Proceedings of Machine Learning Research vol 291:1–104, 2025 38th Annual Conference on Learning Theory

The Pitfalls of Imitation Learning when Actions are Continuous

Max Simchowitz MSIMCHOW@ANDREW.CMU.EDU
Carnegie Mellon University

Daniel Pfrommer DPFROM@MIT.EDU
Massachusetts Institute of Technology

Ali Jadbabaie JADBABAI@MIT.EDU
Massachusetts Institute of Technology

Editors: Nika Haghtalab and Ankur Moitra

Abstract

We study the problem of imitating an expert demonstrator in a discrete-time, continuous state-and-action space control system. We show that there exist stable dynamics (i.e. contracting exponentially quickly) and smooth, deterministic experts such that any smooth, deterministic imitator policy necessarily suffers error on execution that is exponentially larger, as a function of problem horizon, than the error under the distribution of expert training data. Our negative result applies to both behavior cloning and offline-RL algorithms, unless they produce highly *improper* imitator policies — those which are non-smooth, non-Markovian, or which exhibit highly state-dependent stochasticity — or unless the expert trajectory distribution is sufficiently spread. We provide preliminary evidence of the benefits of these more complex policy parameterizations, explicating the benefits of today's popular policy parameterizations in robot learning (e.g. action-chunking and diffusion-policies). We also establish a host of complementary negative and positive results for imitation in control systems.

Keywords: Imitation Learning, Control Theory, Stability

1. Introduction

Imitation Learning (IL), or learning a multi-step behavior from demonstration, encompasses both the earliest-introduced and most currently popular methodologies for training autonomous robotic systems with machine learning techniques (Ross et al., 2011; Ho and Ermon, 2016; Teng et al., 2023; Zhao et al., 2023). These successes have been buoyed a host of new innovations: the uses of generative models (e.g. Diffusion Policies (Chi et al., 2023)) to represent robotic behavior, the practice of "chunking" sequences of predicted actions, and various means of data augmentation beyond raw expert demonstrations (Ke et al., 2021; Jia et al., 2023). At the same time, with the rise of large language models (LLMs), IL also has become increasingly more prevalent in settings in which the agent predicts *discrete tokens*, such as steps on a chess board, lines on a math proof, or words in a sentence (Chen et al., 2021). For robot applications, in contrast, the state and action variables are continuous (but for convenience, time may still be treated discretely). We ask,

> *What are the fundamental differences between imitating continuous-actions and discrete behaviors? How do these differences explain the necessity of common techniques observed in today's robot learning pipelines?*

We consider control systems with continuous-valued states $\mathbf{x} \in \mathbb{X} = \mathbb{R}^d$, control inputs $\mathbf{u} \in \mathbb{U} = \mathbb{R}^m$ and dynamics $\mathbf{x}_{t+1} = f(\mathbf{x}_t, \mathbf{u}_t)$, where t denotes timestep. We assume f is **unknown** to the learner. The key parameter in our study is the task horizon, denoted by $H \in \mathbb{N}$, or number of steps of behavior to be imitated. The expert provides n length-H demonstration trajectories $(\mathbf{x}_1, \mathbf{u}_1, \dots, \mathbf{x}_H, \mathbf{u}_H)$, determined by the *expert policy* $\pi^\star : \mathbb{X} \to \mathbb{U}$ via $\mathbf{u}_t = \pi^\star(\mathbf{x}_t)$ with some initial state distribution of $\mathbf{x}_1$. The learner observes these trajectories and selects a policy $\hat{\pi} : \mathbb{X} \to \mathbb{U}$, deployed under the same dynamics, with the goal of emulating the expert: $\hat{\pi} \approx \pi^\star$. For clarity, we will use **imitation learning** (IL) to refer to learning from expert demonstration in which the agent cannot interact further with its environment or the expert after demonstrations are given. We will (colloquially) refer to as **behavior cloning** (BC) those methods which perform IL by fitting the data with pure supervised learning.

As learning is imperfect, the learner makes small errors which may add together over time, forcing the learner to stray off-course. Ultimately, the difference between the trajectories deployed by the learner and the expert trajectories may be much larger than the errors of learning the experts actions under the distribution of demonstration trajectories, typically by a multiplicative factor depending on H. This is the **compounding error problem.** While much attention has been devoted to circumventing compounding action via additional interaction with the expert (Ross and Bagnell, 2010) or with the environment (Ho and Ermon, 2016), we aim to understand when imitation learning is possible **without interactive access to either the environment or the expert**; what we deem the "non-interactive setting." Even without interaction, existing theoretical literature shows that compounding error is benign in **discrete** problem domains: it scales most polynomially in the problem horizon, H (Ross and Bagnell, 2010) and can even be eliminated entirely in some situations, via an appropriate loss function (e.g. the log-loss, (Foster et al., 2024)). However, these results are contingent on being able to estimate expert behavior in certain very strong error metrics (e.g. the $\{0, 1\}$-loss) which, while feasible for discrete problems, we show are **unattainable when actions are continuous.** Prior theoretical literature studying IL in continuous-action control systems has required additional assumptions and algorithmic modifications (e.g. expert-interactions, stabilization oracles and score-matching oracles). Hence, a systematic theoretical understanding of the difficulty of non-interactive, continuous-action IL remains absent. Due to space constraints, a complete discussion of **related work is deferred to Appendix B.**

Contributions. We show that imitation learning where both the expert $\pi^\star$ and learned policy $\hat{\pi}$ are "simple" suffers from **exponential-in-horizon compounding error, even in seemingly benign continuous-state-and-action control systems.** This contrasts discrete-token behavior cloning, in which compounding error is at most polynomial-in-horizon. We also provide evidence that exponential compounding can be mitigated by more sophisticated policy representations. While it has been popular to motivate more sophisticated policies (e.g. action-chunked Transformers and Diffusion policies) by the need to fit "multi-modal " expert data (Chi et al., 2023) (expert demonstrations with multiple *modes*, or strategies, to solve a given task), our findings suggests that **even the imitation of simple, deterministic, and hence uni-modal experts may benefit from complex policy parameterizations.** Our negative results depend only on the structure of $\hat{\pi}$, but are agnostic to the learning algorithm . In particular, our results apply even to offline reinforcement learning approaches (e.g. Kumar et al. (2020); Kostrikov et al. (2021)). More specifically, we show:

1. Theorems 1 and 2: There exists a smooth, deterministic expert policies and **stable** dynamical systems (in open- and closed-loop) such that, if the imitator policy is smooth and deterministic, or more generally, can be written as the sum of a smooth deterministic policy with state-independent noise (we call these "simply-stochastic), then the learner's execution error is exponentially-in-H larger than the training error under the expert distribution.

2. Theorem 3: Large compounding error occurs for more general stochastic policies, but potentially substantially less than for the "simply-stochastic" policies described above. Moreover, we show that host of more complex policy classes suffice to ameliorate compounding error for our lower bound and valid this finding with numerical simulations (Section 5 and Appendix D).

3. Theorem 4: Exponential compounding error is unconditionally unavoidable if system dynamics may be unstable (even if the dynamical system is smooth, Lipschitz and deterministic). Consequently, **observation of expert trajectories alone does not suffice for learning in these control systems, no matter what algorithm or policy class is used.**

4. Theorem 5: If expert data are sufficiently "spread" (i.e. anti-concentrated), even pure behavior cloning avoids compounding error. Hence, certain conditions on data quality suffice to avoid the pathologies above.

Proof Idea. The learner faces two candidate pairs of policies and dynamics, $(f_i, \pi_i^\star)$, $i \in \{1, 2\}$ (recall; dynamics are unknown). Each $\pi_i^\star$ stabilizes its corresponding f_i, but not the alternative system $f_j, j \neq i$. The learner is given insufficient data to determine the true index i. Were the goal to stabilize the unknown f_i, the zero policy $\hat{\pi}(\mathbf{x}) = \mathbf{0}$ would suffice because each f_i is stable. Yet we show that a smooth policy cannot both stabilize f_i for unknown $i \in \{1, 2\}$, and to simultaneously emulate expert policy under the expert's demonstration data distribution (for example, acting according to $\hat{\pi}(\mathbf{x}) = \mathbf{0}$) would cause large imitation error). Thus, for whatever the learner chooses, one of the systems is de-stabilized, and small errors compound exponentially into large ones. For those familiar with the theoretical RL literature, our result can be interpreted in terms of the Lipschitz constants certain classes of Q-functions (see Remark 3.1). For the control theorist, our argument is related to, but differs importantly from, celebrated gap-metric (Zames and El-Sakkary, 1981), as discussed in Remark C.1. We provide a more detailed proof sketch in Section 4.

Organization. Section 2 contains all preliminaries and notation. Section 3 provides formal statements of all main results. Section 4 provides the broad brushstrokes of the proof of our key result, the lower bounds against imitation in stable systems with "simple" experts (Theorems 1 and 2). Finally, Section 5 describes how our lower bound construction be circumvented by more complex policy parameterizations. The main paper concludes with a discussion in Section 6. For space, related work is deferred to Appendix B. The organization of the appendix is explained in Appendix A. One substantial part of the appendix is Part II, which reformulates our results (Appendix E) and provides more detailed theorem statements (Appendix F) in the language of minimax risks favored by the statistical learning community (Wainwright, 2019).

2. Preliminaries

Consider a control system with states $\mathbf{x} \in \mathbb{X} := \mathbb{R}^d$ and control actions $\mathbf{u} \in \mathbb{U} := \mathbb{R}^m$. Dynamics evolve deterministically, via dynamical maps $f : \mathbb{X} \times \mathbb{U} \to \mathbb{X}$, $\mathbf{x}_{t+1} = f(\mathbf{x}_t, \mathbf{u}_t)$, $t \geq 1$. Unless otherwise stated, we consider time-invariant, Markovian (static) policies that are mappings of states to distributions over actions $\pi : \mathbf{x} \to \mathbf{\Delta}(\mathbb{U})$. When π is deterministic, we simply write $\mathbf{u} = \pi(\mathbf{x})$. A triple (π, f, D) of policy π, dynamics f, and initial distribution $D \in \mathbf{\Delta}(\mathbb{X})$ over states $\mathbf{x}$, define a distribution $\mathbb{P}_{\pi,f,D}$ over trajectories where $\mathbf{x}_1 \sim D$, $\mathbf{u}_t \mid \mathbf{x}_t \sim \pi(\mathbf{x}_t)$, and $\mathbf{x}_{t+1} = f(\mathbf{x}_t, \mathbf{u}_t)$. An imitation learning problem is specificed by a tuple $(\pi^\star, f, D, H)$ with expert policy $\pi^\star$, dynamics f, and inital state distribution,and problem horizon $H \in \mathbb{N}$. Throughout, we take the expert $\pi^\star$ to be deterministic. The learner has access to a sample $\mathrm{S}_{n,H}$ consisting of n trajectories $\mathbf{traj}_{i,1:H} = (\mathbf{x}_{i,1:H}, \mathbf{u}_{i,1:H})$, $1 \le i \le n$, drawn i.i.d. from $\mathbb{P}_{\pi^\star,f,D}$. A (non-interactive) IL algorithm, denoted alg, is a possibly randomized mapping from $\mathrm{S}_{n,H}$ to the space of imitator policies $\hat{\pi}$. We denote these as $\mathrm{S}_{n,H} \sim [\pi^\star, f, D]$ and $\hat{\pi} \sim \mathrm{alg}(\mathrm{S}_{n,H})$, and let $\mathbb{E}_{[\mathrm{alg},\pi^\star,f,n,H]}$ denote expectation over both of these sources of randomness. Importantly, the dynamics f are **not known** to the learner. Given a $\mathrm{cost}(\cdot) : \mathbb{X}^H \times \mathbb{U}^H \to \mathbb{R}$, the **execution error** under $\mathrm{cost}(\cdot)$ is the difference

$$\mathbf{R}_{\mathrm{cost}}(\hat{\pi}; \pi^\star, f, D, H) := \mathbb{E}_{\hat{\pi},f,D}\left[\mathrm{cost}(\mathbf{x}_{1:H}, \mathbf{u}_{1:H})\right] - \mathbb{E}_{\pi^\star,f,D}\left[\mathrm{cost}(\mathbf{x}_{1:H}, \mathbf{u}_{1:H})\right]$$

We focus on the class of additive costs $\mathcal{C}_{\mathrm{lip}}$ comprised of $\mathrm{cost}(\mathbf{x}_{1:H}, \mathbf{u}_{1:H}) = \sum_{h=1}^H \tilde{\mathrm{cost}}(\mathbf{x}_h, \mathbf{u}_h)$, where $\tilde{\mathrm{cost}}(\cdot,\cdot)$ is 1-Lipschitz and bounded in $[0,1]$. Our lower bounds will show impossibility of imitating in $\mathbf{R}_{\mathrm{cost}}$ for a fixed cost. Upper bounds extend to a stronger metric, $\mathbf{R}_{\mathrm{traj},L_1}$, defined in Appendix I.1, that upper bounds $\sup_{\mathrm{cost}\in\mathcal{C}_{\mathrm{Lip}}} \mathbf{R}_{\mathrm{cost}}$.

Further Notation. Throughout, $\|\cdot\|$ denotes the Euclidean norm. A function $g : \mathbb{R}^{d_1} \to \mathbb{R}$ is *L-Lipschitz* if $|g(\mathbf{z}) - g(\mathbf{z}')| \le L\|\mathbf{z} - \mathbf{z}'\|$; g is *M-smooth* if it is twice-continuously differentiable and $\|\nabla^2 g\|_{\mathrm{op}} \le M$; $g : \mathbb{R}^{d_1} \to \mathbb{R}^{d_2}$ is L-Lipschitz (resp. M- smooth) if $\langle \mathbf{v}, g\rangle$ is L-Lipschitz (resp. M-smooth) for all unit vectors $\mathbf{v} \in \mathbb{R}^{d_2}$. The mean of stochastic policy π is the deterministic policy $\mathrm{mean}[\pi](\mathbf{x}) := \mathbb{E}_{\mathbf{u}\sim\pi(\mathbf{x})}[\mathbf{u}]$ We use $\mathbf{e}_i$ as shorthand for the i-th canonical basis vector, where dimension is clear from context. $\mathcal{B}_d(r)$ denotes the ball of radius r in $\mathbb{R}^d$, and $\mathcal{S}^{d-1}$ the sphere. C is a "universal constant" if it does not depend on any problem parameters; $a = \mathrm{O}\left(b\right)$ if $a \le Cb$ for a universal constant C, and $a = o_\star(b)$ means "$a \le c \cdot b$ for a *sufficiently small* universal constant c."

Compounding Error Compounding error is the phenomenon by which small errors in estimation of $\pi^\star$ during training compound, leading to deviations between $\pi^\star$ and $\hat{\pi}$ when deployed on horizon H. We measure this by comparing $\mathbf{R}_{\mathrm{cost}}$ to a natural measure of error under the distribution of expert data collected:

$$\mathbf{R}_{\mathrm{expert},L_p}(\hat{\pi}; \pi^\star, f, D, H) = \sum_{t=1}^H \mathbb{E}_{\pi^\star,f,D}\mathbb{E}_{\hat{\mathbf{u}}_t\sim\hat{\pi}(\mathbf{x}_t)}\left[\|\hat{\mathbf{u}}_t - \pi^\star(\mathbf{x}_t)\|^p\right]^{1/p}. \tag{2.1}$$

Note that, while $\pi^\star$ is deterministic, $\mathbf{R}_{\mathrm{expert},L_p}$ is well-defined even if $\hat{\pi}$ is stochastic. For reasons of technical convenience, we focus on $\mathbf{R}_{\mathrm{expert},L_2}$ (see Appendix I.5), but qualitatively similar results hold for other choices of p (e.g. $\mathbf{R}_{\mathrm{expert},L_1}$). Our paper argues that there exist natural, seemingly benign settings for continuous action IL where, for some choice of cost, imitating a simple expert with a simple policy renders $\mathbf{R}_{\mathrm{cost}}$ exponentially larger than $\mathbf{R}_{\mathrm{expert},L_2}$.

Control-Theoretic Stability. Adopting a control-theoretic perspective (e.g. Pfrommer et al. (2022)), our notion of "benign-ness" is defined in terms of **exponential incremental stability**. In general, stability is a control-theoretic property of a dynamical system that describe the sensitivity of the dynamics to perturbations of the state or input (c.f. Kirk (2004)). We focus on an incredibly strong form of stability that we call exponential incremental stability, which corresponds to a dynamic system in which the effects of perturbations on future dynamics diminish exponentially in time.

Definition 2.1 (Exponential Incremental Input-to-State Stability) *Let $C \ge 1$ and $\rho \in (0,1)$. We say $f : \mathbb{X} \times \mathbb{U} \to \mathbb{X}$ is (C,ρ)-exponentially incrementally input-to-state stable (E-IISS) if for any two states $\mathbf{x}_1, \mathbf{x}_1'$ and sequences of inputs $\{\mathbf{u}_k, \mathbf{u}_k'\}_{k=1}^t$, the resulting trajectories $\mathbf{x}_{t+1} = f(\mathbf{x}_t, \mathbf{u}_t)$*

and $\mathbf{x}'_{t+1} = f(\mathbf{x}'_t, \mathbf{u}'_t)$ *satisfy* $\|\mathbf{x}_{t+1} - \mathbf{x}'_{t+1}\| \le C\rho^t \|\mathbf{x}_1 - \mathbf{x}'_1\| + \sum_{1 \le k \le t} C\rho^{t-k}\|\mathbf{u}_k - \mathbf{u}'_k\|$. *Let* $\pi^\star : \mathbb{X} \to \mathbb{U}$ *be a deterministic policy. We say* $(\pi^\star, f)$ *are* (C, ρ)*-E-IISS if the "closed loop" system* $f^\pi(\mathbf{x}, \mathbf{u}) := f(\mathbf{x}, \pi(\mathbf{x}) + \mathbf{u})$ *is* (C, ρ)*-E-IISS.*

In all of our lower bound constructions, the origin will be a fixed point of both the open-loop and closed-loop system dynamics: $f^\pi(\mathbf{0}, \mathbf{0}) = f(\mathbf{0}, \mathbf{0}) = \mathbf{0}$. Thus, for these cases, Definition 2.1 stipulates that the dynamics exponentially contract towards the origin. E-IISS is essentially the strongest form of incremental stability, a term originally due to Agrachev et al. (2008). Yet, **despite its strength**, we demonstrate that exponential-in-horizon compounding error can occur even when the dynamics f and expert $(\pi^\star, f)$ satisfy E-IISS. We complement these results with upper bounds that hold under much weaker conditions.

2.1. The RL Perspective on Imitation Learning

Given that policies can be measured in terms of total-cost incurred, it has been popular adopt the formalism of reinforcement learning to study performance of IL methods. Focusing on additive costs $\mathrm{cost}(\mathbf{x}_{1:H}, \mathbf{u}_{1:H}) = \sum_{h=1}^H \mathrm{cost}_h(\mathbf{x}_h, \mathbf{u}_h)$, define the Q- and value functions $Q_{h;f,\hat{\pi},\mathrm{cost},H}(\mathbf{x}, \mathbf{u}) := \mathrm{cost}_h(\mathbf{x}, \mathbf{u}) + \sum_{h'>h}^H \mathbb{E}_{\hat{\pi},f}\left[\mathrm{cost}_{h'}(\mathbf{x}_{h'}, \mathbf{u}_{h'}) \mid (\mathbf{x}_h, \mathbf{u}_h) = (\mathbf{x}, \mathbf{u})\right]$. The Q-function formalism gives two natural conditions under which $\mathsf{R}_{\mathrm{eval}}$ can be controlled by training risk.

First, if $\mathbf{u} \mapsto Q_{h;f,\hat{\pi},\mathrm{cost},H}(\mathbf{x}, \mathbf{u})$ is L-Lipschitz for each h, then Lemma I.3 yields $\mathsf{R}_{\mathrm{cost}}(\hat{\pi}; \pi^\star, f, D, H) \le L \cdot \mathsf{R}_{\mathrm{expert},L_1}(\hat{\pi}; \pi^\star, f, D, H)$. Second, if each $Q_{h;f,\hat{\pi},\mathrm{cost},H}(\mathbf{x}, \mathbf{u}) \in [0, B]$, then Lemma I.4 ensures $\mathsf{R}_{\mathrm{cost}}(\hat{\pi}; \pi^\star, f, D, H) \le B \cdot \mathsf{R}_{\mathrm{expert},\{0,1\}}(\hat{\pi}; \pi^\star, f, D, H)$. where $\mathsf{R}_{\mathrm{expert},\{0,1\}}(\hat{\pi}; \pi^\star, f, D, H)$, defined as $\sum_{h=1}^H \mathbb{E}_{\pi^\star,f,D}\mathbb{E}_{\hat{\mathbf{u}} \sim \hat{\pi}(\mathbf{x}^\star_h)} I\left\{\mathbf{u}^\star_h \ne \hat{\mathbf{u}}_h\right\}$, is the $\{0,1\}$-loss analogue of $\mathsf{R}_{\mathrm{expert},L_p}$. Both statements are proved in Appendix I.2 via the celebrated performance difference lemma (Kakade, 2003). Thus, IL (even pure behavior cloning!) exhibits limited compounding error provided that either (a) relevant Q functions are Lipschitz, or (b) it is feasible to minimize the $\{0,1\}$ risk $\mathsf{R}_{\mathrm{expert},\{0,1\}}$.

2.2. The Control v.s. RL perspectives, and Limitations of the Latter

The control perspectives focuses on the properties of the dynamical map f, and the closed loop dynamics between f and the expert policy $\pi^\star$. The RL perspective places assumptions directly on the Q-functions; these depend implicity on the dynamics and choice of cost, and, when arguing via the performance difference lemma, on the learner policy $\hat{\pi}$. One connection between the two viewpoints is that, when the learned policy $\hat{\pi}$ is such that $(\hat{\pi}, f)$ E-IISS, the resulting Q functions are Lipschitz, an hence compounding error is avoided (see Appendix I.3 for proof):

Lemma 2.1 *Suppose that* $(f, \hat{\pi})$ *is* (C, ρ)*-E-IISS and* $\hat{\pi}$ *is* $L_{\hat{\pi}}$*-Lipschitz. Then, for any* $\mathrm{cost} \in \mathcal{C}_{\mathrm{lip}}$*,* $Q_{h;f,\hat{\pi},\mathrm{cost}}$ *is* $\frac{C}{1-\rho}(2+L_{\hat{\pi}})$*-Lipschitz. Moreover, for any* $D \in \mathbf{\Delta}(\mathbb{X})$ *and* $H \ge 1$*, and any* $\mathrm{cost} \in \mathcal{C}_{\mathrm{Lip}}$*,* $\mathsf{R}_{\mathrm{cost}}(\hat{\pi}; \pi^\star, f, D, H) \le \frac{C}{1-\rho}(2 + L_{\hat{\pi}}) \cdot \mathsf{R}_{\mathrm{expert},L_1}(\hat{\pi}; \pi^\star, f, D, H)$.

The above bound extends to $\mathsf{R}_{\mathrm{expert},L_2}$ via Hölder's inequality, and resembles classical equivalences between controllability to the origin and existence of Lyapunov functions (Sontag, 1983). Still, when the dynamics f are unknown, it is not clear how to ensure that that the closed-loop learned dynamics with $\hat{\pi}$ are E-IISS. Indeed:

Lemma 2.2 *There exists a pair of linear, deterministic time-invariant policies and dynamics* (f_i, π_i)*,* $i \in \{1, 2\}$*, such that* f_1*,* f_2*,* (π_1, f_1) *and* (π_2, f_2) *are all* (C, ρ)*-E-IISS for some* $C \geq 1$ *and* $\rho \in (0, 1)$*. However, neither* (π_1, f_2) *nor* (π_2, f_1) *are E-IISS for any choice of* $C' \geq 1, \rho' \in \{0, 1\}$*.*

The above follows from Proposition 4.1, and this insight lies at the heart of our lower bound. Lemma 2.2 cautions against placing overly optimistic assumptions on the class of learners' Q-functions, or claiming that, if such assumptions fail, the problem faced is unrealistically pathological. Instead, the control-theoretic lens suggests that there are seemingly benign problem regimes in which uniform Lipschitzness of Q functions is itself **too coarse** an assumption.

Remark 2.1 *Recall from Section 2.1 that imitating in the* $\{0, 1\}$ *loss yields at most* $\mathrm{poly}(H)$ *compounding error, a now-classical argument present, e.g., in the the seminal* DAGGER *paper Ross and Bagnell (2010). In Appendix I.4, we show that non-vacuous* $\{0, 1\}$ *imitation is impossible in continuous action spaces, exposing the limitations of this analysis in such settings.*

3. Main Results

Organization. This section presents our main results in their most concrete forms: Theorems 1 and 2 are lower bounds against "simple" policies (defined below), Theorem 3 is a lower bound for more general policies, and Theorem 4 is lower bounds arbitrary policies when dynamics are unstable. Finally, Theorem 5 shows that compounding error can be avoided when expert data provides sufficient coverage. Each theorem has a corresponding result, labeled as "Theorem #.A" given in Appendix F, which is more granular and formulated in the language of minimax risks better suited to the expert reader. These show that arbitrary families of L_2 regression problems can be embedded into imitation learning problems which witness the same degree of compounding error. All lower bounds instantiate a common proof schematic, given in Appendix G.

Recall from Lemma 2.1 that if $(\hat{\pi}, f)$ is E-IISS, compounding error is avoided. Hence, our negative results necessarily leverage that the learner has uncertainty over the true dynamics f, and thus cannot ensure $(\hat{\pi}, f)$ is incrementally stable. Indeed, if f is known and $(\pi^\star, f)$ is guaranteed to be stable, then the (possibly inefficient) algorithm which optimizes only over policies $\hat{\pi}$ for which $(\hat{\pi}, f)$ is stable avoids compounding error. To this end, we establish lower bounds against problem families defined as follows.

Definition 3.1 (Problem Class) *An* $(\mathbb{R}^d, \mathbb{R}^m)$*-**IL problem family*** $(\mathcal{P}, D)$ *is specified by state space* $\mathbb{X} = \mathbb{R}^d$*, input space* $\mathbb{U} = \mathbb{R}^m$*, and an **instance class*** $\mathcal{P} = \{(\pi^\star, f)\}$ *of pairs of candidate expert policies* $\pi^\star$ *and ground-truth dynamics* f*, as well as a distribution* D *over initial states.*

Given an instance class $\mathcal{P}$, we define its constituent policies $\Pi(\mathcal{P}) := \{\pi^\star : \exists f \text{ for which } (\pi^\star, f) \in \mathcal{P}\}$ and dynamics $\mathcal{F}(\mathcal{P}) := \{f : \exists \pi^\star \text{ for which } (\pi^\star, f) \in \mathcal{P}\}$. Our lower bound constructions are "regular," with experts and dynamics being deterministic, Lipschitz, and smooth.

Definition 3.2 (Regularity Conditions) *We say* $(\mathcal{P}, D)$ *is* (R, L, M)*-regular for all* $(\pi^\star, f) \in \mathcal{P}$*, if (a)* $\pi^\star$ *is deterministic, (b)* $\mathbf{x} \mapsto \pi^\star(\mathbf{x})$ *and* $(\mathbf{x}, \mathbf{u}) \mapsto f(\mathbf{x}, \mathbf{u})$ *is* L*-Lipschitz and* M*-smooth, and (c) with probability 1 under* $\mathbb{P}_{\pi^\star, f, D}$*, it holds that* $\max_t \max\{\|\mathbf{x}_t\|, \|\mathbf{u}_t\|\} \leq R$*. We say* $(\mathcal{P}, D)$ *is* (R, L, M) *is* $O(1)$*-regular if we can take* R, L, M *to be at most universal constants.*

3.1. "Simple" Policies and Algorithms

We define **simple IL policies** as a slight generalization of the smooth, deterministic expert policies considered above. **Simple algorithms** are those that return simple policies.

Definition 3.3 (Simple Policies and Algorithms) *A policy $\hat{\pi}$ is **simply-stochastic** if the distribution of deviations from the mean, $\hat{\pi}(\cdot \mid \mathbf{x}) - \mathrm{mean}[\hat{\pi}](\mathbf{x})$, does not depend on $\mathbf{x}$. We say $\hat{\pi}$ is (L, M)**-simple** if $\hat{\pi}$ is simply-stochastic, and $\mathrm{mean}[\hat{\pi}]$ is L-Lipschitz and M-smooth. An IL algorithm alg is (L, M)**-simple** if, for any sample $\mathrm{S}_{n,H}$, with probability one over $\hat{\pi} \sim \mathrm{alg}(\mathrm{S}_{n,H})$, $\hat{\pi}$ is (L, M)-simple. We let $\mathbb{A}_{\mathrm{simple}}(L, M)$ denote the class of (L, M)-simple IL algorithms, and denote by $\mathbb{A}_{\mathrm{simple}}(\mathrm{O}(1))$ a class $\mathbb{A}_{\mathrm{simple}}(L, M)$ for some sufficiently large $L, M = \mathrm{O}(1)$.*

The simply-stochastic requirement permits both deterministic policies, as well as popular Gaussian policies, where $\hat{\pi}(\mathbf{x}) = \mathcal{N}(\boldsymbol{\mu}(\mathbf{x}), \boldsymbol{\Sigma})$, where $\boldsymbol{\Sigma}$ is fixed for all $\mathbf{x}$. For the "regular" IL problem families above, restricting to simple policies subsumes the classical notion of proper learning.

Definition 3.4 (Proper Algorithms) *Given an instance class $\mathcal{P} = \{(\pi^\star, f)\}$, we say that alg is $\mathcal{P}$-proper if, for any sample $\mathrm{S}_{n,H}$, with probability one over $\hat{\pi} \sim \mathrm{alg}(\mathrm{S}_{n,H})$, it holds that $\hat{\pi} \in \Pi(\mathcal{P})$. We denote the set of $\mathcal{P}$-proper algorithms $\mathbb{A}_{\mathrm{proper}}(\mathcal{P})$.*

In particular, if $(\mathcal{P}, D)$ is $\mathrm{O}(1)$-regular, all deterministic expert policies are simple: $\mathbb{A}_{\mathrm{proper}}(\mathcal{P}) \subset \mathbb{A}_{\mathrm{simple}}(\mathrm{O}(1))$. Thus, lower bounds against simple algorithms imply lower bounds against proper algorithms as a special case. At the same time, they rule out the potential benefits of adding state-independent noise, or of inflating smoothness and Lipschitzness constraints by constant factors.

3.2. Simple Policies Fail to Imitate Simple Experts

Recall the notation $\mathbb{E}_{[\mathrm{alg},\pi^\star,f,n,H]}$ denoting expectation under a sample $\mathrm{S}_{n,H}$ drawn from $[\pi^\star, f, D]$, and policy $\hat{\pi} \sim \mathrm{alg}(\mathrm{S}_{n,H})$. Our main result states that, for any desired fractional rate of estimation, there exists regular problem families with stable open and closed loop dynamics for which execution error is exponentially larger than training error.

Theorem 1 *Fix an $k, s \in \mathbb{N}$ with $s \geq 2$ and define $\boldsymbol{\epsilon}_n = n^{-s/k}$, and let constants $C \geq 1$, $\rho \in (0, 1)$ be universal constants, and let C_1, C_2 be constants depending only on (k, s). There exists a $(\mathbb{R}^d, \mathbb{R}^d)$-IL problem family $(\mathcal{P}, D)$, with $d = k + 2$ and a $\mathrm{cost}(\cdot, \cdot) \in \mathcal{C}_{\mathrm{Lip}}$, such that that $(\mathcal{P}, D)$ is $O(1)$-regular, f and $(\pi^\star, f)$ are (C, ρ)-E-IISS for all $(\pi^\star, f) \in \mathcal{P}$, and for all $H \geq 2, n \geq 1$:*

(a) There exists a IL $\mathrm{alg} \in \mathbb{A}_{\mathrm{proper}}(\mathcal{P}) \subset \mathbb{A}_{\mathrm{simple}}(O(1))$ such that for all $(\pi^\star, f) \in \mathcal{P}$, it holds that $\mathbb{E}_{[\mathrm{alg},\pi^\star,f,n,H]}[\mathbf{R}_{\mathrm{expert},L_2}(\hat{\pi}; \pi^\star, f, D, H)] \leq C_1 \boldsymbol{\epsilon}_n$.

(b) Let $L, M \geq 1$. For any $\mathrm{alg} \in \mathbb{A}_{\mathrm{simple}}(L, M)$, there exists $(\pi^\star, f) \in \mathcal{P}$ for which it holds that $\mathbb{E}_{[\mathrm{alg},\pi^\star,f,n,H]}[\mathbf{R}_{\mathrm{cost}}(\hat{\pi}; \pi^\star, f, D, H)]$ is at least $C_2 \min\left\{1.05^H \boldsymbol{\epsilon}_n,\ 1/(ML^2)\right\}$.

In words, our bound states there are problem instances where it is possible to make an L_2 supervised learning loss small, but the error incurred on deployment is exponential in horizon, up to a threshold which shrinks gracefully as the smoothness and Lipschitz constants grow. By comparing $\mathbf{R}_{\mathrm{expert},L_2}$ and $\mathbf{R}_{\mathrm{eval}}$, we show that the imitation learning problem is challenging even if the underlying supervised learning problem is statistically tractable.

Crucially, our lower bound does not prescribe *how* the learner uses the expert demonstration data, only that the returned policy $\hat{\pi}$ is simple. In particular, the learner need not attempt to minimize Eq. (2.1), or conduct any form of behavior cloning. Indeed, our lower bound is entirely agnostic to the learning algorithm. Therefore, our lower bound applies to algorithms which attempt to imitate in some integral probability metric via inverse reinforcement learning (Ho and Ermon, 2016), provided that they do not interact further with the dynamics f. Moreover, because our bound holds for a fixed cost across all instances, it applies **even to cost/reward-aware algorithms, such as those based on offline reinforcement learning**.

The above result is strengthened as Theorem 1.A in Appendix F, where we further show that (a) all optimal algorithms for minimize $\mathbf{R}_{\mathrm{expert},L_2}$ are proper (and thus simple) algorithms; that is, non-simple policies confer *no advantage* on the expert data distribution; and (b) the dynamics $f \in \mathcal{F}(\mathcal{P})$ are one-step controllable with good condition number. Finally, our lower bound can be strengthened so that exponential compounding occurs on a constant-probability event.

Theorem 2 *In the setting of Theorem 1, and the same cost* $\mathrm{cost}(\cdot,\cdot) \in \mathcal{C}_{\mathrm{Lip}}$. *There exist constants* $C_3, C_4 > 0$ *depending only on* (k, s) *for which* $\mathbb{P}_{\pi^\star,f,D}[\mathrm{cost}(\mathbf{x}_{1:H}, \mathbf{u}_{1:H}) = 0] = 1$ *for all* $(\pi^\star, f) \in \mathcal{P}$, *but for any* $\mathrm{alg} \in \mathbb{A}_{\mathrm{simple}}(L, M)$, *there exists* $(\pi^\star, f) \in \mathcal{P}$ *such that*

$$\mathbb{E}_{[\mathrm{alg},\pi^\star,f,n,H]}\,\mathbb{P}_{\hat{\pi},f}\left[\mathrm{cost}(\mathbf{x}_{1:H}, \mathbf{u}_{1:H}) \geq C_3 \min\left\{1.05^H \boldsymbol{\epsilon}_n, L^{-2}M^{-1}\right\}\right] \geq C_4 \tag{3.1}$$

Theorems 1 and 2 are derived from Theorem 1.A, whose proof is given in Appendix L. These results rely on three properties of simple policies: smoothness, simple-stochasticity, and (tacitly) that $\hat{\pi} : \mathbb{X} \to \boldsymbol{\Delta}(\mathbb{X})$ is Markovian (static). In Section 5, we illustrate (and in Appendix O, we formally prove) that removing any of the three restrictions breaks our lower bound construction.

Remark 3.1 *From an RL-theoretic perspective, open loop stability ensures the existence of a nominal policy* $\pi_0(\mathbf{x}) \equiv \mathbf{0}$ *which ensures that the Lipschitz constant of all* Q *functions, induced by* π_0 *and the unknown dynamics* f, *is bounded. Thus, our lower bound ensures the mere existence of* some known policy *with bounded Lipschitz Q-functions is not sufficient to avoid exponential compounding error. Importantly, our result uses a class* $\mathcal{P}$ *which is* not *a product-set (i.e.* $\mathcal{P} \neq \Pi(\mathcal{P}) \times \mathcal{F}(\mathcal{P})$*); otherwise, either (a)* $\mathcal{P}$ *would contain some pair* $(\pi^\star, f)$ *which is either unstable in closed loop or (b) all pairs* $(\pi^\star, f)$ *would be stable, obviating a lower bound.*

3.3. Lower Bounds Against More Complex Policies

We now give two lower bounds for possibly non-simple policies. The first relaxes the simple-stochastic requirement, at the expense of a weaker result, and the second holds unconditionally, but considers unstable open-loop (as opposed to E-IISS) dynamics.

The next result requires two new objects. First, the class $\mathbb{A}_{\mathrm{gen,smooth}}(L, M, \alpha, p) \supset \mathbb{A}_{\mathrm{simple}}(L, M)$ of algorithms which return policies with L-Lipschitz, M-smooth means, and whose stochasticity satisfies a mild anti-concentration condition parameterized by $(\alpha, p) \in (0, 1]$. For suitable constants α, p bounded away from zero, this class includes all simply-stochastic, Gaussian, and most mixture-policies as special cases. The second is an L_2-variant of $\mathbf{R}_{\mathrm{cost}}$, denoted $\mathbf{R}_{\mathrm{cost},L_2}$. Formal definitions are deferred to Appendix F.2. Once supplied, the following theorem is entirely formal.

Theorem 3 (Lower Bound beyond Simple-Stochasticity) *Consider the setting and problem family* $(\mathcal{P}, D)$ *of Theorem 1, with integers* $k, s \geq 2$, *and* $\boldsymbol{\epsilon}_n := n^{-s/k}$. *Given parameters,* $\alpha, p \in (0, 1]$.

Consider now the algorithm class $\mathbb{A} = \mathbb{A}_{\text{gen,smooth}}(L, M, \alpha, p)$. Moreover, suppose that n is larger than a sufficiently large polynomial in $(LMk/\alpha p)^{k/s}$. Then, there exists a constant C depending only on (k, s) and universal $C' \geq 1$ such that for any alg $\in \mathbb{A}$ *and $n, H \geq 2$, $\mathbb{E}_{[\text{alg},\pi^\star,f,n,H]}[\mathbf{R}_{\text{cost},L_2}(n; \mathcal{P}, D, H)]$ is at least C times the minimum of $\epsilon_n \cdot 1.05^H$ and $\epsilon_n^{1-\frac{1}{C'(1+\log(1/(\alpha p)))}}$.*

Theorem 3 is a consequence of Theorem 3.A, proven in Appendix M. Because $\mathbb{A} = \mathbb{A}_{\text{gen,smooth}}(L, M, \alpha, p) \supset \mathbb{A}_{\text{simple}}(L, M)$, and Theorem 3 uses the same problem family as Theorem 1, ϵ_n upper bounds the best attainable expert-distribution error in Theorem 3 as well. Under $\mathbf{R}_{\text{cost},L_2}(n; \mathcal{P}, D, H)$, we suffer exponential compounding error to a threshold that is $\epsilon_n^{1-\Omega(1)}$. That is, the scaling of the error for large H has a strictly worse exponent than of ϵ_n. $\mathbf{R}_{\text{cost},L_2}$ is an L_2 analogue of $\mathbf{R}_{\text{cost}}$ which places greater emphasis on the upper tails of the cost. We suspect that, with a sharper analysis, we may be able to obtain the same bound on expected (i.e. L_1) cost, $\mathbf{R}_{\text{cost}}$.

Unlike Theorem 2, compounding error in Theorem 3 occurs on a very low-probability event and this is responsible for the at most $\epsilon_n^{1-\Omega(1)}$ rate of error. In Appendix O, we show that both the low-probability of compounding error and $\epsilon_n^{1-\Omega(1)}$ error rates are qualitatively unimprovable for the construction used in Theorem 1/Theorem 3. This is a reflection of the Benign Gambler's Ruin phenomenon described in Section 5.

Lastly, if the dynamics are stable in closed loop but possibly unstable in open-loop, exponential compounding occurs with zero restriction on the learned policies $\hat{\pi}$.

Theorem 4 (Unstable Dynamics) *Fix integers $1 \leq k$ and $s \geq 2$; set $\epsilon_n = n^{-s/k}$. For $d \geq k$, there exists an $O(1)$-regular IL class $\mathcal{P}$ such that each $(f, \pi^\star) \in \mathcal{P}$ is $(0, 1)$-E-IISS, constants C_1, C_2 depending only k, s, and a* cost $\in \mathcal{C}_{\text{Lip}}$ *such that for all $2 \leq H \leq \frac{1}{2}e^{d/8}$ and $n \geq 1$ There is an algorithm* alg $\in \mathbb{A}_{\text{proper}}(\mathcal{P})$ *such that, for all instances $(\pi^\star, f) \in \mathcal{P}$, $\mathbb{E}_{[\text{alg},\pi^\star,f,n,H]}[\mathbf{R}_{\text{expert},L_2}(\hat{\pi}; \pi^\star, f, D, H)] \leq C_1\epsilon_n$, but (b) for* any *IL algorithm* alg, *including those permitted to return policies $\hat{\pi}$ which are arbitrarily non-smooth, stochastic, and even history-dependent, there exist some $(\pi^\star, f) \in \mathcal{P}$ for which $\mathbb{E}_{[\text{alg},\pi^\star,f,n,H]}[\mathbf{R}_{\text{cost}}(\hat{\pi}; \pi^\star, f, D, H)] \geq C_2 \min\{2^H\epsilon_n, 1\}$ Moreover, a constant-probability variant analogous to that in Theorem 2 holds.*

Theorem 4 is derived from Theorem 4.A, whose proof is given in Appendix N. While exponential compounding is intuitive when dynamics are unstable, a rigorous and unconditional proof is subtle. For example, in a certain unstable *scalar* system, exponential compounding error can be mitigated by a number of strategies (Section 5). Indeed, our bound requires sufficiently large dimension $d = \Omega(\log H)$, and this is likely sharp if one combines the concentric stabilization strategy (Section 5) with a covering argument. Importantly, the dimension d in Theorem 4 can be made arbitrarily large, while the constants depend only on s and k, which can be taken to be fixed.

3.4. Simple Policies Avoid Compounding Error with Sufficient Coverage

Theorem 1 relies on the indistinguishability of different stabilizing system dynamics from the perspective of the learner. In Theorem 5 which follows, we show that this can be circumvented via E-IISS in addition to a strong data coverage requirement, which we term *well-spreadness* (Definition 3.5). Well-spread distributions can arise naturally, for instance via additive Gaussian exploration noise in the context of fully controllable systems. Our result, proved in Appendix P, can be interpreted as a polynomial upper bound for experts whose own trajectories induce sufficient exploration (see Remark P.1 for a more careful explanation).

Definition 3.5 *A distribution P over $\mathbb{R}^d$ is (L, ϵ, σ_0)-**well-spread** if P has a density $p(\cdot)$ with respect to the Lebesgue measure, and if there exists a convex, compact set $\mathcal{K} \subset \mathbb{R}^d$ such that (a) score function $\mathbf{x} \mapsto \log p(\mathbf{x})$ is is L-Lipschitz on $\mathcal{K}$, and (b) $\mathbb{P}_{\mathbf{x}\sim P}[\mathrm{dist}(\mathbf{x}, \mathcal{K}^c) \le \sigma_0] \le \epsilon$.*

Theorem 5 (Smooth Training Distribution) *Consider any (d, m)-BC instance $(\mathcal{P}, D)$. Provided for any $(\pi^\star, f) \in \mathcal{P}$, $h \in [H]$, the distribution of $\mathbf{x}_h^\star$ under $\mathbb{P}_{\pi^\star,f,D}$ is (L, ϵ, σ_0)-well-spread for $h > 1$ and $\pi^\star, \hat{\pi}$ are deterministic, β-second-order-smooth, L_π-Lipschitz, and B-bounded, and $\pi^\star$ is (C, ρ) incrementally input-to-state stabilizing (Definition 2.1), the following holds. Then, provided that $\mathbf{R}_{\mathrm{expert},L_2}(\hat{\pi}; \pi^\star, f, D, H) \le \min\{\rho_0, 1/L\}$, $\mathbf{R}_{\mathrm{eval}}(\hat{\pi}; \pi^\star, f, D, H)$ is upper bounded by $cHd\frac{C^2}{(1-\rho)^2}\left[\mathbf{R}_{\mathrm{expert},L_2}(\hat{\pi}; \pi^\star, f, D, H) + \sqrt{\epsilon}\right]$, where $c := d(8 + 16B^2 + 16\beta^2)$.*

4. Proof Overview

Here, we focus on providing the core intuitions behind proof of Theorems 1 and 2. The formal proof is given in Appendix L, which instantiates a general schematic in Appendix G.

The crux of the proof is to construct two pairs of policies and dynamical systems $(\pi_i, f_i)_{i\in\{1,2\}}$ which (a) open- and closed-loop stable, (b) π_i destablizes f_j for $i \ne j$, and (c) (π_i, f_i) look indistinguishable on the distribution of expert data. We accomplish this first by constructing a pair of 2×2 linear dynamical systems with these properties:

Definition 4.1 (Challenging Pair) *Fix a parameter $\mu \in (0, 1/2]$. Define $c_\mu = \frac{3}{2}\mu$. The* challenging pair *of instances $(\mathbf{A}_i, \mathbf{K}_i)_{i\in\{1,2\}}$ are the matrices in $\mathbb{R}^{2\times 2}$ given by $\mathbf{A}_1 = \begin{bmatrix} 1+\mu & c_\mu \\ -c_\mu & 1-2\mu \end{bmatrix}$, $\mathbf{A}_2 = \begin{bmatrix} -(1-\frac{1}{4}\mu) & c_\mu \\ 0 & 1-2\mu \end{bmatrix}$, $\mathbf{K}_1 = \begin{bmatrix} -(1+\mu) & -c_\mu \\ c_\mu & 0 \end{bmatrix}$, $\mathbf{K}_2 = \begin{bmatrix} (1-\frac{1}{4}\mu) & -c_\mu \\ 0 & 0 \end{bmatrix}$.*

Defining the linear dynamics $\bar{f}_i(\mathbf{x}, \mathbf{u}) = \mathbf{A}_i\mathbf{x} + \mathbf{u}_i$ and policies $\bar{\pi}_i(\mathbf{x}) = \mathbf{K}_i\mathbf{x}$, $(\bar{f}_i, \bar{\pi}_i)_{i\in\{1,2\}}$ satisfy points (a) and (b) above, and satify (c) for expert data on the $\mathrm{span}(\mathbf{e}_2)$ subspace of $\mathbb{R}^2$.

Definition 4.2 *A matrix $\mathbf{A} \in \mathbb{R}^{d\times d}$ is (C, ρ)-stable if $\|\mathbf{A}^s\|_{\mathrm{op}} \le C\rho^s\|\mathbf{v}\|$.*

Proposition 4.1 (The Challenging Pair induces exponential compounding error.) *Consider associated challenging pair as in Definition 4.1 with parameter $\mu \in (0, \frac{1}{2}]$. Also, set We $\mathbf{A}_{\mathrm{cl},i} := \mathbf{A}_i + \mathbf{K}_i$. Then*

(a) $\mathbf{A}_1, \mathbf{A}_2, \mathbf{A}_{\mathrm{cl},1}, \mathbf{A}_{\mathrm{cl},2}$ are all (C_μ, ρ_μ) stable for some $C_\mu > 0$, $\rho_\mu \in (0, 1)$ depending on on μ.

(b) Fix $\hat{\mathbf{K}} \in \mathbb{R}^{2\times 2}$ satisfying $\hat{\mathbf{K}}\mathbf{e}_2 = \mathbf{K}_1\mathbf{e}_2 (= \mathbf{K}_2\mathbf{e}_2)$. Then $\max_{i\in\{1,2\}} \|(\mathbf{A}_i + \mathbf{B}\hat{\mathbf{K}})^H\mathbf{e}_1\| \ge (1 + \frac{\mu}{4})^H$.

(c) The values of $\mathbf{A}_i\mathbf{e}_2, \mathbf{K}_i\mathbf{e}_2, \mathbf{A}_{\mathrm{cl},i}\mathbf{e}_2$ do not depend on $i \in \{1, 2\}$, and $V = \mathrm{span}(\mathbf{e}_2)$ is an invariant subspace of both $\mathbf{A}_{\mathrm{cl},1}$ and $\mathbf{A}_{\mathrm{cl},2}$. Hence, $(\bar{f}_i, \bar{\pi}_i)$ yield indistinguishable trajectories for any starting state $\mathbf{x}_1 \in \mathrm{span}(\mathbf{e}_2)$.

Proposition 4.1 is proven in Appendix C. Recall that we consider **noiseless** expert demonstrations. Thus, purely linear dynamics and policies do not suffice for a lower bound as such policies can be imitated *exactly* given a sufficient number of expert trajectories. Instead, we embed a *nonlinear* supervised learning problem into our linear construction. Our constructions are parametrized by the unknown index i of the challenging pair, and an unknown, nonlinear function $g^\star : \mathbb{R}^d \to \mathbb{R}$, belonging to a class $\mathcal{G}$ whose rate of estimation in L_2 under a suitable distribution D_{est} matches $\boldsymbol{\epsilon}_n$.

Our construction combines both $g^\star$ and A_i by carving the state space $\mathbb{X}$ into two distinct regions, $R_1, R_2 \subset \mathbb{X}$, where R_1 is a unit ball around 0 and R_2 is a unit ball around $3\mathbf{e}_1$. For $\mathbf{x} \in R_1$, we let the dynamics and policy be given by $f(\mathbf{x}, \mathbf{u}) = A_i\mathbf{x} + \mathbf{u}$, $\pi^\star(\mathbf{x}) = K_i\mathbf{x}$. Conversely for $\mathbf{x} \in R_2$ we use the aforementioned $g^\star \in \mathcal{G}$ to define $f(\mathbf{x}, \mathbf{u}) = g^\star(\mathbf{x})\mathbf{e}_1 - \mathbf{u}$, $\pi^\star(\mathbf{x}) = g^\star(\mathbf{x})\mathbf{e}_1$. For our initial state distribution, we consider a mixture which samples half of the initial states from D_{est} over R_2 and half drawn uniformly on the segment between $-\mathbf{e}_2$ and $+\mathbf{e}_2$.

This construction ensures that the learner errors on R_2 scale with $\boldsymbol{\epsilon}_n$ by the choice of $(\mathcal{G}, D_{\text{est}})$, meaning that $\hat{\pi}, \pi^\star$ must diverge at $t = 2$ when the initial condition is sampled from R_2, with the learner perturbed in the $\mathbf{e}_1$ direction. For the chosen initial state distribution, trajectories under $\pi^\star$ give no information regarding $\mathbf{K}_i\mathbf{e}_1$, as $\mathbf{x}_t = 0$ under $\pi^\star$ for $t \geq 2$ and hence the learner has no way of disambiguating between $\mathbf{K}_1, \mathbf{K}_2$. Furthermore, by our choice of distribution over R_1, any learner with smooth mean that matches the expert $\mathbf{K}_i$ on $\mathrm{span}(\{\mathbf{e}_2\})$ can be written $\mathrm{mean}[\hat{\pi}](\mathbf{x}) \approx \hat{\mathbf{K}}\mathbf{x}$, where $\hat{\mathbf{K}}$ satisfies the conditions of Proposition 4.1(b). In this case, by Proposition 4.1(b), the $\boldsymbol{\epsilon}_n$-magnitude errors in the $\mathbf{e}_1$ space are then magnified exponentially in the horizon H. Crucially, the argument requires the "simple stochastic" noise as more intelligent noise distributions can cancel the compounding error. Alternatively, $\hat{\pi}$ may deviate from the expert on the $\mathbf{e}_2$ subspace, as we do not restrict alg to behavior-cloning-like-algorithms. However, in which case, the learner must incurs at least $C_2/(ML^2)$ error from $\pi^\star$ in order to prevent the exponential instability.

4.1. Overview of Additional Proof Techniques.

Statistical Learning. The functions $g^\star$ defining the "R_2 policy" $\pi^\star(\mathbf{x}) = g^\star(\mathbf{x})\mathbf{e}_1$ must be chosen from a class that is (a) smooth (to preserve overall system smoothness), and (b) has non-trivial statistical error when learned from *noiseless training examples* $(\mathbf{x}_0, g^\star(\mathbf{x}_0))$. In particular, linear $g^\star$ does not suffice. A key subtley is that (c) we require the estimation error of $g^\star$ to be large with constant probability; otherwise, the large errors can only compound by a limited amount before saturating the bound on the cost magnitude. In Proposition E.1, we show that non-parametric function classes of $\{g\}$ satisfy requirements (a), (b), and (c). This requires operating in the "interpolation," or noise-free, setting of nonparameteric regression (Kohler and Krzyżak, 2013).

Bump functions. We use bump functions to stitch together the aforementioned R_1, R_2 regions in a smooth manner. Doing so requires care to ensure that the system remains globally stable, and we accomplish this by making the magnitude of the nonlinear terms sufficiently small, so that they are dominated by the stable linear dynamics.

Enforcing $\hat{\mathbf{K}}\mathbf{v} \approx \mathbf{K}_i\mathbf{v}$ for $\mathbf{v} \perp \mathbf{e}_1$. With some probability, the initial state distribution randomizes over $\mathbf{v} \sim \Delta\mathcal{B}_V$, the uniform distribution on the unit ball of radius Δ on the subspace $V = \mathrm{span}(\mathbf{e}_1)^\perp$. Thus for any policy with low imitation error, $\mathbb{E}_{\mathbf{v}\sim\Delta\mathcal{B}_V}\|\hat{\pi}(\mathbf{x}) - \mathbf{K}_i\mathbf{x}\|^2$ is small. By smoothness of $\hat{\pi}$, and by making Δ sufficiently small, classical arguments for zero-order gradient estimation ensure $\nabla\mathrm{mean}[\hat{\pi}(\mathbf{0})]\mathbf{P}_V \approx \mathbf{K}_i\mathbf{P}_V$, where $\mathbf{P}_V$ is the projection onto V (note: $\mathbf{K}_1\mathbf{P}_V = \mathbf{K}_2\mathbf{P}_V$). To ensure compounding error with constant probability, we leverage anti-concentration due to the Carbery-Wright (Carbery and Wright, 2001) and Paley-Zygmud inequalities; these use the convenient fact that the uniform distribution on the unit ball is log-concave.

Compounding error in nonlinear systems. The most significant technical obstacle is generalizing our compounding error argument from linear to non-linear systems. First, consider deterministic policies $\hat{\pi}$. Define the autonomous dynamical system $F_i(\mathbf{x}) = f_i(\mathbf{x}, \hat{\pi}(\mathbf{x})) = \mathbf{A}_i + \hat{\pi}(\mathbf{x})$, and $\hat{\mathbf{K}} := \nabla\hat{\pi}(\mathbf{0})$, we see that $\nabla F_i(\mathbf{0}) = \mathbf{A}_i + \hat{\mathbf{K}}$ must be unstable alone the $\mathbf{e}_1$ direction for one $i \in \{1, 2\}$, by Proposition 4.1. To show that the resulting *nonlinear* system is unstable, we adopt an argument used due to Jin et al. (2017) to bound the rate at which gradient-based optimizers escape strict saddle points. When policies are simply stochastic, their randomness can be coupled such that the joint distribution $(\hat{\mathbf{u}}, \hat{\mathbf{u}}') \sim (\pi(\mathbf{x}), \pi(\mathbf{x}'))$ ensures the differences $\hat{\mathbf{u}} - \hat{\mathbf{u}}' = \mathrm{mean}[\pi](\mathbf{x}) - \mathrm{mean}[\pi](\mathbf{x}')$ are deterministic. Beyond simply stochastic policies, as in the proof of Theorem 3, we use a considerably more subtle coupling to witnesses our stipulated anti-concentration condition, described in Appendix F.2.

5. Potential benefits of complex policy parameterizations.

Here, we provide an informal discussion of how non-simple policies can circumvent exponential compounding error. Each strategy can be applied to the construction underpinning our main theorem, Theorem 1, and show that the lower bounds based on that construction can be circumvented (see Appendix O). In particular, this shows that Theorem 3 is qualitatively unimprovable without appealing to a different construction. In Appendix D, we provide experimental evidence that more sophisticated policy parameterizations do indeed ameliorate compounding error.

For simplicity, consider a scalar analogue of the two-time step construction of Section 4 with dimension $d = m = 1$. We take $f_0(\mathbf{x}, \mathbf{u}) = \mathbf{u} - g^\star(\mathbf{x})$, but the dynamics at steps $t \geq 1$ are

$$\mathbf{x}_{t+1} = \xi \cdot \rho\mathbf{x}_{t+1} + \mathbf{u}_t, \quad \mathbf{x}_1 = \boldsymbol{\epsilon}, \tag{5.1}$$

where $\rho > 1$ is known , but $\xi \in \{-1, 1\}$ is an unknown sign. We begin in state $\mathbf{x}_1 = \boldsymbol{\epsilon}$, representing some initial learner error. The goal is to select a policy π which keeps $|\mathbf{x}_t|$ small, without prior knowledge of the sign ξ.

Any smooth, deterministic policy $\hat{\pi}$ suffers from exponential compounding error on this problem: approximating $\hat{\pi}(\mathbf{x}) \approx k\mathbf{x} + \mathbf{u}_0$, where $k \in \mathbb{R}$, around the origin $\mathbf{x} \approx \mathbf{0}$, we see that the dynamics compound with either $(\rho + k)^t$ or $(\rho - k)^t$, one of which must have an exponent of base > 1. This same pathology extends to simply stochastic policies by considering *differences* in trajectories, and coupling them so that their randomness cancels. We further recall from Theorem 4 that in $d = \Omega(\log H)$-dimensions, compounding error is unconditionally unavoidable. However, for the one-dimensional described here, removing the constraints of either Markovianity, simple-stochasticity, smoothness can evade this challenge.

Action-Switching. Consider a time-dependent strategy $\pi(\mathbf{x}, t)$, with $\pi(\mathbf{x}, t) = -\rho\mathbf{x}$ if t is odd and $\pi(\mathbf{x}, t) = \rho\mathbf{x}$ if t is even. By time-step $t = 3$, the system will have converged to state $\mathbf{x}_3 = 0$, and will remain at rest there. This strategy uses time-dependence to hedge over dynamical uncertainty; time-dependence can be replaced by stochasticity as shown below.

Benevolent Gambler's Ruin . Gambler's ruin is the classical paradox where a Gambler's wealth either doubles or is zeroed over rounds t; the expected wealth remains constant, but is non-zero with vanishing-in-t probability. While gambling ultimately ruins the gambler in finite time, a stochastic policy can enact the same strategy to its benefit. For $t \geq 1$, we consider the benign Gambler's Ruin policy $\pi_{\mathrm{BGR}}(\mathbf{x})$ which selects $\rho\mathbf{x}$ with probability $1/2$ and $-\rho\mathbf{x}$ with remaining probability. Crucially, such a policies randomization depends on the state, and therefore is **not simply stochastic**. This policy has an identically zero, and therefore smooth, mean $\mathrm{mean}[\pi_{\mathrm{BGR}}](\mathbf{x}) \equiv \mathbf{0}$.

Under this policy and dynamics Eq. (5.1), $\mathbb{P}[\mathbf{x}_{t+1} \neq 0] = 2^t \to 0$. With remaining probability $|\mathbf{x}_{t+1}| = (2\rho)^t \boldsymbol{\epsilon}$, so in expectation, $\mathbb{E}[|\mathbf{x}_{t+1}|] = \rho^t \boldsymbol{\epsilon}$. The clipped error is then $\mathbb{E}[\min\{1, |\mathbf{x}_{t+1}|\}] = 2^{-t} \min\{1, (2\rho)^t \boldsymbol{\epsilon}\}$. By balancing these two terms over t, this quantity is only ever at most $\boldsymbol{\epsilon}^p$, where $p = \log \rho / \rho(2\rho) \in (0, 1)$. In other words, the clipped expectation of state magnitude $\mathbb{E}[\min\{1, |\mathbf{x}_{t+1}|\}] \leq \min\{\boldsymbol{\epsilon}^p, (2\rho)^t \boldsymbol{\epsilon}\}$ grows at most *sublinearly in the initial error* $\boldsymbol{\epsilon}$.

Concentric Stabilization . Concentric stabilization swaps randomization/alternation for non-smoothness. For $j \in \mathbb{Z}$, define intervals $\mathcal{I}_j = ((2\rho)^{-2j}, (2\rho)^{-2(j-1)}]$. For any $\mathbf{x} \in \mathbb{R} \setminus \{0\}$, there exists a unique $j(\mathbf{x})$ such that $|\mathbf{x}| \in \mathcal{I}_{j(\mathbf{x})}$. We define the concentric stabilization policy, $\pi_{\mathrm{CS}}(\mathbf{x})$ which selects $\rho\mathbf{x}$ if $j(\mathbf{x})$ is even, and $-\rho\mathbf{x}$ if $j(\mathbf{x})$ is odd. This policy is deterministic, but highly non-smooth as $\mathbf{x} \to 0$.

Let $f^{\pi_{\mathrm{cs}}}(\mathbf{x}) = \xi\rho\mathbf{x} + \pi_{\mathrm{CS}}(\mathbf{x})$ denote the induced closed-loop dynamics. For any $\mathbf{x}_1$, and either choice of $\xi \in \{-1, 1\}$, consider the sequence induced by $\mathbf{x}_{t+1} = f^{\pi_{\mathrm{cs}}}(\mathbf{x}_t)$. One can compute that $\mathbf{x}_t = 0$ for $t > 3$, and that $\max\{|\mathbf{x}_1|, |\mathbf{x}_2|, |\mathbf{x}_3|\} \leq (2\rho)^2|\mathbf{x}_1|$. By leveraging non-smoothness, concentric stabilization limits the state's growth by at most a constant factor.

6. Discussion

We demonstrate that imitation learning in a continuous-action control system can exhibit exponential-in-horizon compounding error, even if the dynamics are stable in both open- and closed-loop. We provide preliminary evidence that more complex policy parameterizations may be able to avoid this pitfall, and that expert data with good coverage avoids compounding error even under unstable dynamics. There are many exciting questions for future work: (a) When precisely can complex policies mitigate compounding error? (b) How can the expert provide optimal agents from suboptimal states? (c) What is the sample complexity of offline RL, e.g. from *suboptimal data*, in control systems. A final pressing question is understanding the benefits and limitations of online environment interaction (e.g. RL finetuning) in continuous-action control.

Lastly, our work corroborates a provocative empirical finding from Block et al. (2023): what makes behavior cloning challenging is not instability in the dynamics themselves, but rather instabilities arising from the closed-loop feedback between dynamics and an imperfect imitation policy. As shown in Section 5, the design choices in the behavior cloning policy (Diffusion, data-augmentation (see replica-noising) lead to meaningful differences in performance; Block et al. (2023) finds similarly that the choice of *optimizer* can have similar effects on downstream performance as well. Thus, better understanding the interactions between the design space of algorithms, optimizers, and data is an important direction for future theoretical, empirical, and methodological work.

