WHAT DATA IS NEEDED FOR OFFLINE RL IN CONTINUOUS CONTROL? A GEOMETRIC PERSPECTIVE

Anonymous authors
Paper under double-blind review

ABSTRACT

Learning from high-quality data forms the foundation of most modern AI applications, ranging from large language models to generalist robotic systems. However, collecting sufficient expert data to satisfy the increasing demands of robust performance is infeasible in general. In this paper, we study the question: what properties must offline data exhibit to enable near-optimal behavior in continuous control? We present a rigorous theoretical analysis of the offline RL problem, and outline the inability of common notions of coverage to discriminate between benign and intractable problem instances. To this end, we propose new notions of coverage which capture the inherent geometric nature of continuous control. Via novel analyses leveraging stability-theoretic tools and these geometric notions of coverage, we identify regimes where offline RL is tractable versus difficult, and we demonstrate how popular model-based algorithmic and data practices can achieve polynomial regret bounds under these regimes. We perform experiments on manipulation and locomotion tasks and find strong alignment with our theoretical predictions, as well as empirical evidence that our recommended practices benefit a wider range of model-based and model-free frameworks.

1 INTRODUCTION

Learning from demonstration data is the bedrock of many modern AI applications, ranging from large language models to emerging robotic frontier models. However, collecting sufficient quantities of expert-quality data alone to elicit maximal model performance is infeasible in general. In digital domains like language, models are subsequently post-trained on their own generations to reduce additional errors that arise from supervised pre-training. While similar efforts are underway in robotics, deploying suboptimal robotic policies in the physical world is often slow, impractical, costly, or dangerous.

As a consequence, recent attention has transitioned to understanding how to learn from large-scale suboptimal robotic data (Mandlekar et al., 2020; Zolna et al., 2020). This data may contain trajectories of robotic successes and mistakes. Moreover, even professional human demonstration data may show behaviors that are suboptimal relative to what a robot can do, e.g., in terms of motion speed, smoothness, and or throughput. Here focus on offline reinforcement learning (offline RL), a family of methods for learning from suboptimal data that optimize policy performance without recourse to online interaction with the environment. In this work, we ask the basic question:

> *In continuous control tasks, such as those encountered in robot learning, what mathematical properties must suboptimal data exhibit so that near-optimal behavior can be learned from them in a purely offline manner?*

The problem of coverage in continuous control. The central challenge in offline RL is that of **coverage**: do the suboptimal data contain sufficient information such that optimal behavior can be extracted from them? In traditional analyses, whose assumptions we prove are applicable in discrete settings, like language, but *not* in continuous control domains like robotics, coverage is when a technical quantity called *concentrability* is bounded. This term measures the ratio of the likelihood of dynamical states and actions that would be visited under expert-quality trajectories, divided by the likelihood of the same under the training data, and must remain appropriately bounded. Intuitively, bounded concentrability can be enforced as long as some fraction of your suboptimal data was generated by an expert. In particular, offline RL is capable of "data filtering" optimal behaviors from suboptimal ones. In fact, offline RL is capable of the stronger property of "stitching together" suboptimal behavior (Hepburn & Montana, 2022).

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

In this work, we show that the aforementioned conventional measures of concentrability *do not suffice* in the continuous world. In particular, without direct expert demonstrations in the data, concentrability is almost always *vacuous* in continuous settings. Worse, even mixing in optimal demonstrations is no longer enough to ensure that offline RL can learn near-optimal policies. Therefore, we set out to understand what alternative notions of coverage are required in the physical world.

Conceptual Contribution: Geometric Coverage. To this end, we introduce a notion of geometric coverage, which revises concentrability based on the geometric nature of continuous control. In particular, while concentrability solely measures how *likely* optimal actions are under the data distribution, geometric coverage concerns how *close* reachable directions in the dataset distribution are to salient ones for offline learning—which may not be from an optimal policy, or any directly considered under concentrability.

Offline RL is Feasible under Geometric Coverage. In this work, we analyze the so-called pessimistic, model-based RL framework widely studied in the theoretical literature. Here, *model-based* means that the algorithm explicitly estimates environment dynamics, in contrast to *model-free* algorithms, more popular in practice, that solely rely on a *critic function* that estimates future returns. As with the overwhelming majority of the theoretical offline RL literature, we analyze the model-based setting due to known mathematical obstacles surrounding model-free offline RL. Nevertheless, our experiments corroborate that the same requisite data conditions do enable effective learning from model-free methods as well, supporting the broader validity of our findings.

1.1 Contributions

We provide a **rigorous theoretical analysis** that characterizes the *difficult* and *feasible* regimes of the continuous state-action space model-based offline RL problem, providing a clear delineation for when the problem can be provably hard, and what is needed for pessimistic algorithms to achieve universal regret upperbounds.

- Offline RL frameworks developed for the discrete setting are unsuited for the continuous space problem, and often produce vacuous guarantees. In addition, common notions of *coverage* fail to characterize when an offline RL instance is benign or hard.
- Even when using continuous standard practices, we show that settings exist where low error on an offline dataset is insufficient to rule out models which can produce exponentially bad policy evaluations, making policy selection intractable. This provides a *general performance lower bound* for offline RL in continuous control.

This is the offline RL counterpart to the compounding errors problem explored in Simchowitz et al. (2025) and Zhang et al. (2025), and the continuous space consequence of the model exploitation problem explored in other model-based work. Next, we identify complementary conditions under which offline RL is feasible, characterized by the *stability* properties of the instance.

- When the underlying system is *stable* (to be concretely defined below), delaying how often your policy re-plans, e.g. via the popular practice of **action-chunking**, suffices to prevent model feedback from destabilizing the system and causing bad evaluations.
- When the underlying system is *unstable*, we identify new notions of **geometric coverage** that delineate the difficulty of offline RL in continuous spaces. When satisfied, error on the offline dataset can reliably rule out bad models and allow algorithms to select good policies.

Finally, we conduct experiments for a variety of model-free and model-based algorithms in both robotic manipulation and locomotion tasks, and demonstrate that our insights lead to practical benefits even under settings not explicitly analyzed in our theoretical results.

2 Preliminaries

Whereas traditional analyses of offline RL consider MDPs with arbitrary stochastic transitions, our analysis exposes how the structure of dynamics in continuous control alters the feasibility of learning. We formulate our problem by considering a discrete-time finite-horizon dynamical system with continuous state and action spaces $\mathcal{S} = \mathbb{R}^{d_x}$ and $\mathcal{A} = \mathbb{R}^{d_u}$ respectively, following conventions from prior work on imitation learning in continuous control systems (Pfrommer et al., 2022; Simchowitz et al., 2025; Zhang et al., 2025). Following these works, for simplicity we consider learning deterministic control policies $\pi(s_{1:h}, a_{1:h-1}, h) = a_h$. If we can simply express $a_h = \pi(s_h)$, we say that π is *Markovian and time-invariant*.

From an initial state s_1 sampled from a fixed initial state distribution d_1, we define the **true dynamics** $f^\star$ to be the function mapping the current state and action to the next state $s_{h+1} = f^\star(s_h, a_h)$.

We will refer to s_h^{π,f,s_1} and a_h^{π,f,s_1} as denoting the states and actions obtained from rolling out the policy π on dynamics or model f from some initial state $s_1 \sim d_1$; we omit the initial state $s_h^{\pi,f}$ and $a_h^{\pi,f}$ when it is not of relevance. We will additionally denote the distribution induced by rolling out a policy π on some dynamics f from initial state distribution d_1 as (π, f, d_1), where $(\pi, f, d_1)_h$ refers to the marginal distribution of $(s_h^{\pi,f}, a_h^{\pi,f})$ induced at timestep $h \geq 1$.

Offline RL setting. In offline RL we assume access to a dataset distribution $\mu_h \triangleq (\pi_b, f^\star, d_1)$ representing the joint state-action distribution induced by some suboptimal *stochastic* behavior policy π_b, and a known reward function $\mathrm{R} : \mathcal{S} \times \mathcal{A} \to [0, 1]$. Given this data and a policy class Π, our goal is to find a learned policy $\hat{\pi}$ which will accumulate high reward when deployed on the true dynamics $f^\star$; we define this true cumulative reward $J_H(\hat{\pi})$ as follows: $J_H(\pi) \triangleq \mathop{\mathbb{E}}_{s_1 \sim d_1} \sum_{h=1}^{H} \left[\mathrm{R}(s_h^{\pi,f^\star,s_1}, a_h^{\pi,f^\star,s_1}) \right]$. We focus on selecting a policy that performs closely to the **best Markovian comparator policy** $\pi^\star$ in our policy class Π, which we refer to as the *expert comparator* or simply the *expert* for short. As the expert rollout distribution appears frequently in our analysis, we denote $\mathbb{P}_h^\star \triangleq (\pi^\star, f^\star, d_1)_h$. In offline RL, we are unable to query $f^\star$ during training, so we instead optimize over an estimated future reward $Q_H^{\pi,f}$, called the Q-function, as a proxy for $J_H(\pi)$: $Q_H^{\pi,f} \triangleq \mathop{\mathbb{E}}_{s_1 \sim d_1} \sum_{h=1}^{H} \left[\mathrm{R}(s_h^{\pi,f,s_1}, a_h^{\pi,f,s_1}) \right]$. Critically, accurate assessment of a policy's optimality relies on this estimate being close to its true reward $J_H(\pi)$.

Pessimistic algorithms in Offline RL. A common abstraction for offline RL is that of *pessimism*; see e.g., Rashidinejad et al. (2021); Xie et al. (2021); Uehara & Sun (2021). In such works, optimal policies are selected by picking the policy with the *best* reward achievable by the *worst-case* candidate satisfying low loss on the dataset, e.g., in the model-based setting, this gives:

$$\hat{\pi} = \arg\max_{\pi \in \Pi} \min_{f \in \mathcal{F}(\epsilon)} Q_H^{\pi,f}, \tag{1}$$

where $\mathcal{F}(\epsilon)$ is the subset of dynamics f achieving low loss $\mathbb{E}_\mu[L(f, f^\star)] \leq \epsilon$ for some loss L. Pessimistic algorithms benefit by reducing RL to *expert evaluation error*: $J_H(\pi^\star) - J_H(\hat{\pi}) \leq J_H(\pi^\star) - \min_{f \in \mathcal{F}(\epsilon)} Q_H^{\pi^\star,f}$ As a result, analyzing pessimistic algorithms reduces to estimating the expert's behavior on the true versus the candidate dynamics.

2.1 Traditional Notions of Coverage are Insufficient in Continuous Control

Standard approaches characterize the feasibility of offline RL and the utility of the offline data via the **concentrability coefficient**. Concentrability concerns a Radon-Nikodym derivative (i.e. the *density ratio*) between the dataset distribution μ_h the rollout distribution of some policy π: $\mathscr{C}(\pi; \mu) \triangleq \max_h \|d(\pi, f^\star, d_1)_h / d\mu_h\|_\infty$. Broadly speaking, concentrability controls the likelihood that quantities derived from μ, such as $Q_H^{\pi,f}$, are accurate on the true dynamics $(\pi, f^\star, d_1)$. Informally, this measures the **coverage** of trajectories induced by π under data sampled by μ.

Existing analyses of online RL choose one of a few assumptions, each of which we argue is not suitable for continuous control: **(1)** *all-policy concentrability*, namely that $\sup_\pi \mathscr{C}(\pi; \mu) \leq C_{\mathrm{cov}}$ is uniformly bounded, and error rates grow linearly in C_{cov}. **(2)** Only *single-policy concentrability*, where $C(\pi^\star, \mu) \leq C_{\mathrm{cov}}$. **(3)** *Bellman completeness*, a subtle technical conditioning roughly meaning that for any *changing* sequence of dynamics $f_1, \ldots, f_H \in \mathcal{F}$ and $\pi \in \Pi$, the set of Q functions associated with $f_{1:H}$ can be estimated with reasonable statistical efficiency.

All-policy concentrability is ill-suited in control since it may be unbounded for deterministic policies π even if under purely expert data $\mu_h = \mathbb{P}_h^\star$, and in general density ratios grow exponentially in state-space dimension. In Section B.1, we provide an intuitive explanation of why Bellman completeness is also too much to ask for, as switching between dynamical systems can cause wild oscillations in Q-values. As such, we shall focus on analyzing the single-policy concentrability assumption.

2.2 Jacobian Linearizations and Dynamical Stability

In this section that follows, we introduce a control-theoretic perspective of offline RL to explain how offline RL can suffer from errors exponential in the problem horizon H, motivating alternative assumptions which make offline RL tractable. Throughout, we make the following assumption.

Assumption 2.1 (Regularity conditions)**.** We assume that the true dynamics $f^\star$, every $f \in \mathcal{F}$, and every $\pi \in \Pi$ are Lipschitz and smooth; i.e., their first and second derivatives are globally bounded. Finally, we assume that the reward function R is Lipschitz.

