 + 0.5|x_t|^p}$. From Proposition 3.5, the incremental stability of the closed-loop system degrades as p increases.

p	BC	CMILe	CMILe+Agg	DAgger
1	0.615 ± 0.154	0.247 ± 0.071	0.239 ± 0.038	0.474 ± 0.131
2	1.194 ± 0.130	0.737 ± 0.059	0.578 ± 0.095	0.865 ± 0.142
3	1.637 ± 0.220	1.115 ± 0.066	0.868 ± 0.065	1.199 ± 0.130
4	1.976 ± 0.106	1.409 ± 0.080	1.111 ± 0.080	1.441 ± 0.126
5	2.107 ± 0.079	1.570 ± 0.055	1.240 ± 0.091	1.594 ± 0.146

Table 1: Final $\|x_T^{\text{expert}} - x_T^{\text{IL}}\|_2$ of imitation learning algorithms on (5.2).

p	BC	CMILe	CMILe+Agg	DAgger
1	0.161 ± 0.088	0.113 ± 0.019	0.021 ± 0.005	0.050 ± 0.017
2	0.324 ± 0.097	0.128 ± 0.013	0.021 ± 0.003	0.052 ± 0.021
3	0.498 ± 0.120	0.151 ± 0.023	0.028 ± 0.007	0.064 ± 0.018
4	0.672 ± 0.139	0.163 ± 0.019	0.031 ± 0.007	0.061 ± 0.023
5	0.905 ± 0.154	0.154 ± 0.030	0.035 ± 0.007	0.058 ± 0.015

Table 2: Final average closed-loop imitation loss error of imitation learning algorithms on (5.2).

In this experiment, we vary $p \in \{1, ..., 5\}$ to see the effect of p on the final task goal error and imitation loss. We compare four different algorithms. **BC** is standard behavior cloning. **CMILe** is Algorithm 1 with the practical modifications as described above. **CMILe+Agg** is **CMILe**, except that at epoch k, the data from previous epochs $j \in \{0, ..., k-1\}$ is also used in training. **DAgger** is the imitation learning algorithm from Ross et al. [4]. For all algorithms, we fix the number of trajectories m from (5.2) to be $m = 250$. The horizon length is $T = 100$. The distribution $\mathcal{D}$ over initial condition is set as $N(0, I)$. We set the policy class Π to be two layer MLPs with hidden width 64 and tanh activations. Each algorithm minimizes the imitation loss using 300 epochs of Adam with learning rate 0.01 and batch size 512. For all algorithms except **BC**, we use $E = 25$ epochs with $\alpha = 0.15$ (in DAgger's notation, we set $\beta_k = 0.85^k$), resulting in 10 trajectories per epoch.

In Table 1, we track the difference in norm $\|x_T^{\text{expert}} - x_T^{\text{IL}}\|_2$ between the expert's final state (x_T^{expert}) and the IL algorithm's final state (x_T^{IL}), both seeded seeded from the same initial conditions. In Table 2, we track the final average closed-loop imitation loss error $\frac{1}{T}\mathbb{E}_{\xi \sim \mathcal{D}}\ell_{\pi_E}(\xi; \pi_E, \pi_\star)$. The entries in the tables are computed by rolling out 500 test trajectories and computing the median quantity over the test trajectories. Each algorithm is repeated for 50 trials, and the median quantity $\pm \max(\text{80th percentile} - \text{median}, \text{median} - \text{20th percentile})$ (over the 50 trials) is shown. In Table 1, we see that as p decreases, the goal deviation error decreases, showing that the task becomes fundamentally easier. This trend is also reflected in all the imitation learning algorithms. Table 2 provides insight into why the goal deviation error decreases with p, and also shows that our main theorems are indeed predictive: as p decreases, the closed-loop average imitation loss $\frac{1}{T}\mathbb{E}_{\xi \sim \mathcal{D}}\ell_{\pi_E}(\xi; \pi_E, \pi_\star)$ generally decreases for all algorithms.

5.4 Unitree Laikago

We now study IL on the Unitree Laikago robot, an 18-dof quadruped with 3-dof of actuation per leg. We use PyBullet [36] for our simulations. The goal of this experiment is to demonstrate, much like for the previous tuneable family of Ψ-IGS systems, that increasing the stability of the underlying closed-loop expert decreases the sample-complexity of imitation learning. We do this qualitatively by studying a sideways walking task where the robot tracks a constant sideways linear velocity. By increasing the desired linear velocity, the resulting expert closed-loop becomes more unstable.

Our expert controller is a model-based predictive controller using a simplified center-of-mass dynamics as described in Di Carlo et al. [37]. The stance and swing legs are controlled separately. The swing leg controller is based on a proportional-derivative (PD) controller. The stance leg controller solves for the desired contact forces to be applied at the foot using a finite-horizon constrained linear-quadratic optimal control problem; the linear model is computed from linearizing the center-of-mass dynamics. The desired contact forces are then converted to hip motor torques using the body Jacobian. More details about the expert controller can be found in the appendix.

We restrict our imitation learning to the stance leg controller, as it is significantly more complex than the swing leg controller. Furthermore, instead of randomizing over initial conditions, we inject randomization into the environment by subjecting the Laikago to a sequence of random push forces throughout the entire trajectory. As in the previous experiment, we compare **BC**, **CMILe**, **CMILe+Agg** (omitting **DAgger** for space reasons as its performance is comparable to **CMILe+Agg**). We set the horizon length to $T = 1000$. We featurized the robot state into a 14-dimensional feature vector; the exact features are given in the appendix. The output of the policy is a 12-dimensional vector (x, y, z contact forces for each of the 4 legs). We used a policy class of two layer MLPs of hidden width 64 with ReLU activations. For training, we ran 500 epochs of Adam with a batch size of 512 and step size of 0.001. Furthermore, we tried to overcome the effect of overfitting in **BC** by using the following heuristic: we used 5% of the training data as a holdout set, and we stopped training when either the holdout risk increased $h = 50$ times or 500 epochs were completed, whichever came first. To assess the effect of the number of samples on imitation learning, we vary the number of rollouts per epoch $S \in \{1, ..., 5\}$. For **CMILe** and **CMILe+Agg**, we fix $\alpha = 0.3$ and $E = 12$. We provide **BC** with $S \times E$ total trajectories.

Figure 2 shows the result of our experiments. In the top row, we plot the *deviation* $|y_T^{\text{exp}} - y_T^{\text{IL}}|$ between the expert's final y position (y_T^{exp}) and the imitation learning algorithm's final y position (y_T^{IL}). We observe that as the target linear speed decreases, the deviation between the expert and IL
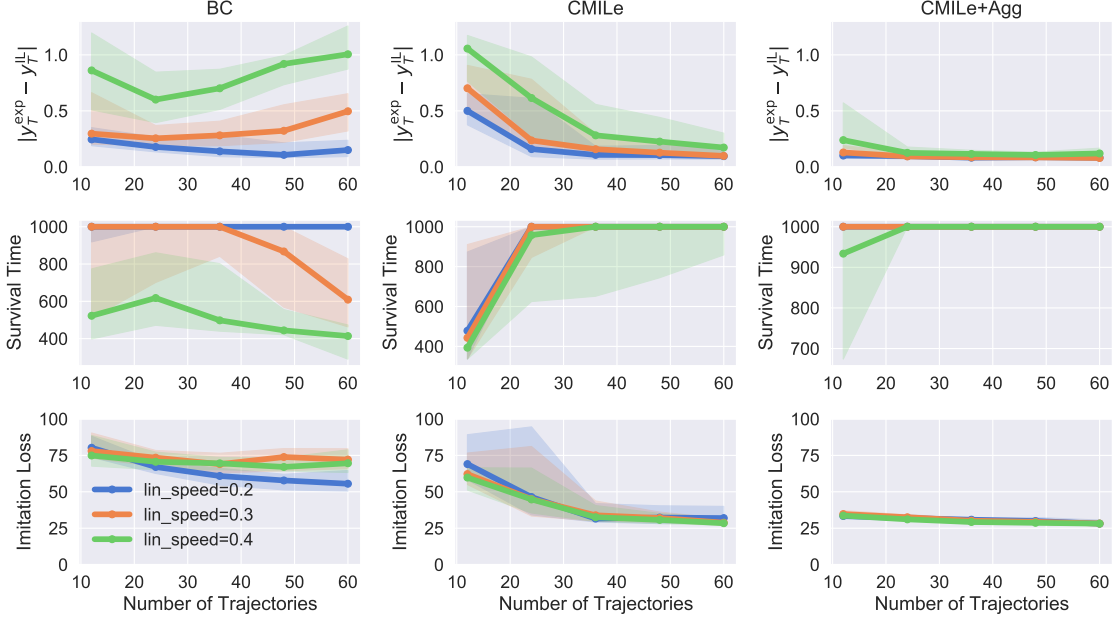
Figure 2: Imitation learning on a sideways walking task. The top row shows the deviation error $|y_T^{\exp} - y_T^{\text{IL}}|$ of the various algorithms, the middle row shows the survival times, and the bottom row plots the average closed-loop imitation loss $\frac{1}{T}\mathbb{E}_{\xi \sim \mathcal{D}}\ell_{\pi_E}(\xi; \pi_E, \pi_\star)$.

algorithms also decreases; this qualitative trend is consistent with Theorem 4.3 and Theorem 4.4. Note that the y positions are computed by subjecting the Laikago to the same sequence of random force pushes for both the expert and IL algorithm. For the IL algorithm, if the rollout terminates early, then the last y position (before termination) is used in place of y_T^{IL}. In the middle row, we plot the survival times for each of the algorithms, which is the number of simulation steps (out of 1000) that the robot successfully executes before a termination criterion triggers which indicates the robot is about to fall. We see that for all algorithms, by decreasing the sideways linear velocity, the resulting learned policy is able to avoid falling more. Note that for **CMILe+Agg**, the learned policy does not fall for linear speeds $0.2, 0.3$. In the bottom row, we plot the average closed-loop imitation loss $\frac{1}{T}\mathbb{E}_{\xi \sim \mathcal{D}}\ell_{\pi_E}(\xi; \pi_E, \pi_\star)$. We see that for **BC**, the imitation loss shows improvement with increased samples for linear speed of 0.2, but no improvements occur for the more difficult linear speeds of $0.3, 0.4$. This trend is less apparent for **CMILe** and **CMILe+Agg**, but is most prominently seen in the deviation error $|y_T^{\exp} - y_T^{\text{IL}}|$.

6 Conclusions & Future Work

We showed that IGS-constrained IL algorithms allow for a granular connection between the stability properties of an underlying expert system and the resulting sample-complexity of an IL task. Our future work will focus on two complementary directions. First, CMILe and DAgger significantly outperform BC in our experiments, but our bounds are not yet able to capture this: future work will look to close this gap. Second, although our focus in this paper has been on imitation learning, we have developed a general framework for reasoning about learning over trajectories in continuous

state and action spaces. We will look to apply our framework in other settings, such as safe exploration and model-based reinforcement learning.

Acknowledgements

References

[1] 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.

[2] 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.

[3] 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.

[4] 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.

[5] 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.

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

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

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

[9] 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.

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

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

[12] 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.

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

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

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

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

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

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

[19] Nicholas M. Boffi, Stephen Tu, and Jean-Jacques E. Slotine. Regret bounds for adaptive nonlinear control. In *3rd Annual Learning for Dynamics & Control Conference*, 2021.

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

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

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

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

[24] Patrick M. Wensing and Jean-Jacques E. Slotine. Beyond convexitycontraction and global convergence of gradient descent. *PLOS ONE*, 15(8):1–29, 2020.

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

[26] 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.

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

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

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

[30] Luiz Chamon and Alejandro Ribeiro. Probably approximately correct constrained learning. In *Neural Information Processing Systems*, 2020.

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

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

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

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

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

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

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

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

[39] Quang-Cuong Pham. Analysis of discrete and hybrid stochastic systems by nonlinear contraction theory. In *2008 10th International Conference on Control, Automation, Robotics and Vision*, 2008.

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

Contents

A   Stability Study: Constructing a Robust Lyapunov function

In order to compute a robust Lyapunov function for use in the stability experiments of Section 5.2, we use the following approach. Let $A_{\mathrm{lqr}} := A + BK_{\mathrm{lqr}}$ and consider the following Lyapunov equation:

$$A_{\mathrm{lqr}}^{\mathsf{T}} X A_{\mathrm{lqr}} - \gamma^2 X + \varepsilon I = 0, \tag{A.1}$$

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 $\gamma \geqslant \rho(A_{\mathrm{lqr}})$ [35].

Then note that we can rewrite the Lyapunov equation (A.1) as

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

for $Q = (1 - \gamma^2)X + \varepsilon I$. 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 [38].

B   Laikago Experimental Details

B.1   More Details on Expert Controller

The expert controller contains multiple components: the swing controller, the stance controller, and the gait generator. The gait generator uses the clock source to generate a desired gait pattern, where a pair of diagonal legs are synchronized and are out of phase with the other pair. Throughout our experiments, we fixed the gait to be a trotting gait. The swing controller generates the aerial trajectories of the feet when they lift up and controls the landing positions based on the desired moving speed. The stance leg controller is based on model predictive control (MPC) using centroidal dynamics [37]. Recall that in our experiments, we only perform imitation learning for the stance leg controller.

We now describe the centroidal dynamics model. We treat the whole robot as a single rigid body, and assume that the inertia contribution from the leg movements is negligible (cf. Figure 3). With these assumptions, the system dynamics can be simply written using the Newton-Euler equations:

$$m\ddot{\mathbf{x}} = \sum_{i=1}^{4} \mathbf{F}_i - \mathbf{g},$$

$$\frac{d}{dt}(\mathbf{I}\omega) = \sum_{i=1}^{4} \mathbf{r}_i \times \mathbf{F}_i,$$
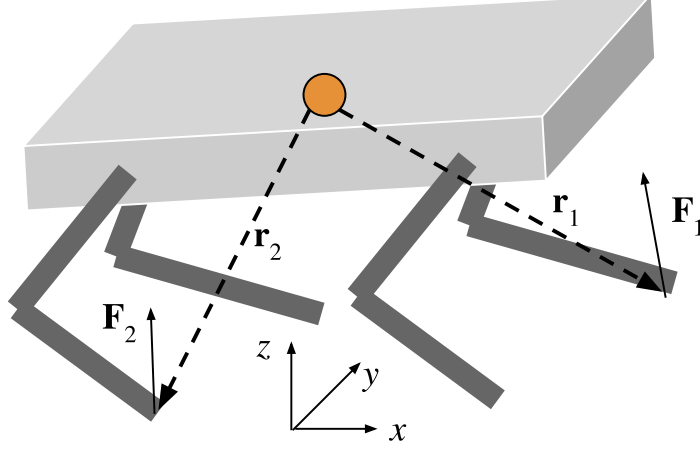
Figure 3: The centroidal dynamics model used to formulate the model predictive control problem.

where $\mathbf{x} = (x, y, z, \Phi, \Theta, \Psi)$ denotes the center of mass (CoM) translation and rotation, $\mathbf{F}_i = (f_x, f_y, f_z)_i$ is the contact force applied on the i-th foot (set to zero if the i-th foot is not in contact with the ground), and $\mathbf{r}_i$ is the displacement from the CoM to the contact point. We used the same $Z - Y - X$ Euler angle conventions in [37] to represent the CoM rotation. Since the robot operates in a regime where its base is close to flat, singularity from the Euler angle representation is not an concern.