Acknowledgements

The authors are deeply grateful to a number who contributed to shaping this manuscript: John Miller for the incredibly invaluable conversations regarding framing and scoping the formalism; Nati Srebro for the insightful early discussions about the results; Dylan Foster, for the initial encouragement to pursue a more systematic treatment of control-theoretic decision making; and Reese Pathak, for helping us navigate the noise-free-regression literature.

DP and AJ acknowledge support from the Office of Naval Research under ONR grant N00014-23-1-2299. DP additionally acknowledges support from a MathWorks Research Fellowship.

References

Andrei A Agrachev, A Stephen Morse, Eduardo D Sontag, Héctor J Sussmann, Vadim I Utkin, and Eduardo D Sontag. Input to state stability: Basic concepts and results. *Nonlinear and optimal control theory: lectures given at the CIME summer school held in Cetraro, Italy June 19–29, 2004*, pages 163–220, 2008.

Mayank Bansal, Alex Krizhevsky, and Abhijit Ogale. Chauffeurnet: Learning to drive by imitating the best and synthesizing the worst. *arXiv preprint arXiv:1812.03079*, 2018.

Benedikt Bauer, Luc Devroye, Michael Kohler, Adam Krzyżak, and Harro Walk. Nonparametric estimation of a function from noiseless observations at random points. *Journal of Multivariate Analysis*, 160:93–104, 2017.

Kevin Black, Noah Brown, Danny Driess, Adnan Esmail, Michael Equi, Chelsea Finn, Niccolo Fusai, Lachy Groom, Karol Hausman, Brian Ichter, et al. pi_0: A vision-language-action flow model for general robot control. *arXiv preprint arXiv:2410.24164*, 2024.

Adam Block, Dylan J Foster, Akshay Krishnamurthy, Max Simchowitz, and Cyril Zhang. Butterfly effects of sgd noise: Error amplification in behavior cloning and autoregression. *arXiv preprint arXiv:2310.11428*, 2023.

Adam Block, Ali Jadbabaie, Daniel Pfrommer, Max Simchowitz, and Russ Tedrake. Provable guarantees for generative behavior cloning: Bridging low-level stability and high-level behavior. *Advances in Neural Information Processing Systems*, 2024.

Mariusz Bojarski, Davide Del Testa, Daniel Dworakowski, Bernhard Firner, Beat Flepp, Prasoon Goyal, Lawrence D Jackel, Mathew Monfort, Urs Muller, Jiakai Zhang, et al. End to end learning for self-driving cars. *arXiv preprint arXiv:1604.07316*, 2016.

Anthony Carbery and James Wright. Distributional and l^q norm inequalities for polynomials over convex bodies in $\mathbb{R}^n$. *Mathematical research letters*, 8(3):233–248, 2001.

Lili Chen, Kevin Lu, Aravind Rajeswaran, Kimin Lee, Aditya Grover, Misha Laskin, Pieter Abbeel, Aravind Srinivas, and Igor Mordatch. Decision transformer: Reinforcement learning via sequence modeling. *Advances in neural information processing systems*, 34:15084–15097, 2021.

Cheng Chi, Siyuan Feng, Yilun Du, Zhenjia Xu, Eric Cousineau, Benjamin Burchfiel, and Shuran Song. Diffusion policy: Visuomotor policy learning via action diffusion. *arXiv preprint arXiv:2303.04137*, 2023.

Ayoub El Hanchi, Chris Maddison, and Murat Erdogdu. Minimax linear regression under the quantile risk. In *The Thirty Seventh Annual Conference on Learning Theory*, pages 1516–1572. PMLR, 2024.

Chelsea Finn, Tianhe Yu, Tianhao Zhang, Pieter Abbeel, and Sergey Levine. One-shot visual imitation learning via meta-learning. In *Conference on robot learning*, pages 357–368. PMLR, 2017.

Abraham D Flaxman, Adam Tauman Kalai, and H Brendan McMahan. Online convex optimization in the bandit setting: gradient descent without a gradient. *arXiv preprint cs/0408007*, 2004.

Dylan J Foster, Adam Block, and Dipendra Misra. Is behavior cloning all you need? understanding horizon in imitation learning. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024.

László Györfi, Michael Kohler, Adam Krzyzak, and Harro Walk. *A distribution-free theory of nonparametric regression*. Springer Science & Business Media, 2006.

Philippe Hansen-Estruch, Ilya Kostrikov, Michael Janner, Jakub Grudzien Kuba, and Sergey Levine. Idql: Implicit q-learning as an actor-critic method with diffusion policies. *arXiv preprint arXiv:2304.10573*, 2023.

Jonathan Ho and Stefano Ermon. Generative adversarial imitation learning. *Advances in neural information processing systems*, 29, 2016.

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

Ahmed Hussein, Eyad Elyan, Mohamed Medhat Gaber, and Chrisina Jayne. Deep imitation learning for 3d navigation tasks. *Neural computing and applications*, 29:389–404, 2018.

Michael Janner, Yilun Du, Joshua B Tenenbaum, and Sergey Levine. Planning with diffusion for flexible behavior synthesis. *arXiv preprint arXiv:2205.09991*, 2022.

Mingxi Jia, Dian Wang, Guanang Su, David Klee, Xupeng Zhu, Robin Walters, and Robert Platt. Seil: simulation-augmented equivariant imitation learning. In *2023 IEEE International Conference on Robotics and Automation (ICRA)*, pages 1845–1851. IEEE, 2023.

Chi Jin, Rong Ge, Praneeth Netrapalli, Sham M Kakade, and Michael I Jordan. How to escape saddle points efficiently. In *International conference on machine learning*, pages 1724–1732. PMLR, 2017.

Sham Machandranath Kakade. *On the sample complexity of reinforcement learning*. University of London, University College London (United Kingdom), 2003.

Liyiming Ke, Jingqiang Wang, Tapomayukh Bhattacharjee, Byron Boots, and Siddhartha Srinivasa. Grasping with chopsticks: Combating covariate shift in model-free imitation learning for fine manipulation. In *2021 IEEE International Conference on Robotics and Automation (ICRA)*, pages 6185–6191. IEEE, 2021.

Michael Kelly, Chelsea Sidrane, Katherine Driggs-Campbell, and Mykel J Kochenderfer. Hg-dagger: Interactive imitation learning with human experts. In *2019 International Conference on Robotics and Automation (ICRA)*, pages 8077–8083. IEEE, 2019.

Donald E Kirk. *Optimal control theory: an introduction*. Courier Corporation, 2004.

Michael Kohler and Adam Krzyżak. Optimal global rates of convergence for interpolation problems with random design. *Statistics & Probability Letters*, 83(8):1871–1879, 2013.

Ilya Kostrikov, Ashvin Nair, and Sergey Levine. Offline reinforcement learning with implicit q-learning. *arXiv preprint arXiv:2110.06169*, 2021.

David Krieg, Erich Novak, and Mathias Sonnleitner. Recovery of sobolev functions restricted to iid sampling. *Mathematics of Computation*, 91(338):2715–2738, 2022.

Aviral Kumar, Aurick Zhou, George Tucker, and Sergey Levine. Conservative q-learning for offline reinforcement learning. *Advances in Neural Information Processing Systems*, 33:1179–1191, 2020.

Michael Laskey, Jonathan Lee, Roy Fox, Anca Dragan, and Ken Goldberg. Dart: Noise injection for robust imitation learning. In *Conference on robot learning*, pages 143–156. PMLR, 2017.

I Loshchilov. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017.

Ilya Loshchilov and Frank Hutter. Sgdr: Stochastic gradient descent with warm restarts. *arXiv preprint arXiv:1608.03983*, 2016.

Tianyi Ma, Kabir A Verchand, and Richard J Samworth. High-probability minimax lower bounds. *arXiv preprint arXiv:2406.13447*, 2024.

Tim Pearce, Tabish Rashid, Anssi Kanervisto, Dave Bignell, Mingfei Sun, Raluca Georgescu, Sergio Valcarcel Macua, Shan Zheng Tan, Ida Momennejad, Katja Hofmann, et al. Imitating human behaviour with diffusion models. *arXiv preprint arXiv:2301.10677*, 2023.

Ethan Perez, Florian Strub, Harm De Vries, Vincent Dumoulin, and Aaron Courville. Film: Visual reasoning with a general conditioning layer. In *Proceedings of the AAAI conference on artificial intelligence*, volume 32, 2018.

Daniel Pfrommer, Thomas Zhang, Stephen Tu, and Nikolai Matni. Tasil: Taylor series imitation learning. *Advances in Neural Information Processing Systems*, 35:20162–20174, 2022.

Stéphane Ross and Drew Bagnell. Efficient reductions for imitation learning. In *Proceedings of the thirteenth international conference on artificial intelligence and statistics*, pages 661–668. JMLR Workshop and Conference Proceedings, 2010.

Stéphane Ross, Geoffrey Gordon, and Drew Bagnell. A reduction of imitation learning and structured prediction to no-regret online learning. In *Proceedings of the fourteenth international conference on artificial intelligence and statistics*, pages 627–635. JMLR Workshop and Conference Proceedings, 2011.

Nur Muhammad Shafiullah, Zichen Cui, Ariuntuya Arty Altanzaya, and Lerrel Pinto. Behavior transformers: Cloning k modes with one stone. *Advances in neural information processing systems*, 35:22955–22968, 2022.

Eduardo D Sontag. A lyapunov-like characterization of asymptotic controllability. *SIAM journal on control and optimization*, 21(3):462–471, 1983.

Xiatao Sun, Shuo Yang, and Rahul Mangharam. Mega-dagger: Imitation learning with multiple imperfect experts. *arXiv preprint arXiv:2303.00638*, 2023.

Gokul Swamy, Sanjiban Choudhury, J Andrew Bagnell, and Steven Wu. Of moments and matching: A game-theoretic framework for closing the imitation gap. In *International Conference on Machine Learning*, pages 10022–10032. PMLR, 2021.

Siyu Teng, Xuemin Hu, Peng Deng, Bai Li, Yuchen Li, Yunfeng Ai, Dongsheng Yang, Lingxi Li, Zhe Xuanyuan, Fenghua Zhu, et al. Motion planning for autonomous driving: The state of the art and future perspectives. *IEEE Transactions on Intelligent Vehicles*, 8(6):3692–3711, 2023.

A.B. Tsybakov. *Introduction to Nonparametric Estimation*. Springer Series in Statistics. Springer, 2009. ISBN 9780387790510. URL https://books.google.com/books?id=JwzYNAEACAAJ.

Alexandre B Tsybakov. On nonparametric estimation of density level sets. *The Annals of Statistics*, 25(3):948–969, 1997.

Sara A van der Geer. *Empirical Processes in M-estimation*, volume 6. Cambridge university press, 2000.

Roman Vershynin. *High-dimensional probability: An introduction with applications in data science*, volume 47. Cambridge university press, 2018.

Martin J Wainwright. *High-dimensional statistics: A non-asymptotic viewpoint*, volume 48. Cambridge university press, 2019.

Runzhe Wu, Yiding Chen, Gokul Swamy, Kianté Brantley, and Wen Sun. Diffusing states and matching scores: A new framework for imitation learning. *arXiv preprint arXiv:2410.13855*, 2024.

G Zames and A El-Sakkary. Uncertainty in unstable systems: the gap metric. *IFAC Proceedings Volumes*, 14(2):149–152, 1981.

Tianhao Zhang, Zoe McCarthy, Owen Jow, Dennis Lee, Xi Chen, Ken Goldberg, and Pieter Abbeel. Deep imitation learning for complex manipulation tasks from virtual reality teleoperation. In *2018 IEEE International Conference on Robotics and Automation (ICRA)*, pages 5628–5635. IEEE, 2018.

Tony Z Zhao, Vikash Kumar, Sergey Levine, and Chelsea Finn. Learning fine-grained bimanual manipulation with low-cost hardware. *arXiv preprint arXiv:2304.13705*, 2023.

Brianna Zitkovich, Tianhe Yu, Sichun Xu, Peng Xu, Ted Xiao, Fei Xia, Jialin Wu, Paul Wohlhart, Stefan Welker, Ayzaan Wahid, Quan Vuong, Vincent Vanhoucke, Huong Tran, Radu Soricut, Anikait Singh, Jaspiar Singh, Pierre Sermanet, Pannag R. Sanketi, Grecia Salazar, Michael S. Ryoo, Krista Reymann, Kanishka Rao, Karl Pertsch, Igor Mordatch, Henryk Michalewski, Yao Lu, Sergey Levine, Lisa Lee, Tsang-Wei Edward Lee, Isabel Leal, Yuheng Kuang, Dmitry Kalashnikov, Ryan Julian, Nikhil J. Joshi, Alex Irpan, Brian Ichter, Jasmine Hsu, Alexander Herzog, Karol Hausman, Keerthana Gopalakrishnan, Chuyuan Fu, Pete Florence, Chelsea Finn, Kumar Avinava Dubey, Danny Driess, Tianli Ding, Krzysztof Marcin Choromanski, Xi Chen, Yevgen Chebotar, Justice Carbajal, Noah Brown, Anthony Brohan, Montserrat Gonzalez Arenas, and Kehang Han. Rt-2: Vision-language-action models transfer web knowledge to robotic control. In Jie Tan, Marc Toussaint, and Kourosh Darvish, editors, *Proceedings of The 7th Conference on Robot Learning*, volume 229 of *Proceedings of Machine Learning Research*, pages 2165–2183. PMLR, 06–09 Nov 2023.

Contents

Appendix A. Organization of the Appendix

The Appendix is divided into three parts. Part I contains discussions deferred from the maint text: related work (Appendix B), a proof of Proposition 4.1 (Appendix C), and experimental details (Appendix D).

In Part II, we present general formulations of our results in the language of minimax theory from statistical learning (e.g. Wainwright (2019)), which may be preferrable to experts in the theoretical reinforcement learning and learning theory communities. After describing these minimax notions in detail (Appendix J), Appendix F states results that generalize those in Section 3, revealing that compounding error can be realized by embedding *near-arbitrary* L_2-regression problems into a specific family of control systems. Finally, we give a detailed statement of our overal proof schematic in Appendix G, which is instantiated in the proof of all negative results. The discussion at the beginning of Part II provides a more detailed overview of the contents of these sections.

Finally, Part III provides the proofs of all results in both the main text and Part II, including all lower and upper bounds. This includes Appendix I, which contains the proofs for the preliminaries in Section 2, such as the equivalence between Q-function-Lipschitzness and incremental-input-to-state-stability. The lower-bound proofs can be found in Appendix L, Appendix M, and Appendix N.

- Appendix L contains the proofs for Theorem 1.A (which generalizes Theorem 1, Theorem 2).
- Appendix M proves the non-simply-stochastic variant, Theorem 3.A (from which Theorem 3 follows).
- Appendix N presents Theorem 4, which demonstrates exponentially poor generalization for all policies (including non-simply stochastic) for an open-loop unstable system.

We demonstrate ways to circumvent these constructions in Appendix O, via chunked-policies and time-varying policies, which deviate from the smooth and/or simple policies on which our lower bounds are predicated.

Lastly, the proof of the upper bound (Theorem 5) can be found in Part III as part of Appendix P. Additionally, we demonstrate that our lower-bound construction can be circumvented in **??** and analyze both chunked-policies and time-varying-policies, which deviate from the smooth and/or simple policies. We direct the reader to the discussion at the start of Part III for a directory of the various sections contained therein.

Part I
Deferred Discussion

Appendix B. Related Work

Imitation from expert demonstration has emerged as a pre-eminent technique for learning in robotic control tasks; applications have included self-driving vehicles (Hussein et al., 2017; Bojarski et al., 2016; Bansal et al., 2018), visuomotor policies (Finn et al., 2017; Zhang et al., 2018), and navigation tasks (Hussein et al., 2018), and large-scale robotic decision making models (Zitkovich et al., 2023; Black et al., 2024). These advances have been accelerated by the introduction of generative neural network architectures parameterizing the robotic policy, including diffusion and flow-based models

(Janner et al., 2022; Chi et al., 2023; Pearce et al., 2023; Hansen-Estruch et al., 2023; Black et al., 2024), and Transformer architectures with appropriate tokenization of actions (Zhao et al., 2023; Chen et al., 2021; Shafiullah et al., 2022). The common rationale for these models is that they may represent a rich and varied distribution of expert strategies, or *modes*, for solving a given task (Chi et al., 2023; Shafiullah et al., 2022). Our contributions suggests that these models may enjoy benefits even for deterministic and smooth (i.e., unimodal!) expert policies.

The compounding error problem — that is, the possibility that execution error can be significantly larger than error on the training data distribution — has been widely acknowledge in imitation learning (Ross and Bagnell, 2010; Ross et al., 2011). The seminal work of (Ross and Bagnell, 2010) proposes the DAGGER algorithm for *interactive data* collection to circumvent this challenge, which has seen widespread adoption (Sun et al., 2023; Kelly et al., 2019). Other approaches have focused on modifying the distribution of data collected by the expert to provide sufficient coverage of failure models (Laskey et al., 2017; Ke et al., 2021).

On the theoretical side, however, the challenge of compounding error appears more benign: for example, Ross and Bagnell (2010) show that without interventions, the discrepancy between training and execution error is at most polynomial in the horizon. Further, recent work by Foster et al. (2024) demonstrates that, by minimzing the log-loss (as is common in discrete imitation learning applications, such as text), horizon length may have no adverse affect on performance of imitation learning. However, both of these works operate in settings that are not well-suited for control settings: Ross and Bagnell (2010) and Foster et al. (2024) assume the ability to learn the expert policy in the total-variation and Hellinger distances, respectively, which is not feasible for deterministic policies in continuous action spaces (see Proposition I.1). Though these purely probabilistic distances can be relaxed to integral probability metrics (IPMs) induced by relevant classes of Q-functions (see e.g. Swamy et al. (2021) or the discussion in Section 2.1), we explain how these metrics may be too stringent in the worst case as well (Section 2.2).

Recent work has established mathematical guarantees for imitation specifically for control-theoretic settings. Unfortunately, these required either interactive access to the expert demonstrator, or rather a rather complex procedure involve generative imitators Pfrommer et al. (2022), multiple steps of environment interaction, (Wu et al., 2024) or a complex recipe of hierarchical trajectory stabilization, and targeted data augmentation (Block et al., 2024). Hence, the theoretical understanding of **non-interactive imitation learning in control systems** has remained entirely open.

(Pfrommer et al., 2022; Block et al., 2024) propose incremental input-to-state stability (Agrachev et al., 2008) as the natural regularity condition governing the possibility of imitation in these settings. Section 2.2 connects this notion to formalisms more commonly studied in the theoretical reinforcement learning literature, arguing how traditional assumptions in the latter community may be insufficiently delicate for control-theoretic settings. Our negative results draw connections to yet-more-classical principle in control theory, namely the gap metric due to Zames and El-Sakkary (1981) (Remark C.1). Finally, our lower bounds also involve a range of other technical tools, including log-concave anti-concentration (Carbery and Wright, 2001), nonparametric regression in the zero-noise (interpolation) setting (Kohler and Krzyżak, 2013; Bauer et al., 2017; Krieg et al., 2022), and quantitative variants of the unstable manifold theorem applied in the study of saddle-point escape in non-convex optimization (Jin et al., 2017).

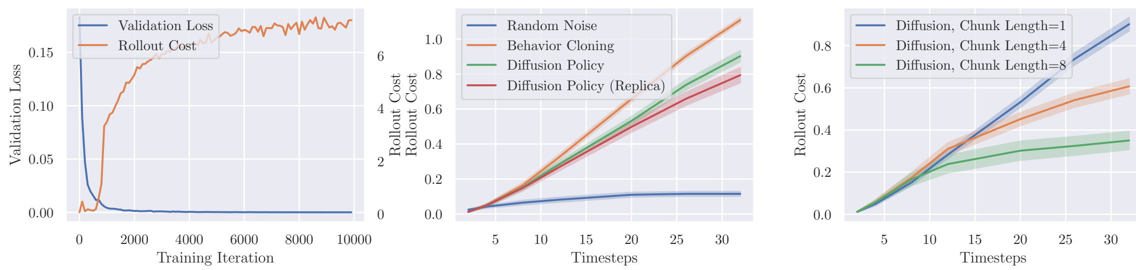

Figure 1: We benchmark the performance of different methods on Construction L.1. See Appendix D for details. **Left**: Validation loss and Rollout Cost ($\max_t \langle e_1, \mathbf{x}_t \rangle$) of Behavior Cloning using $H = 32$. Center: Performance of Behavior Cloning, Diffusion Policy (Chi et al., 2023), replica noising (Block et al., 2024), and random noise $\mathbf{u}_h \sim \mathcal{N}(\mathbf{0}, \frac{1}{6}\mathbf{I})$. Right: Diffusion Policy with action-chunking.

Appendix C. Proof of Proposition 4.1

Remark C.1 (Connection to the gap metric) *The gap-metric (Zames and El-Sakkary, 1981) in control theory allows one to measure the extent to which two different dynamical systems can be stabilized by the same control law. In our case, both transition matrices $\mathbf{A}_i$ stable in the classical sense (see also Definition 4.2 below), and thus, as noted above, are simultaneously stabilized by the indentically-zero control law. However, neither system can be stabilized by any linear feedback which coincides with the $\mathbf{K}_i$'s on the subspace $V = \mathrm{span}(\mathbf{e}_2)$.*

Proof We prove the stability and instability, properties (a) and (b). Property (c) follows directly from observation.

Proof of (a). A standard fact is that $\mathbf{A}$ is (C, ρ) stable for some $C > 0$ and $\rho < 1$ if and only if $\rho(\mathbf{A}) < 1$, where $\rho(\mathbf{A})$ denotes the spectral radius, or largest-magnitude eigenvalue, of $\mathbf{A}$. The eigenvalues of $\mathbf{A}_1, \mathbf{A}_2, \mathbf{A}_{\mathrm{cl},i}, i \in \{1, 2\}$ are $\{1 - \frac{\mu}{2}\}$, $\{1 - 2\mu, \frac{1}{4}\mu - 1\}$, and $\{0, 1 - 2\mu\}$ respectively, and are strictly less than one. The eigenvalues for $\mathbf{A}_2, \mathbf{A}_{\mathrm{cl},1}, \mathbf{A}_{\mathrm{cl},2}$ can be read directly off their upper triangular form. For $\mathbf{A}_1$, we observe that its spectrum are the roots of the characteristic polynomial $(1 + \mu - \lambda)(1 - 2\mu - \lambda) + \frac{9}{4}\mu^2 = (\lambda - (1 - \frac{\mu}{2}))^2$, both of which are $1 - \mu/2 \in (0, 1)$.

Proof of (b). Any $\hat{\mathbf{K}}$ satisfying the stipulated constraint is of the form $\hat{\mathbf{K}} = \begin{bmatrix} a & -c_\mu \\ b & 0 \end{bmatrix}$. Then,

$$\mathbf{A}_1 + \mathbf{B}\hat{\mathbf{K}} = \begin{bmatrix} 1 + \mu + a & 0 \\ b - c_\mu & 1 - 2\mu \end{bmatrix}, \quad \mathbf{A}_2 + \mathbf{B}\hat{\mathbf{K}} = \begin{bmatrix} -(1 - \frac{1}{4}\mu) + a & 0 \\ b & 1 - 2\mu \end{bmatrix}$$

Using the lower triangular structure of the above matrices, we can check $\mathrm{err}(H)$ is at least the maximum of $|1 + \mu + a|^H$ and $|-(1 - \frac{1}{4}\mu) + a|^H\}$. Since $\min_a \max\{|1 + \mu + a|, |-(1 - \frac{1}{4}\mu) + a|\} \geq 1 + \frac{\mu}{4}$, the bound follows. ■

Appendix D. Experiments

We conduct a series of experiments using the open-loop stable construction Construction L.1 underlying Theorem 1, demonstrating that our construction can be used as a benchmark for common behavior cloning pipelines. See Appendix D for details. We visualize in Figure 1 the cost $\max_t \langle \mathbf{e}_1, \mathbf{x}_t \rangle$

(a) for different checkpoints over the course of a single training run, (b) as a function of the number of rollout timesteps for different methods, and (c) on diffusion policy with larger action-chunks. The experiments highlight several counterintuitive aspects of our construction:

1. The rollout cost increases although validation loss decreases throughout training.
2. Random noise outperforms all policy learning methods and avoids exponential-in-time error, due to the E-IISS open-loop stability of the dynamics.
3. More complex techniques such as Diffusion Policy (Chi et al., 2023), replica noising (Block et al., 2024), and action chunking improve performance.

Notably, action chunking does not suffer from exponential error, which we attribute to the open loop stability of each chunk. These results affirm our theory and suggest that imitators must non-simple in order to avoid exponential error.

Dynamics Details. For the experiments in Appendix D, we use the construction Construction L.1, with $d = 4$ and visualize the performance on the A_1, K_1 matrices. We use $\mu = 1/8$ (instead of $1/4$) to slightly reduce the instability of the system so that we can visualize the effect of larger H. This does not affect the key properties of the construction beyond slightly reducing the instability.

For the nonlinear perturbation function g used in the construction of the dynamics and expert of Construction L.1, we used a randomly initialized 3-layer MLP with 16 hidden units in each layer and tanh activations. The weights and biases were initialized using a truncated normal and a uniform distribution over $[-1, 1]$, respectively.

Model Details. The behavior cloning policies were parameterized by 4-layer MLPs of similar design to the g network to ensure feasibility of the learning problem. For all diffusion policy experiments, we used a 3-layer MLP with 16 hidden units with FiLM conditioning Perez et al. (2018). We used a 256-dimensional sinusoidal time embedding, concatenated with the observation, as an input to the FiLM embedding.

Training Details. We used a batch size of 512 for the behavior cloning and 128 for the diffusion policy. All policies were trained for $10,000$ iterations using $N = 8192$ training trajectories. For all expertments we use the AdamW optimizer (Loshchilov, 2017) with a cosine decay schedule (Loshchilov and Hutter, 2016). For the behavior cloning experiments, we use an initial learning rate and weight decay of 1×10^{-3} and for diffusion policy we use an initial learning rate of 1×10^{-4} and weight decay of 1×10^{-5}.

Evaluation Details. All models were evaluated over 16 initial conditions across 5 different training seeds (for a total of 80 unique $\mathbf{x}_1$). For the action chunking experiments, we trained models with chunk lengths $h \in [1, 2, 4, 8]$. For the replica noising experments, we used a noise parameter of $\sigma = 0.1$. We show the performance of the different policies over rollouts of length $H \in [2, 4, 8, 12, 20, 26, 32]$.

Part II
Minimax Formulations and Generalized Results

The next three sections that follow are intended for readers either familiar with, or curious about, the statistical learning and decision-making literature. The main focus is framing all results in the language of minimax risks which, while natural and expedient to those familiar with statistical learning, may be cumbersome for those less familiar (hence, the decision to defer these sections).

Appendix E introduces a systematic treatment of compounding error in imitation learning problems via minimax risks. Compounding error is then the distuation when the minimax risk under the expert distribution (we focus on $\mathsf{M}_{\mathrm{expert},L_2}$) and the risk under evaluation on a cost ($\mathsf{M}_{\mathrm{cost}}$ in expectation, and $\mathsf{M}_{\mathrm{cost,prob}}$ in probability) differ by large amounts. This section also introduces language for minimax risks of standard supervised learning problems with the L_2 loss.

Appendix F provides more general, more detailed versions of the results in Section 3, stated in terms of the minimax risks from the preceding section. The idea is to show that **any** L_2 **regression problem** satisfying the appropriate regularity conditions can be **embedded** into an imitation learning problem in which compounding error occurs. The results in Section 3 can be instantiated by using Proposition E.1 in the previous section, which guarantees that for any rational $q \in \mathbb{Q}$, there exist a sufficiently regular L_2-regression problem whose error decays like n^{-q}. This section also describes additional features and strengthenings of the results.

Finally, Appendix G describes the formal, general schematic used to prove of the aforementioned results. It then briefly discusses how this schematic is specialized to each particular result. The proof of the main results for simple policies, Theorems 2 and 1.A, adopts the strategy already described in Section 4. Theorem 3, our guarantee for non-simple policies, uses the same construction but a somewhat more intricate proof strategy, and the lower bound with unstable dynamics, Theorem 4, uses a construction based on random orthogonal matrices.

Appendix E. Minimax Imitation Learning Risks

In this section, we introduce a more systematic formulation of the results stated in Section 3. We adopt the language of *minimax risks*, which cast statistical decision problems as zero-sum games between learning algorithms (the "$\min$ player") and adversaries selecting the unknown problem parameter (the $\max$ player). The cost (or negative payoff) in the game is the risk function to be minimized by the learner. The minimax risk thus characterizes the best attainable expected value of the function, over all randomness involved, on the worst-case problem instance. For a comprehensive treatment of minimax risks in statistical estimation and decision making, consult the works Wainwright (2019); van der Geer (2000); Györfi et al. (2006); Tsybakov (1997), and references therein. Furthermore, all proofs in this section are deferred to Appendix J.

The remainder of the section has the following organization. First, in Appendix E.1, we introduce the standard formalism of the minimax risk specialized to IL problems. Next, we introduce a notion of "in-probability" minimax risk in Appendix E.2, which gives provides a more granular characteriztion of the compounding error behavior. Our lower bounds follow by embedding supervised learning problems. To this end, we introduce minimax risks for standard supervised learning problems in Appendix E.3. This includes the stipulation of an important *typicality assumption*, which we show holds for a very general family of regression problems (Proposition E.1).

E.1. IL Minimax Risks

Recall that a $(\mathbb{R}^d, \mathbb{R}^m)$-**IL problem family** is a tuple $(\mathcal{P}, D)$ of instances $\mathcal{P} = \{(\pi^\star, f)\}$ with $\mathbb{X} = \mathbb{R}^d$ and $\mathbb{U} = \mathbb{R}^m$, and initial distribution D on $\mathbb{X}$. .

Definition E.1 (IL Minimax Risk) *Let $(\mathcal{P}, D)$ be an $(\mathbb{R}^d, \mathbb{R}^m)$-IL problem family. Further, let $\mathbb{A}$ be a class of IL estimation algorithms mapping samples $\mathrm{S}_{n,H}$ to (distributions over) policies. For $n \in \mathbb{N}$ trajectories and horizon $H \in \mathbb{N}$, the minimax risk of $\mathbb{A}$ under a risk function $\mathbf{R}(\hat{\pi}; \pi^\star, f, D, H)$ is*

$$\mathbf{M}^{\mathbb{A}}(n, \mathbf{R}; \mathcal{P}, D, H) := \inf_{\mathrm{alg} \in \mathbb{A}} \sup_{(\pi^\star, f) \in \mathcal{P}} \mathbb{E}_{[\mathrm{alg}, \pi^\star, f, n, H]} \left[\mathbf{R}(\hat{\pi}; \pi_\star, f, D, H) \right]. \tag{E.1}$$

As described above, the minimax risk admits a game-theoretic interpretation: a learner's move is their selection of algorithm alg, and an *adversary* selects an instance $(\pi^\star, f) \in \mathcal{P}$. The learner's penalty is then the expected risk over all sources of randomness $\mathbb{E}_{\mathrm{S}_{n,H} \sim [\pi^\star, f, D]} \mathbb{E}_{\hat{\pi} \sim \mathrm{alg}(\mathrm{S}_{n,H})} [\mathbf{R}(\hat{\pi}; \pi_\star, f, D, H)]$. Minimax risk thus measures the minimal penalty the learner can suffer in such a game. Notice that our formalism treats D as fixed, which can be interpreted as given the learner foreknowledge of the initial state distribution. This foreknowledge only makes lower bounds stronger.

Thoughout, we adopt the shorthand for validation and evaluation risks:

$$\begin{aligned} \mathbf{M}^{\mathbb{A}}_{\mathrm{expert}, L_2}(n; \mathcal{P}, D, H) &:= \mathbf{M}^{\mathbb{A}}(n, \mathbf{R}_{\mathrm{expert}, L_2}; \mathcal{P}, D, H), \\ \mathbf{M}^{\mathbb{A}}_{\mathrm{cost}}(n; \mathcal{P}, D, H) &:= \mathbf{M}^{\mathbb{A}}(n, \mathbf{R}_{\mathrm{cost}}; \mathcal{P}, D, H) \end{aligned} \tag{E.2}$$

While most of our lower bounds focus on restricted algorithm classes $\mathbb{A}$, some lower bounds: they hold even without restriction to a particular class of algorithms.

Definition E.2 (Unrestricted Minimax Risk) *We define the* unrestricted *minimax risk $\mathbf{M}(n, \mathbf{R}; \mathcal{P}, D, H)$ as $\mathbf{M}^{\mathbb{A}_\star}(n, \mathbf{R}; \mathcal{P}, D, H)$, where $\mathbb{A}_\star$ contains all IL algorithms* alg *mapping $\mathrm{S}_{n,H}$ to (distributions over) policies $\hat{\pi}$. We even include in $\mathbb{A}_\star$ algorithms* alg *which can return a $\hat{\pi}$ for which $\hat{\pi}$ may depend on time-step t and past; i.e. $\hat{\pi}$ maps $(t, \mathbf{x}_{1:t}, \mathbf{u}_{1:t-1})$ to distributions over $\mathbf{u}_t$. We define the unrestricted minimax validation and evaluation risks $\mathbf{M}_{\mathrm{expert}, L_2}$ and $\mathbf{M}_{\mathrm{cost}}$ by direct analogy to Eq. (E.2).*

Lower bounds against unrestricted algorithm classes are often called *information-theoretic*, in that they leverage the learners incomplete information about the ground-truth problem instance moreso than any algorithmic limitation imposed on the learner (or on the policies $\hat{\pi}$).

E.2. In-Probability Minimax Risks

It may be objected that lower bounds on expected costs may be misleading, because compounding error may be large on rare events (as, for example, observed in the case of benevolent gambler's ruin in Section 5). In what follows, we present a fixed-cost, in-probability risk, $\mathbf{M}_{\mathrm{eval,prob}}$, which leads to more stringent lower bounds that rule out rare-event compounding error. We shall also show that this notion implies lower-bounds on the $\mathbf{M}_{\mathrm{cost}}$ defined above.

It is most convenient to state our definition of in-probability risk for a $\mathrm{cost} : \mathbb{X}^H \times \mathbb{U}^H \to \mathbb{R}$ that vanishes on expert trajectories:

Definition E.3 *We say a* $\mathrm{cost} : \mathbb{X}^H \times \mathbb{U}^H \to \mathbb{R}$ *"vanishes on* $(\mathcal{P}, D)$*" if for all* $(\pi^\star, f) \in \mathcal{P}$*,*

$$\mathbb{P}_{\pi^\star, f, D}[\mathrm{cost}(\mathbf{x}_{1:H}, \mathbf{u}_{1:H}) = 0] = 1.$$

We define the set of such costs as $\mathcal{C}_{\mathrm{vanish}}(\mathcal{P}, D)$*.*

We define a fixed-cost "in-probability risk" as the probability p as the smallest ϵ such that the cumulative probability over exceeding ϵ, under all randomness of validation and evaluation of the policy, is at most p.

Definition E.4 (In-Probability Risk) *Given* $n \geq 1$ *and* $p \in (0, 1]$*, and a* $\mathrm{cost} \in \mathcal{C}_{\mathrm{vanish}}(\mathcal{P}, D)$*, we define the in-probability risk as*

$$\mathsf{M}^{\mathbb{A}}_{\mathrm{cost,prob}}(n, \delta; \mathcal{P}, D, H) := \inf \left\{ \epsilon : \inf_{\mathrm{alg} \in \mathbb{A}} \sup_{(\pi^\star, f) \in (\mathcal{P}, D)} \mathbb{E}_{[\mathrm{alg}, \pi^\star, f, n, H]}[\mathbb{P}_{\hat{\pi}, f, D}[\mathrm{cost}(\mathbf{x}_{1:H}, \mathbf{u}_{1:H}) \geq \epsilon]] \leq \delta \right\}.$$

We define unrestricted minimax risks by analogy to Definition E.2.

We remark that the above risks are are equivalent to quantile risks considered in recent work in the statistical learning community (El Hanchi et al., 2024; Ma et al., 2024). However, while those works are concerned with establishes larger lower bounds for estimation with high-probability guarantees, the focus in this work is simply showing that large error occurs with constant probability.

Note that, by Markov's inequality, it holds that (c.f. Proposition J.2(c))

$$\forall \mathrm{cost} \in \mathcal{C}_{\mathrm{vanish}}(\mathcal{P}, D), \quad \delta \cdot \mathsf{M}^{\mathbb{A}}_{\mathrm{cost,prob}}(n, \delta; \mathcal{P}, D, H) \leq \mathsf{M}^{\mathbb{A}}_{\mathrm{cost}}(n; \mathcal{P}, D, H), \tag{E.3}$$

Thus, a lower bound $\mathsf{M}_{\mathrm{cost,prob}}$ suffices for lower bounds on $\mathsf{M}_{\mathrm{cost}}$. Further variants of the above risks are discussed in Appendix J.3.

E.3. Embedding Regression Problems

We will derive lower bounds on the minimax risk by embedding in more standard supervised regression problems over classes of functions $\mathcal{G}$, which can be viewed as 1-step IL problems.

Definition E.5 (Supervised Learning Minimax Risks) *A-*$\mathbb{R}^k$ *regression problem family is a pair* $(\mathcal{G}, D_{\mathrm{reg}})$ *consisting of a distribution* D_{reg} *on* $\mathbb{R}^k$ *and a class of scalar-valued functions* $\mathcal{G} = \{g : \mathbb{R}^k \to \mathbb{R}\}$*. Given such a regression problem family* $(\mathcal{G}, D_{\mathrm{reg}})$*, its minimax risk is*

$$\mathsf{M}_{\mathrm{reg}, L_2}(n; \mathcal{G}, D_{\mathrm{reg}}) = \inf_{\mathrm{alg}_{\mathrm{reg}}} \sup_{g^\star \in \mathcal{G}} \mathbb{E}_{S_{n,\mathrm{reg}}} \mathbb{E}_{\hat{g} = \mathrm{alg}_{\mathrm{reg}}(S_{n,\mathrm{reg}})} \left(\mathbb{E}_{\mathbf{z} \sim D_{\mathrm{reg}}}[|\hat{g}(\mathbf{z}) - g^\star(\mathbf{z})|^2] \right)^{1/2}. \tag{E.4}$$

where $\mathbb{E}_{S_{n,\mathrm{reg}}}$ *denotes expectation over samples* $S_{n,\mathrm{reg}} = (\mathbf{z}^{(i)}, g^\star(\mathbf{z}^{(i)}))_{1 \leq i \leq n}$ *for* $\mathbf{z}^{(i)} \overset{\mathrm{i.i.d}}{\sim} D_{\mathrm{reg}}$*, and* $\mathrm{alg}_{\mathrm{reg}}$ *is any measurable function mapping* $S_{n,\mathrm{reg}}$ *to functions* $\hat{g} : \mathbb{R}^k \to \mathbb{R}$*. Given* $p \in (0, 1]$*, we define an in-probability risk*

$$\mathsf{M}_{\mathrm{reg,prob}}(n, \delta; \mathcal{G}, D_{\mathrm{reg}}) := \inf \left\{ \epsilon : \inf_{\mathrm{alg}_{\mathrm{reg}}} \sup_{g^\star \in \mathcal{G}} \mathbb{E}_{S_{n,\mathrm{reg}}} \mathbb{E}_{\hat{g} = \mathrm{alg}_{\mathrm{reg}}(S_{n,\mathrm{reg}})} \mathbb{P}_{\mathbf{z} \sim D_{\mathrm{reg}}, \hat{\mathbf{y}} \sim \hat{g}(\mathbf{z})}[|\hat{\mathbf{y}} - g^\star(\mathbf{z})| > \epsilon] \leq \delta \right\}.$$

Remark E.1 *Note that, in full generality, both* $\mathrm{alg}_{\mathrm{reg}}$ *may be randomized, and the function* $\hat{g}$ *may be a stochastic function of its input:* $\hat{\mathbf{y}} \sim \hat{g}(\mathbf{z})$*. However, for the* L_2 *regression risk, Jensen's inequality implies that randomized regression estimators do not improve the minimax regression risk.*

By a Chebyshev's inequality argument, we always have the inequality

$$\mathsf{M}_{\mathrm{reg},L_2}(n;\mathcal{G},D_{\mathrm{reg}}) \ge \sqrt{\delta}\mathsf{M}_{\mathrm{reg,prob}}(n,\delta;\mathcal{G},D_{\mathrm{reg}}). \tag{E.5}$$

E.4. Regularity and "Typicality" Conditions for Regression

Because we consider imitation of an deterministic expert, the regression problems considered are noiseless. This is often referred to the *interpolation setting* in statistical learning. For further discussion, see e.g. Kohler and Krzyżak (2013) and the references therein.

First, we codify some more standard regularity conditions

Definition E.6 (Regular Regression Instances) *We say* $(\mathcal{G},D_{\mathrm{reg}})$ *is* R-bounded *if with probability one over* $\mathbf{z} \sim D_{\mathrm{reg}}$*,* $\|\mathbf{z}\| \le R$*, and for all* $\mathbf{z} : \|\mathbf{z}\| \le R$*,* $|g(\mathbf{z})| \le R$*. We say* $(\mathcal{G},D_{\mathrm{reg}})$ *is* (R,L,M)*-regular of if each* $(g,D_{\mathrm{reg}}), g \in \mathcal{G}$ *is* R*-bounded, and* g *are* L*-Lipschitz and* M*-smooth, and the class* $\mathcal{G}$ *is closed under convex combination.*

Next, recall that we must show compounding error occurs with good probability; otherwise, if large errors occur with low probability, then for the losses bounded in $[0,1]$ in $\mathcal{C}_{\mathrm{Lip}}$, the contributions of these errors are insignificant in expectation. To this end, we introduce technical condition ensuring that if the minimax risk scales like $\boldsymbol{\epsilon}_n$, then a similar lower bound on the risk holds with constant probability as well. Because it is common to derive in-expectation lower bounds from in-probability ones (see, e.g. Tsybakov (1997)), we denote this condition "typical"-ity.

Condition E.1 (Typical Problem Class) *Let* $\kappa,\delta \in (0,1)$*. We say that* $(\mathcal{G},D_{\mathrm{reg}})$ *is* (κ,δ)*-typical if*

$$\mathsf{M}_{\mathrm{reg,prob}}(n,\delta;\mathcal{G},D_{\mathrm{reg}}) \ge \kappa\mathsf{M}_{\mathrm{reg},L_2}(n;\mathcal{G},D_{\mathrm{reg}}), \quad \forall n \ge 1. \tag{E.6}$$

Up to κ and δ, Condition E.1 is the converse of the inequality Eq. (E.5). Finally, we say $\mathcal{G}$ is *convex* if $g_1, g_2 \in \mathcal{G}$ implies $\alpha g_1 + (1-\alpha)g_2 \in \mathcal{G}$ for any $\alpha \in [0,1]$. In Appendix J.2, we verify that a large, classical families of regression problems are smooth, typical, and realize any desired fractional rate of estimation. Specifically, we establish the following result.

Proposition E.1 *For any integers* $s \ge 2, k \ge 1$*, there exist constants* $\kappa,\delta \in (0,1)$ *and* $C, C' > 0$ *depending only on* s *and* k*, and an* $\mathbb{R}^k$ *-regression problem family* $(\mathcal{G},D_{\mathrm{reg}})$ *which is* (κ,δ)*-typical,* $(1,1,1)$*-regular, such that* $\mathcal{G}$ *is convex and for all* $n \ge 1$*,*

$$C\mathsf{M}_{\mathrm{reg},L_2}(n;\mathcal{G},D) \le n^{-s/k} \le C'\mathsf{M}_{\mathrm{reg},L_2}(n;\mathcal{G},D). \tag{E.7}$$

Appendix F. Minimax Lower Bounds for Imitation Learning

This section presents detailed statements of our lower bounds, stated in the language of minimax risks developed in Appendix E. These results demonstrate that compounding error is a phenomena that occurs independent of the statistical difficulty of minimizing the training risk, in the following sense that any typical statistical learning problem (Condition E.1) can be embedded into a IL problem with exponential compounding erorr.

More specifically, we assume we are given an $\mathbb{R}^k$-regression problem family $(\mathcal{G}, D_{\mathrm{reg}})$ which is (κ, δ)-typical (Condition E.1), and such that D_{reg} is 1-bounded (Definition E.6). We use the following shorthand for the minimax risk of this regression problem

$$\boldsymbol{\epsilon}_n := \mathbf{M}_{\mathrm{reg},L_2}(n; \mathcal{G}, D_{\mathrm{reg}}).$$

For the first two results, we also assume $(\mathcal{G}, D_{\mathrm{reg}})$ is $(1,1,1)$-regular (Definition E.6) and $\mathcal{G}$ is convex. We will then show that such classes can be embedded into Behavior Cloning problems such that

(a) the restricted and unrestricted minimax training risks coincide, and are close to the supervised learning minimax risk $\mathbf{M}^{\mathbb{A}}_{\mathrm{expert},L_2}(n; \mathcal{P}, D, H) = \mathbf{M}_{\mathrm{expert},L_2}(n; \mathcal{P}, D, H) \approx \boldsymbol{\epsilon}_n$.

(b) There exists a cost $\in \mathcal{C}_{\mathrm{Lip}}$ such that the in-probability risks are considerably large. Specifically, $\mathbf{M}^{\mathbb{A}}_{\mathrm{cost,prob}}(n; \mathcal{P}, D, H) \gg \boldsymbol{\epsilon}_n$, and often $\mathbf{M}_{\mathrm{cost,prob}}(n, \Omega(\delta); \mathcal{P}, D, H) \ge \exp(\Omega(H))\boldsymbol{\epsilon}_n$.

We proceed to state three formal lower bounds. First, Theorem 1.A (Appendix F.1) demonstrates that the class of simple IL algorithms (Definition 3.3) with smooth means and simply-stochastic noise incur exponential-in-H compoundinger error. Next, Theorem 3.A (Appendix F.2) shows that exponential-in-H compounding occurs even for a much larger class of algorithms with anti-concentrated noise (Definition F.3), but this is capped to a rate of $\boldsymbol{\epsilon}_n^{1-\Omega(1)}$. The illustrative benevolvent gambler's ruin policy in Section 5 provides weak evidence that non-simply stochastic may indeed be able to enjoy at most $\boldsymbol{\epsilon}_n^{1-\Omega(1)}$ error due to clever randomization. Finally, Theorem 4.A (Appendix F.3), shows that for problem families where the expert-dynamics pairs $(\pi^\star, f)$ are closed-loop E-IISS, but the open-loop dynamics may be unstable, the unrestricted minimax rates exhibits exponential-in-H compounding error.

Before continuing to the statements of these results, we describe some additional further features of the lower bounds that follow.

Proper learning is optimal on the expert distribution. In all results that follow, we show that proper learning is *optimal* from the perspective of minimizing the loss under the distribution of the expert. Hence, while improperness may be of benefit when the policy is deployed, it confers no benefit when imitating expert data. The exact optimality of proper algorithms requires our consideration of L_2 expert distribution error (see Appendix I.5 for discussion).

Controllability. In addition to the all the regularity conditions (smoothness, boundedness, stability) promise above, we will also ensure that our constructions satisfy yet another desirable property: the dynamics $f \in \mathcal{F}(\mathcal{P})$ are 1-step controllable.

Definition F.1 *Let $f : \mathbb{X} \times \mathbb{U} \to \mathbb{X}$ be a dynamical map. We say that f is C-one-step controllable if, for all $\mathbf{x}, \mathbf{x}' \in \mathbb{X}$, there exists some $\mathbf{u} \in \mathbb{U}$ for which $f(\mathbf{x}, \mathbf{u}) = \mathbf{x}'$, and $\|\mathbf{u}\| \le C(1 + \|\mathbf{x}\| + \|\mathbf{x}'\|)$. We say that f is $O(1)$-one-step-controllable if the above holds for some universal constant $C = O(1)$.*

In fact, with a little additional effort, one can show that for the dynamics f in our construction, the equation $\mathbf{x}' = f(\mathbf{x}, \mathbf{u})$ admits a unique solution $\mathbf{u}^\star(\mathbf{x}', \mathbf{x})$ for each $\mathbf{x}', \mathbf{x} \in \mathbb{X}$, and $\mathbf{u}^\star$ depends smoothly on $(\mathbf{x}', \mathbf{x})$. This means that neither a lack of controllability, nor the an innability to control the system smoothly, are to blame for the lower bounds.

Horizon scale invariance. All the bounds that follow also hold when the cost function, cost, is the maximum over time steps H over 1-Lipschitz costs, rather than the sum. This gives a normalization of the total cost which is horizon independent, whereas the sum of costs typically grows linearly in H. See Appendix J.3 for further discussion.

Longer horizon demonstrations do not help. Each of the lower bounds hold in the regime where the learner has access to a sample $S_{n,H'}$, where $H' \ge H$ is any *arbitrarily long* problem horizon (even infinitely long $H' = \infty$, measure-theoretic considerations permitting). This rules out the possibility that longer problem horizons may make the behavior cloning problem easier.

F.1. Minimax Compounding Error for IL with Simple Policies

This section states our lower bound against simple IL algorithms ($\mathbb{A}_{\mathrm{simple}}$, Definition 3.3), which we recall are those algorithms which return simply-stochastic policies with smooth and Lipschitz means. Our lower bounds follow from embedding regular, typical regression problems satisfying the assumption that follows.

Assumption F.1 *We assume that $(\mathcal{G}, D_{\mathrm{reg}})$ is $(1,1,1)$-regular (recall Definition E.6) and is (κ, δ)-typical (Condition E.1), and that $\mathcal{G}$ is convex. In particular, the classes of regression problems whose existence is guaranteed by Proposition E.1 all satisfy this condition.*

We now state the main theorem:

Theorem 1.A (Lower Bound for Stable Systems, Detailed Version) *Let $0 < c \le 1 \le C$ be universal constants, let system dimension $k \in \mathbb{N}$, and consider any k-dimensional regression problem family $(\mathcal{G}, D_{\mathrm{reg}})$ satisfying Assumption F.1. Then, for $d = k+2$, there is a (d,d)-dimensional IL problem family $(\mathcal{P}, D)$ which is $\mathrm{O}(1)$-regular (Definition 3.2), and cost function $\mathrm{cost} \in \mathcal{C}_{\mathrm{Lip}} \cap \mathcal{C}_{\mathrm{vanish}}(\mathcal{P}, D)$, such that for any $L, M \ge C$, the the class of estimators $\mathbb{A} = \mathbb{A}_{\mathrm{simple}}(L, M)$ contains $\mathbb{A}_{\mathrm{proper}}(\mathcal{P})$, and satisfies the following:*

$$\begin{aligned} \mathsf{M}_{\mathrm{expert},L_2}(n;\mathcal{P},D,H) &= \mathsf{M}^{\mathbb{A}}_{\mathrm{expert},L_2}(n;\mathcal{P},D,H) = \mathsf{M}^{\mathbb{A}_{\mathrm{proper}}(\mathcal{P})}_{\mathrm{expert},L_2}(n;\mathcal{P},D,H) \\ &\in \left[\frac{\tau}{2}\boldsymbol{\epsilon}_n, \quad \boldsymbol{\epsilon}_{n/3} + Ce^{-cn}\right] \end{aligned} \tag{F.1}$$

and

$$\mathsf{M}^{\mathbb{A}}_{\mathrm{cost,prob}}(n, c\delta; \mathcal{P}, D, H) \ge c \min\left\{\boldsymbol{\epsilon}_n \cdot \kappa \left(\frac{17}{16}\right)^{H-2}, \frac{1}{L^2 M d}\right\}, \tag{F.2}$$

where τ, c, C can be chosen to be universal constants, and δ, κ are as in Assumption F.1. Finally, for every $(\pi, f) \in \mathcal{P}$, are both f and (π, f) are (C, ρ)-E-IISS, where $\rho \in (0,1)$ is a universal constant strictly less than 1, and f is $O(1)$-one-step-controllable.

The proof of Theorem 1.A is given in Appendix L, based on the high-level schematic in Appendix G. The result consists of four statements. First, the minimax expert distribution minimax risk of the IL problem is, up to constants, exponentially small additive terms, and constant scalings of the sample size, the same as that of the embedded regression problem. Second, the minimax rates of proper IL algorithms, unrestricted IL algorithms, and simple algorithms are identical when measured in terms of the expert distribution (note: the equivalence of the first two implies equivalence to the third, due to $\mathbb{A}_{\mathrm{proper}}(\mathcal{P}) \subset \mathbb{A}_{\mathrm{simple}}(O(1)) \subset \{\text{unrestricted algorithms}\}$). The third is that the in-probability minimax risk of the IL problem is exponentially-in-H larger.

The final statement checks all desired regularity conditions. As mentioned above, cost and D are fixed for all n and H; thus, neither unsupervised knowledge of the initial state distribution nor knowledge of the cost (as in, say, an offline RL framework) suffice to avoid exponentially compounding error.

Proof [Deriving Theorems 1 and 2] Theorems 1 and 2 are both readily derived from Theorem 1.A. By Proposition E.1, for each s, k, we can take an $\mathbb{R}^k$ regression class $\mathcal{G}$ for which $\boldsymbol{\epsilon}_n = \epsilon^{-s/k}$, and κ, δ to be constants depending only on (s, k), and $d = O(k)$. Further, $\boldsymbol{\epsilon}_n \gg \exp(-cn)$, but to constants. Thus, Eq. (F.1) implies (a) of Theorem 1, whereas Eq. (F.2) implies Theorem 2. Finally, Theorem 2 implies (b) in Theorem 1 via the Markov's inequality statement, Eq. (E.5). ■

F.2. Minimax Compounding for Smooth, Non-Simply-Stochastic Policies

Generalizing from simply-stochastic policies, we now establish lower bounds against algorithms which return policies that need not be simply stochastic, but satisfy a mild and broadly applicable anti-concentration condition. As noted above, the lower bound is somewhat weaker: compounding error occurs, but only up until an $\boldsymbol{\epsilon}_n^{1-\Theta(1)}$ threshold. Moreover, compounding error is measured in L_2, which exacerbates the contribution of heavy-tailed errors. Specifically, for $\mathrm{cost} \in \mathcal{C}_{\mathrm{vanish}}(\mathcal{P}, D)$, we define an L_2-analogue of $\mathbf{M}_{\mathrm{cost}}$, namely:

$$\begin{aligned} \mathbf{R}_{\mathrm{cost},L_2}(\hat{\pi}; \pi^\star, f, D, H) &:= \mathbb{E}_{\hat{\pi}, f_{g,\xi}, D}\left[|\mathrm{cost}(\mathbf{x}_{1:H}, \mathbf{u}_{1:H})|^2\right]^{1/2}, \quad \mathrm{cost} \in \mathcal{C}_{\mathrm{vanish}}(\mathcal{P}, D) \\ \mathbf{M}_{\mathrm{cost},L_2}(n; \mathcal{P}, D, H) &:= \mathbf{M}^{\mathbb{A}}\left(n, \mathbf{R}_{\mathrm{cost},L_p}; \mathcal{P}, D, H\right) \end{aligned} \tag{F.3}$$

These differences aside, our lower bounds shows that the benevolent gambler's ruin strategy of Section 5 is qualitatively unimprovable in general. Our lower bound pertains to algorithms which return policies statisfying a mild anti-concentration condition, stated first for general random variables.

Definition F.2 (Quantitative Anti-Concentration) *Let $\alpha, p \in (0, 1]$. We say that a scalar random variable Z is (α, p)-anti-concentrated if it satisfies*

$$\mathbb{P}[|Z - \mathbb{E}[Z]| \geq \alpha \mathbb{E}[|Z - \mathbb{E}[Z]|^2]^{1/2}] \geq p. \tag{F.4}$$

We say that a random vector $\mathbf{z} \in \mathbb{R}^d$ if (c, p)-anti-concentrated if $\langle \mathbf{v}, \mathbf{z} \rangle$ is (α, p)-anti-concentrated for any vector $\mathbf{v} \in \mathbb{R}^d$ (equivalently, for any unit vector).

Importantly, our definition of anti-concentration is relative to the random variable's own variance. In particular, **deterministic** random variables $(1, 1)$-anti-concentrated according to the above definition. Next, we extend our notion of anti-concentration to policies.

Definition F.3 (Anti-Concentrated Policy) *We say that a policy π is (α, p) anti-concentrated if, for any $\mathbf{x}, \mathbf{x}' \in \mathbb{R}^d$, there exists a coupling $P(\mathbf{x}, \mathbf{x}')$ of $\pi(\mathbf{x}), \pi(\mathbf{x}')$[1] such that if $(\mathbf{u}, \mathbf{u}') \sim P(\mathbf{x}, \mathbf{x}')$, the random vector $\mathbf{u} - \mathbf{u}'$ is (α, p)-anti-concentrated.*

The ability to choose any coupling P implies that anti-concentration holds for very general classes of policies, including: all simply-stochastic policies (in particular, determinisitic policies), all Gaussian policies $\pi(\mathbf{x}) = \mathcal{N}(\mu(\mathbf{x}), \mathbf{\Sigma}(\mathbf{x}))$, and policies which are mixtures of anti-concentrated policies (e.g. Gaussian mixtures or mixtures of deterministic policies) with components of constant-magnitude probability. In particular, the benevolent gambler's ruin policy (Section 5) is anti-concentrated. We verify these claims in Appendix M.2.

Generalized Smooth Policies Motivated by these examples, we define the class of "generalized smooth policies" as those which are anti-concentrated, and which have Lipschitz and smooth means.

Definition F.4 (Generalized Smooth Policies) *Let $\mathbb{A}_{\mathrm{gen,smooth}}(L, M, \alpha, p)$ denote the class of algorithms which, with probability one, return stochastic, Markovian policies π for which $\mathrm{mean}[\pi](\mathbf{x})$ is L-Lipschitz and M-smooth, and π is (α, p)-anti-concentrated.*

We are now ready to state our main result. Recall the L_2 minimax risks defined in Eq. (F.3) above. We also establish a convenient asymptotic notation.

Definition F.5 (poly-$o^\star$-notation) *Given $b_1, b_2, \dots \le 1$, we use the notation $a = \mathrm{poly}\text{-}o^\star(b_1, b_2, \dots, b_k)$ to denote that $a \le c_1(b_1 \cdot b_2 \cdot b_k)^{c_2}$, c_1 is a sufficiently small universal constant, and c_2 a sufficiently large universal constant.*

Our main theorem is as follows.

Theorem 3.A (Lower Bound for Non-Simply Stochastic Systems, Detailed Version) *Consider the setting of Theorem 1.A with $d = k + 2$, and let $(\mathcal{P}, D)$ be the corresponding problem family from that theorem. Further, recall $\boldsymbol{\epsilon}_n := \mathbf{M}_{\mathrm{reg},L_2}(n; \mathcal{G}, D_{\mathrm{reg}})$. For $L, M \ge 1$ and $\alpha, p \in (0, 1]$, now consider the class of algorithms $\mathbb{A} = \mathbb{A}_{\mathrm{gen,smooth}}(L, M, \alpha, p)$. Then Eq. (F.1) still applies to this choice of $\mathbb{A}$. Moreover, suppose that $\boldsymbol{\epsilon}_n \le \mathrm{poly}\text{-}o^\star(1/L, 1/M, 1/d, \alpha, p, \kappa, \delta)$. Then, for all $n \ge 1$,*

$$\mathbf{M}^{\mathbb{A}}_{\mathrm{cost},L_2}(n; \mathcal{P}, D, H) \ge c\kappa \cdot \delta \cdot \min\left\{\boldsymbol{\epsilon}_n \cdot 1.05^{H-2}, \boldsymbol{\epsilon}_n^{1 - \frac{1}{C'(1+\log(1/(\alpha p)))}}\right\}. \tag{F.5}$$

The proof of Theorem 3.A is given in Appendix M, again based on the high-level schematic in Appendix G. In words, this result shows that the same construction from Theorem 1.A provides a challenging distribution for non-simply-stochastic, but exponential-in-H compounding error occurs only up to a threshold which is $\boldsymbol{\epsilon}_n^{1-\Omega(1)}$. Note that, because the construction is the same, Eq. (F.1) with $\mathbb{A} = \mathbb{A}_{\mathrm{simple}}(L, M)$ implies the same for $\mathbb{A} = \mathbb{A}_{\mathrm{gen,smooth}}(L, M, \alpha, p)$, as the latter is a large algorithm class.

Proof [Deriving Theorem 3] Theorem 3 follows from Theorem 3.A exactly the same way as Theorems 1 and 2 follow from Theorem 1.A. That is, we by Proposition E.1, for each s, k, we can use an $\mathbb{R}^k$ regression class $\mathcal{G}$ for which $\boldsymbol{\epsilon}_n = \epsilon^{-s/k}$, and κ, δ to be constants depending only on (s, k), and $d = O(k)$. Theorem 3.A gives an in-probabiliy bound, whilst Eq. (E.5) converts this to a bound in expectation. ■

1. Recall that a coupling of $\pi(\mathbf{x}), \pi(\mathbf{x}')$ is a joint distribution over $(\mathbf{u}, \mathbf{u}')$ with marginals $\mathbf{u} \sim \pi(\mathbf{x})$ and $\mathbf{u}' \sim \pi(\mathbf{x}')$.

Remark F.1 (Is anti-concentration necessary?) *The anti-concentration requirement is a consequence of our choice to define policy smoothness in terms of its mean. Without this condition, policies which appear highly non-smooth with constant probability can be "smoothed" by adding low-probability, large-mass components to balance them out the means. We conjecture that by replacing mean-smoothness with a more careful notion of smoothness, based either on smoothness of densities (provided dominating measures exists), or based on classes of smooth test functions, the anti-concentration can be removed from the class* $\mathbb{A}_{\mathrm{gen,smooth}}$.

F.3. Minimax Compounding Error for Unstable Dynamics

We round out the section by proving entirely unconditional lower bounds against compounding error when the dynamics are permitted to be smooth and Lipschitz, but unstable.

Theorem 4.A (Lower Bound with Unstable Dynamics, Detailed Version) *Consider a* (κ, δ)*-typical* $\mathbb{R}^k$*-regression problem family* $(\mathcal{G}, D_{\mathrm{reg}})$*, and let* $\boldsymbol{\epsilon}_n := \mathbf{M}_{\mathrm{reg},L_2}(n; \mathcal{G}, D_{\mathrm{reg}})$*. For any integer* $d \geq k$*, and any* $\rho > 1$*, there is an* $(\mathbb{R}^d, \mathbb{R}^d)$*-IL problem family* $(\mathcal{P}, D)$ *and* $\mathrm{cost} \in \mathcal{C}_{\mathrm{Lip}} \cap \mathcal{C}_{\mathrm{vanish}}(\mathcal{P}, D)$ *, such that for all* $2 \leq H \leq \frac{1}{2}e^{d(1-\rho^{-1})^2/2}$*,*

$$\mathbf{M}_{\mathrm{expert},L_2}(n; \mathcal{P}, D, H) = \boldsymbol{\epsilon}_n \tag{F.6}$$

$$\mathbf{M}_{\mathrm{cost,prob}}\left(n, \frac{\delta}{2}; \mathcal{P}, D, H\right) \geq \min\left\{\kappa \cdot \boldsymbol{\epsilon}_n \cdot \rho^{(H-1)/2},\, c_0\right\} \tag{F.7}$$

Above, c_0 *is a universal constant. Moreover, the construction ensures that each* $(f, \pi^\star) \in \mathcal{P}$ *is* $(0,1)$*-E-IISS, and if* $(\mathcal{G}, D_{\mathrm{reg}})$ *is* (R, L, M)*-regular, then* $(\mathcal{P}, D)$ *is* (R, L', M')*-regular for* $L' = O(L + \rho)$ *and* $M' = O(M + L + \rho)$*, and each* $f \in \mathcal{F}(\mathcal{P})$ *is* $O(L + \rho)$*-one-step-controllable.*

Again, the theorem is based on the schematic outlined in (Appendix G), with the formal proof deferred to Appendix N. Note that, in the above theorem, $\mathbf{M}_{\mathrm{eval,prob}}$ is the *unrestricted minimax risk* (Definition E.2). That is, even history-dependent, non-smooth policies with arbitrary stochastic policies fail to elude the $\exp(H)$ compounding error.