These assumptions enable us to study the dynamics by its first-order Taylor approximation. Let $\tau = (s_1, a_1, \ldots, s_H, a_H)$ denote a given trajectory. Given a trajectory τ and (smooth, deterministic) policy π, we define the **Jacobian Linearization** as the matrices $(A_{h|\tau}, B_{h|\tau}, K_{h|\tau})$

$$A_{h|\tau} = \nabla_s f(s_h, a_h), \quad B_{h|\tau} = \nabla_a f(s_h j.a_h), \quad K_{h|\tau} := \nabla_s \pi_h(s_j) \tag{2}$$

These matrices imply that, if we create a new trajectory $\tilde{\tau}$ by perturbing the actions to $\tilde{a}_h = a_h + \delta a_h$ for small δa_h, then to first order the differences δs_h between the perturbed state $\tilde{s}_h$ and s_h satisfy $\delta s_{h+1} \triangleq (\tilde{s}_{h+1} - s_{h+1}) \approx (A_{h|\tau} + B_{h|\tau} K_{h|\tau}) \delta s_{h+1} + B_{h|\tau} \delta a_h$. Unfolding the recursion, we see that to first order in δa, $\delta s_{h+1} \approx \sum_{h'=1}^{h} A^{\mathrm{cl}}_{h'+1:h|\tau} B_{h'} \delta a_{h'}$, $A^{\mathrm{cl}}_{k:h} \triangleq \prod_{j=k}^{h-1} (A_{j|\tau} + B_{j|\tau} K_{j|\tau})$. We note both equations above are exact when π, f are linear. These equations reveal that, to first order, sensitivity of the dynamics is controlled by the matrices $A^{\mathrm{cl}}_{k:h}$, which we call the **closed-loop** transfer matrices. If these matrix products grow exponentially with the number of terms in the product, we expect growing sensitivity to error with problem horizon, including harder learning. In the alternative, we call the system *stable*, meaning errors are not amplified (Angeli, 2002; Tu et al., 2022). More formally, we introduce the following condition:

Definition 2.2 (Stability)**.** We say that (π, f) are (C, ρ) **closed-loop stable** if for all trajectories τ under (π, f), we have that $\max\{\|A^{\mathrm{cl}}_{k:h|\tau} B_k\|, \|A^{\mathrm{cl}}_{k:h|\tau}\|\} \leq C\rho^{h-k}$, for all $1 \leq k \leq h$. We say f is (C, ρ)-**open loop stable** if the above holds when $\pi(s) \equiv 0$ is the identically zero policy.

This is equivalent to the notion of *exponential incremental stability* studied in Tu et al. (2022); Pfrommer et al. (2022); Simchowitz et al. (2025); we defer the formal derivation to Definition B.1.

Assumption 2.3. Following prior work Simchowitz et al. (2025); Zhang et al. (2025), we assume that the expert policy $(\pi^\star, f^\star)$ is (C, ρ) closed-loop stable for some $C > 0$, $\rho \in (0, 1)$.

The above assumption is a formalization of expert robustness: the expert $\pi^\star$ policy can eventually recover from bounded perturbations on the true dynamics $f^\star$. Importantly, as shown in Section D, Assumption 2.3 does not mean that it is easy to learn some other $f \in \mathcal{F}$ such that $(\pi^\star, f)$ is similarly stable. This challenge means learned models can be **exponentially sensitive** to learning errors. This insight leads to the following negative results, established in Section C.

Theorem 2.4. *There exist continuous control instances where Equation* (1) *yields constant regret in model error, yet even refined notions of concentrability scale exponentially in horizon and action dimension. In other words,* **concentrability identifies instances as intractable, when they are in fact benign.** *Conversely, there exist instances with perfect concentrability, yet any offline algorithm yields suboptimality scaling exponentially in horizon. In other words,* **concentrability identifies instances as benign, when they are in fact intractable.**

Proof Sketch. The first direction comes naturally, as density ratios tend to grow exponentially in horizon and dimension. The second direction comes because $\mathbb{P}^\star_h$ is unable to rule out unstable closed-loop systems, so error gets compounded exponentially through unstable $A^{\mathrm{cl}}_{k:h}$. □

Further, for a policy π, we say that our problem instance suffers **exponential policy evaluation** if there exists some constant $\Gamma > 1$ such that, for every $H \geq 2$, there exists loss $\epsilon_{\max}(H)$ such that for all $0 < \epsilon \leq \epsilon_{\max}(H)$, $\sup_{f \in \mathcal{F}(\epsilon)} J_H(\pi) - Q^{\pi,f}_H \gtrsim \Gamma^H \epsilon$. If a problem suffers from exponential policy evaluation, reducing model error will be insufficient to rule out models that evaluate reward exponentially badly in horizon compared to their one-step model error.

We show that in the general case and absent any *new* notions of coverage, policy evaluation can still pose an intractable problem: inherently subverting any good regret upperbound for offline RL.

Theorem 2.5. *There exists a setting satisfying Assumption B.2 where $(\pi, f^\star)$ is closed-loop stable that suffers from the exponential policy evaluation problem* ***under any dataset distribution*** *μ_h reachable under $f^\star$ at timestep h from the initial state distribution.*

Proofs and formal statements can be found in Section D. Clearly, additional conditions are required to inform problem feasibility, and additional notions of badness are required during model learning.

Key Takeaways

- Most offline RL work relies on a concentrability assumption or on refinements of the concept.
- Theorem 2.4 shows that concentrability is **orthogonal to problem difficulty** in the continuous control setting, identifying *benign cases as hard* and *hard cases as benign*.
- Theorem 2.5 implies that **policy-evaluation is hard** in general under current frameworks, demonstrating the need for an alternative framework for characterizing offline RL instances.

3 A GEOMETRIC DEFINITION OF COVERAGE.

The failure of concentrability is that it is only concerned with *likelihood*, and does not contend with the geometric nature of continuous control. For a benign instance in Theorem 2.4, while the *likelihood* of seeing optimal data in μ_h may be low, the states and actions in the dataset may be *close* to optimal ones and may therefore be informative, e.g., by virtue of Lipschitzness.

We argue that stability better describes the difficulty of offline RL in continuous control. In particular, the failures of concentrability in Theorem 2.4 arise from: a) stable instances (for which offline RL is easy) being able to induce poor concentrability, and b) instances where any learned policy is destabilizing can exhibit ideal concentrability. To incorporate the geometric nature of stability we consider geometric transport metrics instead of likelihood ratios considered for concentrability, for example the Wasserstein distance (given formally in Definition B.3).

By incorporating geometric structure, Wasserstein distance complements stability and avoids the pitfalls of concentrability: a) stable instances induce bounded, horizon-independent Wasserstein distance, b) destabilizing policies induce exponential-in-horizon Wasserstein distance (see Section C.1 for proofs and formalization). This allows us to make the following expert coverage assumption.

Assumption 3.1. The dataset distribution satisfies $W_2(\mathbb{P}^\star_h, \mu_h) \le W^\star_{\mathbb{P}}$ for all $1 \le h \le H$.

This motivates an analysis of difficulty that focuses on a discussion of stability and geometric coverage instead of concentrability, as this would give way to a more informative problem decomposition.

3.1 AN INTUITIVE DECOMPOSITION FROM GEOMETRIC COVERAGE

Recall that pessimism reduces offline RL to analyzing the reward difference between the expert's trajectories under dynamics f and $f^\star$. Let us consider the illustrative linear setting such that the Taylor approximation is exact, where the Jacobian linearizations are simply the system matrices: $s_{h+1} = f_h(s_h, a_h) = A_h s_h + B_h a_h$, $a_h = \pi^\star_h(s_h) = K^\star_h s_h$. Take $A^{\mathrm{cl}\star}$ and A^{cl} to be the true and candidate transition matrices induced by the expert policy $\pi^\star$, such that $s^\star_{h+1} = A^{\mathrm{cl}\star}_h s^\star_h$ and $s_{h+1} = A^{\mathrm{cl}}_h s_h$. If we expand out the trajectory difference, the state error decomposes as:,

$$s^\star_{h+1} - s_{h+1} = \underbrace{A^{\mathrm{cl}}_h(s^\star_h - s_h)}_{\text{error propagation}} + \underbrace{(A^{\mathrm{cl}\star} - A^{\mathrm{cl}}_h)s^\star_h}_{\text{model error}}$$

This decomposition highlights the crux of argument throughout this work. Model fitting on the expert trajectory will control the errors present in the second term $A^{\mathrm{cl}\star}_k s^\star_k - A^{\mathrm{cl}}_k s^\star_k$; however, these errors then propagate through the system as governed by $A^{\mathrm{cl}}_{k:h}$. Two cases clearly appear: the case where the closed-loop system under the learned model A^{cl}_h is **stable**, and when it isn't.

When the candidate system is closed-loop stable, *controlling the model error is sufficient* as Definition 2.2 gives us that $\|A^{\mathrm{cl}}_{k:h}\| \le C\rho^{h-k}$, thus the error propagation term quickly shrinks. Unfortunately, without access to the expert $\pi^\star$, establishing that the candidate system is closed-loop stable under the expert is nontrivial. This difficulty is demonstrated in Simchowitz et al. (2025), where two open-loop dynamics functions produce identical closed-loop stable behavior when paired with one of two policies; however, both systems are destabilized if the policies are swapped. In Section 4.2, we delineate the natural question of when stability of the candidate system can be enforced.

$$s^\star_{h+1} - s_{h+1} = \underbrace{A^{\mathrm{cl}\star}_h(s^\star_h - s_h)}_{f^\star \text{ error propagation}} + \underbrace{(A^{\mathrm{cl}\star} - A^{\mathrm{cl}}_h)s^\star_h}_{\text{model error}} + \underbrace{(A^{\mathrm{cl}}_h - A^{\mathrm{cl}\star}_h)(s^\star_h - s_h)}_{f - f^\star \text{ error propagation}}$$

When only the true closed-loop system $A^{\mathrm{cl}\star}_h$ is stable, candidate models with *unstable model mismatch* $A^{\mathrm{cl}}_h - A^{\mathrm{cl}\star}_h$ can make the error propagation exponential even if $A^{\mathrm{cl}\star}_h$ is stable. This is unavoidable in the general case, as proven by Theorem 2.5. the Therefore, a natural inquiry is how geometric coverage conditions identify tractable offline RL instances, described in Section 4.3.

Key Takeaways

- **Stability** provides a more meaningful *geometric* decomposition of problem difficulty than concentrability, denoting when a system will converge over time.
- Stable problems are **naturally benign**, but enforcing stability is challenging.
- Unstable problems **require additional geometric coverage conditions** to be feasible.
- **Wasserstein distance** naturally complements stability and is bounded in stable instances and explodes in unstable instances, lending itself to a notion of **geometric coverability** of systems.

4 Main Results

4.1 Algorithm: Pessimistic Model Based Offline RL

As mentioned earlier, we select a model class $\mathcal{F}$ of candidate dynamics models, and consider a refined subset $\mathcal{F}(\epsilon) \subseteq \mathcal{F}$ where every $f \in \mathcal{F}(\epsilon)$ has low model error on the dataset distribution:

$$\mathcal{F}(\epsilon) \triangleq \{f \in \mathcal{F} : \mathrm{J}_{\mathrm{model}}(f; \mu_h) \leq \epsilon,\ \forall h \in [H]\},\ \mathrm{J}_{\mathrm{model}}(f; \mu_h) \triangleq \sqrt{\mathop{\mathbb{E}}_{\mu_h} \|f(s_h, a_h) - f^\star(s_h, a_h)\|^2}$$

This refined model class will be used in the pessimistic algorithm defined in Equation (1). Finally, we make a realizablity assumption:

Assumption 4.1. We assume the true dynamics is realizable $f^\star \in \mathcal{F}$.

We compile and formalize all these preliminary assumptions in Assumption B.2.

4.2 Open-Loop Stable Data: Policy Evaluation and Action Chunking

When your data comes from an open-loop stable system, we show that the **policy class** determines whether offline RL is feasible. Simchowitz et al. (2025) proves that even when your data is open-loop stable and an optimal Markovian policy exists, expert data and optimal concentrability can be insufficient to rule out models with *inherently poor geometric coverage*. **Versions of this property can be found in many real-world systems**, as open-loop incremental stability is a natural product of *low-level tracking controllers* as found in many robotic manipulation settings (Block et al., 2023).

Assumption 4.2. We assume that the true dynamics $f^\star$ are (C, ρ) open-loop EISS, and we restrict our model learning to a class of (C, ρ) open-loop EISS dynamics functions, without loss of generality with $\rho \geq 1/e$.

We further note that this open-loop model class parametrization has precedent in recurrent equilibrium network (Wang & Manchester, 2022; Revay et al., 2023) and contracting neural network (Revay & Manchester, 2020; Jaffe et al., 2024) literature, where models are constructed to satisfy contraction guarantees across timesteps.

We show that a simple recipe called *action chunking* suffices to leverage the open-loop properties of the underlying system. We inherit our definition of an induced chunking policy from Zhang et al. (2025), considering how this abstraction has proven useful in the imitation learning setting.

Definition 4.3 (Induced Chunking Policy)**.** A chunking policy is specified by a chunk-length ℓ, and mappings $\mathsf{chunk}_i[\pi] : \mathcal{S} \to \mathcal{A}$, $i \in [\ell]$ such that, for $k \in \mathbb{Z}_{\geq 0}$ and $i \in [\ell]$ and $h = k\ell + i$, $\pi(s_{1:h}, a_{1:h-1}, h) = \mathsf{chunk}_i[\pi](s_{k\ell+1})$. We also write $\mathsf{chunk}[\pi](s) = (\mathsf{chunk}_1[\pi](s), \cdots, \mathsf{chunk}_\ell[\pi](s))$. For simplicity, we always assume ℓ divides $H - 1$.