The MPC module solves an optimization problem over a finite horizon H to track a desired pose and velocity $\mathbf{q} = (\mathbf{x}, \dot{\mathbf{x}})$. The system dynamics can be linearized around the desired state and discretized:

$$\mathbf{q}_{t+1} = A(\mathbf{q}^d)\mathbf{q}_t + B(\mathbf{q}^d)\mathbf{u}_t,$$

where $\mathbf{u}_t = (\mathbf{F}_{1,t}, \mathbf{F}_{2,t}, \mathbf{F}_{3,t}, \mathbf{F}_{4,t})$, is the concatenated force vectors from all feet. We then formally write the optimization target:

$$\min_{\mathbf{u}_t} \quad \sum_{t=1}^{H} [(\mathbf{q}_t - \mathbf{q}_t^d)^T \mathbf{Q}(\mathbf{q}_t - \mathbf{q}_t^d) + \mathbf{u}_t^T \mathbf{R} \mathbf{u}_t],$$

$$\text{s.t.} \quad \mathbf{q}_{t+1} = A\mathbf{q}_t + B\mathbf{u}_t,$$

$$\mathbf{F}_{i,t} = 0 \text{ if } i\text{-th foot is not in contact,}$$

$$0 \leqslant f_{z,t} \leqslant f_{max}, \quad \text{contact normal force for each foot,}$$

$$-\mu f_{z,t} \leqslant f_{x,t} \leqslant \mu f_{z,t},$$

$$-\mu f_{z,t} \leqslant f_{y,t} \leqslant \mu f_{z,t}, \quad \text{friction cone,}$$

where we used a diagonal $\mathbf{Q}$ and $\mathbf{R}$ matrix with the weights detailed in [37]. At runtime, we apply the feet contact forces from the first step by converting them to motor torques using the Jacobian matrix.

B.2 Featurization

The inputs to the MPC algorithm is a 28 dimension vector $(\mathbf{q}, \mathbf{q}^d, \mathbf{c}, \mathbf{r})$, where $\mathbf{c}$ is a binary vector indicating feet contact states, and $\mathbf{r} = (\mathbf{r}_1, \mathbf{r}_2, \mathbf{r}_3, \mathbf{r}_4)$ represent the relative displacement from the CoM to each feet. This representation contains redundant information, since one can infer the body height z from the contact state, and local feet displacements when the quadruped is walking on flat ground. Also, since in our experiments the desired pose and speed of the robot are fixed (moving along y direction at constant speed without body rotation), they are not passed as inputs to the imitation policy. Furthermore, the current body linear velocities of the robot are omitted from the inputs as well, since they are not directly measurable on a legged robots without motion capture systems or state estimators. As a result, the inputs to the imitation policy are condensed to a 14 dimensional vector $(\Phi, \Theta, \mathbf{r} \cdot \mathbf{c})$, i.e., the roll, pitch angle of the CoM, and the contact state masked feet positions.

C Incremental Gain Stability Proofs

C.1 Preliminaries

We first prove a simple proposition which we use repeatedly.

Proposition C.1. *Let $a_1, ..., a_k \in [1, \infty)$. Then for any $x \in \mathbb{R}$, we have:*

$$\min\{|x|^{a_1}, ..., |x|^{a_k}\} = \min\{|x|^{\min\{a_1,...,a_k\}}, |x|^{\max\{a_1,...,a_k\}}\},$$
$$\max\{|x|^{a_1}, ..., |x|^{a_k}\} = \max\{|x|^{\min\{a_1,...,a_k\}}, |x|^{\max\{a_1,...,a_k\}}\}.$$

Proof. We only prove the result for min since the proof for max is nearly identical. First, suppose that $|x| \leqslant 1$, then since $a_i \geqslant 1$,

$$\min\{|x|^{a_1}, ..., |x|^{a_k}\} = |x|^{\max\{a_1,...,a_k\}}.$$

Now, suppose that $|x| > 1$. Again, since $a_i \geqslant 1$,

$$\min\{|x|^{a_1}, ..., |x|^{a_k}\} = |x|^{\min\{a_1,...,a_k\}}.$$

The claim now follows. $\qquad\square$

We now derive some basic consequences of the definition of incremental gain stability. The following helper proposition will be useful for what follows.

Proposition C.2. *Fix any initial conditions $\xi_1, \xi_2 \in \mathbb{R}^n$, and any signal $\{u_t\}_{t \geqslant 0}$. Let*

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

For any integers $0 \leqslant T_1 \leqslant T_2$, we have:

$$\sum_{t=T_1}^{T_2} \|\Delta_t\|_X \leqslant \left(\sum_{t=T_1}^{T_2} \min\{\|\Delta_t\|_X^{a \wedge a_0}, \|\Delta_t\|_X^{a \vee a_1}\} \right)^{\frac{1}{a \wedge a_0}} (T_2 - T_1 + 1)^{1 - \frac{1}{a \wedge a_0}}$$

$$+ \left(\sum_{t=T_1}^{T_2} \min\{\|\Delta_t\|_X^{a \wedge a_0}, \|\Delta_t\|_X^{a \vee a_1}\} \right)^{\frac{1}{a \vee a_1}} (T_2 - T_1 + 1)^{1 - \frac{1}{a \vee a_1}}.$$

Proof. Let the index set $I \subseteq \{T_1, ..., T_2\}$ be defined as:

$$I := \{t \in \{T_1, ..., T_2\} : \|\Delta_t\|_X^{a \wedge a_0} \leqslant \|\Delta_t\|_X^{a \vee a_1}\}.$$

By Hölder's inequality,

$$
\begin{aligned}
\sum_{t=T_1}^{T_2} \|\Delta_t\|_X &= \sum_{t \in I} \|\Delta_t\|_X + \sum_{t \in I^c} \|\Delta_t\|_X \\
&\leqslant \left(\sum_{t \in I} \|\Delta_t\|_X^{a \wedge a_0} \right)^{\frac{1}{a \wedge a_0}} |I|^{1 - \frac{1}{a \wedge a_0}} + \left(\sum_{t \in I^c} \|\Delta_t\|_X^{a \vee a_1} \right)^{\frac{1}{a \vee a_1}} |I^c|^{1 - \frac{1}{a \vee a_1}} \\
&= \left(\sum_{t \in I} \min\{\|\Delta_t\|_X^{a \wedge a_0}, \|\Delta_t\|_X^{a \vee a_1}\} \right)^{\frac{1}{a \wedge a_0}} |I|^{1 - \frac{1}{a \wedge a_0}} + \left(\sum_{t \in I^c} \min\{\|\Delta_t\|_X^{a \wedge a_0}, \|\Delta_t\|_X^{a \vee a_1}\} \right)^{\frac{1}{a \vee a_1}} |I^c|^{1 - \frac{1}{a \vee a_1}} \\
&\leqslant \left(\sum_{t=T_1}^{T_2} \min\{\|\Delta_t\|_X^{a \wedge a_0}, \|\Delta_t\|_X^{a \vee a_1}\} \right)^{\frac{1}{a \wedge a_0}} (T_2 - T_1 + 1)^{1 - \frac{1}{a \wedge a_0}} \\
&\quad + \left(\sum_{t=T_1}^{T_2} \min\{\|\Delta_t\|_X^{a \wedge a_0}, \|\Delta_t\|_X^{a \vee a_1}\} \right)^{\frac{1}{a \vee a_1}} (T_2 - T_1 + 1)^{1 - \frac{1}{a \vee a_1}}.
\end{aligned}
$$

$\square$

Next, we compare the autonomous trajectories between two different initial conditions (both trajectories are not driven by any input).

Proposition C.3. *Suppose that f is Ψ-incrementally-gain-stable. Fix a pair of initial conditions $\xi_1, \xi_2 \in X$ and define for $t \in \mathbb{N}$:*

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

We have for all $t \in \mathbb{N}$:

$$\|\Delta_t\|_X \leqslant (\zeta \vee 1)^{\frac{1}{a \wedge a_0}} \max\left\{ \|\xi_1 - \xi_2\|_X^{\frac{a}{a \wedge a_0}}, \|\xi_1 - \xi_2\|_X^{\frac{a}{a \vee a_1}} \right\}. \tag{C.1}$$

Furthermore, for any horizon $T \in \mathbb{N}$,

$$\sum_{t=0}^{T-1} \|\Delta_t\|_X \leqslant 2(\zeta \vee 1)^{\frac{1}{a \wedge a_0}} T^{1 - \frac{1}{a \vee a_1}} \max\left\{ \|\xi_1 - \xi_2\|_X^{\frac{a}{a \wedge a_0}}, \|\xi_1 - \xi_2\|_X^{\frac{a}{a \vee a_1}} \right\}. \tag{C.2}$$

Proof. First, we observe that because $a_0 \leqslant a_1$:

$$a \wedge a_0 \leqslant a \vee a_1 \iff \frac{1}{a \vee a_1} \leqslant \frac{1}{a \wedge a_0} \iff 1 - \frac{1}{a \wedge a_0} \leqslant 1 - \frac{1}{a \vee a_1}.$$

Therefore,

$$T^{1 - \frac{1}{a \wedge a_0}} \leqslant T^{1 - \frac{1}{a \vee a_1}}, \quad (\zeta \vee 1)^{\frac{1}{a \vee a_1}} \leqslant (\zeta \vee 1)^{\frac{1}{a \wedge a_0}}.$$

22

We will utilize these inequalities in what follows.

Now we turn to (C.1). Fix a $t \geqslant 0$. First, suppose that $\|\Delta_t\|_X^{a \wedge a_0} \leqslant \|\Delta_t\|_X^{a \vee a_1}$. Then by Ψ-incremental-gain-stability,

$$\|\Delta_t\|_X^{a \wedge a_0} = \min\{\|\Delta_t\|_X^{a \wedge a_0}, \|\Delta_t\|_X^{a \vee a_1}\} \leqslant \sum_{k=0}^{t} \min\{\|\Delta_k\|_X^{a \wedge a_0}, \|\Delta_k\|_X^{a \vee a_1}\} \leqslant \zeta \|\xi_1 - \xi_2\|_X^a.$$

This implies that:

$$\|\Delta_t\|_X \leqslant [\zeta \|\xi_1 - \xi_2\|_X^a]^{\frac{1}{a \wedge a_0}}.$$

Now, suppose that $\|\Delta_t\|_X^{a \wedge a_0} > \|\Delta_t\|_X^{a \vee a_1}$. A similar argument as before yields:

$$\|\Delta_t\|_X \leqslant [\zeta \|\xi_1 - \xi_2\|_X^a]^{\frac{1}{a \vee a_1}}.$$

Combining these inequalities yields the desired inequality (C.1):

$$\|\Delta_t\|_X \leqslant \max\left\{ [\zeta \|\xi_1 - \xi_2\|_X^a]^{\frac{1}{a \wedge a_0}}, [\zeta \|\xi_1 - \xi_2\|_X^a]^{\frac{1}{a \vee a_1}} \right\}$$

$$\leqslant (\zeta \vee 1)^{\frac{1}{a \wedge a_0}} \max\left\{ \|\xi_1 - \xi_2\|_X^{\frac{a}{a \wedge a_0}}, \|\xi_1 - \xi_2\|_X^{\frac{a}{a \vee a_1}} \right\}.$$

Now we turn to (C.2). By Proposition C.2 and Ψ-incremental-gain-stability, we have:

$$\sum_{t=0}^{T-1} \|\Delta_t\|_X \leqslant [\zeta \|\xi_1 - \xi_2\|_X^a]^{\frac{1}{a \wedge a_0}} T^{1 - \frac{1}{a \wedge a_0}} + [\zeta \|\xi_1 - \xi_2\|_X^a]^{\frac{1}{a \vee a_1}} T^{1 - \frac{1}{a \vee a_1}}.$$

Therefore:

$$\sum_{t=0}^{T-1} \|\Delta_t\|_X \leqslant 2(\zeta \vee 1)^{\frac{1}{a \wedge a_0}} T^{1 - \frac{1}{a \vee a_1}} \max\left\{ \|\xi_1 - \xi_2\|_X^{\frac{a}{a \wedge a_0}}, \|\xi_1 - \xi_2\|_X^{\frac{a}{a \vee a_1}} \right\}.$$

$\square$

The next result compares two trajectories starting from the same initial condition, but one being driven by an input sequence $\{u_t\}$ whereas the other is autonomous.

Proposition C.4. *Suppose that f is Ψ-incrementally-gain-stable. Then, for all $T \in \mathbb{N}$, all initial conditions $\xi \in X$ and all input signals $\{u_t\}_{t \geqslant 0} \subseteq U$, letting*

$$\Delta_t := \varphi_t(\xi, \{u_t\}_{t \geqslant 0}) - \varphi_t(\xi, \{0\}_{t \geqslant 0}),$$

we have:

$$\sum_{t=0}^{T} \|\Delta_t\|_X \leqslant 4 (\gamma \vee 1)^{\frac{1}{a \wedge a_0}} T^{1 - \frac{1}{a \vee a_1}} \max\left\{ \left(\sum_{t=0}^{T-1} \|u_t\|_U \right)^{\frac{b_0}{a \vee a_1}}, \left(\sum_{t=0}^{T-1} \|u_t\|_U \right)^{\frac{b_1}{a \wedge a_0}} \right\}.$$

Hence if $b_1 \leqslant a \wedge a_0$, then we furthermore have:

$$\mathbb{E} \sum_{t=0}^{T} \|\Delta_t\|_X \leqslant 4 (\gamma \vee 1)^{\frac{1}{a \wedge a_0}} T^{1 - \frac{1}{a \vee a_1}} \max\left\{ \left(\mathbb{E} \sum_{t=0}^{T-1} \|u_t\|_U \right)^{\frac{b_0}{a \vee a_1}}, \left(\mathbb{E} \sum_{t=0}^{T-1} \|u_t\|_U \right)^{\frac{b_1}{a \wedge a_0}} \right\}.$$

23

Proof. By Proposition C.2, the fact that $\Delta_0 = 0$, and Ψ-incremental-gain-stability,

$$\sum_{t=0}^{T}\|\Delta_t\|_X \leqslant \left(\gamma\sum_{t=0}^{T-1}\max\{\|u_t\|_U^{b_0},\|u_t\|_U^{b_1}\}\right)^{\frac{1}{a\wedge a_0}}T^{1-\frac{1}{a\wedge a_0}} + \left(\gamma\sum_{t=0}^{T-1}\max\{\|u_t\|_U^{b_0},\|u_t\|_U^{b_1}\}\right)^{\frac{1}{a\vee a_1}}T^{1-\frac{1}{a\vee a_1}}.$$