Proof [Deriving Theorem 4] As with the proofs of Theorems 1 to 3 above , the result follows by instantiating Theorem 4.A with Proposition E.1. Details are the same as in the other cases. ■

Appendix G. Proof Schematic

All three lower bounds, Theorems 1.A to 4.A, all follow from the same schematic. We describe this schematic here, and then remark on how the arguments specialize at the end of the section. Throughout, fix $H \in \mathbb{N}$. Let $(\mathcal{G}, D_{\mathrm{reg}})$ be $\mathbb{R}^k$-regression problem family, and consider an $(\mathbb{R}^d, \mathbb{R}^m)$-IL problem families $(\mathcal{P}, D)$, where the instances take the form

$$\mathcal{P} = \{(\pi_{g,\xi}, f_{g,\xi}) : g \in \mathcal{G}, \xi \in \Xi\}, \tag{G.1}$$

indexed by $g \in \mathcal{G}$, and auxilliary parameter ξ. The function $g \in \mathcal{G}$ parameterizes a "first-step" of a regression problem that the learner needs to solve (as in Section 4), and ξ parameterizes some remaining residual uncertainty over the dynamics.

We assume that each $(\pi^\star, f) \in \mathcal{G}$ are deterministic. However (for convenience), we consider a slight generalization of Section 2 in which $\pi^\star(\mathbf{x}, t)$ and $f(\mathbf{x}, \mathbf{u}, t)$ are allowed depend on a time argument t. Moreover, we allow $\hat{\pi}(\mathbf{x}_1, t = 1)$ to depend on time and arbitrarily on the past $\hat{\pi}(\mathbf{x}_{1:t}, \mathbf{u}_{1:t-1}, t)$; indeed, the schematica arguments that follow hold for time-varying, non-Markov policies. Rather, it is the **instantiation** of the schematic in the proofs of Theorems 1.A and 3.A in which Markovianity plays an essential role.

Our results show that if the IL family $(\mathcal{P}, D)$ satisfies three key properties vis-a-vis the regression family $(\mathcal{G}, D_{\mathrm{reg}})$, then a general result template holds. These properties are as follows.

Property G.1 *We say the* τ-orthogonal embedding property *holds if there exists a unit vector* $\mathbf{v} \in \mathbb{R}^d : \|\mathbf{v}\| = 1$*, a mapping* $\pi_0 : \mathbb{X} \to \mathbb{U}$*, and mapping* $\mathrm{proj} : \mathbb{X} \to \mathcal{Z}$*, and a probability kernel* $\mathcal{K} : \mathcal{Z} \to \mathbf{\Delta}(\mathbb{X})$ *such that*

- *The distribution of* $\mathbf{z} = \mathrm{proj}(\mathbf{x})$ *under* $\mathbf{x} \sim D$ *is* D_{reg}*, and the distribution of* $\mathbf{x} \sim \mathcal{K}(\mathbf{z})$ *under* $\mathbf{z} \sim D_{\mathrm{reg}}$ *is* D*, and satisfies* $\mathrm{proj}(\mathbf{x}) = \mathbf{z}$*.*
- *With probability* 1 *over* $\mathbf{x} \sim D$*,* $\pi_{g,\xi}(\mathbf{x}, 1) = \pi_0(\mathbf{x}) + \tau g(\mathrm{proj}(\mathbf{x}))\mathbf{v}$*, where again* π_0 *is fixed across instances. In particular,* $\pi_{g,\xi}(\mathbf{x}, 1)$ *does not depend on* ξ*.*

Property G.2 *We say the* single step property *holds if if the conditional distribution* $\mathbb{P}^{\pi,f,D}_{\mathbf{traj}_H | (\mathbf{x}_1, \mathbf{u}_1)}$ *of the trajectory given* $(\mathbf{x}_1, \mathbf{u}_1)$ *is identical for all* $(\pi, f) \in \mathcal{P}$*, for all* $(\mathbf{x}_1, \mathbf{u}_1)$ *which are in the support of the distribution of* $\mathbb{P}^{\pi,f,D}$*.*

Property G.3 *We say the* ξ-indistinguishable *property holds if, under* D *if* $(\pi_{g,\xi}, f_{g,\xi})$ *and* $(\pi_{g,\xi'}, f_{g,\xi'})$ *induces the same distribution over trajectories for all* ξ, ξ' *(notice* g *is the same).*

Effectively, Property G.1 says that the first-step of behavior cloning in $(\mathcal{P}, D)$ is equivalent to the regression problems in $(\mathcal{G}, D_{\mathrm{reg}})$. Property G.2 says that all information about g can be gleaned only from the $t = 1$ time steps in the available sample $\mathrm{S}_{n,H}$, and Property G.3 says that $\mathrm{S}_{n,H}$ does not provide any information about the auxiliary vector ξ.

Our lower bounds will all be established by checking Properties G.1 to G.3. Once verified, the following proposition can be invoked, whose proof is given in Appendix K.

Proposition G.1 *Suppose* $(\mathcal{P}, D)$ *satisfy Properties G.1 to G.3 with parameter* τ *vis-a-vis* $(\mathcal{G}, D_{\mathrm{reg}})$*. Then,*

(a) We have the equality

$$\mathbf{M}_{\mathrm{expert},L_2}(n; \mathcal{P}, D) = \mathbf{M}_{\mathrm{expert},h=1}(n; \mathcal{P}, D) = \tau \mathbf{M}_{\mathrm{reg},L_2}(n; \mathcal{G}, D_{\mathrm{reg}}), \tag{G.2}$$

where

$$\mathbf{M}_{\mathrm{expert},h=1}(n; \mathcal{P}, D) := \inf_{\mathrm{alg}} \sup_{(\pi,f)\in\mathcal{P}} \mathbb{E}_{\mathrm{S}_{n,H}} \mathbb{E}_{\mathbf{x}_1 \sim D} \mathbb{E}_{\mathbf{u} \sim \hat{\pi}(\mathbf{x},1)} \left[\|\pi(\mathbf{x}_1, t = 1) - \mathbf{u}\|^2\right]^{1/2},$$

considers the training minimax risk associated with errors at time step $t = 1$*.*

(b) If $\mathcal{G}$ *is convex, and* $\mathbb{A} \supseteq \mathbb{A}_{\mathrm{proper}}(\mathcal{P})$*, then*

$$\mathbf{M}_{\mathrm{expert},L_2}(n; \mathcal{P}, D) = \mathbf{M}^{\mathbb{A}}_{\mathrm{expert},L_2}(n; \mathcal{P}, D) = \mathbf{M}^{\mathbb{A}_{\mathrm{proper}}(\mathcal{P})}_{\mathrm{expert},L_2}(n; \mathcal{P}, D) \tag{G.3}$$

(c) Let $\mathbb{A}$ be a class of estimators satisfying Eq. (G.3), and let $\Pi_{\mathbb{A}}$ denote some class of policies such that every $\mathrm{alg}(\mathrm{S}_{n,H}) \in \mathbb{A}$ returns a policy $\hat{\pi} \in \Pi_{\mathbb{A}}$ with probability one, for any sample $\mathrm{S}_{n,H}$.

Set $\boldsymbol{\epsilon}_n := \mathbf{M}_{\mathrm{reg},L_2}(n; \mathcal{G}, D_{\mathrm{reg}})$. Let P be any distribution over ξ, and choose a risk $\mathbf{R}(\hat{\pi}; g, \xi) = \mathbf{R}(\hat{\pi}; \pi_{g,\xi}, f_{g,\xi}, D, H)$ satisfying, for all $\hat{\pi} \in \Pi_{\mathbb{A}}$, the inequality

$$\mathbb{E}_{\xi \sim P} \mathbf{R}(\hat{\pi}; g, \xi) \geq K(\boldsymbol{\epsilon}_n, H) \cdot \mathbb{P}_{\mathbf{x} \sim D, \mathbf{u} \sim \hat{\pi}(\mathbf{x}, t=1)}[|\langle \pi_{g,\xi_0}(\mathbf{x}, t=1) - \mathbf{u}, \mathbf{v} \rangle| \geq \kappa \tau \boldsymbol{\epsilon}_n], \quad \text{(G.4)}$$

for some $K(H, \boldsymbol{\epsilon}_n) > 0$, where we note that the term on the right-hand side does depend on ξ_0, in view of Property G.3. Then

$$\mathbf{M}^{\mathbb{A}}(n, \mathbf{R}; \mathcal{P}, D, H) = \inf_{\mathrm{alg} \in \mathbb{A}} \sup_{g,\xi} \mathbb{E}_{\mathrm{S}_{n,H} \sim [\pi_{g,\xi}, f_{g,\xi}, D]} \mathbb{E}_{\hat{\pi} \sim \mathrm{alg}(\mathrm{S}_{n,H})} \mathbf{R}(\hat{\pi}; g, \xi) \geq K(\boldsymbol{\epsilon}_n, H)\delta. \quad \text{(G.5)}$$

Part (a) of the above proposition establishes equivalence of the minimax training risks and minimax regression risks, and shows both are equivalent to the risk incurred at the first time-step of the observed trajectories. Part (b) shows that, if $\mathbb{A}$ contains all proper algorithms, restricting to $\mathbb{A}$ does not worsen the IL training risk.

The "meat" of the proposition is in part (c). The condition states if the condition Eq. (G.4) holds for some risk $\mathbf{R}$ of interest, then the minimax risk under $\mathbf{R}$ admits a lower bound. Eq. (G.4) can be thought of a compounding error condition, which says that the average risk, over the uncertainty of the dynamics parameterized by $\xi \sim P$, scaled up by the compounding factor $K(\boldsymbol{\epsilon}_n, H)$, is at least as large as the probability that the learner makes some mistake at time $t = 1$. We note that we can simply just write $K = K(\boldsymbol{\epsilon}_n, H)$ (we don't need any uniform quantification over H and $\boldsymbol{\epsilon}_n$), but the expressing $K(\boldsymbol{\epsilon}_n, H)$ as a function of these terms clarifies its intended use. Lastly, we note that magnitude of the mistake considered inside the probability operator scales with $\tau\kappa\boldsymbol{\epsilon}_n$, where again $\boldsymbol{\epsilon}_n$ is the regression minimax risk, and the parameter κ comes from Definition F.3.

The key challenge in all of our lower bounds is to construct families of instances obeying Properties G.1 to G.3, and where there is enough variety over the dynamics (as parameterized by ξ) to force compondition Eq. (G.4) for $K(\boldsymbol{\epsilon}_n, H) \approx \boldsymbol{\epsilon}_n \cdot \exp(\Omega(H))$. We summarize here, deferring full proofs to the Appendix.

- Theorem 1.A creates instances $(\pi^\star, f)$ resembling the construction in Section 4. Here, we use ξ to encode whether or not the expert/dynamics are the system $(\mathbf{A}_1, \mathbf{K}_1)$ or $(\mathbf{A}_2, \mathbf{K}_2)$. As show in that section, uncertainty over these cases is enough to force error to compound exponentially in the horizon. The formal construction and proof are given in Appendix L, which explains the other subtleties of the argument.

- Theorem 3.A uses the same construction as Theorem 1.A. The main difference is that, for general anti-concentrated policies, only a weaker form of Eq. (G.4) can be established: namely, one of the form $K(\boldsymbol{\epsilon}_n, H) \approx \min\{\exp(H)\boldsymbol{\epsilon}_n, \boldsymbol{\epsilon}_n^{1-\Theta(1)}\}$ The argument is delicate, and given in Appendix M. In view of the benevolent gambler's ruin policy (Eq. (5.1)), we cannot hope for a larger compounding error factor $K(\boldsymbol{\epsilon}_n, H)$ when relaxing from simply-stochastic to general policies.

- Theorem 4.A, permitting unstable dynamics, uses bump-functions to embed a time-varying dynamical system, where the state-transition matrices are orthogonal matrices $\mathbf{O}_t \in \mathbb{O}(d)$, scaled by a factor $\rho > 1$. When these are drawn from a uniform prior, there is no choice of control actions which can cancel the exponential growth, because any control action will be approximately orthogonal to a randomly rotated state with high probability. The use of rotation matrices in $d > 1$ is essential. Otherwise, if only scalar systems are considered, the "non-simple" policies of Section 5 can be used to thwart compounding error. The full proof is given in Appendix N.

Part III
Deferred Proofs

This appendix begins with statements and proofs of all fundamental technical tools in Appendix H. Appendices I and J provided additional material for Section 2 and Appendix E, respectively. The remaining appendices are each dedicated to the proof of a single result. Appendix K establishes the general proof schematic, Proposition G.1, underying all results. Appendix L proves the lower bounds for simple policies (Theorem 1.A, from which Theorems 1 and 2 are stable). The proofs of Theorems 3.A and 4.A, from which Theorems 3 and 4.A are derived, are given in Appendices M and N, respectively. Appendix O demonstrates how the use of non-simple policies provably overcomes our lower bound construction. Lastly, Appendix P establishes the upper bounds (Theorem 5).

Appendix H. Technical Tools

This section outlines our technical tools. The most unique to this work are the first three sections. Appendix H.1 gives quantitative compounding error guarantees for smooth nonlinear dynamical systems with (Hurwitz)-unstable Jacobians. This generalizes arguments in Jin et al. (2017) to non-symmetric Jacobians. Appendix H.2 contains useful results regarding the stability of products of matrices. Building on these, Appendix H.3 provides convenient sufficient conditions for incremental stability of nonlinear control systems.

Moving to more standard results, Appendix H.5 recalls the seminal Paley-Zygmund and Carbery-Wright anti-concentration inequalities. These are applied to derive anti-concentration results for polynomials under the uniform distribution on the unit ball in Appendix H.6. Finally, Appendix H.7 recalls the construction of bump functions, verifying that the construction allows their derivatives to have norms which do not grow with ambient dimension.

H.1. Exponential Compounding in Unstable Systems

Definition H.1 *Given parameters $\gamma > 1, \mu \in (0,1), L \geq 1, r > 0$, we say $\mathbf{A}$ is a (γ, μ, L, r)-*matrix *if $\mathbf{A}$ admits the following block decomposition, where $\mathbf{Y}_1$ and $\mathbf{Y}_2$ are square matrices:*

$$\mathbf{A} = \begin{bmatrix} \mathbf{Y}_1 & \mathbf{W}^\top \\ \tilde{\mathbf{W}} & \mathbf{Y}_2 \end{bmatrix},$$

where, for parameters (γ, μ, L, ν), $\|\mathbf{Y}_2\|_{\mathrm{op}} \leq 1 - \mu < 0$, $\|\tilde{\mathbf{W}}\|_{\mathrm{op}} \leq L$, and $\sigma_{\min}(\mathbf{Y}_1) \geq 1 + \gamma > 1$, and $\|\mathbf{W}\|_{\mathrm{op}} \leq r$.

Proposition H.1 (Exponential Compounding for (μ, γ, L)-matrices) *Let $r > 0$, and let $F(\mathbf{x}, t)$ be a time-varying, M-smooth dynamical map such that each*

$$\mathbf{A}_t := \nabla_{\mathbf{x}} F(\mathbf{x}, t)\big|_{\mathbf{x}=0}$$

is a (γ, μ, L, r)-matrix with $\gamma \leq 1$, with the same block structure across t, and where $r = o_\star(L/\gamma\mu)$. Then, for any $\mathbf{x}_1 \in \mathbb{R}^d$, then

$$\mathbf{x}_{t+1} = F(\mathbf{x}_t, t), \quad \tilde{\mathbf{x}}_{t+1} = F(\tilde{\mathbf{x}}_t, t), \quad \tilde{\mathbf{x}}_1 = \mathbf{x}_1 \pm \epsilon \mathbf{e}_1$$

then either

$$\max_{1 \le t \le H} |\mathbf{e}_1^\top(\mathbf{x}_t - \tilde{\mathbf{x}}_t)| \ge \left(1 + \frac{\gamma}{2}\right)^{H-1} \epsilon \tag{H.1}$$

or

$$\max_{1 \le t \le H} \max\{\|\mathbf{x}_t\|, \|\mathbf{x}'_t\|\} \ge o_\star\left(\frac{1}{\mu\gamma \cdot LM}\right) \tag{H.2}$$

The proof of the above proposition is based on the following elementary recursion.

Lemma H.1 (Core Recursion) *Let α_t, β_t be two sequences satisfying $\alpha_1 = \epsilon, \beta_1 = 0$ and, for $\gamma, \mu > 0, L, r \ge 0$:*

$$\alpha_{t+1} \ge (1+\gamma)\alpha_t - r\beta_t, \quad \beta_{t+1} \le (1-\mu)\beta_t + L\alpha_t.$$

Then, if $\eta = \frac{rL}{\gamma\mu} \le 1$, we have that $\alpha_{t+1} \ge (1 + (1-\eta)\gamma)\alpha_t \ge (1 + (1-\eta)\gamma)^t\epsilon$.

Proof [Proof of Lemma H.1] We assume the inductive hypothesis that $t \mapsto \alpha_t$ is non-decreasing. Under this hypothesis, we have

$$\beta_{t+1} \le (1-\mu)\beta_t + L\alpha_t \le \underbrace{(1-\mu)^t\beta_1}_{=0} + \sum_{k=1}^{t} L(1-\mu)^{t-k}\alpha_k \le \frac{L}{\mu}\alpha_t.$$

Then,

$$\alpha_{t+1} = (1+\gamma)\alpha_t - r\beta_t \ge (1 + \gamma(1 - \frac{rL}{\gamma\mu}))\alpha_t = (1 + \gamma(1-\eta))\alpha_t.$$

which concludes the proof after recursing. ∎

We now turn to proving Proposition H.1.

Proof [Proof of Proposition H.1] Consider two sequences $\mathbf{x}_t, \tilde{\mathbf{x}}_t$ with $\delta\mathbf{x} = \mathbf{x}_t - \tilde{\mathbf{x}}_t$. Set $\nabla F(\mathbf{x})\big|_{\mathbf{x}=\mathbf{x}_0} = \mathbf{D}$. Then,

$$\begin{aligned}
\|\delta\mathbf{x}_{t+1} - \mathbf{A}_t\delta\mathbf{x}_t\| &= \|F(\mathbf{x}_t, t) - F(\tilde{\mathbf{x}}_t, t) - \nabla F(\mathbf{0}, t)\delta\mathbf{x}_t\| \\
&\le \|F(\mathbf{x}_t) - F(\tilde{\mathbf{x}}_t) - \nabla F(\mathbf{x}_t)\delta\mathbf{x}_t\| + \|\nabla F(\mathbf{x}_t, t) - \nabla F(\mathbf{0}, t)\|\|\delta\mathbf{x}_t\| \\
&\le \frac{M}{2}\|\delta\mathbf{x}_t\|^2 + \|\nabla F(\mathbf{0}, t) - \nabla F(\mathbf{x}_t, t)\|\|\delta\mathbf{x}_t\| \\
&\le \frac{M}{2}\|\delta\mathbf{x}_t\|^2 + M\|\mathbf{x}_t\|\|\delta\mathbf{x}_t\|.
\end{aligned}$$

Assume $\max\{\|\mathbf{x}_t\|, \|\delta\mathbf{x}_t\|\} \le r_0 := \frac{2r}{3M}$ for $1 \le t \le H$. Then, $\frac{M}{2}\|\delta\mathbf{x}_t\|^2 + M\|\mathbf{x}_t\|\|\delta\mathbf{x}_t\| \le \frac{3Mr}{2}\|\delta\mathbf{x}_t\|$, so there exists a matrix $\boldsymbol{\Delta}_t$ with $\|\boldsymbol{\Delta}_t\| \le \frac{3Mr_0}{2} = r$ for which

$$\delta\mathbf{x}_{t+1} = (\mathbf{A}_t + \boldsymbol{\Delta}_t)\delta\mathbf{x}_t. \tag{H.3}$$

Let $\mathbf{P}$ denote the projection onto the coordinates contained in $(\mathbf{A}_t)_{[1]}$ (recall: we assume shared block-structure across t), and define $\alpha_t = \|\mathbf{P}\delta\mathbf{x}_t\|$ and $\beta_t := \|(\mathbf{I}-\mathbf{P})^\top\delta\mathbf{x}_t\|$. Then, using the block-structure of $\mathbf{A}_t$ and conditions of Definition H.1,

$$\alpha_{t+1} \geq (1+\gamma-r)\alpha_t - 2r\beta_t, \quad \beta_{t+1} \leq (1-\mu+r)\beta_t + (r+L)\alpha_t.$$

Let us make an inductive hypothesis that α_t is non-decreasing. Then, given $r \leq \min\{\mu/2, L, \gamma/4\}$, the above simplifies to

$$\alpha_{t+1} \geq (1+\frac{3\gamma}{4})\alpha_t - 2r\beta_t, \quad \beta_{t+1} \leq (1-\mu/2)\beta_t + 2L\alpha_t.$$

The result now follows from Lemma H.1, provided that

$$\eta = \frac{2r\cdot 2L}{(\mu/2)(3\gamma/4)} \leq \frac{1}{4}, \tag{H.4}$$

which requires $r = o_\star(1/\mu\gamma L)$. ■

H.2. Stability of Products of Matrices

Definition H.2 *We say a sequence of matrices $(\mathbf{X}_1, \mathbf{X}_2, \dots)$ is (C,ρ)-stable if, for any n, $\|\mathbf{X}_n \cdot \mathbf{X}_{n-1}\cdot \mathbf{X}_j\|_{\mathrm{op}} \leq C\rho^{n-j}$ for all $1 \leq j \leq n$. Recall $\mathbf{X}$ is (C,ρ)-stable if the sequence $(\mathbf{X},\mathbf{X},\dots)$ is.*

Lemma H.2 *Let $(\mathbf{A}_i)_{i\geq 1}$ be a (C,ρ)-stable sequence of matrices. Let $(\mathbf{X}_i)_{i\geq 1}$ be a sequence of matrices such that, for each i, for which $\|\mathbf{X}_i - \mathbf{A}_i\| \leq \epsilon$. Then, $(\mathbf{X}_i)_{i\geq 1}$ is $(C, \rho + C\epsilon)$-stable. In particular, if $\rho = 1-2\gamma$ and $\epsilon = \frac{\gamma}{C}$, then $(\mathbf{X}_i)_{i\geq 1}$ is $(C, 1-\gamma)$-stable.*

Proof Throughout, let $\|| \cdot \|$ denote the operator norm. First, let us prove our lemma in the case where for all i, we have $\|\mathbf{X}_i - \nu_i\mathbf{A}_i\| \leq \epsilon$ for some $0 \leq \nu_i \leq 1$.

$$\mathbf{X}_n\mathbf{X}_{n-1}\dots\mathbf{X}_1 = \sum_{S\subset[n]} \mathbf{T}_S, \quad \mathbf{T}_S := \prod_{i=1}^t (\mathbf{I}\{i \notin S\}\mathbf{A}_i + \mathbf{I}\{i\in S\}\mathbf{\Delta}_i). \tag{H.5}$$

For $|S| = k$, this means that there are at most $k_0 \leq k+1$ (integer) subintervals of $[n]$, denoted whose endpoints we denote $a_j, b_j, 1 \leq j \leq k_0$, for which $a_j, a_{j+1}, \dots, b_j \notin S$. Furthermore, we must have $\sum_{j=1}^{k_0}(b_j - a_j) = n-k$. Lastly, we have that

$$\|\mathbf{A}_{b_j}\cdot \mathbf{A}_{b_j-1}\dots\mathbf{A}_{a_j}\| \leq C\rho^{b_j-a_j}. \tag{H.6}$$

Indeed, We therefore conclude that for $|S| = k$,

$$\begin{aligned}
\|\mathbf{T}_S\| &= \|\prod_{i=1}^t (\mathbf{I}\{i\notin S\}\mathbf{A}_i + \mathbf{I}\{i\in S\}\mathbf{\Delta}_i)\| \\
&\leq \prod_{i\in S}\|\mathbf{\Delta}_i\| \prod_{j=1}^{k_0}\|\mathbf{A}_{b_j}\cdot\mathbf{A}_{b_j-1}\dots\mathbf{A}_{a_j}\| \\
&\leq \epsilon^k C^{k_0}\prod_{j=1}^{k_0}\rho^{b_j-a_j} \\
&\leq \epsilon^k C^{k_0}\rho^{n-k} \leq C(C\epsilon)^k\rho^{n-k},
\end{aligned}$$

where above we use $k_0 \le k+1$ and $\nu_i \in [0,1]$. Therefore,

$$\|\mathbf{X}_n \mathbf{X}_{n-1} \dots \mathbf{X}_1\| = \sum_{S \subset [n]} \|\mathbf{T}_S\| \le \sum_{S \subset [n]} C(C\epsilon)^{|S|} \rho^{n-|S|} = C(\rho + C\epsilon)^n. \tag{H.7}$$

■

H.3. Stability of Linearizations Implies Incremental Stability

Lemma H.3 *Let $\rho, \epsilon > 0$ and $\rho + \epsilon < 1$. Let $(\delta\mathbf{x}_t, \delta\mathbf{u}_t)$ be any sequence for which there exist a (C, ρ)-strongly stable sequence $(\mathbf{A}_t)_{t \ge 1}$, for which*

$$\|\delta\mathbf{x}_{t+1} - \mathbf{A}_t \delta\mathbf{x}_t\| \le L\|\delta\mathbf{u}_t\| + \epsilon\|\delta\mathbf{x}_t\|. \tag{H.8}$$

Then,

(a) There exists matrices $(\mathbf{B}_t)$ and $(\mathbf{X}_t)$ with $\|\mathbf{B}_t\| \le L$ and $\|\mathbf{X}_t - \mathbf{A}_t\| \le \epsilon$ such that

$$\delta\mathbf{x}_{t+1} = \mathbf{X}_t \delta\mathbf{x}_t + \mathbf{B}_t \delta\mathbf{u}_t. \tag{H.9}$$

(b) We have

$$\|\delta\mathbf{x}_{t+1}\| \le C(\rho + C\epsilon)^t + L \sum_{1 \le j \le t} C(\rho + C\epsilon)^{t-j} \|\delta\mathbf{u}_j\|.$$

Proof We first prove part (a) at each time step t. If $\delta\mathbf{x}_{t+1} - \mathbf{A}_t\delta\mathbf{x}_t = 0$, this is holds for $\mathbf{X}_t = \mathbf{A}_t$ and $\mathbf{B}_t = 0$. By similar reasoning, it suffices to prove the case when $\delta\mathbf{x}_t, \delta\mathbf{u}_t \ne 0$. Hence, let $\mathbf{z}_t$ be a unit vector in the direction of $\delta\mathbf{x}_{t+1} - \mathbf{A}_t\delta\mathbf{x}_t$, $\mathbf{v}_t$ in the direction of $\delta\mathbf{x}_t$ and let $\mathbf{w}_t$ a unit vector in the direction of $\delta\mathbf{u}_t$ (arbitrary if $\mathbf{u}_t = 0$). Then, for some $\gamma_t \in [0,1]$, $\delta\mathbf{x}_{t+1} - \mathbf{A}_t\delta\mathbf{x}_t = \|\delta\mathbf{x}_{t+1} - \mathbf{A}_t\delta\mathbf{x}_t\|\mathbf{z}_t\gamma_t L\|\delta\mathbf{u}_t\| + \epsilon\|\delta\mathbf{x}_t\| = \gamma L\mathbf{z}_t\mathbf{w}_t^\top\delta\mathbf{u}_t + \gamma\epsilon\mathbf{z}_t\mathbf{v}_t^\top\delta\mathbf{x}_t$. Choosing $\mathbf{B}_t = \gamma L\mathbf{z}_t\mathbf{w}_t^\top$ and $\mathbf{X}_t - \mathbf{A}_t = \gamma\epsilon\mathbf{z}_t\mathbf{v}_t^\top$ proves the claim.

We now turn to part (b). Define $\mathbf{Y}_{t+1,s} := \mathbf{X}_t \cdot \mathbf{X}_{t-1} \dots \mathbf{X}_s$. with the convention $\mathbf{Y}_{t+1,t+1} = \mathbf{I}$. Part (a) implies

$$\delta\mathbf{x}_{t+1} = \mathbf{Y}_{y+1,1}\delta\mathbf{x}_1 + \sum_{i=1}^{t} \mathbf{Y}_{t+1,i+1}\mathbf{B}_i\delta\mathbf{u}_i \tag{H.10}$$

Taking the norm of each side and using Holder's inequality for ℓ_1 and ℓ_∞,

$$\|\delta\mathbf{x}_{t+1}\| \le \|\mathbf{Y}_{t+1,1}\|\|\delta\mathbf{x}_1\| + \left(\sum_{i=1}^{t} \|\mathbf{Y}_{t+1,i+1}\|\|\mathbf{B}_i\|\right)\|\delta\mathbf{u}_j\|. \tag{H.11}$$

Using Lemma H.2, we have $\|\mathbf{Y}_{y+1,s}\| \le C(\rho + C\epsilon)^{t+1-s}$, and by the above, $\|\mathbf{B}_i\| \le L$. Hence,

$$\|\delta\mathbf{x}_{t+1}\| \le C(\rho + C\epsilon)^t + L \sum_{1 \le j \le t} C(\rho + C\epsilon)^{t-j} \|\delta\mathbf{u}_j\|.$$

■

Lemma H.4 *Let $\rho, \epsilon > 0$, $L \geq 1$, and $\rho + C\epsilon < 1$. Suppose that there exists a (C, ρ)-stable matrix $\mathbf{A}$ such that*

$$\sup_{\mathbf{x},\mathbf{u}} \|\nabla_{\mathbf{x}} f(\mathbf{x}, \mathbf{u}) - \mathbf{A}\| \leq \epsilon, \quad \sup_{\mathbf{x},\mathbf{u}} \|\nabla_{\mathbf{u}} f(\mathbf{x}, \mathbf{u})\| \leq L.$$

Then, $f(\mathbf{x}, \mathbf{u})$ is (C', ρ') stable such that $\rho' = \rho + C\epsilon$ and $C' = CL$.

Proof [Proof of Lemma H.4] Let $(\mathbf{x}_i, \mathbf{u}_i)_{i \geq 1}$ and $(\mathbf{x}'_i, \mathbf{u}'_i)_{i \geq 1}$ be two sequences. Define $\delta\mathbf{x}_t := \mathbf{x}'_t - \mathbf{x}_t$ and $\delta\mathbf{u}_t$ similarly.

$$\begin{aligned}
\delta\mathbf{x}_{t+1} = \mathbf{x}'_{t+1} - \mathbf{x}_{t+1} &= f(\mathbf{x}'_t, \mathbf{u}'_t) - f(\mathbf{x}_t, \mathbf{u}_t) \\
&= \mathbf{A}\delta\mathbf{x}_t + \underbrace{\int_{\alpha=0}^{1} \nabla_{\mathbf{u}} f(\mathbf{x}_t, \alpha\mathbf{u}'_t + (1-\alpha)\mathbf{u}_t)\delta\mathbf{u}_t \mathrm{d}\alpha}_{\|\cdot\| \leq L\|\delta\mathbf{u}_t} \\
&\quad + \underbrace{\int_{\alpha=0}^{t} (\nabla_{\mathbf{x}} f(\alpha\mathbf{x}'_t + (1-\alpha)\mathbf{x}_t, \mathbf{u}'_t) - \mathbf{A}_x)\delta\mathbf{x}}_{\|\cdot\| \leq \epsilon\|\delta\mathbf{x}\|}.
\end{aligned}$$

Thus, we obtain

$$\|\delta\mathbf{x}_{t+1} - \mathbf{A}\delta\mathbf{x}_t\| \leq L\|\delta\mathbf{u}_t\| + \epsilon\|\delta\mathbf{x}_t\|.$$

The result now follows from Lemma H.3. ■

H.4. Sufficient Conditions for One-Step Controllability

Lemma H.5 *Consider a control system with $\mathbb{R}^d = \mathbb{R}^m$ and dynamics $f(\mathbf{x}, \mathbf{u}) = \phi(\mathbf{x}) + \mathbf{u} + \psi(\mathbf{u}, \mathbf{x})$, where (a) $\mathbf{x} \mapsto \phi(\mathbf{x})$ is L-Lipschitz, (b) $\psi(\mathbf{u}, \mathbf{x})\big|_{\mathbf{u}=\mathbf{0}} = \mathbf{0}$ for all $\mathbf{u}$, and for for some $\nu \in [0, 1)$, $\mathbf{u} \mapsto \psi(\mathbf{x}, \mathbf{u})$ for all $\mathbf{x}$. Then, f is $C := (1-\nu)^{-1} \max\{1, \phi(\mathbf{0}), L\}$-one-step controllable. The same also holds for dynamics $f(\mathbf{x}, \mathbf{u}) = \phi(\mathbf{x}) - \mathbf{u} + \psi(\mathbf{u}, \mathbf{x})$.*

Proof Given $\mathbf{x}, \mathbf{x}'$, define $\mathbf{x}^+ := \mathbf{x}' - \phi(\mathbf{x})$. Consider $F(\mathbf{u}) := \|\mathbf{x}^+ - \psi(\mathbf{x}, \mathbf{u}) - \mathbf{u}\|$. First, we have $F(\mathbf{u}) \geq \|\mathbf{u}\|(1-\nu) - \|\mathbf{x}'\|$, and $F(\mathbf{0}) = \|\mathbf{x}'\|$. Hence, $F(\mathbf{u})$ has all global minimizers in the set $U := \{\mathbf{u} : \|\mathbf{u}\| \leq (1-\nu)^{-1}\|\mathbf{x}^+\|\}$. Now let $\mathbf{u}^\star$ be a global minizer of $F(\mathbf{u})$ If $F(\mathbf{u}^\star) \neq 0$, then $\mathbf{v} := \mathbf{x}^+ - \psi(\mathbf{x}, \mathbf{u}) - \mathbf{u} \neq 0$. But then for $\eta \in (0, 1)$,

$$\begin{aligned}
F(\mathbf{u}^\star + \eta\mathbf{v}) &:= \|\mathbf{x}^+ - \psi(\mathbf{x}, \mathbf{u}^\star + \eta\mathbf{v}) - \mathbf{u}^\star - \eta\mathbf{v}\| \\
&= \|\mathbf{x}^+ - \psi(\mathbf{x}, \mathbf{u}^\star) - \mathbf{u} - \eta\mathbf{v}\| - \|\psi(\mathbf{x}, \mathbf{u}) - \psi(\mathbf{x}, \mathbf{u}^\star + \eta\mathbf{v})\| \\
&\leq (1-\eta)\|\mathbf{v}\| + \eta\nu\|\mathbf{v}\| \qquad (\nu\text{-Lipschitzness of } \psi \text{ in } \mathbf{u}) \\
&\leq (1 - (1-\nu)\eta)\|\mathbf{v}\| > \|\mathbf{v}\| = F(\mathbf{u}^\star),
\end{aligned}$$

contradicting optimal of $\mathbf{u}^\star$. Hence, we find that $F(\cdot)$ has a global minimum for which $F(\mathbf{u}^\star) = 0$, $\|\mathbf{u}^\star\| \leq (1-\nu)^{-1}\|\mathbf{x}^+\| = (1-\nu)^{-1}\|\mathbf{x}' - \phi(\mathbf{x})\| \leq (1-\nu)^{-1}(\phi(\mathbf{0}) + L\|\mathbf{x}\| + \|\mathbf{x}'\|)$. This concludes the proof of $f(\mathbf{x}, \mathbf{u}) = \phi(\mathbf{x}) + \mathbf{u} + \psi(\mathbf{u}, \mathbf{x})$, and it is easy to see the same argument holds for $f(\mathbf{x}, \mathbf{u}) = \phi(\mathbf{x}) - \mathbf{u} + \psi(\mathbf{u}, \mathbf{x})$. ■

H.5. Anti-Concentration Tools

Lemma H.6 (Paley-Zygmud Inequality) *Let Z be a non-negative scalar. random varible. Then,*

$$\mathbb{P}[Z \ge \theta\mathbb{E}[Z]] \ge (1-\theta)^2\frac{\mathbb{E}[Z]^2}{\mathbb{E}[Z^2]}, \quad \theta \in (0,1)$$

Lemma H.7 (Carbery-Wright Inequality, Carbery and Wright (2001)) *Let $(B, \|\cdot\|)$ be a Banach Space (e.g. $B = \mathbb{R}$ and $\|\cdot\| = |\cdot|$), and let $P : \mathbb{R}^d \to B$ be a polynomial of degree at most s. Then, for any log-concave probability measure μ on $\mathbb{R}^d$, and any $0 \le r \le q < \infty$, we have*

$$\mathbb{E}_{\mathbf{x}\sim\mu}[\|P(\mathbf{x})\|^{\frac{q}{s}}]^{\frac{1}{q}} \le C\frac{\max\{q,1\}}{\max\{r,1\}}\mathbb{E}_{\mathbf{x}\sim\mu}[\|P(\mathbf{x})\|^{\frac{r}{s}}]^{\frac{1}{r}}$$

Lemma H.8 *Let $\mathbf{x}$ be uniformly distribution on the unit ball of radius 1 in dimension d. Then $\mathbb{E}[\mathbf{x}\mathbf{x}^\top] \succeq \mathbf{I}/3$.*

Proof By rotation invariance, we have $\mathbb{E}[\mathbf{x}\mathbf{x}^\top] = \mathbb{E}[\|\mathbf{x}\|^2]\mathbf{I}$. In one dimension, $\mathbb{E}[\|\mathbf{x}\|^2] = \mathbb{E}_{U\sim[-1,1]}U^2 = \frac{1}{2}\int_{-1}^{1} x^2 \mathrm{d}x = \frac{1}{3}$. In higher dimensions, a more involved computation shows this integral is larger than $1/3$: this is the concentration of measure phenomenon, where the $\|\mathbf{x}\|^2$ concentrates more strongly around one in large dimensions. ■

H.6. Expectation-to-Uniform Bounds on the Sphere

Lemma H.9 *Let $\bar{\pi} : \mathbb{R}^d \to \mathbb{R}^m$ be M-smooth and deterministic. Let $\mathbf{x}', \mathbf{x}'$ be drawn i.i.d. on the ball of radius Δ supported on a subspace $V \subset \mathbb{R}^d$. Then, there exists a universal constant $c_\star$ such that, if*

$$\mathbb{P}_{\mathbf{x},\mathbf{x}'}[|\langle \mathbf{v}, \mathbf{K}(\mathbf{x}'-\mathbf{x}) - \mathrm{mean}[\hat{\pi}](\mathbf{x}') - \mathrm{mean}[\hat{\pi}](\mathbf{x})\rangle| \ge M\Delta^2] \le c_\star \tag{H.12}$$

Then, $\|(\mathbf{K} - \nabla\mathrm{mean}[\hat{\pi}](\mathbf{0}))\mathbf{P}_V\|_{\mathrm{op}} \le 6M\Delta\sqrt{d}$.

Proof [Proof of Lemma H.9] Let $\mathbf{K}_0 = \nabla\bar{\pi}(\mathbf{0})$. Then, by a Taylor expansion, we have $\|\mathrm{mean}[\hat{\pi}](\mathbf{x}') - \mathrm{mean}[\hat{\pi}](\mathbf{x}) - \mathbf{K}_0(\mathbf{x}'-\mathbf{x})\| \le M\Delta^2$. Hence,

$$\mathbb{P}_{\mathbf{x},\mathbf{x}'}[|\langle \mathbf{v}, (\mathbf{K}-\mathbf{K}_0)(\mathbf{x}'-\mathbf{x})\rangle| \ge 2M\Delta^2] \le c_\star. \tag{H.13}$$

By the Paley-Zygmund and Carbery-Wright Inequalities (Lemmas H.6 and H.7), $\mathbb{P}_{\mathbf{x},\mathbf{x}'}[|\langle \mathbf{v}, (\mathbf{K} - \mathbf{K}_0)(\mathbf{x}'-\mathbf{x})\rangle| \ge \frac{1}{2}\mathbb{E}[|\langle \mathbf{v}, (\mathbf{K}-\mathbf{K}_0)(\mathbf{x}'-\mathbf{x})\rangle|^2]^{1/2} \ge c_0$ for some universal constant c_0. Hence if $c_\star \le c_0$, we must have that $\mathbb{E}[|\langle \mathbf{v}, (\mathbf{K}-\mathbf{K}_0)(\mathbf{x}'-\mathbf{x})\rangle|^2]^{1/2} \le 4M\Delta^2$. Because $\mathbf{x}, \mathbf{x}'$ are uniformly distributed on the unit ball restricted to V, by rescaling and invoking Lemma H.8, their covariances are at least $\Delta\mathbf{P}_V/3d$. Adding these covariances of the independent variables, we find $\|\mathbf{v}^\top(\mathbf{K}-\mathbf{K}_0)\mathbf{P}_V\| \le 6\Delta\sqrt{d}$. Taking the supremum over $\mathbf{v}$ concludes. ■

Lemma H.10 (Consequence of Carbery-Wright) *Let $\mathcal{D}$ be a log-concave distribution on $\mathbb{R}^d$, and let $G(\mathbf{x})$ be a polynomial of degree at most p. Then, for some universal constant $C \ge 1$,*

$$\mathbb{E}_{\mathbf{x}\sim\mathcal{D}}[\|G(\mathbf{x})\|^2] \le C^p\left(\mathbb{E}_{\mathbf{x}\sim\mathcal{D}}[\|G(\mathbf{x})\|]\right)^2 \tag{H.14}$$

Proof The Carberry-Wright inequality, Lemma H.7, with $q = 2p$ and $r = p$ yields

$$\mathbb{E}_{\mathbf{x}\sim\mathcal{D}}[\|G(\mathbf{x})\|^2]^{1/(2p)} \le C\frac{\max\{p,1\}}{\max\{q,1\}}\mathbb{E}_{\mathbf{x}\sim\mathcal{D}}[\|G(\mathbf{x})\|]^{1/p} = 2C\mathbb{E}_{\mathbf{x}\sim\mathcal{D}}[\|G(\mathbf{x})\|]^{1/p}, \tag{H.15}$$

where C is a universal constant. Taking the $2p$-th power of both sides and multiplying C by a factor of two concludes. ■

Lemma H.11 (Derivative Bounds on the Ball) *Let $p \ge 2$, and let G be a function satisfying*

$$\|G(\delta\mathbf{x}) - \sum_{\ell=0}^{p-1}\nabla^{(\ell)}G(\mathbf{0})\circ(\delta\mathbf{x}^{\otimes\ell})\| \le \frac{M}{p!}\|\delta\mathbf{x}\|^p. \tag{H.16}$$

and suppose that

$$\mathbb{E}_{\mathbf{x}\sim\Delta\cdot\mathcal{B}_d(1)}\|G(\mathbf{x})\| \le \epsilon. \tag{H.17}$$

Then, for all $0 \le \ell \le d-1$, $\|\nabla^{(\ell)}G(\mathbf{0})\|_{\mathrm{F}} \le C^p(2d)^{\ell/2}(\epsilon\Delta^{-\ell} + \Delta^{p-\ell}/(p!))$. In particular, if p is taken to be a universal constant, we have

$$\|\nabla^{(\ell)}G(\mathbf{0})\|_{\mathrm{F}} \le O(d^{\ell/2}(\epsilon\Delta^{-\ell} + \Delta^{p-\ell})) \tag{H.18}$$

The result generalizes, up to universal constant multiplicative factors, to the case when $\mathbf{x}$ has the distribution of $\mathbf{x}^1 - \mathbf{x}^2$, where $\mathbf{x}^1, \mathbf{x}^2$ are drawn independently from $\Delta\mathcal{B}_1(d)$.

Proof The two facts about the distribution we use on the sphere are that it is log concave, enabling the use of Carbery-Wright, that the signs of each coordinate are independent and symmetric, and that its covariance has a particular form. For the final statement of the lemma, we note that $\mathbf{x}^1 - \mathbf{x}^2$ is log concave (log concavity is preserved under convolution), its coordinates still have independent signs, that and its covariance is equal to twice that of $\mathbf{x} \sim \Delta\mathcal{B}_1(d)$. Hence, we prove the statement only for the distribution of $\mathbf{x}$.

Define the function

$$G_0(\mathbf{x}) = \sum_{\ell=0}^{p-1}\nabla^{(\ell)}G(\mathbf{0})\circ(\delta\mathbf{x}^{\otimes\ell}), \quad \epsilon_0 = \epsilon + \frac{M}{p!}\Delta^p. \tag{H.19}$$

Then, G_0 is a polynomial of degree at most $p-1$, and $\mathbb{E}[\|G_0(\mathbf{x})\|] \le \epsilon_0$. From Lemma H.10 and the fact that the uniform measure on the convex body $\mathcal{B}_d(1)$ is log-concanve, it follows that $\mathbb{E}[\|G_0(\mathbf{x})\|^2] \le C^{p-1}\epsilon_0^2 \le C^p\epsilon_0^2$ for some universal $C \ge 1$. Notice further that if $\mathbf{x} \sim \mathcal{B}_d(1)$, $\mathbb{E}[\mathbf{x}_i^2] \ge 1/(2d)$. Thus Lemma H.13 with $\nu = 2$ below implies that $\|\nabla^{(\ell)}G(\mathbf{0})\|_{\mathrm{F}} = \|\nabla^{(\ell)}G_0(\mathbf{0})\|_{\mathrm{F}} \le C^p\epsilon_0\Delta^{-\ell}(2d)^{\ell/2} = C^p(2d)^{\ell/2}(\epsilon\Delta^{-\ell} + \Delta^{p-\ell}/(p!))$. This concludes the proof. ■

Lemma H.12 *Let $p \ge 2$, and let G be a function satisfying*

$$\|G(\delta\mathbf{x}) - \sum_{\ell=0}^{p-1}\nabla^{(\ell)}G(\mathbf{0})\circ(\delta\mathbf{x}^{\otimes\ell})\| \le \frac{M}{p!}\|\delta\mathbf{x}\|^p. \tag{H.20}$$

and suppose that

$$\mathbb{P}[\|G(\mathbf{x})\| \geq \epsilon] < \frac{1}{4C^{2p}}, \tag{H.21}$$

where C is the universal constant as in Carberry Wright. Then, the results of Lemma H.11 hold up to universal multiplicative constants. Note further that if p is a universal constant, then we can take $\frac{1}{4C^{2p}}$ to be as well.

Proof [Proof of Lemma H.12] Again

$$G_0(\mathbf{x}) = \sum_{\ell=0}^{p-1} \nabla^{(\ell)} G(\mathbf{0}) \circ (\delta\mathbf{x}^{\otimes \ell}), \quad \epsilon_0 = \epsilon + \frac{M}{p!}\Delta^p. \tag{H.22}$$

Then $\mathbb{P}[\|G_0(\mathbf{x})\| \geq \epsilon_0] < \frac{1}{4C^{2p}},$. Now, suppose that $\mathbb{E}[\|G_0(\mathbf{x})\|] \geq 2\epsilon_0$. By Carbery Wright (Lemma H.10), $\mathbb{E}[\|G_0(\mathbf{x})\|^2]^{1/2} \leq C^p\epsilon_1$. By the Paley-Zygmud inequality (Lemma H.6),

$$\frac{1}{4C^{2p}} > s \geq \mathbb{P}[\|G_0(\mathbf{x})\| \geq \epsilon_0] \geq \frac{1}{4}\frac{\mathbb{E}[\|G_0(\mathbf{x})\|^2]}{\mathbb{E}[\|G_0(\mathbf{x})\|]^2} \geq \frac{1}{4C^{2p}}. \tag{H.23}$$

Hence, it must follow that in fact $\mathbb{E}[\|G_0(\mathbf{x})\|] \leq 2\epsilon_0$. The bound now follows by repeating the arguments of Lemma H.11.

■

Lemma H.13 *Let $G : \mathbb{R}^d \to \mathbb{R}^m$ be a polynomial satisfying*

$$G(\delta\mathbf{x}) = \sum_{\ell=0}^{p-1} \nabla^{(\ell)} G(\mathbf{0}) \circ (\delta\mathbf{x}^{\otimes \ell}). \tag{H.24}$$

where $\nabla^{(\ell)}$ is the ℓ-th order derivative. Let $\mathcal{D}$ be a distribution supported on $\mathcal{B}_d(1)$, such that $\mathbb{E}_{\mathbf{x}\sim\mathcal{D}}[(\mathbf{x}_i)^2] \geq 1/(\nu d)$, and such that the signs of each of its coordinates are symmetric and independent. Suppose further that

$$\mathbb{E}_{\delta\mathbf{x}\sim\Delta\mathcal{D}}\|G(\delta\mathbf{x})\|^2 \leq \epsilon^2. \tag{H.25}$$

Then, letting $\|\cdot\|_{\mathrm{F}}$ denote the (tensor) Frobenius norm,

$$\sum_{\ell=0}^{p-1} \Delta^{2\ell} d^{-\ell} \|\nabla^{(\ell)} G(\mathbf{0})\|_{\mathrm{F}}^2 \leq \epsilon^2. \tag{H.26}$$

From this, it follows that

- *For all $0 \leq \ell \leq p-1$, $\|\nabla^{(\ell)} G(\mathbf{0})\|_{\mathrm{F}} \leq \epsilon\Delta^{-\ell}(\nu d)^{\ell/2}$*
- *For all $\delta\mathbf{x} \in \Delta \cdot \mathcal{S}^{d-1}$, $\|G(\delta\mathbf{x})\| \leq \epsilon \sum_{\ell=0}^{p-1} (\nu d)^{\ell/2}$.*

Proof We have

$$\begin{aligned}
\epsilon^2 \geq \mathbb{E}\left[\left\|\sum_{\ell=0}^{p-1} \nabla^{(\ell)} G(\mathbf{0}) \circ (\delta\mathbf{x}^{\otimes \ell})\right\|^2\right] &= \mathbb{E}\left[\sum_{\ell,\ell'=0}^{p-1} \left\langle (\nabla^{(\ell)} G(\mathbf{0}) \circ (\delta\mathbf{x}^{\otimes \ell})), (\nabla^{(\ell')} G(\mathbf{0}) \circ (\delta\mathbf{x}^{\otimes \ell'}))\right\rangle\right] \\
&= \mathbb{E}\left[\sum_{\ell}^{p-1} \left\langle (\nabla^{(\ell)} G(\mathbf{0}) \circ (\delta\mathbf{x}^{\otimes \ell})), (\nabla^{(\ell)} G(\mathbf{0}) \circ (\delta\mathbf{x}^{\otimes \ell}))\right\rangle\right] \\
&= \sum_{\ell}^{p-1} \mathbb{E}\left[\|\nabla^{(\ell)} G(\mathbf{0}) \circ (\delta\mathbf{x}^{\otimes \ell}))\|^2\right] \\
&= \sum_{\ell}^{p-1} \mathbb{E}\left[\|\nabla^{(\ell)} G(\mathbf{0}) \circ (\delta\mathbf{x}^{\otimes \ell}))\|^2\right] \\
&= \sum_{\ell}^{p-1} \sum_{s=1}^{m} \mathbb{E}\left[\left(\mathbf{e}_s^\top (\nabla^{(\ell)} G(\mathbf{0})) \circ (\delta\mathbf{x}^{\otimes \ell}))\right)^2\right],
\end{aligned}$$

For each ℓ, s, $\mathbf{e}_s^\top \nabla^{(\ell)} G(\mathbf{0}))$ is some tensor T with entries $T_{i_{1:\ell}}$, where $i_{1:\ell} = (i_1, \ldots, i_\ell) \in [d]^\ell$. Denote the entries of $\delta\mathbf{x}$ by $\delta\mathbf{x}[j]$. Then $\mathbb{E}[(T \circ \delta\mathbf{x}^{\otimes \ell}))^2] = \sum_{i_{1:\ell}, i'_{1:\ell}} (T_{i_{1:\ell}})(T_{i'_{1:\ell}}) \mathbb{E}[\prod_{j=i_1,\ldots,i_\ell,i'_1,\ldots,i'_\ell} \delta\mathbf{x}[j]]$. Because $\delta\mathbf{x}[j] \mid \delta\mathbf{x}[j'], j' \neq j$ is symmetric, each term $\mathbb{E}[\prod_{j=i_1,\ldots,i_\ell,i'_1,\ldots,i'_\ell} \delta\mathbf{x}[j]]$ either vanishes, or is positive. Consequently, we can lower bound $\mathbb{E}[(T \circ \delta\mathbf{x}^{\otimes \ell}))^2]$ by the sum over only terms where $i_{1:\ell} = i'_{1:\ell}$, and for these terms, $\mathbb{E}[\prod_{j=i_1,\ldots,i_\ell,i'_1,\ldots,i'_\ell} \delta\mathbf{x}[j]] = \mathbb{E}[\prod_{j=i_1,\ldots,i_\ell} \delta\mathbf{x}[j]^2] \geq \prod_{j=i_1,\ldots,i_\ell} \mathbb{E}[\delta\mathbf{x}[j]^2] = \Delta^{2\ell}(\nu d)^{-\ell}$. We conclude that

$$\mathbb{E}[(T \circ \delta\mathbf{x}^{\otimes \ell}))^2] \geq \Delta^{2\ell}(\nu d)^{-\ell} \sum_{i_{1:\ell}} (T_{i_{1:\ell}})^2. \tag{H.27}$$

Thus, we conclude that

$$\epsilon^2 \geq \sum_{\ell}^{p-1} \sum_{s=1}^{m} d^{-\ell} \sum_{i_{1:\ell}} (\mathbf{e}_s^\top \nabla^{(\ell)} G(\mathbf{0}))_{i_{1:\ell}}^2 = \sum_{\ell}^{p-1} \Delta^{2\ell}(\nu d)^{-\ell} \|\nabla^{(\ell)} G(\mathbf{0})\|_{\mathrm{F}}^2. \tag{H.28}$$

The first consequence statement follows from the above, the fact that all summands are non-negative, and the elementary inequality $\sqrt{x+y} \leq \sqrt{x} + \sqrt{y}$. To prove the second statement of the lemma, we use the Taylor remainder bound of G and

$$\begin{aligned}
\|G(\delta\mathbf{x})\| &= \left\|\sum_{\ell=0}^{p-1} \nabla^{(\ell)} G(\mathbf{0}) \circ (\delta\mathbf{x}^{\otimes \ell})\right\| \\
&\leq \sum_{\ell=0}^{p-1} \left\|\nabla^{(\ell)} G(\mathbf{0})\right\|_{\mathrm{F}} \Delta^\ell \\
&\leq \sum_{\ell=0}^{p-1} \sqrt{\left\|\nabla^{(\ell)} G(\mathbf{0})\right\|_{\mathrm{F}} \Delta^\ell (\nu d)^{-\ell})} (\nu d)^{\ell/2} \\
&\leq \epsilon \sum_{\ell=0}^{p-1} (\nu d)^{\ell/2}.
\end{aligned}$$

■

H.7. The existence of bump functions.

Lemma H.14 (Existence of Bump Functions) *For any $k \in \mathbb{N}$, there exists a C^∞ function* $\mathrm{bump}_k(\mathbf{z}) : \mathbb{R}^k \to \mathbb{R}$*, called an* bump function*, sastisfying* $\mathrm{bump}_k(\mathbf{z}) = 1$ *if and only if* $\|\mathbf{z}\| \le 1$*,* $\mathrm{bump}_k(\mathbf{z}) = 0$ *if and only if* $\|\mathbf{z}\| \ge 2$*. And, for each* $p \ge 1$*,* $\|\nabla^p \mathrm{bump}_k(\mathbf{z})\|_{\mathrm{op}} \le c_p$*, where* $\|\cdot\|_{\mathrm{op}}$ *denotes the tensor-operator norm, and* c_p *is a constant independent of* k *but depending on* p*. Finally,* $\nabla^p \mathrm{bump}_k(\mathbf{z}) = 0$ *for all* $\mathbf{z} : \|\mathbf{z}\| \ge 2$*.*

Proof [Proof of Lemma H.14] The proof is standard, and included for completeness. Consider the function $\phi(u) = \exp(1 - \frac{1}{u})$ defined on $(0, 1)$, and define

$$\psi(u) = \begin{cases} 0 & u \le 0 \\ 1 & u \ge 1 \\ (1 - \phi(1-u))\phi(u) & u \in (0,1) \end{cases} \tag{H.29}$$

We define

$$\mathrm{bump}_k(\mathbf{z}) := \psi(2 - \|\mathbf{z}\|^2). \tag{H.30}$$

By construction $\mathrm{bump}_k(\mathbf{z}) = 1$ if and only if $\|\mathbf{z}\| \le 1$, $\mathrm{bump}_k(\mathbf{z}) = 0$ if and only if $\|\mathbf{z}\| \ge 2$. For the second, clearly $\psi(u)$ is C^∞ for $u > 0, u < 0$ and $u \in (0, 1)$. It is easy to check continuity at $u \in \{0, 1\}$, and by using the fact that the derivatives of $\phi(u)$ take the form $g(1/u)\phi(u)$, where $g(u)$ is a polynomial, one can check that all derivatives of $\psi(u)$ vanish at $u \in \{0, 1\}$; this establishes that ψ is C^∞. As $\mathbf{z} \mapsto 2 - \|\mathbf{z}\|^2$ is also C^∞, we obtain that $\mathrm{bump}_k(\mathbf{z})$ is as well.

To bound $\|\nabla^p \mathrm{bump}_k(\mathbf{z})\|_{\mathrm{op}}$, we observe that $\nabla^p \mathrm{bump}_k(\mathbf{z})$ is a symmetric p-tensor, and hence its operator norm is equal to the largest value of $|\langle \nabla^p \mathrm{bump}_k(\mathbf{z}), \mathbf{v}^{\otimes p}\rangle|$ where $\mathbf{v} \in \mathcal{B}_k(1)$. Note that $\langle \nabla^p \mathrm{bump}_k(\mathbf{z}), \mathbf{v}^{\otimes p}$ is just the order-p directional derivative in the direction p, and thus

$$\begin{aligned} \|\nabla^p \mathrm{bump}_k(\mathbf{z})\|_{\mathrm{op}} &\le \sup_{\mathbf{v} \in \mathcal{B}_k(1)} \frac{\mathrm{d}}{\mathrm{d}s^p} (\psi \circ (1 - \|\mathbf{z} + u\mathbf{v}\|^2)) \\ &\le \sup_{\mathbf{v} \in \mathcal{B}_k(1)} \frac{\mathrm{d}}{\mathrm{d}s^p} (\psi(1 - \|\mathbf{z}\|^2 + 2u\langle \mathbf{z}, \mathbf{v}\rangle + u^2\|\mathbf{v}\|^2)). \end{aligned}$$

Using this expression, one can show that the maximial derivative does not depend on the dimension k. Note that it is also uniformly bounded because the derivatives of ψ are. ■

Appendix I. Appendix for Section 2

I.1. Trajectory Distance

We begin by defining a canonical coupling (joint distribution) between $\hat{\pi}$ and $\pi^\star$ trajectories.

Definition I.1 (Canonical Coupling) *Let* $\hat{\pi}$ *be arbitrary and* $\pi^\star, f$ *be deterministic. We define the canonical coupling of* $(\mathbb{P}_{\hat{\pi},f,D}, \mathbb{P}_{\pi^\star,f,D})$*, denoted by* $\mathbb{P}_{\hat{\pi},\pi^\star,f,D}$ *(resp.* $\mathbb{E}_{\hat{\pi},\pi^\star,f,D}$*) as the distribution of (resp. expectation over) the random variables* $(\mathbf{x}^\star_{1:H}, \mathbf{u}^\star_{1:H}, \hat{\mathbf{x}}_{1:H}, \hat{\mathbf{u}}_{1:H})$*, where*

(a) *Both trajectories have same initial state* $\mathbf{x}_1^\star = \hat{\mathbf{x}}_1 \sim D$

(b) *Inputs* $\mathbf{u}_t^\star = \pi^\star(\mathbf{x}_t^\star)$ *are chosen according to* $\pi^\star$*, and inputs* $\hat{\mathbf{u}}_t \sim \hat{\pi}(\hat{\mathbf{x}}_t)$ *are chosen by* $\hat{\pi}$ *with independent randomness at each time step*

(c) *Both* $\mathbf{x}_{t+1}^\star = f(\mathbf{x}_t^\star, \mathbf{u}_t^\star)$ *and* $\hat{\mathbf{x}}_{t+1} = f(\hat{\mathbf{x}}_t, \hat{\mathbf{u}}_t)$ *evolve according to (deterministic) the system dynamics.*

In terms of this, we define the L_1-trajectory risk as

$$\mathbf{R}_{\mathrm{traj},L_1}(\hat{\pi}; \pi^\star, f, D, H) = \mathbb{E}_{\hat{\pi},\pi^\star,f,D}\left[\sum_{t=1}^{H} \min\left\{\|\mathbf{x}_t^\star - \hat{\mathbf{x}}_t\| + \|\mathbf{u}_t^\star - \hat{\mathbf{u}}_t\|,\ 1\right\}\right]. \tag{I.1}$$

Above, we clip the expectation to a maximum of one to avoid pathologies of unbounded rewards. We now show that $\mathbf{R}_{\mathrm{traj},L_1} \ge \sup_{\mathrm{cost}\in\mathcal{C}_{\mathrm{Lip}}} \mathbf{R}_{\mathrm{cost}}$.

Lemma I.1 $\mathbf{R}_{\mathrm{traj},L_1}(\hat{\pi}; \pi^\star, f, D, H) \ge \sup_{\mathrm{cost}\in\mathcal{C}_{\mathrm{Lip}}} \mathbf{R}_{\mathrm{cost}}(\hat{\pi}; \pi^\star, f, D, H)$. *This bound holds even if we inflate* $\mathcal{C}_{\mathrm{Lip}}$ *to include all time-varying costs of the form* $\mathrm{cost}(\mathbf{x}_{1:H}, \mathbf{u}_{1:H}) = \sum_h \mathrm{cost}_h(\mathbf{x}_h, \mathbf{u}_h)$*, where each* cost_h *is* 1*-Lipschitz and bounded in* $[0, 1]$.