We define a chunked policy as taking in a Markovian deterministic policy $\pi : \mathcal{S} \to \mathcal{A}$, a dynamics map $f : \mathcal{S} \times \mathcal{A} \to \mathcal{S}$, and a chunk length $\ell \in \mathbb{Z}_{\geq 1}$ such that $\tilde{\pi} = \mathsf{chunked}(\pi, f, \ell)$, $\tilde{\pi} : \mathcal{S} \to \mathcal{A}^\ell$, defined with $\mathsf{chunk}_i[\tilde{\pi}](s) \triangleq \pi((f\pi)^{i-1}(s))$ where $(f\pi)(s) \triangleq f(s, \pi(s))$ and $(f\pi)^i$ is understood as repeated composition.

Action chunking means policy actions are telegraphed by planning under a closed-loop stable system (even though these dynamics need not be correct), which ensures the policy actions will not blow up. This leverages the open-loop stability of the true system by guaranteeing a reasonable sequence of actions will *eventually* lead to reasonable policy evaluation error, whereas a Markovian policy may induce closed-loop unstable behavior due to erroneous dynamics and re-planning too frequently for open-loop stability to kick in.

Zhang et al. (2025, Proposition 3.1) (restated in Proposition E.3) proved that for sufficiently long chunk lengths, action chunking suffices to induce closed-loop stability across the candidate model class. Thus, we restrict ourselves to such policies, and exploit the resulting stability of the deployed system to provide guarantees on reward evaluation and policy performance.

Proposition 4.4 (Chunking makes reward evaluation feasible, informal)**.** *Let $\tilde{\pi}$ be a policy $\tilde{\pi} = \mathsf{chunked}(\pi, f_\pi, \ell)$ where (π, f_π) is a (C, ρ) EISS pair and $\ell > \log(1/\rho)^{-1} \cdot \log(\mathrm{poly}(L_\pi, C))$, and let f be a (C, ρ) open-loop EISS dynamics function. For true dynamics $f^\star$, it holds that:*

$$|Q_H^{\tilde{\pi}, f^\star} - Q_H^{\tilde{\pi}, f}| \leq O(H^2) \max_{h \leq H} \mathrm{J}_{\mathrm{model}}(f; (\tilde{\pi}, f^\star, d_1)_h).$$

All formal statements and proofs can be found in Section E.

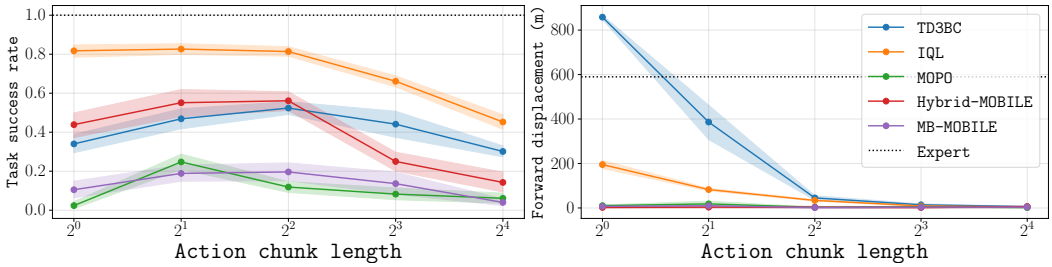


Figure 1: Visualization of the benefits of *action-chunking* under a variety of model-free and model-based algorithms. **Left:** Lift environment from `robomimic`. We see that under end-effector control, action chunking improves performance up to a stability threshold after which performance degrades. **Right:** Halfcheetah-v5 environment from MuJoCo. In open-loop unstable systems, we observe no beneficial effect from action chunking.

If we use a dynamics model class that is constrained such that every model is open-loop stable, we are now guaranteed that the salient data is geometrically coverable, as $(\tilde{\pi}, f^\star)$ will be closed-loop stable. Notably, the expert policy $\pi^\star$ and the true dynamics $f^\star$ themselves consist of just such a closed-loop stable policy-dynamics pair such that $\mathsf{chunked}(\pi^\star, f^\star, \ell)$ would be stable over any of these candidate models. These facts come together for our positive result:

Theorem 4.5. *Let us have a dynamics class $\mathcal{F}$ such that each dynamics function f is (C, ρ) open-loop EISS, and let Π be a class of length-ℓ chunked policies such that for all $\tilde{\pi} \in \Pi$, $\tilde{\pi} = \mathsf{chunked}(\pi, f_\pi, \ell)$ where (π, f_π) is a (C, ρ) EISS pair and $\ell > \log(1/\rho)^{-1} \cdot \log(\mathrm{poly}(L_\pi, C))$. We assume that ℓ is fixed independently of H. It holds that $\hat{\pi} \in \arg\max_{\tilde{\pi} \in \Pi} \inf_{f \in \mathcal{F}(\epsilon)} Q^{\tilde{\pi}, f}$ satisfies*

$$J_H(\pi^\star) - J_H(\hat{\pi}) \leq \max_{h \leq H} O(H^2) \left(\sup_{f \in \mathcal{F}(\epsilon)} \mathrm{J}_{\mathrm{model}}(f; \mu_h) + W^\star_{\mathbb{P}} \right).$$

We note that this open-loop stability of the data is crucial for action chunking to suffice. Without it, open-loop instability means erroneous sequences of actions generally lead to unstable closed-loop behavior. In fact, Simchowitz et al. (2025) reveals that absent open-loop stability, there exist systems for which no modification of the policy or dynamics class alone suffices to ensure good offline RL under perfect coverage of the expert, with $W^\star_{\mathbb{P}} = 0$. This once again motivates exploration into the notions of geometric offline coverage required in order to allow near-optimal policy selection in the open-loop unstable setting.

Key Findings

- When your data comes from an open-loop stable system, *policy evaluation remains difficult*.
- In this setting, simply modifying the policy class, i.e. via action chunking, removes **geometrically uncoverable** closed-loop behavior: this makes offline RL feasible!
- Experimental results support our assertion that **open-loop stability of the data is crucial** for this policy class modification to be sufficient.

4.3 Geometric Coverage in the Unstable Setting

Without open-loop stability of the true dynamics, we can no longer rely on action chunking to guarantee stability across the candidate model class. Instead, we must ask what additional offline data coverage conditions are needed to subvert the lowerbound setting proposed in Theorem 2.5. Our first result makes the intuition of the key analysis in Section 3.1 explicit, proving that controlling the Jacobian mismatch was the missing step required to control exponential policy evaluation.

Proposition 4.6. *Let $\pi^\star, f^\star$ be a (C, ρ) EISS pair, and let us denote $s^\star_h \triangleq s_h^{\pi^\star, f^\star, s_1}$ and $a^\star_h \triangleq \pi^\star(s^\star_h)$ for some initial state $s_1 \sim d_1$. For any $1 \leq \ell < H$ and for some $C_{\mathrm{sys}} \geq 1$, we have:*

$$D^\star_h(f) \triangleq \nabla_s (f - f^\star)(s^\star_h, a^\star_h) + \nabla_s \pi^\star(s^\star_h) \nabla_a (f - f^\star)(s^\star_h, a^\star_h)$$

$$J_H(\pi^\star) - Q_H^{\pi^\star, f} \lesssim H \ell^2 C^\ell_{\mathrm{sys}} \max_{1 \leq h \leq \ell} \mathrm{J}_{\mathrm{model}}(f; \mathbb{P}^\star_h) + (H - \ell)^2 \left(\max_{\ell+1 \leq h \leq H} \mathrm{J}_{\mathrm{model}}(f; \mathbb{P}^\star_h) + \max_{\ell+1 \leq h \leq H} \sqrt{\mathbb{E}_{\mathbb{P}^\star_h} \|D^\star_h(f)\|^2} \right)$$

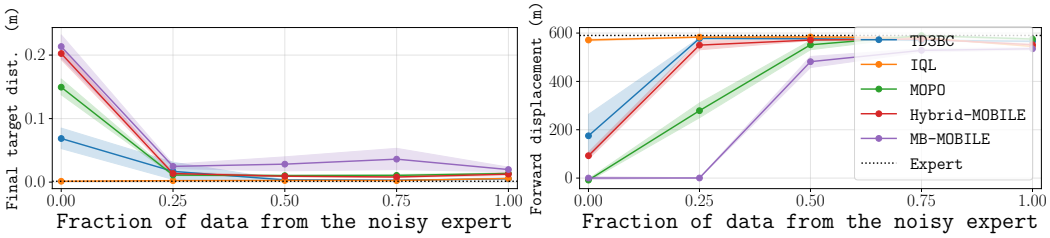


Figure 2: Visualization of the benefits of *exploratory data collection* under a variety of model-free and model-based algorithms, for noise scale $\sigma_u = 0.5$. **Left:** Reacher-v5 environment from MuJoCo. **Right:** Halfcheetah-v5 environment from MuJoCo. We find that all algorithms experienced strongly increased performance even under low levels of noise-injected data inclusion, with the effect plateauing afterwards. Additional discussion and details are provided in Section G.

This proposition clearly delineates the two regimes of matching discussed in Section 3.1: when we only have *zeroth order matching*, as for timesteps $1 \le h \le \ell$, we absorb an **exponential dependence on ℓ in the worst case**; however, if we're able to control the *first order errors* $D_h^\star(f)$, a **polynomial guarantee arises**. This is precisely the issue exploited in Theorem 2.5: controlling zeroth order error is possible on any distribution that covers the desired trajectories, but first order model error can only be controlled if the data satisfies **sufficient excitation properties in the state space**.

To formalize this notion of data coverage, we introduce two measures of closed-loop dynamics sensitivity to perturbations to the expert trajectory.

Definition 4.7. For a given initial state s_1, let us take B_h and $A^{\mathrm{cl}}_{k:h}$ for the Jacobian linearizations of the expert trajectory $(\pi^\star, f^\star)$ defined in Section 3. We define the **action controllability Gramian** $W^u_{1:h}(s_1)$ and the **state controllability Gramian** $W^x_{1:h}(s_1)$ as follows: $W^u_{1:h}(s_1) \triangleq \sum_{k=1}^{h-1} A^{\mathrm{cl}}_{k+1:h} B_k B_k^\top A^{\mathrm{cl}\,\top}_{k+1:h}, \quad W^x_{1:h}(s_1) \triangleq \sum_{k=1}^{h-1} A^{\mathrm{cl}}_{k+1:h} A^{\mathrm{cl}\,\top}_{k+1:h}$.

These Gramians capture the sensitive directions of the closed-loop dynamics to action perturbations for $W^u_{1:h}$ or state perturbations for $W^x_{1:h}$. In order for policy rollouts, and importantly behavior policy rollouts, to express all the possible states that model errors could induce, the true system would have to satisfy the domination condition of $W^u_{1:h}(s_1) \gtrsim W^x_{1:h}(s_1)$. While such guarantees have precedent in many settings such as controllable LTI systems, we can make two key observations:

First, our system does not need to express every state perturbation: our focus is on **first-order error directions expressible by our model class $\mathcal{F}$**. As such, we may constrain our Gramian domination condition to be model-class aware. This constraint could be exploited by recent work on representation learning for world models (Park et al., 2022) and physics-informed dynamics models (Lutter et al., 2019; Zhong et al., 2019; Greydanus et al., 2019). This first-order error projector can be formalized as follows.

Definition 4.8. For model class $\mathcal{F}$ and first-order difference $D_h^\star(f) \triangleq \nabla_s (f - f^\star)(s_h^{\pi^\star, f^\star}, a_h^{\pi^\star, f^\star}) + \nabla_s \pi^\star(s_h^{\pi^\star, f^\star}) \nabla_a (f - f^\star)(s_h^{\pi^\star, f^\star}, a_h^{\pi^\star, f^\star})$, we define $\mathcal{R}_h^{\mathcal{F}} \triangleq \mathrm{span}\{\mathrm{range}(D_h^\star(f)) \ : \ f \in \mathcal{F}\}$, and the corresponding orthogonal projection matrix $\mathcal{P}_{\mathcal{R}_h^{\mathcal{F}}}$.

Second, we can relax the controllability assumption by allowing directions to only become excitable after some finite number of timesteps. This is salient in *higher-order systems*, such as locomotion systems with thrust or acceleration inputs where actions only affect position after a delay. This lends us to consider **ℓ-step action excitation**, where the condition $W^u_{1:h}(s_1) \gtrsim W^x_{1:h}(s_1)$ holds only for $h > \ell$. We note that $\ell = 1$ captures the special case of direct excitation, whereas the lack of such a finite ℓ lends itself to the impossibility results from Theorem 2.5.

These intuitions allow us to formalize our assumption.