Now we observe that because $a_0 \leqslant a_1$, we have that

$$a\wedge a_0 \leqslant a\vee a_1 \iff \frac{1}{a\vee a_1} \leqslant \frac{1}{a\wedge a_0} \iff 1-\frac{1}{a\wedge a_0} \leqslant 1-\frac{1}{a\vee a_1}.$$

Therefore,

$$T^{1-\frac{1}{a\wedge a_0}} \leqslant T^{1-\frac{1}{a\vee a_1}}, \quad (\gamma\vee 1)^{\frac{1}{a\vee a_1}} \leqslant (\gamma\vee 1)^{\frac{1}{a\wedge a_0}}.$$

From this, we conclude,

$$\sum_{t=0}^{T}\|\Delta_t\|_X \leqslant (\gamma\vee 1)^{\frac{1}{a\wedge a_0}}T^{1-\frac{1}{a\vee a_1}}\left[\left(\sum_{t=0}^{T-1}\max\{\|u_t\|_U^{b_0},\|u_t\|_U^{b_1}\}\right)^{\frac{1}{a\wedge a_0}} + \left(\sum_{t=0}^{T-1}\max\{\|u_t\|_U^{b_0},\|u_t\|_U^{b_1}\}\right)^{\frac{1}{a\vee a_1}}\right].$$

Next, we observe that:

$$\sum_{t=0}^{T-1}\max\{\|u_t\|_U^{b_0},\|u_t\|_U^{b_1}\} \leqslant \sum_{t=0}^{T-1}\|u_t\|_U^{b_0} + \sum_{t=0}^{T-1}\|u_t\|_U^{b_1} \leqslant 2\max\left\{\sum_{t=0}^{T-1}\|u_t\|_U^{b_0},\sum_{t=0}^{T-1}\|u_t\|_U^{b_1}\right\}.$$

Therefore,

$$\left(\sum_{t=0}^{T-1}\max\{\|u_t\|_U^{b_0},\|u_t\|_U^{b_1}\}\right)^{\frac{1}{a\wedge a_0}} + \left(\sum_{t=0}^{T-1}\max\{\|u_t\|_U^{b_0},\|u_t\|_U^{b_1}\}\right)^{\frac{1}{a\vee a_1}}$$

$$\leqslant 2^{\frac{1}{a\wedge a_0}}\max\left\{\left(\sum_{t=0}^{T-1}\|u_t\|_U^{b_0}\right)^{\frac{1}{a\wedge a_0}},\left(\sum_{t=0}^{T-1}\|u_t\|_U^{b_1}\right)^{\frac{1}{a\wedge a_0}}\right\}$$

$$+ 2^{\frac{1}{a\vee a_1}}\max\left\{\left(\sum_{t=0}^{T-1}\|u_t\|_U^{b_0}\right)^{\frac{1}{a\vee a_1}},\left(\sum_{t=0}^{T-1}\|u_t\|_U^{b_1}\right)^{\frac{1}{a\vee a_1}}\right\}$$

$$\leqslant 4\max\left\{\left(\sum_{t=0}^{T-1}\|u_t\|_U\right)^{\frac{b_0}{a\wedge a_0}},\left(\sum_{t=0}^{T-1}\|u_t\|_U\right)^{\frac{b_1}{a\wedge a_0}},\left(\sum_{t=0}^{T-1}\|u_t\|_U\right)^{\frac{b_0}{a\vee a_1}},\left(\sum_{t=0}^{T-1}\|u_t\|_U\right)^{\frac{b_1}{a\vee a_1}}\right\}$$

$$= 4\max\left\{\left(\sum_{t=0}^{T-1}\|u_t\|_U\right)^{\frac{b_0}{a\vee a_1}},\left(\sum_{t=0}^{T-1}\|u_t\|_U\right)^{\frac{b_1}{a\wedge a_0}}\right\}.$$

Above, the last equality follows from Proposition C.1. The first claimed inequality now follows by combining the inequalities. The second claimed inequality follows from Jensen's inequality. $\square$

C.2 Proof of Proposition 3.3

Let ξ_1, ξ_2 and $\{u_t\}_{t \geqslant 0}$ be arbitrary. Fix a $T \geqslant 1$. Define two dynamics, for $t = 0, ..., T-1$:

$$x_{t+1} = f(x_t, u_t), \quad x_0 = \xi_1,$$
$$y_{t+1} = f(y_t, 0), \quad y_0 = \xi_2.$$

Now define $V_t := V(x_t, y_t)$. Then for $t \in \{0, ..., T-1\}$, by the assumed inequality,

$$\begin{aligned}
V_{t+1} = V(x_{t+1}, y_{t+1}) &= V(f(x_t, u_t), f(y_t, 0)) \\
&\leqslant V(x_t, y_t) - \mathfrak{a} \min\{\|x_t - y_t\|_X^{a_0}, \|x_t - y_t\|_X^{a_1}\} + \mathfrak{b} \max\{\|u_t\|_U^{b_0}, \|u_t\|_U^{b_1}\} \\
&= V_t - \mathfrak{a} \min\{\|x_t - y_t\|_X^{a_0}, \|x_t - y_t\|_X^{a_1}\} + \mathfrak{b} \max\{\|u_t\|_U^{b_0}, \|u_t\|_U^{b_1}\}.
\end{aligned}$$

Therefore, we have:

$$V_T + \mathfrak{a} \sum_{t=0}^{T-1} \min\{\|x_t - y_t\|_X^{a_0}, \|x_t - y_t\|_X^{a_1}\} \leqslant V_0 + \mathfrak{b} \sum_{t=0}^{T-1} \max\{\|u_t\|_U^{b_0}, \|u_t\|_U^{b_1}\}.$$

But this implies:

$$\underline{\alpha}\|x_T - y_T\|_X^a + \mathfrak{a} \sum_{t=0}^{T-1} \min\{\|x_t - y_t\|_X^{a_0}, \|x_t - y_t\|_X^{a_1}\} \leqslant \overline{\alpha}\|\xi_1 - \xi_2\|_X^a + \mathfrak{b} \sum_{t=0}^{T-1} \max\{\|u_t\|_U^{b_0}, \|u_t\|_U^{b_1}\}.$$

Next, observe that by Proposition C.1:

$$\underline{\alpha}\|x_T - y_T\|_X^a + \mathfrak{a} \sum_{t=0}^{T-1} \min\{\|x_t - y_t\|_X^{a_0}, \|x_t - y_t\|_X^{a_1}\} \geqslant (\underline{\alpha} \wedge \mathfrak{a}) \sum_{t=0}^{T} \min\{\|x_t - y_t\|_X^{a \wedge a_0}, \|x_t - y_t\|_X^{a \vee a_1}\}.$$

Therefore,

$$\sum_{t=0}^{T} \min\{\|x_t - y_t\|_X^{a \wedge a_0}, \|x_t - y_t\|_X^{a \vee a_1}\} \leqslant \frac{\overline{\alpha}}{\underline{\alpha} \wedge \mathfrak{a}}\|\xi_1 - \xi_2\|_X^a + \frac{\mathfrak{b}}{\underline{\alpha} \wedge \mathfrak{a}} \sum_{t=0}^{T-1} \max\{\|u_t\|_U^{b_0}, \|u_t\|_U^{b_1}\}.$$

D Examples of Incremental Gain Stability Proofs

D.1 Contraction

Recall the following definition of autonomously contracting in Proposition 3.4, which we duplicate below for convenience.

Definition D.1. *Consider the dynamics $x_{t+1} = f(x_t, u_t)$. We say that f is autonomously contracting if there exists a positive definite metric $M(x)$ and a scalar $\gamma \in (0, 1)$ such that:*

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

Now for what follows, we let d_M denote the geodesic distance under the metric $M(x)$:

$$d_M(x,y) := \inf_{\gamma \in \Gamma(x,y)} \sqrt{\int_0^1 \frac{\partial \gamma}{\partial s}(s)^\mathsf{T} M(\gamma(s)) \frac{\partial \gamma}{\partial s}(s)\, ds}.$$

Here, $\Gamma(x,y)$ is the set of smooth curves $\gamma : [0,1] \to \mathbb{R}^n$ satisfying $\gamma(0) = x$ and $\gamma(1) = y$. The next result shows that distances contract in the metric d_M under an application of the dynamics f:

Proposition D.2 (cf. Lemma 1 of Pham [39]). *Suppose that f is autonomously contracting. Then for all $x, y \in \mathbb{R}^n$ we have:*

$$d_M^2(f(x,0), f(y,0)) \leqslant \gamma d_M^2(x,y).$$

The next result shows that the Euclidean norm lower and upper bounds the geodesic distance under M as long as M is uniformly bounded.

Proposition D.3 (cf. Proposition D.2 of Boffi et al. [19]). *Suppose that $\underline{\mu} I \preccurlyeq M(x) \preccurlyeq \overline{\mu} I$. Then for all $x, y \in \mathbb{R}^n$ we have:*

$$\sqrt{\underline{\mu}}\|x - y\|_2 \leqslant d_M(x,y) \leqslant \sqrt{\overline{\mu}}\|x - y\|_2.$$

We now restate and prove Proposition 3.4.

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

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

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

Proof. Fix initial conditions ξ_1, ξ_2 and an input sequence $\{u_t\}_{t \geqslant 0}$. Consider two systems:

$$x_{t+1} = f(x_t, u_t), \ \ x_0 = \xi_1,$$
$$y_{t+1} = f(y_t, 0), \ \ y_0 = \xi_2.$$

Now fix a $t \geqslant 0$. We have:

$$
\begin{aligned}
d_M(x_{t+1}, y_{t+1}) &= d_M(f(x_t, u_t), f(y_t, 0)) \\
&\overset{(a)}{\leqslant} d_M(f(x_t, u_t), f(x_t, 0)) + d_M(f(x_t, 0), f(y_t, 0)) \\
&\overset{(b)}{\leqslant} \sqrt{\gamma} d_M(x_t, y_t) + \sqrt{\overline{\mu}}\|f(x_t, u_t) - f(x_t, 0)\|_2 \\
&\overset{(c)}{\leqslant} \sqrt{\gamma} d_M(x_t, y_t) + L_u \sqrt{\overline{\mu}}\|u_t\|_2.
\end{aligned}
$$

Above, (a) is triangle inequality, (b) follows from Proposition D.2 and Proposition D.3, and (c) follows from the Lipschitz assumption. Now unroll this recursion, to yield for all $t \geqslant 0$

$$d_M(x_t, y_t) \leqslant \gamma^{t/2} d_M(x_0, y_0) + L_u \sqrt{\bar{\mu}} \sum_{k=0}^{t-1} \gamma^{(t-1-k)/2} \|u_k\|_2.$$

Using the upper and lower bounds on d_M from Proposition D.3, we obtain for all $t \geqslant 0$:

$$\sqrt{\underline{\mu}} \|x_t - y_t\|_2 \leqslant \gamma^{t/2} \sqrt{\bar{\mu}} \|x_0 - y_0\|_2 + L_u \sqrt{\bar{\mu}} \sum_{k=0}^{t-1} \gamma^{(t-1-k)/2} \|u_k\|_2.$$

Now dividing both sides by $\sqrt{\underline{\mu}}$ and summing the LHS from $t = 0$ to T,

$$\sum_{t=0}^{T} \|x_t - y_t\|_2 \leqslant \sqrt{\frac{\bar{\mu}}{\underline{\mu}}} \frac{1}{1 - \sqrt{\gamma}} \|x_0 - y_0\|_2 + L_u \sqrt{\frac{\bar{\mu}}{\underline{\mu}}} \frac{1}{1 - \sqrt{\gamma}} \sum_{t=0}^{T-1} \|u_t\|_2.$$

$\square$

D.2   Scalar Example with $p \in (0, \infty)$

Recall we are interested in the family of systems:

$$x_{t+1} = x_t - \eta x_t \frac{|x_t|^p}{1 + |x_t|^p} + \eta u_t, \ \ p \in (0, \infty).$$

We define for convenience the function $h : \mathbb{R} \to \mathbb{R}$ as:

$$h(x) := x \frac{|x|^p}{1 + |x|^p},$$

and observe that $h(-x) = -h(x)$ for all $x \in \mathbb{R}$. Our first proposition gives us a lower bound which we will utilize later on.

Proposition D.4. *For every $x, y \in \mathbb{R}$,*

$$\operatorname{sgn}(x - y) \left[x \frac{|x|^p}{1 + |x|^p} - y \frac{|y|^p}{1 + |y|^p} \right] \geqslant \frac{1}{2^{2+p}} \min\{|x - y|, |x - y|^{1+p}\}.$$

Proof. Define

$$z(x, y) := \operatorname{sgn}(x - y)(h(x) - h(y)).$$

It is straightforward to see that the following properties of z hold for all $x, y \in \mathbb{R}$:

1. $z(x, y) = z(y, x)$.

2. $z(-x, -y) = z(x, y)$.

We want to show that for $x, y \in \mathbb{R}$,

$$z(x, y) \geqslant \frac{1}{2^{2+p}} \min\{|x - y|, |x - y|^{1+p}\}. \tag{D.1}$$

Observe that

$$z(0, y) = \frac{|y|^{p+1}}{1 + |y|^p} \geqslant \frac{1}{2} \min\{|y|, |y|^{1+p}\},$$

so (D.1) holds for $x = 0$, $y \in \mathbb{R}$. Next, by symmetry, $z(x, 0) = z(0, x)$, so (D.1) also holds for $x \in \mathbb{R}$, $y = 0$. Furthermore, (D.1) holds for $(x, y) = (0, 0)$ trivially. Finally, we can assume that $x \leqslant y$ since $z(x, y) = z(y, x)$. Therefore, for remainder of the proof, we may assume that $x \neq 0$, $y \neq 0$, and $x < y$.

Case 1: $0 < x < y$. Since $z \mapsto z/(1 + z)$ is monotonically increasing on $\mathbb{R}_{\geqslant 0}$,

$$z(x, y) = \frac{y^{1+p}}{1 + y^p} - \frac{x^{1+p}}{1 + x^p} = y \frac{y^p}{1 + y^p} - x \frac{x^p}{1 + x^p} \geqslant (y - x) \frac{y^p}{1 + y^p}.$$

If $y > 1$, then we observe that $z(x, y) \geqslant (y - x)/2 = \frac{1}{2}|x - y|$. Now we assume $y \leqslant 1$. Then, because $z \mapsto z^p$ is monotonically increasing on $\mathbb{R}_{\geqslant 0}$,

$$(y - x) \frac{y^p}{1 + y^p} \geqslant \frac{1}{2}(y - x)y^p \geqslant \frac{1}{2}(y - x)(y - x)^p = \frac{1}{2}(y - x)^{1+p} = \frac{1}{2}|x - y|^{1+p}.$$

Thus, (D.1) holds when $0 < x < y$.