Proof Suppose that $\mathrm{cost}(\mathbf{x}_{1:H}, \mathbf{u}_{1:H}) = \sum_h \mathrm{cost}_h(\mathbf{x}_h, \mathbf{u}_h)$, where each cost_h is 1-Lipschitz, and bounded in $[0, 1]$. We have

$$\begin{aligned}
\mathbf{R}_{\mathrm{cost}}(\hat{\pi}; \pi^\star, f, D, H) &:= \mathbb{E}_{\hat{\pi},f,D}\left[\mathrm{cost}(\mathbf{x}_{1:H}, \mathbf{u}_{1:H})\right] - \mathbb{E}_{\pi^\star,f,D}\left[\mathrm{cost}(\mathbf{x}_{1:H}, \mathbf{u}_{1:H})\right] \\
&= \mathbb{E}_{\hat{\pi},\pi^\star,f,D}\left[\mathrm{cost}(\hat{\mathbf{x}}_{1:H}, \hat{\mathbf{u}}_{1:H}) - \mathrm{cost}(\mathbf{x}_{1:H}^\star, \mathbf{u}_{1:H}^\star)\right] \\
&= \sum_{h=1}^{H} \mathbb{E}_{\hat{\pi},\pi^\star,f,D}\left[\mathrm{cost}_h(\hat{\mathbf{x}}_h, \hat{\mathbf{u}}_h) - \mathrm{cost}(\mathbf{x}_h^\star, \mathbf{u}_h^\star)\right] \\
&= \sum_{h=1}^{H} \mathbb{E}_{\hat{\pi},\pi^\star,f,D}\left[\min\{1, \mathrm{cost}_h(\hat{\mathbf{x}}_h, \hat{\mathbf{u}}_h) - \mathrm{cost}_h(\mathbf{x}_h^\star, \mathbf{u}_h^\star)\}\right] \quad (\mathrm{cost}_h \in [0, 1]) \\
&= \sum_{h=1}^{H} \mathbb{E}_{\hat{\pi},\pi^\star,f,D}\left[\min\{1, \|\mathbf{x}_h^\star - \hat{\mathbf{x}}_h\| + \|\mathbf{u}_h^\star - \hat{\mathbf{u}}_h\|\}\right] \quad (\text{cost is 1-Lipschitz}) \\
&=: \mathbf{R}_{\mathrm{traj},L_1}(\hat{\pi}; \pi^\star, f, D, H).
\end{aligned}$$

■

I.2. Guarantees under Q-function regularity

Recall the definition of the Q-functions,

$$Q_{h;\hat{\pi},f,\mathrm{cost},H}(\mathbf{x}, \mathbf{u}) := \mathrm{cost}_h(\mathbf{x}, \mathbf{u}) + \sum_{h'>h} \mathbb{E}_{\hat{\pi},f}\left[\mathrm{cost}_{h'}(\mathbf{x}_{h'}, \mathbf{u}_{h'}) \mid (\mathbf{x}_h, \mathbf{u}_h) = (\mathbf{x}, \mathbf{u})\right].$$

In what follows, we fix $(\mathrm{cost}, \hat{\pi}, f, H)$, and adopt the shorthand $Q_h := Q_{h;f,\hat{\pi},\mathrm{cost},H}$. The evaluation performance of a policy $\hat{\pi}$ can be evaluated via the celebrated *performance difference lemma* (Kakade, 2003):

Lemma I.2 *Fix an additive cost* $\mathrm{cost}(\mathbf{x}_{1:H}, \mathbf{u}_{1:H}) = \sum_{h=1}^{H} \mathrm{cost}_h(\mathbf{x}_h, \mathbf{u}_h)$*, and let* $Q_h := Q_{h;f,\hat{\pi},\mathrm{cost},H}$*. Then,*

$$\mathbf{R}_{\mathrm{cost}}(\hat{\pi}; \pi^\star, f, D, H) = \sum_{h=1}^{H} \mathbb{E}_{\pi^\star, f, D} \mathbb{E}_{\hat{\mathbf{u}}_h \sim \hat{\pi}(\mathbf{x}_h^\star)} \left[Q_h(\mathbf{x}_h^\star, \hat{\mathbf{u}}_h) - Q_h(\mathbf{x}_h^\star, \mathbf{u}_h^\star) \right].$$

We use the performance difference lemma to establish the claims of Section 2.1.

Lemma I.3 *Suppose each* Q_h *is* L*-Lipschitz. Then,*

$$\mathbf{R}_{\mathrm{eval}}(\hat{\pi}; \pi^\star, f, D, H) \le L \cdot \mathbf{R}_{\mathrm{expert}, L_1}(\hat{\pi}; \pi^\star, f, D, H) \le L \cdot \mathbf{R}_{\mathrm{expert}, L_p}(\hat{\pi}; \pi^\star, f, D, H), \quad \forall p \ge 1.$$

Proof From Lemma I.2,

$$\begin{aligned}
\mathbf{R}_{\mathrm{cost}}(\hat{\pi}; \pi^\star, f, D, H) &= \sum_{h=1}^{H} \mathbb{E}_{\pi^\star, f, D} \mathbb{E}_{\hat{\mathbf{u}}_h \sim \hat{\pi}(\mathbf{x}_h^\star)} \left[Q_h(\mathbf{x}_h^\star, \hat{\mathbf{u}}_h) - Q_h(\mathbf{x}_h^\star, \mathbf{u}_h^\star) \right] \\
&\le \sum_{h=1}^{H} \mathbb{E}_{\pi^\star, f, D} \mathbb{E}_{\hat{\mathbf{u}}_h \sim \hat{\pi}(\mathbf{x}_h^\star)} \left[L \cdot \|\hat{\mathbf{u}}_h - \mathbf{u}_h^\star\| \right] && (Q_h \text{ is } L\text{-Lipschitz}) \\
&= L \mathbf{R}_{\mathrm{expert}, L_1}(\hat{\pi}; \pi^\star, f, D, H).
\end{aligned}$$

The second inequality follows from Jensen's inequality. ■

Lemma I.4 *Recall* $\mathbf{R}_{\mathrm{train},\{0,1\}}(\hat{\pi}; \pi^\star, f, H) := \sum_{h=1}^{H} \mathbb{E}_{\pi^\star, f, D} \mathbb{E}_{\hat{\mathbf{u}} \sim \hat{\pi}(\mathbf{x}_h^\star)} I\left\{ \mathbf{u}_h^\star \ne \hat{\mathbf{u}}_h \right\}$*. Then, if each* $Q_h \in [0, B]$*, we have*

$$\mathbf{R}_{\mathrm{cost}}(\hat{\pi}; \pi^\star, f, D, H) \le B \cdot \mathbf{R}_{\mathrm{train},\{0,1\}}(\hat{\pi}; \pi^\star, f, D, H).$$

Proof Appealing again to the performance difference lemma,

$$\begin{aligned}
\mathbf{R}_{\mathrm{cost}}(\hat{\pi}; \pi^\star, f, D, H) &= \sum_{h=1}^{H} \mathbb{E}_{\pi^\star, f, D} \mathbb{E}_{\hat{\mathbf{u}}_h \sim \hat{\pi}(\mathbf{x}_h^\star)} \left[Q_h(\mathbf{x}_h^\star, \hat{\mathbf{u}}_h) - Q_h(\mathbf{x}_h^\star, \mathbf{u}_h^\star) \right] \\
&\le \sum_{h=1}^{H} \mathbb{E}_{\pi^\star, f, D} \mathbb{E}_{\hat{\mathbf{u}}_h \sim \hat{\pi}(\mathbf{x}_h^\star)} \left[B \cdot \mathbf{I}\{\hat{\mathbf{u}}_h \ne \mathbf{u}_h^\star\} \right] && (Q_h \in [0, B]) \\
&= B \cdot \mathbf{R}_{\mathrm{train},\{0,1\}}(\hat{\pi}; \pi^\star, f, D, H).
\end{aligned}$$

■

I.3. Proof of Lemma 2.1

Lemma 2.1 *Suppose that* $(f, \hat{\pi})$ *is* (C, ρ)*-E-IISS and* $\hat{\pi}$ *is* $L_{\hat{\pi}}$*-Lipschitz. Then, for any* $\mathrm{cost} \in \mathcal{C}_{\mathrm{lip}}$*,* $Q_{h;f,\hat{\pi},\mathrm{cost}}$ *is* $\frac{C}{1-\rho}(2+L_{\hat{\pi}})$*-Lipschitz. Moreover, for any* $D \in \mathbf{\Delta}(\mathbb{X})$ *and* $H \ge 1$*, and any* $\mathrm{cost} \in \mathcal{C}_{\mathrm{Lip}}$*,* $\mathbf{R}_{\mathrm{cost}}(\hat{\pi}; \pi^\star, f, D, H) \le \frac{C}{1-\rho}(2 + L_{\hat{\pi}}) \cdot \mathbf{R}_{\mathrm{expert}, L_1}(\hat{\pi}; \pi^\star, f, D, H)$.

Proof Consider any $\mathbf{x}_h$ and perturbation $\delta\mathbf{u}$. Let $\mathbf{x}'_h = \mathbf{x}_h$ and,

$$\begin{aligned} \mathbf{x}_{h+1} &= f(\mathbf{x}_h, \pi^\star(\mathbf{x}_h^\star)), \quad \mathbf{x}'_{h+1} = f(\mathbf{x}'_h, \hat{\pi}(\mathbf{x}'_h) + \delta\mathbf{u}), \\ \mathbf{x}'_{h'+1} &= f(\mathbf{x}'_{h'}, \pi^\star(\mathbf{x}'_{h'})), \quad \mathbf{x}'_{h'+1} = f(\mathbf{x}'_{h'}, \hat{\pi}(\mathbf{x}'_{h'})), \qquad \forall h' > h. \end{aligned}$$

Using that $\hat{\pi}$ is (C, ρ)-E-IISS we have

$$\begin{aligned} \sum_{h' \geq h} \|\mathbf{x}'_{h'} - \mathbf{x}_{h'}\| &\leq \sum_{h' \geq h} C\rho^h \|\delta\mathbf{u}\| \\ &\leq \frac{C}{1-\rho}\|\delta\mathbf{u}\|. \end{aligned}$$

Then, provided $\hat{\pi}$ is $L_{\hat{\pi}}$-Lipschitz, expanding the definition of Q and applying triangle inequality,

$$\begin{aligned} |Q_{h;f,\hat{\pi},\mathrm{cost},H}(\mathbf{x}, \mathbf{u} + \delta\mathbf{u}) - Q_{h;f,\hat{\pi},\mathrm{cost},H}(\mathbf{x}, \mathbf{u})| &\leq \|\delta\mathbf{u}\| + \sum_{h'=h+1}^{H} |\mathrm{cost}(\mathbf{x}'_h, \hat{\pi}(\mathbf{x}'_h)) - \mathrm{cost}(\mathbf{x}_h, \hat{\pi}(\mathbf{x}_h))| \\ &\leq \|\delta\mathbf{u}\| + (1 + L_{\hat{\pi}}) \sum_{h'>h} \|\mathbf{x}'_h - \mathbf{x}_h\| \\ &\leq \frac{C}{1-\rho}(2 + L_{\hat{\pi}})\|\delta\mathbf{u}\|. \end{aligned}$$

The result for $\mathbf{R}_{\mathrm{cost}}$ follows by Lemma I.3. ■

I.4. Impossibility of Estimation in the $\{0, 1\}$-Loss (Section 2.2)

Remark I.1 (Hellinger Distance, Total Variation Distnace, and the $\{0, 1\}$ loss) . *Let P, Q be probability distributions over the same probability space Ω, with common densities p, q with respect to a common base measure μ. We recall that the Hellinger and total variation distances, respectively, are given by*

$$\mathrm{d}_{\mathrm{HEL}}(P, Q)^2 = \frac{1}{2}\int (\sqrt{p(\omega)} - \sqrt{q(\omega)})^2 \mathrm{d}\mu(\omega), \quad \mathrm{d}_{\mathrm{TV}}(P, Q) = \frac{1}{2}\int |p(\omega) - q(\omega)|\mathrm{d}\mu(\omega). \tag{I.2}$$

By the LeCam's inequality (Tsybakov, 1997, Lemma 2.4), the above are qualitatively equivalent.

$$\frac{1}{2}\mathrm{d}_{\mathrm{HEL}}(P, Q)^2 \leq \mathrm{d}_{\mathrm{TV}}(P, Q) \leq \mathrm{d}_{\mathrm{HEL}}(P, Q) \tag{I.3}$$

Moreover, in the special case where P is a Dirac distribution supported on $\omega_p \in \Omega$, we have that

$$\mathrm{d}_{\mathrm{TV}}(P, Q) = \mathbb{P}_{\omega_q \sim Q}[\omega_q \neq \omega_p] = \mathbb{E}_{\omega_q \sim Q}[\mathbf{I}\{\omega_q \neq \omega_p\}] \tag{I.4}$$

In the case where P is the conditional distribution of $\pi^\star(x)$ given x, which is a Dirac distribution supported at $\pi^\star(x)$, we therefore see that the $\{0, 1\}$ loss is precisely equal to the total variation distance $\mathbb{E}_{u\sim\hat{\pi}(x)}[\mathbf{I}\{u \neq \pi^\star(x)\}] = \mathrm{d}_{\mathrm{TV}}(\hat{\pi}(x), \pi^\star(x))$.

Proposition I.1 (Impossibility of 0/1 and Information-Theoretic Estimation) *Let $\mathcal{G}$ denote the class of* 1*-Lipschitz functions from* $[0,1] \to [-1,1]$. *Then, and let* D_{reg} *denote the uniform distribution on* $[0,1]$. *Then, by Proposition J.1, it holds that* $\mathsf{M}_{\mathrm{reg},L_2}(\mathcal{G}, D_{\mathrm{reg}}) \lesssim \frac{1}{n}$. *However, the minimax* $\{0,1\}$*-risk is:*

$$\forall n \in \mathbb{N}, \quad \inf_{\mathrm{alg}_{\mathrm{reg}}} \sup_{g^\star \in \mathcal{G}} \mathbb{E}_{S_{n,\mathrm{reg}}} \mathbb{E}_{\hat{g} \sim \mathrm{alg}_{\mathrm{reg}}(S_{n,\mathrm{reg}})} \mathbb{E}_{z \sim D_{\mathrm{reg}}} \mathbb{E}_{\hat{y} \sim \hat{g}(\cdot | z} [g^\star(z) \neq \hat{y}] = 1, \tag{I.5}$$

where above, we permit randomized estimators $\hat{g}(\cdot \mid z)$ *In particular, this means that*

$$\begin{aligned} &\forall n \in \mathbb{N}, \quad \inf_{\mathrm{alg}_{\mathrm{reg}}} \sup_{g^\star \in \mathcal{G}} \mathbb{E}_{S_{n,\mathrm{reg}}} \mathbb{E}_{\hat{g} \sim \mathrm{alg}_{\mathrm{reg}}(S_{n,\mathrm{reg}})} \mathbb{E}_{z \sim D_{\mathrm{reg}}} \mathrm{D}(\boldsymbol{\delta}_{g^\star(z)}, \hat{g}(\cdot \mid z)) \\ &= \begin{cases} 1 & \mathrm{D} = \textit{Total Variation, Hellinger Distance} \\ \infty & \mathrm{D} = \textit{KL Divergence, Reverse KL Divergence}, \end{cases} \end{aligned}$$

where above $\boldsymbol{\delta}_{g^\star(z)}$ *is the Dirac distribution supported at* $y = g^\star(z)$.

Remark I.2 *The proof below generalizes to the case where* $\mathcal{G} = \mathcal{G}_{\mathrm{smooth}}(s, L; \mathcal{B}_k(1))$, $L > 0$ *(see Definition J.1). Indeed, the hard functions below are convergent sums over cosine functions with exponentially decaying weights, so these can be renormalized to have all* s *first derivatives bounded by any desired constant. In particular, by Proposition J.1 and setting* $k = 1$*, the same result holds even when* $\mathsf{M}_{\mathrm{reg},L_2}(\mathcal{G}, D_{\mathrm{reg}}) \lesssim C(s) n^{-s}$ *for any integral exponent* $s \in \mathbb{N}$.

Proof Define the embedding $\phi(x) = (\cos(2\pi i z))_{1 \le i \le D} : [0,1] \to \ell_1([D]) = \ell_2([D])$, and for vectors $\mathbf{w} \in \ell_2([D])$,

$$g_{\mathbf{w}}(z) = \langle \mathbf{w}, \phi(z) \rangle.$$

We first establish a claim which states that $g_{\mathbf{w}}(z)$ typically behaves like a continuous random variable.

Claim I.5 *Let* $\mathbf{w} = \mathbf{w}_0 + \mathbf{w}_1$*, where* $\mathbf{w}_0$ *is deterministic, and* $\mathbf{w}_1$ *has Lebesgue density w.r.t. to a subspace of dimension* $k \ge 1$. *Then, for almost every* $z \in [0,1]$, $g_{\mathbf{w}}(z)$ *has density with respect to the Lebesgue measure.*

Proof Let $\mathbf{P}$ denote the projection onto the subspace on which $\mathbf{w}_1$ has density. Then, for $g_0(z) = \langle \mathbf{w}_0, \phi(z) \rangle$, there is a random vector $\mathbf{w}'$ with density with respect to the Lebesgue meaure on $\mathbb{R}^D$ such that $g_{\mathbf{w}}(z) = g_0(z) + \langle \mathbf{P}\mathbf{w}', \phi(z) \rangle = g_0(z) + \langle \mathbf{w}', \mathbf{P}\phi(z) \rangle$. We claim that $\mathbf{P}\phi(z)$ vanishes on a set of of measure zero. Indeed, $\|\mathbf{P}\phi(z)\|^2$ is an analytic function, so if$\{z \in [0,1] : \phi(z)\}$ has positive measure, $\|\mathbf{P}\phi(z)\|^2$ vanishes on all of $z \in [0,1]$. Yet, at the same time, $\int_{z=0}^1 \|\mathbf{P}\phi(z)\|^2 \mathrm{d}z = \mathrm{tr}(\mathbf{P} \int_0^1 \phi(z)\phi(z)^\top \mathrm{d}z) = \mathrm{tr}(\mathbf{P}) > 0$, ans the entries of $\phi(z)$ are orthogonal on $[0,1]$, and their square expectation is nonvanishing. Finally, of all $z : \mathbf{P}\phi(z) \neq \mathbf{0}$, we observe that $\langle \mathbf{w}', \mathbf{P}\phi(z) \rangle$ has density with respect to the Lebesgue measure. ■

We now construct a Bayesian problem where $\mathbf{w}$ is drawn from a prior (supported on 1-bounded 1-Lipschitz functions) such that, for any $S_{n,\mathrm{reg}}$, the posterior $\mathbf{w} \mid S_{n,\mathrm{reg}}$ can be decomposed as in

Claim I.5. To do so, sample coordinates of $\mathbf{w}$ independently as $w_i \overset{\text{unif}}{\sim} [-1,1]/16^{-i}$, $1 \le i \le D$; we let $P([D])$ denote this prior. A simple computation reveals that

$$|\nabla g_{\mathbf{w}}(z)|^2 = \sum_{i=1}^{D} w_i^2 |\nabla \cos(2\pi i z)| \le \sum_{i=1}^{D} \frac{(2\pi)^2 i^2}{16^2} \le 1, \tag{I.6}$$

so $g_{\mathbf{w}}$ is supported on 1-Lipschitz functions. Finally, we take $D_{\text{reg}} = \text{Unif}([0,1])$.

Then, an sample $S_{n,\text{reg}} = (z_i, g_{\mathbf{w}}(z_i))_{1 \le i \le n}$ corresponds to taking n measurements of $\mathbf{w}$ with vectors $\phi(z_i) \in \mathbb{R}^D$. For $D > n$ it follows that, conditioned on $S_{n,\text{reg}}$, the distribution of $\mathbf{w}$ has density with respect to the Lebesgue measure supported on the subspace orthogonal to the span of $\phi(z_i)$, and is ortherwise deterministic on that subspace. Hence, by Claim I.5, the posterior distribution of $g_{\mathbf{w}}(z) \mid S_{n,\text{reg}}$ has density with respect to the Lebesgue measure. Thus, for any conditional distribution $\mu(\hat{y} \mid z, S_{n,\text{reg}})$,

$$\mathbb{E}_{\mathbf{w}|S_{n,\text{reg}}} \mathbb{E}_{z \sim [0,1]} \mathbb{E}_{\hat{y} \sim \mu(\cdot|z, S_{n,\text{reg}})} \mathbf{I}\{y \neq g_{\mathbf{w}}(x)\} = \int_0^1 \left(\mathbb{E}_{\hat{y} \sim \mu(\cdot|z, S_{n,\text{reg}})} \underbrace{\mathbb{E}_{\mathbf{w}|S_{n,\text{reg}}}[y = g_{\mathbf{w}}(x)]}_{=0 \text{ for almost all } z} \right) \mathrm{d}z = 0. \tag{I.7}$$

Hence, we have established a prior $P([D])$ such that

$$\begin{aligned}
&\inf_{\text{alg}_{\text{reg}}} \sup_{g^\star \in \mathcal{G}} \mathbb{E}_{S_{n,\text{reg}}} \mathbb{E}_{\hat{g} \sim \text{alg}_{\text{reg}}(S_{n,\text{reg}})} \mathbb{E}_{z \sim D_{\text{reg}}} \mathbb{E}_{\hat{y} \sim \hat{g}(\cdot|z)} [g^\star(z) \neq \hat{y}] \\
&\ge \inf_{\text{alg}_{\text{reg}}} \sup_{D \in \mathbb{N}} \mathbb{E}_{\mathbf{w} \sim P([D])} \mathbb{E}_{S_{n,\text{reg}}} \mathbb{E}_{\hat{g} \sim \text{alg}_{\text{reg}}(S_{n,\text{reg}})} \mathbb{E}_{z \sim D_{\text{reg}}} \mathbb{E}_{\hat{y} \sim \hat{g}(\cdot|z)} [g_{\mathbf{w}}(z) \neq \hat{y}] \\
&= 1.
\end{aligned}$$

■

I.5. Why the L_2 validation risk

Here, we justify our choice of focusing on the L_2 validation risk $\mathsf{R}_{\text{expert},L_2}$. There are three main reasons. First, L_2 validation risks are commonplace in both empirical and theoretical studies of regression problems. Moreover, as L_2 risks are large than L_1 risks by Jensen's inequality, showing that compounding error occurs *even if* the L_2 validation risk is bounded yields a stronger lower bound than showing the same given a bound only on the L_1 analogue.

Second, we shall establish lower bounds with a convenient feature: restriction to the algorithm class $\mathbb{A}$ does not harm the validation risk, in the sense that the restricted an unrestricted minimax risks are identical: $\mathsf{M}^{\mathbb{A}}_{\text{expert},L_2} = \mathsf{M}_{\text{expert},L_2}$. Establishing this equality relies on the fact that a Pythagorean theorem holds in L_2 space, which renders proper estimators optimal (recall the definition of proper estimators in Definition 3.4). We will show that the algorithm classes $\mathbb{A}$ of interest contain proper estimators, rendering the inequality $\mathsf{M}^{\mathbb{A}}_{\text{expert},L_2} = \mathsf{M}_{\text{expert},L_2}$.

Finally, our lower bounds for non-simply stochastic algorithms hold against an L_2-validation risk, defined in Appendix F.2. This is because the L_2-risk emphasizes the tails of the errors more significantly. For these results, an L_2 validation risk is preferrable for consistency.

Appendix J. Additional material for minimax formulations in Appendix E

J.1. The necessity of the typical regresion classes, Condition E.1

Below, we provide an example of regression problem classes where optimal estimators make errors of magnitude at least 1, but as $n \to \infty$, the probabilty of these errors decays as $1/n$, and consequently, the minimax risk still decays to 0 as $n \to \infty$.

Example 1 (An example where Condition E.1 fails) *Consider regression with a distribution D_{reg} supported on the dyadic set $\mathbb{D} := \{2^{-k}, k \in \mathbb{N} \cup \{0\}\}$, with $\mathbb{P}_{D_{\mathrm{reg}}}[z = 2^{-k}] \propto k^{-2}$. Consider $\mathcal{G}$ to be the class of all binary functions $g : \mathbb{D} \to \{-1, 1\}$. It is easy to check that the minimax optimal estimator predicts $g(z)$ for all $z \in \mathbb{D}$ seen in the sample, and predicts $z = 0$ otherwise; from this estimator, one can check that $\mathsf{M}_{\mathrm{reg},L_2}(n, \mathcal{G}; \mathcal{D}) \propto 1/n$. On the other hand, Condition E.1 only holds for $c = 1/n$. This occurs because all errors have magnitude at least 1, but are make with increasingly lower probability as n grows larger.*

We now illustrate why Example 1 is unsuitable for a compounding error construction. Consider the following formulation of minimax risk for $C \geq 1, B > 0$:

$$\mathsf{M}_{\mathrm{reg},L_2}(n; \mathcal{G}, D_{\mathrm{reg}}, [C, B]) = \inf_{\mathrm{alg}_{\mathrm{reg}}} \sup_{g^\star \in \mathcal{G}} \mathbb{E}_{S_{n,\mathrm{reg}}} \mathbb{E}_{\hat{g} \sim \mathrm{alg}_{\mathrm{reg}}(S_{n,\mathrm{reg}})} \mathbb{E}_{\mathbf{z} \sim D_{\mathrm{reg}}} [\min\{B, C|g^\star(\mathbf{z}) - \hat{g}(\mathbf{z})|^2\}]. \tag{J.1}$$

This measures the risk, magnitude by a factor of $C \geq 1$, but clipped by B. Effectively, our arguments lower bound $\mathsf{M}_{\mathrm{reg},L_2}(n; \mathcal{G}, D_{\mathrm{reg}}, [C, B])$ when $C \sim \exp(\Omega(H))$. We observe that the mininimax risk of the problem in Example 1 does not meaningfully change $B = 1$ and we increase C, because all errors made are saturated at magnitude $\Omega(1)$. Hence, Example 1 provides a problem which *cannot* be embedded into a compounding error construction.

J.2. Verifying Condition E.1 (Proof of Proposition E.1)

This section demonstrates that Condition E.1 holds for a natural class of non-parametric functions. Before continuing, recall $\mathcal{B}_k(r)$ is the radius-r ball in $\mathbb{R}^k$. Given a set $\Omega \subset \mathbb{R}^k$ of nonzero Lebesgue measure, we let $D_{\mathrm{unif}}(\Omega)$ denote the uniform distribution on that set.

Definition J.1 (Smooth Functions) *For $k, s \in \mathbb{N}$, and an open, bounded domain $\Omega \subset \mathbb{R}^k$, define $\mathcal{G}_{\mathrm{smooth}}(s, L; \Omega)$ as the set of functions $g : \Omega$ which are s-times continuously differentiable, and such*

$$0 \leq j \leq s, \quad \|\nabla^j g(\mathbf{z})\|_{\mathrm{op}} \leq L \tag{J.2}$$

where ∇^j is the j-th order derivative tensor (with $\nabla^0 g \equiv g$), and $\|\cdot\|_{\mathrm{op}}$ the tensor operator norm.

The above definition of smooth functions corresponds to the space of functions whose L_∞, order-s Sobolev norm (denoted $W^s_\infty(\Omega)$) is bounded, and in fact the results in this section extend to all L_p, order-s Sobolev norms (the space $W^s_p(\Omega)$) for $p \geq 2$. We refer the reader to Krieg et al. (2022) for these generalizations. It is clear that the class $\mathcal{G}_{\mathrm{smooth}}(s, L; \Omega)$ (even if the domain Ω is nonconvex).

Our main result is a more constructive statement of Proposition E.1.

Proposition J.1 *For $k, s \in \mathbb{N}$, let $\mathcal{G}_{s,k} := \mathcal{G}_{\mathrm{smooth}}(s, 1; \mathcal{B}_k(1))$ denote the space of s-order 1-smooth functions on the unit ball in R^k, and let $D_k := D_{\mathrm{unif}}(\mathcal{B}_k(1))$. Note that this class is $(1,1,1)$-regular (recall Definition E.6) for $s \geq 2$. Then, there exists constants $C_1(s,k) > 0$ and $C_2(s,k) > 0$, depending only on s_k, and a universal constant $c > 0$, such that for all $n \geq 1$,*

$$\mathsf{M}_{\mathrm{reg},L_2}(n; \mathcal{G}_{s,k}, D_k) \leq C_1(s,k) n^{-\frac{s}{k}}, \quad \mathsf{M}_{\mathrm{reg,prob}}(n, c^k; \mathcal{G}_{s,k}, D_k) \geq C_2(s,k) n^{-\frac{s}{k}}$$

In particular, there exists a constants $\kappa(s,k)$ and $\delta(k)$ such that $(\mathcal{G}_{s,k}, D_k)$ is $(\kappa(s,k), \delta(k))$-typical. Furthermore, in view of Eq. (E.5),the above bounds imply that there exists some other $C_3(s,k)$ for which

$$\mathsf{M}_{\mathrm{reg},L_2}(n; \mathcal{G}_{s,k}, D_k) \geq C_3(s,k) n^{-\frac{s}{k}}.$$

The upper bound is a direct consequence of Krieg et al. (2022, Theorem 1), taking s as is, $d \leftarrow k$, $p = \infty$ and $q = 2$. [2] Here, we focus on the lower bounds. Again, the arguments are somewhat standard (see e.g. (Krieg et al., 2022; Tsybakov, 2009; Bauer et al., 2017)).

J.2.1. PROOFS OF PROPOSITION J.1

As noted above, the upper bound follows from Krieg et al. (2022, Theorem 1). The lower bound follows from standard construction (see, e.g. Kohler and Krzyżak (2013); Tsybakov (1997), but where we take care to lower-bound the in-probability minimax risk, $\mathsf{M}_{\mathrm{reg,prob}}$.

Definition J.2 (Packing, see e.g. Section 4 in Vershynin (2018)) *We say that $(\mathbf{z}_1, \dots, \mathbf{z}_m)$ forms an ϵ-packing of a set Ω if each $\mathbf{z}_i \in \Omega \subset \mathbb{R}^k$, and $\|\mathbf{z}_i - \mathbf{z}_j\| \geq \epsilon$ for $i \neq j$.*

Definition J.3 *Let $\mathcal{G}$ be a function class supported on $\mathcal{B}_k(1)$. We say that a function $r(\cdot) : (0,1) \to \mathbb{R}_{>0}$ is a ϵ_0-bandwith function for $\mathcal{G}$ if for all $\epsilon \leq \epsilon_0$, there is exists a function $g_\epsilon : \mathbb{R}^k \to \mathbb{R}_{\geq 0}$ for which (a) $g_\epsilon(\mathbf{z}) = 0$ for all $\mathbf{z} : \|\mathbf{z}\| \geq \epsilon$, (b) if $\mathbf{z}_1, \mathbf{z}_2, \dots, \mathbf{z}_m$ are the centers of an ϵ-packing of $\mathcal{B}_k(1)$, for $i \neq j$, that $\mathbf{z} \mapsto \sum_i g_\epsilon(\mathbf{z} - \mathbf{x}_i) \in \mathcal{G}$, and (c), if $\|\mathbf{z}\| \leq \frac{\epsilon}{2}$, then $|g_\epsilon(\mathbf{z})| \geq 2r(\epsilon)$.*

Lemma J.1 *There exists a universal constant $c \in (0,1)$ such that, for all $k \in \mathbb{N}$, the following is true. Let $\mathcal{G}$ be a function class supported on $\mathcal{B}_k(1)$, and suppose that $r(\cdot) : (0,1) \to \mathbb{R}_{\geq 0}$ is a ϵ_0-bandwith function for $\mathcal{G}$. Then, then, for $n \geq (\epsilon_0)^k$, we have*

$$\mathsf{M}_{\mathrm{reg,prob}}\left(n, c^k; \mathcal{G}, D_{\mathrm{unif}}(\mathcal{B}_k(1))\right) \geq r\left(n^{-\frac{1}{k}}\right) \tag{J.3}$$

Before proving Lemma J.1, we prove the main results of this section. To instantiate the lower bound, recall the definition of bump functions:

Lemma H.14 (Existence of Bump Functions) *For any $k \in \mathbb{N}$, there exists a C^∞ function* $\mathrm{bump}_k(\mathbf{z}) : \mathbb{R}^k \to \mathbb{R}$, *called an* bump function, *sastisfying* $\mathrm{bump}_k(\mathbf{z}) = 1$ *if and only if* $\|\mathbf{z}\| \leq 1$, $\mathrm{bump}_k(\mathbf{z}) = 0$ *if and only if $\|\mathbf{z}\| \geq 2$. And, for each $p \geq 1$, $\|\nabla^p \mathrm{bump}_k(\mathbf{z})\|_{\mathrm{op}} \leq c_p$, where $\|\cdot\|_{\mathrm{op}}$ denotes the tensor-operator norm, and c_p is a constant independent of k but depending on p. Finally, $\nabla^p \mathrm{bump}_k(\mathbf{z}) = 0$ for all $\mathbf{z} : \|\mathbf{z}\| \geq 2$.*

2. Note that their normalization of the (s, ∞)-Sobolev norm is in fact slightly larger than ours (consult the third equation on Krieg et al. (2022, Page 3)).

We use the following lemma to construct bandwidth functions.

Lemma J.2 *Let c_s denote the constant given in Lemma H.14, and define $c'_s := \max_{1\le j\le s} c_j$. Given $\epsilon \in (0,1]$, define the function $\phi_{\epsilon,s}(\mathbf{z}) = \frac{\epsilon^s}{2^s c'_s}\mathrm{bump}_k(2\mathbf{z}/\epsilon)$. Then, $\phi_\epsilon(\mathbf{z}) = 0$ for $\|\mathbf{z}\| \ge \epsilon$, $\phi_\epsilon(\mathbf{z}) \ge \frac{\epsilon^s}{2^s c'_s}$ for $\|\mathbf{z}\| \le \epsilon/2$, and for any ϵ-packing $\mathbf{z}_1,\dots,\mathbf{z}_m$, the function*

$$g(\mathbf{z}) := \sum_{i=1}^m \phi_{\epsilon,s}(\mathbf{z}-\mathbf{z}_i)$$

satisfies $g(\mathbf{z}) \in [0,1]$, and $\sup_{\boldsymbol{\alpha}\in\mathbb{N}^k:|\boldsymbol{\alpha}|\le s} |\mathrm{D}^{\boldsymbol{\alpha}} g(\mathbf{z})| \le \max_{0\le j\le s}\|\nabla^j g(\mathbf{z})\| \le 1$.

Proof The inequalities $\phi_{\epsilon,s}(\mathbf{z}) = 0$ for $\|\mathbf{z}\| \ge \epsilon$ and $\phi_{\epsilon,s}(\mathbf{z}) \ge \epsilon^s/c'_s$ for $\|\mathbf{z}\| \le \epsilon/2$ follow directly from the definition of the bump function. In addition, we have that $\nabla^j\phi_{\epsilon,s}(\mathbf{z}) = 0$ for all $\mathbf{z} : \|\mathbf{z}\| \ge \epsilon$ as well. Given an ϵ-packing $\mathbf{z}_1,\dots,\mathbf{z}_m$, there is at most one index $i_\star$ such that $\|\mathbf{z}_{i_\star} - \mathbf{z}\| < \epsilon$. If no such index $i_\star$ exists, than $g(\mathbf{z}) := \sum_{i=1}^m \phi_{\epsilon,s}(\mathbf{z}-\mathbf{z}_i)$ and all its derivatives vanish. Otherwise, for $\epsilon \le 1$,

$$\|\nabla^j g(\mathbf{z})\| = \|\nabla^j \phi_\epsilon(\mathbf{z}-\mathbf{z}_{i_\star})\| = \frac{\epsilon^s}{2^s c'_s}\cdot(2/\epsilon)^j\|\nabla^j \mathrm{bump}_k(\mathbf{z}')\big|_{\mathbf{z}'=2(\mathbf{z}-\mathbf{z}_{i_\star})/\epsilon}\| \le \frac{c_j}{c'_s}\cdot(2/\epsilon)^{j-s} \le 1.$$

■

Together, Lemmas J.1 and J.2 imply the result.
Proof [Proof of lower bound in Proposition J.1] By Lemma J.2, we see that $r(\epsilon) = \frac{\epsilon^s}{C}$ is $(\epsilon_0 = 1)$-bandwidth function for the class $\mathcal{G}_{\mathrm{smooth}}(s,1;\mathcal{B}_k(1))$ for some constant $C = C(s)$. The lower bound on $\mathsf{M}_{\mathrm{reg,prob}}$ now follows directly from Lemma J.1. ■

J.2.2. PROOF OF LEMMA J.1

Proof Let N_ϵ denote the maximal cardinality of an ϵ-packing of $\mathcal{B}_k(1)$. Pick a one such packing, and enumerate the center of the balls in the packing $\mathbf{z}_1,\mathbf{z}_2,\dots,\mathbf{z}_{N_\epsilon}$.

Now, let us consider estimation against a prior over functions $\sum_{i=1}^{N_\epsilon}\xi_i g_i(\mathbf{z}-\mathbf{z}_i)$,where ξ_i are i.i.d. Bernoulli random variables. For any estimator $\mathrm{alg}_{\mathrm{reg}}$, the Bayesian probability of an error of magnitude $r(\epsilon)$, which lower bounds the worst-case probability, under this prior is

$$\begin{aligned}
&\inf_{\mathrm{alg}_{\mathrm{reg}}} \mathbb{E}_\xi \mathbb{E}_{S_{n,\mathrm{reg}}}\mathbb{E}_{\mathbf{z}\sim\mathcal{B}_1}\mathbb{E}_{y\sim\mathrm{alg}_{\mathrm{reg}}(S_{n,\mathrm{reg}},\mathbf{z})}[\mathbf{I}\{|y-\sum_{i=1}^{N_\epsilon}\xi_i g_\epsilon(\mathbf{z}-\mathbf{z}_i)| > r(\epsilon)\}].\\
&\ge \sum_{j=1}^{N_\epsilon}\frac{1}{\mathrm{vol}(\mathcal{B}_k(1))}\inf_{\mathrm{alg}_{\mathrm{reg}}}\mathbb{E}_\xi\mathbb{E}_{S_{n,\mathrm{reg}}}\int_{x_0\in x_j+\mathbf{B}_k(\epsilon)}\mathbb{E}_{y\sim\mathrm{alg}_{\mathrm{reg}}(S_{n,\mathrm{reg}},x_0)}[\mathbf{I}\{|y-\sum_{i=1}^{N_\epsilon}\xi_i g_\epsilon(\mathbf{z}-\mathbf{z}_i)| > \epsilon\}]\mathrm{d}x_0\\
&= \sum_{j=1}^{N_\epsilon}\frac{1}{\mathrm{vol}(\mathcal{B}_k(1))}\inf_{\mathrm{alg}_{\mathrm{reg}}}\mathbb{E}_\xi\mathbb{E}_{S_{n,\mathrm{reg}}}\int_{x_0\in x_j+\mathbf{B}_k(\epsilon)}\mathbb{E}_{y\sim\mathrm{alg}_{\mathrm{reg}}(S_{n,\mathrm{reg}},x_0)}[\mathbf{I}\{|y-\xi_j g_\epsilon(\mathbf{z}-\mathbf{z}_j)| > \epsilon\}]\mathrm{d}x_0.
\end{aligned}$$

Let $\mathcal{E}_j$ be the probability that $S_{n,\mathrm{reg}}$ is contains no elements in the set $x_j + \mathbf{B}_k(\epsilon)$; since samples are uniform on $\mathbf{B}_k$, $p(n,\epsilon) := \mathbb{P}_{S_{n,\mathrm{reg}}}[\mathcal{E}_j]$ is independent of j. On $\mathcal{E}_j$, we have no information about

ξ_j, so its posterior is uniform. Thus, for any chosen index j, and conditional measure $\mu(\cdot \mid x_0)$, the above is at least

$$\begin{aligned}
&\frac{N_\epsilon p(n,\epsilon)}{2\mathrm{vol}(\mathcal{B}_k(1))}\int_{\mathbf{z}\in\mathbf{z}_j+\mathbf{B}_k(\epsilon)} \mathbb{E}_{y\sim\mu(\cdot|x_0)}[\mathbf{I}\{|y-g_\epsilon(\mathbf{z}-\mathbf{z}_j)|>r(\epsilon)\}+\mathbf{I}\{|y|>r(\epsilon)\}]\mathrm{d}\mathbf{z}\\
&\geq \frac{N_\epsilon p(n,\epsilon)}{2\mathrm{vol}(\mathcal{B}_k(1))}\int_{\mathbf{z}\in\mathbf{z}_j+\mathbf{B}_k(\epsilon)} \mathbf{I}\{|g_\epsilon(\mathbf{z}-\mathbf{z}_j)|>2r(\epsilon)\}\mathrm{d}\mathbf{z}\\
&= \frac{N_\epsilon p(n,\epsilon)}{2\mathrm{vol}(\mathcal{B}_k(1))}\int_{\mathbf{z}\in\mathbf{z}_j+\mathbf{B}_k(\epsilon)} \mathbf{I}\{|g_\epsilon(\mathbf{z})|>2r(\epsilon)\}\mathrm{d}\mathbf{z}\\
&> \frac{N_\epsilon p(n,\epsilon)}{2\mathrm{vol}(\mathcal{B}_k(1))}\int_{\mathbf{z}\in\mathbf{z}_j+\mathbf{B}_k(\epsilon)} \mathbf{I}\left\{\|\mathbf{z}\|<\frac{1}{2}\epsilon\right\}\mathrm{d}\mathbf{z} \qquad \text{(Definition of a bandwidth function)}\\
&= \frac{(\mathrm{vol}(\mathcal{B}_k(\frac{\epsilon}{2}))N_\epsilon p(n,\epsilon)}{2\mathrm{vol}(\mathcal{B}_k(1))}\\
&= \frac{\epsilon^k}{2^{k+1}}\cdot N_\epsilon p(n,\epsilon).
\end{aligned}$$

By a standard estimate (see, e.g. Vershynin (2018, Section 4)), $\epsilon^k N_\epsilon \gtrsim (c_1)^k$ for $\epsilon \leq 1/4$ and universal, dimension-independent constant $c_1 \in (0,1)$. Moreover, for $n \geq 1/\epsilon^k$,

$$p(n,\epsilon) = 1 - \left(1-\frac{\mathrm{vol}(\mathcal{B}_k(\epsilon))}{\mathcal{B}_k(1)}\right)^n = 1-(1-\epsilon^k)^n \geq 1-\exp(-n\epsilon^{-k}) \geq \frac{1}{2}.$$

Hence, by setting $c = c_1/4$,

$$\inf_{\mathrm{alg}_{\mathrm{reg}}} \mathbb{E}_\xi \mathbb{E}_{S_{n,\mathrm{reg}}}\mathbb{E}_{\mathbf{z}\sim\mathcal{B}_1}\mathbb{E}_{y\sim\mathrm{alg}_{\mathrm{reg}}(S_{n,\mathrm{reg}},\mathbf{z})}[\mathbf{I}\{|y-\sum_{i=1}^{N_\epsilon}\xi_i g_\epsilon(\mathbf{z}-\mathbf{z}_i)|>r(\epsilon)\}] \geq c^k.$$

By choosing $\epsilon = n^{-1/k}$, we conclude. ■

J.3. General (Minimax) Risks and Comparisons Between Them

In this section, we provide comparisons between general families of costs and their associated minimax risks. We recall the cannonical coupling between $(\mathbb{P}_{\hat{\pi},f,D}, \mathbb{P}_{\pi^\star,f,D},)$ over random variables $(\hat{\mathbf{x}}_{1:H}, \hat{\mathbf{u}}_{1:H}) \sim (\mathbb{P}_{\hat{\pi},f,D}$ and $(\mathbf{x}^\star_{1:H}, \mathbf{u}^\star_{1:H}) \sim (\mathbb{P}_{\pi^\star,f,D})$ in Definition I.1.

We begin by defining general notions of L_p-style risks:

$$\begin{aligned}
\mathsf{R}_{\mathrm{cost},L_p}(\hat{\pi};\pi^\star,f,D,H) &:= \mathbb{E}_{\hat{\pi},\pi^\star,f_{g,\xi},D}\left[|\mathrm{cost}(\mathbf{x}_{1:H},\mathbf{u}_{1:H}) - \mathrm{cost}(\mathbf{x}_{1:H},\mathbf{u}_{1:H})|^p\right]^{1/p}.\\
\mathsf{R}_{\mathrm{traj},L_p}(\hat{\pi};\pi^\star,f,D,H) &:= \mathbb{E}_{\hat{\pi},\pi^\star,f,D}\left[\sum_{t=1}^H \min\left\{\|\mathbf{x}^\star_t-\hat{\mathbf{x}}_t\|+\|\mathbf{u}^\star_t-\hat{\mathbf{u}}_t\|,\ 1\right\}^p\right]^{1/p}.
\end{aligned} \tag{J.4}$$

In the special case that cost vanishes on $(\mathcal{P}, D)$, $\mathsf{R}_{\mathrm{cost},L_p}$ takes a simpler form (coinciding with)

$$\mathsf{R}_{\mathrm{cost},L_2}(\hat{\pi};\pi^\star,f,D,H) := \mathbb{E}_{\hat{\pi},f_{g,\xi},D}\left[|\mathrm{cost}(\mathbf{x}_{1:H},\mathbf{u}_{1:H})|^p\right]^{1/p}, \quad \mathrm{cost}\in\mathcal{C}_{\mathrm{vanish}}(\mathcal{P},D),$$

coinciding with Eq. (F.3).

We define the associated minimax risks

$$\mathbf{M}_{\mathrm{cost},L_p}(n;\mathcal{P},D,H) := \mathbf{M}^{\mathbb{A}}\left(n,\mathbf{R}_{\mathrm{cost},L_p};\mathcal{P},D,H\right), \quad \mathbf{M}_{\mathrm{traj},L_p} := \mathbf{M}^{\mathbb{A}}\left(n,\mathbf{R}_{\mathrm{traj},L_p};\mathcal{P},D,H\right). \tag{J.5}$$

Finally, for the sake of completeness, we propose a generalization of the in-probability risk for non-vanishing costs:

$$\begin{aligned}&\mathbf{M}^{\mathbb{A}}_{\mathrm{cost,prob}}(n,\delta;\mathcal{P},D,H)\\ &:= \inf\left\{\epsilon : \inf_{\mathrm{alg}\in\mathbb{A}} \sup_{(\pi^\star,f)\in(\mathcal{P},D)} \mathbb{E}_{\mathrm{S}_{n,H}}\mathbb{E}_{\hat\pi\sim\mathrm{alg}(\mathrm{S}_{n,H})}\mathbb{P}_{\hat\pi,\pi^\star,f,D}[\mathrm{cost}(\hat{\mathbf{x}}_{1:H},\hat{\mathbf{u}}_{1:H}) - \mathrm{cost}(\mathbf{x}^\star_{1:H},\mathbf{u}^\star_{1:H}) \ge \epsilon] \le \delta\right\},\end{aligned}$$

which coincides with Definition E.4 in the case that $\mathrm{cost} \in \mathcal{C}_{\mathrm{vanish}}(\mathcal{P},D)$.

We now state a proposition consisting of elementary relations between the risks thus defined.

Proposition J.2 *Fix an IL problem class $(\mathcal{P},D)$, let n,δ,H and algorithm class $\mathbb{A}$ be arbitrary.*

(a) ***Monotonicty:*** *$\mathbf{R}(\hat\pi;\pi^\star,f,D,H) \ge \mathbf{R}'(\hat\pi;\pi^\star,f,D,H)$ for all $(\pi^\star,f)\in\mathcal{P}$ and all $\hat\pi$. Then $\mathbf{M}^{\mathbb{A}}(n;\mathbf{R},\mathcal{P},D,H) \ge \mathbf{M}^{\mathbb{A}}(n;\mathbf{R}',\mathcal{P},D,H)$*

(b) ***Markov's Inequality:*** *For any cost, $\mathbf{M}^{\mathbb{A}}_{\mathrm{cost},L_p}(n;\mathcal{P},D,H) \ge \delta^{1/p}\mathbf{M}^{\mathbb{A}}_{\mathrm{cost,prob}}(n,\delta;\mathcal{P},D,H)$.*

(c) ***Simplification for Vanishing Costs:*** *For any nonnegative cost $\in\mathcal{C}_{\mathrm{vanish}}(\mathcal{P},D)$ and $(\pi^\star,f)\in\mathcal{P}$, $\mathbf{R}_{\mathrm{cost}}(\hat\pi;\pi^\star,f,D,H) = \mathbf{R}_{\mathrm{cost},L_1}(\hat\pi;\pi^\star,f,D,H)$. Thus, $\mathbf{M}^{\mathbb{A}}_{\mathrm{cost}}(n;\mathcal{P},D,H) = \mathbf{M}_{\mathrm{cost},L_1}(n;\mathcal{P},D,H)$. In particular,*

$$\mathbf{M}^{\mathbb{A}}_{\mathrm{cost}}(n;\mathcal{P},D,H) \ge \delta\mathbf{M}^{\mathbb{A}}_{\mathrm{cost,prob}}(n,\delta;\mathcal{P},D,H).$$

(d) ***Trajectory Risk Dominates Lipschitz Cost:*** *For any cost $\in\mathcal{C}_{\mathrm{Lip}}$, $\mathbf{R}_{\mathrm{traj},L_p}(\hat\pi;\pi^\star,f,D,H) \ge \mathbf{R}_{\mathrm{cost},L_p}(\hat\pi;\pi^\star,f,D,H)$ and $\mathbf{M}^{\mathbb{A}}_{\mathrm{traj},L_p}(n;\mathcal{P},D,H) \ge \mathbf{M}^{\mathbb{A}}_{\mathrm{cost},L_p}(n,\delta;\mathcal{P},D,H)$.*

(e) ***Monotonicity in p:*** *$\mathbf{R}_{\mathrm{traj},L_p}, \mathbf{R}_{\mathrm{cost},L_p}, \mathbf{M}_{\mathrm{traj},L_p}$ and $\mathbf{M}_{\mathrm{cost},L_p}$ are nondecreasing in p.*

Point (c) also holds when $\mathcal{C}_{\mathrm{Lip}}$ is replaced by the set of non-stationary additive costs, $\mathrm{cost}(\mathbf{x}_{1:H},\mathbf{u}_{1:H}) = \sum_{h=1}^H \mathrm{cost}_h(\mathbf{x}_h,\mathbf{u}_h)$, where each $\mathrm{cost}_h(\cdot,\cdot)$ is 1-Lipschitz and bounded in $[0,1]$.

Proof The points are straightforward to verify. Point (a) is immediate from the definition of minimax risk. Point (b) uses the fact that, for a nonnegative random-variable X, $\mathbb{E}[X^p]^{1/p} \ge \epsilon\,\mathbb{P}[X\ge\epsilon]^{1/p}$ by Markov's inequality. Point (c) is simply uses $|x-y| = x$ for $y = 0$ and x nonnegative. Point (d) directly generalizes the proof of Lemma I.1. Finally, point (e) is Jensen's inequality: $\mathbb{E}[|X|^p]^{1/p} \le \mathbb{E}[|X|^q]^{1/q}$ for any $p \le q$. ■

We conclude the section with a subtle point that may be of interest to experts. The class $\mathcal{C}_{\mathrm{Lip}}$ considers additive costs, and hence typically will scale linearly in H. We may instead consider a class $\mathcal{C}_{\mathrm{lip,max}}$ of costs normalized by their maximum, which ensures total cost stays bounded.

Definition J.4 ($\max$-Lipschitz Cost Family) *We define $\mathcal{C}_{\mathrm{lip,max}} := \{\mathrm{cost}(\mathbf{x}_{1:H},\mathbf{u}_{1:H}) = \max_{h\ge1}\tilde{\mathrm{cost}}(\mathbf{x}_h,\mathbf{u}_h) : \tilde{\mathrm{cost}}$ is $1-$Lipschitz and takes values in $[0,1]\}$.*

Lower bounds on $\mathcal{C}_{\text{lip,max}}$ implies those on $\mathcal{C}_{\text{Lip}}$.

Lemma J.3 (An alternate horizon normalization) *For each* $\text{cost} \in \mathcal{C}_{\text{lip,max}} \cap \mathcal{C}_{\text{vanish}}(\mathcal{P}, D)$*, there exists a* $\text{cost}' \in \mathcal{C}_{\text{Lip}} \cap \mathcal{C}_{\text{vanish}}(\mathcal{P}, D)$ *such that* $\mathsf{M}^{\mathbb{A}}_{\text{cost}',\text{prob}}(n, \delta; \mathcal{P}, D, H) \geq \mathsf{M}^{\mathbb{A}}_{\text{cost},\text{prob}}(n, \delta; \mathcal{P}, D, H)$*, and similarly for the* L_p*risks and minimax risks.*

Proof Let $\text{cost}(\mathbf{x}_{1:H}, \mathbf{u}_{1:H}) = \max_h \tilde{\text{cost}}(\mathbf{x}_h, \mathbf{u}_h) \in \mathcal{C}_{\text{lip,max}} \cap \mathcal{C}_{\text{vanish}}(\mathcal{P}, D)$. It is straightforward to check that $\text{cost}'(\mathbf{traj}) = \sum_h \tilde{\text{cost}}(\mathbf{x}_h, \mathbf{u}_h)$ satisfies the desired conditions. Note that this argument extends to non-stationary costs as well. ■

Lastly, we remark that we can remove the restriction of the costs $\mathcal{C}_{\text{lip,max}}$ to the range $[0, 1]$ with the following trick.

Remark J.1 *Let* $\tilde{\mathcal{C}}_{\text{lip,max}} := \{\text{cost}(\mathbf{x}_{1:H}, \mathbf{u}_{1:H}) = \max_{h \geq 1} \tilde{\text{cost}}(\mathbf{x}_h, \mathbf{u}_h) : \tilde{\text{cost}}$ *is* $1-$*Lipschitz and nonnegative*$\}$*, by analogy to* $\mathcal{C}_{\text{lip,max}}$ *but without the* $[0, 1]$ *restriction. Then, if* $\text{cost} \in \tilde{\mathcal{C}}_{\text{lip,max}}$*,* $\text{cost}' = \min\{1, \text{cost}\} \in \mathcal{C}_{\text{lip,max}}$*, and* $\min\{1, \mathsf{M}^{\mathbb{A}}_{\text{cost},\text{prob}}(n, \delta; \mathcal{P}, D, H)\} = \mathsf{M}^{\mathbb{A}}_{\text{cost}',\text{prob}}(n, \delta; \mathcal{P}, D, H)$

Appendix K. Proof of Proposition G.1

We recall

$$\mathsf{M}_{\text{expert},h=1}(n; \mathcal{P}, D) := \inf_{\text{alg}} \sup_{(\pi,f)\in\mathcal{P}} \mathbb{E}_{\mathrm{S}_{n,H}} \mathbb{E}_{\mathbf{x}_1 \sim D} \mathbb{E}_{\mathbf{u} \sim \hat{\pi}(\mathbf{x},1)} \left[\|\pi(\mathbf{x}_1, t = 1) - \mathbf{u}\|^2\right]^{1/2}.$$

Definition K.1 (Shorthand Notation) *We use the shorthand* $S_{n,\text{reg}} \sim \text{law}_{\text{reg}}(g)$ *to denote the law of samples* $S_{n,\text{reg}}$ *from the regression problem with ground truth* $g \in \mathcal{G}$*. We let* $\mathrm{S}_{n,H} \sim \text{law}(g)$ *to denote the law of samples* $\mathrm{S}_{n,H}$ *under the instance* $(\pi_{g,\xi}, f_{g,\xi})$*. Notice that under the* ξ*-indistinguishable property (Property G.3),* $\text{law}(g)$ *is well-defined as it does not depend on* ξ*. Finally, for IL algorithms* alg *and regression estimators* alg_{reg}*, we define*

$$\begin{aligned}
\mathsf{R}_{\text{train},h=1}(\text{alg}; g) &= \mathbb{E}_{\mathrm{S}_{n,H} \sim \text{law}(g)} \mathbb{E}_{\hat{\pi} \sim \text{alg}(\mathrm{S}_{n,H})} \left(\mathbb{E}_{\mathbf{x} \sim D} \mathbb{E}_{\mathbf{u} \sim \hat{\pi}(\mathbf{x},1)} \|\mathbf{u} - \pi(\mathbf{x}, t = 1)\|^2\right)^{1/2} \\
\mathsf{R}_{\text{expert},L_1}(\text{alg}; g) &= \mathbb{E}_{\mathrm{S}_{n,H} \sim \text{law}(g)} \mathbb{E}_{\hat{\pi} \sim \text{alg}(\mathrm{S}_{n,H})} \mathsf{R}_{\text{expert},L_2}(\hat{\pi}; \pi_{g,\xi}, f_{\xi,g}) \\
\mathsf{R}_{\text{reg}}(\text{alg}_{\text{reg}}; g) &= \mathbb{E}_{S_{n,\text{reg}} \sim \text{law}_{\text{reg}}(g)} \left(\mathbb{E}_{\hat{g} \sim \text{alg}_{\text{reg}}(\mathrm{S}_{n,H})} \mathbb{E}_{\mathbf{z} \sim D_{\text{reg}}} \|\hat{g}(\mathbf{z}) - g(\mathbf{z})\|^2\right)^{1/2}
\end{aligned}$$

Again, indistinguishability implies the above are well-defined (e.g. independent of ξ*).*

Lemma K.1 (Reduction from regression to IL) *For every IL algorithm* alg*, there exists a regression algorithm* alg_{reg} *such that for all* $g \in \mathcal{G}$*,*

$$\tau \mathsf{R}_{\text{reg}}(\text{alg}_{\text{reg}}; g) \leq \mathsf{R}_{\text{train},h=1}(\text{alg}; g) \leq \mathsf{R}_{\text{expert},L_1}(\text{alg}; g) \tag{K.1}$$

Proof Let us construct the transformation from regression algorithms alg_{reg} to IL algorithms alg. Our first step is to construct a (stochastic) transformation Φ from regression samples to IL trajectories defined by

$$\Phi : (\mathbf{z}, y) \in (\mathbb{R}^{d'} \times \mathbb{R})^n \mapsto (\mathbf{x}, y\mathbf{v} + \pi_0(\mathbf{x}_1), \mathbf{x}_2, \mathbf{u}_2, \dots), \quad \mathbf{x} \sim \mathcal{K}(\mathbf{z}), \mathbf{x}_2, \mathbf{u}_2 \sim \mathbb{P}[\cdot \mid \mathbf{x}_1, \mathbf{u}_1], \tag{K.2}$$

where $\mathbb{P}$ is instance-independent measure governing the remainder of the trajectory conditioned on $(\mathbf{x}_1, \mathbf{u}_1)$. We extend Φ as mapping from samples $S_{n,\mathrm{reg}}$ of regression pairs to samples $\mathrm{S}_{n,H}$ of trajectories by independently applying Φ to each pair. Further recall the definition $\mathrm{mean}[\pi](\mathbf{x}, t) = \mathbb{E}_{\mathbf{u}\sim\pi(\mathbf{x},t)}[\mathbf{u}]$. Then,

$$\hat{g}(\mathbf{z}; \hat{\pi}) = \frac{1}{\tau}\mathbf{v}^\top \left(\mathbb{E}_{\mathbf{x}\sim\mathcal{K}(\mathbf{z})}\left[\mathrm{mean}[\hat{\pi}](\mathbf{x}, t=1) - \pi_0(\mathbf{x})\right]\right), \tag{K.3}$$

and finally define the regression estimator $\mathrm{alg}_{\mathrm{reg}}(S_{n,\mathrm{reg}})$ via

$$\mathrm{S}_{n,H} \sim \Phi(S_{n,\mathrm{reg}}), \quad \hat{\pi} \sim \mathrm{alg}(\mathrm{S}_{n,H}), \quad \hat{g} = \hat{g}(\mathbf{z}; \hat{\pi}).$$

Let us now show that $\mathbf{R}_{\mathrm{expert},L_1}(\mathrm{alg}_{\mathrm{reg}}; g) \le \mathbf{R}_{\mathrm{reg}}(\mathrm{alg}; g)$. For any ξ, we have

$$\begin{aligned}
&\tau\mathbf{R}_{\mathrm{reg}}(\mathrm{alg}_{\mathrm{reg}}; g)\\
&= \tau\mathbb{E}_{S_{n,\mathrm{reg}}\sim\mathrm{law}_{\mathrm{reg}}(g)}\mathbb{E}_{\hat{g}\sim\mathrm{alg}_{\mathrm{reg}}(S_{n,\mathrm{reg}})}\left(\mathbb{E}_{\mathbf{z}\sim D_{\mathrm{reg}}}\|\hat{g}(\mathbf{z}) - g(\mathbf{z})\|^2\right)^{1/2}\\
&= \mathbb{E}_{\mathrm{S}_{n,H}\sim\mathrm{law}(g)}\mathbb{E}_{\hat{\pi}\sim\mathrm{alg}(\mathrm{S}_{n,H})}\left(\mathbb{E}_{\mathbf{z}\sim D_{\mathrm{reg}}}\left\|\mathbf{v}^\top\left(\mathbb{E}_{\mathbf{x}\sim\mathcal{K}(\mathbf{z})}\left[\mathrm{mean}[\hat{\pi}](\mathbf{x}, t=1) - \pi_0(\mathbf{x})\right]\right) - g(\mathbf{z})\right\|^2\right)^{1/2}\\
&\le \mathbb{E}_{\mathrm{S}_{n,H}\sim\mathrm{law}(g)}\mathbb{E}_{\hat{\pi}\sim\mathrm{alg}(\mathrm{S}_{n,H})}\left(\mathbb{E}_{\mathbf{z}\sim D_{\mathrm{reg}}}\mathbb{E}_{\mathbf{x}\sim\mathcal{K}(\mathbf{z})}\left\|\mathbf{v}^\top\left(\mathrm{mean}[\hat{\pi}](\mathbf{x}, t=1) - \pi_0(\mathbf{x})\right) - g(\mathbf{z})\right\|^2\right)^{1/2}\\
&\le \mathbb{E}_{\mathrm{S}_{n,H}\sim\mathrm{law}(g)}\mathbb{E}_{\hat{\pi}\sim\mathrm{alg}(\mathrm{S}_{n,H})}\left(\mathbb{E}_{\mathbf{z}\sim D_{\mathrm{reg}}}\mathbb{E}_{\mathbf{x}\sim\mathcal{K}(\mathbf{z})}\left\|\mathbf{v}^\top\mathrm{mean}[\hat{\pi}](\mathbf{x}, t=1) - \mathbf{v}^\top\pi_{g,\xi}(\mathbf{x})\right\|^2\right)^{1/2}\\
&= \mathbb{E}_{\mathrm{S}_{n,H}\sim\mathrm{law}(g)}\mathbb{E}_{\hat{\pi}\sim\mathrm{alg}(\mathrm{S}_{n,H})}\left(\mathbb{E}_{\mathbf{x}\sim D}\left\|\mathbf{v}^\top\mathrm{mean}[\hat{\pi}](\mathbf{x}, t=1) - \mathbf{v}^\top\pi_{g,\xi}(\mathbf{x})\right\|^2\right)^{1/2}
\end{aligned}$$