Assumption 4.9 (Restricted Controllability)**.** Fix $\ell < H$. We assume there exists $c_\ell > 0$ such that for all $s_1 \in \mathrm{supp}(d_1)$ and for every $h \in [\ell + 1, H]$, $\mathcal{P}_{\mathcal{R}_h^{\mathcal{F}}} W^u_{1:h}(s_1) \mathcal{P}_{\mathcal{R}_h^{\mathcal{F}}} \succeq c_\ell \mathcal{P}_{\mathcal{R}_h^{\mathcal{F}}} W^x_{1:h}(s_1) \mathcal{P}_{\mathcal{R}_h^{\mathcal{F}}}$.

This is inherently a geometric coverability condition that must be satisfied by the system in order to allow collected data to be sufficiently informative as to make offline RL feasible.

We use action-noise injection as a sufficient construction of a reference distribution with the required local excitation. Although a broader class of distributions is sufficient to provide the required excitation, we focus on this simple construction for clarity.

Definition 4.10 (Expert action-noise injection)**.** Fix $\sigma_u > 0$. For $s_1 \sim d_1$, let $\tilde{s}_1 = s_1$, and let $z_1, \ldots, z_H \overset{\text{i.i.d.}}{\sim} \mathrm{Unif}(\mathbb{B}^{d_u}(1))$ be drawn independently of s_1. Define the action-noised expert trajectory by $\tilde{a}_h \triangleq \pi^\star(\tilde{s}_h) + \sigma_u z_h$ and $\tilde{s}_{h+1} \triangleq f^\star(\tilde{s}_h, \tilde{a}_h)$ for all $h \in [H]$. Let $\tilde{\nu}_h^{\sigma_u}$ denote the joint distribution of $(\tilde{s}_h, \tilde{a}_h)$ induced by $s_1, z_1, \ldots, z_h$. When σ_u is clear from context, we denote the distribution as $\tilde{\nu}_h$. We generally assume for simplicity that $\pi^\star(\tilde{s}_h) + \sigma_u z_h \in \mathcal{A}$.

Under restricted controllability, this exploratory data can rule out models with high first order error.

Proposition 4.11. *Let $\pi^\star, f^\star$ be a (C, ρ) EISS pair satisfying Assumption 4.9. For $s_h^\star, a_h^\star$ to be generated from trajectories of $\pi^\star, f^\star$ given $s_1 \sim d_1$, for every candidate $f \in \mathcal{F}$, and for sufficiently small σ_u, for every $h > \ell$, $\|D_h^\star(f)\|^2 \le O\left(\frac{d_u}{c_\ell \sigma_u^2}\right) \mathop{\mathbb{E}}_{\tilde{\nu}_h | s_1} \|(f - f^\star)(\tilde{s}_h, \tilde{a}_h)\|^2 + O\left(\frac{d_u \sigma_u^2}{c_\ell}\right)$.*

Finally, pessimistic policy optimization translates this difference between modeled and real expert trajectories into a polynomial-in-horizon upperbound on regret.

Theorem 4.12. *Let us take a system satisfying Assumption 4.9, and let us take $\tilde{\nu}_h$ from Definition 4.10 for sufficiently small $\sigma_u > 0$. It holds that $\hat{\pi} = \arg\max_{\pi \in \Pi} \inf_{f \in \mathcal{F}(\epsilon)} Q^{\pi, f}$ satisfies, for some $C_{\mathrm{sys}} \ge 1$,*

$$J(\pi^\star) - J(\hat{\pi}) \lesssim \sup_{f \in \mathcal{F}(\epsilon)} H \ell^2 C_{\mathrm{sys}}^\ell \max_{1 \le h \le \ell} (\mathrm{J}_{\mathrm{model}}(f; \mu_h) + W_{\mathbb{P}}^\star) + (H - \ell)^2 \max_{\ell + 1 \le h \le H} \left(\left(1 + \frac{1}{\sigma_u} \sqrt{\frac{d_u}{c_\ell}}\right) \mathrm{J}_{\mathrm{model}}(f; \mu_h) + W_{\mathbb{P}}^\star + W_2(\tilde{\nu}_h^{\sigma_u}, \mu_h) \right)$$

The first term inevitably scales exponentially with the controllability horizon as we can't control the intermediate steps (Mania et al., 2019), whereas after $h > \ell$ geometric coverage allows us to control exponential compounding.

Key Findings

- Unstable systems require **exploratory, suboptimal data** to rule out bad models.
- **Restricted controllability** informs which error directions *must* be covered.
- Experiments on locomotion and end-effector control support our theoretical findings.

4.4 Experimental Evidence

We run experiments on a suite of model-free (TD3+BC, IQL) and model-based (MOPO, MOBILE) algorithms; we include two configurations of MOBILE: MB-MOBILE is run with no offline data use during policy training, while Hybrid-MOBILE uses half offline and half model-generated data.

Action Chunking. We run complementary experiments in the `robomimic` Lift and the MuJoCo HalfCheetah-v5 tasks, holding all parameters constant across chunk lengths except the necessary training targets and input/output dimensions. In line with our theoretical results, we observe low chunk lengths (2-4) exhibit the best performance in the open-loop stable Lift task, while HalfCheetah-v5 fails catastrophically with any action chunking.

Geometric Coverage. We run experiments in the MuJoCo Reacher-v5 and HalfCheetah-v5 tasks while varying the fraction of trajectories that were collected from the expert policy versus expert action-noise injection, holding all parameters constant across runs. We find that even small levels of exploratory data collection produce increased performance, demonstrating the efficacy of geometric coverage in practice. Further ablations and discussion could be found in Section G.

5 Discussion

In this work, we demonstrate how standard notions of data coverage are uninformative for continuous space. We demonstrate that when provided with open-loop stable data, the practice of action chunking allows the dynamics to stabilize the policy class in closed loop and obtain polynomial policy-evaluation guarantees. Without open-loop stability, we provide a minimal geometric coverability assumption which ensures that action excitation can identify how candidate models respond to perturbations, once again overcoming the lower-bound obstruction. With these theoretical results in hand, we conduct action chunking experiments on the robomimic Lift environment and noise injection experiments on the MuJoCo Reacher-v5 and HalfCheetah-v5 environments. These all support our theoretical findings, and indicate that our practices provide beneficial effects to a wide variety of both model-based and model-free algorithms.

AI DISCLOSURE

In this work, we used generative AI tools for assistance in providing critical ingredients for proving mathematical claims, assistance in the writing of proofs, and implementing methods. We have not used generative AI tools for help developing theoretical models, conceptual frameworks, formulating mathematical claims, or proposing or refining hypotheses, and the rest of the required disclosure tasks are not applicable to this work. We have extensively reviewed and verified all AI-assisted work, with any mathematical claims being rigorously re-proven and code tested for correctness. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

REFERENCES

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

Kavosh Asadi, Dipendra Misra, and Michael Littman. Lipschitz continuity in model-based reinforcement learning. In Jennifer Dy and Andreas Krause (eds.), *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pp. 264–273. PMLR, 10–15 Jul 2018. URL https://proceedings.mlr.press/v80/asadi18a.html.

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*, 36:48534–48547, 2023.

Jinglin Chen and Nan Jiang. Information-theoretic considerations in batch reinforcement learning. In *International conference on machine learning*, pp. 1042–1051. PMLR, 2019.

Simon S Du, Sham M Kakade, Ruosong Wang, and Lin F Yang. Is a good representation sufficient for sample efficient reinforcement learning? *arXiv preprint arXiv:1910.03016*, 2019.

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. URL https://openreview.net/forum?id=8KPyJm4gt5.

Justin Fu, Aviral Kumar, Ofir Nachum, George Tucker, and Sergey Levine. D4rl: Datasets for deep data-driven reinforcement learning. *arXiv preprint arXiv:2004.07219*, 2020.

Scott Fujimoto and Shixiang Shane Gu. A minimalist approach to offline reinforcement learning. *Advances in neural information processing systems*, 34:20132–20145, 2021.

Scott Fujimoto, David Meger, and Doina Precup. Off-policy deep reinforcement learning without exploration. In *International conference on machine learning*, pp. 2052–2062. PMLR, 2019.

Samuel Greydanus, Misko Dzamba, and Jason Yosinski. Hamiltonian neural networks. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett (eds.), *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019. URL https://proceedings.neurips.cc/paper_files/paper/2019/file/26cd8ecadce0d4efd6cc8a8725cbd1f8-Paper.pdf.

David Ha and Jürgen Schmidhuber. World models. *arXiv preprint arXiv:1803.10122*, 2(3):440, 2018.

Tuomas Haarnoja, Aurick Zhou, Pieter Abbeel, and Sergey Levine. Soft actor-critic: Off-policy maximum entropy deep reinforcement learning with a stochastic actor. In *International conference on machine learning*, pp. 1861–1870. Pmlr, 2018.

Danijar Hafner, Timothy Lillicrap, Jimmy Ba, and Mohammad Norouzi. Dream to control: Learning behaviors by latent imagination. In *International Conference on Learning Representations*, 2020. URL https://openreview.net/forum?id=S1lOTC4tDS.

Charles A Hepburn and Giovanni Montana. Model-based trajectory stitching for improved offline reinforcement learning. *arXiv preprint arXiv:2211.11603*, 2022.

Sean Jaffe, Alexander Davydov, Deniz Lapsekili, Ambuj K Singh, and Francesco Bullo. Learning neural contracting dynamics: Extended linearization and global guarantees. *Advances in Neural Information Processing Systems*, 37:66204–66225, 2024.

Michael Janner, Justin Fu, Marvin Zhang, and Sergey Levine. When to trust your model: Model-based policy optimization. *Advances in neural information processing systems*, 32, 2019.

Michael Kearns and Satinder Singh. Near-optimal reinforcement learning in polynomial time. *Machine learning*, 49(2):209–232, 2002.

Rahul Kidambi, Aravind Rajeswaran, Praneeth Netrapalli, and Thorsten Joachims. Morel: Model-based offline reinforcement learning. *Advances in neural information processing systems*, 33: 21810–21823, 2020.

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

Qiyang Li, Zhiyuan (Paul) Zhou, and Sergey Levine. Reinforcement learning with action chunking. In D. Belgrave, C. Zhang, H. Lin, R. Pascanu, P. Koniusz, M. Ghassemi, and N. Chen (eds.), *Advances in Neural Information Processing Systems*, volume 38, Main Conference, pp. 55518–55553. Curran Associates, Inc., 2025. doi: 10.52202/085713-1856. URL https://proceedings.neurips.cc/paper_files/paper/2025/file/50348e8f9aef984abe0ea1ec2a326f78-Paper-Conference.pdf.

Cong Lu, Philip J. Ball, Tim G. J. Rudner, Jack Parker-Holder, Michael A Osborne, and Yee Whye Teh. Challenges and opportunities in offline reinforcement learning from visual observations. *Transactions on Machine Learning Research*, 2023. ISSN 2835-8856. URL https://openreview.net/forum?id=1QqIfGZOWu.

Fan-Ming Luo, Tian Xu, Hang Lai, Xiong-Hui Chen, Weinan Zhang, and Yang Yu. A survey on model-based reinforcement learning. *Science China Information Sciences*, 67(2):121101, 2024.

Michael Lutter, Kim Listmann, and Jan Peters. Deep lagrangian networks for end-to-end learning of energy-based control for under-actuated systems. In *2019 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pp. 7718–7725. IEEE, 2019.

Ajay Mandlekar, Fabio Ramos, Byron Boots, Silvio Savarese, Li Fei-Fei, Animesh Garg, and Dieter Fox. Iris: Implicit reinforcement without interaction at scale for learning control from offline robot manipulation data. In *2020 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 4414–4420, 2020. doi: 10.1109/ICRA40945.2020.9196935.

Ajay Mandlekar, Danfei Xu, Josiah Wong, Soroush Nasiriany, Chen Wang, Rohun Kulkarni, Li Fei-Fei, Silvio Savarese, Yuke Zhu, and Roberto Martín-Martín. What matters in learning from offline human demonstrations for robot manipulation. *arXiv preprint arXiv:2108.03298*, 2021.

Horia Mania, Stephen Tu, and Benjamin Recht. Certainty equivalence is efficient for linear quadratic control. *Advances in neural information processing systems*, 32, 2019.

Tatsuya Matsushima, Hiroki Furuta, Yutaka Matsuo, Ofir Nachum, and Shixiang Gu. Deployment-efficient reinforcement learning via model-based offline optimization. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=3hGNqpI4WS.

Jung Yeon Park, Ondrej Biza, Linfeng Zhao, Jan Willem van de Meent, and Robin Walters. Learning symmetric embeddings for equivariant world models. *arXiv preprint arXiv:2204.11371*, 2022.

Kwanyoung Park, Seohong Park, Youngwoon Lee, and Sergey Levine. Scalable offline model-based rl with action chunks. In *International Conference on Learning Representations*, volume 2026, pp. 9599–9623, 2026.

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

Rafael Rafailov, Tianhe Yu, Aravind Rajeswaran, and Chelsea Finn. Offline reinforcement learning from images with latent space models. In *Learning for dynamics and control*, pp. 1154–1168. PMLR, 2021.

Antonin Raffin, Ashley Hill, Adam Gleave, Anssi Kanervisto, Maximilian Ernestus, and Noah Dormann. Stable-baselines3: Reliable reinforcement learning implementations. *Journal of Machine Learning Research*, 22(268):1–8, 2021. URL http://jmlr.org/papers/v22/20-1364.html.