Case 2: $x < 0 < y$. Let us assume wlog that $y \geqslant |x|$, otherwise we can swap x, y by considering $z(-y, -x) = z(x, y)$ instead. Now, we have

$$z(x, y) = \frac{y^{1+p}}{1 + y^p} + \frac{|x|^{1+p}}{1 + |x|^p} \geqslant \frac{y^{1+p}}{1 + y^p}.$$

If $y > 1$, then we lower bound:

$$z(x, y) \geqslant y/2 \geqslant (y + |x|)/4 = \frac{1}{4}|x - y|.$$

On the other hand, if $y \leqslant 1$, then

$$z(x, y) \geqslant y^{1+p}/2 = \max\{y^{1+p}, |x|^{1+p}\}/2 \overset{(a)}{\geqslant} \frac{1}{2^{2+p}}(y + |x|)^{1+p} = \frac{1}{2^{2+p}}|x - y|^{1+p}.$$

Here, (a) holds because for non-negative $a, b \in \mathbb{R}$, we have

$$(a + b)^{1+p} \leqslant (2 \max\{a, b\})^{1+p} = 2^{1+p} \max\{a^p, b^p\}.$$

Therefore, (D.1) holds when $x < 0 < y$.

28

Case 3: $x < y < 0$. In this case, we have $z(x,y) = z(-y,-x)$. By swapping x, y, we reduce to Case 1 where we know that (D.1) already holds. $\qquad\square$

Next, we show that the sign of $x - y$ is preserved under a perturbation $x - y - \eta(h(x) - h(y))$, as long as $\eta \geqslant 0$ is small enough.

Proposition D.5. *Let η satisfy $0 \leqslant \eta < \frac{4}{5+p}$. We have that:*

$$\operatorname{sgn}\left((x - y) - \eta\left[x\frac{|x|^p}{1 + |x|^p} - y\frac{|y|^p}{1 + |y|^p}\right]\right) = \operatorname{sgn}(x - y).$$

Proof. Define

$$z(x,y) := (x - y) - \eta(h(x) - h(y)).$$

We want to show that for all $x, y \in \mathbb{R}$,

$$\operatorname{sgn}(z(x,y)) = \operatorname{sgn}(x - y). \tag{D.2}$$

It is straightforward to see that for all $x, y \in \mathbb{R}$,

1. $z(x, y) = -z(y, x)$,

2. $z(-x, -y) = -z(x, y)$.

Therefore, we have for all $x, y \in \mathbb{R}$:

$$\operatorname{sgn}(z(x,y)) = \operatorname{sgn}(x - y) \implies \{\operatorname{sgn}(z(y,x)) = \operatorname{sgn}(y - x)\} \bigwedge \{\operatorname{sgn}(z(-x,-y)) = \operatorname{sgn}((-x) - (-y))\}.$$

Observe that (D.2) holds trivially when $x = y$. Furthermore, when $y > 0$:

$$z(0, y) = -y + \eta\frac{y^{1+p}}{1 + y^p} \leqslant -y + \eta y = -(1 - \eta)y < 0.$$

Therefore, (D.2) holds when $x = 0$ and $y > 0$, which implies it also holds when $x = 0$ and $y < 0$ (and hence it holds when $x = 0$ and $y \in \mathbb{R}$). But this also implies that (D.2) holds when $x \in \mathbb{R}$ and $y = 0$. Hence, as we did in Proposition D.4, we will assume that $x \neq 0$, $y \neq 0$, and $x < y$.

Case 1: $0 < x < y$. In this case we have:

$$z(x, y) = (x - y) - \eta\left[\frac{x^{1+p}}{1 + x^p} - \frac{y^{1+p}}{1 + y^p}\right].$$

The derivative of $h(x)$ when $x > 0$ is:

$$h'(x) = \frac{x^p(x^p + p + 1)}{(x^p + 1)^2}.$$

Clearly $h'(x) \geqslant 0$ when $x > 0$. Furthermore, one can check that $\sup_{x \geqslant 0} \frac{x}{(1+x)^2} = 1/4$. Therefore:

$$h'(x) = \left(\frac{x^p}{1 + x^p}\right)^2 + (p + 1)\frac{x^p}{(1 + x^p)^2} \leqslant 1 + (p + 1)/4.$$

29

Therefore:

$$(x - y) - \eta \left[\frac{x^{1+p}}{1 + x^p} - \frac{y^{1+p}}{1 + y^p} \right] = (x - y) + \eta(h(y) - h(x))$$

$$= (x - y) + \eta(y - x) \int_0^1 h'(ty + (1 - t)x)\, dt$$

$$\leqslant (x - y) + \eta(1 + (p + 1)/4)(y - x)$$

$$= (1 - \eta(1 + (p + 1)/4))(x - y).$$

By assumption, we have that $1 - \eta(1 + (p + 1)/4) > 0$ and hence $z(x, y) < 0$. This shows that $\mathrm{sgn}(z(x, y)) = \mathrm{sgn}(x - y)$, and hence (D.2) holds in this case.

Case 2: $x < 0 < y$. Here we have:

$$z(x, y) = -(|x| + y) + \eta \left[\frac{|x|^{1+p}}{1 + |x|^p} + \frac{y^{1+p}}{1 + y^p} \right]$$

$$\leqslant -(|x| + y) + \eta(|x| + y) = -(1 - \eta)(|x| + y) < 0.$$

Hence, we have $\mathrm{sgn}(z(x, y)) = \mathrm{sgn}(x - y)$, and hence (D.2) holds in this case.

Case 3: $x < y < 0$. By a reduction to Case 1 when $0 < x < y$:

$$\mathrm{sgn}(z(x, y)) = -\mathrm{sgn}(z(y, x)) = -\mathrm{sgn}(y - x) = \mathrm{sgn}(x - y).$$

$\square$

Proposition D.6. *Let η satisfy $0 \leqslant \eta < \frac{4}{5+p}$. Put $V(x, y) = |x - y|$ and $f(x, u) := x - \eta \frac{x|x|^p}{1+|x|^p} + \eta u$. Then we have for all $x, y, u \in \mathbb{R}$:*

$$V(f(x, u), f(y, 0)) - V(x, y) \leqslant -\frac{\eta}{2^{2+p}} \min\{|x - y|, |x - y|^{1+p}\} + \eta|u|.$$

Proof. Because $\mathrm{sgn}(x)$ is an element of $\partial|x|$, by convexity of $|\cdot|$,

$$|a| - |b| \leqslant \mathrm{sgn}(a)(a - b).$$

Therefore:

$$V(f(x, u), f(y)) - V(x, y) = \left| (x - y) - \eta \left[x\frac{|x|^p}{1 + |x|^p} - y\frac{|y|^p}{1 + |y|^p} \right] + \eta u \right| - |x - y|$$

$$\leqslant \left| (x - y) - \eta \left[x\frac{|x|^p}{1 + |x|^p} - y\frac{|y|^p}{1 + |y|^p} \right] \right| - |x - y| + \eta|u|$$

$$\leqslant \mathrm{sgn}\left((x - y) - \eta \left[x\frac{|x|^p}{1 + |x|^p} - y\frac{|y|^p}{1 + |y|^p} \right] \right) \left\{ -\eta \left[x\frac{|x|^p}{1 + |x|^p} - y\frac{|y|^p}{1 + |y|^p} \right] \right\} + \eta|u|$$

$$\overset{(a)}{=} -\eta\, \mathrm{sgn}(x - y) \left[x\frac{|x|^p}{1 + |x|^p} - y\frac{|y|^p}{1 + |y|^p} \right] + \eta|u|$$

$$\overset{(b)}{\leqslant} -\frac{\eta}{2^{2+p}} \min\{|x - y|, |x - y|^{1+p}\} + \eta|u|.$$

Above, (a) is Proposition D.5 and (b) is Proposition D.4. $\square$

Note that Proposition 3.5 is an immediate consequence of Proposition D.6 with Proposition 3.3.

E   Proof of Theorem 4.3 and Theorem 4.4

In this section we prove our main technical contribution, which is a theoretical analysis of Algorithm 1.

Recall the simplifying assumptions in Assumption 4.2 for $\Psi = (a, a_0, a_1, b_0, b_1, \zeta, \gamma)$ appearing in the definition of incremental gain stability: $a = a_0$, $b = b_0 = b_1$, both $\zeta \geqslant 1$ and $\gamma \geqslant 1$, and $a \geqslant b$. Also, recall that Π is our policy class, $\mathcal{S}_\Psi$ is the set of all policies π such that $(f_{\mathrm{cl}}^\pi, \mathsf{Id})$ is Ψ-incrementally-gain-stable, and $\Pi_\Psi = \Pi \cap \mathcal{S}_\Psi$. Finally, for any $\pi_d \in \Pi_\Psi$ and $\pi_1, \pi_2 \in \Pi$, recall the definition of ℓ_{π_d}:

$$\ell_{\pi_d}(\xi; \pi_1, \pi_2) = \sum_{t=0}^{T-1} \|\Delta_{\pi_1, \pi_2}(\varphi_t^{\pi_d}(\xi))\|_2.$$

E.1   Uniform Convergence Toolbox

Our main tool will be the following uniform convergence result.

Proposition E.1. *Define B_ℓ to be the constant:*

$$B_\ell := \sup_{\pi_d \in \Pi_\Psi} \sup_{\pi_1, \pi_2 \in \Pi} \operatorname*{ess\,sup}_{\xi \sim \mathcal{D}} \ell_{\pi_d}(\xi; \pi_1, \pi_2). \tag{E.1}$$

Next, define the following two notions of Rademacher complexity for the policy class Π:

$$\mathcal{R}_m(\Pi) := \sup_{\pi_d \in \Pi_\Psi} \sup_{\pi_g \in \Pi} \mathbb{E}_{\{\xi_i\}} \mathbb{E}_{\{\varepsilon_i\}} \left[\sup_{\pi \in \Pi} \frac{1}{m} \sum_{i=1}^m \varepsilon_i \ell_{\pi_d}(\xi_i; \pi, \pi_g) \right], \tag{E.2}$$

$$\mathcal{R}_m^{\mathrm{incr}}(\Pi) := \mathbb{E}_{\{\xi_i\}} \mathbb{E}_{\{\varepsilon_i\}} \left[\sup_{\pi_d \in \Pi_\Psi} \sup_{\pi_1, \pi_2 \in \Pi} \frac{1}{m} \sum_{i=1}^m \varepsilon_i \ell_{\pi_d}(\xi; \pi_1, \pi_2) \right]. \tag{E.3}$$

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

$$\sup_{\pi \in \Pi} \left| \mathbb{E}_{\xi \sim \mathcal{D}} \ell_{\pi_d}(\xi; \pi, \pi_g) - \frac{1}{m} \sum_{i=1}^m \ell_{\pi_d}(\xi_i; \pi, \pi_g) \right| \leqslant 2\mathcal{R}_m(\Pi) + B_\ell \sqrt{\frac{\log(2/\delta)}{m}}, \tag{E.4}$$

Furthermore, with probability at least $1 - \delta$ (over $\xi_1, ..., \xi_m$), we have:

$$\sup_{\pi_d \in \Pi_\Psi} \sup_{\pi_1, \pi_2 \in \Pi} \left| \mathbb{E}_{\xi \sim \mathcal{D}} \ell_{\pi_d}(\xi; \pi_1, \pi_2) - \frac{1}{m} \sum_{i=1}^m \ell_{\pi_d}(\xi_i; \pi_1, \pi_2) \right| \leqslant 2\mathcal{R}_m^{\mathrm{incr}}(\Pi) + B_\ell \sqrt{\frac{\log(2/\delta)}{m}}, \tag{E.5}$$

Proof. This follows from standard uniform convergence results, see e.g., Wainright [40]. □

In order to use Proposition E.1, we need to have upper bounds on the constants B_ℓ and $\mathcal{R}_m(\Pi)$. We first give an upper bound on B_ℓ.

Proposition E.2. *Under Assumption 4.1 and Assumption 4.2, we have that:*

$$B_\ell = \sup_{\pi_d \in \Pi_\Psi} \sup_{\pi_1, \pi_2 \in \Pi} \operatorname*{ess\,sup}_{\xi \sim \mathcal{D}} \ell_{\pi_d}(\xi; \pi_1, \pi_2) \leqslant 2\zeta^{1/a} B_0 L_\Delta T^{1-1/a_1}.$$

Proof. Let $\pi_d \in \Pi_\Psi$ and $\pi_1, \pi_2 \in \Pi$. Since $\Delta_{\pi_1, \pi_2}(0) = 0$, by the assumption Δ_{π_1, π_2} is L_Δ-Lipschitz, we have:

$$\ell_{\pi_d}(\xi; \pi_1, \pi_2) = \sum_{t=0}^{T-1} \|\Delta_{\pi_1, \pi_2}(\varphi_t^{\pi_d}(\xi))\|_2 \leqslant L_\Delta \sum_{t=0}^{T-1} \|\varphi_t^{\pi_d}(\xi)\|_2 \leqslant 2\zeta^{1/a} B_0 L_\Delta T^{1-1/a_1}.$$

Above, the last inequality follows from Proposition C.3. □

We now give a bounds on the Rademacher complexities $\mathcal{R}_m(\Pi)$ and $\mathcal{R}_m^{\mathrm{incr}}(\Pi)$. For our main proof, we will only utilize the uniform convergence guarantee of (E.4). Hence, we only require a bound on $\mathcal{R}_m(\Pi)$. However, we also state and prove a bound for $\mathcal{R}_m^{\mathrm{incr}}(\Pi)$, as we believe it is of independent interest and could be used to provide an analysis of more practical variations of CMILe.