The steps used in each line are as follows: definition of $\mathbf{R}_{\mathrm{reg}}$; the definition of the estimator $\mathrm{alg}_{\mathrm{reg}}$ constructed from alg, and that $\mathrm{S}_{n,H} \sim \mathrm{law}(g)$ in that construction; Jensen's inequality; the formula for $\hat{\pi}_{g,\xi}$ given by Property G.1; and the fact the pushforward of D_{reg} under $\mathcal{K}$ is D. Continuing,

$$\begin{aligned}
&\le \mathbb{E}_{\mathrm{S}_{n,H}\sim\mathrm{law}(g)}\mathbb{E}_{\hat{\pi}\sim\mathrm{alg}(\mathrm{S}_{n,H})}\left(\mathbb{E}_{\mathbf{x}\sim D}\|\mathrm{mean}[\hat{\pi}](\mathbf{x}, t=1) - \pi_{g,\xi}(\mathbf{x})\|^2\right)^{1/2}\\
&\le \mathbb{E}_{\mathrm{S}_{n,H}\sim\mathrm{law}(g)}\mathbb{E}_{\hat{\pi}\sim\mathrm{alg}(\mathrm{S}_{n,H})}\left(\mathbb{E}_{\mathbf{x}\sim D}\mathbb{E}_{\hat{\mathbf{u}}\sim\hat{\pi}(\mathbf{x},1)}\|\hat{\mathbf{u}} - \pi_{g,\xi}(\mathbf{x})\|^2\right)^{1/2}\\
&\le \mathbf{R}_{\mathrm{train},h=1}(\mathrm{alg}; g) \le \mathbf{R}_{\mathrm{expert},L_1}(\mathrm{alg}; g).
\end{aligned}$$

where we use that $\mathbf{v}^\top(\cdot)$ is an orthogonal projection; Jensen's inequality again; the fact that the IL training risk is at least the ℓ_2 loss on the first prediction. ■

Lemma K.2 (Reduction from IL to regression) *For every regression algorithm* $\mathrm{alg}_{\mathrm{reg}}$*, there exists a regression algorithm* alg *such that for all* $g \in \mathcal{G}$*,*

$$\mathbf{R}_{\mathrm{expert},L_1}(\mathrm{alg}; g) = \tau\mathbf{R}_{\mathrm{reg}}(\mathrm{alg}_{\mathrm{reg}}; g) \tag{K.4}$$

Moreover, if $\mathrm{alg}_{\mathrm{reg}}$ *is proper, i.e. with probability* 1 *over its randomness* $\hat{g} \sim \mathrm{alg}_{\mathrm{reg}}$ *lies in* $\mathcal{G}$*, then so is* alg*. Lastly, it holds that if*

$$\inf_{\mathrm{alg}_{\mathrm{reg}}}\sup_{g_\star\in\mathcal{G}}\mathbb{E}_{S_{n,\mathrm{reg}}}\mathbb{E}_{\hat{g}\sim\mathrm{alg}_{\mathrm{reg}}(S_{n,\mathrm{reg}})}\mathbb{P}_{\mathbf{z}\sim D_{\mathrm{reg}},\mathbf{y}\sim\hat{g}(\mathbf{y})}[|g^\star(\mathbf{z}) - \mathbf{y}| \ge \epsilon] \ge \delta, \tag{K.5}$$

then, for any IL algorithm alg *and any* ξ*,*

$$\sup_{g\in\mathcal{G}} \mathbb{E}_{\mathrm{S}_{n,H}\sim(\pi_{\xi,g},f_{\xi,g})}\mathbb{E}_{\hat{\pi}\sim\mathrm{alg}(\mathrm{S}_{n,H})}\,\mathbb{P}_{\mathbf{x}\sim D,\mathbf{u}\sim\hat{\pi}(\mathbf{x})}[|\langle\pi_{g,\xi}(\mathbf{x})-\mathbf{u},\mathbf{v}\rangle|\geq\tau\cdot\epsilon]\geq\delta, \tag{K.6}$$

Proof We prove the first statement; the "Lasty," statement follows from a similar argument. Consider the map

$$\begin{aligned}\Phi:(\mathbf{x}_1,\mathbf{u}_1,\mathbf{x}_2,\mathbf{u}_2,\dots)&\mapsto(\mathrm{proj}(\mathbf{x}_1),\mathbf{v}^\top\mathbf{u}_1)\\ \hat{\pi}(\mathbf{x};\hat{g})&=\mathbf{v}\hat{g}(\mathrm{proj}(\mathbf{x}))+\pi_0(\mathbf{x}).\end{aligned}$$

We let alg be the algorithm which, given $\mathrm{S}_{n,H}\sim\mathrm{law}(g)$, construct $S_{n,\mathrm{reg}}=\Phi(\mathrm{S}_{n,H})$ and selects $\hat{g}=\mathrm{alg}_{\mathrm{reg}}(S_{n,\mathrm{reg}})$, and returns $\hat{\pi}$ such that $\hat{\pi}(\mathbf{x},1)=\hat{\pi}(\mathbf{x};\hat{g})$. By the definition of the one-step problem, we can define $\hat{\pi}(\mathbf{x},t)$ for $t>1$ in a manner independent of the instance $(\pi,f)\in\mathcal{P}$ and such that with probability one over $\mathbf{x}_2,\mathbf{x}_3,\dots$ under $\mathbb{P}_{\pi,f,D}$, $\hat{\pi}(\mathbf{x},t)=\pi(\mathbf{x},t)$. Doing so yields

$$\begin{aligned}\mathbf{R}_{\mathrm{train},h=1}(\mathrm{alg};g)^2&=\mathbb{E}_{\mathbf{x}\sim D}\|\hat{\pi}(\mathbf{x},t=1)-\pi_{g,\xi}(\mathbf{x},t=1)\|^2\\&=\tau^2\mathbb{E}_{\mathbf{x}\sim D}\|\mathbf{v}\hat{g}(\mathrm{proj}(\mathbf{x}))+\pi_0(\mathbf{x})-(\mathbf{v}g(\mathrm{proj}(\mathbf{x}))+\pi_0(\mathbf{x}))\|^2\\&=\tau^2\mathbb{E}_{\mathbf{x}\sim D}\|\hat{g}(\mathrm{proj}(\mathbf{x}))-g(\mathrm{proj}(\mathbf{x}))\|^2\\&=\tau^2\mathbb{E}_{\mathbf{z}\sim D_{\mathrm{reg}}}\|\hat{g}(\mathbf{z})-g(\mathbf{z})\|^2=\tau^2\mathbf{R}_{\mathrm{reg}}(\hat{g};g)^2\end{aligned}$$

Hence,

$$\begin{aligned}\mathbf{R}_{\mathrm{expert},L_1}(\mathrm{alg};g)&=\mathbb{E}_{\mathrm{S}_{n,H}\sim\mathrm{law}(g)}\mathbb{E}_{\hat{\pi}\sim\mathrm{alg}(\mathrm{S}_{n,H})}\mathbf{R}_{\mathrm{expert},L_2}(\hat{\pi};g)\\&=\tau\mathbb{E}_{S_{n,\mathrm{reg}}\sim\mathrm{law}_{\mathrm{reg}}(g)}\mathbb{E}_{\hat{g}\sim\mathrm{alg}(S_{n,\mathrm{reg}})}\mathbf{R}_{\mathrm{reg}}(\hat{g};g)=\tau\mathbf{R}_{\mathrm{reg}}(\mathrm{alg}_{\mathrm{reg}};g).\end{aligned}$$

Moreover, if $\mathrm{alg}_{\mathrm{reg}}$ is proper, note that $\hat{g}\in\mathcal{G}$ with probability one. Thus $\hat{\pi}(\mathbf{x};\hat{g})=\mathbf{v}\hat{g}(\mathrm{proj}(\mathbf{x}))+\pi_0(\mathbf{x})$ is equal to some $\pi_{\hat{g},\xi}(\mathbf{x},t=1)\in\Pi_{\mathcal{P}}$; moreover, by choosing $\hat{\pi}$ above to be equal to such a $\pi_{\hat{g},\xi}(\mathbf{x},t=1)$, we can verify that $\hat{\pi}$ incurs no IL training error for $t>1$, and $\mathbf{R}_{\mathrm{expert},L_2}(\hat{\pi};g)=\mathbf{R}_{\mathrm{train},h=1}(\hat{\pi};g)=\tau\mathbf{R}_{\mathrm{reg}}(\hat{g};g)$, The algorithm alg constructed this way is now proper, and satisfies meaning that $\mathbf{R}_{\mathrm{expert},L_1}(\mathrm{alg};g)=\tau\mathbf{R}_{\mathrm{reg}}(\mathrm{alg}_{\mathrm{reg}};g)$. ■

Proof [Proof of Proposition G.1] We begin with **part (a)**. Recall the notation $\mathbf{R}_{\mathrm{expert},L_1}(\mathrm{alg};g)$ from Definition K.1, which reflects the fact that the IL training risk is the same for all instances of the form$(\pi_{g,\xi},f_{g,\xi})$ for the same g but differing ξ. By Lemma K.2, we have

$$\begin{aligned}\mathbf{M}_{\mathrm{expert},L_2}(n;\mathcal{P},D)&=\inf_{\mathrm{alg}}\sup_{g\in\mathcal{G}}\mathbf{R}_{\mathrm{expert},L_1}(\mathrm{alg};g)\\&\leq\inf_{\mathrm{alg}_{\mathrm{reg}}}\sup_{g\in\mathcal{G}}\mathbf{R}_{\mathrm{reg}}(\mathrm{alg}_{\mathrm{reg}};g)=\mathbf{M}_{\mathrm{reg},L_2}(n;\mathcal{G},D_{\mathrm{reg}}).\end{aligned}$$

The reverse follows from from Lemma K.1, which establishes in fact that $\mathbf{M}_{\mathrm{reg},L_2}(n;\mathcal{G},D_{\mathrm{reg}})\leq\mathbf{M}_{\mathrm{expert},h=1}(n;\mathcal{P},D)$. As $\mathbf{M}_{\mathrm{expert},h=1}(n;\mathcal{P},D)\leq\mathbf{M}_{\mathrm{expert},L_2}(n;\mathcal{P},D)$ (the former only considers loss from $h=1$), we conclude that all three terms under consideration are equal.

Part (b). Lemma K.1 gives

$$\begin{aligned}\mathbf{M}^{\mathbb{A}}_{\mathrm{expert},L_2}(n;\mathcal{P},D) &= \inf_{\mathrm{alg}\in\mathbb{A}}\sup_{g\in\mathcal{G}}\mathbf{R}_{\mathrm{expert},L_1}(\mathrm{alg};g)\\ &\le \inf_{\mathrm{alg\ proper}}\sup_{g\in\mathcal{G}}\mathbf{R}_{\mathrm{expert},L_1}(\mathrm{alg};g)\\ &\le \inf_{\mathrm{alg}_{\mathrm{reg}}\ \mathrm{proper}}\sup_{g\in\mathcal{G}}\mathbf{R}_{\mathrm{reg}}(\mathrm{alg}_{\mathrm{reg}};g).\end{aligned}$$

When $\mathcal{G}$ is convex, restriction to proper estimators does not change the minimax rate:

$$\inf_{\mathrm{alg}_{\mathrm{reg}}\ \mathrm{proper}}\sup_{g\in\mathcal{G}}\mathbf{R}_{\mathrm{reg}}(\mathrm{alg}_{\mathrm{reg}};g) = \inf_{\mathrm{alg}_{\mathrm{reg}}}\sup_{g\in\mathcal{G}}\mathbf{R}_{\mathrm{reg}}(\mathrm{alg}_{\mathrm{reg}};g) = \mathbf{M}_{\mathrm{reg},L_2}(n;\mathcal{G},D_{\mathrm{reg}}).$$

This follows because on can always project the estimated function $\hat{g}$ on $\mathcal{G}$ in the metric $\|\cdot\|_{L_2(D_{\mathrm{reg}})}$, which by the Pythagorean theorem and convexity of $\mathcal{G}$ will never increase the loss. On the other hand, $\mathbf{M}^{\mathbb{A}}_{\mathrm{expert},L_2}(n;\mathcal{P},D) \ge \mathbf{M}_{\mathrm{expert},L_2}(n;\mathcal{P},D)$, and $\mathbf{M}_{\mathrm{expert},L_2}(n;\mathcal{P},D) = \mathbf{M}_{\mathrm{reg},L_2}(n;\mathcal{G},D_{\mathrm{reg}})$ by the first statement of this lemma. Thus,

$$\mathbf{M}_{\mathrm{reg},L_2}(n;\mathcal{G},D_{\mathrm{reg}}) = \mathbf{M}_{\mathrm{expert},L_2}(n;\mathcal{P},D) \le \mathbf{M}^{\mathbb{A}}_{\mathrm{expert},L_2}(n;\mathcal{P},D) = \mathbf{M}_{\mathrm{reg},L_2}(n;\mathcal{G},D_{\mathrm{reg}}),$$

proving the desired equality.

Part (c). We start by using the fact that distribution of $\mathrm{S}_{n,H}$ does not depend the realization of ξ. Hence, setting $\epsilon = \tau\kappa\boldsymbol{\epsilon}_n$,

$$\begin{aligned}&\sup_{g,\xi}\mathbb{E}_{\mathrm{S}_{n,H}\sim(\pi_{g,\xi},f_{g,\xi})}\mathbb{E}_{\hat{\pi}\sim\mathrm{alg}(\mathrm{S}_{n,H})}\mathbf{R}(pi;g,\xi)\\ &\ge \sup_{g}\mathbb{E}_{\xi\sim P}\mathbb{E}_{\mathrm{S}_{n,H}\sim(\pi_{g,\xi},f_{g,\xi})}\mathbb{E}_{\hat{\pi}\sim\mathrm{alg}(\mathrm{S}_{n,H})}\mathbf{R}(\pi;g,\xi)\\ &\overset{(i)}{=} \sup_{g}\mathbb{E}_{\xi\sim P}\mathbb{E}_{\mathrm{S}_{n,H}\sim(\pi_{g,\xi_0},f_{g,\xi_0})}\mathbb{E}_{\hat{\pi}\sim\mathrm{alg}(\mathrm{S}_{n,H})}\mathbf{R}(\pi;g,\xi)\\ &\overset{(ii)}{=} \sup_{g}\mathbb{E}_{\mathrm{S}_{n,H}\sim(\pi_{g,\xi_0},f_{g,\xi_0})}\mathbb{E}_{\hat{\pi}\sim\mathrm{alg}(\mathrm{S}_{n,H})}\mathbb{E}_{\xi\sim P}\mathbf{R}(\pi;g,\xi),\end{aligned}$$

where (i) uses the ξ-indistinguishability property (Property G.3) , and (ii) is a consequence of Fubini's theorem.

Next, by Eq. (G.4), we may lower bound (ii) via

$$\begin{aligned}&\inf_{\mathrm{alg}\in\mathbb{A}}\sup_{g}\mathbb{E}_{\mathrm{S}_{n,H}\sim(\pi_{g,\xi_0},f_{g,\xi_0})}\mathbb{E}_{\hat{\pi}\sim\mathrm{alg}(\mathrm{S}_{n,H})}\mathbb{E}_{\xi\sim P}\mathbf{R}_{\boldsymbol{\epsilon}_n\kappa\tau}(\pi;g,\xi)\\ &\quad\ge K\inf_{\mathrm{alg}\in\mathbb{A}}\sup_{g}\mathbb{E}_{\mathrm{S}_{n,H}\sim(\pi_{g,\xi_0},f_{g,\xi_0})}\mathbb{E}_{\hat{\pi}\sim\mathrm{alg}(\mathrm{S}_{n,H})}\,\mathbb{P}_{\mathbf{x}\sim D,\mathbf{u}\sim\hat{\pi}(\mathbf{x})}[|\langle\pi_{g,\xi_0}(\mathbf{x},t=1)-\mathbf{u},\mathbf{v}\rangle|\ge\kappa\tau\boldsymbol{\epsilon}_n]\\ &\quad\ge K\inf_{\mathrm{alg}_{\mathrm{reg}}}\sup_{g_\star\in\mathcal{G}}\mathbb{E}_{S_{n,\mathrm{reg}}}\mathbb{E}_{\hat{g}\sim\mathrm{alg}_{\mathrm{reg}}(S_{n,\mathrm{reg}})}\,\mathbb{P}_{\mathbf{z}\sim D_{\mathrm{reg}},\mathbf{y}\sim\hat{g}(\mathbf{y})}[|g^\star(\mathbf{z})-\mathbf{y}|\ge\kappa\boldsymbol{\epsilon}_n],\end{aligned}$$

where the last line follows from Lemma K.2, using convexity of $\mathcal{G}$ and the fact that $\mathbb{A}$ contains all proper algorithms. Finally, Condition E.1 implies that the above is at least $K\delta$. ■

Appendix L. Proof for Simple Policies, Theorems 1, 2 and 1.A

In this section, we prove Theorem 1.A. As noted below the statement of Theorem 1.A in Appendix F.1, Theorems 1 and 2 are direct consequences. Our aim is to make rigorous the intuitive proof sketched outlined in Section 4, by carefully instantiating the reduction given in Proposition G.1. We encourage the review to review that proposition before continuing to read this section.

We recall our asymptotic notation: $a = \mathrm{O}(b)$ to denote $a \le Cb$ for some universal constant C, and $a = o_\star(b)$ to mean "$a \le c \cdot b$ for c sufficiently small." We will also write $a = \Theta(b)$ do denote that there exists universal constants $c_1, c_2 > 0$ such that $c_1 a \le b \le c_2 b$. Finally, we will use the notation $a = \Theta_\star(b)$ to denote $a = o_\star(b)$ and $a = \Theta(b)$, that is, a is smaller than a sufficiently small universal constant times b, but no more than a constant smaller.

In what follows, Appendix L.1 provides the construction for the lower bound, explaining key simplifications and motivations. Appendix L.2 proceeds with a proof strategy. With this context, that section concludes by outlining the remainder of this Appendix, and describing the roles that the subsections that follow play in the overall proof.

L.1. Lower Bound Construction

As in Section 4, our lower bound forces the learner to make a single error at step $h = 1$, and shows that this error compounds exponentially in H. To do so, we effectively "patch together" a region of space in which the learner needs to learn the embedded regression family $(\mathcal{G}, D_{\mathrm{reg}})$, and a region where the dynamics and optimal policy follow the linear construction detailed in Definition 4.1. We separate these regions via *bump functions*, a construction widespread in mathematical analysis and statistical learning. We recall the salient properties of the bump function here.

Lemma H.14 (Existence of Bump Functions) *For any $k \in \mathbb{N}$, there exists a C^∞ function $\mathrm{bump}_k(\mathbf{z}) : \mathbb{R}^k \to \mathbb{R}$, called an* bump function, *sastisfying $\mathrm{bump}_k(\mathbf{z}) = 1$ if and only if $\|\mathbf{z}\| \le 1$, $\mathrm{bump}_k(\mathbf{z}) = 0$ if and only if $\|\mathbf{z}\| \ge 2$. And, for each $p \ge 1$, $\|\nabla^p \mathrm{bump}_k(\mathbf{z})\|_{\mathrm{op}} \le c_p$, where $\|\cdot\|_{\mathrm{op}}$ denotes the tensor-operator norm, and c_p is a constant independent of k but depending on p. Finally, $\nabla^p \mathrm{bump}_k(\mathbf{z}) = 0$ for all $\mathbf{z} : \|\mathbf{z}\| \ge 2$.*

Next, we introduce our formal construction. Recall that, in line with Proposition G.1, we parametrize our instance class with instances of the form $(\pi_{g,\xi}, f_{g,\xi})$, where g encodes the function to be estimated at the first time step, and ξ parametrizes remaining uncertainty.

Construction L.1 (Embedding Construction) *Let $\tau, \Delta \in (0,1)$ be parameters to be chosen. We shall choose $\tau = \Theta_\star(1)$ and $\Delta = \Theta_\star\left(\frac{1}{ML\sqrt{d}}\right)$. Define the matrices $\bar{\mathbf{A}}_i, \bar{\mathbf{K}}_i$ via*

$$\bar{\mathbf{A}}_i := \begin{bmatrix} \mathbf{A}_i & \mathbf{0}_{2\times d} \\ \mathbf{0}_{d\times 2} & \mathbf{0}_{d\times d} \end{bmatrix}, \quad \bar{\mathbf{K}}_i = \begin{bmatrix} \mathbf{K}_i & \mathbf{0}_{2\times d} \\ \mathbf{0}_{d\times 2} & \mathbf{0}_{d\times d} \end{bmatrix}, \qquad \text{(Dynamical Matrices)}$$

where above $\mathbf{A}_i$ and $\mathbf{K}_i$ are the matrices in Definition 4.1, with $\mu \leftarrow 1/4$. Furthermore, let $\mathrm{Proj}_{\ge 3}$ denote the cannonical mapping from $\mathbb{R}^d$ to $\mathbb{R}^{d-2}$ which removes the first two coordinates. Define the function $\mathrm{restrict}(\mathbf{x})$ and transformation $\mathcal{T}[g]$ via

$$\mathrm{restrict}(\mathbf{x}) := \mathrm{bump}_d(\mathbf{x} - \mathbf{x}_{\mathrm{offset}}),\ \mathbf{x}_{\mathrm{offset}} := 3\mathbf{e}_3, \qquad \mathcal{T}[g](\mathbf{x}) := g(\mathrm{Proj}_{\ge 3}(\mathbf{x} - \mathbf{x}_{\mathrm{offset}})),$$

Let ξ denote pairs $\xi = (i, \omega)$, where $i \in \{1, 2\}$, $\omega \in \{-1, 1\}$. We define the instances $(\pi_{g,\xi}, f_{g,\xi})$ via

$$\begin{aligned} \pi_{g,\xi}(\mathbf{x}) &= \bar{\mathbf{K}}_i \mathbf{x} + \tau \cdot \mathrm{restrict}(\mathbf{x}) \cdot \mathcal{T}[g](\mathbf{x})\mathbf{e}_1 \\ f_{g,\xi}(\mathbf{x}, \mathbf{u}) &= \bar{\mathbf{A}}_i \mathbf{x} + \mathbf{u} - \tau \cdot \mathrm{restrict}(\mathbf{x}) \cdot \mathcal{T}[g](\mathbf{x})\mathbf{e}_1 \\ &\quad + \omega \cdot \tau^2 \cdot \mathrm{restrict}(\mathbf{x}) \cdot \mathbf{e}_1 \cdot (\mathcal{T}[g](\mathbf{x}) - \langle \mathbf{e}_1, \mathbf{u} \rangle \mathrm{bump}_d(\mathbf{u})/\tau)\,. \end{aligned} \tag{Instance Class}$$

Finally, given a 1-bounded distribution D_{reg} on $\mathbb{R}^{d-2}$, define a distribution on $\mathbb{R}^d$ via,

$$D = D(D_{\mathrm{reg}}) \overset{d}{=} \mathbf{I}\{Z = 0\} \cdot (\mathbf{x}_{\mathrm{offset}} + (0, 0, \mathbf{z})) + \mathbf{I}\{Z = 1\}(\Delta \cdot Y \cdot \mathbf{w}) \tag{Initial State Distribution}$$

where $Z \sim \mathrm{Bernoulli}(1/2)$, $\mathbf{z} \in \mathbb{R}^{d-2} \sim D_{\mathrm{reg}}$, $\mathbf{w}$ is drawn uniformly on the unit ball supported on coordinates 2-through-d: $\{\mathbf{w} : \sum_{i=2}^d (\mathbf{e}_i^\top \mathbf{w})^2 \le 1\}$, Y is a nonnegative random variable with $\mathbb{P}[Y = 1] = 1/2$ and $\mathbb{P}[Y = 2^{-k}] \propto 1/k^2$ for $k \ge 1$, and where $(Z, Y, \mathbf{w}, \mathbf{z})$ are independent random variables.

The difference between $Z = 0$ and $Z = 1$ cases is essential in the argument, which warrants us establishing a convenient shorthand.

Definition L.1 *We define the shorthand $D_{\{Z=z\}}$, $z \in \{0, 1\}$ to denote the conditional distribution of D given $Z = z$.*

Explanation of Construction L.1 . The construction involves a number of daunting and complicated-seeming terms designed to carefully restrict the dynamics to ensure various global smoothness and stability properties, detailed in Appendix L.6. However, much is simplified by considering behavior of the dynamics at an initial state $\mathbf{x}_1$ drawn from D.

Claim L.1 *Consider instance $(\pi_{g,\xi}, f_{g,\xi})$ from Construction N.1, with $\xi = (i, \omega) \in \{1, 2\} \times \{-1, 1\}$. Let $\mathbf{x} \sim D$. If $Z = 0$ and $\mathbf{x} = \mathbf{x}_{\mathrm{offset}} + (0, 0, \mathbf{z})$ for $\mathbf{z} \in \mathbb{R}^{d-2}$, and let $\|\mathbf{u}\| \le 1$. Then,*

$$\pi_{g,\xi}(\mathbf{x}) = \tau g(\mathbf{z})\mathbf{e}_1, \quad f_{g,\xi}(\mathbf{x}, \mathbf{u}) = \mathbf{u} - (1 - \omega\tau)\mathbf{e}_1 \cdot (\pi_{g,\xi}(\mathbf{x}) - \langle \mathbf{e}_1, \mathbf{u} \rangle).$$

On the other hand, if $Z = 1$, then

$$\pi_{g,\xi}(\mathbf{x}) = \bar{\mathbf{K}}_i \mathbf{x}, \quad f_{g,\xi}(\mathbf{x}, \mathbf{u}) = \mathbf{u} + \bar{\mathbf{A}}_i \mathbf{x}.$$

Proof When $Z = 0$, the $\mathrm{restrict}(\mathbf{x})$ term is equal to 1, $\mathcal{T}[g](\mathbf{x}) = g(\mathbf{z})$. And when $\|\mathbf{u}\| \le 1$, $\mathrm{bump}_d(\mathbf{u}) = 1$. When $Z = 0$, the $\mathrm{restrict}(\mathbf{x})$ term is equal to zero. Applying these simplifications to Construction L.1 establishes the claim. ■

It is now more transparent to see how Construction L.1 implies the plan described in Section 4. The case $Z = 0$ is responsible for introducing statistical error which is to be compounded, and the case $Z = 1$ provides information about the linear regime of the expert and dynamics, but only along the subspace perpendicular to $\mathbf{e}_1$ (recall, $\mathbf{x}_1 \mid Z = 1$ is distributed uniformly on the sphere on coordinates 2-d). This forces the Jacobian of the mean of the learner policy to correspond to the $\bar{\mathbf{K}}_i$ matrices on that subspace. For the proof of the present theorem (Theorem 1.A), we only leverage the $Y = 1$ subcase of $Z = 1$ to make this argument, but the $Y > 1$ cases are useful in the proof of

Theorem 3.A, and to simplify the statements all theorems, we opted to allow the distribution D to be the same for both results.

Even with these simplifications, there is the additional parameter τ, and indices i and ω, that arise. The parameter τ is chosen to be sufficiently small that the nonlinear terms in the dynamics are overwhelmed by the linear terms. This ensures global exponentiall incremental stability. The indices $i \in \{1, 2\}$ induce uncertainty over the challenging pair of dynamical system $(\bar{\mathbf{A}}_i, \bar{\mathbf{K}}_i)$, which embed the $(\mathbf{A}_i, \mathbf{K}_i)$ defined in Definition 4.1. Finally, the parameter $\omega \in \{-1, 1\}$ gives uncertainty over the sign of the error made along the $\mathbf{e}_1$ access when $Z = 0$. Before continuing, we verfy that the matrices $\bar{\mathbf{A}}_i$ and $\bar{\mathbf{A}}_i + \bar{\mathbf{K}}_i$ are indeed stable.

Lemma L.2 *There exists some $C \geq 1$ and $\rho \in (0, 1)$ such that both $\bar{\mathbf{A}}_i$ and $\bar{\mathbf{A}}_i + \bar{\mathbf{K}}_i$ are (C, ρ) stable for $i \in \{1, 2\}$.*

Proof Recall that a matrix $\mathbf{A}$ is (C, ρ) stable if $\|\mathbf{A}^k\|_{\mathrm{op}} \leq C\rho^k$. As block-diagonal matrices are preserved under matrix powers, and operator norms decompose as maxima across blocks, we see that a block diagonal matrix $\mathbf{A}$ is (C, ρ)-stable if and only if its blocks are. The top blocks of $\bar{\mathbf{A}}_i$ and $\bar{\mathbf{A}}_i + \bar{\mathbf{K}}_i$ are the matrices $\mathbf{A}_i$ and $\mathbf{A}_i + \mathbf{K}_i$, whose stability is ensured by Proposition 4.1. The remaining block is the zero matrix, which is clearly (C, ρ) for any $C, \rho \geq 0$. ■

Cost functions. We will use a *single*, *time-invariant* cost function which witnesses the separation between expert and imitator policies. The cost is constructed to carefully vanish on expert trajectories, whilst exposing large errors along the $\mathbf{e}_1$ direction. In view of Remark J.1, we only need the cost to the maximum of costs which are nonnegative and Lipschitz, but not necessarily bounded above by 1.

Construction L.2 (Challenging Cost) *Let C_Δ be the universal constant in Lemma L.8. We define*

$$\begin{aligned}\mathrm{cost}_{\mathrm{hard}}(\mathbf{x}, \mathbf{u}) = C_{\mathrm{cost}}|\langle \mathbf{e}_1, \mathbf{x}\rangle| &+ C_{\mathrm{cost}}\left(\|\mathbf{u} - \bar{\mathbf{K}}_1\mathbf{x}\| + \|\mathbf{u} - \bar{\mathbf{K}}_2\mathbf{x}\|\right)\mathrm{bump}\left(\frac{\mathbf{x}}{2}\right) \\ &+ C_{\mathrm{cost}}\Delta\left(1 - \mathrm{bump}(\mathbf{x} - \mathbf{x}_{\mathrm{offset}})\right)\left(1 - \mathrm{bump}\left(\frac{\mathbf{x}}{C_\Delta \Delta}\right)\right) \\ &+ \tau C_{\mathrm{cost}}(1 - \mathrm{bump}(\mathbf{u}/\tau)) \\ &+ C_{\mathrm{cost}}\left(\mathrm{bump}(\mathbf{x} - \mathbf{x}_{\mathrm{offset}})\right)\|(\mathbf{I} - \mathbf{e}_1\mathbf{e}_1^\top)\mathbf{u}\|.\end{aligned}$$

In terms of this, we we define

$$\overline{\mathrm{cost}}_{\mathrm{hard}}(\mathbf{x}_{1:H}, \mathbf{u}_{1:H}) := \max_{1 \leq t \leq H} \mathrm{cost}_{\mathrm{hard}}(\mathbf{x}_t, \mathbf{u}_t)$$

We now show that the cost vanishes under the experts demonstration distribution, and is Lipschitz.

Lemma L.3 *For $\Delta = o_\star(\tau), \tau = o_\star(1)$, it holds that $\overline{\mathrm{cost}}_{\mathrm{hard}} \in \mathcal{C}_{\mathrm{vanish}}(\mathcal{P}, D)$, i.e., vanished on $(\mathcal{P}, D)$:*

$$\sup_{(\pi^\star, f) \in \mathcal{P}} \mathbb{P}_{\pi^\star, f, D}[\overline{\mathrm{cost}}_{\mathrm{hard}}(\mathbf{x}_{1:H}, \mathbf{u}_{1:H}) \neq 0] = \sup_{(\pi^\star, f) \in \mathcal{P}} \mathbb{P}_{\pi^\star, f, D}[\exists t : \mathrm{cost}_{\mathrm{hard}}(\mathbf{x}_t, \mathbf{u}_t) \neq 0] = 0. \tag{L.1}$$

Proof Observe that $\mathrm{cost}_{\mathrm{hard}}(\mathbf{x}, \mathbf{u})$ with $\mathbf{u} = \pi_{g,\xi}(\mathbf{x})$ vanishes whenever either $\mathbf{x}$ is supported on coordinates $3, \dots, d$ (all the linear terms vanish) and in a unit ball around $\mathbf{x}_{\mathrm{offset}} = 3\mathbf{e}_3$ (the bump function term vanishes, and on that ball, $\mathbf{u}$ lies only in the $\mathbf{e}_1$ direction), or is supported on coordinates $2, \dots, d$ (the $\langle \mathbf{e}_1, \mathbf{x} \rangle$ term vanishes) and lies in of radius $\min\{C_\Delta \Delta, o_\star(\tau)\}$ around the origin (the bump function terms vanish, and $\bar{\mathbf{K}}_i \mathbf{x}$ is the same for both i, and $\|\mathbf{u}\| \le \|\bar{\mathbf{K}}_i \mathbf{x}\| \le O(\tau)$ when $\|\mathbf{x}\| \le O(\tau) \le o_\star(1)$). Lemma L.5 ensures the former situation under $Z = 0$ and time step 1, and the latter under time steps $t > 1$, and Lemma L.8 ensures the latter under $Z = 1$ or $Z = 2$, for all timesteps. ■

Lemma L.4 *There is a choice of $C_{\mathrm{cost}} = \Theta(1)$ for which $\mathrm{cost}_{\mathrm{hard}}$ is 1-Lipschitz, and nonnegative. Hence, $\overline{\mathrm{cost}}_{\mathrm{hard}} \in \tilde{\mathcal{C}}_{\mathrm{lip,max}}$.*

Proof This follows from the fact that bump functions are $O(1)$-Lipschitz. ■

L.2. Overall Proof Strategy

Recall that $D_{\{Z=z\}}$ denoste the distribution of D conditioned on the event $\{Z = z\}$. Our proof strategy is as follows:

- The distribution $D_{\{Z=1\}}$ forces any $\hat{\pi}$ with low error to satisfy $\hat{\pi}(\mathbf{x}) \approx \pi_{g,\xi}(\mathbf{x}) = \bar{\mathbf{K}}_i \mathbf{x}$ on average over the Δ-ball. By smoothness of $\hat{\pi}$, and by taking Δ to be of appropriate magnitude, this forces the projection of the Jacobian of $\hat{\pi}$ along the directions spanned by the coordinates $\{2, 3, \dots, d\}$ to match those of $\pi^\star$. Technical details for this section are derived in Appendix L.3, and rely on smoothness of $\mathrm{mean}[\hat{\pi}]$, as well as some convenient properties of expectations under the uniform distribution on the unit ball (notably, anti-concentration, which is slightly stronger than necessary for this argument, but ends up being useful in the proof of Theorem 3.A). This argument is similar in spirit to popular zero-order gradient estimators (see, e.g. Flaxman et al. (2004)), and role of the parameter Δ is to trade of between the quality of the Taylor approximation (which improves for smaller Δ) and effective variance of the Jacobian estimate (which, after appropriate normalizatiom, degrades with Δ small).

  Fixing the Jacobian of $\mathrm{mean}[\hat{\pi}]$ ensures that $\nabla \mathrm{mean}[\hat{\pi}](\mathbf{0})$ takes the form

  $$\nabla \mathrm{mean}[\hat{\pi}](\mathbf{0}) \approx \begin{bmatrix} \star & -c_\mu & \mathbf{0} \\ \star & 0 & \mathbf{0} \\ \star & \mathbf{0} & \mathbf{0} \end{bmatrix}. \tag{L.2}$$

  Following the argument of Proposition 4.1, this implies that, for at least one of $i \in \{1, 2\}$,

  $$\mathbf{A}_i + \nabla \mathrm{mean}[\hat{\pi}](\mathbf{0}) \approx \begin{bmatrix} a & 0 & \mathbf{0} \\ \star & 1 - 2\mu & \mathbf{0} \\ \star & \mathbf{0} & \mathbf{0} \end{bmatrix}, |a| \ge 1 + \frac{\mu}{4}, \tag{L.3}$$

  which is a matrix with single unstable eigenvector $\mathbf{e}_1$.

- The distribution $D_{\{Z=0\}} = D\big|_{Z=0}$ embeds the supervised learning problem associated with the class $\mathcal{G}$. It is designed such that it conveys no further information about the parameters $\xi = (i, \omega)$. Moreover, by randomizing over the ω, we force errors along the $\mathbf{e}_1$ direction. Specifically, for $\mathbf{x} = (0, 0, \mathbf{z}) \sim D_{\{Z=1\}}$ and for $\|\mathbf{u}\| \le 1$ (otherwise, $\overline{\mathrm{cost}}_{\mathrm{hard}}$ is large), Claim L.1 shows that

$$f_{g,\xi=(i,1)}(\mathbf{x}, \mathbf{u}) - f_{g,\xi=(i,-1)}(\mathbf{x}, \mathbf{u}) = 2\tau^2 \cdot (g(\mathbf{z}) - \langle \mathbf{e}_1, \mathbf{u} \rangle)\, \mathbf{e}_1, \tag{L.4}$$

  Hence, we make statistical errors along the $\mathbf{e}_1$ direction proportional to our mis-estimation of $g(\mathbf{z})$.

  Moreover, our construction ensures that the time step $t = 2$ is in the region in which the dynamics f are given by the linear function. $f(\mathbf{x}, \mathbf{u}) = \mathbf{A}_i + \mathbf{u}$. These arguments are given in Appendix L.3.

- We now invoke a quantitative variant of the unstable manifold theorem (Proposition H.1), applying an argument similar to an efficient saddle-point escape introduced in (Jin et al., 2017) (but generalized to account for non-symmetric Jacobians and stripped of inessential details). This shows that for the choice of i for which Eq. (L.3) holds, the autonomous dynamical system $F(\mathbf{x}) = \mathbf{A}_i\mathbf{x} + \hat{\pi}(\mathbf{x})$ is exponentially unstable to perturbations along the $\mathbf{e}_1$ direction. Consequently, when $Z = 0$, either the $(i, \omega = -1)$ and $(i, \omega = +1)$ dynamics divergence proportional to the estimation error of g, in view of Eq. (L.4). We emphasize that Proposition H.1 is the technical cornerstone of the entire lower bound argument. Building upon this argument, we establish a comprehensive statement of compounding error, whose presentation and proof are given in Appendix L.4

- To conclude, Appendix L.5 applies the reduction in Proposition G.1 to show that, the error at time step $t = 1$ along $\mathbf{e}_1$ when $\{Z = 0\}$, which is proportional to $|\mathcal{T}[g](\mathbf{x}) - \langle \mathbf{e}_1, \mathbf{u} \rangle|$, scales with the error of the embed regression problem, $\Omega(\mathsf{M}_{\mathrm{reg},L_2}(n; \mathcal{G}, D_{\mathrm{reg}}))$. We apply other ideas in that same reduction to relate the minimax risks under BC training, regression, and BC training restricted to estimators in $\mathbb{A}$.

- Finally, Appendix L.6 we verifies the various regularity conditions (smoothness, boundedness, stability). Here, the parameter τ plays a role in ensuring that the nonlinear terms do not overwhelm the stability guaranteed by the linear terms.

L.3. Analysis of the $Z \in \{0, 1\}$ cases

Here, we establish essential properties of the demonstration distribution, according to the value of the Bernoulli variable Z.

L.3.1. CASE $Z = 0$.

On these trajectories, the learner sees samples $\mathbf{z}$ from the regression distribution D_{reg}, embedded into dimension d by appending two zero coordinates, and shifting by $3\mathbf{e}_1$. These $\mathbf{x}_1$ take the form $\mathbf{x}_1 = (0, 0, \mathbf{z})$: their first two coordinates are zero, which implies that $\bar{\mathbf{K}}_i\mathbf{x}_1 = 0$ and that expert policy selects $\pi_{g,\xi}(\mathbf{x}) = \mathcal{T}[g](\mathbf{x})\mathbf{e}_1 = g(\mathbf{z})\mathbf{e}_1$. Thus, the event $\{Z = 0\}$ embeds the regression problem.

. Notice that the expert action is exactly canceled by the dynamics, as $\bar{\mathbf{A}}_i \mathbf{x}_1 = 0$ (again, the first two coordinates of $\mathbf{x}_1$ vanish). As $|g(\mathbf{z})| \le 1$, $\mathbf{u}_1 = \pi_{g,\xi}(\mathbf{x}_1)\mathbf{e}_1$ also satisfies $\mathrm{bump}_d(\mathbf{u}_1) = 1$, and thus we find that for $\mathbf{x} \leftarrow \mathbf{x}_1$ and $\mathbf{u} \leftarrow \pi_{g,\xi}(\mathbf{x}_1)\mathbf{e}_1$,

$$
\begin{aligned}
f_{g,\xi}(\mathbf{x}, \mathbf{u}) &= \bar{\mathbf{A}}_i \mathbf{x} + \underbrace{\mathbf{u} - \tau \cdot \mathrm{restrict}(\mathbf{x}) \cdot \mathcal{T}[g](\mathbf{x})\mathbf{e}_1}_{=0} \\
&\qquad + \omega \cdot \tau^2 \cdot \mathrm{restrict}(\mathbf{x}) \cdot \mathbf{e}_1 \cdot \left(\underbrace{\mathcal{T}[g](\mathbf{x}) - \langle \mathbf{e}_1, \mathbf{u} \rangle \mathrm{bump}_d(\mathbf{u}/4)/\tau}_{=0} \right) \\
&= \bar{\mathbf{A}}_i \mathbf{x} = 0,
\end{aligned}
$$

where the last line uses the fact that $\mathbf{x}$ is supported on the last $d-2$ coordinates, and the block structure of $\bar{\mathbf{A}}_i$. This establishes the following:

Lemma L.5 *Conditioned on $Z = 0$, the expert trajectories take the form $\mathbf{x}_1 = \mathbf{x}_{\mathrm{offset}} + (0, 0, \mathbf{z})$, $\mathbf{z} \sim D_{\mathrm{reg}}$, $\mathbf{u}_1 = \tau \mathbf{e}_1 g(\mathbf{z})$, $\mathbf{x}_h = \mathbf{u}_h \equiv \mathbf{0}$ for $h > 1$.*

In addition to characterizing the expert behavior on these trajectories, we also cheeck that unless $\overline{\mathrm{cost}}_{\mathrm{hard}}$ grows large, the dynamics conditioned on $Z = 0$ are linear.

Lemma L.6 *Let $(\mathbf{x}_{1:H}, \mathbf{u}_{1:H})$ be a trajectory under dynamics $f_{g,i,\xi}$ for which $\mathbf{x}_1 \in \mathrm{support}(D \mid Z = 0)$. Suppose that*

$$
\overline{\mathrm{cost}}_{\mathrm{hard}}(\mathbf{x}_{1:H}, \mathbf{u}_{1:H}) \le \epsilon \le o_\star(\tau), \tag{L.5}
$$

Then, for all $2 \le t \le H$, we have

$$
\mathbf{x}_{t+1} = \bar{\mathbf{A}}\mathbf{x}_t + \mathbf{u}_t, \quad \max\{\|\mathbf{x}_t\|, \|\mathbf{u}_t\|\} \le O(\epsilon)
$$

Proof Assume Eq. (L.5). Define $\epsilon' = \epsilon / C_{\mathrm{cost}} \ge \epsilon$. If $\epsilon' \le \tau \le 1$, then, from the definition of $\mathrm{cost}_{\mathrm{hard}}$, $\|\mathbf{u}_1\| \le \tau$, so that (using $\bar{\mathbf{A}}_i \mathbf{x}_1 = 0$ when $\mathbf{x}_1 \in \mathrm{support}(D \mid Z = 0)$)

$$
\begin{aligned}
\mathbf{x}_2 &= \bar{\mathbf{A}}_i \mathbf{x}_1 + \mathbf{u}_1 + \omega \cdot \tau \cdot \mathrm{restrict}(\mathbf{x}_1) \cdot \mathbf{e}_1 \cdot (\tau \mathcal{T}[g](\mathbf{x}_1) - \langle \mathbf{e}_1, \mathbf{u}_1 \rangle), \\
&= \mathbf{u}_1 + \omega \tau \mathbf{e}_1 \cdot (\tau \cdot \mathcal{T}[g](\mathbf{x}_1) - \langle \mathbf{e}_1, \mathbf{u}_1 \rangle),
\end{aligned}
$$

We also have $\|(\mathbf{I} - \mathbf{e}_1\mathbf{e}_1^\top)\mathbf{x}_2\| = \|(\mathbf{I} - \mathbf{e}_1\mathbf{e}_1^\top)\mathbf{u}_1\| \le \epsilon'$, and $|\langle \mathbf{e}_1, \mathbf{x}_2 \rangle| \le \mathrm{cost}_{\mathrm{hard}}(\mathbf{x}_2, \mathbf{u}_2)/C_{\mathrm{cost}} \le \epsilon'$. Hence, $\|\mathbf{x}_2\| \le 2\epsilon'$. Lastly, set $\|\delta \mathbf{u}_t\| = \|\mathbf{u}_t - \bar{\mathbf{K}}_i \mathbf{x}_t\| \le \epsilon \le \epsilon'$. Then, if $\|\mathbf{x}_2\|, \dots, \|\mathbf{x}_t\|, \|\mathbf{u}_2\|, \dots, \|\mathbf{u}_t\| \le 1/2$ we have

$$
\mathbf{x}_{t+1} = (\bar{\mathbf{A}}_i + \bar{\mathbf{K}}_i)\mathbf{x}_t + \delta \mathbf{u}_t = \left(\sum_{i=2}^{t} ((\bar{\mathbf{A}}_i + \bar{\mathbf{K}}_i))^{t-i} \delta \mathbf{u}_i \right) + (\bar{\mathbf{A}}_i + \bar{\mathbf{K}}_i))^{t-1} \mathbf{x}_2. \tag{L.6}
$$

By Lemma L.2 $(\bar{\mathbf{A}}_i + \bar{\mathbf{K}}_i)$ is (C, ρ)-stable for universal $C, \rho \in (0, 1)$. Thus, bounding the geometric series and using the fact that the magnitude of $\mathbf{x}_2, \delta \mathbf{u}_i$ are atmost $2\epsilon'$ and ϵ', respectively, we find

$$
\|\mathbf{x}_{t+1}\| = O(\epsilon') = O(\epsilon), \tag{L.7}
$$

where the $O(\epsilon')$ hides a constant of $C/(1-\rho)$, not depending on t. Hence, for $\epsilon = o_\star(1)$, we conclude that $\|\mathbf{x}_t\| = O(\epsilon)$ for all $2 \le t \le H$. Similarly, we have $\mathbf{u}_t = \bar{\mathbf{K}}_i \mathbf{x}_i + \delta \mathbf{u}_t = O(\epsilon)$ for all t. Taking $\epsilon = o_\star(1)$ enures that $\|\mathbf{x}_{t+1}\|, \|\mathbf{x}_{t+1} \le 1/2$, completing the induction. ∎

L.3.2. CASE $Z = 1$

The purpose of the $Z = 1$ case is to force the Jacobian of the mean of the learner's policy to approximate $\bar{\mathbf{K}}_i$ on the subspace spanned by the cannonical basis vectors $\mathbf{e}_2, \ldots, \mathbf{e}_d$. The following lemma makes this precise:

Lemma L.7 *Let $\mathrm{Proj}_{\geq 2}$ denote the projection onto coordinates 2-through-d, and let $\hat{\pi}$ be any M-smooth simply-stochastic policy. Then, if*

$$\mathbb{P}_{\hat{\pi}, f_{g,(i,\omega)}}[\overline{\mathrm{cost}}_{\mathrm{hard}}(\mathbf{x}_{1:H}, \mathbf{u}_{1:H}) \geq M\Delta^2/2] \leq o_\star(1), \tag{L.8}$$

we have the bound $\|(\hat{\mathbf{K}} - \bar{\mathbf{K}}_i)\mathrm{Proj}_{\geq 2}\|_{\mathrm{F}} \leq 6M\Delta\sqrt{d}$.

Proof Suppose $\mathbb{P}_{\hat{\pi}, f_{g,(i,\omega)}, D}[\overline{\mathrm{cost}}_{\mathrm{hard}}(\mathbf{x}_{1:H}, \mathbf{u}_{1:H}) \geq \epsilon] \leq c_0$, where c_0 is a sufficiently small constant to be chosen. Let $\mathcal{D}_{\{Z=1,Y=k\}}$ is the distribution of $\mathbf{x} \mid Z = 1, Y = k$. Because $\mathbb{P}[Z = 1, Y = 1] = 1/4$,

$$\mathbb{E}_{\mathbf{x}_1 \sim \mathcal{D}_{\{Z=1,Y=1\}}} \mathbb{E}_{\mathbf{u} \sim \hat{\pi}(\mathbf{x}_1)}[\|\bar{\mathbf{K}}_i \mathbf{x}_1 - \mathbf{u}\| \geq \epsilon] \leq 4c_0. \tag{L.9}$$

By simple-stochasticity, there is a coupling $\hat{P}(\mathbf{x}', \mathbf{x})$ over random inputs $\mathbf{u}' \sim \hat{\pi}(\mathbf{x}'), \mathbf{u} \sim \hat{\pi}(\mathbf{x})$ where $\mathbf{u}' - \mathbf{u} = \mathrm{mean}[\hat{\pi}](\mathbf{x}') - \mathrm{mean}[\hat{\pi}](\mathbf{x})$. Thus, by the triangle inequality and a union bound, we can symmetrize to obtain

$$\begin{aligned}
&\mathbb{P}_{\mathbf{x}, \mathbf{x}' \overset{\text{i.i.d}}{\sim} \mathcal{D}_{\{Z=1,Y=1\}}}[\|\bar{\mathbf{K}}_i(\mathbf{x}' - \mathbf{x}) - (\mathrm{mean}[\pi](\mathbf{x}') - \mathrm{mean}[\pi](\mathbf{x}))\| \geq 2\epsilon] \\
&= \mathbb{E}_{\mathbf{x}', \mathbf{x} \overset{\text{i.i.d}}{\sim} \mathcal{D}_{\{Z=1,Y=1\}}} \mathbb{P}_{\mathbf{u}', \mathbf{u} \sim \hat{P}(\mathbf{x}', \mathbf{x})}[\|\bar{\mathbf{K}}_i(\mathbf{x}' - \mathbf{x}) - (\mathbf{u}' - \mathbf{u})\| \geq 2\epsilon] \\
&\leq 2\mathbb{E}_{\mathbf{x}_1 \mid Z=1} \mathbb{P}_{\mathbf{u} \sim \hat{\pi}(\mathbf{x}_1)}[\|\bar{\mathbf{K}}_i \mathbf{x}_1 - \mathbf{u}\| \geq \epsilon] \leq 8c_0.
\end{aligned}$$

Since $\mathcal{D}_{\{Z=1,Y=1\}}$ has $\mathbf{x}_1$ drawn from the uniform distribution on the unit ball over coordinates 2-through-d, the result now follows from a technical Lemma H.9 by taking $\epsilon = M\Delta^2/2$, and using the assumption that $\mathbf{x} \mapsto \mathrm{mean}[\pi](\mathbf{x})$ is M-smooth. ■

Whilst the $Z = 1$ case forces the learner's policy to resemble $\bar{\mathbf{K}}_i$ on appropriate coordinate, it does so *without conveying any information about the instance*.

Lemma L.8 *There is a universal and dimension-independent constant Δ_0 such that, if $\Delta \leq \Delta_0$, the distribution of $(\mathbf{x}_1, \ldots, \mathbf{x}_H)$ under $\mathbb{P}_{\pi_{g,\xi}, f_{g,\xi}, D}[\cdot \mid Z = 1]$ does not depend on (g, ξ), and moreover, $\max_t \|\mathbf{x}_t\| \leq C_\Delta \cdot \Delta$, where C_Δ is a universal constant.*

Proof The "Moreover," part is clear from the construction. For the first part, there exists (C, ρ) such that $(\mathbf{A}_i + \mathbf{K}_i)$ is (C, ρ)-stable for some $\rho < 1$ and $C < \infty$, and both of $i \in \{1, 2\}$. Using the block structure, this implies the same for $(\bar{\mathbf{A}}_i, \bar{\mathbf{K}}_i)$. By inflating C if necessary (note $\|\bar{\mathbf{K}}_i\| = \|\mathbf{K}_i\|$ is dimension independent) we may ensure that $\sup_{n \geq 0}\{\|(\bar{\mathbf{A}}_i + \bar{\mathbf{K}}_i)^n \bar{\mathbf{K}}_i)\|, \|(\bar{\mathbf{A}}_i + \bar{\mathbf{K}}_i)^n \bar{\mathbf{K}}_i)\bar{\mathbf{K}}_i\| \leq C$. Hence, if we start at a state $\mathbf{x}_1$ with $\|\mathbf{x}_1\| \leq 1/C$, we have that for either choosing of i, the linear dynamics $\tilde{\mathbf{u}}_h = \bar{\mathbf{K}}_i \tilde{\mathbf{x}}_h$, $\tilde{\mathbf{x}}_{h+1} = \bar{\mathbf{A}}_i \tilde{\mathbf{x}}_{h+1} + \tilde{\mathbf{u}}_h$, $\tilde{\mathbf{x}}_1 = \mathbf{x}_1$ satisfy $\sup_{h \geq 1} \max\{\|\tilde{\mathbf{x}}_h\|, \|\tilde{\mathbf{u}}_h\|\} \leq C\|\mathbf{x}_1\| \leq \Delta$. By construction, $\mathbf{x}_{h+1} = \bar{\mathbf{A}}_i \mathbf{x}_h + \mathbf{u}_h$ and $\mathbf{u}_h = \mathbf{K}_i \mathbf{x}_h$ obeys these same linear dynamics under the expert trajectory when $\max\{\|\mathbf{x}_h\|, \|\mathbf{u}_h\| \leq 1\}$. In particular, when Δ is chosen to be less than $1/C$, we ensure these linear dynamics hold starting from $\mathbf{x}_1 \mid Z = 1$. Note that such $\mathbf{x}_1$ is also supported coordinates 2-through-d, one can check inductively that $\mathbf{x}_h, h > 1$ are also supported on these same coordinates, and that $(\bar{\mathbf{A}}_i + \bar{\mathbf{K}}_i)\mathbf{x}_h$ and $\bar{\mathbf{K}}_i \mathbf{x}_h$ does not depend on i. ■

L.4. The compounding error argument.

The goal of this section is to establish the following proposition. It establishes that, up to a threshold over $1/\mathrm{poly}(L, M, d)$, the probability of experiences exponential in H compounding error is at least a constant times the probability that, under $D_{\{Z=0\}}$, the learner makes a large mistake in the $\mathbf{e}_1$ direction. This nonlinear formalizes the heuristic argument given in Section 4.

Proposition L.1 *Fix an ϵ_0 and simply-stochastic, L-Lipschitz, M-smooth policy $\hat{\pi}$ (with $L, M \geq 1$). Suppose $\tau = o_\star(1), \Delta = \Theta_\star\left(\frac{1}{ML\sqrt{d}}\right)$. Fix an $\epsilon > 0$ and $g \in \mathcal{G}$. In terms of these, define*

$$\epsilon_\star = \epsilon_\star(\epsilon_0) := \min\left\{o_\star\left(\frac{1}{L^2 M d}\right), \left(\frac{17}{16}\right)^{H-2} 2\tau\epsilon_0\right\}$$
$$p_\star = p_\star(\epsilon_0, g) := \mathbb{P}_{\mathbf{x}_1 \sim D_{\{Z=0\}}, \mathbf{u} \sim \hat{\pi}}\left[\left|\pi^\star_{\xi_0, g}(\mathbf{x}_1) - \langle \mathbf{e}_1, \mathbf{u}\rangle\right| \geq \epsilon_0\right],$$

where above we note that $p_\star$ does not depend on ξ_0 because $\pi^\star_{\xi_0,g}(\mathbf{x}_1)$ does not depend on ξ_0 when $\mathbf{x}_1$ lies in the support of $D_{\{Z=0\}}$. Then,

$$\mathbb{E}_{(i,\omega)\sim P}\, \mathbb{P}_{\hat{\pi}, f_{g,(i,\omega)}, \mathcal{D}}\left[\overline{\mathrm{cost}}_{\mathrm{hard}}(\mathbf{x}_{1:H}, \mathbf{u}_{1:H}) \geq \epsilon_\star\right] \geq p_\star / C \tag{L.10}$$

for some appropriate constant C.

We prove Proposition L.1 via a Lemma L.9 below. The remainder of the subsection will be dedicated to the proof of that lemma.

In what follows, we define two objects, parameterizerized in terms of the initial state $\mathbf{x}_1$, and deviations $(\boldsymbol{\zeta}_t)$ from the mean.

Definition L.2 *The trajectory induced by $\hat{\pi}$ conditioned on the random terms $\boldsymbol{\zeta}_{1:H}$:*

$$\mathsf{traj}_{g,(i,\omega)}(\boldsymbol{\zeta}_{1:H}, \mathbf{x}_1) = (\mathbf{x}_{1:H}, \mathbf{u}_{1:H}), \quad \mathbf{u}_h = \bar{\pi}(\mathbf{x}_h) + \boldsymbol{\zeta}_h, \quad \mathbf{x}_{h+1} = f_{g,(i,\omega)}(\mathbf{x}_h, \mathbf{u}_h) \tag{L.11}$$

Definition L.3 (First Stage Error) *We define*

$$\mathsf{err}(\mathbf{x}_1, g, \boldsymbol{\zeta}_1) = \left|\tau \cdot \mathcal{T}[g](\mathbf{x}_1) - \langle \mathbf{e}_1, \bar{\pi}(\mathbf{x}_1) + \boldsymbol{\zeta}_1\rangle\right|.$$

Lemma L.9 *Let $\mathbf{x}_1 \in \mathrm{support}(D_{\{Z=0\}})$. Suppose $\tau = o_\star(1), \Delta = o_\star\left(\frac{1}{ML\sqrt{d}}\right)$, and*

$$\max_{i,\omega} \mathbb{P}_{\hat{\pi}, f_{g,(i,\omega)}}[\overline{\mathrm{cost}}_{\mathrm{hard}}(\mathbf{x}_{1:H}, \mathbf{u}_{1:H}) \geq M\Delta^2/2] \leq o_\star(1), \tag{L.12}$$

Then, for any choice of g and any sequence $\boldsymbol{\xi}$, we have

$$\max_{i,\omega} \overline{\mathrm{cost}}_{\mathrm{hard}}(\mathsf{traj}_{g,(i,\omega)}(\boldsymbol{\zeta}_{1:H}, \mathbf{x}_1)) \geq \min\left\{o_\star(\min\{\tau, \sqrt{d}\Delta\}), \left(\frac{17}{16}\right)^{H-2} 2\tau\epsilon(\mathbf{x}_1, g, \boldsymbol{\zeta}_1)\right\} \tag{L.13}$$

Proof [Proof of Proposition L.1 assuming Lemma L.9] Observe that under our parameter choices and $L, M \geq 1$, we can ensure

$$\epsilon_\star = \epsilon_\star(\epsilon_0, g) = \min\left\{\frac{M\Delta^2}{2}, o_\star\left(\min\left\{\tau, \sqrt{d}\Delta\right\}\right), \left(1 + \frac{\gamma}{2}\right)^{H-2} 2\tau\epsilon_0\right\}.$$

We have two cases. First, let $c_0 = o_\star(1)$ be the constant implicit on the right hand side of Eq. (L.12).

Case 1: $\mathbb{E}_{(i,\omega)\sim P}\,\mathbb{P}_{\hat{\pi},f_{g,(i,\omega)}}[\overline{\mathrm{cost}}_{\mathrm{hard}}(\mathbf{x}_{1:H},\mathbf{u}_{1:H}) \ge M\Delta^2/2] \ge c_0/4$. Then,

$$\mathbb{E}_{(i,\omega)\sim P}\,\mathbb{P}_{\hat{\pi},f_{g,(i,\omega)},\mathcal{D}}\left[\overline{\mathrm{cost}}_{\mathrm{hard}}(\mathbf{x}_{1:H},\mathbf{u}_{1:H}) \ge \epsilon_\star\right] \ge c_0/4 \ge p_\star c_0/4 \ge p_\star/C$$

for $C = 4/c_0$.

Case 2: In the second case, we can assume that $\mathbb{E}_{(i,\omega)\sim P}\,\mathbb{P}_{\hat{\pi},f_{g,(i,\omega)}}[\overline{\mathrm{cost}}_{\mathrm{hard}}(\mathbf{x}_{1:H},\mathbf{u}_{1:H}) \ge M\Delta^2/2] \le c_0/4$, so that $\max_{(i,\omega)}\mathbb{P}_{\hat{\pi},f_{g,(i,\omega)}}[\overline{\mathrm{cost}}_{\mathrm{hard}}(\mathbf{x}_{1:H},\mathbf{u}_{1:H}) \ge M\Delta^2/2] \le c_0$. Let us adopt the shorthand

$$\phi(\epsilon_0) = \min\left\{o_\star(\min\{\tau,\sqrt{d}\Delta\}),\left(1+\frac{\gamma}{2}\right)^{H-2}2\tau\epsilon_0\right\}$$

It then follows from Lemma L.9 that, for any fixed $\mathbf{x}_1$ in the support of $D_{\{Z=0\}}$ and noise sequence $\boldsymbol{\zeta}_{1:H}$ we have

$$\mathbb{P}_{(i,\omega)\sim P}\left[\overline{\mathrm{cost}}_{\mathrm{hard}}(\mathbf{traj}_{g,(i,\omega)}(\boldsymbol{\zeta}_{1:H},\mathbf{x}_1)) \ge \phi(\epsilon_0)\right] \ge \frac{1}{4}\text{whenever } \mathbf{err}(\mathbf{x}_1,g,\boldsymbol{\zeta}_1) \ge \epsilon_0.. \tag{L.14}$$

Thus,

$$\begin{aligned}
&\mathbb{E}_{(i,\omega)\sim P}\,\mathbb{P}_{\hat{\pi},f_{g,(i,\omega)},\mathcal{D}}\left[\overline{\mathrm{cost}}_{\mathrm{hard}}(\mathbf{x}_{1:H},\mathbf{u}_{1:H}) \ge \phi(\epsilon_0)\right]\\
&\ge \frac{1}{2}\mathbb{E}_{(i,\omega)\sim P}\,\mathbb{P}_{\mathbf{x}_1\sim D_{\{Z=0\}},\zeta_{1:H}}\left[\overline{\mathrm{cost}}_{\mathrm{hard}}(\mathbf{traj}_{g,(i,\omega)}(\boldsymbol{\zeta}_{1:H},\mathbf{x}_1)) \ge \phi(\epsilon_0)\right]\\
&\ge \frac{1}{8}\,\mathbb{P}_{\mathbf{x}_1\sim D_{\{Z=0\}},\zeta_1}\left[\mathbf{err}(\mathbf{x}_1,g,\boldsymbol{\zeta}_1) \ge \epsilon_0\right] && \text{(by Eq. (L.14) and Bayes' rule)}\\
&:= \frac{1}{8}\,\mathbb{P}_{\mathbf{x}_1\sim D_{\{Z=0\}},\mathbf{u}\sim\hat{\pi}}\left[|\tau\cdot\mathcal{T}[g](\mathbf{x}_1) - \langle\mathbf{e}_1,\mathbf{u}\rangle| \ge \epsilon_0\right] && \text{(definition of } \mathbf{err}(\mathbf{x}_1,g,\boldsymbol{\zeta}_1))\\
&= \frac{p_\star(\epsilon_0,g)}{8} && \text{(Definition of } p_\star)
\end{aligned}$$

Hence, the result holds for $C = 1/8$. ■

L.4.1. PROOF OF LEMMA L.9

Our proof strategy is to adopt a quantitative variant of the stable manifold theorem, adapting an argument to due to Jin et al. (2017), and extending it to handle dynamical maps with non-symmetric gradients[3] Informally, the stable manifold theorem considers a smooth dynamics map $F : \mathbb{X} \to \mathbb{X}$, whose gradient ∇F exhibits an eigenvalue strictly greater than one at the origin. The smoothness of the dynamics F allow the approximation $F(\mathbf{x}) \approx F(\mathbf{0}) + \nabla F(\mathbf{0})\mathbf{x}$ near the origin, which entails that the k-fold compositions $F^k(\mathbf{x})$ scale with $(\nabla F(\mathbf{0}))^k$ which, due to the unstable value, causes the state to grow exponentially.