Aravind Rajeswaran, Igor Mordatch, and Vikash Kumar. A game theoretic framework for model based reinforcement learning. In *International conference on machine learning*, pp. 7953–7963. PMLR, 2020.

Paria Rashidinejad, Banghua Zhu, Cong Ma, Jiantao Jiao, and Stuart Russell. Bridging offline reinforcement learning and imitation learning: A tale of pessimism. *Advances in Neural Information Processing Systems*, 34:11702–11716, 2021.

Max Revay and Ian Manchester. Contracting implicit recurrent neural networks: Stable models with improved trainability. In *Learning for Dynamics and Control*, pp. 393–403. PMLR, 2020.

Max Revay, Ruigang Wang, and Ian R Manchester. Recurrent equilibrium networks: Flexible dynamic models with guaranteed stability and robustness. *IEEE Transactions on Automatic Control*, 69(5):2855–2870, 2023.

Stephane Ross and Drew Bagnell. Efficient reductions for imitation learning. In Yee Whye Teh and Mike Titterington (eds.), *Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics*, volume 9 of *Proceedings of Machine Learning Research*, pp. 661–668, Chia Laguna Resort, Sardinia, Italy, 13–15 May 2010. PMLR. URL https://proceedings.mlr.press/v9/ross10a.html.

Max Simchowitz, Daniel Pfrommer, and Ali Jadbabaie. The pitfalls of imitation learning when actions are continuous. In Nika Haghtalab and Ankur Moitra (eds.), *Proceedings of Thirty Eighth Conference on Learning Theory*, volume 291 of *Proceedings of Machine Learning Research*, pp. 5248–5351. PMLR, 30 Jun–04 Jul 2025. URL https://proceedings.mlr.press/v291/simchowitz25a.html.

Yihao Sun, Jiaji Zhang, Chengxing Jia, Haoxin Lin, Junyin Ye, and Yang Yu. Model-Bellman inconsistency for model-based offline reinforcement learning. In Andreas Krause, Emma Brunskill, Kyunghyun Cho, Barbara Engelhardt, Sivan Sabato, and Jonathan Scarlett (eds.), *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 33177–33194. PMLR, 23–29 Jul 2023. URL https://proceedings.mlr.press/v202/sun23q.html.

Emanuel Todorov, Tom Erez, and Yuval Tassa. Mujoco: A physics engine for model-based control. In *2012 IEEE/RSJ International Conference on Intelligent Robots and Systems*, pp. 5026–5033. IEEE, 2012. doi: 10.1109/IROS.2012.6386109.

Mark Towers, Ariel Kwiatkowski, John Balis, Gianluca De Cola, Tristan Deleu, Manuel Goulão, Kallinteris Andreas, Markus Krimmel, Arjun KG, Rodrigo Perez-Vicente, J Terry, Andrea Pierré, Sander Schulhoff, Jun Jet Tai, Hannah Tan, and Omar G. Younis. Gymnasium: A standard interface for reinforcement learning environments. In D. Belgrave, C. Zhang, H. Lin, R. Pascanu, P. Koniusz, M. Ghassemi, and N. Chen (eds.), *Advances in Neural Information Processing Systems*, volume 38, Main Conference. Curran Associates, Inc., 2025. doi: 10.52202/085713-4916. URL https://proceedings.neurips.cc/paper_files/paper/2025/file/d7ff1795e8527f6443371c3933bdb52b-Paper-Datasets_and_Benchmarks_Track.pdf.

Stephen Tu, Alexander Robey, Tingnan Zhang, and Nikolai Matni. On the sample complexity of stability constrained imitation learning. In Roya Firoozi, Negar Mehr, Esen Yel, Rika Antonova,

Jeannette Bohg, Mac Schwager, and Mykel Kochenderfer (eds.), *Proceedings of The 4th Annual Learning for Dynamics and Control Conference*, volume 168 of *Proceedings of Machine Learning Research*, pp. 180–191. PMLR, 23–24 Jun 2022. URL https://proceedings.mlr.press/v168/tu22a.html.

Masatoshi Uehara and Wen Sun. Pessimistic model-based offline reinforcement learning under partial coverage. *arXiv preprint arXiv:2107.06226*, 2021.

Ruigang Wang and Ian R Manchester. Youla-ren: Learning nonlinear feedback policies with robust stability guarantees. In *2022 American Control Conference (ACC)*, pp. 2116–2123. IEEE, 2022.

Ruosong Wang, Dean P Foster, and Sham M Kakade. What are the statistical limits of offline rl with linear function approximation? *arXiv preprint arXiv:2010.11895*, 2020.

Ruosong Wang, Yifan Wu, Ruslan Salakhutdinov, and Sham Kakade. Instabilities of offline rl with pre-trained neural representation. In *International Conference on Machine Learning*, pp. 10948–10960. PMLR, 2021.

Tengyang Xie and Nan Jiang. Batch value-function approximation with only realizability. In *International Conference on Machine Learning*, pp. 11404–11413. PMLR, 2021.

Tengyang Xie, Ching-An Cheng, Nan Jiang, Paul Mineiro, and Alekh Agarwal. Bellman-consistent pessimism for offline reinforcement learning. *Advances in neural information processing systems*, 34:6683–6694, 2021.

Zhilong Zhang Yihao Sun. Offlinerl-kit: An elegant pytorch offline reinforcement learning library. https://github.com/yihaosun1124/OfflineRL-Kit, 2023.

Omar G. Younis, Rodrigo Perez-Vicente, John U. Balis, Will Dudley, Alex Davey, and Jordan K Terry. Minari, September 2024. URL https://doi.org/10.5281/zenodo.13767625.

Tianhe Yu, Garrett Thomas, Lantao Yu, Stefano Ermon, James Y Zou, Sergey Levine, Chelsea Finn, and Tengyu Ma. Mopo: Model-based offline policy optimization. *Advances in neural information processing systems*, 33:14129–14142, 2020.

Wenhao Zhan, Baihe Huang, Audrey Huang, Nan Jiang, and Jason Lee. Offline reinforcement learning with realizability and single-policy concentrability. In *Conference on Learning Theory*, pp. 2730–2775. PMLR, 2022.

Thomas T Zhang, Daniel Pfrommer, Chaoyi Pan, Nikolai Matni, and Max Simchowitz. Action chunking and exploratory data collection yield exponential improvements in behavior cloning for continuous control. *The Fourteenth International Conference on Learning Representations*, 2025.

Yaofeng Desmond Zhong, Biswadip Dey, and Amit Chakraborty. Symplectic ode-net: Learning hamiltonian dynamics with control. *arXiv preprint arXiv:1909.12077*, 2019.

Yuke Zhu, Josiah Wong, Ajay Mandlekar, Roberto Martín-Martín, Abhishek Joshi, Kevin Lin, Abhiram Maddukuri, Soroush Nasiriany, and Yifeng Zhu. robosuite: A modular simulation framework and benchmark for robot learning. *arXiv preprint arXiv:2009.12293*, 2020.

Konrad Zolna, Alexander Novikov, Ksenia Konyushkova, Caglar Gulcehre, Ziyu Wang, Yusuf Aytar, Misha Denil, Nando De Freitas, and Scott Reed. Offline learning from demonstrations and unlabeled experience. *arXiv preprint arXiv:2011.13885*, 2020.

CONTENTS

1 Introduction **1**

1.1 Contributions 2

2 Preliminaries **2**
2.1 Traditional Notions of Coverage are Insufficient in Continuous Control 3
2.2 Jacobian Linearizations and Dynamical Stability 3

3 A Geometric Definition of Coverage. **5**
3.1 An Intuitive Decomposition from Geometric Coverage 5

4 Main Results **6**
4.1 Algorithm: Pessimistic Model Based Offline RL 6
4.2 Open-Loop Stable Data: Policy Evaluation and Action Chunking 6
4.3 Geometric Coverage in the Unstable Setting 7
4.4 Experimental Evidence 9

5 Discussion **9**

A Related Work **14**

B Preliminaries **15**
B.1 The Difficulty of Bellman Completeness in Continuous Space 16
B.2 Applying the Discrete Analysis to the Continuous Control Setting 16

C Coverage in the Continuous Context **17**
C.1 Towards a Geometric Notion of Coverage: Wasserstein and Stability 21

D When Offline RL Fails **25**

E Extended Proofs for Section 4.2 **31**

F Extended Proofs for Section 4.3 **38**

G Experimental Validation **49**
G.1 Implementation Details 51

A Related Work

Model-based RL has been found to be highly sample efficient in both the online (Janner et al., 2019; Rajeswaran et al., 2020) and offline (Matsushima et al., 2021) settings, as the dynamics model is trained through direct supervision and, once trained, can be reused for synthetic data generation and planning. The application area of world models in particular has sought to make use of the generalization and reuse abilities of model-based offline methods in complex tasks and settings Rafailov et al. (2021); Lu et al. (2023).

Since offline data may come from a variety of sources and learned policy rollouts are not witnessed, offline RL contends with distribution shift issues (Fujimoto et al., 2019; Wang et al., 2020; 2021). For model-based methods, these problems often emerge as *model exploitation*, where naive planning under a learned model can produce poor results in areas of low dataset coverage. These challenges are relatively well-studied in the discrete setting, and theoretical work has explored both negative results delineating the necessity of good representations and notions of "coverage" (Du et al., 2019;

Wang et al., 2020; Chen & Jiang, 2019) as well as settings and algorithms that provably scale nicely with sample complexity, Bellman or model error, and discounted horizon (Xie et al., 2021; Xie & Jiang, 2021; Zhan et al., 2022). In the model-based setting, pessimistic and conservative penalties are often applied in practice to prevent learned policies from deviating from high-confidence areas, with impressive results (Sun et al., 2023; Kidambi et al., 2020).

We focus on the continuous control setting. In the behavior cloning setting, a special case of offline RL where the offline dataset contains rollouts of a desirable policy and enjoys perfect data "coverage", benign guarantees have been established in the discrete setting (Ross & Bagnell, 2010; Foster et al., 2024). However, it is well-known that BC can suffer *exponential-in-horizon compounding errors* (Tu et al., 2022; Pfrommer et al., 2022). Simchowitz et al. (2025) further shows that this exponential suboptimality can hold unconditionally on the learning algorithm even for benign dynamics and expert behavior. Notably, considering data coverage is ideal in BC, these works suggest a blindspot in existing offline RL frameworks. On the other hand, recent work by Zhang et al. (2025) demonstrates certain simple interventions on the BC policy deployment or data collection provably and empirically mitigate exponential compounding error.

Prior work provides guarantees for model-based RL in continuous spaces, but leaves unresolved fundamental difficulties arising in the deterministic continuous space offline setting. Yu et al. (2020) formulated bounds relating guarantees to model errors evaluated along candidate trajectories; however, they do not handle the question of when this is a controllable quantity. The analysis of Asadi et al. (2018) made a stronger uniformly bounded model error assumption, and formulated its bounds in terms of either the same model error along candidate trajectories or by controlling amplification along the candidate dynamics. We study when the required control is satisfied for this amplification to be benign. These are nontrivial issues which aren't guaranteed by realistic data settings: in fact, this discrepancy provides the basis for our Section D. In particular, the work of Uehara & Sun (2021) tackles a related problem and setting; however, they make the assumption that the true dynamics evolve through probabilistic transition kernels, constructing dynamics candidate sets through maximum likelihood and total variation distance constraints: this is incompatible with our setting, as the total variation distance induced by our deterministic candidate and true dynamics is necessarily maximal even under arbitrarily small model mismatch in the state space. Due to their probabilistic dynamics, they maintain the benign behavior described in discrete settings, where an ϵ probability of making a mistake is considered instead of an ϵ mistake in state space, allowing the usage of traditional techniques to handle the problem.

B PRELIMINARIES

We first provide an extended definition of exponential incremental input-to-state stability, which characterizes the concept of global versus local stability.

Definition B.1. A system $s_{h+1} = f(s_h, a_h)$ is (C, ρ)-*exponentially incrementally input-to-state stable* (EISS) if for all pairs of initial conditions (s_1, s_1') and input sequences $(\{a_h\}_{h\geq 1}, \{a_h'\}_{h\geq 1})$, there exist constants $C \geq 1$, $\rho \in (0, 1)$ such that for any $h \geq 1$:

$$\|s_h - s_h'\| \leq C\rho^{h-1}\|s_1 - s_1'\| + C\sum_{k=1}^{h-1} \rho^{h-1-k}\|a_k - a_k'\|, \quad h \geq 1.$$

We say a policy-dynamics pair (π, f) is (C, ρ)-EISS if the induced closed-loop dynamics are (C, ρ)-EISS. We also denote the shorthand $C_{\mathrm{stab}} = \frac{C}{1-\rho}$.

Assumption B.2 (Collected preliminary assumptions)**.** We define a function $h : \mathbb{R}^d \to \mathbb{R}^p$ as L-Lipschitz if for all $x, x' \in \mathbb{R}^d$, $\|h(x) - h(x')\| \leq L\|x - x'\|$, and M-smooth if for all $x, x' \in \mathbb{R}^d$, $\|\nabla_x h(x) - \nabla_x h(x')\|_2 \leq M\|x - x'\|$.