Proposition E.3. *Let* $\Pi = \{\pi(x, \theta) : \theta \in \mathbb{R}^q, \|\theta\|_2 \leqslant B_\theta\}$ *for a fixed twice continuously differentiable map* π. *Define the constant* $L_{\partial^2 \pi}$ *to be:*

$$L_{\partial^2 \pi} = 1 \vee \sup_{\|x\|_2 \leqslant \zeta^{1/a} B_0, \|\theta\|_2 \leqslant B_\theta} \left\| \frac{\partial^2 \pi}{\partial \theta \partial x} \right\|_{\ell^2(\mathbb{R}^q) \to M(\mathbb{R}^{d \times n})}.$$

Here, $M(\mathbb{R}^{d \times n})$ *is the Banach space of* $d \times n$ *real-valued matrices equipped with the operator norm. Under Assumption 4.1 and Assumption 4.2, we have that:*

$$\mathcal{R}_m(\Pi) \leqslant 65\zeta^{1/a} B_0 B_g B_\theta L_{\partial^2 \pi} T^{1-1/a_1} \sqrt{\frac{q}{m}}. \tag{E.6}$$

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:

$$\pi(x, \theta_1) - \pi(x, \theta_2) = \left(\int_0^1 \left[\frac{\partial \pi}{\partial x}(s_1 x, \theta_1) - \frac{\partial \pi}{\partial x}(s_1 x, \theta_2) \right] ds_1 \right) x$$
$$= \left(\int_0^1 \left[\int_0^1 \frac{\partial^2 \pi}{\partial \theta \partial x}(s_1 x, s_2 \theta_1 + (1 - s_2)\theta_2)(\theta_1 - \theta_2) ds_2 \right] ds_1 \right) x.$$

Now supposing $\|x\|_2 \leqslant \zeta^{1/a} B_0$ and $\|\theta_i\|_2 \leqslant B_\theta$ for $i = 1, 2$, then:

$$\|\pi(x; \theta_1) - \pi(x; \theta_2)\|_2 \leqslant L_{\partial^2 \pi} \|x\|_2 \|\theta_1 - \theta_2\|_2. \tag{E.7}$$

Next, we have that:

$$|\ell_{\pi_d}(\xi; \pi_1, \pi_g) - \ell_{\pi_d}(\xi; \pi_2, \pi_g)| \leqslant \sum_{t=0}^{T-1} |\|\Delta_{\pi_1, \pi_g}(\varphi_t^{\pi_d}(\xi))\|_2 - \|\Delta_{\pi_2, \pi_g}(\varphi_t^{\pi_d}(\xi))\|_2|$$
$$\leqslant \sum_{t=0}^{T-1} \|\Delta_{\pi_1, \pi_2}(\varphi_t^{\pi_d}(\xi))\|_2$$
$$\overset{(a)}{\leqslant} B_g \sum_{t=0}^{T-1} \|\pi_1(\varphi_t^{\pi_d}(\xi)) - \pi_2(\varphi_t^{\pi_d}(\xi))\|_2$$
$$\overset{(b)}{\leqslant} B_g L_{\partial^2 \pi} \left(\sum_{t=0}^{T-1} \|\varphi_t^{\pi_d}(\xi)\|_2 \right) \|\theta_1 - \theta_2\|_2$$
$$\overset{(c)}{\leqslant} 2\zeta^{1/a} B_0 B_g L_{\partial^2 \pi} T^{1-1/a_1} \|\theta_1 - \theta_2\|_2.$$

32

Here, (a) by bound on $\|g(x)\|_{\mathrm{op}} \leqslant B_g$, (b) by Equation (E.7) and the fact that $\|\varphi_t^{\pi_d}(\xi)\|_2 \leqslant \zeta^{1/a} B_0$, and (c) by using Proposition C.3 to bound $\sum_{t=0}^{T-1} \|\varphi_t^{\pi_d}(\xi)\|_2$. 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 $\|\cdot\|_{\mathbb{P}_m}$ metric over Π as:

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

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

$$\|\pi_1 - \pi_2\|_{\mathbb{P}_m} \leqslant 2\zeta^{1/a} B_0 B_g L_{\partial^2 \pi} T^{1-1/a_1} \|\theta_1 - \theta_2\|_2.$$

Therefore by Dudley's entropy integral (cf. Wainwright [40]):

$$\mathcal{R}_m^{\mathrm{basic}}(\Pi) \leqslant 24 \sup_{\pi_d \in \Pi_\Psi} \sup_{\pi_g \in \Pi} \mathbb{E}_{\{\xi_i\}} \frac{1}{\sqrt{m}} \int_0^\infty \sqrt{\log N(\varepsilon; \Pi, \|\cdot\|_{\mathbb{P}_m})} \, d\varepsilon$$

$$\leqslant 48 \zeta^{1/a} B_0 B_g L_{\partial^2 \pi} T^{1-1/a_1} \frac{1}{\sqrt{m}} \int_0^\infty \sqrt{\log N(\varepsilon; \Theta, \|\cdot\|_2)} \, d\varepsilon$$

$$\leqslant 65 \zeta^{1/a} B_0 B_g B_\theta L_{\partial^2 \pi} T^{1-1/a_1} \sqrt{\frac{q}{m}}.$$

Above, the last inequality follows by bounding

$$\frac{1}{\sqrt{m}} \int_0^\infty \sqrt{\log N(\varepsilon; \Theta, \|\cdot\|_2)} \, d\varepsilon \leqslant \frac{B_\theta}{\sqrt{m}} \int_0^\infty \sqrt{\log N(\varepsilon; \mathbb{B}_2(1), \|\cdot\|_2)} \, d\varepsilon$$

$$\leqslant B_\theta \sqrt{\frac{q}{m}} \int_0^1 \sqrt{\log(1 + 2/\varepsilon)} \, d\varepsilon$$

$$\leqslant 1.3526 B_\theta \sqrt{\frac{q}{m}}.$$

This yields (E.6). $\qquad \square$

For the next proposition, we bound $\mathcal{R}_m^{\mathrm{incr}}(\Pi)$. As mentioned before, we provide this result for independent interest; the reader who is primarily interested in only the proof of the main results may safely skip this next proposition.

Proposition E.4. *Let $\Pi = \{\pi(x, \theta) : \theta \in \mathbb{R}^q, \|\theta\|_2 \leqslant B_\theta\}$ for a fixed twice continuously differentiable map π. Define the constant $L_{\partial^2 \pi}$ to be:*

$$L_{\partial^2 \pi} = 1 \vee \sup_{\|x\|_2 \leqslant \zeta^{1/a} B_0, \|\theta\|_2 \leqslant B_\theta} \left\| \frac{\partial^2 \pi}{\partial \theta \partial x} \right\|_{\ell^2(\mathbb{R}^q) \to M(\mathbb{R}^{d \times n})}.$$

Here, $M(\mathbb{R}^{d \times n})$ is the Banach space of $d \times n$ real-valued matrices equipped with the operator norm. Under Assumption 4.1 and Assumption 4.2, we have that:

$$\mathcal{R}_m^{\mathrm{incr}}(\Pi) \leqslant 3456 \sqrt{\frac{a_1}{b}} \gamma^{1/a} \zeta^{b/a^2} L_\Delta B_\theta B_0 B_g L_{\partial^2 \pi} T^{(1-1/a_1)(1+b/a)} \sqrt{\frac{q}{m}}$$

$$+ 432 \zeta^{1/a} B_\theta B_0 B_g L_{\partial^2 \pi} T^{1-1/a_1} \sqrt{\frac{q}{m}}. \tag{E.8}$$

Proof. Let $\pi_d, \pi'_d \in \Pi_\Psi$ and $\pi_1, \pi_2, \pi'_1, \pi'_2 \in \Pi$. Let the parameters $\theta_d, \theta_1, \theta_2$ correspond to π_d, π_1, π_2 and the parameters ν_d, ν_1, ν_2 correspond to π'_d, π'_1, π'_2. Let $\ell_{\pi_d}(\pi_1, \pi_2)$ be shorthand for $\ell_{\pi_d}(\xi; \pi_1, \pi_2)$ (the ξ will be implicit). We decompose:

$$\begin{aligned}
\ell_{\pi_d}(\pi_1, \pi_2) - \ell_{\pi'_d}(\pi'_1, \pi'_2) &= \ell_{\pi_d}(\pi_1, \pi_2) - \ell_{\pi'_d}(\pi_1, \pi_2) \\
&\quad + \ell_{\pi'_d}(\pi_1, \pi_2) - \ell_{\pi'_d}(\pi'_1, \pi_2) \\
&\quad + \ell_{\pi'_d}(\pi'_1, \pi_2) - \ell_{\pi'_d}(\pi'_1, \pi'_2).
\end{aligned}$$

We first bound:

$$\begin{aligned}
|\ell_{\pi_d}(\xi; \pi_1, \pi_2) - \ell_{\pi'_d}(\xi; \pi_1; \pi_2)| &\leqslant \sum_{t=0}^{T-1} |\|\Delta_{\pi_1, \pi_2}(\varphi_t^{\pi_d}(\xi))\|_2 - \|\Delta_{\pi_1, \pi_2}(\varphi_t^{\pi'_d}(\xi))\|_2| \\
&\leqslant L_\Delta \sum_{t=0}^{T-1} \|\varphi_t^{\pi_d}(\xi) - \varphi_t^{\pi'_d}(\xi)\|_2.
\end{aligned}$$

Now, we write:

$$\begin{aligned}
f_{\mathrm{cl}}^{\pi_d}(x) &= f(x) + g(x)\pi_d(x) \\
&= f(x) + g(x)\pi'_d(x) + g(x)(\pi_d(x) - \pi'_d(x)) \\
&= f_{\mathrm{cl}}^{\pi'_d}(x) + \Delta_{\pi_d, \pi'_d}(x).
\end{aligned}$$

Because $(f_{\mathrm{cl}}^{\pi'_d}, \mathsf{Id})$ is Ψ-incrementally-gain-stable, by Proposition C.4,

$$\begin{aligned}
&\sum_{t=0}^{T-1} \|\varphi_t^{\pi_d}(\xi) - \varphi_t^{\pi'_d}(\xi)\|_2 \\
&\leqslant 4\gamma^{1/a} T^{1-1/a_1} \left[\left(\sum_{t=0}^{T-2} \|\Delta_{\pi_d, \pi'_d}(\varphi_t^{\pi'_d}(\xi))\|_2 \right)^{b/a_1} + \left(\sum_{t=0}^{T-2} \|\Delta_{\pi_d, \pi'_d}(\varphi_t^{\pi'_d}(\xi))\|_2 \right)^{b/a} \right].
\end{aligned}$$

Next, by (E.7), we have

$$\sum_{t=0}^{T-1} \|\Delta_{\pi_d, \pi'_d}(\varphi_t^{\pi'_d}(\xi))\|_2 \leqslant 2\zeta^{1/a} B_0 B_g L_{\partial^2 \pi} T^{1-1/a_1} \|\theta_d - \nu_d\|_2.$$

Therefore:

$$\begin{aligned}
&|\ell_{\pi_d}(\xi; \pi_1, \pi_2) - \ell_{\pi'_d}(\xi; \pi_1; \pi_2)| \\
&\leqslant 4\gamma^{1/a} L_\Delta T^{1-1/a_1} \left(2\zeta^{1/a} B_0 B_g L_{\partial^2 \pi} T^{1-1/a_1} \|\theta_d - \nu_d\|_2 \right)^{b/a_1} \\
&\quad + 4\gamma^{1/a} L_\Delta T^{1-1/a_1} \left(2\zeta^{1/a} B_0 B_g L_{\partial^2 \pi} T^{1-1/a_1} \|\theta_d - \nu_d\|_2 \right)^{b/a} \\
&\leqslant 8\gamma^{1/a} \zeta^{b/a^2} L_\Delta (B_0 B_g L_{\partial^2 \pi})^{b/a} T^{(1-1/a_1)(1+b/a)} \left(\|\theta_d - \nu_d\|_2^{b/a_1} + \|\theta_d - \nu_d\|_2^{b/a} \right).
\end{aligned}$$

34

Next, we have that:

$$|\ell_{\pi'_d}(\xi; \pi_1, \pi_2) - \ell_{\pi'_d}(\xi; \pi'_1, \pi_2)| \leqslant \sum_{t=0}^{T-1} |\|\Delta_{\pi_1,\pi_2}(\varphi_t^{\pi'_d}(\xi))\|_2 - \|\Delta_{\pi'_1,\pi_2}(\varphi_t^{\pi'_d}(\xi))\|_2|$$

$$\leqslant \sum_{t=0}^{T-1} \|\Delta_{\pi_1,\pi'_1}(\varphi_t^{\pi'_d}(\xi))\|_2$$

$$\leqslant 2\zeta^{1/a} B_0 B_g L_{\partial^2\pi} T^{1-1/a_1} \|\theta_1 - \nu_1\|_2.$$

The identical bound holds for $|\ell_{\pi'_d}(\pi'_1, \pi_2) - \ell_{\pi'_d}(\pi'_1, \pi'_2)|$. Therefore, we have shown:

$$|\ell_{\pi_d}(\xi; \pi_1, \pi_2) - \ell_{\pi_d}(\xi; \pi'_1, \pi'_2)|$$

$$\leqslant 8\gamma^{1/a}\zeta^{b/a^2} L_\Delta (B_0 B_g L_{\partial^2\pi})^{b/a} T^{(1-1/a_1)(1+b/a)} \left(\|\theta_d - \nu_d\|_2^{b/a_1} + \|\theta_d - \nu_d\|_2^{b/a}\right)$$

$$+ 2\zeta^{1/a} B_0 B_g L_{\partial^2\pi} T^{1-1/a_1} \left(\|\theta_1 - \nu_1\|_2 + \|\theta_2 - \nu_2\|_2\right).$$

Now put $\Lambda = \Pi_\Psi \times \Pi \times \Pi$. For a fixed $\xi_1, ..., \xi_m$, define the $\|\cdot\|_{\mathbb{P}_m}$ metric over Λ as:

$$\|\lambda - \lambda'\|_{\mathbb{P}_m}^2 := \frac{1}{m} \sum_{i=1}^m (\ell_{\pi_d}(\xi_i; \pi_1, \pi_2) - \ell_{\pi'_d}(\xi_i; \pi_1, \pi_2))^2,$$

with $\lambda = (\pi_d, \pi_1, \pi_2)$ and $\lambda' = (\pi'_d, \pi'_1, \pi'_2)$. The calculation above shows that for all $\lambda, \lambda' \in \Lambda$,

$$\|\lambda - \lambda'\|_{\mathbb{P}_m}$$

$$\leqslant 8\gamma^{1/a}\zeta^{b/a^2} L_\Delta (B_0 B_g L_{\partial^2\pi})^{b/a} T^{(1-1/a_1)(1+b/a)} \left(\|\theta_d - \nu_d\|_2^{b/a_1} + \|\theta_d - \nu_d\|_2^{b/a}\right)$$

$$+ 2\zeta^{1/a} B_0 B_g L_{\partial^2\pi} T^{1-1/a_1} \left(\|\theta_1 - \nu_1\|_2 + \|\theta_2 - \nu_2\|_2\right)$$

$$=: L_1 \left(\|\theta_d - \nu_d\|_2^{b/a_1} + \|\theta_d - \nu_d\|_2^{b/a}\right) + L_2 \left(\|\theta_1 - \nu_1\|_2 + \|\theta_2 - \nu_2\|_2\right).$$

Dudley's entropy integral (cf. Wainwright [40]) states that:

$$\mathcal{R}_m^{\mathrm{incr}}(\Pi) \leqslant 24\mathbb{E}_{\{\xi_i\}} \frac{1}{\sqrt{m}} \int_0^\infty \sqrt{\log N(\varepsilon; \Lambda, \|\cdot\|_{\mathbb{P}_m})} \, d\varepsilon.$$