For simply stochastic policies, we take (up to the additive noise $\boldsymbol{\zeta}_t$)

$$F_i(\mathbf{x}) := \bar{\mathbf{A}}_i\mathbf{x} + \bar{\pi}(\mathbf{x}), \tag{L.15}$$

3. In their paper Jin et al. (2017), the dynamical map in question arises from the gradient-descent update of a scalar-valued function say $h(\mathbf{x})$, so the gradient of the induced dynamical map is proportional the Hessian of $h(\mathbf{x})$, which is symmetric.

which is equal to $f_{\xi,g}(\mathbf{x}, \bar{\pi}(\mathbf{x}))$ is within the unit ball around $\mathbf{x} = 0$. As $\mathbf{A} := \nabla F_i(\mathbf{x})$ is a non-symmetric in general, its eigenvectors may be poorly conditioned. Thus, we argue that $\mathbf{A}$ is approximately lower triangular, with small top-right block, not-too-large bottom-left block, stable bottom-right block, and finally, an unstable (magnitude > 1 entry) in the $(1, 1)$-position. We define this structure as follows:

Definition H.1 *Given parameters $\gamma > 1, \mu \in (0, 1), L \geq 1, r > 0$, we say $\mathbf{A}$ is a (γ, μ, L, r)-* matrix *if $\mathbf{A}$ admits the following block decomposition, where $\mathbf{Y}_1$ and $\mathbf{Y}_2$ are square matrices:*

$$\mathbf{A} = \begin{bmatrix} \mathbf{Y}_1 & \mathbf{W}^\top \\ \tilde{\mathbf{W}} & \mathbf{Y}_2 \end{bmatrix},$$

where, for parameters (γ, μ, L, ν), $\|\mathbf{Y}_2\|_{\mathrm{op}} \leq 1 - \mu < 0$, $\|\tilde{\mathbf{W}}\|_{\mathrm{op}} \leq L$, and $\sigma_{\min}(\mathbf{Y}_1) \geq 1 + \gamma > 1$, and $\|\mathbf{W}\|_{\mathrm{op}} \leq r$.

There is at least one choice of $i \in \{1, 2\}$ for which $\nabla F_i(\mathbf{x})$ is of this form.

Claim L.10 *Let $\hat{\mathbf{K}} := \nabla\bar{\pi}(\mathbf{x})\big|_{\mathbf{x}=\mathbf{0}}$. Suppose that $\|(\hat{\mathbf{K}} - \bar{\mathbf{K}}_i)\mathrm{Proj}_{\geq 2}\|_{\mathrm{F}} \leq 6M\Delta\sqrt{d}$, which by Lemma L.7 holds under the condition Eq. (L.12). Then, for $\Delta = o_\star(1/LM\sqrt{d})$, there exists at least one of $i \in \{1, 2\}$ for which $\nabla F_i(\mathbf{x})\big|_{\mathbf{x}=\mathbf{0}}$, is a $(1/8, 1/2, 2L, r)$-matrix in Definition H.1, where $r = o_\star(1/L)$.*

Proof [Proof of Claim L.10] The argument generalizes the matrix argument given in Proposition 4.1. Formally, set $\alpha = \mathbf{e}_1^\top (\nabla\bar{\pi}(\mathbf{x}))\, \mathbf{e}_1$ and set $\boldsymbol{\beta} = \mathbf{Q}_{\geq 2}\nabla\bar{\pi}(\mathbf{x})\mathbf{e}_1$, where $\mathbf{Q}_{\geq 2}$ denote the projection from $\mathbb{R}^d \to \mathbb{R}^{d-1}$ which zeros out the first coordinate. Using the block structure of $\bar{\mathbf{A}}_i, \bar{\mathbf{K}}_i$ and the computation in Proposition 4.1, we have

$$\begin{aligned}
\nabla F_i(\mathbf{x}) &= \bar{\mathbf{A}}_i + \bar{\mathbf{K}}_i\mathrm{Proj}_{\leq 2} + \alpha\mathbf{e}_1\mathbf{e}_1^\top + \begin{bmatrix} 0 \\ \boldsymbol{\beta} \end{bmatrix} + \left(\nabla\bar{\pi}(\mathbf{x}) - \bar{\mathbf{K}}_i\right)\mathrm{Proj}_{\geq 2} \\
&= \begin{bmatrix} (\mathbf{A}_i + \mathbf{K}_i)_{11} + \alpha & \boldsymbol{\Delta}_{21} \\ \begin{bmatrix} (\mathbf{A}_i + \mathbf{K}_i)_{21} \\ \mathbf{0} \end{bmatrix} + \boldsymbol{\beta} & \begin{bmatrix} 1 - 2\mu & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{bmatrix} + \boldsymbol{\Delta}_{22} \end{bmatrix},
\end{aligned}$$

where

$$\begin{aligned}
&\max\{\|\boldsymbol{\Delta}_{21}\|, \|\boldsymbol{\Delta}_{22}\|\} \leq \|\left(\nabla\bar{\pi}(\mathbf{x}) - \bar{\mathbf{K}}_i\right)\mathrm{Proj}_{\geq 2}\| \leq 6M\Delta\sqrt{d}, \\
&\|\boldsymbol{\beta}\| \leq \|\nabla\bar{\pi}(\mathbf{x})\|_{\mathrm{op}} \leq L.
\end{aligned}$$

Arguing as in Proposition 4.1, we have that $(\mathbf{A}_i + \mathbf{K}_i)_{11} \in \{1 + \mu, -(1 - \frac{1}{4}\mu)\}$, so that there exists one of $i \in \{1, 2\}$ for which $|(\bar{\mathbf{A}}_i + \bar{\mathbf{K}}_i)_{11} + \alpha| \geq 1 + \frac{\mu}{4}$. Moreover, $|(\mathbf{A}_i + \mathbf{K}_i)_{11}| \in \{|-c_\mu|, |0|\} \leq |c_\mu| = \frac{3}{2}\mu \leq 2\mu$, and $\|\boldsymbol{\beta}\| \leq L$ when $\bar{\pi}(\mathbf{x})$ is L-Lipschitz. Finally, as $\mu \leq 1$, assuming $6M\Delta\sqrt{d} \leq \mu$ and $L \geq 2\mu$, we conclude that$\nabla F_t(\mathbf{x})$ (which again, does not depend on t) admits the following block decomposition:

$$\nabla F_i(\mathbf{x}) := \begin{bmatrix} y_1 & \mathbf{W}^\top \\ \tilde{\mathbf{W}} & \mathbf{Y}_{[2]} \end{bmatrix},$$

where $|y_1| \ge 1 + \frac{\mu}{4}$, $\|\tilde{\mathbf{W}}\|_{\mathrm{op}} \le 2L$, $\|\mathbf{Y}_{[2]}\|_{\mathrm{op}} \le 1 - \mu$ and $\|\mathbf{W}\|_{\mathrm{op}} \le 6M\Delta\sqrt{d}$. We conclude by setting $\gamma \leftarrow \mu/4$, $\mu \leftarrow \mu$, and $L \leftarrow 2L$ in the definition of a (γ, μ, L)-matrix. To ensure the bound on r, we use Lemma L.7, which ensures that $6M\Delta\sqrt{d} \le o_\star(1/L)$. ∎

Next, we recall our major technical tool, which is a quantitative statement of an unstable-manifold theorem for dynamics whose Jacobian is a (μ, γ, L, Δ)-matrix.

Proposition H.1 (Exponential Compounding for (μ, γ, L)-matrices) *Let $r > 0$, and let $F(\mathbf{x}, t)$ be a time-varying, M-smooth dynamical map such that each*

$$\mathbf{A}_t := \nabla_{\mathbf{x}} F(\mathbf{x}, t)\big|_{\mathbf{x}=0}$$

is a (γ, μ, L, r)-matrix with $\gamma \le 1$, with the same block structure across t, and where $r = o_\star(L/\gamma\mu)$. Then, for any $\mathbf{x}_1 \in \mathbb{R}^d$, then

$$\mathbf{x}_{t+1} = F(\mathbf{x}_t, t), \quad \tilde{\mathbf{x}}_{t+1} = F(\tilde{\mathbf{x}}_t, t), \quad \tilde{\mathbf{x}}_1 = \mathbf{x}_1 \pm \epsilon \mathbf{e}_1$$

then either

$$\max_{1 \le t \le H} |\mathbf{e}_1^\top (\mathbf{x}_t - \tilde{\mathbf{x}}_t)| \ge \left(1 + \frac{\gamma}{2}\right)^{H-1} \epsilon \tag{H.1}$$

or

$$\max_{1 \le t \le H} \max\{\|\mathbf{x}_t\|, \|\mathbf{x}'_t\|\} \ge o_\star\left(\frac{1}{\mu\gamma \cdot LM}\right) \tag{H.2}$$

We may now conclude the proof of Lemma L.9.

Proof [Proof of Lemma L.9] Fix $\boldsymbol{\zeta}_{1:H}$. We may assume that $\max_{i,\omega} \overline{\mathrm{cost}}_{\mathrm{hard}}(\mathbf{traj}_{g,(i,\omega)}(\boldsymbol{\zeta}_{1:H}, \mathbf{x}_1)) \le \min\{\tau, \Delta\} C_{\mathrm{cost}}$, otherwise the result is immediate. Let $(\mathbf{x}_{i,\omega;t})$ denote the sequence of iterates given by the dynamics in Eq. (L.11), namely by $\mathbf{u}_h = \bar{\pi}(\mathbf{x}_h) + \boldsymbol{\zeta}_h, \mathbf{x}_{h+1} = f_{g,(i,\omega)}(\mathbf{x}_h, \mathbf{u}_h)$.

By Lemma L.6 and $\tau = o_\star(1), \Delta = o_\star(1/ML\sqrt{d}) = o_\star(1)$, $\max_{i,\omega} \overline{\mathrm{cost}}_{\mathrm{hard}}(\mathbf{traj}_{g,(i,\omega)}(\boldsymbol{\zeta}_{1:H}, \mathbf{x}_1)) \le \min\{o_\star(\mathbf{traj}), o_\star(\sqrt{d}\Delta)\}$ implies that

$$\max_{2 \le t \le H} \|\mathbf{x}_{i,\omega;t}\| \le o_\star(\sqrt{d}\Delta). \tag{L.16}$$

By Claim L.10, we may choose an $i \in \{1, 2\}$ for which $\mathbf{A}_i + \nabla\bar{\pi}(\mathbf{x})\big|_{\mathbf{x}=\mathbf{0}}$ is a $(\gamma, \mu, L, r) = (1/8, 1/2, L, o_\star(1/L))$-matrix. Then, using the linearity of dynamics ensured by Lemma L.6,

$$\begin{aligned} \mathbf{x}_{i,\omega;t+1} &= \bar{\mathbf{A}}_i + \bar{\pi}(\mathbf{x}_{i,\omega;t}), \quad \omega \in \{-1, 1\} \\ \mathbf{x}_{i,1;2} - \mathbf{x}_{i,-1;2} &= \pm 2\tau\epsilon(\mathbf{x}_1, g, \boldsymbol{\zeta}_1), \end{aligned}$$

where $\pm$ denotes an arbitrary choice of sign. Proposition H.1 implies and the fact that $\mathbf{A}_i + \nabla\bar{\pi}(\mathbf{x})\big|_{\mathbf{x}=\mathbf{0}}$ is a $(1/8, 1/2, L, o_\star(1/L))$-matrix implies $2 \le t \le H$ for which either $\max_{\omega \in \{-1,1\}} |\mathbf{e}_1^\top \mathbf{x}_{i,\omega;t+1}| \ge (\frac{17}{16})^{H-2} 2\tau\epsilon(\mathbf{x}_1, g, \boldsymbol{\zeta}_1)$, or $\max_{\omega \in \{-1,1\}} \max_{2 \le t \le H} \|\mathbf{x}_{i,\omega;t+1}\| \ge \Omega(\frac{1}{LM})$. By making $\Delta = o_\star(1/LM\sqrt{d})$, the second case cannot occur without contradicting Eq. (L.16). ∎

L.5. Proof of Minimax Risk Bounds in Theorem 1.A

In what follows, and in keeping with Construction L.1, we let again $\xi = \{i, \omega\}$ denote the hidden parameter, so policies and dynamics are of the form $f_{g,\xi}, \pi_{g,\xi}$. As established in Lemmas L.5 and L.8, the distribution over samples does not depend on the instance label (g, ξ) for $Z = 1$, and one can verify that the properties, Properties G.1 to G.3 all hold. We use these properties in what follows. Lastly, we remark that our arguments extends to the class of proper algorithms by noticing that $\mathbb{A} = \mathbb{A}_{\mathrm{smooth}}(L, M)$ contains all proper estimators, provided $L \geq L_0$ and $M \geq M_0$; this is a consequence of the fact our construction uses deterministics policies and dynamics, and the smoothness/Lipschitzness computations of Appendix L.6.

Lemma L.11 *Let $\mathcal{P}, D$ be as in Construction L.1, and recall that $D_{\{Z=0\}}$ denotes the conditional of $D \mid \{Z = 0\}$. Then,*

- *(a) The BC problem class $(\mathcal{P}, D_{\{Z=0\}})$ satisfies the general reduction conditions of all part of Proposition G.1: namely Properties G.1 to G.3, with parameter τ, the class $\mathcal{G}$ is convex (by assumption), and $\mathbb{A}$ contains all estimation algorithms.*
- *(b) Let $(\pi^\star, f) \in \mathcal{P}$. Given a sample $\mathrm{S}_{n,H}^{(Z=0)}$ of n, length H trajectories from $\mathbb{P}_{\pi^\star,f,D_{\{Z=0\}}}$, one can simulate a sample $\mathrm{S}_{n,H}$ of n, length H trajectories from $\mathbb{P}_{\pi^\star,f,D}$.*

Proof Part (a) can be easily checked from Lemma L.5 and going through the various conditions. The properness of $\mathbb{A}$ follows from Lemma L.12. Part (b) follows from the fact that, from initial states in the support of $D_{\{Z=1\}}$, the distribution of the trajectories is identical for all $(\pi^\star, f) \in \mathcal{P}$ (Lemma L.8). ∎

L.5.1. LOWER BOUND ON THE TRAINING RISKS.

Let $\hat{n} \sim \mathrm{Binomial}(\frac{1}{2}, n)$. Because we samples collect from $\{Z = 1\}$-trajectories can be simulated without knowledge of the ground truth instance (Lemma L.11(b)), we can decompose

$$
\begin{aligned}
\mathsf{M}_{\mathrm{expert},L_2}(n; \mathcal{P}, D, \mathcal{H}) &= \frac{1}{2}\mathbb{E}_{\hat{n}}[\mathsf{M}_{\mathrm{expert},L_2}(\hat{n}; \mathcal{P}, D_{\{Z=0\}}, \mathcal{H})] \\
\mathsf{M}^{\mathbb{A}}_{\mathrm{expert},L_2}(n; \mathcal{P}, D, \mathcal{H}) &= \frac{1}{2}\mathbb{E}_{\hat{n}}[\mathsf{M}_{\mathrm{expert},L_2}(\hat{n}; \mathcal{P}, D_{\{Z=0\}}, \mathcal{H})], \\
\mathsf{M}^{\mathbb{A}}_{\mathrm{eval},h,B}(n; \mathcal{P}, D, \mathcal{H}) &= \frac{1}{2}\mathbb{E}_{\hat{n}}[\mathsf{M}^{\mathbb{A}}_{\mathrm{eval},h,B}(n; \mathcal{P}, D, \mathcal{H})],
\end{aligned}
$$

which follows by conditioning on the number sampled trajectories for which $Z = 0$, which is distributed as $\hat{n}$. From Lemma L.11(a), the conclusion of Proposition G.1 holds with parameter τ. Invoking that proposition,

$$
\forall \hat{n} \in \mathbb{N}, \quad \mathsf{M}_{\mathrm{expert},L_2}(\hat{n}; \mathcal{P}, D_{\{Z=0\}}, \mathcal{H}) = \mathsf{M}^{\mathbb{A}}_{\mathrm{expert},L_2}(\hat{n}; \mathcal{P}, D_{\{Z=0\}}, \mathcal{H}) = \tau \mathsf{M}_{\mathrm{reg},L_2}(\hat{n}; \mathcal{G}, D_{\mathrm{reg}}).
$$

Taking an expectation over $\hat{n}$ and using the previous display proves the equality of $\mathsf{M}_{\mathrm{expert},L_2}(n; \mathcal{P}, D, \mathcal{H})$ and $\mathsf{M}^{\mathbb{A}}_{\mathrm{eval},h,B}(n; \mathcal{P}, D, \mathcal{H})$. The same also holds any $\mathbb{A}' \supseteq \mathbb{A}_{\mathrm{proper}}(\mathcal{P})$, and in particular, for $\mathbb{A}_{\mathrm{proper}}(\mathcal{P})$.

To upper bound the relevant terms, a chernoff bound on $\hat{n} \sim \mathrm{Binomial}(1/2, n)$ implies that that with probability $1 - e^{-c'n}$ for some $c' > 0$, we have $\hat{n} \ge n/3$. And, when $\hat{n} \ge n/3$, then because an estimator with fewer samples can always be simulated via an estimator with more samples, we have $\mathsf{M}^{\mathbb{A}}_{\mathrm{eval},h,B}(\hat{n}; \mathcal{P}, D_{\{Z=0\}}, \mathcal{H}) \le \mathsf{M}^{\mathbb{A}}_{\mathrm{eval},h,B}(n/3; \mathcal{P}, D_{\{Z=0\}}, \mathcal{H})$ when $\hat{n} \ge n/3$. On the other hand, when $\hat{n} < n/3$, because all terms in Construction L.1 remain uniformly bounded, there exists an estimator which makes at most constant error, say $C' > 0$. Hence,

$$\begin{aligned}
\mathbb{E}_{\hat{n}}[\mathsf{M}^{\mathbb{A}}_{\mathrm{expert},L_2}(\hat{n}; \mathcal{P}, D, \mathcal{H})] &\le \mathbb{E}_{\hat{n}}[\mathsf{M}^{\mathbb{A}}_{\mathrm{expert},L_2}(\hat{n}; \mathcal{P}, D_{\{Z=0\}}, \mathcal{H}]) \\
&\le \mathbb{P}_{\hat{n}}[\hat{n} \ge n/3] \mathsf{M}^{\mathbb{A}}_{\mathrm{expert},L_2}(n/3; \mathcal{P}, D_{\{Z=0\}}, \mathcal{H}) + C' \, \mathbb{P}[\hat{n} < n/3] \\
&\le \mathsf{M}^{\mathbb{A}}_{\mathrm{expert},L_2}(n/3; \mathcal{P}, D_{\{Z=0\}}, \mathcal{H}) + C' e^{-c'n} \\
&= \tau \mathsf{M}_{\mathrm{reg},L_2}(n/3; \mathcal{G}, D_{\mathrm{reg}}) + C' e^{-c'n}.
\end{aligned}$$

For a lower bound on the training risk, we see that

$$\begin{aligned}
\mathsf{M}^{\mathbb{A}}_{\mathrm{expert},L_2}(n; \mathcal{P}, D, \mathcal{H}) &= \frac{1}{2} \mathbb{E}_{\hat{n}}[\mathsf{M}^{\mathbb{A}}_{\mathrm{expert},L_2}(\hat{n}; \mathcal{P}, D_{\{Z=0\}}, \mathcal{H}]) \\
&\ge \frac{1}{2} \mathsf{M}^{\mathbb{A}}_{\mathrm{expert},L_2}(n; \mathcal{P}, D_{\{Z=0\}}, \mathcal{H}) = \frac{\tau}{2} \mathsf{M}_{\mathrm{reg},L_2}(n; \mathcal{G}, D_{\mathrm{reg}}), \qquad \text{(L.17)}
\end{aligned}$$

where we use the fact that $\hat{n} \le n$ and that (because one can always neglect to use some samples) minimax risks are nonincreasing in n. We conclude this section by noting that $\tau = \Theta_\star(1)$ in Construction L.1.

L.5.2. LOWER BOUND ON THE EVALUATION RISK.

We apply Proposition L.1 with $\epsilon_0 \leftarrow \tau \kappa \boldsymbol{\epsilon}_n$, where τ is as in the construction, and, κ and $\boldsymbol{\epsilon}_n$ are as in Condition E.1 on the problem class $(\mathcal{G}, D_{\mathrm{reg}})$. Let P denote the uniform prior on $(i, \omega) \in \{1, 2\} \times \{-1, 1\}$Assume $\tau = o_\star(1)$ and $\Delta = o_\star\left(\frac{1}{ML\sqrt{d}}\right)$. and for the terms

$$\epsilon_{\mathrm{compound}} := \min\left\{ o_\star\left(\frac{1}{L^2 M d}\right), \left(1 + \frac{\gamma}{2}\right)^{H-2} 2\tau^2 \boldsymbol{\epsilon}_n \kappa \right\}$$

Finally, introduce the risk $\mathsf{R}_\star(\hat{\pi}, g, \xi) := \mathbb{P}_{\hat{\pi}, f_{g,\xi}, \mathcal{D}}\left[\overline{\mathrm{cost}}_{\mathrm{hard}}(\mathbf{x}_t, \mathbf{u}_t) \ge \epsilon_{\mathrm{compound}}\right]$. Then, Proposition L.1 implies

$$\mathbb{E}_{\xi \sim P} \mathsf{R}_\star(\hat{\pi}, g, \xi) \ge \frac{1}{C} \mathbb{P}_{\mathbf{x}_1 \sim D_{\{Z=0\}}, \mathbf{u} \sim \hat{\pi}}\left[|\hat{\pi}_{g,\xi_0}(\mathbf{x}_1) - \langle \mathbf{e}_1, \mathbf{u} \rangle| \ge \tau \kappa \boldsymbol{\epsilon}_n\right].$$

for some appropriate constant C. Above, we note $\hat{\pi}_{g,\xi_0}(\mathbf{x}_1) = \tau \cdot \mathcal{T}[g](\mathbf{x}_1)$ for any choice of ξ_0 when $\mathbf{x}_1 \sim D_{\{Z=0\}}$. In light of Lemma L.11(a), we may apply Proposition G.1 to the problem instance $(\mathcal{P}, D_{\{Z=0\}})$. This implies that

$$\inf_{\mathrm{alg}} \sup_{g,\xi} \mathbb{E}_{\mathrm{S}_{n,H} \sim (\pi_{g,\xi}, f_{g,\xi}, D_{\{Z=0\}})} \mathbb{E}_{\hat{\pi} \sim \mathrm{alg}(\mathrm{S}_{n,H})} \mathsf{R}_\star(\pi; g, \xi) \ge \delta / C. \qquad \text{(L.18)}$$

Finally, from Lemma L.11(b), the n samples from $D_{\{Z=0\}}$ can simulate n samples from the unconditioned distribution, D. Thus, any estimator can do no better taking samples from D:

$$\inf_{\mathrm{alg}} \sup_{g,\xi} \mathbb{E}_{\mathrm{S}_{n,H} \sim (\pi_{g,\xi}, f_{g,\xi}, D_{\{Z=0\}})} \mathbb{E}_{\hat{\pi} \sim \mathrm{alg}(\mathrm{S}_{n,H})} \mathsf{R}_\star(\pi; g, \xi) \ge \delta / C.$$

Substituting in our definition of $\mathbf{R}_\star(\hat{\pi}, g, \xi) := \mathbb{P}_{\hat{\pi}, f_{g,\xi}, \mathcal{D}}\left[\overline{\mathrm{cost}}_{\mathrm{hard}}(\mathbf{x}_t, \mathbf{u}_t) \ge \epsilon_{\mathrm{compound}}\right]$ and the definition of $\epsilon_{\mathrm{compound}}$ implies that

$$\mathbf{M}_{\mathrm{eval,prob}}\left(n, \frac{\delta}{C}; \mathcal{P}, D, H\right) \ge \min\left\{o_\star\left(\frac{1}{L^2 M d}\right), \left(\frac{17}{16}\right)^{H-2} 2\tau^2 \boldsymbol{\epsilon}_n \kappa\right\},$$

Tuning $\tau = \Theta_\star(1)$ and absorbing constants concludes the demonstration.

L.6. Regularity Conditions

The goal of this section is to establish the following:

Lemma L.12 *Suppose $\mathcal{G}$ satisfies Assumption F.1. Then, provided τ is smaller than a universal constant, we have that $(\mathcal{P}, D)$ is $(O(1), O(1), O(1))$-regular, and for all $(\pi, f) \in \mathcal{P}$, f and (π, f) are (C', ρ')-E-IISS for some $C' \ge 1, \rho' \in (0,1)$. In particular, for $\mathbb{A} = \mathbb{A}_{\mathrm{simple}}(L, M)$ for some sufficiently large constants $L, M \ge 1$, $\mathbb{A}$ is proper. Lastly, all f as in Construction L.1 are $O(1)$-one-step-controllable.*

This result is a consequence of the arguments that follow, with controllability deferred to the end of this section. Recall from Construction L.1 the functions $[\mathcal{T}(g)](\mathbf{x}) := g(\mathrm{Proj}_{\ge 3}\mathbf{x})$ and $\mathrm{restrict}(\mathbf{x}) := \mathrm{bump}_d(\mathbf{x} - 3\mathbf{e}_1)$ Lets introduce the shorthand

$$\psi_g(\mathbf{x}) := \mathrm{restrict}(\mathbf{x}) \cdot \mathcal{T}[g](\mathbf{x}), \quad \psi_u(\mathbf{u}, \mathbf{x}) := \langle \mathbf{e}_1, \mathbf{u}\rangle \mathrm{bump}_d(\mathbf{u}/4)\mathrm{restrict}(\mathbf{x}).$$

Then, we can write

$$\begin{aligned} \pi_{g,\xi}(\mathbf{x}) &= \bar{\mathbf{K}}_i \mathbf{x} + \tau\psi_g(\mathbf{x})\mathbf{e}_1 \\ f_{g,\xi}(\mathbf{x}, \mathbf{u}) &= \bar{\mathbf{A}}_i \mathbf{x} + \mathbf{u} - \tau\psi_g(\mathbf{x})\mathbf{e}_1 + \omega \cdot \mathbf{e}_1(\tau^2\psi_g(\mathbf{x}) - \tau\psi_u(\mathbf{u}, \mathbf{x})). \end{aligned} \tag{L.19}$$

Claim L.13 *Suppose that each $g \in \mathcal{G}$ is L_0-Lipschitz, M_0-smooth, and 1-bounded on the ball of radius 2 on $\mathbb{R}^{d-2}$. Then, letting $O(\cdot)$ hide universal constants,*

- *$\psi_g(\mathbf{x})$ is $O(L_0 + 1)$-Lipschitz and $O(1 + L_0 + M_0)$-smooth.*
- *$\psi_u(\mathbf{x}, \mathbf{u})$ is $O(1)$-Lipschitz and $O(1)$-smooth.*

Proof Recall that the bump-functions have derivatives bounded by universal constants (Lemma H.14). Hence, the desired bounds follow from the product rule, and the fact that $\mathcal{T}[g]$ inherits the smoothness/Lipschitzness of $\mathcal{G}$, and the fact that $\mathrm{restrict}(\mathbf{x})$ constrains to a ball of radius 2. ■

The following lemma gives us the desired regularity guarantee.

Lemma L.14 (Regularity) *Let $\tau \le 1$. Suppose that each $g \in \mathcal{G}$ is L_0-Lipschitz, M_0-smooth, and 1-bounded on the ball of radius 2 on $\mathbb{R}^{d-2}$. Then, every $(\pi, f) \in \mathcal{P}$ are are $O(L_0 + 1)$-Lipschitz and $\tau \cdot O(1 + L_0 + M_0)$-smooth. Hence, $(\mathcal{P}, D)$ is $(O(L_0 + 1), O(1 + L_0 + M_0), O(1))$-regular.*

Proof Follows from Eq. (L.19), Claim L.13, the chain rule, and the fact that $\|\bar{\mathbf{K}}_i\|, \|\bar{\mathbf{A}}_i\|$ are bounded by universal consants. The regularity statement requires further verifying that all trajectories remain bounded, which follows from Lemmas L.5 and L.8. ■

L.6.1. STABILITY OF THE CONSTRUCTION

We use the following result, whose proof is deferred to Appendix H.3.

Lemma H.4 *Let $\rho, \epsilon > 0$, $L \geq 1$, and $\rho + C\epsilon < 1$. Suppose that there exists a (C, ρ)-stable matrix $\mathbf{A}$ such that*

$$\sup_{\mathbf{x},\mathbf{u}} \|\nabla_{\mathbf{x}} f(\mathbf{x}, \mathbf{u}) - \mathbf{A}\| \leq \epsilon, \quad \sup_{\mathbf{x},\mathbf{u}} \|\nabla_{\mathbf{u}} f(\mathbf{x}, \mathbf{u})\| \leq L.$$

Then, $f(\mathbf{x}, \mathbf{u})$ is (C', ρ') stable such that $\rho' = \rho + C\epsilon$ and $C' = CL$.

Lemma L.15 *There exists universal constants $c' > 0$, $C \geq 1$ and $\rho \in (0, 1)$ such that, if each g is L_0-Lipschitz, and $\tau \leq c' \min\{1, 1/L_0\}$, then for all $(\pi, f) \in \mathcal{P}$, f and (π, f) are globally IISS with $\beta(r, k) = r \cdot C\rho^k$ and $\gamma(r) = Cr$.*

Proof Let $(\pi, f) \in \mathcal{P}$

$$\begin{aligned} f(\mathbf{x}, \mathbf{u}) &= \bar{\mathbf{A}}_i \mathbf{x} + \mathbf{u} - \tau\psi_g(\mathbf{x})\mathbf{e}_1 + \omega \cdot \tau \cdot \mathbf{e}_1(\psi_g(\mathbf{x}) - \psi_u(\mathbf{u}, \mathbf{x})) \\ f^\pi(\mathbf{x}, \mathbf{u}) &= (\bar{\mathbf{A}}_i + \bar{\mathbf{K}}_i)\mathbf{x} + \mathbf{u} + \omega\mathbf{e}_1(\tau^2\psi_g(\mathbf{x}) - \tau\psi_u(\bar{\mathbf{K}}_i\mathbf{x} + \tau\psi_g(\mathbf{x})\mathbf{e}_1 + \mathbf{u}, \mathbf{x})). \end{aligned}$$

Following the proof of Lemma L.14, we surmise that

$$\begin{aligned} \|\nabla_{\mathbf{x}} f(\mathbf{x}, \mathbf{u}) - \bar{\mathbf{A}}_i\| \vee \|\nabla_{\mathbf{x}} f^\pi(\mathbf{x}, \mathbf{u}) - (\bar{\mathbf{A}}_i + \bar{\mathbf{K}}_i)\| &\leq \epsilon_{\nabla,\mathbf{x}} = O(\tau(1 + L_0)) \\ \|\nabla_{\mathbf{u}} f(\mathbf{x}, \mathbf{u})\| \vee \|\nabla_{\mathbf{u}} f^\pi(\mathbf{x}, \mathbf{u})\| &\leq L_{\nabla,\mathbf{u}} = O(1 + \tau) \leq O(1). \end{aligned}$$

The result now follows by observing that $\bar{\mathbf{A}}_i$ and $(\bar{\mathbf{A}}_i + \bar{\mathbf{K}}_i)$ are both (C, ρ)-stable for some $C \geq 1, \rho \in (0, 1)$. Hence, chosing $\tau \leq o_\star(1/(1 + L_0))$, we ensure that have $C\epsilon_{\nabla,\mathbf{x}} \leq (1 + \rho)/2 < 1$. The result now follows from Lemma H.4. ∎

L.6.2. CONTROLLABILITY

For functions $\phi(\mathbf{x}), \psi(\mathbf{x}, \mathbf{u})$ different than those defined above, we can still express $f(\mathbf{x}, \mathbf{u})$ in Construction L.1, as $f(\mathbf{x}, \mathbf{u}) = \phi(\mathbf{x}) + \psi(\mathbf{x}, \mathbf{u}) + \mathbf{u}$, where $\phi(\mathbf{x})$ is $O(1)$-Lipschitz by the computations above, and $\psi(\mathbf{x}, \mathbf{u}) = \omega \cdot \tau \cdot \mathrm{restrict}(\mathbf{x}) \cdot \mathbf{e}_1 \cdot (\langle \mathbf{e}_1, \mathbf{u} \rangle \mathrm{bump}_d(\mathbf{u}))$.

Note that $\omega \in \{-1, 1\}$, and as $\mathrm{bump}_d(\mathbf{u})$ is $O(1)$-Lipschitz and $\mathrm{restrict}(\mathbf{x})$ is $O(1)$-bounded, we can make $\psi(\mathbf{x}, \mathbf{u})$, say, $1/2$-Lipschitz by taking $\tau = o_\star(1)$. Moreover, we clearly also have $\psi(\mathbf{x}, \mathbf{u} = \mathbf{0}) = \mathbf{0}$. Hence, the conditions of Lemma H.5 are met to ensure $O(1)$-one-step-controllability.

Appendix M. Proof for Non-Simple Policies, Theorems 3 and 3.A

In this section, we prove Theorem 3.A. As noted below the statement of Theorem 3.A in Appendix F.2, Theorem 3 follows as as direct consequence.

We begin by recalling the asymptotic notation in Definition F.5. Given $b_1, b_2, \dots \leq 1$, we use the notation $a = \mathrm{poly}\text{-}o^\star(b_1, b_2, \dots, b_k)$ to denote that $a \leq c_1(b_1 \cdot b_2 \cdot b_k)^{c_2}$, c_1 is a sufficiently small universal constant, and c_2 a sufficiently large universal constant. We also recall that we consider the class $\mathbb{A} = \mathbb{A}_{\mathrm{gen,smooth}}(L, M, \alpha, p)$ (Definition F.4) of algorithms which, with probability one, return stochastic, Markovian policies π for which $\mathrm{mean}[\pi](\mathbf{x})$ is L-Lipschitz and M-smooth, and π is (α, p)-anti-concentrated.

Orgnization of the section. In the section below, we give an overview of the proof of Theorem 3.A. We then give natural examples of anti-concentrated policies in Appendix M.2. We then turn in to proving the truncation lemma, Appendix M.3, and establishing useful consequences. We then briefly generalize the Jacobian estimation lemma, Lemma L.7, in Appendix M.4. Penultimately, we provide a statement and proof of compounding error with anti-concentrated policies in Appendix M.5. Finally, in Appendix M.6, we rigorously conclude the proof of Theorem 3.A. Theorem 3 is a corollary of Theorem 3.A, as noted in Appendix J.

M.1. Proof Overview

The construction is identical to the Construction L.1 used in the proof of Theorem 1.A. In particular, the regularity conditions all hold, as do the relations between $\mathsf{M}_{\mathrm{reg},L_2}$, $\mathsf{M}_{\mathrm{expert},L_2}$, and $\mathsf{M}^{\mathbb{A}}_{\mathrm{expert},L_2}$ established in Theorem 1. Our aim is to establish instead the compounding error guarantee, Eq. (F.5), which we restate here for convenience.

$$\mathsf{M}^{\mathbb{A}}_{\mathrm{eval},L_2}(n;\mathcal{P},D,H) \geq c\kappa\cdot\delta\boldsymbol{\epsilon}_n\cdot\min\left\{1.05^{H-2},(1/\boldsymbol{\epsilon}_n)^{\frac{1}{C'(1+\log(1/(\alpha p)))}}\right\}. \qquad \text{(Eq. (F.5))}$$

To this end, we need to modify the two arguments from the proof of Theorem 1.A which required to simply-stochasticity. We instead replace these with arguments that rely on the more general anti-concentration condition (Definition F.3). For convenience, we recall the relevant definitions here.

Definition F.2 (Quantitative Anti-Concentration) *Let $\alpha,p\in(0,1]$. We say that a scalar random variable Z is (α,p)-anti-concentrated if it satisfies*

$$\mathbb{P}[|Z-\mathbb{E}[Z]|\geq\alpha\mathbb{E}[|Z-\mathbb{E}[Z]|^2]^{1/2}]\geq p. \tag{F.4}$$

We say that a random vector $\mathbf{z}\in\mathbb{R}^d$ if (c,p)-anti-concentrated if $\langle\mathbf{v},\mathbf{z}\rangle$ is (α,p)-anti-concentrated for any vector $\mathbf{v}\in\mathbb{R}^d$ (equivalently, for any unit vector).

Definition F.3 (Anti-Concentrated Policy) *We say that a policy π is (α,p) anti-concentrated if, for any $\mathbf{x},\mathbf{x}'\in\mathbb{R}^d$, there exists a coupling $P(\mathbf{x},\mathbf{x}')$ of $\pi(\mathbf{x}),\pi(\mathbf{x}')$*[4] *such that if $(\mathbf{u},\mathbf{u}')\sim P(\mathbf{x},\mathbf{x}')$, the random vector $\mathbf{u}-\mathbf{u}'$ is (α,p)-anti-concentrated.*

To reiterate, there are two arguments in need of ammending. Both arguments appeal to the following property of anti-concentrated random variables, whose proof and useful consequences are deferred to Appendix M.3. This property states that if if a random variable X' dominates in magnitude the sum of anti-concentrated random variable Z and any constant offset, then the expectation of a sufficiently lenient truncation of X' is still large in expectation.

Lemma M.1 (Truncation) *Suppose that Z is scalar, mean zero and (α,p)-anti-concentrated random variable, x a deterministic scalar, and X' a random scalar satisfying, with probability one,*

$$|X'|\geq|x+Z|.$$

Then, for any $\eta\in(0,1)$, setting $B(\eta)=\frac{5}{\eta\alpha^2p^2}$, we have

$$\mathbb{E}[\min\{B(\eta)|x|,|X'|\}]\geq(1-\eta)|x|$$

4. Recall that a coupling of $\pi(\mathbf{x}),\pi(\mathbf{x}')$ is a joint distribution over $(\mathbf{u},\mathbf{u}')$ with marginals $\mathbf{u}\sim\pi(\mathbf{x})$ and $\mathbf{u}'\sim\pi(\mathbf{x}')$.

Next, the first argument to ammend is the one that forces $\nabla \mathrm{mean}[\hat{\pi}](\mathbf{0})\mathrm{Proj}_{\geq 2} \approx \bar{\mathbf{K}}_i \mathrm{Proj}_{\geq 2}$ (Lemma L.7). Building on Lemma M.1, it is is straightforward to generalize this to the anti-concentrated setting, and this step is carried out by Lemma M.6 in Appendix M.4.

The more challenging argument to generalize is the compounding error argument. Our new proof here generalizes mirrors the what occurs in the benevolent gamblers ruin example in Section 5. Leveraging Lemma M.1, we carefully truncate the sequence $(\mathbf{x}_1, \mathbf{x}_2, \dots)$ to form a sequence $(\mathbf{y}_1, \mathbf{y}_2, \dots)$ such that $\mathbf{y}_t \equiv \mathbf{x}_t$ with good probability, that $\mathbb{E}[|\langle \mathbf{e}_1, \mathbf{y}_{t+1}\rangle|] \geq \rho_1 |\langle \mathbf{e}_1, \mathbf{y}_t\rangle|$, and at the same time, $|\langle \mathbf{e}_1, \mathbf{y}_{t+1}\rangle| \leq \rho_2 |\langle \mathbf{e}_1, \mathbf{y}_t\rangle|$, where $1 < \rho_1 < \rho_2$. In particular, if $|\langle \mathbf{e}_1, \mathbf{y}_t\rangle| = \epsilon$, we must have

$$\mathbb{E}|\langle \mathbf{e}_1, \mathbf{y}_{t+1}\rangle| \geq \rho_1^t \epsilon, \quad \text{and} \quad |\langle \mathbf{e}_1, \mathbf{y}_{t+1}\rangle| \leq \rho_2^t \epsilon \text{ w.p. } 1. \tag{M.1}$$

These two bounds imply that $|\langle \mathbf{e}_1, \mathbf{y}_{t+1}\rangle| \geq \rho_1^t \epsilon$ with some probability roughly $(\rho_1/\rho_2)^t$. By a Markov's inequality argument, this yields that $\mathbb{E}[|\langle \mathbf{e}_1, \mathbf{y}_{t+1}\rangle|] \geq \epsilon(\rho_1^2/\rho_2)^t$. Unfortunately, this argugment does not quite work as is because, in general $\rho_2 \gg \rho_1^2$. However, we show a careful modification applies, provided that we can instead lower bound $\mathbb{E}[|\langle \mathbf{e}_1, \mathbf{y}_{t+1}\rangle|^2]^{1/2}$, which can better take advantage of the heavy tails of $|\langle \mathbf{e}_1, \mathbf{y}_{t+1}\rangle|$. The argument is carried out in Appendix M.5.

M.2. Examples of Anti-Concentrated Policies

Before providing examples, we establish a few useful facts about anti-concentrated random variables.

Lemma M.2 (Anti-Concentration via Tail Bounds) *Let Z be a mean-zero scalar random variable satisfying $\mathbb{E}[Z^4] \leq c\mathbb{E}[Z^2]^2$. Then, Z is $(\frac{1}{\sqrt{2}}, \frac{1}{4c})$-anti-concentrated.*

We note the next three lemmas use the 'Z' notation we have been using for scalar random variables, but apply to vector-valued ones by taking projections along vector-directions.

Proof The Paley-Zygmund inequality (Lemma H.6) implies that $\mathbb{P}[Z^2 \geq \theta \mathbb{E}[Z^2]] \geq (1-\theta)^2 \frac{\mathbb{E}[Z^2]^2}{\mathbb{E}[Z^4]}]$ for any $\theta \in (0,1)$. Taking $\theta = 1/2$ proves the statement. ■

Lemma M.3 *Any Gaussian random vector is $(\frac{1}{\sqrt{2}}, \frac{1}{12})$-anti-concentrated.*

Proof For Gaussian random vectors, it suffices to establish the case where $Z \sim \mathcal{N}(0,1)$ (by taking vector directions, scaling, and re-centering). In this case, $3\mathbb{E}[Z^2]^2 = 3 = \mathbb{E}[Z^4]$, so Lemma M.2 applies with $c = 3$. ■

Lemma M.4 *Let Z be discretely distributed on set $\{z_1, z_2, \dots, z_m\}$, and let $p = \min_{1 \leq i \leq m} \mathbb{P}[Z = z_i]$. Then, Z is $(1, p)$-anti-concentrated. In particular, a Dirac-delta is $(1,1)$-anti-concentrated.*

Proof We may assume without loss of generality that $\mathbb{E}[Z] = 0$. For this recentering, let $i_\star := \arg\max_{1 \leq i \leq m} |z_i|$. Then, $\mathbb{E}[Z^2]^{1/2} = \sqrt{\sum_i \mathbb{P}[Z = z_i] z_i^2} \leq |z_{i_\star}|$, and $\mathbb{P}[Z = z_{i_\star}] \geq p$. ■

Lemma M.5 *Generalizing Lemma M.4, let Z be drawn from a discrete mixture of random variables Z_i with mixture weights p_i, each satisfying $p_i \ge p_{\min}$, and which each Z_i (α, p)-anti-concentrated for some $\alpha \le 1$, and is either mean-zero, or symmetric about its mean. Then, Z is $(\alpha, p \cdot p_{\min}/2)$ anti-concentrated.*

Proof Again, by taking projections along unit vectors, we may assume the variables are scalar and centered such that Z has mean zero, and set $i_\star := \arg\max_i \mathbb{E}[|Z_i|^2]$. Then,

$$\mathbb{E}[Z^2] \le \mathbb{E}[|Z_{i_\star}|^2]. \tag{M.2}$$

If $Z_{i_\star}$ has mean zero, then $\mathbb{P}[|Z_{i_\star}| \ge \alpha \mathbb{E}[Z_{i_\star}^2]^{1/2}] \ge p$ as $Z_{i_\star}$ is (α, p) anti-concentrated. Otherwise, suppose without loss of generality that $\mathbb{E}[Z_{i_\star}] > 0$, let $\tilde{Z}_{i_\star} = Z_{i_\star} - \mathbb{E}[Z_{i_\star}]$. By the assume of the lemma, we may take $\tilde{Z}_{i_\star}$ to be symmetric. Then, $\mathbb{P}[\tilde{Z}_{i_\star} \ge \alpha \mathbb{E}[\tilde{Z}_{i_\star}^2]^{1/2}] = \frac{1}{2}\mathbb{P}[|\tilde{Z}_{i_\star}| \ge \alpha \mathbb{E}[\tilde{Z}_{i_\star}^2]^{1/2}] \ge p/2$. Thus,

$$\begin{aligned}
\mathbb{P}[Z_{i_\star} \ge \alpha\sqrt{\mathbb{E}[Z_{i_\star}^2]}] &= \mathbb{P}[\tilde{Z}_{i_\star} + \mathbb{E}[Z_{i_\star}] \ge \alpha\sqrt{\mathbb{E}[\tilde{Z}_{i_\star}^2 + \mathbb{E}[Z_{i_\star}]^2]}] \\
&\ge \mathbb{P}[\tilde{Z}_{i_\star} + \mathbb{E}[Z_{i_\star}] \ge \alpha\sqrt{\mathbb{E}[\tilde{Z}_{i_\star}^2]} + \alpha|\mathbb{E}[Z_{i_\star}]|] && (\sqrt{x+y} \le \sqrt{x} + \sqrt{y}) \\
&\ge \mathbb{P}[\tilde{Z}_{i_\star} \ge \alpha\sqrt{\mathbb{E}[\tilde{Z}_{i_\star}^2]}] && (\alpha \le 1, \mathbb{E}[Z_{i_\star}] > 0 \text{ by assumption}) \\
&\ge p/2. && \text{(established above)}
\end{aligned}$$

In both cases, we obtain $\mathbb{P}[Z_{i_\star} \ge \alpha\sqrt{\mathbb{E}[Z_{i_\star}^2]}] \ge p/2$. Hence, $\mathbb{P}[Z \ge \sqrt{\mathbb{E}[Z^2]}] \ge \mathbb{P}[Z \ge \sqrt{\mathbb{E}[Z_{i_\star}^2]}] \ge \mathbb{P}[Z = Z_{i_\star}]\,\mathbb{P}[Z_{i_\star} \ge \sqrt{\mathbb{E}[Z_{i_\star}^2]}] \ge p_{\min} p/2$. ■

We now list a number of examples of anti-concentrated properties, illustrating that the condition is natural and easy to meet.

Example 2 (Simply Stochastic Policies) *Any simply stochastic policy is $(1,1)$ anti-concentrated, because there exists a coupling P of $\pi(\mathbf{x})$ and $\pi(\mathbf{x}')$ under which $(\mathbf{u}, \mathbf{u}') \sim P(\mathbf{x}, \mathbf{x}')$ ensures $\mathbf{u} - \mathbf{u}'$ is deterministic. This is the coupling which sets $\mathbf{u} = \mathrm{mean}[\pi](\mathbf{x}) + \boldsymbol{\zeta}$ and $\mathbf{u}' = \mathrm{mean}[\pi](\mathbf{x}') + \boldsymbol{\zeta}$, where $\boldsymbol{\zeta}$ is the noise distribution. Implicitly, this is the coupling we use in the proof of Theorem 1.A. In particular, discrete policies are anti-concentrated.*

Example 3 (Gaussian Policies) *Gaussian policies are also anti-concentrated. Consider any π of the form $\pi(\mathbf{x}) = \mathrm{Normal}(\mathrm{mean}[\pi](\mathbf{x}), \Sigma(\mathbf{x}))$, and let $P(\mathbf{x}, \mathbf{x}') = \pi(\mathbf{x}) \otimes \pi(\mathbf{x}')$ denote the independent coupling. Then, $(\mathbf{u}, \mathbf{u}') \sim P(\mathbf{x}, \mathbf{x}')$ is jointly Gaussian, and thus so is $\mathbf{u} - \mathbf{u}'$. Hence, it is $(\frac{1}{\sqrt{2}}, \frac{1}{12})$-anti-concentrated by Lemma M.3.*

Example 4 (Benevolent Gambler's Ruin Policy) *Recall the benevolent gambler's ruin policy from Section 5. At each point, the policy is a mixture of two Dirac-distributions, each with probability $1/2$. Hence, under the independent coupling, $P(\mathbf{x}, \mathbf{x}') = \pi(\mathbf{x}) \otimes \pi(\mathbf{x}')$, $\mathbf{u} - \mathbf{u}'$ is a mixture of at most 4 Dirac-deltas, each with probability at least $1/2$. Hence, it is $(1, 1/4)$-anti-concentrated by Lemma M.4*

Example 5 (Mixture of Gaussian Policies) *If $\pi(\mathbf{x})$ is point-wise a mixture of Gaussians, with minimal probability of each component p, then under the independent coupling $P(\mathbf{x}, \mathbf{x}') = \pi(\mathbf{x}) \otimes \pi(\mathbf{x}')$, $\mathbf{u} - \mathbf{u}'$ is a mixture of Gaussians with minimal component probability at least p^2. Moreover, each component distribution, being a sum of two Gaussians, is Gaussian and thus both symmetric and $(\frac{1}{\sqrt{2}}, \frac{1}{12})$-anti-concentrated by Lemma M.3. Thus, the mixture is $(\frac{1}{\sqrt{2}}, \frac{p^2}{24})$-anti-concentrated by Lemma M.5.*

M.3. The Truncation Lemma (Lemma M.1) and Its Consequences

We prove the core truncation lemma, and then state and prove two useful corollaries.
Proof [Proof of Lemma M.1] Let $\Delta = \mathrm{Var}[Z]$, and assume $x > 0$ without loss of generality (the $x < 0$ follows by symmetry, and $x = 0$ case can be checked directly). We consider two cases. First, assume $\Delta \geq C|x|$, where we pick $C = \frac{2}{cp}$. Let $\mathcal{E} = \{|Z| \geq \alpha C x\}$. On $\mathcal{E}$, we have

$$|X'| \geq |Z| - x - \epsilon \geq \alpha C x - (x) \geq (\frac{2}{p}x - x) \geq x/p.$$

Therefore, $\mathbb{E}[\min\{|X'|, x/p] \geq \mathbb{P}[E]x/p \geq x$.

Next, assume $\Delta \leq \frac{2(1+\gamma)x}{\alpha p}$. Then,

$$\begin{aligned}
\mathbb{E}[\min\{|X'|, Bx + x\}] &= \mathbb{E}[\min\{|x + \sigma Z|, Bx + x\}] \\
&\geq \mathbb{E}[\mathbf{I}\{|Z| \leq Bx] \min\{|x + \sigma Z|, B + x\}] \\
&\geq \mathbb{E}[\mathbf{I}\{|Z| \leq Bx]\}((1+\gamma)x + \sigma Z)] \\
&= x + \sigma \mathbb{E}[\mathbf{I}\{|Z| \leq Bx]\}Z] \\
&= x - \sigma \mathbb{E}[\mathbf{I}\{|Z| > Bx]\}Z] \\
&\geq (x - \mathbb{E}[\mathbf{I}\{|Z| > Bx]\}|Z|].
\end{aligned}$$

We bound $\mathbb{E}[|Z|\mathbf{I}\{Z > Bx\}] \geq \int_{Bx}^{\infty} \mathbb{P}[|Z| \geq t] \leq \int_{Bx}^{\infty} \frac{\mathbb{E}[Z^2]}{t^2} = \frac{\mathbb{E}[Z^2]}{Bx} \leq \frac{\Delta^2}{Bx}$. Subsituting in $\Delta \leq \frac{2x}{\alpha p}$, we get

$$\mathbb{E}[|Z|\mathbf{I}\{Z > Bx\}] \leq \frac{4x}{\alpha^2 p^2 B}.$$

If we take $B = \frac{4}{\eta \alpha^2 p^2}$ for $\eta \leq 1$, we get $\mathbb{E}[|Z|\mathbf{I}\{Z > B\}] \leq \eta x$, and hence

$$\mathbb{E}[\min\{|X'|, Bx + x\}] \geq (1 - \eta)x.$$

substituting $Bx + x \leq \frac{5x}{\eta \alpha^2 p^2}$ concludes. ■

Corollary M.1 *Suppose that Z is a mean zero and (c, p)-anti-concentrated scalar random variable, x a deterministic scalar, and X' a scalar random variable. Suppose further that for $\gamma > 0$ and $\epsilon \geq 0$, the following holds with probability one:*

$$|X'| \geq |x(1 + \gamma) + Z| - \epsilon$$

Then, we have

$$\mathbb{E}\left[\min\left\{\left(\frac{40\max\{\gamma, \gamma^{-1}\}}{\alpha^2 p^2}\right)|x|, |X'|\right\}\right] \geq (1 + \gamma/2)|x|\epsilon$$

Proof By applying Lemma M.1 to the random variable $|X'| + \epsilon$ and setting $B \leftarrow \frac{5}{\eta\alpha^2 p^2}$, then

$$\begin{aligned}\epsilon + \mathbb{E}[\min\{B(1+\gamma)|x|, |X'|\}] &= \mathbb{E}[\min\{B(1+\gamma)|x| + \epsilon, |X'| + \epsilon\}] \\ &\geq \mathbb{E}[\min\{B(1+\gamma)|x|, |X'| + \epsilon\}] \geq (1+\gamma)(1-\eta)|x|,\end{aligned}$$

or rearranging,

$$\mathbb{E}[\min\{B(1+\gamma)|x|, |X'|\}] = \mathbb{E}[\min\{B(1+\gamma)|x| + \epsilon, |X'| + \epsilon\}] \geq (1+\gamma)(1-\eta)|x| - \epsilon.$$

Take η to be such that $(1+\gamma)(1-\eta) = (1+\gamma/2)$, or $\eta = 1 - \frac{1+\gamma/2}{1+\gamma} = \frac{\gamma}{2(1+\gamma)}$. Then, $B(1+\gamma) = \frac{10(1+\gamma)^2}{\alpha^2 p^2 \gamma} \leq \frac{20(\gamma^2+1)}{\alpha^2 p^2 \gamma} = \frac{40\max\{\gamma,\gamma^{-1}\}}{\alpha^2 p^2}$. ∎

Corollary M.2 *Suppose that Z is a mean zero and (c,p)-anti-concentrated scalar random variable, x a deterministic scalar, and X' a scalar random variable. Furthers suppose that, with probability one,*

$$|X'| \geq |x + Z|,$$

Then, $\mathbb{P}[|X'| \geq |x|/4] \geq \alpha^2 p^2/40$.

Proof From Lemma M.1, we have $(1-\eta)|x| \leq \mathbb{P}[|X'| \geq t|x|]$. We have $\mathbb{E}[\min\{B|x|, |X'|\}] \leq B|x|\,\mathbb{P}[|X'| \geq t|x|] + t|x|\,\mathbb{P}[|X'| \geq t|x|] \leq B|x|\,\mathbb{P}[|X'| \geq t|x|] + t|x|$. Setting $t = \eta$, we have

$$(1-2\eta)|x| \leq B|x|\,\mathbb{P}[|X'| \geq t|x|], \quad \mathbb{P}[|X'| \geq \eta|x|] \geq \frac{(1-2\eta)}{B} = \frac{(1-2\eta)\eta c^2 p^2}{5}.$$

Taking $\eta = 1/4$, the above probability is at least $\alpha^2 p^2/40$. ∎

M.4. Derivative Estimation under Anti-Concentration (Case $Z = 1$)

In this section, we generalize the derivative estimation arguments of Lemma L.7 from simply-stochastic policies to anti-concentrated ones.

Lemma M.6 *Let $\mathrm{Proj}_{\geq 2}$ denote the projection onto coordinates 2-through-d, and let $\hat{\pi}$ be any policy with M-smooth which is (α, p) anti-concentrated (recall Definition F.3) satisfying*

$$\mathbb{P}_{\hat{\pi}, f_{g,(i,\omega)}}[\overline{\mathrm{cost}}_{\mathrm{hard}}(\mathbf{x}_{1:H}, \mathbf{u}_{1:H}) \geq M(2^{-k}\Delta)^2/8] \leq o_\star(\alpha^2 p^2/k^2), \tag{M.3}$$

we have the bound $\|(\hat{\mathbf{K}} - \bar{\mathbf{K}}_i)\mathrm{Proj}_{\geq 2}\|_{\mathrm{F}} \leq 8\sqrt{d}M\Delta 2^{-k}$.

Proof Recall the distribution $\mathcal{D}_{\{Z=1,Y=k\}}$ asthe distribution of $\mathbf{x} \mid Z = 1, Y = k$. Because $\mathbb{P}[Z = 1, Y = k] \propto 1/k^2$, then if $\mathbb{P}_{\hat{\pi}, f_{g,(i,\omega)}, D}[\overline{\mathrm{cost}}_{\mathrm{hard}}(\mathbf{x}_{1:H}, \mathbf{u}_{1:H}) \geq \epsilon] \leq c_0/k^2$. Then, arguing as in Lemma L.7, we can start with

$$\left(\mathbb{E}_{\mathbf{x}_1 \sim \mathcal{D}_{\{Z=1,Y=k\}}}\right) \mathbb{E}_{\mathbf{u} \sim \hat{\pi}(\mathbf{x}_1)}[[\|\bar{\mathbf{K}}_i \mathbf{x}_1 - \mathbf{u}\| \geq \epsilon] \leq O(c_0)$$

Consider the coupling $(\mathbf{x},\mathbf{u})$ and $(\mathbf{x}',\mathbf{u}')$ with $\mathbf{x},\mathbf{x}' \sim D_{\{Z=1\}}$ and $\mathbf{u},\mathbf{u}' \sim \pi(\mathbf{x}),\pi(\mathbf{x}')$ where $\mathbf{x},\mathbf{x}'$ are independent and $\mathbf{u},\mathbf{u}' \sim \hat{P}(\mathbf{x}_1,\mathbf{x}_1')$. By the triangle inequality and a union bound, we can symmetrize to obtain

$$\mathbb{E}_{\mathbf{x},\mathbf{x}'\sim D_{\{Z=1\}}}\mathbb{E}_{\mathbf{u}',\mathbf{u}\sim\hat{P}(\mathbf{x}',\mathbf{x})}[\|\bar{\mathbf{K}}_i(\mathbf{x}'-\mathbf{x})-(\mathbf{u}'-\mathbf{u})\|\geq 2\epsilon]\leq O(c_0)$$

And thus, for all unit vectors $\mathbf{v}$,

$$\mathbb{E}_{\mathbf{x},\mathbf{x}'\sim D_{\{Z=1\}}}\mathbb{E}_{\mathbf{u}',\mathbf{u}\sim\hat{P}(\mathbf{x}',\mathbf{x})}[|\langle\mathbf{v},\bar{\mathbf{K}}_i(\mathbf{x}'-\mathbf{x})-(\mathbf{u}'-\mathbf{u})\rangle|\geq 2\epsilon]\leq O(c_0)$$

We may write $\bar{\mathbf{K}}_i(\mathbf{x}'-\mathbf{x})-(\mathbf{u}'-\mathbf{u}) = \bar{\mathbf{K}}_i(\mathbf{x}'-\mathbf{x}) - \mathrm{mean}[\hat{\pi}](\mathbf{x}') - \mathrm{mean}[\hat{\pi}](\mathbf{x}) + \mathbf{z}$, where $\langle\mathbf{v},\mathbf{z}\rangle$ is (α,ρ) anti-concentrated. It follows from Corollary M.2, a corollary of the main truncation lemma Lemma M.1, that

$$\mathbb{E}_{\mathbf{x},\mathbf{x}'\sim D_{\{Z=1\}}}[|\langle\mathbf{v},\bar{\mathbf{K}}_i(\mathbf{x}'-\mathbf{x})-\mathrm{mean}[\hat{\pi}](\mathbf{x}')-\mathrm{mean}[\hat{\pi}](\mathbf{x})\rangle|\geq 8\epsilon]\leq \mathrm{O}\left(\frac{c_0}{\alpha^2p^2}\right).$$

The result now follows by taking $\epsilon \leq M(2^{-k}\Delta)^2/8$ and invoking Lemma H.9, whose conditions are met as soon as $\frac{c_0}{\alpha^2p^2} = o_\star(1)$, i.e. $c_0 = o_\star(\alpha^2p^2)$.

■

M.5. The Compounding Error Argument (Proposition M.1)

This section establishes a general compounding error argument for anti-concentrated policies. We recall $\overline{\mathrm{cost}}_{\mathrm{hard}}$ as the cost from Construction L.2 in Appendix L. We show that the probability $\overline{\mathrm{cost}}_{\mathrm{hard}}$ exceeds some threshold is sufficiently small (otherwise, of course, large error occurs), then we still observe a compounding error phenomenon.

Condition M.1 *Let P be the uniform distribution over $\xi = (i,\omega)\in\{1,2\}\times\{-1,1\}$. For a given $g\in\mathcal{G}$, we will assume that*

$$\mathbb{E}_{\xi\sim P}\,\mathbb{P}_{\hat{\pi},f_{\xi,g},D}\left[\overline{\mathrm{cost}}_{\mathrm{hard}}(\mathbf{x}_{1:H},\mathbf{u}_{1:H})\geq\epsilon^{.9}\right]\leq\epsilon^{.18}/4. \tag{M.4}$$

We will further assume that $\epsilon = \text{poly-}o^\star(\alpha,p,1/L,1/M,\tau,1/d,\kappa,\delta)$ (recall: this means that ϵ is smaller than some polynomial of sufficiently high degree and with sufficiently small coefficients in these terms).

The goal of this section is to establish the following.

Proposition M.1 *Suppose Condition M.1 holds. Define*

$$K(\epsilon,H) := \min\left\{(1.05)^{H-2},\epsilon^{-\frac{1}{C'(1+\log(1/(\alpha p)))}}\right\}.$$

Then, we have

$$\begin{aligned}&\mathbb{E}_{\xi\sim P}\mathbb{E}_{\hat{\pi},f_{g,\xi},D}\left[\epsilon^{0.85}\wedge\overline{\mathrm{cost}}_{\mathrm{hard}}(\mathbf{x}_{1:H},\mathbf{u}_{1:H})\right]\\&\quad\geq\frac{C_{\mathrm{cost}}}{4}K(\epsilon,H)\mathbb{E}_{\xi\sim P}\mathbb{E}_{\hat{\pi},f_{g,\xi},D_{\{Z=0\}}}\left[\epsilon^{.9}\wedge 2\tau\left|\left\langle\mathbf{e}_1,\mathbf{u}_1-\hat{\pi}_{g,(\cdot)}(\mathbf{x}_1)\right\rangle\right|\right]-4\epsilon^{1.03}.\end{aligned}\tag{M.5}$$

In what follows, for our given policy $\hat{\pi}$, we set

$$\hat{\mathbf{K}} := \nabla\mathrm{mean}[\hat{\pi}](\mathbf{z})\big|_{\mathbf{z}=\mathbf{0}}. \tag{M.6}$$

Properties of the linearized closed-loop system. We apply Lemma M.6 with $k = \log_2(6\sqrt{d}M\Delta/\epsilon^{.4})$. Using $\Delta = \Theta_\star\left(\frac{1}{ML\sqrt{d}}\right)$ from Construction L.1, taking $\boldsymbol{\epsilon} = \text{poly-}o^\star(\text{problem parameters})$ to be sufficiently small, and invoking Condition M.1, we can make the following hold:

Claim M.7 *Under Condition M.1, we have that*

$$\|\mathbf{e}_1^\top(\bar{\mathbf{A}}_i + \hat{\mathbf{K}})\mathrm{Proj}_{\le 2}\| \le \epsilon^{0.4}.$$

Following the proof of Claim L.10, there exists an index i for which the $(1,1)$-entry of the closed loop linearized system $(\bar{\mathbf{A}}_i + \hat{\mathbf{K}})$ has magnitude greater than one. This will be the entry responsible for the large compounding error.