Regularity assumptions:

- We assume that the true dynamics are L_f-Lipschitz and M_f-smooth, and we consider a model class $\mathcal{F}$ such that for all $f \in \mathcal{F}$, f is also L_f-Lipschitz and M_f-smooth.
- We consider a policy class Π such that for all $\pi \in \Pi$, π is L_π-Lipschitz and M_π-smooth.
- We consider a reward function R that is bounded in $[0, 1]$ and is L_{R}-Lipschitz.

Realizability and stability assumptions:

- We assume that the true dynamics satisfy $f^\star \in \mathcal{F}$.
- We define the best policy in our policy class the expert comparator such that necessarily $\pi^\star \in \Pi$, and we assume that $(\pi^\star, f^\star)$ is (C, ρ) closed-loop EISS.

We assume that $W_2(\mathbb{P}_h^\star, \mu_h) \leq W_{\mathbb{P}}^\star$ for all timesteps $1 \leq h \leq H$, for some $W_{\mathbb{P}}^\star$ independent of h.

Definition B.3. We use the Wasserstein 2-distance, defined between joint state-action distributions ν_1 and ν_2 as

$$W_2(\nu_1, \nu_2) = \inf_{\gamma \in \Gamma(\nu_1, \nu_2)} \sqrt{\mathop{\mathbb{E}}_{((s,a),(s',a')) \sim \gamma} [\|(s,a) - (s',a')\|^2]}$$

where $\Gamma(\nu_1, \nu_2)$ is the set of all couplings between ν_1 and ν_2, and the joint draw $((s,a),\ (s',a')) \sim \gamma$ satisfies $(s,a) \sim \nu_1$ and $(s',a') \sim \nu_2$.

We also borrow the bump function definition from Simchowitz et al. (2025), as it is helpful in the construction of smooth and Lipschitz functions.

Lemma B.4 (Lem A.15 in Simchowitz et al. (2025), Existence of Bump Functions)**.** *For any $d \in \mathbb{N}$, there exists a C^∞ function $\mathrm{bump}_d(x) : \mathbb{R}^d \to [0,1]$, called a bump function, satisfying $\mathrm{bump}_d(x) = 1$ if and only if $\|x\| \leq 1$, $\mathrm{bump}_d(x) = 0$ if and only if $\|x\| \geq 2$. For each $p \geq 1$, $\|\nabla^p \mathrm{bump}_d(x)\|_{\mathrm{op}} \leq c_p$, where $\|\cdot\|_{\mathrm{op}}$ denotes the tensor-operator norm, and c_p is a constant independent of d but depending on p. Finally, $\nabla^p \mathrm{bump}_d(x) = 0$ for all $x : \|x\| \geq 2$. For $d = 1$, I will denote $\mathrm{bump}_1(x) \triangleq \mathrm{bump}(x)$ for brevity.*

We note that while the original statement of this lemma did not explicitly specify that $\mathrm{bump}_d(\cdot) \in [0,1]$, this was proven as part of the result.

B.1 The Difficulty of Bellman Completeness in Continuous Space

Here we provide intuition behind our assertion that Bellman completeness can unrealistic in the continuous control setting. This is generally because the statistical complexity of a Bellman complete class can be prohibitively large, especially when considering Q-functions induced by unstable closed-loop systems. Intuitively, these systems can deviate exponentially in horizon from small initial state differences, resulting in substantial variation in long-horizon Q-functions. Since these Q-functions can correspond to unstable policies, the model class has to be discriminative over an *exponential-in-horizon* large domain. Further, a sufficiently discriminative class of Q-functions, by standard covering arguments, also induces exponential-in-d_x discretization. This means a sufficiently expressive Bellman Complete class will have doubly-exponential complexity, for which statistical generalization is infeasible.

B.2 Applying the Discrete Analysis to the Continuous Control Setting

A large portion of offline reinforcement learning theory literature focuses itself upon the setting of discrete or tabular state and action spaces. In these contexts, support-sensitive distances where difference is measured through support mismatches between distributions provide the basis of natural notions of loss reduction; however, these notions become brittle in the continuous space context, and often render prior analyses inapplicable to our setting.

Model-based reinforcement learning commonly treats dynamics estimation as a conditional density-estimation task, using maximum likelihood or an equivalent supervised prediction objective in order to fit a transition model (Luo et al., 2024). When the true and candidate dynamics both belong to a regular dominated stochastic transition family, as is commonly assumed by discrete literature, maximum likelihood controls KL divergence, controlling total variation by Pinsker's inequality: this is often critical, as total variation control allows policy-value guarantees through the simulation-lemma (Kearns & Singh, 2002). Unfortunately, this route is ill-suited to the deterministic continuous dynamics setting, since TV distance and KL divergence reduce to an exact-match notion of prediction: this results in any non-zero displacement producing maximal total variation distance and infinite

KL divergence. This impossibility of non-vacuous exact-match estimation was formalized in Simchowitz et al. (2025, Proposition B.5).

Our analysis applies the continuous deterministic standard practice of fitting models through reducing Euclidean next-state model error (Luo et al., 2024). This is similar to what occurs when matching model classes with isotropic Gaussian process noise through maximum likelihood. Maximum likelihood on a Dirac would reduce to regression on the negative log likelihood of the candidate's distribution, which results in squared Euclidean regression for model classes with isotropic Gaussian process noise. Our analysis gains geometric interpretability; however, we necessarily lose the automatic simulation-lemma guarantee supplied by total variation.

C Coverage in the Continuous Context

This Radon-Nikodym density ratio construction has been weakened in some works to only transfer guarantees for certain function classes (Xie et al., 2021); however, we can show that even when attempting to transfer a model error guarantee for a smooth and stable system between two distributions centered at the same expert rollout, probabilistic coverage can deteriorate exponentially in horizon.

Setting C.1 (Exponential function-class concentrability setting)**.** We consider a setting where $\mathcal{S} = \mathcal{A} = \mathbb{R}^d$ for $d \geq 1$, and d_1 is a point mass at e_1. We define the fixed true dynamics and the behavior policy in question as

$$\pi(s) = 0, \qquad f^\star(s,a) = \lambda \operatorname{bump}_d(16a) s \operatorname{bump}_d(s) \quad \text{for } 0 < \lambda \leq \frac{1}{4(1+2c_1)}$$

and for evaluation we define the reward function and candidate dynamics for $0 < \epsilon \leq 1$.

$$\mathrm{R}(s,a) = \frac{1}{2}\min\left\{2, \max\{0, s^\top e_1\}\right\}, \qquad f_\epsilon(s,a) = \lambda \operatorname{bump}_d(16a) s \operatorname{bump}_d(s) + \epsilon\lambda s.$$

I will also define action-noise injected joint state action distributions $\tilde{\nu}_h^\sigma$ for $\sigma \geq 1/8$, where starting from $s_1^\sigma \sim d_1$, states evolve as $s_{h+1}^\sigma = f^\star(s_h^\sigma, a_h^\sigma)$ and $a_h^\sigma = \pi(s_h^\sigma) + \sigma z_h$ for $z_h \overset{\text{i.i.d.}}{\sim} \mathrm{Unif}(\mathcal{B}_d(1))$.

Remark C.2. Before beginning any proofs, I will remark on the behavior of this system. First, the true dynamics has three conditions under which the system will go to the absorbing state of 0: first of course being if the state ever hits $s = 0$, second being that if $\|s\| \geq 2$ then $\operatorname{bump}_d(s) = 0$, and third being that if $\|a\| \geq 1/8$, the other $\operatorname{bump}_d(16a) = 0$. When $\|s_h\| \leq 1$ and $\|a_h\| \leq 1/16$ the true dynamics simplifies down to $s_{h+1} = \lambda s_h$, and when $\frac{1}{16} \leq \|a_h\| < \frac{1}{8}$ this only expands to $s_{h+1} = \lambda \operatorname{bump}_d(16a_h) s_h$: since $\lambda < 1$ and $\operatorname{bump}(\cdot) \leq 1$, the magnitude of the state contracts in both settings and so $\|s_{h+1}\| \leq 1$. From here, we can see the reason for $\sigma \geq \frac{1}{8}$: any smaller wouldn't significantly impact the system. Since $\|s_1\| \leq 1$, we can see that $\|s_h\|$ will never actually leave the unit ball; however, $\operatorname{bump}_d(s)$ is still necessary for ensuring that the function is regular.

Lemma C.3. *The setting provided in Setting C.1 satisfies our assumptions of all dynamics and policies being globally Lipschitz and smooth, the reward being Lipschitz and bounded in* $[0,1]$*, and* $(\pi, f^\star)$ *is stable.*

Proof. We first establish that $f^\star$ is smooth and Lipschitz. Let us define the state factor $f_s^\star(s) = s \operatorname{bump}_d(s)$ and the action factor $f_a^\star(a) = \operatorname{bump}_d(16a)$, such that $f^\star(s,a) = \lambda f_a^\star(a) f_s^\star(s)$; note that $\|f_s^\star(s)\| \leq 2$ where $f_s^\star(s) = 0$ whenever $\|s\| \geq 2$ and $\|f_a^\star(a)\| \leq 1$ for all states and actions by Lemma B.4. Taking the derivatives, we see that

$$\nabla f_s^\star(s) = \operatorname{bump}_d(s) I_d + s \nabla \operatorname{bump}_d(s)^\top, \qquad \|\nabla f_s^\star(s)\| \leq 1 + 2c_1$$
$$\nabla f_a^\star(a) = 16 \nabla \operatorname{bump}_d(16a), \qquad \|\nabla f_a^\star(a)\| \leq 16 c_1.$$

As such, $\|\nabla_s f^\star(s,a)\| \leq \lambda \|f_a^\star(a)\| \|\nabla f_s^\star(s)\| \leq \frac{1}{4}$ and $\|\nabla_a f^\star(s,a)\| \leq \lambda \|f_s^\star(s)\| \|\nabla f_a^\star(a)\| \leq 32\lambda c_1 \leq 4$ for all s, a, establishing that $f^\star$ is Lipschitz. We can similarly give that $f^\star$ is smooth by taking the second derivatives, noting that $\|s\| \leq 2$ wherever the derivatives of the state bump are

nonzero.

$$\begin{aligned}
\|\nabla^2 f_s^\star(s)\| &\le 2\|\nabla \operatorname{bump}_d(s)\| + \|s\|\|\nabla^2 \operatorname{bump}_d(s)\| \le 2c_1 + 2c_2 \\
\|\nabla^2 f_a^\star(a)\| &= 2^8\|\nabla^2 \operatorname{bump}_d(16a)\| \le 256c_2 \\
\|\nabla_s \nabla_s f^\star(s,a)\| &= \lambda\|\nabla^2 f_s^\star(s)\|\|f_a^\star(a)\| \le \frac{2c_1 + 2c_2}{4(1+2c_1)} \\
\|\nabla_a \nabla_a f^\star(s,a)\| &= \lambda\|f_s^\star(s)\|\|\nabla^2 f_a^\star(a)\| \le \frac{512c_2}{4(1+2c_1)} = \frac{128c_2}{1+2c_1} \\
\|\nabla_s \nabla_a f^\star(s,a)\| &= \lambda\|\nabla f_s^\star(s)\|\|\nabla f_a^\star(a)\| \le 16\lambda c_1(1+2c_1) \le 4c_1.
\end{aligned}$$

From here, f is also smooth and Lipschitz: this can be seen as it simply adds the affine function $\epsilon\lambda s$ to $f^\star(s,a)$. The policy π is smooth and Lipschitz as a constant policy.