Fix a $\varepsilon > 0$. Define

$$\bar{\varepsilon} := \min\left\{ \left(\frac{\varepsilon}{3L_1 B_\theta}\right)^{a_1/b}, \left(\frac{\varepsilon}{3L_1 B_\theta}\right)^{a/b} \right\},$$

and let $\{\theta_d^{(i)}\}$ be an $\bar{\varepsilon}$-covering of $\mathbb{B}_2(1)$. Let $\{\theta_1^{(i)}\}$ and $\{\theta_2^{(i)}\}$ be $\bar{\varepsilon}' := \varepsilon/(6L_2 B_\theta)$ coverings of $\mathbb{B}_2(1)$. Let $\lambda \in \Lambda$ with associated parameters $(\theta_d, \theta_1, \theta_2)$. Let $(B_\theta\theta_d^{(i)}, B_\theta\theta_1^{(j)}, B_\theta\theta_2^{(q)})$ be the nearest points to θ_d, θ_1, and θ_2 in their respective covers, and let $\tilde{\lambda}$ denote the associated element in Λ. Then:

$$\|\lambda - \tilde{\lambda}\|_{\mathbb{P}_m} \leqslant L_1 \left(\|\theta_d - B_\theta\theta_d^{(i)}\|_2^{b/a_1} + \|\theta_d - B_\theta\theta_d^{(i)}\|_2^{b/a}\right) + L_2(\|\theta_1 - B_\theta\theta_1^{(j)}\|_2 + \|\theta_2 - B_\theta\theta_2^{(q)}\|_2)$$

$$\leqslant L_1 B_\theta \left(\bar{\varepsilon}^{b/a_1} + \bar{\varepsilon}^{b/a}\right) + 2L_2 B_\theta\bar{\varepsilon}'$$

$$\leqslant \varepsilon.$$

This shows that the Cartesian product $\{\theta_d^{(i)}\} \times \{\theta_1^{(i)}\} \times \{\theta_2^{(i)}\}$ can be used to construct an ε-cover of Λ, which yields the bound:

$$\log N(\varepsilon; \Lambda, \|\cdot\|_{\mathbb{P}_m}) \leqslant q \log\left(1 + \frac{2}{\frac{\varepsilon}{6L_2 B_\theta}}\right) + q \begin{cases} \log\left(1 + \frac{2}{\left(\frac{\varepsilon}{3L_1 B_\theta}\right)^{a/b}}\right) & \text{if } 0 < \varepsilon \leqslant 3L_1 B_\theta \\ \log\left(1 + \frac{2}{\left(\frac{\varepsilon}{3L_1 B_\theta}\right)^{a_1/b}}\right) & \text{o.w.} \end{cases}.$$

Therefore:

$$\int_0^\infty \sqrt{\log N(\varepsilon; \Lambda, \|\cdot\|_{\mathbb{P}_m})}\, d\varepsilon$$

$$\leqslant \int_0^{\frac{1}{6L_2 B_\theta}} \sqrt{q \log\left(1 + \frac{2}{\frac{\varepsilon}{6L_2 B_\theta}}\right)}\, d\varepsilon$$

$$+ \int_0^{\frac{1}{3L_1 B_\theta}} \sqrt{q \log\left(1 + \frac{2}{\left(\frac{\varepsilon}{3L_1 B_\theta}\right)^{a/b}}\right)}\, d\varepsilon + \int_0^{\frac{1}{3L_1 B_\theta}} \sqrt{q \log\left(1 + \frac{2}{\left(\frac{\varepsilon}{3L_1 B_\theta}\right)^{a_1/b}}\right)}\, d\varepsilon$$

$$= 6L_2 B_\theta \sqrt{q} \int_0^1 \sqrt{\log(1 + 2/\varepsilon)}\, d\varepsilon + 3L_1 B_\theta \sqrt{q}\left(\int_0^1 \sqrt{\log(1 + 2/\varepsilon^{a/b})}\, d\varepsilon + \int_0^1 \sqrt{\log(1 + 2/\varepsilon^{a_1/b})}\, d\varepsilon\right)$$

$$\leqslant 9L_2 B_\theta \sqrt{q} + 18 L_1 B_\theta \sqrt{q} \sqrt{\frac{a_1}{b}}.$$

Above, the last inequality follows since any $p \geqslant 1$,

$$\int_0^1 \sqrt{\log(1 + 2/\varepsilon^p)}\, d\varepsilon \leqslant \int_0^1 \sqrt{\log(3/\varepsilon^p)}\, d\varepsilon = \int_0^1 \sqrt{\log 3 + p \log(1/\varepsilon)}\, d\varepsilon$$

$$\leqslant \sqrt{3} + \sqrt{p} \int_0^1 \sqrt{\log(1/\varepsilon)}\, d\varepsilon = \sqrt{3} + \sqrt{p}\sqrt{\pi}/2 \leqslant 3\sqrt{p}.$$

This yields (E.8). $\qquad\square$

E.2 Proof

We first state our main meta-theorem, from which we deduce our rates.

Theorem E.5. *Suppose that Assumption 4.1 and Assumption 4.2 hold. Suppose that E divides m. Define $\Gamma(m, E, \delta)$ as:*

$$\Gamma(m, E, \delta) := 2\mathcal{R}_{m/E}(\Pi) + B_\ell \sqrt{\frac{\log(4E/\delta)}{m/E}}. \tag{E.9}$$

Fix a $\delta \in (0, 1)$. Assume that for all $k \in \{0, ..., E - 2\}$:

$$c_k + \Gamma(m, E, \delta) \leqslant 1. \tag{E.10}$$

Suppose that $\alpha \in (0, 1]$ satisfies

$$\frac{(1-\alpha)^E}{\alpha} \leqslant 1.$$

For $k \in \{0, ..., E-1\}$, define $\beta_k(m, E, \delta)$ as:

$$\beta_k(m, E, \delta) := 2\alpha k\Gamma(m, E, \delta) + 8L_\Delta\gamma^{1/a}T^{1-1/a_1}\sum_{i=0}^{k-1}(\alpha c_i + \alpha\Gamma(m, E, \delta))^{\frac{b}{a_1}}. \tag{E.11}$$

With probability at least $1 - \delta$ (over $\{\xi_i^k\}_{i=1,k=0}^{m/E,E-1}$ drawn i.i.d. from $\mathcal{D}$), we have that the following inequalities simultaneously hold for the policies $\pi_1, ..., \pi_E$ produced by Algorithm 1:

$$\mathbb{E}_{\xi\sim\mathcal{D}}\ell_{\pi_k}(\xi; \pi_k, \pi_\star) \leqslant \beta_k(m, E, \delta), \quad k = 1, ..., E-1, \tag{E.12}$$

and furthermore,

$$\mathbb{E}_{\xi\sim\mathcal{D}}\ell_{\pi_E}(\xi; \pi_E, \pi_\star)$$
$$\leqslant \frac{1}{1-(1-\alpha)^E}\beta_{E-1}(m, E, \delta) + \frac{2\alpha}{1-(1-\alpha)^E}\Gamma(m, E, \delta)$$
$$+ 4L_\Delta\gamma^{1/a}T^{1-1/a_1}\left(\frac{2(1-\alpha)^E}{1-(1-\alpha)^E}\beta_{E-1}(m, E, \delta) + \frac{2\alpha}{1-(1-\alpha)^E}\Gamma(m, E, \delta)\right)^{b/a_1}$$
$$+ 4L_\Delta\gamma^{1/a}T^{1-1/a_1}\left(\frac{2(1-\alpha)^E}{1-(1-\alpha)^E}\beta_{E-1}(m, E, \delta) + \frac{2\alpha}{1-(1-\alpha)^E}\Gamma(m, E, \delta)\right)^{b/a}. \tag{E.13}$$

Proof. We proceed by induction on k.

Base case $k = 0$: We first argue the base case. As $\pi_\star \in \Pi$ and $(f_{cl}^{\pi_\star}, \mathsf{Id})$ is Ψ-incrementally-gain-stable by Assumption 4.2, we have that the optimization problem defining $\hat{\pi}_0$ is feasible. By Proposition E.1, there exists an event $\mathcal{E}_0$ such that $\mathbb{P}(\mathcal{E}_0) \geqslant 1 - \delta/E$ and on $\mathcal{E}_0$,

$$\mathbb{E}_{\xi\sim\mathcal{D}}\ell_{\pi_0}(\xi; \hat{\pi}_0, \pi_\star) \stackrel{(a)}{\leqslant} \frac{1}{m/E}\sum_{i=1}^{m/E}\ell_{\pi_0}(\xi_i^0; \hat{\pi}_0, \pi_\star) + \Gamma(m, E, 2\delta)$$

$$\stackrel{(b)}{\leqslant} \frac{1}{m/E}\sum_{i=1}^{m/E}\ell_{\pi_0}(\xi_i^0; \pi_\star, \pi_\star) + \Gamma(m, E, 2\delta)$$

$$\stackrel{(c)}{=} \Gamma(m, E, 2\delta),$$

where the first inequality (a) follows from Proposition E.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 (4.1), 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:

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

37

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

$$
\begin{aligned}
\mathbb{E}_{\xi \sim \mathcal{D}} \ell_{\pi_1}(\xi; \pi_1, \pi_\star) &= \mathbb{E}_{\xi \sim \mathcal{D}} \sum_{t=0}^{T-1} \|\Delta_{\pi_1, \pi_\star}(\varphi_t^{\pi_1}(\xi))\|_2 = \alpha \mathbb{E}_{\xi \sim \mathcal{D}} \sum_{t=0}^{T-1} \|\Delta_{\hat{\pi}_0, \pi_\star}(\varphi_t^{\pi_1}(\xi))\|_2 \\
&\leqslant \alpha \mathbb{E}_{\xi \sim \mathcal{D}} \sum_{t=0}^{T-1} \|\Delta_{\hat{\pi}_0, \pi_\star}(\varphi_t^{\pi_0}(\xi))\|_2 + \alpha L_\Delta \mathbb{E}_{\xi \sim \mathcal{D}} \sum_{t=0}^{T-1} \|\varphi_t^{\pi_1}(\xi) - \varphi_t^{\pi_0}(\xi)\|_2 \\
&= \alpha \mathbb{E}_{\xi \sim \mathcal{D}} \ell_{\pi_0}(\xi; \hat{\pi}_0, \pi_\star) + \alpha L_\Delta \mathbb{E}_{\xi \sim \mathcal{D}} \sum_{t=0}^{T-1} \|\varphi_t^{\pi_1}(\xi) - \varphi_t^{\pi_0}(\xi)\|_2.
\end{aligned}
$$

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 $(f_{\mathrm{cl}}^{\pi_1}, \mathsf{Id})$ is Ψ-incrementally-gain-stable as a result of constraint (4.1c), by Proposition C.4, we have that for all ξ:

$$
\begin{aligned}
\sum_{t=0}^{T-1} \|\varphi_t^{\pi_1}(\xi) &- \varphi_t^{\pi_0}(\xi)\|_2 \\
&\leqslant 4\gamma^{1/a} T^{1-1/a_1} \max\left\{ \left(\alpha \sum_{t=0}^{T-2} \|\Delta_{\hat{\pi}_0, \pi_\star}(\varphi_t^{\pi_0}(\xi))\|_2 \right)^{\frac{b}{a_1}}, \left(\alpha \sum_{t=0}^{T-2} \|\Delta_{\hat{\pi}_0, \pi_\star}(\varphi_t^{\pi_0}(\xi))\|_2 \right)^{\frac{b}{a}} \right\} \\
&\leqslant 4\gamma^{1/a} T^{1-1/a_1} \left[(\alpha \ell_{\pi_0}(\xi; \hat{\pi}_0, \pi_\star))^{\frac{b}{a_1}} + (\alpha \ell_{\pi_0}(\xi; \hat{\pi}_0, \pi_\star))^{\frac{b}{a}} \right].
\end{aligned}
$$

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

$$
\begin{aligned}
\mathbb{E}_{\xi \sim \mathcal{D}} \ell_{\pi_1}&(\xi; \pi_1, \pi_\star) \\
&\leqslant \alpha \Gamma(m, E, 2\delta) + 4\alpha L_\Delta \gamma^{1/a} T^{1-1/a_1} \left[(\alpha \Gamma(m, E, 2\delta))^{\frac{b}{a_1}} + (\alpha \Gamma(m, E, 2\delta))^{\frac{b}{a}} \right] \\
&\leqslant \alpha \Gamma(m, E, 2\delta) + 8\alpha L_\Delta \gamma^{1/a} T^{1-1/a_1} (\alpha \Gamma(m, E, 2\delta))^{\frac{b}{a_1}} \\
&=: \beta_1(n, E, \delta).
\end{aligned}
$$

The inequality above uses Jensen's inequality to move the expectation inside $x \mapsto x^{b/a_1}$ and $x \mapsto x^{b/a}$.

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 $i = 1, ..., m/E$, and $(f_{\mathrm{cl}}^{\pi_k}, \mathsf{Id})$ is Ψ-incrementally-gain-stable by constraint (4.1c). Observe that on $\mathcal{E}_{0:k-1}$, we have that

$$
\mathbb{E}_{\xi \sim \mathcal{D}} \ell_{\pi_k}(\xi; \pi_k, \pi_\star) \leqslant \beta_k(m, E, \delta). \tag{E.14}
$$

By Proposition E.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:

$$\max_{\pi_t \in \{\pi_k, \pi_\star\}} \sup_{\pi \in \Pi} \left| \mathbb{E}_{\xi \sim \mathcal{D}} \ell_{\pi_k}(\xi; \pi, \pi_t) - \frac{1}{m/E} \sum_{i=1}^{m/E} \ell_{\pi_k}(\xi_i^k; \pi, \pi_t) \right| \leqslant \Gamma(m, E, \delta). \tag{E.15}$$

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

$$\frac{1}{m/E} \sum_{i=1}^{m/E} \ell_{\pi_k}(\xi_i^k; \pi_k, \pi_\star) \overset{(a)}{\leqslant} \mathbb{E}_{\xi \sim \mathcal{D}} \ell_{\pi_k}(\xi; \pi_k, \pi_\star) + \Gamma(m, E, \delta)$$

$$\overset{(b)}{\leqslant} \beta_k(m, E, \delta) + \Gamma(m, E, \delta). \tag{E.16}$$

Above, (a) follows from (E.15) and (b) follows from (E.14).