Claim M.8 *Under Condition M.1, there exists an index $i_{\mathrm{bad}} \in \{1,2\}$ for which $|\mathbf{e}_1^\top(\bar{\mathbf{A}}_{i_{\mathrm{bad}}} + \hat{\mathbf{K}})\mathbf{e}_1| := 1+\gamma$, where $\gamma = 1/16$, and $1+\gamma \le 2+L$.*

Proof The first part follows from an argument as in Claim L.10. We also notice that $(1+\gamma) \le |\bar{\mathbf{A}}_i[1]| + \|\nabla \mathrm{mean}[\hat{\pi}](\mathbf{x})\big|_{\mathbf{x}=0}\|_{\mathrm{op}} \le 2 + L$ by Lipschitzness of $\mathrm{mean}[\hat{\pi}]$. ■

Trajectory Coupling. The next step is to define a coupling of two trajectories generated by $\hat{\pi}$ on the $\{Z=0\}$ case, both under the dynamics associated with i_{bad}, but under a the different values of $\omega = \pm 1$.

Definition M.1 (The "plus-and-minus" sequence) *Given index $i \in \{1,2\}$ chosen above, and $g \in \mathcal{G}$ fixed, let $(\mathbf{x}_t^+, \mathbf{x}_t^-)$ denote a joint sequence defined as follows:*

$$\begin{aligned}&\mathbf{x}_1^+ \equiv \mathbf{x}_1^- \sim D_{\{Z=0\}}, \quad \mathbf{u}_1^+ \equiv \mathbf{u}_1^- \sim \hat{\pi}(\mathbf{x}_1^+), \quad \mathbf{u}_t^+, \mathbf{u}_t^- \sim \hat{P}(\mathbf{x}_t^+, \mathbf{u}_t^+), t > 1\\ &\mathbf{x}_{t+1}^+ = f_{g,(i_{\mathrm{bad}},\omega=+1)}(\mathbf{x}_t^+, \mathbf{u}_t^+), \quad \mathbf{x}_{t+1}^- = f_{g,(i_{\mathrm{bad}},\omega=-1)}(\mathbf{x}_t^-, \mathbf{u}_t^-).\end{aligned}$$

We let $\mathscr{T}_{\mathbf{x}}$ denote the random variable with distribution $(\mathbf{x}_{2:H}^+, \mathbf{x}_{2:H}^-)$.

The trajectories defined above make the same initial mistake at $t=1$ but, due to differences in ω, these mistakes are multiplied by opposite directions. See Construction L.1 to that, when $\|\mathbf{u}_1\| \le 1$, we have

$$\langle \mathbf{e}_1, \mathbf{x}_2^+ - \mathbf{x}_2^- \rangle = 2\tau \langle \mathbf{e}_1, \mathbf{u}_1 - \hat{\pi}_{g,(\cdot)}(\mathbf{x}_1)\rangle, \tag{M.7}$$

where $\mathbf{x}_1 = \mathbf{x}_1^+ \equiv \mathbf{x}_1^-$, $\mathbf{u}_1 = \mathbf{u}_1^+ \equiv \mathbf{u}_1^-$, and where $(\cdot)$ above follows from the fact that, when $\mathbf{x}_1 \sim D_{\{Z=0\}}$, $\hat{\pi}_{g,\xi}(\mathbf{x}_1)$ does not depend on ξ.

The truncated sequence. We now introduce another stochastic process which serves as a surrogate for the coupled process defined in Definition M.1, but is truncated in such a way as to facillitate analysis. We will denote random variables from these truncated processd with the letter $\mathbf{y}$. To start, define the stochastic map

$$F(\mathbf{y}, \mathbf{y}') \overset{d}{=} (\bar{\mathbf{A}}_{i_{\mathrm{bad}}}\mathbf{x} + \mathbf{u}, \bar{\mathbf{A}}_{i_{\mathrm{bad}}}\mathbf{y}' + \mathbf{u}'), \quad (\mathbf{u}, \mathbf{u}') \sim \hat{P}(\mathbf{y}, \mathbf{y}'), \tag{M.8}$$

where $\hat{P}(\mathbf{y}, \mathbf{y}')$ is the coupling between $\hat{\pi}(\mathbf{y})$ and $\hat{\pi}(\mathbf{y}')$ for which $\mathbf{u} - \mathbf{u}'$ is $(\mathbf{u}, \mathbf{u}') \sim \hat{P}(\mathbf{y}, \mathbf{y}')$-anti-concentrated (Definition F.3). Before continuing, let us introduce two bits of notation used throughout. We define the clipping operator, which projects onto the ball of radius B:

$$\mathrm{clip}_B(\mathbf{z}) = \begin{cases}\mathbf{z} & \|\mathbf{z}\| \le B\\ B\frac{\mathbf{z}}{\|\mathbf{z}\|} & \|\mathbf{z}\| \ge B_1\end{cases}$$

Definition M.2 (Truncated Process Process) *We define the sequence $B_1 \le B_2 \le \dots$ as follows. For a constant C_{trunc} defined in Lemma M.9, set*

$$B_1 = 8C_{\mathrm{trunc}}\epsilon^{0.9}, \quad B_{t+1} = \rho_\star B_t = 8C_{\mathrm{trunc}}\rho_\star^t\epsilon^{0.9}, \quad \rho_\star = \frac{8 \cdot 40 \max\{\gamma, \gamma^{-1}\}}{\alpha^2 p^2}. \tag{M.9}$$

Let $(\mathbf{x}_2^+, \mathbf{x}_2^-)$ be as Definition M.1. Define the sequence $\tilde{\mathbf{y}}_1 = \mathrm{clip}_{B_1/8}(\mathbf{x}_2^+)$, and $\mathbf{y}_1 = \mathrm{clip}_{B_1/8}(\mathbf{x}_2^-)$. Further, define $(\tilde{\mathbf{y}}_t^{\mathrm{next}}, \mathbf{y}_t^{\mathrm{next}}) \sim F(\tilde{\mathbf{y}}_t, \mathbf{y}_t)$ as follows:

$$\begin{aligned}
\mathbf{y}_{t+1} &= \mathrm{clip}_{B_{t+1}/8}(\mathbf{y}_t^{\mathrm{next}}) \\
\tilde{\mathbf{y}}_{t+1}[1] &= \mathbf{y}_{t+1}[1] + \mathrm{clip}_{B_{t+1}/4}(\tilde{\mathbf{y}}_t^{\mathrm{next}}[1] - \mathbf{y}_t^{\mathrm{next}}[1]) \\
\tilde{\mathbf{y}}_{t+1}[2:d] &= \mathrm{clip}_{B_{t+1}/8}\tilde{\mathbf{y}}_{t+1}^{+}[2:d],
\end{aligned}$$

we use following indexing conventions in popular programming languages such as NumPy, albeit with indexing starting at 1. Let $\mathscr{T}_{\mathbf{y}} = (\tilde{\mathbf{y}}_1, \dots, \tilde{\mathbf{y}}_{H-1}, \mathbf{y}_1, \dots, \mathbf{y}_{H-1})$.

Comparing the coupled sequence and its truncated analogue. Because the coupled $\mathbf{x}$-sequence and truncated $\mathbf{y}$-sequence differ only when $\mathbf{y}$ is subject to clipping, and clipping only arises when sequences exceed a certain magnitude, we can use Condition M.1 to control the TV-distance between $\mathscr{T}_{\mathbf{x}}$ and $\mathscr{T}_{\mathbf{y}}$.

Lemma M.9 *There exists a constant C_{trunc} such that, for our definition $B_1 := 8C_{\mathrm{trunc}}\epsilon^{0.9}$, we have under Condition M.1*

$$\mathrm{TV}(\mathscr{T}_{\mathbf{x}}, \mathscr{T}_{\mathbf{y}}) \le \epsilon^{0.18}/2. \tag{M.10}$$

Proof [Proof of Lemma M.9] From their definitions, we can couple together the $\mathscr{T}_{\mathbf{x}}$ and $\mathscr{T}_{\mathbf{y}}$ trajectory such that, when the clipping operation is never activated, we have

$$\tilde{\mathbf{y}}_t = \mathbf{x}_{t+1}^+, \quad \mathbf{y}_t = \mathbf{x}_{t+1}^-, \quad 1 \le t \le H-1. \tag{M.11}$$

The clipping operator is only ever activated when there is some t for which $\max\{\|\tilde{\mathbf{y}}_t\|, \|\|\mathbf{y}_t\|\} \ge B_{t+1}/8$ (the triangle inequality addresses $\|\tilde{\mathbf{y}}_t - \mathbf{y}_t\| \ge B_{t+1}/4$). As $B_{t+1}/8 \ge B_t \ge B_1$, we that Eq. (M.11) can fail at least only when $\max_{2 \le t \le H} \max\{\|\mathbf{x}_t^+\|, \|\mathbf{x}_t^-\|\} > B_1$. For $B_1 = o_\star(\tau)$, Lemma L.6 ensures that there is a universal constant C_{trunc} such that this occurs only on the event

$$\mathcal{E} = \{\overline{\mathrm{cost}}_{\mathrm{hard}}(\mathbf{x}_{1:H}^+, \mathbf{u}_{1:H}^+) \ge B_1/C_{\mathrm{trunc}}\} \cup \{\overline{\mathrm{cost}}_{\mathrm{hard}}(\mathbf{x}_{1:H}^-, \mathbf{u}_{1:H}^-) \ge B_1/C_{\mathrm{trunc}}\}$$

Note that the condition $B_1 = o_\star(\tau)$ is implied by $\epsilon^{.9} = o_\star(\tau)$ as C_{trunc} is universal.

By a union bound, we can bound

$$\begin{aligned}
\mathrm{TV}(\mathscr{T}_{\mathbf{x}}, \mathscr{T}_{\mathbf{y}}) &= \inf_{\text{couplings}} \mathbb{P}[\mathscr{T}_{\mathbf{x}} \neq \mathscr{T}_{\mathbf{y}}] && \text{(variation representation of TV)}\\
&\le \mathbb{P}[\mathcal{E}] && \text{(argument above)}\\
&\le \mathbb{P}[\{\overline{\mathrm{cost}}_{\mathrm{hard}}(\mathbf{x}^{+}_{1:H}, \mathbf{u}^{+}_{1:H}) \ge B_1/C_{\mathrm{trunc}}\}] + \mathbb{P}[\{\overline{\mathrm{cost}}_{\mathrm{hard}}(\mathbf{x}^{-}_{1:H}, \mathbf{u}^{-}_{1:H}) \ge B_1/C_{\mathrm{trunc}}\}]\\
& && \text{(follows form a union bound)}\\
&= \sum_{\omega \in \{+1,-1\}} \mathbb{P}_{\hat{\pi}, f_{g,(i,\omega)}, D_{\{Z=0\}}}[V(\mathbf{x}_{1:H}, \mathbf{u}_{1:H}) \ge B_1/C_{\mathrm{trunc}}]\\
& && \text{(construnction of coupled sequences, Definition M.1)}\\
&\le \sum_{\omega \in \{+1,-1\}} \mathbb{P}_{\hat{\pi}, f_{g,(i,\omega)}, D_{\{Z=0\}}}[V(\mathbf{x}_{1:H}, \mathbf{u}_{1:H}) \ge \epsilon^{0.9}] && \text{(Definition of } B_1)\\
&\le \epsilon^{.18}/2. && \text{(Condition M.1)}
\end{aligned}$$

■

Establishing compounding error of the truncated sequence. The heart of the argument is now to establish compounding error on the $(\mathbf{y}_t, \tilde{\mathbf{y}}_t)$ sequence. This is achieved by the following lemma, whose proof is deferred to the following subsection below. The key idea is to use show, via the truncation lemma Corollary M.1, that in expectation, the magnitude of $\mathbf{y}_t - \tilde{\mathbf{y}}_t$ along the $\mathbf{e}_1$ axis grows, even after the clipping operation is applied. The application of Corollary M.1 hinges crucially on the anti-concentration of the deviation of the policy $\hat{\pi}$ from its mean. We then use the clipping to ensure that $\mathbf{y}_t, \tilde{\mathbf{y}}_t$ are small enough to ensure the Taylor approximation by the linear system, as well as a certain "off-diagonal term", remain controlled. These allow us to establish a one-step recursion which, when iterated yields the desired lemma.

Lemma M.10 *Suppose that $B_t \le \epsilon^{.8}$. Then, it holds that*

$$(1+\gamma/2)|\langle \mathbf{e}_1, \mathbf{y}_t - \tilde{\mathbf{y}}_t\rangle| - \epsilon_{\mathrm{small}} \le \mathbb{E}[|\langle \mathbf{e}_1, \mathbf{y}_{t+1} - \tilde{\mathbf{y}}_{t+1}\rangle| \mid \mathbf{y}_t, \tilde{\mathbf{y}}_t].$$

In particular, by recursing,

$$\mathbb{E}[|\langle \mathbf{e}_1, \mathbf{y}_{t+1} - \tilde{\mathbf{y}}_{t+1}\rangle| \mid \mathbf{y}_1, \tilde{\mathbf{y}}_1] \ge (1+\gamma/2)^t \left(|\langle \mathbf{e}_1, \mathbf{y}_1 - \tilde{\mathbf{y}}_1\rangle| - \frac{\epsilon_{\mathrm{small}}}{1-(1/(1+\gamma/2))}\right).$$

Establishing Compounding Error in the original sequence. Proceeding from Lemma M.10, we establish compounding error on the $(\mathbf{x}^{+}_t, \mathbf{x}^{-}_t)$ sequence.

Lemma M.11 *Suppose that t is such that $B_{t+1} \le \epsilon^{.85}$ and $t \le H-2$. Then for $\epsilon = \mathrm{poly}\text{-}o^{\star}(1/M)$. we have*

$$\mathbb{E}\left[\epsilon^{0.85} \wedge \left|\left\langle \mathbf{e}_1, \mathbf{x}^{+}_{t+2} - \mathbf{x}^{-}_{t+2}\right\rangle\right|\right] \ge (1+\gamma/2)^t \left(\mathbb{E}\left[B_1 \wedge \left|\left\langle \mathbf{e}_1, \mathbf{x}^{+}_2 - \mathbf{x}^{-}_2\right\rangle\right|\right]\right) - 3\epsilon^{1.03}.$$

Proof [Proof of Lemma M.11] Assume we have that as long as $B_t \le B_{t+1} \le \epsilon^{.8}$. Taking expectations of Lemma M.10, we have

$$\mathbb{E}\left[|\langle \mathbf{e}_1, \mathbf{y}_{t+1} - \tilde{\mathbf{y}}_{t+1}\rangle|\right] \ge (1+\gamma/2)^t \left(\mathbb{E}\left[|\langle \mathbf{e}_1, \mathbf{y}_1 - \tilde{\mathbf{y}}_1\rangle|\right] - \frac{\epsilon_{\mathrm{small}}}{1-(1/(1+\gamma/2))}\right). \tag{M.12}$$

By Claim M.12, we have $|\langle \mathbf{e}_1, \mathbf{y}_{t+1} - \tilde{\mathbf{y}}_{t+1}\rangle| \le B_{t+1}$ and $|\langle \mathbf{e}_1, \mathbf{y}_1 - \tilde{\mathbf{y}}_1\rangle| \le B_1$. Hence,

$$\mathbb{E}[B_{t+1} \wedge |\langle \mathbf{e}_1, \mathbf{y}_{t+1} - \tilde{\mathbf{y}}_{t+1}\rangle|] \ge (1+\gamma/2)^t \left(\mathbb{E}[B_1 \wedge |\langle \mathbf{e}_1, \mathbf{y}_1 - \tilde{\mathbf{y}}_1\rangle|] - \frac{\epsilon_{\text{small}}}{1 - (1/(1+\gamma/2))}\right). \tag{M.13}$$

We may now perform a change-of-measure to the $\mathbf{x}_t^+, \mathbf{x}_t^-$ sequence of Definition M.1. This yields

$$\begin{aligned}
&\mathbb{E}\left[B_{t+1} \wedge \left|\left\langle \mathbf{e}_1, \mathbf{x}_{t+2}^+ - \mathbf{x}_{t+2}^-\right\rangle\right|\right] \\
&\quad \ge (1+\gamma/2)^t \left(\mathbb{E}[B_1 \wedge \left|\left\langle \mathbf{e}_1, \mathbf{x}_2^+ - \mathbf{x}_2^-\right\rangle\right|] - \frac{\epsilon_{\text{small}}}{1 - (1/(1+\gamma/2))}\right). \\
&\quad - B_{t+1}\left(\mathrm{TV}\left((\mathbf{x}_{t+2}^+, \mathbf{x}_{t+2}^-), (\tilde{\mathbf{y}}_{t+1}, \mathbf{y}_{t+1})\right) - B_1(1+\gamma/2)^t \mathrm{TV}\left((\mathbf{x}_2^+, \mathbf{x}_2^-), (\tilde{\mathbf{y}}_1, \mathbf{y}_1)\right)\right).
\end{aligned}$$

Recall the definition of $\mathscr{T}_{\mathbf{x}} = (\mathbf{x}_{2:H}^+, \mathbf{x}_{2:H}^-)$ and $\mathscr{T}_{\mathbf{y}} = (\tilde{\mathbf{y}}_{1:H-1}, \tilde{\mathbf{y}}_{1:H-1})$. Then, both $\mathrm{TV}(\dots)$ terms in the above display are at most $\mathrm{TV}(\mathscr{T}_{\mathbf{x}}, \mathscr{T}_{\mathbf{y}})$. Furthermore, examining the definition of the sequence B_t (Eq. (M.9)), we have $B_1(1+\gamma/2)^t \le B_{t+1}$. In therefore follows that

$$\begin{aligned}
&\mathbb{E}\left[B_{t+1} \wedge \left|\left\langle \mathbf{e}_1, \mathbf{x}_{t+2}^+ - \mathbf{x}_{t+2}^-\right\rangle\right|\right] \\
&\quad \ge (1+\gamma/2)^t \left(\mathbb{E}\left[B_1 \wedge \left|\left\langle \mathbf{e}_1, \mathbf{x}_2^+ - \mathbf{x}_2^-\right\rangle\right|\right] - \frac{\epsilon_{\text{small}}}{1 - (1/(1+\gamma/2))}\right) - 2B_{t+1}\mathrm{TV}\left(\mathscr{T}_{\mathbf{x}}, \mathscr{T}_{\mathbf{y}}\right).
\end{aligned}$$

Recalling $\epsilon_{\text{small}} := \epsilon^{1.2} + M\epsilon^{1.8}$, we have for $\epsilon = \text{poly-}o^\star(1/M)$, that $\epsilon_{\text{small}} \le 2\epsilon^{1.2}$. Notice that $\gamma \ge 1/8$ (Claim M.8), we have $\frac{\epsilon_{\text{small}}}{1-(1/(1+\gamma/2))} \le \frac{\epsilon_{\text{small}}}{1-(16/17)} = 17\epsilon_{\text{small}} \le 34\epsilon^{1.2} \le 5\epsilon^{.3}B_1$. And hence, $(1+\gamma/2)^t \frac{\epsilon_{\text{small}}}{1-(1/(1+\gamma/2))} \le 5\epsilon^{.3}B_{t+1}$. With this simplification, and bounding $\mathrm{TV}(\mathscr{T}_{\mathbf{x}}, \mathscr{T}_{\mathbf{y}}) \le \epsilon^{.18}$, and using $\epsilon = o_\star(1)$, we can bound the above by

$$\begin{aligned}
&\mathbb{E}\left[B_{t+1} \wedge \left|\left\langle \mathbf{e}_1, \mathbf{x}_{t+2}^+ - \mathbf{x}_{t+2}^-\right\rangle\right|\right] \\
&\quad \ge (1+\gamma/2)^t \left(\mathbb{E}\left[B_1 \wedge \left|\left\langle \mathbf{e}_1, \mathbf{x}_2^+ - \mathbf{x}_2^-\right\rangle\right|\right]\right) - \underbrace{B_{t+1}\left(2\mathrm{TV}\left(\mathscr{T}_{\mathbf{x}}, \mathscr{T}_{\mathbf{y}}\right) + 5\epsilon^{.3}\right)}_{\le 3B_{t+1}\epsilon^{0.18}},
\end{aligned}$$

By assumption, $B_{t+1} \le \epsilon^{.85}$, which concludes the proof.

■

Concluding the proof of Proposition M.1. Finally, we derive Proposition M.1 from Lemma M.11. The key steps are to relate errors in the difference between the $(\mathbf{x}^+, \mathbf{x}^-)$ sequence to the magnitude of $\overline{\mathrm{cost}}_{\text{hard}}(\mathbf{x}_{1:H}, \mathbf{u}_{1:H})$, and to select t as large as possibily so as to satisfy $B_{t+1} \le \epsilon^{.85}$.

Proof [Proof of Proposition M.1] Let P denote the uniform distribution on $(i,\omega) \in \{1,2\} \times \{-1,+1\}$. Then,

$$\begin{aligned}
\mathbb{E}\left[\epsilon^{0.85} \wedge \left|\left\langle \mathbf{e}_1, \mathbf{x}_{t+2}^{+} - \mathbf{x}_{t+2}^{-}\right\rangle\right|\right] &\le \mathbb{E}\left[\epsilon^{0.85} \wedge \left|\left\langle \mathbf{e}_1, \mathbf{x}_{t+2}^{+}\right\rangle\right|\right] + \mathbb{E}[\epsilon^{0.85} \wedge \left|\left\langle \mathbf{e}_1, \mathbf{x}_{t+2}^{-}\right\rangle\right|] \\
&\le \left(\mathbb{E}_{\hat{\pi}, f_{g,(i_{\mathrm{bad}},1)}, D_{\{Z=0\}}} + \mathbb{E}_{\hat{\pi}, f_{g,(i_{\mathrm{bad}},-1)}, D_{\{Z=0\}}}\right)[\epsilon^{0.85} \wedge |\langle \mathbf{e}_1, \mathbf{x}_{t+2}\rangle|] \\
&\le 2\mathbb{E}_{\xi\sim P}\mathbb{E}_{\hat{\pi}, f_{g,\xi}, D_{\{Z=0\}}}[\epsilon^{0.85} \wedge |\langle \mathbf{e}_1, \mathbf{x}_{t+2}\rangle|] \\
&\le 4\mathbb{E}_{\xi\sim P}\mathbb{E}_{\hat{\pi}, f_{g,\xi}, D}\left[\epsilon^{0.85} \wedge |\langle \mathbf{e}_1, \mathbf{x}_{t+2}\rangle|\right] \\
&\le 4\mathbb{E}_{\xi\sim P}\mathbb{E}_{\hat{\pi}, f_{g,\xi}, D}\left[\epsilon^{0.85} \wedge \frac{1}{C_{\mathrm{cost}}}\overline{\mathrm{cost}}_{\mathrm{hard}}(\mathbf{x}_{1:H}, \mathbf{u}_{1:H})\right] \\
&\qquad\qquad (\text{Definition of } \overline{\mathrm{cost}}_{\mathrm{hard}} \text{ in Appendix L.1}) \\
&\le \frac{4}{C_{\mathrm{cost}}}\mathbb{E}_{\xi\sim P}\mathbb{E}_{\hat{\pi}, f_{g,\xi}, D}[\epsilon^{0.85} \wedge \overline{\mathrm{cost}}_{\mathrm{hard}}(\mathbf{x}_{1:H}, \mathbf{u}_{1:H})] \\
&\qquad\qquad (C_{\mathrm{cost}} \le 1)
\end{aligned}$$

Moreover, from Eq. (M.7), we have

$$\begin{aligned}
\mathbb{E}\left[B_1 \wedge \left|\left\langle \mathbf{e}_1, \mathbf{x}_2^{+} - \mathbf{x}_2^{-}\right\rangle\right|\right] &= \mathbb{E}_{\hat{\pi}, f_{g,(\cdot)}, D_{\{Z=0\}}}\left[B_1 \wedge 2\tau\left|\left\langle \mathbf{e}_1, \mathbf{u}_1 - \hat{\pi}_{g,(\cdot)}(\mathbf{x}_1)\right\rangle\right| \cdot \mathbf{I}\{\|\mathbf{u}_1\| \le 1\}\right] \\
&\ge \mathbb{E}_{\hat{\pi}, f_{g,(\cdot)}, D_{\{Z=0\}}}\left[B_1 \wedge 2\tau\left|\left\langle \mathbf{e}_1, \mathbf{u}_1 - \hat{\pi}_{g,(\cdot)}(\mathbf{x}_1)\right\rangle\right|\right] - B_1\,\mathbb{P}_{\hat{\pi}, f_{g,(\cdot)}, D_{\{Z=0\}}}[\|\mathbf{u}_1\| > 1]
\end{aligned}$$

where above we used Construction L.1 and where $(\cdot)$ denotes a lack of dependence on the ξ argument in $f_{g,\xi}, \hat{\pi}_{g,\xi}$. Thus, we have that $B_1\,\mathbb{P}_{\hat{\pi}, f_{g,(\cdot)}, D_{\{Z=0\}}}[\|\mathbf{u}_1\| > 1] = \inf_{\xi} B_1\,\mathbb{P}_{\hat{\pi}, f_{g,\xi}, D_{\{Z=0\}}}[\|\mathbf{u}_1\| > 1] \le \inf_{\xi} B_1\,\mathbb{P}_{\hat{\pi}, f_{g,\xi}, D_{\{Z=0\}}}[\overline{\mathrm{cost}}_{\mathrm{hard}}(\mathbf{x}_{1:H}, \mathbf{u}_{1:H}) \ge C_{\mathrm{cost}}] \le B_1\epsilon^{.18}$, where the last inequality uses Condition M.1. Finally, we bound $B_1\epsilon^{.18} \le 8C_{\mathrm{trunc}}\epsilon^{1.08} \le \epsilon^{.103}$ for $\epsilon = o_\star(1)$. Thus,

$$\begin{aligned}
\mathbb{E}[B_1 \wedge \left|\left\langle \mathbf{e}_1, \mathbf{x}_2^{+} - \mathbf{x}_2^{-}\right\rangle\right|] &= \mathbb{E}_{\hat{\pi}, f_{g,(\cdot)}, D_{\{Z=0\}}}\left[B_1 \wedge 2\left|\left\langle \mathbf{e}_1, \mathbf{u}_1 - \hat{\pi}_{g,(\cdot)}(\mathbf{x}_1)\right\rangle\right| \cdot \mathbf{I}\{\|\mathbf{u}_1\| \le 1\}\right] \\
&\ge \mathbb{E}_{\hat{\pi}, f_{g,(\cdot)}, D_{\{Z=0\}}}\left[B_1 \wedge 2\left|\left\langle \mathbf{e}_1, \mathbf{u}_1 - \hat{\pi}_{g,(\cdot)}(\mathbf{x}_1)\right\rangle\right|\right] - \epsilon^{1.03} \\
&= \mathbb{E}_{\xi\sim P}\mathbb{E}_{\hat{\pi}, f_{g,\xi}, D_{\{Z=0\}}}\left[B_1 \wedge 2\left|\left\langle \mathbf{e}_1, \mathbf{u}_1 - \hat{\pi}_{g,(\cdot)}(\mathbf{x}_1)\right\rangle\right|\right] - \epsilon^{1.03}.
\end{aligned}$$

Finally, using $B_1 \ge \epsilon^{.9}$, and combining these results with Lemma M.11 yields

$$\begin{aligned}
&\frac{4}{C_{\mathrm{cost}}}\mathbb{E}_{\xi\sim P}\mathbb{E}_{\hat{\pi}, f_{g,\xi}, D}[\epsilon^{0.85} \wedge \overline{\mathrm{cost}}_{\mathrm{hard}}(\mathbf{x}_{1:H}, \mathbf{u}_{1:H})] \\
&\quad \ge (1+\gamma/2)^t\mathbb{E}_{\xi\sim P}\mathbb{E}_{\hat{\pi}, f_{g,\xi}, D_{\{Z=0\}}}[\epsilon^{.9} \wedge 2\left|\left\langle \mathbf{e}_1, \mathbf{u}_1 - \hat{\pi}_{g,(\cdot)}(\mathbf{x}_1)\right\rangle\right|] - 4\epsilon^{1.03},
\end{aligned}$$

again provided $B_{t+1} \le \epsilon^{0.85}$, as well as $t \le H-2$. For the first constraint on t, we require that $B_{t+1} = 8\epsilon^{.9}C_{\mathrm{trunc}}\rho_\star^t \le \epsilon^{0.85}$, it suffices to take

$$t = \min\left\{H-2, \lfloor \log\left(\frac{\epsilon^{-.05}}{8C_{\mathrm{trunc}}}\right) / \log(\rho_\star) \rfloor\right\} \ge \min\left\{H-2, \frac{1}{2}\log\left(\frac{\epsilon^{-.05}}{8C_{\mathrm{trunc}}}\right) / \log(\rho_\star)\right\}, \tag{M.14}$$

where the inequality follows by checking that $\log\left(\frac{\epsilon^{-.05}}{8C_{\mathrm{trunc}}}\right) / \log(\rho_\star) \ge 1$ for $\epsilon = \text{poly-}o^\star(1/L, \alpha, p) = \text{poly-}o^\star(1/\log(\rho_\star))$. If the $H-2$ in the above minimum is the smaller term, then $(1+\gamma/2)^t = (1+\gamma/2)^{H-2} \ge (1.05)^{H-2}$.

Otherwise,

$$(1+\gamma/2)^t \geq \exp\left(\frac{1}{2}\frac{\log(1+\gamma/2)}{\log(\rho_\star)} \cdot \frac{\epsilon^{-.05}}{8C_{\mathrm{trunc}}}\right). \tag{M.15}$$

As $\gamma \geq 1/8$, one can show that $\log(\rho_\star) = \log(\gamma + \gamma^{-1}) + \log(\mathrm{const} \cdot 1/(\alpha^2 p^2)) \leq \log(1+\gamma/2) + \log(\mathrm{const}) + \log(1/\alpha^2 p^2)$, and thus $\frac{\log(1+\gamma/2)}{\log(\rho_\star)} \geq \frac{1}{C+\log(1/\alpha^2 p^2)}$ for an appropriately large constant C. Hence, for some other C' (using $C_{\mathrm{trunc}} = \mathrm{O}\left(1\right)$), we find that

$$(1+\gamma/2)^t \geq \epsilon^{-\frac{1}{C'(1+\log(1/(\alpha p)))}}. \tag{M.16}$$

This concludes the proof. ■

M.5.1. PROOF OF LEMMA M.10

Throughout, in view of Claim M.7, we assume $\|\mathbf{e}_1^\top(\bar{\mathbf{A}}_i + \bar{\mathbf{K}})\mathrm{Proj}_{\geq 2}\| \leq \epsilon^{0.4}$. Our first step is to show that the $(B_t)_{t\geq 1}$ sequences dominates the terms in $\|\mathbf{y}_t\|, \|\tilde{\mathbf{y}}_t\|$ in magnitude.

Claim M.12 *For all t, we have*

$$\|\mathbf{y}_t\| \vee \|\tilde{\mathbf{y}}_t\| \vee \|\tilde{\mathbf{y}}_t - \mathbf{y}_t\| \leq B_t/2. \tag{M.17}$$

Proof [Proof of Claim M.12] By construction, $\|\mathbf{y}_1\| \vee \|\tilde{\mathbf{y}}_1\| \leq B_1/8$. In general, we have $\|\mathbf{y}_{t+1}\| \leq B_{t+1}/8$, $|\tilde{\mathbf{y}}_{t+1}[1]| \leq B_{t+1}/4 + |\mathbf{y}_{t+1}[1]| \leq B_{t+1}/4 + B_{t+1}/8 \leq 3B_{t+1}/8$, and thus $\|\tilde{\mathbf{y}}_{t+1}\| \leq |\tilde{\mathbf{y}}_{t+1}[1]| + \|\tilde{\mathbf{y}}_{t+1}[2:d]\| \leq B_{t+1}/2$. The bound on $\|\tilde{\mathbf{y}}_t - \mathbf{y}_t\|$ follows by noting $\|\tilde{\mathbf{y}}_t - \mathbf{y}_t\| \leq |\tilde{\mathbf{y}}_t^{\mathrm{next}}[1] - \mathbf{y}_t^{\mathrm{next}}[1]| + \|\tilde{\mathbf{y}}_{t+1}[2:d]\| + \|\mathbf{y}_{t+1}[2:d]\| \leq \frac{B_t}{2}$. ■

The next claim described shows that, on a single time-step, the magnitude of the distance between $(\mathbf{y}^{\mathrm{next}}, \tilde{\mathbf{y}}^{\mathrm{next}}) \sim G(\mathbf{y}, \tilde{\mathbf{y}})$ along the $\mathbf{e}_1$-axis increases, even if subject to truncation.

Claim M.13 *Suppose that $\mathbf{y}, \tilde{\mathbf{y}} \in \mathbb{R}^d$ satisfy $\|\mathbf{y}\|, \|\tilde{\mathbf{y}}\| \leq \epsilon^{0.8}$. Then, there exists a (α, p)-anti-concentrated scalar random variable Z such that $(\mathbf{y}^{\mathrm{next}}, \tilde{\mathbf{y}}^{\mathrm{next}}) \sim G(\mathbf{y}, \tilde{\mathbf{y}})$ can satisfies the inequality*

$$\mathbb{E}[|\langle \mathbf{e}_1, \tilde{\mathbf{y}}^{\mathrm{next}} - \mathbf{y}^{\mathrm{next}} \rangle|] \geq \mathbb{E}[|Z + (1+\gamma)\langle \mathbf{e}_1, \mathbf{y} - \tilde{\mathbf{y}} \rangle|] - \epsilon_{\mathrm{small}}.$$

where $\epsilon_{\mathrm{small}} := \epsilon^{1.2} + M\epsilon^{1.8}$. In particular, by Lemma M.1

$$\mathbb{E}\left[\min\left\{|\langle \mathbf{e}_1, \tilde{\mathbf{y}}_t^{\mathrm{next}} - \mathbf{y}_t^{\mathrm{next}} \rangle|, \ \frac{\rho_\star}{8}|\langle \mathbf{e}_1, \tilde{\mathbf{y}}_t - \mathbf{y}_t \rangle|\right\}\right] \geq (1+\gamma/2)|\langle \mathbf{e}_1, \tilde{\mathbf{y}} - \mathbf{y} \rangle| - \epsilon_{\mathrm{small}}.$$

Proof [Proof of Claim M.13] Note that $\mathbb{E}\langle \mathbf{e}_1, F(\tilde{\mathbf{y}}, \mathbf{y}) \rangle = \langle \mathbf{e}_1, \bar{\mathbf{A}}_{i_{\mathrm{bad}}}(\tilde{\mathbf{y}} - \mathbf{y}) + \mathrm{mean}[\hat{\pi}](\tilde{\mathbf{y}}) - \mathrm{mean}[\hat{\pi}](\mathbf{y}) \rangle$. Hence,

$$\langle \mathbf{e}_1, \tilde{\mathbf{y}}_t^{\mathrm{next}} - \mathbf{y}_t^{\mathrm{next}} \rangle = \langle \mathbf{e}_1, \bar{\mathbf{A}}_{i_{\mathrm{bad}}}(\tilde{\mathbf{y}} - \mathbf{y}) + \mathrm{mean}[\hat{\pi}](\tilde{\mathbf{y}}) - \mathrm{mean}[\hat{\pi}](\mathbf{y}) \rangle + Z,$$

We recall from the construction of F, $Z = \langle \mathbf{e}_1, F(\tilde{\mathbf{y}}, \mathbf{y}) \rangle - \mathbb{E}[\langle \mathbf{e}_1, F(\tilde{\mathbf{y}}, \mathbf{y}) \rangle] \overset{\mathrm{dist}}{=} \tilde{\mathbf{u}} - \mathbf{u} - \mathbb{E}[\tilde{\mathbf{u}} - \mathbf{u}]$ is (α, p)-anti-concentrated under $(\tilde{\mathbf{u}}, \mathbf{u}) \sim \hat{P}(\tilde{\mathbf{y}}, \mathbf{y})$. .

Recall $\hat{\mathbf{K}} = \nabla \mathrm{mean}[\hat{\pi}](\mathbf{0})$, we get $\bar{\mathbf{A}}_{i_{\mathrm{bad}}}(\tilde{\mathbf{y}}-\mathbf{y})+\mathrm{mean}[\hat{\pi}](\tilde{\mathbf{y}})-\mathrm{mean}[\hat{\pi}](\mathbf{y}) = \bar{\mathbf{K}}(\tilde{\mathbf{y}}-\mathbf{y})+\mathbf{w}_0$, where $\mathbf{w}_0$ is a Taylor remainder term, and where by M-smoothness of $\mathrm{mean}[\hat{\pi}]$, the remainder term is at most $\|\mathbf{w}_0\| \le \frac{M}{2}(\|\tilde{\mathbf{y}}\|^2 + \|\mathbf{y}\|^2) \le M\epsilon^{1.6}$.

Moreover, by assumption, $|\mathbf{e}_1^\top(\bar{\mathbf{A}}_{i_{\mathrm{bad}}} + \bar{\mathbf{K}})\mathrm{Proj}_{\ge 2}(\tilde{\mathbf{y}} - \mathbf{y})| \le \epsilon^{0.4}\|\tilde{\mathbf{y}} - \mathbf{y}\| \le \epsilon^{1.2}$. Finally, by assumption, $|\mathbf{e}_1(\bar{\mathbf{A}}_{i_{\mathrm{bad}}} + \hat{\mathbf{K}})\mathbf{e}_1| := (1+\gamma)$. Putting things together,

$$\begin{aligned}
\left|\left\langle \mathbf{e}_1, \tilde{\mathbf{y}}_t^{\mathrm{next}} - \mathbf{y}_t^{\mathrm{next}}\right\rangle\right| &= \left|\left\langle \mathbf{e}_1, \bar{\mathbf{A}}_{i_{\mathrm{bad}}}(\tilde{\mathbf{y}} - \mathbf{y}) + \mathrm{mean}[\pi](\tilde{\mathbf{y}}) - \mathrm{mean}[\pi](\mathbf{y})\right\rangle + Z\right| \\
&\ge |Z + \zeta(1+\gamma)\langle \mathbf{e}_1, \tilde{\mathbf{y}} - \mathbf{y}\rangle| - \left(\epsilon^{1.2} + M\epsilon^{1.8}\right) \\
&\ge \left|Z' + (1+\gamma)\langle \mathbf{e}_1, \tilde{\mathbf{y}} - \mathbf{y}\rangle\right| - \underbrace{\left(\epsilon^{1.2} + M\epsilon^{1.8}\right)}_{=:\epsilon_{\mathrm{small}}}
\end{aligned}$$

where ζ is the sign of $\mathbf{e}_1(\bar{\mathbf{A}}_{i_{\mathrm{bad}}} + \hat{\mathbf{K}})\mathbf{e}_1$ and $Z' = \zeta Z$. Note that anti-concentration of Z implies anti-concentration of any scaling of Z, and hence of Z'. Hence, Corollary M.1 implies:

$$\mathbb{E}\left[\min\left\{\underbrace{\left(\frac{40\max\{\gamma,\gamma^{-1}\}}{\alpha^2 p^2}\right)}_{=:\rho_\star/8}|\langle \mathbf{e}_1, \tilde{\mathbf{y}} - \mathbf{y}\rangle|, \left|\left\langle \mathbf{e}_1, \tilde{\mathbf{y}}_t^{\mathrm{next}} - \mathbf{y}_t^{\mathrm{next}}\right\rangle\right|\right\}\right] \ge (1+\gamma/2)\,|\langle \mathbf{e}_1, \tilde{\mathbf{y}} - \mathbf{y}\rangle| - \epsilon_{\mathrm{small}}.$$

■

Proof [Proof Lemma M.10] Let $(\mathcal{F}_t)$ denote the filtration generated by $(\mathbf{y}_t, \tilde{\mathbf{y}}_t)$. We have

$$\begin{aligned}
&(1+\gamma/2)\,|\langle \mathbf{e}_1, \mathbf{y}_t - \tilde{\mathbf{y}}_t\rangle| - \epsilon_{\mathrm{small}} \\
&\le \mathbb{E}\left[\min\left\{\left|\left\langle \mathbf{e}_1, \mathbf{y}_t^{\mathrm{next}} - \tilde{\mathbf{y}}_t^{\mathrm{next}}\right\rangle\right|, \frac{\rho_\star}{8}|\langle \mathbf{e}_1, \mathbf{y}_t - \tilde{\mathbf{y}}_t\rangle| \mid \mathcal{F}_t\right\}\right] && (\text{ Claim M.13}) \\
&\le \mathbb{E}\left[\min\left\{\left|\left\langle \mathbf{e}_1, \mathbf{y}_t^{\mathrm{next}} - \tilde{\mathbf{y}}_t^{\mathrm{next}}\right\rangle\right|, \frac{\rho_\star}{8}B_t\right\} \mid \mathcal{F}_t\right] && (\text{Claim M.12}) \\
&= \mathbb{E}\left[\min\left\{\left|\left\langle \mathbf{e}_1, \mathbf{y}_t^{\mathrm{next}} - \tilde{\mathbf{y}}_t^{\mathrm{next}}\right\rangle\right|, \frac{B_{t+1}}{8}\right\} \mid \mathcal{F}_t\right]. && (\text{Definition for } (B_t), \text{Eq. (M.9)})
\end{aligned}$$

Lets now relate $\mathbf{y}_t^{\mathrm{next}} - \tilde{\mathbf{y}}_t^{\mathrm{next}}$ to $\mathbf{y}_{t+1} - \tilde{\mathbf{y}}_{t+1}$. Now notice that one of two things may either $\left|\left\langle \mathbf{e}_1, \mathbf{y}_t^{\mathrm{next}} - \tilde{\mathbf{y}}_t^{\mathrm{next}}\right\rangle\right| \ge \frac{B_{t+1}}{8}$, in which case $|\,|\langle \mathbf{e}_1, \mathbf{y}_{t+1} - \tilde{\mathbf{y}}_{t+1}\rangle| \ge \frac{B_{t+1}}{8}$, or else $\left|\left\langle \mathbf{e}_1, \mathbf{y}_t^{\mathrm{next}} - \tilde{\mathbf{y}}_t^{\mathrm{next}}\right\rangle\right| \le \frac{B_{t+1}}{8}$, which case $|\langle \mathbf{e}_1, \mathbf{y}_{t+1} - \tilde{\mathbf{y}}_{t+1}\rangle| = \left|\left\langle \mathbf{e}_1, \mathbf{y}_t^{\mathrm{next}} - \tilde{\mathbf{y}}_t^{\mathrm{next}}\right\rangle\right|$. Hence, we have that

$$(1+\gamma/2)\,|\langle \mathbf{e}_1, \mathbf{y}_t - \tilde{\mathbf{y}}_t\rangle| - \epsilon_{\mathrm{small}} \le \mathbb{E}[|\langle \mathbf{e}_1, \mathbf{y}_{t+1} - \tilde{\mathbf{y}}_{t+1}\rangle| \mid \mathcal{F}_t].$$

By recursing, we then above

$$\begin{aligned}
\mathbb{E}\left[|\langle \mathbf{e}_1, \mathbf{y}_{t+1} - \tilde{\mathbf{y}}_{t+1}\rangle| \mid \mathcal{F}_1\right] &\ge (1+\gamma/2)^t\,|\langle \mathbf{e}_1, \mathbf{y}_1 - \tilde{\mathbf{y}}_1\rangle| - \sum_{s=1}^{t}(1+\gamma/2)^{t-s}\epsilon_{\mathrm{small}} \\
&\ge (1+\gamma/2)^t\left(|\langle \mathbf{e}_1, \mathbf{y}_1 - \tilde{\mathbf{y}}_1\rangle| - \frac{\epsilon_{\mathrm{small}}}{1-(1/(1+\gamma/2))}\right).
\end{aligned}$$

■

M.6. Formal Proof of Lower bound Eq. (F.5) in Theorem 3.A

Proof One of two cases can occur. Either, Eq. (M.4) in Condition M.1 holds, in which case Proposition M.1 ensures

$$\begin{aligned}&\mathbb{E}_{\xi\sim P}\mathbb{E}_{\hat{\pi},f_{g,\xi},D}\left[\epsilon^{0.85}\wedge\overline{\mathrm{cost}}_{\mathrm{hard}}(\mathbf{x}_{1:H},\mathbf{u}_{1:H})\right]\\&\quad\geq\frac{C_{\mathrm{cost}}}{4}K(\epsilon,H)\mathbb{E}_{\xi\sim P}\mathbb{E}_{\hat{\pi},f_{g,\xi},D_{\{Z=0\}}}\left[\epsilon^{.9}\wedge 2\tau\left|\left\langle\mathbf{e}_1,\mathbf{u}_1-\hat{\pi}_{g,(\cdot)}(\mathbf{x}_1)\right\rangle\right|\right]-4\epsilon^{1.03}.\end{aligned}$$

Otherwise, Eq. (M.4) fails, so that

$$\begin{aligned}\mathbb{E}_{\xi\sim P}\,\mathbb{P}_{\hat{\pi},f_{\xi,g},D}[\min\{\epsilon^{.9},\overline{\mathrm{cost}}_{\mathrm{hard}}(\mathbf{x}_{1:H},\mathbf{u}_{1:H})\}^2]^{1/2}&\geq\epsilon^{.9}\sqrt{\mathbb{E}_{\xi\sim P}\,\mathbb{P}_{\hat{\pi},f_{\xi,g},D}\left[\overline{\mathrm{cost}}_{\mathrm{hard}}(\mathbf{x}_{1:H},\mathbf{u}_{1:H})\geq\epsilon^{.9}\right]}\\&\geq\epsilon^{.9}\epsilon^{.09}/2=\epsilon^{.99}/2.\end{aligned}$$

By Jensen's inequality

$$\mathbb{E}_{\xi\sim P}\mathbb{E}_{\hat{\pi},f_{g,\xi},D}\left[\epsilon^{0.85}\wedge\overline{\mathrm{cost}}_{\mathrm{hard}}(\mathbf{x}_{1:H},\mathbf{u}_{1:H})\right]\leq\mathbb{E}_{\xi\sim P}\mathbb{E}_{\hat{\pi},f_{g,\xi},D}\left[\epsilon^{1.7}\wedge\overline{\mathrm{cost}}_{\mathrm{hard}}(\mathbf{x}_{1:H},\mathbf{u}_{1:H})^2\right]^{1/2}.$$

Therefore, we find that

$$\begin{aligned}&\mathbb{E}_{\xi\sim P}\mathbb{E}_{\hat{\pi},f_{g,\xi},D}\left[\epsilon^{1.7}\wedge\overline{\mathrm{cost}}_{\mathrm{hard}}(\mathbf{x}_{1:H},\mathbf{u}_{1:H})^2\right]^{1/2}+C_{\mathrm{cost}}\epsilon^{1.03}\\&\quad\geq\min\left\{\epsilon^{.99}/2,\frac{1}{4}C_{\mathrm{cost}}K(\epsilon,H)\mathbb{E}_{\xi\sim P}\mathbb{E}_{\hat{\pi},f_{g,\xi},D_{\{Z=0\}}}\left[\epsilon^{.9}\wedge 2\tau\left|\left\langle\mathbf{e}_1,\mathbf{u}_1-\hat{\pi}_{g,(\cdot)}(\mathbf{x}_1)\right\rangle\right|\right]\right\}.\end{aligned}$$

Moreover, for ϵ sufficiently small as in Condition M.1, we have that

$$\mathbb{E}_{\xi\sim P}\mathbb{E}_{\hat{\pi},f_{g,\xi},D_{\{Z=0\}}}\left[\epsilon^{.9}\wedge 2\tau\left|\left\langle\mathbf{e}_1,\mathbf{u}_1-\hat{\pi}_{g,(\cdot)}(\mathbf{x}_1)\right\rangle\right|\right]\geq 2\tau^2\kappa\epsilon\mathbb{E}_{\xi\sim P}\mathbb{E}_{\hat{\pi},f_{g,\xi},D_{\{Z=0\}}}[\left\langle\mathbf{e}_1,\mathbf{u}_1-\hat{\pi}_{g,(\cdot)}(\mathbf{x}_1)\right\rangle\geq\kappa\tau\epsilon],$$

and by modifying the constant C' in the term

$$K(\epsilon,H):=\min\left\{(1.05)^{H-2},\epsilon^{-\frac{1}{C'(1+\log(1/(\alpha p)))}}\right\},$$

to be at least $C'\geq 100$ and using $C_{\mathrm{cost}}\leq 1$ (see Construction L.2), we may ensure that

$$\begin{aligned}&\min\left\{\epsilon^{.99}/2,\frac{1}{4}C_{\mathrm{cost}}K(\epsilon,H)\mathbb{E}_{\xi\sim P}\mathbb{E}_{\hat{\pi},f_{g,\xi},D_{\{Z=0\}}}\left[\epsilon^{.9}\wedge 2\tau\left|\left\langle\mathbf{e}_1,\mathbf{u}_1-\hat{\pi}_{g,(\cdot)}(\mathbf{x}_1)\right\rangle\right|\right]\right\}\\&\quad\geq\mathbb{E}_{\xi\sim P}\mathbb{E}_{\hat{\pi},f_{g,\xi},D_{\{Z=0\}}}\left[\left\langle\mathbf{e}_1,\mathbf{u}_1-\hat{\pi}_{g,(\cdot)}(\mathbf{x}_1)\right\rangle\geq\kappa\tau\epsilon\right],\end{aligned}$$

Rearranging,

$$\begin{aligned}&\mathbb{E}_{\xi\sim P}\underbrace{\left(\mathbb{E}_{\hat{\pi},f_{g,\xi},D}\left[\epsilon^{1.7}\wedge\overline{\mathrm{cost}}_{\mathrm{hard}}(\mathbf{x}_{1:H},\mathbf{u}_{1:H})^2\right]^{1/2}+C_{\mathrm{cost}}\epsilon^{1.03}\right)}_{=:\mathbf{R}(\hat{\pi},g,\xi)}\\&\quad\geq\left(\frac{C_{\mathrm{cost}}K(\epsilon,H)\tau^2\kappa\epsilon}{2}\right)\mathbb{E}_{\xi\sim P}\mathbb{E}_{\hat{\pi},f_{g,\xi},D_{\{Z=0\}}}\left[\left\langle\mathbf{e}_1,\mathbf{u}_1-\hat{\pi}_{g,(\cdot)}(\mathbf{x}_1)\right\rangle\geq\kappa\tau\epsilon\right].\end{aligned}$$

We now invoke Proposition G.1(d) with $\mathsf{R}(\hat{\pi}, g, \xi)$ as defined above, and following the proof of Theorem 1.A given in Appendix L.5. Here, recall $\epsilon = \boldsymbol{\epsilon}_n = \mathsf{M}_{\mathrm{reg},L_2}(n, \mathcal{G}, D_{\mathrm{reg}})$, and that $(\mathcal{G}, D_{\mathrm{reg}})$ is (κ, δ)-typical. From Proposition G.1(d) and a slight bit of rearranging,

$$\sup_{g\in\mathcal{G},\xi} \mathbb{E}_{\xi\sim P}\mathbb{E}_{\hat{\pi},f_{g,\xi},D}\left[\epsilon^{1.7} \wedge \overline{\mathrm{cost}}_{\mathrm{hard}}(\mathbf{x}_{1:H}, \mathbf{u}_{1:H})^2\right]^{1/2} \ge \left(\frac{C_{\mathrm{cost}} K(\epsilon, H)\tau^2\kappa\epsilon}{2}\right)\delta - C_{\mathrm{cost}}\epsilon^{1.03}.$$

Lastly, we recall that $K(\epsilon, H) \ge 1$, $\tau, C_{\mathrm{cost}} = \Omega(1)$, and and that for $\epsilon = \text{poly-}o^\star(\kappa, \delta)$, there is some small universal constant c such that $\left(\frac{C_{\mathrm{cost}} K(\epsilon,H)\tau^2\kappa\epsilon}{2}\right)\delta - C_{\mathrm{cost}}\epsilon^{1.03} \ge K(\epsilon, H)\epsilon\kappa\delta$. Hence,

$$\begin{aligned}\sup_{g\in\mathcal{G},\xi} \mathbb{E}_{\xi\sim P}\mathbb{E}_{\hat{\pi},f_{g,\xi},D}\left[\epsilon^{1.7} \wedge \overline{\mathrm{cost}}_{\mathrm{hard}}(\mathbf{x}_{1:H}, \mathbf{u}_{1:H})^2\right]^{1/2} &\ge c\cdot\kappa\delta\cdot K(\epsilon, H)\epsilon \\ &=: c\cdot\kappa\delta\min\left\{\epsilon 1.05^{H-2}, \epsilon^{1-\frac{1}{C'(1+\log(1/(\alpha p)))}}\right\},\end{aligned}$$

as needed. ■

Appendix N. Proof for Unstable Dynamics, Theorems 4 and 4.A

In this section, we establish exponential compounding error for unstable dynamical systems, Theorem 4.A. We note that Theorem 4 follows as a direct consequence of Theorem 4.A, as noted below the statement of the latter theorem in Appendix F.3.

We prove Theorem 4.A by first establishing a variant, Theorem 6, which pertains to time-varying systems, and is proven in Appendix N.1 below. A time-varying dynamical system is just a dynamical system $f(\mathbf{x}, \mathbf{u}, t)$, which may depend on the arbitrarily on t. Similarly, we allow the expert $\pi^\star(\mathbf{x}, t)$ to also depend on a t-argument. In Appendix N.2, we proceed to establishing Theorem 4.A my modifying the construction to hold for systems and policies which do not vary with the argument t.

We now turn to the statement of Theorem 6. Below $\mathbb{O}(d)$ denotes the orthogonal group, that is, the set of matrices in $\mathbb{R}^d$ with orthonormal columns, and $\mathbf{P}_{\le k}$ the projection onto the first k cannonical basis elements.

Construction N.1 *Let $(\mathcal{G}, D_{\mathrm{reg}})$ be an (k, ℓ_2)-regression family, and let $\rho > 2$. We define a (d, d)-IL family $(\mathcal{P}, D)$, where*

(a) D draws $\mathbf{z} \sim D_{\mathrm{reg}}$ and appends $d - k$ zeros to $\mathbf{z} \in \mathbb{R}^k$ to form $\mathbf{x} = (\mathbf{z}, \mathbf{0}) \in \mathbb{R}^d$.

(b) Let $\xi = (\mathbf{O}_2, \mathbf{O}_2, \dots)$ denote sequences in $\mathbb{O}(d)$, and let $g \in \mathcal{G}$. We take $\mathcal{P}$ is the set of all instances $(\pi^\star, f)$ of the following form:

$$\pi_{g,\xi}(\mathbf{x}, t) = \begin{cases} g(\mathbf{P}_{\le k}\mathbf{x})\mathbf{e}_1 & t = 1 \\ -\rho\mathbf{O}_t\mathbf{x}, & t > 1 \end{cases}, \quad f_{g,\xi}(\mathbf{x}, \mathbf{u}, t) = \mathbf{u} - \pi_{g,\xi}(\mathbf{x}, t)$$

The above construction follows the schematic of Proposition G.1, and the same proof plan sketched in Section 4: the learner makes a mistake in the first step, due to uncertainty over the class $\mathcal{G}$, and then must contend with uncertainty over the dynamics in the time steps that follow. Recall that we aim for lower bounds that hold in an *unrestricted sense*, and apply even to learner's which select time-varying, history dependent policies $\hat{\pi}$. This renders simpler constructions that do not incorporate rotational uncertainty insufficient:

Remark N.1 (Scaled-Identity Dynamics do not suffice) *One could imagine a simplified construction where either $\pi_{g,\xi}(\mathbf{x}, t) = \sigma\rho\mathbf{I}$ or $\pi_{g,\xi}(\mathbf{x}, t) = \sigma_t\rho\mathbf{I}$ is the identity, scaled by ρ, and multipled by either a fixed sign $\sigma \in \{-1, 1\}$ or a time varying sign $\sigma_t \in \{-1, 1\}$. Noting that $\pm\mathbf{I} \in \mathbb{O}(d)$, these constructions are a restriction of the class in Construction N.1. Unfortunately, these constuctions do yield unconditional lower bounds. For a fixed sign σ, a history-dependent learner can identify the dynamics, whereas for a time-varying sign σ_t, the benevolent gambler's ruin strategy (Section 5) mitigates compounding error.*

We consider the following challenging cost function:

$$\mathrm{cost}_{\mathrm{hard,time\,var}}(\mathbf{x}_{1:H}, \mathbf{u}_{1:H}) = \max_{1\le t\le H} \min\{1, \mathrm{cost}_{\mathrm{hard},t}(\mathbf{x}_t, \mathbf{u}_t)\}, \quad \mathrm{cost}_{\mathrm{hard},t}(\mathbf{x}, \mathbf{u}) = \mathbf{I}\{t \ge 1\}\|\mathbf{x}\| \tag{N.1}$$

A salient property of the construction and associated cost function is the following observation:

Observation N.1 *For any $(\pi, f) \in \mathcal{P}$ and D as above, $\mathbb{P}_{\pi,f,D}[\mathbf{x}_t = \mathbf{0}$ and $\mathbf{u}_t = \mathbf{0}, \quad \forall t \ge 2] = 1$. In particular, $\mathrm{cost}_{\mathrm{hard,time\,var}}$ vanishies on $(\mathcal{P}, D)$.*

This observation ensures that trajectories after time step $t \ge 2$ are uninformative. Using this fact, we will establish the following lower bound:

Theorem 6 (Time-varing analogue of Theorem 4.A) *For any $k, d \in \mathbb{N}$ with $k \le d$ and (k, ℓ_2)-regression family $(\mathcal{G}, D_{\mathrm{reg}})$ satisfying Condition E.1, and $\rho \ge 1$, the construction above is such that such that for all $2 \le H \le \frac{1}{2}\exp((1-\rho^{-1})^2 d/2)$,*

$$\mathsf{M}_{\mathrm{expert},L_2}(n; \mathcal{P}, D, H) = \mathsf{M}_{\mathrm{reg},L_2}(n; \mathcal{G}, D_{\mathrm{reg}}) =: \boldsymbol{\epsilon}_n \tag{N.2}$$

$$\mathsf{M}_{\mathrm{cost}_{\mathrm{hard,time\,var}}}(n, p; \mathcal{P}, D, H)\left(n, \frac{\delta}{2}; \mathcal{P}, D, H\right) \ge \kappa\boldsymbol{\epsilon}_n\rho^{(H-1)/2} \tag{N.3}$$

N.1. Proof of Theorem 6

We begin with a lemma which demonstrates that no control policy, even one which depends arbitrarily on history, can avoid compounding error when faced with time varying dynamics given by random rotation matrices:

Lemma N.2 (Compounding Error with Orthonormal Matrices) *Consider a stochastic dynamical system with*

$$\mathbf{x}_{t+1} = \rho\mathbf{O}_t\mathbf{x}_t + \mathbf{u}_t, \quad \mathbf{O}_t \overset{\mathrm{i.i.d}}{\sim} \mathbb{O}(d),$$

Let $\mathbf{u}_t$ be chosen by any (possibly stochastic) control policy such that the conditional distribution of $\mathbf{u}_t$ depends only on $\mathbf{x}_{1:t}, \mathbf{u}_{1:t-1}$. Then, for all $1 \le t \le H$, it holds that

$$\mathbb{P}[\forall 1 \le t \le H, \|\mathbf{x}_{t+1}\| \ge (\sqrt{1-\alpha}\rho)^t\|\mathbf{x}_1\|] \ge 1 - e^{\frac{d\alpha^2}{2}}H. \tag{N.4}$$

In particular, $\alpha = 1 - 1/\rho$, we obtain

$$\mathbb{P}[\forall 1 \le t \le H, \|\mathbf{x}_{t+1}\| \ge (\rho)^{t/2}\|\mathbf{x}_1\|] \ge 1 - e^{\frac{d(1-\rho^{-1})^2}{2}}H. \tag{N.5}$$

Proof

By a union bound, it suffices to show that, for any $\mathbf{u}_t$ conditioned on the past,

$$\mathbb{P}[\|\mathbf{x}_{t+1}\| \geq \sqrt{1-\alpha}\rho\|\mathbf{x}_t\|] \geq 1 - e^{\frac{d\alpha^2}{2}}.$$

We have that

$$\|\mathbf{x}_{t+1}\|^2 = \|\rho\mathbf{O}_t\mathbf{x}_t + \mathbf{u}_t\|^2 = \rho^2\|\mathbf{x}_t\|^2 + \|\mathbf{u}_t\|^2 + 2\rho\|\mathbf{u}_t\|\|\mathbf{x}_t\|\cos\theta(\mathbf{O}_t\mathbf{x}_t, \mathbf{u}_t),$$

where $\theta(\mathbf{O}_t\mathbf{x}_t, \mathbf{u}_t)$ is the angle between the argument vectors. Using the elementary inequality $ab \leq a^2 + b^2$, we can then lower bound the above by

$$\|\mathbf{x}_{t+1}\|^2 \geq (1 - \cos\theta(\mathbf{O}_t\mathbf{x}_t, \mathbf{u}_t))\left(\rho^2\|\mathbf{x}_t\|^2 + \|\mathbf{u}_t\|^2\right) \geq \rho^2(1 - \cos\theta(\mathbf{O}_t\mathbf{x}_t, \mathbf{u}_t))\|\mathbf{x}_t\|^2. \quad \text{(N.6)}$$

Since $\mathbf{O}_t \sim \mathbb{O}(d)$ and is independent of $\mathbf{u}_t$, $\theta(\mathbf{O}_t\mathbf{x}_t, \mathbf{u}_t)$ has the distribution of the angle between a fixed vector and a uniform vector on the sphere. A standard concentration inequality shows then that

$$\mathbb{P}\left[\cos\theta(\mathbf{O}_t\mathbf{x}_t, \mathbf{u}_t) \geq \frac{t}{\sqrt{d}}\right] \leq e^{-t^2/2}$$

Taking $t = \alpha\sqrt{d}$, we have that $\mathbb{P}[\cos\theta(\mathbf{O}_t\mathbf{x}_t, \mathbf{u}_t) \geq \alpha] \leq \exp(-d\alpha^2/2)$. On this event, the Eq. (N.6) gives

$$\|\mathbf{x}_{t+1}\| \geq \rho\sqrt{1-\alpha}\|\mathbf{x}_t\|,$$

as needed.

■

Continuing proof instantiates the arguments of the general schematic in Appendix G; we encourage the reader to review that section before continuing to read the present. We instantiate Appendix G by introduce the parameter $\xi = \{\mathbf{O}_2, \mathbf{O}_3, \ldots, \mathbf{O}_H\} \in \mathbb{O}(d)^{H-1}$. We also let P denote the uniform distribution of ξ, i.e., where $\mathbf{O}_t$ are drawn i.i.d from the Haar measure on $\mathbb{O}(d)$. A direct consequence of the previous lemma is as follows.