We can see that R is bounded in $[0,1]$, considering how $\min\{2, \max\{0,x\}\} \in [0,2]$ for all x, and we may observe that it is Lipschitz as

$$|\mathrm{R}(s,a) - \mathrm{R}(s',a')| \le \frac{1}{2}|(s-s')^\top e_1| \le \frac{1}{2}\|s - s'\|.$$

Now, we must show that $\pi, f^\star$ is globally EISS. Since $\pi(s) = 0$ for all s, we can roll out action perturbed sequences as $s_{h+1} = f^\star(s_h, u_h)$ for action perturbation u_h at time $h \ge 1$: let us give two trajectories $\{(\tilde{s}_h, u_h)\}_{h\ge1}$ and $\{(\tilde{s}'_h, u'_h)\}_{h\ge1}$ for action perturbation sequences $\{u_h, u'_h\}_{h\ge1}$ and differing initial states $\tilde{s}_1$ and $\tilde{s}'_1$. By the Lipschitz constants we derived earlier, we can see

$$\begin{aligned}
\|\tilde{s}_{h+1} - \tilde{s}'_{h+1}\| &= \|f^\star(\tilde{s}_h, u_h) - f^\star(\tilde{s}'_h, u'_h) \pm f^\star(\tilde{s}_h, u'_h)\| \\
&\le \frac{1}{4}\|\tilde{s}_h - \tilde{s}'_h\| + 4\|u_h - u'_h\| \\
&\le \left(\frac{1}{4}\right)^2 \|\tilde{s}_{h-1} - \tilde{s}'_{h-1}\| + 4\|u_h - u'_h\| + \frac{4}{4}\|u_{h-1} - u'_{h-1}\| \\
&\le \left(\frac{1}{4}\right)^h \|\tilde{s}_1 - \tilde{s}'_1\| + 4\sum_{k=1}^{h}\left(\frac{1}{4}\right)^{h-k}\|u_k - u'_k\|
\end{aligned}$$

This gives us that for $C = \max\{1,4\} = 4$ and $\rho = 1/4$, $(\pi, f^\star)$ is EISS. □

Lemma C.4. *The true system as defined by Setting C.1 will satisfy $s_h = te_1$ for $0 \le t \le 1$ for all $h \ge 1$, where specifically $t = \lambda^{h-1}\left(\prod_{k=1}^{h-1}\operatorname{bump}_d(16a_k)\right)$.*

Proof. This follows from the logic from Remark C.2. We can prove this rigorously through induction: this is satisfied at $h = 1$ for $t = 1$, and if this is satisfied at some timestep h we can see that:

$$s_{h+1} = \lambda \operatorname{bump}_d(16a_h)s_h \operatorname{bump}_d(s_h) = \lambda \operatorname{bump}_d(16a_h)te_1.$$

Since $0 \le \lambda \operatorname{bump}_d(16a_h) \le 1$, we have that $0 \le \lambda \operatorname{bump}_d(16a_h)t \le 1$. This necessarily means that for all $h \ge 1$, we have $\operatorname{bump}_d(s_h) = 1$ and so for any actions a_h, we have $s_{h+1} = \lambda \operatorname{bump}_d(16a_h)s_h = \lambda^h\left(\prod_{k=1}^{h}\operatorname{bump}_d(16a_k)\right)e_1$. □

Proposition C.5 (Probabilistic coverage can deteriorate exponentially in horizon)**.** *Let us take the setting defined in Setting C.1. Let $\frac{1}{8} \le \sigma_1 < \sigma_2$, and let $\{\tilde{\nu}_h^{\sigma_i}\}_{i\in\{1,2\}}$ be the corresponding action-noised expert state-action distributions. Let $\mathcal{J}$ be the class of all nonnegative measurable functions with finite expectations, satisfying $J(s,a) = J(s,0)$ and $J(0,a) = 0$ for all $s \in \mathcal{S}$, $a \in \mathcal{A}$, $J \in \mathcal{J}$. For every $J \in \mathcal{J}$, for all $h \ge 1$,*

$$\mathop{\mathbb{E}}_{\tilde{\nu}_h^{\sigma_1}}[J(s_h,a_h)] = \left(\frac{\sigma_2}{\sigma_1}\right)^{d(h-1)} \mathop{\mathbb{E}}_{\tilde{\nu}_h^{\sigma_2}}[J(s_h,a_h)].$$

Proof. Let $\frac{1}{8} \leq \sigma_1 < \sigma_2$.

Let us consider the general distribution $\tilde{\nu}_h^\sigma$ for $\sigma \geq 1/8$.

Noting that $\mathrm{bump}_d(16a) > 0$ if and only if $\|a\| < 1/8$ and $\mathrm{bump}_d(\cdot) \in [0,1]$, $s_h \neq 0$ if and only if $\{\sigma z_k\}_{k=1}^{h-1}$ are all smaller in magnitude than $\frac{1}{8}$: this gives that since $\mathbb{P}[\|\sigma z_k\| \leq 1/8] = (8\sigma)^{-d}$, we have that

$$\mathbb{P}_{\tilde{\nu}_h^\sigma}[s_h \neq 0] = \left(\frac{1}{8\sigma}\right)^{d(h-1)}$$

and conditional on $s_h \neq 0$, we can say that the actions a_k for $k < h$ were generated from a uniform ball scaled by $1/8$.

Let us now fix some function $J \in \mathcal{J}$ with finite expectations. For $h \geq 2$, it must hold that

$$\mathop{\mathbb{E}}_{\tilde{\nu}_h^\sigma}[J(s_h, a_h) \mid s_h \neq 0] = \mathop{\mathbb{E}}_{\mathrm{Unif}(\mathcal{B}_d(1))}\left[J\left(\lambda^{h-1}\left(\prod_{k=1}^{h-1}\mathrm{bump}_d\left(\frac{16}{8}z_k\right)\right)e_1, 0\right)\right]$$

where each sample of z_k are unconditional IID draws from $\mathrm{Unif}(\mathcal{B}_d(1))$. Notably, the righthandside of this equation is independent of σ: this makes sense, as σ mainly serves to increase the probability that the absorbing state is achieved. If we know that it wasn't, then the uniform noise distribution is identical within $\mathcal{B}_d(1/8)$. This means that $\mathop{\mathbb{E}}_{\tilde{\nu}_h^{\sigma_1}}[J(s_h, a_h) \mid s_h \neq 0] = \mathop{\mathbb{E}}_{\tilde{\nu}_h^{\sigma_2}}[J(s_h, a_h) \mid s_h \neq 0]$ and for $h = 1$ both sides would equal $J(e_1, 0)$. Now expanding from the unconditional expectation, and noting that $J(0, a) = 0$, we can see that

$$\begin{aligned}
\mathop{\mathbb{E}}_{\tilde{\nu}_h^{\sigma_1}}[J(s_h, a_h)] &= \mathbb{P}_{\tilde{\nu}_h^{\sigma_1}}[s_h \neq 0] \mathop{\mathbb{E}}_{\tilde{\nu}_h^{\sigma_1}}[J(s_h, a_h) \mid s_h \neq 0] + J(0,0)\mathbb{P}_{\tilde{\nu}_h^{\sigma_1}}[s_h = 0] \\
&= \left(\frac{1}{8\sigma_1}\right)^{d(h-1)} \mathop{\mathbb{E}}_{\tilde{\nu}_h^{\sigma_2}}[J(s_h, a_h) \mid s_h \neq 0] \\
&= \left(\frac{1}{8\sigma_1}\right)^{d(h-1)} \left(\frac{1}{8\sigma_2}\right)^{-d(h-1)} \mathop{\mathbb{E}}_{\tilde{\nu}_h^{\sigma_2}}[J(s_h, a_h)] \\
&= \left(\frac{\sigma_2}{\sigma_1}\right)^{d(h-1)} \mathop{\mathbb{E}}_{\tilde{\nu}_h^{\sigma_2}}[J(s_h, a_h)].
\end{aligned}$$

□

Corollary C.6. *Let us take the setting defined in Setting C.1, but take π to be the deterministic expert $\pi^\star$. For $\sigma \geq 1/8$, let $\tilde{\nu}_h$ noise injected distribution defined in Definition 4.10. We refine the class of functions $\mathcal{J}$ defined in Proposition C.5 with the condition that there exists some $C \geq 0$ such that $J(te_1, a) = Ct^2$ for all $t \in [0,1]$ and all $a \in \mathcal{A}$. We define this class $\mathcal{J}' \subseteq \mathcal{J}$. It holds that, for every $J \in \mathcal{J}'$, for all $h \geq 1$,*

$$\mathop{\mathbb{E}}_{\mathbb{P}_h^\star}[J(s_h, a_h)] \geq (8\sigma)^{d(h-1)} \mathop{\mathbb{E}}_{\tilde{\nu}_h^\sigma}[J(s_h, a_h)].$$

Proof. Since the deterministic expert $\pi^\star(s) = 0$ would satisfy $\mathrm{bump}_d(16a_h) = 1$ every timestep, Lemma C.4 tells us that the induced true system reduces to $s_h^{\pi^\star, f^\star} = \lambda^{h-1}e_1$ and $\mathbb{E}_{\mathbb{P}_h^\star}\|s_h\|^2 = \lambda^{2(h-1)}$ whereas under $\tilde{\nu}_h$ for $\{z_k\}_{k=1}^{h-1}$ we have

$$\tilde{s}_h = \lambda^{h-1}\left(\prod_{k=1}^{h-1}\mathrm{bump}_d(16\sigma z_k)\right)e_1$$

$$\mathop{\mathbb{E}}_{\tilde{\nu}_h}\|s_h\|^2 = \lambda^{2(h-1)}\left(\mathop{\mathbb{E}}_{z\sim\mathrm{Unif}(\mathcal{B}_d(1))}\left[\mathrm{bump}(16\sigma z)^2\right]\right)^{h-1}$$

For all J in our refined $\mathcal{J}'$, since $s_h^{\pi^\star, f^\star}$ and $\tilde{s}_h$ all exist along e_1, we can see

$$\frac{\mathbb{E}_{\tilde{\nu}_h}[J(s,a)]}{\mathbb{E}_{\mathbb{P}_h^\star}[J(s,a)]} = \left(\mathop{\mathbb{E}}_{z\sim\mathrm{Unif}(\mathcal{B}_d(1))}\left[\mathrm{bump}(16\sigma z)^2\right]\right)^{h-1}$$

For $\sigma \geq 1/8$, by Lemma B.4 we can see

$$(16\sigma)^{-d} = \mathbb{P}\left[\|\sigma z\| \leq \frac{1}{16}\right] \leq \mathop{\mathbb{E}}_{z\sim \mathrm{Unif}(\mathcal{B}_d(1))}\left[\mathrm{bump}_d(16\sigma z)^2\right] \leq \mathbb{P}\left[\|\sigma z\| < \frac{1}{8}\right] = (8\sigma)^{-d}$$

so combining gives us our result. □

Definition C.7. For the following lemma, we define the candidate rollout Q-function as $Q^{\pi,f}_{H,H}(s,a) \triangleq \mathrm{R}(s,a)$ and

$$Q^{\pi,f}_{h,H}(s,a) \triangleq \sum_{k=h}^{H} \mathrm{R}(s^{\pi,f}_k, a^{\pi,f}_k) \quad \text{for } 1 \leq h < H,\, H \geq 2$$

where $s^{\pi,f}_h = s$ and $a^{\pi,f}_h = a$ but for $h < k \leq H$ we have $s^{\pi,f}_k = f(s^{\pi,f}_{k-1}, a^{\pi,f}_{k-1})$ and $a^{\pi,f}_k = \pi(s^{\pi,f}_k)$. We also define the Bellman operator $\mathbf{T}^\pi_h$ such that for $h < H$,

$$\mathbf{T}^\pi_h Q^{\pi,f}_{h+1,H}(s,a) \triangleq \mathrm{R}(s,a) + Q^{\pi,f}_{h+1,H}\left(f^\star(s,a), \pi(f^\star(s,a))\right).$$

Finally, we define the RMS Bellman error for $h < H$ as:

$$J_{\mathrm{Bellman}}(\pi, f; \nu_h) \triangleq \sqrt{\mathop{\mathbb{E}}_{s,a\sim\nu_h}\left(Q^{\pi,f}_{h,H}(s,a) - \mathbf{T}^\pi_h Q^{\pi,f}_{h+1,H}(s,a)\right)^2}.$$

Lemma C.8. *The pointwise squared model error and squared Bellman error induced by candidate dynamics function f_ϵ and reward function R from Setting C.1 are elements of function class $\mathcal{J}'$ defined in Corollary C.6, and thus are necessarily elements of $\mathcal{J}$ defined in Proposition C.5. Specifically, squared model error satisfies $C = \epsilon^2\lambda^2$ and squared Bellman error satisfies $C = \left(\frac{\epsilon\lambda}{2}\sum_{k=0}^{H-h-1}(1+\epsilon)^k\lambda^k\right)^2$.*

Proof. For model error, we can see that $J(s,a) = \|f_\epsilon(s,a) - f^\star(s,a)\|^2 = \epsilon^2\lambda^2\|s\|^2$ satisfies the properties we desire.

Similarly, the Bellman error will satisfy

$$Q^{\pi,f_\epsilon}_{h,H}(s,a) - \mathbf{T}^\pi_h Q^{\pi,f_\epsilon}_{h+1,H}(s,a) = Q^{\pi,f_\epsilon}_{h+1,H}(f(s,a),0) - Q^{\pi,f_\epsilon}_{h+1,H}(f^\star(s,a),0).$$

For $a = 0$ and $\|s\| \leq 1$, the candidate dynamics will progress as $f_\epsilon(s,0) = \lambda s + \epsilon\lambda s = (1+\epsilon)\lambda s$, and since $\lambda \leq 1/4$ this will still contract and retain the property of future states having magnitude bounded by 1. We can use this to expand out the Q-function as $Q^{\pi,f_\epsilon}_{h+1,H}(s,0) = \frac{\max\{0, s^\top e_1\}}{2}\sum_{k=0}^{H-h-1}(1+\epsilon)^k\lambda^k$ for $\|s\| \leq 1$.