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

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

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

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

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

$$\ell_{\pi_{k+1}}(\xi; \pi_{k+1}, \pi_\star) = \sum_{t=0}^{T-1} \|\Delta_{\pi_{k+1}, \pi_\star}(\varphi_t^{\pi_{k+1}}(\xi))\|_2$$

$$\leqslant \sum_{t=0}^{T-1} \|\Delta_{\pi_{k+1}, \pi_\star}(\varphi_t^{\pi_k}(\xi))\|_2 + L_\Delta \sum_{t=0}^{T-1} \|\varphi_t^{\pi_{k+1}}(\xi) - \varphi_t^{\pi_k}(\xi)\|_2$$

$$\leqslant (1 - \alpha)\ell_{\pi_k}(\xi; \pi_k, \pi_\star) + \alpha \ell_{\pi_k}(\xi; \hat{\pi}_k, \pi_\star) + L_\Delta \sum_{t=0}^{T-1} \|\varphi_t^{\pi_{k+1}}(\xi) - \varphi_t^{\pi_k}(\xi)\|_2.$$

Now we write:

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

and therefore given that $(f_{\mathrm{cl}}^{\pi_{k+1}}, \mathsf{Id})$ is Ψ-incrementally-gain-stable by constraint (4.1c), by Proposition C.4, we have that for all ξ:

$$\sum_{t=0}^{T-1} \|\varphi_t^{\pi_{k+1}}(\xi) - \varphi_t^{\pi_k}(\xi)\|_2 \leqslant 4\gamma^{1/a} T^{1-1/a_1} \max \left\{ (\alpha \ell_{\pi_k}(\xi; \hat{\pi}_k, \pi_k))^{\frac{b}{a_1}}, (\alpha \ell_{\pi_k}(\xi; \hat{\pi}_k, \pi_k))^{\frac{b}{a}} \right\}$$

$$\leqslant 4\gamma^{1/a} T^{1-1/a_1} \left[(\alpha \ell_{\pi_k}(\xi; \hat{\pi}_k, \pi_k))^{\frac{b}{a_1}} + (\alpha \ell_{\pi_k}(\xi; \hat{\pi}_k, \pi_k))^{\frac{b}{a}} \right].$$

Combining this inequality with the inequality above,

$$\ell_{\pi_{k+1}}(\xi; \pi_{k+1}, \pi_\star) \leqslant (1 - \alpha)\ell_{\pi_k}(\xi; \pi_k, \pi_\star) + \alpha\ell_{\pi_k}(\xi; \hat{\pi}_k, \pi_\star)$$
$$+ 4L_\Delta \gamma^{1/a} T^{1-1/a_1} \left[(\alpha\ell_{\pi_k}(\xi; \hat{\pi}_k, \pi_k))^{\frac{b}{a_1}} + (\alpha\ell_{\pi_k}(\xi; \hat{\pi}_k, \pi_k))^{\frac{b}{a}} \right].$$

Taking expectations of both sides, applying (E.14), and using Jensen's inequality, we obtain

$$\mathbb{E}_{\xi \sim \mathcal{D}}\ell_{\pi_{k+1}}(\xi; \pi_{k+1}, \pi_\star) \leqslant (1 - \alpha)\beta_k(m, E, \delta) + \alpha\mathbb{E}_{\xi \sim \mathcal{D}}\ell_{\pi_k}(\xi; \hat{\pi}_k, \pi_\star)$$
$$+ 4L_\Delta \gamma^{1/a} T^{1-1/a_1} \left[(\alpha\mathbb{E}_{\xi \sim \mathcal{D}}\ell_{\pi_k}(\xi; \hat{\pi}_k, \pi_k))^{\frac{b}{a_1}} + (\alpha\mathbb{E}_{\xi \sim \mathcal{D}}\ell_{\pi_k}(\xi; \hat{\pi}_k, \pi_k))^{\frac{b}{a}} \right].$$

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

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

where the first inequality (a) follows from (E.15), and the second inequality (b) from $\hat{\pi}_k$ being feasible to the constrained optimization problem (4.1).

Similarly, we have that

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

where (a) follows from (E.15), (b) from using π_k as a feasible point for optimization problem (4.1) and optimality of $\hat{\pi}_k$, (c) from another application (E.15), and (d) follows from (E.14).

Hence, we have:

$$\mathbb{E}_{\xi \sim \mathcal{D}}\ell_{\pi_{k+1}}(\xi; \pi_{k+1}, \pi_\star) \leqslant \beta_k(m, E, \delta) + 2\alpha\Gamma(m, E, \delta)$$
$$+ 4L_\Delta \gamma^{1/a} T^{1-1/a_1} \left[(\alpha c_k + \alpha\Gamma(m, E, \delta))^{\frac{b}{a_1}} + (\alpha c_k + \alpha\Gamma(m, E, \delta))^{\frac{b}{a}} \right]$$
$$\leqslant \beta_k(m, E, \delta) + 2\alpha\Gamma(m, E, \delta) + 8L_\Delta \gamma^{1/a} T^{1-1/a_1} (\alpha c_k + \alpha\Gamma(m, E, \delta))^{\frac{b}{a_1}}$$
$$=: \beta_{k+1}(m, E, \delta).$$

This finishes the inductive step. The recurrence relation for β_k is given for all $k = 1, ..., E - 2$ as:

$$\beta_{k+1}(m, E, \delta) = \beta_k(m, E, \delta) + 2\alpha\Gamma(m, E, \delta) + 8L_\Delta \gamma^{1/a} T^{1-1/a_1} (\alpha c_k + \alpha\Gamma(m, E, \delta))^{\frac{b}{a_1}}.$$

By unrolling the recurrence for $\beta_k(m, E, \delta)$, we conclude that for all $k = 1, ..., E - 1$,

$$\beta_k(m, E, \delta) = 2\alpha k \Gamma(m, E, \delta) + 8L_\Delta \gamma^{1/a} T^{1-1/a_1} \sum_{i=0}^{k-1} (\alpha c_i + \alpha \Gamma(m, E, \delta))^{\frac{b}{a_1}}.$$

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 $k = 1, ..., E - 1$,

$$\mathbb{E}_{\xi \sim \mathcal{D}} \ell_{\pi_k}(\xi; \pi_k, \pi_\star) \leqslant 2\alpha k \Gamma(m, E, \delta) + 8L_\Delta \gamma^{1/a} T^{1-1/a_1} \sum_{i=0}^{k-1} (\alpha c_i + \alpha \Gamma(m, E, \delta))^{\frac{b}{a_1}}$$

$$\leqslant 2\alpha k \Gamma(m, E, \delta) + 8L_\Delta k \gamma^{1/a} T^{1-1/a_1} \alpha^{b/a_1} \Gamma(m, E, \delta)^{b/a_1}$$

$$+ 8L_\Delta \gamma^{1/a} T^{1-1/a_1} \alpha^{b/a_1} \sum_{i=0}^{k-1} c_i^{b/a_1}.$$

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

$$\mathbb{E}_{\xi \sim \mathcal{D}} \ell_{\pi_{E-1}}(\xi; \pi_{E-1}, \pi_\star) \leqslant \beta_{E-1}(m, E, \delta). \tag{E.17}$$

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:

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

$$= \frac{(1-\alpha)^E}{\alpha} g(x)(\pi_\star(x) - \pi_{E-1}(x))$$

$$= \frac{(1-\alpha)^E}{\alpha} \Delta_{\pi_\star, \pi_{E-1}}(x).$$

Therefore:

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

which shows that $\tilde{\pi}_{E-1}$ satisfies constraint (4.1b) 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 (4.1c) since $(f_{\mathrm{cl}}^{\pi_{E-1}}, \mathsf{Id})$ is Ψ-incrementally-gain-stable by constraint (4.1c) from the previous iteration. This shows the optimization problem defining $\hat{\pi}_{E-1}$ is feasible.

Now as in the inductive step, by Proposition E.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:

$$\max_{\pi_t \in \{\pi_{E-1}, \pi_\star\}} \sup_{\pi \in \Pi} \left| \mathbb{E}_{\xi \sim \mathcal{D}} \ell_{\pi_{E-1}}(\xi; \pi, \pi_t) - \frac{1}{m/E} \sum_{i=1}^{m/E} \ell_{\pi_{E-1}}(\xi_i^{E-1}; \pi, \pi_t) \right| \leqslant \Gamma(m, E, \delta). \tag{E.18}$$

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

$$\frac{1}{m/E} \sum_{i=1}^{m/E} \ell_{\pi_{E-1}}(\xi_i^{E-1}; \pi_{E-1}, \pi_\star) \leqslant \beta_{E-1}(m, E, \delta) + \Gamma(m, E, \delta). \tag{E.19}$$

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

$$c_{E-1} \leqslant \frac{(1-\alpha)^E}{\alpha}(\beta_{E-1}(m, E, \delta) + \Gamma(m, E, \delta)). \tag{E.20}$$

We will use this bound in the sequel.

Now we write:

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

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

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

$$\ell_{\pi_E}(\xi; \pi_E, \pi_\star) = \sum_{t=0}^{T-1} \|\Delta_{\pi_E, \pi_\star}(\varphi_t^{\pi_E}(\xi))\|_2$$

$$\leqslant \sum_{t=0}^{T-1} \|\Delta_{\pi_E, \pi_\star}(\varphi_t^{\pi_{E-1}}(\xi))\|_2 + L_\Delta \sum_{t=0}^{T-1} \|\varphi_t^{\pi_E}(\xi) - \varphi_t^{\pi_{E-1}}(\xi)\|_2$$

$$\leqslant \frac{1-\alpha}{1-(1-\alpha)^E}\ell_{\pi_{E-1}}(\xi; \pi_{E-1}, \pi_\star) + \frac{\alpha}{1-(1-\alpha)^E}\ell_{\pi_{E-1}}(\xi; \hat{\pi}_{E-1}, \pi_\star)$$

$$+ L_\Delta \sum_{t=0}^{T-1} \|\varphi_t^{\pi_E}(\xi) - \varphi_t^{\pi_{E-1}}(\xi)\|_2. \tag{E.21}$$

Now we relate $f_{\mathrm{cl}}^{\pi_E}$ to $f_{\mathrm{cl}}^{\pi_{E-1}}$ in the following manner:

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

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

From constraint (4.1c), we have that $(f_{\text{cl}}^{\pi_E}, \text{Id})$ is Ψ-incrementally-gain-stable and therefore by Proposition C.4, we have for all ξ:

$$\sum_{t=0}^{T-1} \|\varphi_t^{\pi_E}(\xi) - \varphi_t^{\pi_{E-1}}(\xi)\|_2$$

$$\leqslant 4\gamma^{1/a} T^{1-1/a_1} \left(\frac{\alpha}{1-(1-\alpha)^E} \ell_{\pi_{E-1}}(\xi; \hat{\pi}_{E-1}, \pi_{E-1}) + \frac{(1-\alpha)^E}{1-(1-\alpha)^E} \ell_{\pi_{E-1}}(\xi; \pi_{E-1}, \pi_\star) \right)^{b/a_1}$$

$$+ 4\gamma^{1/a} T^{1-1/a_1} \left(\frac{\alpha}{1-(1-\alpha)^E} \ell_{\pi_{E-1}}(\xi; \hat{\pi}_{E-1}, \pi_{E-1}) + \frac{(1-\alpha)^E}{1-(1-\alpha)^E} \ell_{\pi_{E-1}}(\xi; \pi_{E-1}, \pi_\star) \right)^{b/a}.$$

Combining this inequality with (E.21), taking expectations and applying Jensen's inequality, we obtain:

$$\mathbb{E}_{\xi \sim \mathcal{D}} \ell_{\pi_E}(\xi; \pi_E, \pi_\star)$$

$$\leqslant \frac{1-\alpha}{1-(1-\alpha)^E} \mathbb{E}_{\xi \sim \mathcal{D}} \ell_{\pi_{E-1}}(\xi; \pi_{E-1}, \pi_\star) + \frac{\alpha}{1-(1-\alpha)^E} \mathbb{E}_{\xi \sim \mathcal{D}} \ell_{\pi_{E-1}}(\xi; \hat{\pi}_{E-1}, \pi_\star)$$

$$+ 4L_\Delta \gamma^{1/a} T^{1-1/a_1} \left(\frac{\alpha}{1-(1-\alpha)^E} \mathbb{E}_{\xi \sim \mathcal{D}} \ell_{\pi_{E-1}}(\xi; \hat{\pi}_{E-1}, \pi_{E-1}) + \frac{(1-\alpha)^E}{1-(1-\alpha)^E} \mathbb{E}_{\xi \sim \mathcal{D}} \ell_{\pi_{E-1}}(\xi; \pi_{E-1}, \pi_\star) \right)^{b/a_1}$$

$$+ 4L_\Delta \gamma^{1/a} T^{1-1/a_1} \left(\frac{\alpha}{1-(1-\alpha)^E} \mathbb{E}_{\xi \sim \mathcal{D}} \ell_{\pi_{E-1}}(\xi; \hat{\pi}_{E-1}, \pi_{E-1}) + \frac{(1-\alpha)^E}{1-(1-\alpha)^E} \mathbb{E}_{\xi \sim \mathcal{D}} \ell_{\pi_{E-1}}(\xi; \pi_{E-1}, \pi_\star) \right)^{b/a}.$$

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

$$\mathbb{E}_{\xi \sim \mathcal{D}} \ell_{\pi_{E-1}}(\xi; \hat{\pi}_{E-1}, \pi_{E-1}) \overset{(a)}{\leqslant} \frac{1}{m/E} \sum_{i=1}^{m/E} \ell_{\pi_{E-1}}(\xi_i^{E-1}; \hat{\pi}_{E-1}, \pi_{E-1}) + \Gamma(m, E, \delta)$$

$$\overset{(b)}{\leqslant} c_{E-1} + \Gamma(m, E, \delta)$$

$$\overset{(c)}{\leqslant} \frac{(1-\alpha)^E}{\alpha} (\beta_{E-1}(m, E, \delta) + \Gamma(m, E, \delta)) + \Gamma(m, E, \delta)$$

$$= \frac{(1-\alpha)^E}{\alpha} \beta_{E-1}(m, E, \delta) + \left(1 + \frac{(1-\alpha)^E}{\alpha} \right) \Gamma(m, E, \delta)$$

$$\leqslant \frac{(1-\alpha)^E}{\alpha} \beta_{E-1}(m, E, \delta) + 2\Gamma(m, E, \delta).$$

Here, (a) follows by (E.18), (b) follows from constraint (4.1b), and (c) follows from (E.20). This allows us to bound:

$$\frac{\alpha}{1-(1-\alpha)^E} \mathbb{E}_{\xi \sim \mathcal{D}} \ell_{\pi_{E-1}}(\xi; \hat{\pi}_{E-1}, \pi_{E-1}) + \frac{(1-\alpha)^E}{1-(1-\alpha)^E} \mathbb{E}_{\xi \sim \mathcal{D}} \ell_{\pi_{E-1}}(\xi; \pi_{E-1}, \pi_\star)$$

$$\leqslant \frac{2(1-\alpha)^E}{1-(1-\alpha)^E} \beta_{E-1}(m, E, \delta) + \frac{2\alpha}{1-(1-\alpha)^E} \Gamma(m, E, \delta).$$

Next, observe that:

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

$$= \left(1 - \frac{(1-\alpha)^E}{\alpha} \right) g(x)(\pi_{E-1}(x) - \pi_\star(x))$$

$$= \left(1 - \frac{(1-\alpha)^E}{\alpha} \right) \Delta_{\pi_{E-1}, \pi_\star}(x).$$

Therefore, for any π, ξ, we have:

$$\ell_\pi(\xi; \tilde{\pi}_{E-1}, \pi_\star) = \left(1 - \frac{(1-\alpha)^E}{\alpha} \right) \ell_\pi(\xi; \pi_{E-1}, \pi_\star). \tag{E.22}$$