Corollary N.1 *For any arbitrary (even stateful, time-dependent) policy $\hat{\pi}$, and P the uniform distribution over ξ, and any $g \in \mathcal{G}$, we have*

$$\begin{aligned}
&\mathbb{E}_{\xi\sim P}\,\mathbb{P}_{\hat{\pi}, f_{g,\xi}, D}[\|\mathbf{x}_H\| \geq \epsilon \cdot \rho^{(H-1)/2} \mid \|\mathbf{x}_2\| \geq \epsilon] \cdot \mathbb{E}_{\xi\sim P} \\
&\geq \left(1 - H\exp\left(\frac{d(1-\rho^{-1})^2}{2}\right)\right) \cdot \mathbb{E}_{\xi\sim P} \geq \frac{1}{2},
\end{aligned}$$

where the last inequality holds for our choice of H in Theorem 6.

We now turn to invoking Proposition G.1. To do so, we begin checking that its conditions hold.

Lemma N.3 *The construction satisfies the three conditions of Appendix G, Properties G.1 to G.3, with $\tau = 1$. Hence, Proposition G.1 applies (with $\tau = 1$).*

Proof Properties G.1 and G.2 can be checked directly. Here, we we prove that, $(\pi_{g,\xi}, f_{g,\xi})$ are on-policy indistinguishable under D (Property G.3). By Observation N.1, $\mathbf{x}_t, \mathbf{u}_t$ vanish with probability one ujnder $\mathbb{P}_{\pi,f,D}$ for all $(\pi, f) \in \mathcal{P}$. Thus, all that remains is to show that $(\pi_{g,\xi}, f_{g,\xi})$ and $(\pi_{g,\xi'}, f_{g,\xi'})$ induces the same distribution over $\mathbf{x}_1, \mathbf{u}_1, \mathbf{x}_2$. By construction,

$$f_{g,\xi}(\mathbf{x}, 1) = f_{g,\xi'}(\mathbf{x}, 1), \quad \pi_{g,\xi}(\mathbf{x}, 1) = \pi_{g,\xi'}(\mathbf{x}, 1) \quad \forall \xi, \xi'.$$

Therefore the distributions over $\mathbf{x}_1, \mathbf{x}_2, \mathbf{u}_1$ are identical for all ξ under a given g, D. This concludes the verification of Property G.3. ■

Directly from Proposition G.1, Eq. (F.6) holds: $\mathbf{M}_{\mathrm{expert},L_2}(n; \mathcal{P}, D, H) = \mathbf{M}_{\mathrm{reg},L_2}(n; \mathcal{G}, D_{\mathrm{reg}})$. Moreover, define the risk

$$\mathbf{R}_\epsilon(\hat{\pi}; g, \xi) = \mathbb{P}_{\hat{\pi}, f_{g,\xi}, D}[\mathrm{cost}_{\mathrm{hard,time\,var}}(\mathbf{x}_{1:H}, \mathbf{u}_{1:H}) \geq \epsilon \cdot \rho^{(H-1)/2}]. \tag{N.7}$$

Letting P be the uniform product distribution over $\xi = (\mathbf{O}_t)_{t \geq 2}$ as in Lemma N.2, we have that for any $g \in \mathcal{G}$, and any fixed ξ_0,

$$\begin{aligned}
&\mathbb{E}_{\xi \sim P} \mathbf{R}_\epsilon(\hat{\pi}; g, \xi) \\
&\geq \mathbb{E}_{\xi \sim P}\, \mathbb{P}_{\hat{\pi}, f_{g,\xi}, D}[\|\mathbf{x}_H\| \geq \epsilon \cdot \rho^{(H-1)/2}] && (\text{Defnition of } \mathrm{cost}_{\mathrm{hard,time\,var}}) \\
&= \mathbb{E}_{\xi \sim P}\, \mathbb{P}_{\hat{\pi}, f_{g,\xi}, D}[\|\mathbf{x}_H\| \geq \epsilon \cdot \rho^{(H-1)/2} \mid \|\mathbf{x}_2\| \geq \epsilon] \cdot \mathbb{E}_{\xi \sim P}\, \mathbb{P}_{\hat{\pi}, f, D}[\|\mathbf{x}_2\| \geq \epsilon] \\
&\geq \frac{1}{2} \mathbb{E}_{\xi \sim P}\, \mathbb{P}_{\hat{\pi}, f_{g,\xi}, D}[\|\mathbf{x}_2\| \geq \epsilon] && (\text{Corollary N.1}) \\
&\geq \frac{1}{2} \cdot \mathbb{P}_{\hat{\pi}, f_{g,\xi_0}, D}[|\langle \mathbf{e}_1, \hat{\mathbf{u}} - \pi_{g,\xi_0}(\mathbf{x}) \rangle| \geq \epsilon]. && (\text{Construction N.1, } \xi_0 \text{ arbitrary})
\end{aligned}$$

To conclude, set $\boldsymbol{\epsilon}_n = \mathbf{M}_{\mathrm{reg},L_2}(\mathcal{G}, D_{\mathrm{reg}})$. By Lemma N.3, and the fact that $(\mathcal{G}, D_{\mathrm{reg}})$ satisfies Condition E.1, we may invokve Proposition G.1(c), from which it follows that

$$\begin{aligned}
&\sup_{g,\xi} \mathbb{E}_{\mathrm{S}_{n,H} \sim (\pi_{g,\xi}, f_{g,\xi})} \mathbb{E}_{\hat{\pi} \sim \mathrm{alg}(\mathrm{S}_{n,H})}\, \mathbb{P}_{\hat{\pi}, f_{g,\xi}, D}[\mathrm{cost}_{\mathrm{hard,time\,var}}(\mathbf{x}_{1:H}, \mathbf{u}_{1:H}) \geq \boldsymbol{\epsilon}_n \kappa \cdot \rho^{(H-1)/2}] \\
&:= \sup_{g,\xi} \mathbb{E}_{\mathrm{S}_{n,H} \sim (\pi_{g,\xi}, f_{g,\xi})} \mathbb{E}_{\hat{\pi} \sim \mathrm{alg}(\mathrm{S}_{n,H})} \mathbf{R}_\epsilon(\hat{\pi}; f_{g,\xi}, D)\big|_{\epsilon = \kappa \boldsymbol{\epsilon}_n} \geq \frac{\delta}{2}.
\end{aligned}$$

Hence, the proof follows after recalling the definition of $\mathbf{M}_{\mathrm{cost}_{\mathrm{hard,time\,var}},\mathrm{prob}}$ from Definition E.4.

N.2. Proof of Theorem 4.A from Theorem 6

For some universal constant radius $r_0 \in (0, 1)$, a standard covering argument Vershynin (2018, Section 4) implies that there exists a set of points $\mathbf{y}_1, \ldots, \mathbf{y}_N$, $N = \exp(d/2)$, such that $\mathcal{B}(\mathbf{y}_i, 3r_0) \cap \mathcal{B}(\mathbf{y}_j, 3r_0)$ are disjoint for any $1 \leq i \neq j \leq N$. We now define the function

$$\psi_i(\mathbf{x}) := \mathrm{bump}_d((\mathbf{x} - \mathbf{y}_i)/r_0), \tag{N.8}$$

where $\mathrm{bump}_d(\cdot)$ is the smooth bump function of Lemma H.14.

Construction N.2 *Let $(\mathbf{y}_i)_{i \geq 1}$ be the packing centers, as above. Let $(\mathcal{G}, D_{\mathrm{reg}})$ be an (k, ℓ_2)-regression family, such that D_{reg} is supported on a ball of radius r_0. Let $\rho > 2$. We define a (d, d)-IL family $(\mathcal{P}, D)$ via* *(a) D draws $\mathbf{z} \sim D_{\mathrm{reg}}$ and appends $d-k$ zeros to form $\mathbf{x} = \mathbf{y}_1 + (r_0\mathbf{z}, \mathbf{0}) \in \mathbb{R}^d$.*

(b) $\mathcal{P}$ is the set of all instances of the following form:

$$\begin{aligned}\pi(\mathbf{x}) &= \psi_1(\mathbf{x}) g\left(\frac{\mathbf{P}_{\le k}\mathbf{x} - \mathbf{y}_1}{r_0}\right)\mathbf{e}_1 + \sum_{t=2}^{N-1} -\rho \mathbf{O}_t(\mathbf{x} - \mathbf{y}_t)\psi_t(\mathbf{x}) \\ f(\mathbf{x}, \mathbf{u}) &= \mathbf{u} - \pi(\mathbf{x}).\end{aligned}$$

where $g \in \mathcal{G}, \mathbf{O}_t \in \mathbb{O}(d)$.

We also define a new hard cost

$$\mathrm{cost}_{\mathrm{hard,tiv}}(\mathbf{x}, \mathbf{u}) := \frac{1}{C_{\mathrm{cost}}} \sum_{t=2}^{N} \psi_t(\mathbf{x}) \|\mathbf{x} - \mathbf{y}_t\|.$$

The following lemma establishes all relevant regularity conditions, including that $\mathrm{cost}(\mathbf{x}_{1:H}, \mathbf{u}_{1:H}) := \max_{1 \le h \le H} \min\{1, \mathrm{cost}_{\mathrm{hard,tiv}}(\mathbf{x}, \mathbf{u})\} \in \mathcal{C}_{\mathrm{lip,max}}$, where we recall from Definition J.4 that $\mathcal{C}_{\mathrm{lip,max}}$ consists of all costs of the form $\max_{h \ge 1} \tilde{\mathrm{cost}}(\mathbf{x}_h, \mathbf{u}_h)$ for which $\tilde{\mathrm{cost}}$ is $1-$Lipschitz and takes values in $[0, 1]$.

Lemma N.4 *Suppose $(\mathcal{G}, D_{\mathrm{reg}})$ is (R, L, M)-regular. For any dimension d, π and f are $O(L+\rho)$-Lipschitz and $O(L + M + \rho)$-smooth. Similarly, $\mathrm{cost}_{\mathrm{hard,tiv}}(\mathbf{x}, \mathbf{u})$ is 1 Lipschitz for some $C_{\mathrm{cost}} = O(1)$. Moreover, (π, f) is $(1, 0)$-EISS. Finally, each f is $O(L + \rho)$-one-step-controllable.*

Proof [Proof of Lemma N.4] Since the $3r_0$-balls around each $\mathbf{y}_i$ are disjoint for differen $\mathbf{y}_i$, we have that for any $\mathbf{x}$, either $\mathbf{x}$ lies in exactly one $\mathcal{B}(\mathbf{y}_i, 3r_0)$ for some i and $\|\mathbf{x} - \mathbf{y}_i\| \le 2.5r_0$, lies in exactly one such ball but $\|\mathbf{x} - \mathbf{y}_i\| \ge 2.5r_0$, or lies in no such ball. In the latter two cases, $\psi_i(\mathbf{x})$ definition of G vanish at $\mathbf{x}$, so $\nabla\psi_i(\mathbf{x}), \nabla^2\psi_i(\mathbf{x})$ vanish. Hence, upper bounding the derivatives and Hessians of the above terms amounts to upper bounding the maximal contribution from any i. This is bounded because each $\|\mathbf{y}_t\| \le 1$, $\|\mathbf{O}_t\| = 1$, and and $\psi_i(\mathbf{x})$ is $O(1)$-Lipschitz and $O(1)$-smooth. The first claim now follows from the chain and product rules, using the fact that $g(\cdot)$ is L-Lipschitz and M-smooth by assumption.The guarantee for $\mathrm{cost}_{\mathrm{hard,tiv}}$ is similar.

To see that (π, f) is $(1, 0)$-EISS, we observe that $f^{\pi}(\mathbf{x}, \mathbf{u}) = \mathbf{u}$. For controllability, we invoke the special case of Lemma H.5 with $\phi(\mathbf{x}) = \pi(\mathbf{x})$ being $O(L + \rho)$-Lipschitz, and $\psi(\mathbf{x}, \mathbf{u}) \equiv \mathbf{0}$. ■

We can now prove Theorem 4.A.

Proof [Proof of Theorem 4.A] Let $(\mathcal{P}, D)$ be as in the time-invariant construction Construction N.2. Consider the time-varying invertible rigid transformation

$$G_t(\mathbf{x}) = \mathbf{x} - \mathbf{y}_t. \tag{N.9}$$

We can directly check that expert trajectories under the time-invariant construction Construction N.2 are equivalent to those under the time varying one Construction N.1, after applying G_t. Hence, because we can invert each $G_t(\cdot)$, the equivalence of the IL training risk and supervised learning risk,

$$\mathsf{M}_{\mathrm{expert},L_2}(n; \mathcal{P}, \mathcal{D}, H) = \mathsf{M}_{\mathrm{reg},L_2}(n; \mathcal{G}, D_{\mathrm{reg}}), \tag{N.10}$$

as in Theorem 6, remains true for Construction N.2.

Moreover, with probability one $\mathbb{P}_{\pi,f,D}$ for (π, f), we have that $\mathbf{x}_1$ lies in a ball of radius r_0 around $\mathbf{y}_1$, and $\mathbf{x}_t = \mathbf{y}_t$ for all $t \geq 2$. Thus, with probability 1, $\mathrm{cost}_{\mathrm{hard,tiv}}$ vanishes on these trajectories.

We also see that outside of the event $\{\max_{1\leq h\leq H} \mathrm{cost}_{\mathrm{hard,tiv}}(\mathbf{x}_h, \mathbf{u}_h) \geq C_{\mathrm{cost}} r_0\}$, the imitiator trajectories under Construction N.2 and Construction N.1 by are also related the transformation G_t. Hence, for $\epsilon \leq C_{\mathrm{cost}} r_0$,

$$\begin{aligned}
&\inf_{\mathrm{alg}\in\mathbb{A}} \sup_{(\pi^\star, f)\in(\mathcal{P},D)} \mathbb{E}_{\mathrm{S}_{n,H}} \mathbb{E}_{\hat{\pi}\sim\mathrm{alg}(\mathrm{S}_{n,H})} \mathbb{P}_{\hat{\pi},f,D,H}[\max_{1\leq t\leq H} \mathrm{cost}_{\mathrm{hard,tiv}}(\mathbf{x}_t, \mathbf{u}_t) \geq C_{\mathrm{cost}}\epsilon] \\
&= \inf_{\mathrm{alg}\in\mathbb{A}} \sup_{(\pi^\star, f)\in(\mathcal{P}',D')} \mathbb{E}_{\mathrm{S}_{n,H}} \mathbb{E}_{\hat{\pi}\sim\mathrm{alg}(\mathrm{S}_{n,H})} \mathbb{P}_{\hat{\pi},f,D',H}[\mathrm{cost}_{\mathrm{hard,time\,var}}(\mathbf{x}_{1:H}, \mathbf{u}_{1:H}) \geq \epsilon]
\end{aligned}$$

where $\mathcal{P}', \mathcal{D}'$ are from the time-varying construction, Construction N.1. Applying the definition of $\mathsf{M}_{\mathrm{cost}}$ in . . . , the above implies

$$\begin{aligned}
\mathsf{M}_{\mathrm{cost}}(n, \frac{\delta}{2}; \mathcal{P}, D, H) &\geq \min\left\{C_{\mathrm{cost}} r_0, \mathsf{M}_{\mathrm{cost}}(n, \frac{\delta}{2}; \mathcal{P}', D', H)\right\} \\
&\geq \min\left\{C_{\mathrm{cost}} r_0, \kappa \mathsf{M}_{\mathrm{reg},L_2}(n; \mathcal{G}, D_{\mathrm{reg}})\right\} \\
&\geq \min\left\{C_{\mathrm{cost}} r_0, \kappa \mathsf{M}_{\mathrm{expert},L_2}(n; \mathcal{P}, \mathcal{D}, H)\right\}. \qquad \text{(Eq. (N.10))}
\end{aligned}$$

To conclude, recall (1) $C_{\mathrm{cost}} r_0$ is a universal constant and (2) from Lemma N.4, cost is the maximum of 1-Lipschitz costs, therefore lying in $\mathcal{C}_{\mathrm{lip,max}}$; hence, $\mathsf{M}_{\mathrm{cost}}(n, \frac{\delta}{2}; \mathcal{P}, D, H) \leq \mathsf{M}_{\mathrm{eval,prob}}(n, \frac{\delta}{2}; \mathcal{P}, D, H)$. ■

Appendix O. Non-Simple Policies Circumvent the Construction in Theorem 1

This section shows that the ideas in Section 5 formally avoid exponential compounding error for the construction used in Theorem 1.

Definition O.1 (Chunked-Policies) *For $\ell \in \mathbb{N}$, let $\mathbb{A}_{\mathrm{chunk}}(L, M, 3)$ consider the set of algorithms which return* action-chunked *determinsitic policies of the form $\mathbf{x}_{\ell h+1} \mapsto (\mathbf{u}_{\ell h+1}, \mathbf{u}_{\ell h+2}, \mathbf{u}_{\ell h+2})$, which predicts sequences of ℓ control actions at each time $t = \ell h + 1$, each executed in open loop, for which the mapping $\mathbf{x}_{\ell h+1} \mapsto (\mathbf{u}_{\ell h+1}, \mathbf{u}_{\ell h+2}, \mathbf{u}_{\ell h+2})$ is L-Lipschitz and M-smooth. Note that $\mathbb{A}_{\mathrm{chunk}}(L, M, \ell = 1) = \mathbb{A}_{\mathrm{smooth}}(L, M)$.*

Definition O.2 (Periodic Time-Varying policies) *Let $\mathbb{A}_{\mathrm{period}}(L, M, 3)$ denote the set of algorithms which return, with probability one, periodically time varying policies, $\hat{\pi}(\cdot, 0), \ldots, \hat{\pi}(\cdot, \ell-1)$, which select $\mathbf{u}_t$ as $\mathbf{u}_t \leftarrow \hat{\pi}(\mathbf{x}_t, (t-1) \mod \ell)$, and $\hat{\pi}(\cdot, i)$ is L-Lipschitz and M-smooth for $0 \leq i \leq \ell$. Note that $\mathbb{A}_{\mathrm{period}}(L, M, \ell = 1) = \mathbb{A}_{\mathrm{smooth}}(L, M)$.*

In what follows, we bound a strong notion of minimax risk, $\mathsf{R}_{\mathrm{traj},L_2}$, which we recall satisfies $\mathsf{R}_{\mathrm{traj},L_2} \geq \mathsf{R}_{\mathrm{traj},L_1} \geq \sup_{\mathrm{cost}\in\mathcal{C}_{\mathrm{Lip}}} \mathsf{R}_{\mathrm{cost}}$.

Proposition O.1 *Consider the construction Construction L.1 of Theorems 1 and 1.A. Let $\mathcal{G}$ be convex, and a regular regression convex (Definition E.6), with $O(1)$ Lipschitzness and smoothness parameters. Finally, take $\epsilon_n := \mathsf{M}_{\mathrm{reg},L_2}(n; \mathcal{G}, D_{\mathrm{reg}})$, where D_{reg} is the regression initial distribution form Construction L.1. Then,* *(a) Let $\mathbb{A} = \mathbb{A}_{\text{simple}}(L, \infty)$. Then,*

$$\mathbf{M}^{\mathbb{A}}(n, \mathbf{R}_{\text{traj},L_2}; \mathcal{P}, D, H) \lesssim \exp(-cn) + \boldsymbol{\epsilon}_{n/3}$$

(b) Let $\mathbb{A} = \mathbb{A}_{\text{gen,smooth}}(L, M, 1/4, 1/4)$ for $L, M = O(1)$ and $\alpha, p = \Omega(1)$. Then, for some universal $q \in (0,1)$,

$$\mathbf{M}^{\mathbb{A}}(n, \mathbf{R}_{\text{traj},L_2}; \mathcal{P}, D, H) \lesssim \exp(-cn) + \boldsymbol{\epsilon}_{n/3}^{1-q}$$

(c) Let $\mathbb{A} = \mathbb{A}_{\text{chunk}}(L, M, 3)$ or $\mathbb{A}_{\text{period}}(L, M, 3)$, denoting the set of either 3-action-chunked or periodic-with-period-3 policies defined in Definitions O.1 and O.2, respectively. Then,

$$\mathbf{M}^{\mathbb{A}}(n, \mathbf{R}_{\text{traj},L_2}; \mathcal{P}, D, H) \leq \exp(-cn) + \boldsymbol{\epsilon}_{n/3}$$

In particular, if we consider the regression classes of Proposition E.1 (which are those used to instantiate Theorem 1), then the above all hold with $\boldsymbol{\epsilon}_{n/3} \leftarrow n^{-s/k}$. This gives a form directly comparable with Theorem 1.

O.1. Proof Sketch of Proposition O.1

Proof For brevity, we keep the proof slightly terser than the others in this paper; still, we make sure to provide all essential details.

For notational convience, we notate history-dependent, possibly stochastic policies $\hat{\pi}(\mathbf{x}_{1:t}, t)$ which may depend on past states, inputs, and the time index t. Note that this subsumes all classes of policies in the proposition we aim to prove. For example, $\mathbb{A}_{\text{smooth}}$ and $\mathbb{A}_{\text{gen,smooth}}$ are attained by ignoring all but $\mathbf{x}_t$, periodic policies depend on onlly $\mathbf{x}_t$ and t. The case of chunked policies will require some minor-modifications, which we defer to the end of the proof.

First, using the definition of minimax risk, and optimality of proper algorithms for convex classes, we can find for any n a regression algorithm $\mathrm{alg}_{\text{reg},n}$, which is proper for $\mathcal{G}$ such that, say,

$$\sup_{g \in \mathcal{G}} \mathbb{E}_{S_{n,\text{reg}} \sim (g, D_{\text{reg}})} \mathbb{E}_{\hat{g} \sim \mathrm{alg}_{\text{reg},n}} \mathbb{E}_{\mathbf{z} \sim D_{\text{reg}}} [|(\hat{g} - g)(\mathbf{z})|^2]^{1/2} \leq 2\boldsymbol{\epsilon}_n,$$

where the factor 2 may be replaced by any constant strictly greater than 1.

Now, let $\mathrm{bump}_d(\cdot)$ denote the bump function, and let π_0 be a "base" policy to be specified later. We then apply the following BC algorithm:

1. Collect the sample of n-trajectories $\mathrm{S}_{n,H}$

2. Count how many correspond to the $Z = 0$ case (this is possible since the initial states on each trajectory for $Z = 0$ and $Z = 1$ have disjoint support). Call this number n_0, and form the set $S_{n_0,\text{reg}} := \{(\mathrm{Proj}_{\leq 3}(\mathbf{x}_1^{(i)} - \mathbf{x}_{\text{offset}}), \frac{1}{\tau}\langle \mathbf{e}_1, \mathbf{u}_1^{(i)} \rangle) : \mathbf{x}_1^{(i)} \text{ corresponds to the } Z = 0 \text{ case}\}$. Note that, conditioned on n_0, and for a BC instance indexed by a given $g \in \mathcal{G}$, $S_{n_0,\text{reg}}$ has the distribution of n_0 pairs $(\mathbf{z}, \mathbf{y})$ from the associated regression problem with regression function g and initial distribution D_{reg}.

3. Call $\mathrm{alg}_{\text{reg},n_0}$ on $S_{n_0,\text{reg}}$ to obtain $\hat{g}$. Note that by convexity of $\mathcal{G}$, we may take $\hat{g} \in \mathcal{G}$, so that $\hat{g}$ can be smooth.

4. For the given base policy π_0 to be described (and specialized for each class of function), return

$$\hat{\pi}(\mathbf{x}_{1:t}, \mathbf{u}_{1:t-1}, t) = \bar{\mathbf{K}}_1(1 - \mathbf{e}\mathbf{e}_1^\top)\mathbf{x}_t \tag{O.1}$$
$$+ \operatorname{bump}_d(\mathbf{x}_t) \cdot \pi_0(\mathbf{x}_{1:t}, \mathbf{u}_{1:t-1}, t) \tag{O.2}$$
$$+ \tau \cdot \operatorname{restrict}(\mathbf{x}) \cdot \mathcal{T}[\hat{g}](\mathbf{x})\mathbf{e}_1, \tag{O.3}$$

where $\operatorname{restrict}(\mathbf{x})$ and $\mathcal{T}[\hat{g}]$ are as in Construction L.1.

Recall the matrices $\bar{\mathbf{A}}_i$, $i \in \{1, 2\}$ in the construction. In each case, π_0 will select some $\bar{\mathbf{A}}_i(\mathbf{I} - \mathbf{e}_1\mathbf{e}_1^\top)\mathbf{x}$, for some appropriately chosen $i \in \{1, 2\}$. By examining Construction L.1, we see that for π_0 of this form, then for initial states sampled on the event $Z = 1$, $\hat{\pi}$ perfectly matches the expect trajectories (see the proof of Lemma L.8, which replices on the fact that $(\bar{\mathbf{K}}_1 - \bar{\mathbf{K}}_2)\mathbf{x} = 0$ for $\mathbf{x}$ perpendicular to $\mathbf{e}_1$). Hence,

$$\mathbf{R}_{\mathrm{traj},L_2}(\hat{\pi}; \pi^\star, \mathcal{D}, D_{\mathrm{reg}}) \lesssim \mathbf{R}_{\mathrm{traj},L_2}(\hat{\pi}; \pi^\star, \mathcal{D}, D_{\{Z=0\}}). \tag{O.4}$$

Let's turn to bounding the right-hand side. Let $(\mathbf{x}_t^\star, \mathbf{u}_t^\star)$ and $(\hat{\mathbf{x}}_t, \hat{\mathbf{u}}_t)$ denote random variables from the canonical coupling of trajectories from $(\pi^\star, f, D_{\{Z=0\}})$ and $(\hat{\pi}, f, D_{\{Z=0\}})$, as in Definition I.1. Observe that $\mathbf{x}_1^\star = \hat{\mathbf{x}}_1$, and $\mathbf{u}_t^\star \equiv \mathbf{x}_t^\star = \mathbf{0}$ for $t > 1$. Hence,

$$\mathbf{R}_{\mathrm{traj},L_2}(\hat{\pi}; \pi_{g,\xi}^\star, f_{g,\xi}, D_{\{Z=0\}}) \lesssim \sqrt{\mathbb{E}[\min\{1, \|\hat{\mathbf{u}}_1 - \mathbf{u}_1^\star\|^2\}]} + \sum_{t \geq 2} \sqrt{\mathbb{E}[\min\{1, \|\hat{\mathbf{x}}_t\|^2 + \|\hat{\mathbf{u}}_t\|^2\}]}, \tag{O.5}$$

where above $\mathbb{E} = \mathbb{E}_{\hat{\pi}, \pi_{g,\xi}^\star, f_{g,\xi}, D_{\{Z=0\}}}$, and all random variables are as in the canonical coupling.

By using a similar argument to that of Appendix L.5 (where, with probability $1 - \exp(-\Omega(n))$, we have at least $n_0 \geq n/3$ samples) used for estimating $\hat{g}$. Let us call this event $\mathcal{E}$ over the sampling. Conditioned on n_0 (and $\mathcal{E}$), there are $n_0 \geq n/3$ i.i.d. samples from $D_{\{Z=0\}}$. Using the embedding of the regression problem into the control problem in Construction L.1 and our choice of $\hat{g}$ estimator, we see that the error at time $t = 1 \mid \{Z = 0\}$ is

$$\mathbb{E}_{\mathrm{S}_{n,H} | n_0 \geq n/3} \sqrt{\mathbb{E}_{\hat{\pi}, \pi_{g,\xi}^\star, f_{g,\xi}, D_{\{Z=0\}}}[\|\hat{\mathbf{u}}_1 - \mathbf{u}_1^\star\|^2]} \lesssim \mathbb{E}_{S_{n_0,\mathrm{reg}} | n_0 \geq 3} \sqrt{\mathbb{E}_{\mathbf{z} \sim D_{\mathrm{reg}}} |\hat{g}(\mathbf{z}) - g(\mathbf{z})|^2} \lesssim \boldsymbol{\epsilon}_{n/3} \tag{O.6}$$

where above we use the Construction L.1 and the embedding of the $\mathcal{G}$-regression problem. By the same token, and again using the structure of the construction and the form of our policy $\hat{\pi}$ as above (for this, given $Z = 0$, we have $\hat{\mathbf{x}}_2$ is in the $\mathbf{e}_1$-span, and $\|\hat{\mathbf{x}}_2\| \propto |\hat{g}(\mathbf{z}) - g(\mathbf{z})|^2$)

$$\mathbb{E}_{D_{\{Z=0\}} | n_0 \geq n/3} \sqrt{\mathbb{E}_{\hat{\pi}, \pi_{g,\xi}^\star, f_{g,\xi}, D_{\{Z=0\}}}[\|\hat{\mathbf{x}}_2\|^2]} \lesssim \boldsymbol{\epsilon}_{n/3}.$$

Thus, it remains to show that, starting from $\hat{\mathbf{x}}_2$ satisfying the above expectation, each of the above policies will mitigate compounding error. We also note that, by Markov's inequality, we can assume that $\|\hat{\mathbf{x}}_2\|^2 \leq 1/C$ for some sufficiently large C which probability at least $1 - O(\boldsymbol{\epsilon}_{n/3})$. Hence, using clipping of errors to 1, we can bound (also accoutning for the $\exp(-\Omega(n))$ event where $n_0 \leq n/3$)

$$\mathbf{R}_{\mathrm{traj},L_2}(\hat{\pi}; \pi_{g,\xi}^\star, f_{g,\xi}, D_{\{Z=0\}})$$
$$\lesssim \boldsymbol{\epsilon}_{n/3} + \exp(-\Omega(n)) + \mathbb{E}_{D_{\{Z=0\}} | n_0 \geq n/3} \sum_{t \geq 2} \sqrt{\mathbb{E}[\min\{1, (\|\hat{\mathbf{x}}_t\|^2 + \|\hat{\mathbf{u}}_t\|^2\} \mathbf{I}\{\|\hat{\mathbf{x}}_2\| = o_\star(1)\}]}.$$

We now handle the various cases, again quite tersely.

(a) For **Part (a)**, we apply the same concentric stabilization trick along the $\mathbf{e}_1$ axis as in Section 5, but now where in each interval in the $\mathbf{e}_1$ direction, we either play $\pi_0(\mathbf{x}) = -\bar{\mathbf{A}}_i(\mathbf{I} - \mathbf{e}_1\mathbf{e}_1)^\top\mathbf{x}$ for $i \in \{1, 2\}$. As we can take $\|\hat{\mathbf{x}}_2\|^2 \le 1/C$ to be small, state magnitudes grow at most by a constant on the first few steps, and we see we still remain within the linear region of the construction. Then, within three at most 3 time-steps, the $\mathbf{e}_1$ component becomes set to 0, and remains at zero by the structure of the $(\bar{\mathbf{A}}_i, \bar{\mathbf{K}}_i)$ matrices. Finally, $\bar{\mathbf{K}}_1$ stabilizes either $\bar{\mathbf{A}}_i$ as long as states are orthogonal to the $\mathbf{e}_1$ direction, which keeps a constant compounding error. This establishes that

$$\mathbb{E}_{D_{\{Z=0\}}|n_0 \ge n/3} \sum_{t \ge 2} \sqrt{\mathbb{E}[\min\{1, (\|\hat{\mathbf{x}}_t\|^2 + \|\hat{\mathbf{u}}_t\|^2\}\mathbf{I}\{\|\hat{\mathbf{x}}_2\| = o_\star(1)\}]} \lesssim \sqrt{\mathbb{E}_{D_{\{Z=0\}}|n_0 \ge n/3}\mathbb{E}[\|\hat{\mathbf{x}}_t\|^2]} \lesssim \boldsymbol{\epsilon}_{n/3}.$$

(b) Rather than using concentric stabilization of π_0, we use the benevolent Gambler's Ruin construction to alternative π_0 between each of $\bar{\mathbf{A}}_i(\mathbf{I} - \mathbf{e}_1\mathbf{e}_1^\top)\mathbf{x}$, $i \in \{1, 2\}$ i.i.d. with probability $1/2$. One can show that, by mirroring the argument Section 5, for some $q \in (0, 1)$,

$$\begin{aligned}&\mathbb{E}_{D_{\{Z=0\}}|n_0 \ge n/3} \sum_{t \ge 2} \sqrt{\mathbb{E}[\min\{1, (\|\hat{\mathbf{x}}_t\|^2 + \|\hat{\mathbf{u}}_t\|^2\}\mathbf{I}\{\|\hat{\mathbf{x}}_2\| = o_\star(1)\}]} \\ &\lesssim \sqrt{\mathbb{E}_{D_{\{Z=0\}}|n_0 \ge n/3}\mathbb{E}[\|\hat{\mathbf{x}}_t\|^{2(1-q)}]} \le \mathbb{E}_{D_{\{Z=0\}}|n_0 \ge n/3}\mathbb{E}[\|\hat{\mathbf{x}}_t\|^2]^{\frac{1-q}{2}} \lesssim \boldsymbol{\epsilon}_{n/3}^{1-q}.\end{aligned}$$

We can check that this resulting policy is in $\mathbb{A}_{\text{gen,smooth}}(L, M, \alpha, p)$ by noting all but the π_0 terms are deterministic and Lipschitz and smooth when $\mathcal{G}$, and that the π_0 component has linear (and thus Lipschitz and smooth) mean, and that, being a mixture policy with even component probabilities, it satisfies the anti-concentration property with $\alpha, p = \Omega(1)$ by the same argument as in Example 4.

(c) With alternating or history dependent policies, we altnerate between $\bar{\mathbf{A}}_i(\mathbf{I} - \mathbf{e}_1\mathbf{e}_1^\top)\mathbf{x}$, $i \in \{1, 2\}$. This kills the $\mathbf{e}_1$ direction with at most 3 steps, and as in part (c), we stabilize the system for the remaining part of the trajectory. This yields the same qualitative bound as in (a).

■

Appendix P. Proof of Upper Bounds, Theorem 5

Remark P.1 *The careful reader may notice that we assume that both the expert distribution is well-spread, but also that the expert policy $\pi^\star$ is deterministic. Both appear to be in tension because, e.g. $\pi^\star$ cannot undertake its own exploration. However, our result can be easily extended to more realistic settings with two modifications:*

1. *If we assume a static, stationary expert policy $\pi^\star(\mathbf{x})$, then we need only assume (up to possibly polynomial factors in horizon H) that the mixture measure over all time-steps h, defined as*

$$\mathbb{P}^{\text{mix}}_{\pi^\star, f, D} = \frac{1}{H}\sum_{h=1}^{H} \mathbb{P}_{\pi^\star, f, D}[\mathbf{x}^\star_h \in \cdot] \tag{P.1}$$

is well spread. This requires that only sufficient exploration can be provided in aggregate over timesteps h, and can therefore better take advantage in randomness from the initial conditions $\mathbf{x}_1 \sim D$.

2. *Our argument should be able to be generalized to settings where the learner is given observations of pairs $(\mathbf{x}, \mathbf{u})$, where $\mathbf{x}$ from a sufficiently "well-spread" distribution that covers the expert distribution in an appropriate sense, and $\mathbf{u} = \pi^\star(\mathbf{x})$ are perfect expert actions. This covers the* DART *algorithm due to Laskey et al. (2017), but we defer formal details to future work.*

Supporting Lemmas. We begin by proving the following supporting lemmas, which give bounds on various relevant properties of well-spread distributions.

The first lemma, Lemma P.1 uses the properties of well-spread distributions to upper bound the expectation of $f(x + \sigma w)$ in terms of $f(x)$ for bounded f, where x, w sampled from a well-spread distribution P and a unit-balled supported distribution D, respectively. This allows us to upper bound the effect of injecting noise on top of any well-spread distribution.

The second supporting lemma, Lemma P.2, shows that for second-order-smooth functions (i.e. bounded hessian), we can bound the expectation under P with σ-magnitude adversarial perturbations in terms of P perturbed some σ-scaled noise distribution D. The combination of this with Lemma P.1 yields a powerful result upper bounding the adversarial error.

We then combine these results with the adversarial bound of Proposition 3.1 of Pfrommer et al. (2022) (restated in a specialized form in Lemma P.4) to yield our final guarantees.

Lemma P.1 (Change of Measure for Well-Spread Distributions) *Let $\mathcal{D}$ be any distribution supported on the unit ball in $\mathbb{R}^d$. If P is (L, ϵ, σ_0)-well-spread (Definition 3.5), then for all $\sigma \le \sigma_0$, and all bounded, nonnegative, measurable functions $f : \mathbb{R}^d \to [0, B]$,*

$$\mathbb{E}_{\mathbf{x}\sim\mathcal{P}}\mathbb{E}_{\mathbf{w}\sim\mathcal{D}}[f(\mathbf{x} + \sigma\mathbf{w})] \le e^{L\sigma}\mathbb{E}_{\mathbf{x}\sim P}[f(\mathbf{x})] + \epsilon B. \tag{P.2}$$

Proof Let $\mathcal{K}_0 := \{\mathbf{x} : \mathrm{dist}(\mathbf{x}, \mathcal{K}^c) \le \sigma_0\}$. We have

$$\begin{aligned}
&\mathbb{E}_{\mathbf{x}\sim P}\mathbb{E}_{\mathbf{w}\sim\mathcal{D}}[f(\mathbf{x} + \sigma\mathbf{w})] \\
&\quad = \underbrace{\mathbb{E}\left[\mathbf{I}\{\mathbf{x} \in \mathcal{K}_0\}\mathbb{E}_{\mathbf{w}\sim\mathcal{D}}[f(\mathbf{x} + \sigma\mathbf{w})]\right]}_{T_1} + \underbrace{\mathbb{E}\left[\mathbf{I}\{\mathbf{x} \notin \mathcal{K}_0\}\mathbb{E}_{\mathbf{w}\sim\mathcal{D}}[f(\mathbf{x} + \sigma\mathbf{w})]\right]}_{T_2}
\end{aligned}$$

As $|f| \le B$, we have $T_2 \le B\,\mathbb{P}[\mathbf{x} \notin \mathcal{K}_0] \le B\epsilon$. Thus, we turn to upper bounding the first term. Note that if $\mathbf{x} \in \mathcal{K}_0 \subset \mathcal{K}$, then $\mathbf{x} + \sigma\mathbf{w} \in \mathcal{K}$, as $\|\sigma\mathbf{w}\| = \sigma\|\mathbf{w}\| \le \sigma \le \sigma_0$ (recall $\mathcal{D}$ is supported on the unit ball). Thus, the first term is equal to

$$\begin{aligned}
T_1 &= \mathbb{E}_{\mathbf{x}\sim P}\mathbb{E}_{\mathbf{w}\sim\mathcal{D}}\left[\mathbf{I}\{\{\mathbf{x}, \mathbf{x} + \sigma\mathbf{w}\} \subset \mathcal{K}\}f(\mathbf{x} + \sigma\mathbf{w})\right] \\
&= \mathbb{E}_{\mathbf{w}\sim\mathcal{D}}\left[\underbrace{\mathbb{E}_{\mathbf{x}\sim P}\left[\mathbf{I}\{\{\mathbf{x}, \mathbf{x} + \sigma\mathbf{w}\} \subset \mathcal{K}\}f(\mathbf{x} + \sigma\mathbf{w})\right]}_{:=T_1(\mathbf{w})}\right],
\end{aligned}$$

where we use that f is non-negative, measurable to apply Tornelli's theorem. Gathering our current progress,

$$\mathbb{E}_{\mathbf{x}\sim P}\mathbb{E}_{\mathbf{w}\sim\mathcal{D}}[f(\mathbf{x} + \sigma\mathbf{w})] \le \epsilon B + \mathbb{E}_{\mathbf{w}\sim\mathcal{D}}[T_1(\mathbf{w})] \tag{P.3}$$

Via a change of variables, we have that the quantity $T_1(\mathbf{w})$ above is equal to

$$\int_{\mathbf{x}\in\mathbb{R}^d} \mathbf{I}\{\{\mathbf{x},\mathbf{x}+\sigma\mathbf{w}\}\subset\mathcal{K}\}f(\mathbf{x}+\sigma\mathbf{w})p(\mathbf{x})\mathrm{d}\mathbf{x} = \int_{\mathbf{u}\in\mathbb{R}^d} \mathbf{I}\{\{\mathbf{u}-\sigma\mathbf{w},\mathbf{u}\}\subset\mathcal{K}\}f(\mathbf{u})p(\mathbf{u}-\sigma\mathbf{w})\mathrm{d}\mathbf{u}$$

Now notice that (i) $\mathcal{K}$ is convex, (ii) $\{\mathbf{u}-\sigma\mathbf{w},\mathbf{u}\}\subset\mathcal{K}$ and (iii) $\log p(\cdot)$ is L-Lipschitz on $\mathcal{K}$. This gives that for any $\mathbf{u},\mathbf{w}$ for which $\mathbf{I}\{\{\mathbf{u}-\sigma\mathbf{w},\mathbf{u}\}\subset\mathcal{K}\}=1$, we have

$$|\log p(\mathbf{u}-\sigma\mathbf{w}) - \log p(\mathbf{u})| \le L\sigma\|\mathbf{w}\| \le L\sigma, \tag{P.4}$$

and thus

$$p(\mathbf{u}-\sigma\mathbf{w}) \le e^{L\sigma}p(\mathbf{u}). \tag{P.5}$$

It follows then that we can bound

$$T_1(\mathbf{w}) = \int_{\mathbf{u}\in\mathbb{R}^d} \mathbf{I}\{\{\mathbf{u}-\sigma\mathbf{w},\mathbf{u}\}\subset\mathcal{K}\}f(\mathbf{u})p(\mathbf{u}-\sigma\mathbf{w})\mathrm{d}\mathbf{u} \tag{P.6}$$

$$\le e^{L\sigma}\int_{\mathbf{u}\in\mathbb{R}^d} \mathbf{I}\{\{\mathbf{u}-\sigma\mathbf{w},\mathbf{u}\}\subset\mathcal{K}\}f(\mathbf{u})p(\mathbf{u})\mathrm{d}\mathbf{u} \tag{P.7}$$

$$\le e^{L\sigma}\mathbb{E}_{\mathbf{u}\sim P}[f(\mathbf{u})] \tag{P.8}$$

Since the above bound holds for all $\mathbf{w}:\|\mathbf{w}\|\le 1$, combining the above display with (P.3) concludes the demonstration. ■

Lemma P.2 (Smooth Functions) *Suppose $\hat{\pi},\pi^\star:\mathbb{R}^d\to\mathbb{R}^m$ are β-second-order-smooth. Then, for $f(\mathbf{x}) := \|\hat{\pi}(\mathbf{x})-\pi^\star(\mathbf{x})\|^2$, zero-mean distribution $\mathcal{D}$ supported on the unit ball, and with $\nu = 1/\lambda_{\min}(\mathbb{E}_{\mathbf{w}\sim\mathcal{D}}[\mathbf{w}\mathbf{w}^\top])$, we have*

$$\sup_{\mathbf{w}\in\mathcal{B}_d}\|(\hat{\pi}-\pi^\star)(\mathbf{x}+\sigma\mathbf{w})\|^2 \le 8\nu\mathbb{E}_{\mathbf{w}\sim\mathcal{D}}\|(\hat{\pi}-\pi^\star)(\mathbf{x}+\sigma\mathbf{w})\|^2 + 16\nu\beta^2\sigma^4. \tag{P.9}$$

Consequently, for any P which is (L,ϵ,σ_0)-well-spread, and if $\max_{\mathbf{x}}\|\hat{\pi}(\mathbf{x})-\pi^\star(\mathbf{x})\|^2\le B$, then for all $\sigma\le\min\{\sigma_0,1/L\}$,

$$\mathbb{E}_{\mathbf{x}\sim P}\left[\sup_{\mathbf{w}\in\mathcal{B}_d}\|(\hat{\pi}-\pi^\star)(\mathbf{x}+\sigma\mathbf{w})\|^2\right] \le 8\nu\mathbb{E}_{\mathbf{x}\sim P}[\|(\hat{\pi}-\pi^\star)(\mathbf{x})\|^2] + 8\nu B\epsilon + 16\nu\beta^2\sigma^4.$$

Specializing to the intermediate distribution $\mathcal{D}=S^{d-1}$ yields $\nu=d$ and the relation:

$$\mathbb{E}_{\mathbf{x}\sim P}\left[\sup_{\mathbf{w}\in\mathcal{B}_d}\|(\hat{\pi}-\pi^\star)(\mathbf{x}+\sigma\mathbf{w})\|^2\right] \le 8d\mathbb{E}_{\mathbf{x}\sim P}[\|(\hat{\pi}-\pi^\star)(\mathbf{x})\|^2] + 8dB\epsilon + 16d\beta^2\sigma^4. \tag{P.10}$$

Proof To simplify matters, it suffices to study a function $\pi(\mathbf{x}) = \hat{\pi}-\pi^\star$ which is 2β-second order smooth. We shall also prove the more general statement for arbitrary $\mathcal{D}$. Define $\nu =$

$1/\lambda_{\min}(\mathbb{E}_{\mathbf{w}\sim\mathcal{D}}[\mathbf{w}\mathbf{w}^\top])$; note that in the case where $\mathcal{D}$ is uniform on the sphere, $\nu = d$, recovering the desired bound. We have

$$\begin{aligned}
\sup_{\mathbf{w}\in\mathcal{B}_d} \|\pi(\mathbf{x}+\sigma\mathbf{w})\|^2 &\le \sup_{\mathbf{w}\in\mathcal{B}_d} 2\|\pi(\mathbf{x}+\sigma\mathbf{w}) - \pi(\mathbf{x}) - \sigma\nabla\pi(\mathbf{x})\cdot\mathbf{w}\|^2 + 2\|\pi(\mathbf{x}) - \sigma\nabla\pi(\mathbf{x})\cdot\mathbf{w}\|^2 \\
&\le 2\|\beta\sigma^2\mathbf{w}\|^2 + 2\sup_{\mathbf{w}\in\mathcal{B}_d}\|\pi(\mathbf{x}) - \sigma\nabla\pi(\mathbf{x})\cdot\mathbf{w}\|^2 \\
&\le 2\beta^2\sigma^4 + 4\|\pi(\mathbf{x})\|^2 + 4\sup_{\mathbf{w}\in\mathcal{B}_d}\|\sigma\nabla\pi(\mathbf{x})\cdot\mathbf{w}\|^2 \\
&= 2\beta^2\sigma^4 + 4\|\pi(\mathbf{x})\|^2 + 4\sigma^2\|\nabla\pi(\mathbf{x})\|_{\mathrm{op}} \qquad \text{(P.11)}
\end{aligned}$$

On the other hand, using the elementary inequality $\|\mathbf{x}+\mathbf{x}'\|^2 \ge \frac{1}{2}\|\mathbf{x}\|^2 - \|\mathbf{x}'\|^2$, we have

$$\begin{aligned}
\mathbb{E}_{\mathbf{w}\sim\mathcal{D}}\|\pi(\mathbf{x}+\sigma\mathbf{w})\|^2 \ge \frac{1}{2}&\mathbb{E}_{\mathbf{w}\sim\mathcal{D}}\|\pi(\mathbf{x}) - \sigma\nabla\pi(\mathbf{x})\cdot\mathbf{w}\|^2 \\
&- \mathbb{E}_{\mathbf{w}\sim\mathcal{D}}\|\pi(\mathbf{x}+\sigma\mathbf{w}) - \pi(\mathbf{x}) - \sigma\nabla\pi(\mathbf{x})\cdot\mathbf{w}\|^2
\end{aligned}$$

Using the same smoothness argument as above, the second term on the right hand side contributes at most $(\frac{1}{2}\cdot 2\beta\sigma^2)^2 = \beta^2\sigma^4$. Moreover, using that $\mathbb{E}[\mathbf{w}] = 0$ and $\mathbb{E}[\mathbf{w}\mathbf{w}^\top] = \frac{1}{\nu}$ by definition, we have

$$\mathbb{E}_{\mathbf{w}\sim\mathcal{D}}\|\pi(\mathbf{x}) - \sigma\nabla\pi(\mathbf{x})\cdot\mathbf{w}\|^2 = \|\pi(\mathbf{x})\|^2 + \frac{\sigma^2}{\nu}\mathrm{tr}(\nabla\pi(\mathbf{x})) \ge \frac{1}{\nu}(\|\pi(\mathbf{x})\|^2 + \sigma^2\|\nabla\pi(\mathbf{x})\|_{\mathrm{op}}).$$

Hence, we have,

$$\begin{aligned}
\mathbb{E}_{\mathbf{w}\sim\mathcal{D}}\|\pi(\mathbf{x}+\sigma\mathbf{w})\|^2 &\ge \frac{1}{2\nu}(\|\pi(\mathbf{x})\|^2 + \sigma^2\|\nabla\pi(\mathbf{x})\|_{\mathrm{op}}) - \beta^2\sigma^4 \\
&= \frac{1}{8\nu}(4\|\pi(\mathbf{x})\|^2 + 4\sigma^2\|\nabla\pi(\mathbf{x})\|_{\mathrm{op}}) - \beta^2\sigma^4 \\
&\ge \frac{1}{8\nu}\left(\sup_{\mathbf{w}\in\mathcal{B}_d}\|\pi(\mathbf{x}+\sigma\mathbf{w})\|^2 - 2\beta^2\sigma^4\right) - \beta^2\sigma^4 \qquad \text{(by (P.11))} \\
&\ge \frac{1}{8\nu}\left(\sup_{\mathbf{w}\in\mathcal{B}_d}\|\pi(\mathbf{x}+\sigma\mathbf{w})\|^2\right) - 2\beta^2\sigma^4
\end{aligned}$$

Rearranging,

$$\sup_{\mathbf{w}\in\mathcal{B}_d}\|\pi(\mathbf{x}+\sigma\mathbf{w})\|^2 \le 8\nu\mathbb{E}_{\mathbf{w}\sim\mathcal{D}}\|\pi(\mathbf{x}+\sigma\mathbf{w})\|^2 + 16\nu\beta^2\sigma^4.$$

■

For compatibility with Lemma P.4, we use Markov's and rearrange the above bound to upper bound the probability of the exceeding a given threshold value.

Lemma P.3 *Suppose that $\hat{\pi}, \pi^\star$ are β-second-order-smooth, B-bounded, and P is (L, ϵ, σ_0)-well-spread. Let $\kappa := \sqrt{\mathbb{E}_{x\sim P}[\|\hat{\pi}(\mathbf{x}) - \pi^\star(\mathbf{x})\|^2]}$, $\kappa_1 := \max\{\kappa, \epsilon^2\}, \kappa_2 := \max\{\kappa, \sqrt{\epsilon}\}$. Provided $\kappa_1 \le \rho_0^2, 1/L^2$, for any $K \ge 0$,*

$$\mathbb{P}_{x\sim P}[\sup_{\mathbf{w}\in\mathcal{B}_d}\|(\hat{\pi} - \pi^\star)(x + \sqrt{\kappa_1}\mathbf{w})\| \ge K\sqrt{\kappa_1}] \le \frac{d(8 + 16B^2 + 16\beta^2)}{K^2}(\kappa + \epsilon).$$

Proof Let $\sigma := \sqrt{\kappa_1}$. Note that $\epsilon \le \sigma \le \min\{\rho_0, 1/L\}$. Since $\hat{\pi}, \pi^\star$ are β-second-order-smooth, B-bounded and P is (L, ϵ, σ_0)-well-spread with $\sigma < 1/L, \sigma_0$,

$$\begin{aligned}
\mathbb{E}_{\mathbf{x}\sim P}\left[\sup_{\mathbf{w}\in\mathcal{B}_d} \|(\hat{\pi}-\pi^\star)(\mathbf{x}+\sigma\mathbf{w})\|^2\right] &\le 8d\kappa^2 + 16B^2 d\epsilon + 16\beta^2\sigma^4 d \\
&\le 8d\kappa_2^2 + 16B^2 d\kappa_2^2 + 16\beta^2 d\kappa_2^2 \\
&\le d(8 + 16B^2 + 16\beta^2)\kappa_2^2.
\end{aligned}$$

By Markov's inequality and using that $\frac{k_2^2}{k_1} \le k_2 \le (k + \sqrt{\epsilon})$,

$$\mathbb{P}_{x\sim P}[\sup_{\mathbf{w}\in\mathcal{B}_d} \|(\hat{\pi}-\pi^\star)(x+\sqrt{k_1}\mathbf{w})\| \ge K\sqrt{\kappa_1}] \le \frac{d(24+16B^2+16\beta^2)}{K^2}(\kappa+\sqrt{\epsilon}).$$

■

Lemma P.4 (TaSIL, Pfrommer et al. (2022)) *Let $(\pi^\star, f)$ be deterministic and $\pi^\star$ be (C,ρ)-E-IISS. For any deterministic policy $\hat{\pi}$ and initial state $\mathbf{x}_1$, let $\hat{\mathbf{x}}_1 = \mathbf{x}_1^\star := \mathbf{x}_1$ and $\mathbf{x}_{t+1}^\star := f_{\mathrm{cl}}^{\pi^\star}(\mathbf{x}_t^\star), \hat{\mathbf{x}}_{t+1} := f_{\mathrm{cl}}^{\pi^\star}(\hat{\mathbf{x}}_t)\,\forall t \ge 2$. Then for any $\epsilon > 0, t > 0$,*

$$\max_{1\le k\le t-1} \sup_{\mathbf{w}\in\mathcal{B}_d} \|(\hat{\pi}-\pi^\star)(\mathbf{x}_k^\star + \epsilon\mathbf{w})\| \le \frac{1-\rho}{C}\epsilon \implies \max_{1\le k\le t} \|\hat{\mathbf{x}}_k - \mathbf{x}_k^\star\| \le \epsilon$$

Proof This is a simple proof using induction. The base case $t = 1$ is true by construction as $\mathbf{x}_1^\star = \hat{\mathbf{x}}_1$. For $t \ge 2$, we assume the statement holds for $t-1$. Then, it follows by the induction hypothesis that

$$\begin{aligned}
&\max_{1\le k\le t-1} \sup_{\mathbf{w}\in\mathcal{B}_d} \|(\hat{\pi}-\pi^\star)(\mathbf{x}_k^\star + \mathbf{w})\| \le \frac{1-\rho}{C}\epsilon \\
&\implies \max_{1\le k\le t-1} \|\hat{\mathbf{x}}_k - \mathbf{x}_k^\star\| \le \epsilon \text{ (from induction hypothesis)} \\
&\implies \max_{1\le k\le t-1} \|\hat{\pi}(\hat{\mathbf{x}}_k) - \pi^\star(\hat{\mathbf{x}}_k)\| \le \max_{1\le k\le t-1} \sup_{\|\delta\|\le\epsilon} \|\hat{\pi}(\mathbf{x}_k^\star+\delta) - \pi^\star(\mathbf{x}_k^\star+\delta)\|.
\end{aligned}$$

We now recall the following property of (C,ρ) incrementally input-to-state-stabilizng policies:

$$\|\hat{\mathbf{x}}_t - \mathbf{x}_t^\star\| \le C\sum_{s=1}^{s} \rho^{t-s}\|\hat{\pi}(\hat{\mathbf{x}}_s) - \pi^\star(\hat{\mathbf{x}}_s)\| \le \frac{C}{1-\rho}\left(\max_{1\le k\le t-1} \|\hat{\pi}(\hat{\mathbf{x}}_s) - \pi^\star(\hat{\mathbf{x}}_s)\|\right).$$

This yields the desired bound,

$$\begin{aligned}
\|\hat{\mathbf{x}}_t - \mathbf{x}_t^\star\| &\le \frac{C}{1-\rho}\left(\max_{1\le k\le t-1} \|\hat{\pi}(\hat{\mathbf{x}}_s) - \pi^\star(\hat{\mathbf{x}}_s)\|\right) \\
&\le \frac{C}{1-\rho}\left(\max_{1\le k\le t-1} \sup_{\|\delta\|\le\epsilon} \|\hat{\pi}(\mathbf{x}_k^\star+\delta) - \pi^\star(\mathbf{x}_k^\star+\delta)\|\right) \\
&\le \epsilon.
\end{aligned}$$

■

Main smoothness result: The main result of Theorem 5 follows by a straightforward combination of Lemma P.3, combined with Lemma P.4. At a high level, Lemma P.4 provides performance bounds for the learned policy given a bound on the adversarial error, whie Lemma P.3 gives precisely a bound on the probability of a small adversarial error occuring for well-spread distributions.

Theorem 5 (Smooth Training Distribution) *Consider any (d, m)-BC instance $(\mathcal{P}, D)$. Provided for any $(\pi^\star, f) \in \mathcal{P}$, $h \in [H]$, the distribution $\mathbb{P}_{\pi,f,D}$ is (L, ϵ, σ_0)-well-spread (Definition 3.5) for $h > 1$ and $\pi^\star, \hat{\pi}$ are deterministic, β-second-order-smooth, L_π-Lipschitz, and B-bounded, and $\pi^\star$ is (C, ρ) incrementally input-to-state stablizing (Definition 2.1), the following holds. Then, provided that $\mathbf{R}_{\mathrm{expert},L_2}(\hat{\pi}; \pi^\star, f, D, H) \le \min\{\rho_0, 1/L\}$,*

$$\mathbf{R}_{\mathrm{eval}}(\hat{\pi}; \pi^\star, f, D, H) \le cHd\frac{C^2}{(1-\rho)^2}\left[\mathbf{R}_{\mathrm{expert},L_2}(\hat{\pi}; \pi^\star, f, D, H) + \sqrt{\epsilon}\right].$$

where $c := 16d(1 + 2B^2 + 2\beta^2)$.

Proof Let $\mathbf{x}_1^\star = \hat{\mathbf{x}}_1 \sim D, \mathbf{x}_{t+1}^\star = f_{\mathrm{cl}}^{\pi^\star}(\mathbf{x}_t^\star), \hat{\mathbf{x}}_{t+1} = f_{\mathrm{cl}}^{\hat{\pi}}(\hat{\mathbf{x}}_t)$ and define $\kappa := \mathbf{R}_{\mathrm{expert},L_2}(\hat{\pi}; \pi^\star, f, D, H)$, $\kappa_1 := \max\{\kappa, \epsilon^2\}, \kappa_2 := \max\{\kappa, \epsilon\}$. We note that since cost is 1-Lipschitz and $\pi^\star, \hat{\pi}$ are L_π-Lipschitz, we can rewrite,

$$\begin{aligned}
\mathbf{R}_{\mathrm{eval}}(\hat{\pi}; \pi^\star, f, D, H) &\le \mathbb{E}_{\hat{\pi},\pi^\star,f,D}\left[\sum_{h=1}^{H}(\min\{\|\mathbf{u}_h^\star - \hat{\mathbf{u}}_h\| + \|\mathbf{x}_h^\star - \hat{\mathbf{x}}_h\|, 1\})\right] \\
&\le (1 + 2L_\pi)\mathbb{E}_{\hat{\pi},\pi^\star,f,D}\left[\sum_{h=1}^{H}\min\{\|\mathbf{x}_h - \mathbf{x}_h^\star\|, 1\}\right] \\
&\le (1 + 2L_\pi)H\mathbb{E}_{\hat{\pi},\pi^\star,f,D}\left[\max_{1\le h\le H}\min\{\|\mathbf{x}_h - \mathbf{x}_h^\star\|, 1\}\right] \\
&= (1 + 2L_\pi)H\int_0^1\mathbb{P}_{\hat{\pi},\pi^\star,f,D}\left[\max_{1\le h\le H}\|\hat{\mathbf{x}}_h - \mathbf{x}_h^\star\| \ge \eta\right]d\eta \\
&\le (1 + 2L_\pi)H\Bigg(\int_0^{\sqrt{\kappa_1}}\mathbb{P}_{\hat{\pi},\pi^\star,f,D}\left[\max_{1\le h\le H}\|\hat{\mathbf{x}}_h - \mathbf{x}_h^\star\| \ge \eta\right]d\eta \\
&\qquad + \mathbb{P}_{\hat{\pi},\pi^\star,f,D}\left[\max_{1\le h\le H}\|\hat{\mathbf{x}}_h - \mathbf{x}_h^\star\| \ge \sqrt{\kappa_1}\right]\Bigg).
\end{aligned}$$

We use Lemma P.4 and Lemma P.3 to bound the tail probability:

$$\begin{aligned}
\mathbb{P}_{\hat{\pi},\pi^\star,f,D}\left[\max_{1\le h\le H}\|\hat{\mathbf{x}}_h - \mathbf{x}_h^\star\| \ge \sqrt{\kappa_1}\right] &\le \mathbb{P}_{\pi^\star,f,D}\left[\max_{1\le h\le H}\sup_{\mathbf{w}\in\mathcal{B}_d}\|(\hat{\pi} - \pi^\star)(\mathbf{x}_h + \sqrt{\kappa}_1\mathbf{w})\| \ge \frac{1-\rho}{C}\sqrt{\kappa}_1\right] \\
&\le \frac{C^2}{(1-\rho)^2}d(8 + 16B^2 + 16\beta^2)(\kappa + \sqrt{\epsilon}).
\end{aligned}$$

We can similarly bound the probability over the bulk,

$$\begin{aligned}
\int_0^{\sqrt{\kappa_1}} \mathbb{P}_{\hat{\pi},\pi^\star,f,D}\left[\max_{1\le h\le H}\|\hat{\mathbf{x}}_h-\mathbf{x}_h^\star\|\ge\eta\right]d\eta &\le \int_0^{\sqrt{\kappa_1}} \mathbb{P}_{\pi^\star,f,D}\left[\max_{1\le h\le H}\sup_{\mathbf{w}\in\mathcal{B}_d}\|(\hat{\pi}-\pi^\star)(\mathbf{x}_x+\eta\mathbf{w})\|\ge\frac{1-\rho}{C}\eta\right]d\eta\\
&\le \int_0^{\sqrt{\kappa_1}} \mathbb{P}_{\pi^\star,f,D}\left[\max_{1\le h\le H}\sup_{\mathbf{w}\in\mathcal{B}_d}\|(\hat{\pi}-\pi^\star)(\mathbf{x}_h+\sqrt{\kappa_1}\mathbf{w})\|\ge\frac{1-\rho}{C}\eta\right]d\eta\\
&\le \frac{C}{1-\rho}\mathbb{E}\left[\max_{1\le h\le H}\sup_{w\in\mathcal{B}_d}\|(\hat{\pi}-\pi^\star)(\mathbf{x}_h+\sqrt{\kappa_1}\mathbf{w})\|\right]\\
&\le \frac{C}{1-\rho}[4\sqrt{d}\kappa+4B\sqrt{d}\sqrt{\epsilon}+4\sqrt{d\beta}\kappa].
\end{aligned}$$

Combining these bounds,

$$\begin{aligned}
\mathsf{R}_{\mathrm{eval}}(\hat{\pi};\pi^\star,f,D,H) &\le 16Hd\frac{C^2}{(1-\rho)^2}(1+2L_\pi)(1+2B^2+2\beta^2)\left[\mathsf{R}_{\mathrm{expert},L_2}(\hat{\pi};\pi^\star,f,D,H)+\sqrt{\epsilon}\right]\\
&= cHd\frac{C^2}{(1-\rho)^2}\left[\mathsf{R}_{\mathrm{expert},L_2}(\hat{\pi};\pi^\star,f,D,H)+\sqrt{\epsilon}\right].
\end{aligned}$$

■