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

$$\mathbb{E}_{\xi \sim \mathcal{D}} \ell_{\pi_{E-1}}(\xi; \hat{\pi}_{E-1}, \pi_\star) \overset{(a)}{\leqslant} \frac{1}{m/E} \sum_{i=1}^{m/E} \ell_{\pi_{E-1}}(\xi_i^{E-1}; \hat{\pi}_{E-1}, \pi_\star) + \Gamma(m, E, \delta)$$

$$\overset{(b)}{\leqslant} \frac{1}{m/E} \sum_{i=1}^{m/E} \ell_{\pi_{E-1}}(\xi_i^{E-1}; \tilde{\pi}_{E-1}, \pi_\star) + \Gamma(m, E, \delta)$$

$$\overset{(c)}{=} \left(1 - \frac{(1-\alpha)^E}{\alpha} \right) \frac{1}{m/E} \sum_{i=1}^{m/E} \ell_{\pi_{E-1}}(\xi_i^{E-1}; \pi_{E-1}, \pi_\star) + \Gamma(m, E, \delta)$$

$$\overset{(d)}{\leqslant} \left(1 - \frac{(1-\alpha)^E}{\alpha} \right) \left(\mathbb{E}_{\xi \sim \mathcal{D}} \ell_{\pi_{E-1}}(\xi; \pi_{E-1}, \pi_\star) + \Gamma(m, E, \delta) \right) + \Gamma(m, E, \delta)$$

$$\overset{(e)}{\leqslant} \beta_{E-1}(m, E, \delta) + 2\Gamma(m, E, \delta).$$

Here, (a) follows by (E.18), (b) follows from the optimality of $\hat{\pi}_{E-1}$ and feasibility of $\tilde{\pi}_{E-1}$, (c) follows from (E.22), (d) follows from another application of (E.18), and (e) follows from (E.17). This allows us to bound:

$$\frac{1-\alpha}{1-(1-\alpha)^E} \mathbb{E}_{\xi \sim \mathcal{D}} \ell_{\pi_{E-1}}(\xi; \pi_{E-1}, \pi_\star) + \frac{\alpha}{1-(1-\alpha)^E} \mathbb{E}_{\xi \sim \mathcal{D}} \ell_{\pi_{E-1}}(\xi; \hat{\pi}_{E-1}, \pi_\star)$$

$$\leqslant \frac{1}{1-(1-\alpha)^E} \beta_{E-1}(m, E, \delta) + \frac{2\alpha}{1-(1-\alpha)^E} \Gamma(m, E, \delta).$$

Together with the previous bound, this yields:

$$\mathbb{E}_{\xi\sim\mathcal{D}}\ell_{\pi_E}(\xi;\pi_E,\pi_\star)$$

$$\leqslant \frac{1}{1-(1-\alpha)^E}\beta_{E-1}(m,E,\delta) + \frac{2\alpha}{1-(1-\alpha)^E}\Gamma(m,E,\delta)$$

$$+ 4L_\Delta\gamma^{1/a}T^{1-1/a_1}\left(\frac{2(1-\alpha)^E}{1-(1-\alpha)^E}\beta_{E-1}(m,E,\delta) + \frac{2\alpha}{1-(1-\alpha)^E}\Gamma(m,E,\delta)\right)^{b/a_1}$$

$$+ 4L_\Delta\gamma^{1/a}T^{1-1/a_1}\left(\frac{2(1-\alpha)^E}{1-(1-\alpha)^E}\beta_{E-1}(m,E,\delta) + \frac{2\alpha}{1-(1-\alpha)^E}\Gamma(m,E,\delta)\right)^{b/a}$$

$$\leqslant \frac{1}{1-(1-\alpha)^E}\left[1 + 8L_\Delta\gamma^{1/a}T^{1-1/a_1}(1-\alpha)^E\right]\beta_{E-1}(m,E,\delta)$$

$$+ \frac{2\alpha}{1-(1-\alpha)^E}\left[1 + 4L_\Delta\gamma^{1/a}T^{1-1/a_1}\right]\Gamma(m,E,\delta)$$

$$+ 8L_\Delta\gamma^{1/a}T^{1-1/a_1}\left(\frac{2(1-\alpha)^E}{1-(1-\alpha)^E}\beta_{E-1}(m,E,\delta) + \frac{2\alpha}{1-(1-\alpha)^E}\Gamma(m,E,\delta)\right)^{b/a_1}.$$

$\square$

With Theorem E.5 in place, we now turn to the proof of our main results, which are immediate consequences of Theorem E.5. We first restate and prove Theorem 4.3.

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

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

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

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

Proof. Theorem E.5 states that if $\Gamma(m,1,\delta)\leqslant 1$, then:

$$\mathbb{E}_{\xi\sim\mathcal{D}}\ell_{\pi_1}(\xi;\pi_1,\pi_\star) \leqslant \Gamma(m,1,\delta) + 8L_\Delta\gamma^{1/a}T^{1-1/a_1}\Gamma(m,1,\delta)^{b/a_1}$$

$$\leqslant (1 + 8L_\Delta\gamma^{1/a}T^{1-1/a_1})\Gamma(m,1,\delta)^{b/a_1}.$$

To complete the proof we simply need to bound $\Gamma(m,1,\delta)$, which has the form:

$$\Gamma(m,1,\delta) = 2\mathcal{R}_m(\Pi) + B_\ell\sqrt{\frac{\log(4/\delta)}{m}}.$$

From Proposition E.2 and Proposition E.3, we have that:

$$B_\ell \leqslant 2\zeta^{1/a}B_0 L_\Delta T^{1-1/a_1},$$

$$\mathcal{R}_m(\Pi) \leqslant 65\zeta^{1/a}B_0 B_g B_\theta L_{\partial^2\pi}T^{1-1/a_1}\sqrt{\frac{q}{m}}.$$

This gives us a bound:

$$\Gamma(m,1,\delta) \lesssim \zeta^{1/a} B_0 B_g B_\theta L_{\partial^2\pi} T^{1-1/a_1} \left(\frac{q}{m}\right)^{1/2} + \zeta^{1/a} B_0 L_\Delta T^{1-1/a_1} \left(\frac{\log(1/\delta)}{m}\right)^{1/2}.$$

Plugging $\delta = e^{-q}$ yields:

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

The claim now follows. $\qquad\square$

We now restate and prove Theorem 4.4.

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

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

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

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

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

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

Proof. Theorem E.5 states that if $c_k + \Gamma(m,E,\delta) \leqslant 1$, then:

$$\mathbb{E}_{\xi\sim\mathcal{D}} \ell_{\pi_E}(\xi; \pi_E, \pi_\star)$$
$$\leqslant \frac{1}{1-(1-\alpha)^E} \left[1 + 8L_\Delta \gamma^{1/a} T^{1-1/a_1} (1-\alpha)^E\right] \beta_{E-1}(m,E,\delta)$$
$$+ \frac{2\alpha}{1-(1-\alpha)^E} \left[1 + 4L_\Delta \gamma^{1/a} T^{1-1/a_1}\right] \Gamma(m,E,\delta)$$
$$+ 8L_\Delta \gamma^{1/a} T^{1-1/a_1} \left(\frac{2(1-\alpha)^E}{1-(1-\alpha)^E} \beta_{E-1}(m,E,\delta) + \frac{2\alpha}{1-(1-\alpha)^E} \Gamma(m,E,\delta)\right)^{b/a_1}.$$

Notice that if $E \geqslant \frac{1}{\alpha} \log(1/\alpha)$, then $\frac{(1-\alpha)^E}{\alpha} \leqslant \exp(-\alpha E)/\alpha \leqslant 1$. Furthermore, if $\alpha \leqslant \min\left\{1/2, \frac{1}{L_\Delta \gamma^{1/a} T^{1-1/a_1}}\right\}$, then

$$(1-\alpha)^E \leqslant \min\left\{1/2, \frac{1}{L_\Delta \gamma^{1/a} T^{1-1/a_1}}\right\},$$

46

and hence the bound simplifies to:

$$\mathbb{E}_{\xi \sim \mathcal{D}} \ell_{\pi_E}(\xi; \pi_E, \pi_\star) \lesssim \beta_{E-1}(m, E, \delta) + \alpha\Gamma(m, E, \delta) + L_\Delta \gamma^{1/a} T^{1-1/a_1} \left(\alpha\beta_{E-1}(m, E, \delta) + \alpha\Gamma(m, E, \delta)\right)^{b/a_1}$$

$$\lesssim \beta_{E-1}(m, E, \delta) + \alpha\Gamma(m, E, \delta) + L_\Delta \gamma^{1/a} T^{1-1/a_1} \alpha^{b/a_1} \beta_{E-1}(m, E, \delta)^{b/a_1}$$
$$+ L_\Delta \gamma^{1/a} T^{1-1/a_1} \alpha^{b/a_1} \Gamma(m, E, \delta)^{b/a_1}$$
$$\lesssim \beta_{E-1}(m, E, \delta) + L_\Delta \gamma^{1/a} T^{1-1/a_1} \alpha^{b/a_1} \beta_{E-1}(m, E, \delta)^{b/a_1}$$
$$+ L_\Delta \gamma^{1/a} T^{1-1/a_1} \alpha^{b/a_1} \Gamma(m, E, \delta)^{b/a_1}.$$

We first bound $\Gamma(m, E, \delta)$, which has the form:

$$\Gamma(m, E, \delta) = 2\mathcal{R}_{m/E}(\Pi) + B_\ell \sqrt{\frac{E \log(4E/\delta)}{m}}.$$

From Proposition E.2 and Proposition E.3, we have that:

$$B_\ell \leqslant 2\zeta^{1/a} B_0 L_\Delta T^{1-1/a_1},$$

$$\mathcal{R}_{m/E}(\Pi) \leqslant 65\zeta^{1/a} B_0 B_g B_\theta L_{\partial^2 \pi} T^{1-1/a_1} \sqrt{\frac{Eq}{m}}.$$

This gives us a bound:

$$\Gamma(m, E, \delta) \lesssim \zeta^{1/a} B_0 T^{1-1/a_1} \max\{B_g B_\theta L_{\partial^2 \pi}, L_\Delta\} \sqrt{\frac{E(q \vee \log E)}{m}}.$$

Therefore, since we assumed $c_k \leqslant \Gamma(m, E, \delta)$, then $c_k + \Gamma(m, E, \delta) \leqslant 1$ holds when:

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

Now we proceed to bound $\beta_{E-1}(m, E, \delta)$:

$$\beta_{E-1}(m, E, \delta) = 2\alpha(E-1)\Gamma(m, E, \delta) + 8L_\Delta \gamma^{1/a} T^{1-1/a_1} \sum_{i=0}^{E-2} (\alpha c_i + \alpha\Gamma(m, E, \delta))^{b/a_1}$$

$$\lesssim \alpha E\Gamma(m, E, \delta) + L_\Delta \gamma^{1/a} T^{1-1/a_1} \alpha^{b/a_1} E\Gamma(m, E, \delta)^{b/a_1}$$
$$\lesssim L_\Delta \gamma^{1/a} T^{1-1/a_1} \alpha^{b/a_1} E\Gamma(m, E, \delta)^{b/a_1}.$$

Collecting the previous inequalities, we now bound:

$$\mathbb{E}_{\xi \sim \mathcal{D}} \ell_{\pi_E}(\xi; \pi_E, \pi_\star)$$

$$\lesssim L_\Delta \gamma^{1/a} T^{1-1/a_1} \alpha^{b/a_1} E\Gamma(m, E, \delta)^{b/a_1} + L_\Delta \gamma^{1/a} T^{1-1/a_1} \alpha^{b/a_1} \left[L_\Delta \gamma^{1/a} T^{1-1/a_1} \alpha^{b/a_1} E\Gamma(m, E, \delta)^{b/a_1}\right]^{b/a_1}$$

$$+ L_\Delta \gamma^{1/a} T^{1-1/a_1} \alpha^{b/a_1} \Gamma(m, E, \delta)^{b/a_1}$$

$$\lesssim L_\Delta \gamma^{1/a} T^{1-1/a_1} \alpha^{b/a_1} E\Gamma(m, E, \delta)^{b/a_1} + L_\Delta \gamma^{1/a} T^{1-1/a_1} \alpha^{b/a_1} \left[L_\Delta \gamma^{1/a} T^{1-1/a_1} \alpha^{b/a_1} E\Gamma(m, E, \delta)^{b/a_1}\right]^{b/a_1}$$

$$\lesssim \left[L_\Delta \gamma^{1/a} T^{1-1/a_1}\right]^{1-b/a_1} E\Gamma(m, E, \delta)^{b/a_1}$$

$$+ \left[L_\Delta \gamma^{1/a} T^{1-1/a_1}\right]^{1-b/a_1} \left[\left[L_\Delta \gamma^{1/a} T^{1-1/a_1}\right]^{1-b/a_1} E\Gamma(m, E, \delta)^{b/a_1}\right]^{b/a_1}$$

$$\asymp \left[L_\Delta \gamma^{1/a} T^{1-1/a_1}\right]^{1-b/a_1} E\Gamma(m, E, \delta)^{b/a_1} + \left[L_\Delta \gamma^{1/a} T^{1-1/a_1}\right]^{(1-b/a_1)(1+b/a_1)} \left[E\Gamma(m, E, \delta)^{b/a_1}\right]^{b/a_1}.$$

We bound $E\Gamma(m, E, \delta)^{b/a_1}$ as:

$$E\Gamma(m, E, \delta)^{b/a_1} \lesssim \zeta^{b/(a \cdot a_1)} B_0^{b/a_1} T^{(1-1/a_1)(b/a_1)} \max\{(B_g B_\theta L_{\partial^2 \pi})^{b/a_1}, L_\Delta^{b/a_1}\} E^{1+b/(2a_1)} \left(\frac{q \vee \log E}{m}\right)^{b/(2a_1)}.$$

Furthermore,

$$\left[E\Gamma(m, E, \delta)^{b/a_1}\right]^{b/a_1} \lesssim \zeta^{b^2/(a \cdot a_1^2)} B_0^{b^2/a_1^2} T^{(1-1/a_1)(b^2/a_1^2)}$$

$$\times \max\{(B_g B_\theta L_{\partial^2 \pi})^{b^2/a_1^2}, L_\Delta^{b^2/a_1^2}\} E^{(1+b/(2a_1))(b/a_1)} \left(\frac{q \vee \log E}{m}\right)^{b^2/(2a_1^2)}.$$

Therefore:

$$\left[L_\Delta \gamma^{1/a} T^{1-1/a_1}\right]^{1-b/a_1} E\Gamma(m, E, \delta)^{b/a_1}$$

$$\lesssim \zeta^{b/(a \cdot a_1)} \gamma^{(1-b/a_1)(1/a)} B_0^{b/a_1} T^{1-1/a_1} L_\Delta^{1-b/a_1} \max\{(B_g B_\theta L_{\partial^2 \pi})^{b/a_1}, L_\Delta^{b/a_1}\} E^{1+b/(2a_1)} \left(\frac{q \vee \log E}{m}\right)^{b/(2a_1)},$$

$$\left[L_\Delta \gamma^{1/a} T^{1-1/a_1}\right]^{1-b^2/a_1^2} \left[E\Gamma(m, E, \delta)^{b/a_1}\right]^{b/a_1}$$

$$\lesssim \zeta^{b/(a \cdot a_1)} \gamma^{(1-b^2/a_1^2)(1/a)} B_0^{b/a_1} T^{1-1/a_1} L_\Delta^{1-b^2/a_1^2}$$

$$\times \max\{(B_g B_\theta L_{\partial^2 \pi})^{b/a_1}, L_\Delta^{b/a_1}\} E^{(1+b/(2a_1))} \left(\frac{q \vee \log E}{m}\right)^{b^2/(2a_1^2)}.$$

Here the last inequality follows by nothing that $b/a_1 \leqslant 1$, that all constants are assumed to be no smaller than 1, and that by assumption $\frac{q \vee \log E}{m} \leqslant 1$. $\qquad \square$