In order to show that squared Bellman error is an element of $\mathcal{J}$, we must show that it satisfies the properties for all state action pairs. For $\|s\| \leq 2$, $\|f^\star(s,a)\| \leq \lambda\|s\| \leq 1/2$ and $\|f_\epsilon(s,a)\| \leq (1+\epsilon)\lambda\|s\| \leq 1$ so for $\pi(s) = 0$ we get

$$Q^{\pi,f_\epsilon}_{h+1,H}(f_\epsilon(s,a),0) - Q^{\pi,f_\epsilon}_{h+1,H}(f^\star(s,a),0) = \epsilon\lambda\frac{\max\{0, s^\top e_1\}}{2}\sum_{k=0}^{H-h-1}(1+\epsilon)^k\lambda^k = Q^{\pi,f_\epsilon}_{h+1,H}(\epsilon\lambda s, 0).$$

For $\|s\| > 2$, $f^\star(s,a) = 0$ and $f_\epsilon(s,a) = \epsilon\lambda s$ so $Q^{\pi,f_\epsilon}_{h+1,H}(f^\star(s,a),0) = 0$ and

$$Q^{\pi,f_\epsilon}_{h+1,H}(f_\epsilon(s,a),0) - Q^{\pi,f_\epsilon}_{h+1,H}(f^\star(s,a),0) = Q^{\pi,f_\epsilon}_{h+1,H}(\epsilon\lambda s, 0).$$

Putting these together, we can see that for all state-action pairs,

$$Q^{\pi,f_\epsilon}_{h,H}(s,a) - \mathbf{T}^\pi_h Q^{\pi,f_\epsilon}_{h+1,H}(s,a) = Q^{\pi,f_\epsilon}_{h+1,H}(\epsilon\lambda s, 0)$$

and so if we define $J(s,a) \triangleq \left(Q^{\pi,f_\epsilon}_{h,H}(s,a) - \mathbf{T}^\pi_h Q^{\pi,f_\epsilon}_{h+1,H}(s,a)\right)^2$ we have that J is nonnegative, independent of the action, $J(0,a) = 0$, and $0 \leq J(s,a) \leq (H-h)^2$. Furthermore, we can see that for $t \in [0,1]$, and $a \in \mathcal{A}$, $C = \left(\frac{\epsilon\lambda}{2}\sum_{k=0}^{H-h-1}(1+\epsilon)^k\lambda^k\right)^2$, $J(te_1, a) = \left(Q^{\pi,f_\epsilon}_{h+1,H}(\epsilon\lambda t e_1, 0)\right)^2 = Ct^2$. As such, $J \in \mathcal{J}'$. □

Lemma C.9. *In the setting defined in Setting C.1, the reward difference $|J_H(\pi^\star) - Q_H^{\pi^\star, f_\epsilon}|$ satisfies*

$$|J_H(\pi^\star) - Q_H^{\pi^\star, f_\epsilon}| \lesssim \max_{h<H} \mathrm{J}_{\mathrm{model}}(f_\epsilon; \tilde{\nu}_h), \qquad |J_H(\pi^\star) - Q_H^{\pi^\star, f_\epsilon}| \lesssim \max_{h<H} J_{\mathrm{Bellman}}(\pi^\star, f_\epsilon; \tilde{\nu}_h)$$

Proof. We first note the true and candidate trajectories $s_h^{\pi^\star, f^\star} = \lambda^{h-1} e_1$ and $s_h^{\pi^\star, f_\epsilon} = \lambda^{h-1}(1+\epsilon)^{h-1} e_1$, where both start at e_1 and satisfy te_1 for some $t \in (0, 1]$ for every timestep h since $\lambda < 1/2$, resulting in $\mathrm{bump}_d(s_h) = 1$. Both trajectories additionally never trigger the max or min in R, resulting in cumulative reward difference:

$$\begin{aligned} Q_H^{\pi^\star, f_\epsilon} - J_H(\pi^\star) &= \frac{1}{2} \sum_{h=1}^{H} \left(\lambda^{h-1}(1+\epsilon)^{h-1} - \lambda^{h-1} \right) \\ &\leq \frac{1}{2} \left(\frac{1 - \lambda - 1 + \lambda(1+\epsilon)}{(1 - \lambda(1+\epsilon))(1-\lambda)} \right) \\ &= \frac{\lambda\epsilon}{2(1-\lambda)(1-\lambda-\lambda\epsilon)} \\ &\leq \frac{\lambda\epsilon}{2(1-\lambda)(1-2\lambda)} \end{aligned}$$

By Lemma C.8 and Lemma C.4, we have that both squared model error and squared Bellman error satisfy $J(s,a) = C\|s\|^2$ for all reachable s (all of whom satisfy $\|s\| \leq 1$), where $C = \lambda^2\epsilon^2$ for model error and $C = \left(\frac{\epsilon\lambda}{2} \sum_{k=0}^{H-h-1} (1+\epsilon)^k \lambda^k \right)^2$ for Bellman. Translating this into $\mathrm{J}_{\mathrm{model}}$ and J_{Bellman}, noting that $s_1 = e_1$,

$$\begin{aligned} \max_{h<H} \mathrm{J}_{\mathrm{model}}(f_\epsilon; \tilde{\nu}_h^\sigma) &= \lambda\epsilon, \\ \max_{h<H} J_{\mathrm{Bellman}}(\pi^\star, f_\epsilon; \tilde{\nu}_h^\sigma) &= \frac{\epsilon\lambda}{2} \sum_{k=0}^{H-h-1} (1+\epsilon)^k \lambda^k > \frac{\epsilon\lambda}{2} \sum_{k=0}^{H-h-1} \lambda^k. \end{aligned}$$

Therefore, both of these scale with the cumulative reward difference. □

Corollary C.10 (Concentrability is bad in benign cases)**.** *Let us take the setting defined in Setting C.1, and let us fix arbitrary $\sigma > 1/8$. Let $\pi^\star \triangleq \pi$, and let us take $\mu_h \triangleq \tilde{\nu}_h^\sigma$. For all $h \geq 1$,*

$$\begin{aligned} \mathrm{J}_{\mathrm{model}}(f_\epsilon; \mathbb{P}_h^\star) &\geq \left(\sqrt{8\sigma} \right)^{d(h-1)} \mathrm{J}_{\mathrm{model}}(f_\epsilon; \mu_h), \\ J_{\mathrm{Bellman}}(\pi^\star, f_\epsilon; \mathbb{P}_h^\star) &\geq \left(\sqrt{8\sigma} \right)^{d(h-1)} J_{\mathrm{Bellman}}(\pi^\star, f_\epsilon; \mu_h). \end{aligned}$$

Simultaneously, we have that

$$|J_H(\pi^\star) - Q_H^{\pi^\star, f_\epsilon}| \lesssim \max_{h<H} \mathrm{J}_{\mathrm{model}}(f_\epsilon; \mu_h), \quad |J_H(\pi^\star) - Q_H^{\pi^\star, f_\epsilon}| \lesssim \max_{h<H} J_{\mathrm{Bellman}}(\pi^\star, f_\epsilon; \mu_h).$$

Proof. The proof follows by combining the conclusions of Lemma C.8, Corollary C.6, and Lemma C.9. □

In order to support our claim that maximal coverage, as defined through concentrability, can still lead to intractable cases, we reference the results from Simchowitz et al. (2025, Theorem 4). In the imitation learning setting that they describe, the considered dataset is solely composed of expert trajectories $\mu_h = \mathbb{P}_h^\star$: despite this, they show that when the true dynamics are unstable in open-loop, exponential compounding error occurs regardless of policy class or algorithm used.

C.1 Towards a Geometric Notion of Coverage: Wasserstein and Stability

This motivates considering geometric notions of distributional proximity. While those probabilistic notions required absolute continuity by definition, geometrically close state-action pairs suffice under regularity conditions. Wasserstein distance provides a complementary comparison that respects

the geometry of the state-action space: because of our Lipschitz assumption on the dynamics functions, $\mathrm{J}_{\mathrm{model}}$ is necessarily $2L_f$-Lipschitz and satisfies, for any two joint state-action distributions ξ_h and ν_h,

$$\mathrm{J}_{\mathrm{model}}(f;\xi_h) \le \mathrm{J}_{\mathrm{model}}(f;\nu_h) + 2L_f W_2(\xi_h,\nu_h).$$

This geometric mismatch is additive, unlike the multiplicative probabilistic coverage definitions; however, it can cleanly handle settings with support mismatches and it doesn't suffer an exponential in horizon lowerbound in benign settings as we saw for the probabilistic notions. In fact, any distribution generated from actions that are second-moment bounded from the expert will have a constant W_2 distance from the expert distribution: no absolute bound, zero-mean, independence, or measure-domination assumptions necessary.

Lemma C.11.

$$\max_{k\le h} \mathop{\mathbb{E}}_{(s_k,a_k)\sim\mu_k} \|a_k - \pi^\star(s_k)\|^2 \le \sigma^2 \implies W_2\left(\mathbb{P}_h^\star, \mu_h\right) \lesssim \sigma$$

Proof. Let me denote $(s_h,a_h)\sim\mu_h$ and $(s_h^\star,a_h^\star)\sim\mathbb{P}_h^\star$. By the stability of $(\pi^\star,f^\star)$, we have the following.

$$\begin{aligned}
\sqrt{\mathop{\mathbb{E}}_{\mu_h,\mathbb{P}_h^\star}[\|s_h - s_h^\star\|^2]} &\le \sqrt{\mathop{\mathbb{E}}_{\mu_h,\mathbb{P}_h^\star}\left[\left(C\rho^{h-1}\|s_1 - s_1^\star\| + \sum_{k=1}^{h-1} C\rho^{h-1-k}\|a_k - \pi^\star(s_k)\|\right)^2\right]} \\
&\le \sqrt{2C^2 \mathop{\mathbb{E}}_{\mu_h,\mathbb{P}_h^\star}\left[\sum_{k=1}^{h-1}\sum_{j=1}^{h-1} \rho^{h-1-k}\rho^{h-1-j}\|a_k - \pi^\star(s_k)\|\|a_j - \pi^\star(s_j)\|\right]} \\
&\le C\sqrt{2\sum_{k=1}^{h-1}\sum_{j=1}^{h-1} \rho^{h-1-k}\rho^{h-1-j}\sqrt{\mathop{\mathbb{E}}_{\mu_k}[\|a_k - \pi^\star(s_k)\|^2]\mathop{\mathbb{E}}_{\mu_j}[\|a_j - \pi^\star(s_j)\|^2]}} \\
&\le \sqrt{2}C\sum_{k=1}^{h-1}\rho^{h-1-k}\sqrt{\mathop{\mathbb{E}}_{\mu_k}[\|a_k - \pi^\star(s_k)\|^2]} \\
&\le \sqrt{2}C_{\mathrm{stab}}\max_{k\le h-1}\sqrt{\mathop{\mathbb{E}}_{\mu_k}[\|a_k - \pi^\star(s_k)\|^2]} \\
&\le \sqrt{2}C_{\mathrm{stab}}\sigma
\end{aligned}$$

This may now be applied to the action difference, making use of $\pi^\star$ being L_π-Lipschitz.

$$\begin{aligned}
\sqrt{\mathop{\mathbb{E}}_{\mu_h,\mathbb{P}_h^\star}[\|a_h - a_h^\star\|^2]} &\le \sqrt{2\mathop{\mathbb{E}}_{\mu_h}[\|a_h - \pi^\star(s_h)\|^2] + 2\mathop{\mathbb{E}}_{\mu_h,\mathbb{P}_h^\star}[\|\pi^\star(s_h) - a_h^\star\|^2]} \\
&\le \sqrt{2}\sigma + 2L_\pi C_{\mathrm{stab}}\sigma
\end{aligned}$$

Finally putting this all together, we can see:

$$W_2(\mathbb{P}_h^\star,\mu_h) \le \sigma\sqrt{2C_{\mathrm{stab}}^2 + \left(\sqrt{2} + 2L_\pi C_{\mathrm{stab}}\right)^2} \lesssim \sigma.$$

□

It also follows that Wasserstein between two noise injected distributions is similarly bounded.

Lemma C.12. *Given a (C,ρ) EISS system, the Wasserstein distance between two action-noise injected distributions is bounded with*

$$W_2(\tilde{\nu}_h^{\sigma_1},\tilde{\nu}_h^{\sigma_2}) \lesssim |\sigma_1 - \sigma_2|$$

Proof. Taking that (π, f) is EISS, and given an initial state $s_1 \sim d_1$ and noise sequence $\{z_k\}_{k=1}^h$,

$$\|\tilde{s}_h^{\sigma_1} - \tilde{s}_h^{\sigma_2}\| \leq C\rho^{h-1}\|s_1 - s_1\| + \sum_{k=1}^{h-1} C\rho^{h-1-k}\|\sigma_1 z_k - \sigma_2 z_k\| \leq |\sigma_1 - \sigma_2| \sum_{k=1}^{h-1} C\rho^{h-1-k}\|z_k\|$$

$$\|\tilde{a}_h^{\sigma_1} - \tilde{a}_h^{\sigma_2}\| \leq \|\pi(\tilde{s}_h^{\sigma_1}) - \pi(\tilde{s}_h^{\sigma_2})\| + \|\sigma_1 z_h - \sigma_2 z_h\| \leq |\sigma_1 - \sigma_2|\|z_h\| + L_\pi|\sigma_1 - \sigma_2| \sum_{k=1}^{h-1} C\rho^{h-1-k}\|z_k\|$$

This gives that $W_2(\tilde{\nu}_h^{\sigma_1}, \tilde{\nu}_h^{\sigma_2}) \leq \sqrt{(C_{\text{stab}}^2 + (1 + L_\pi C_{\text{stab}})^2)\mathbb{E}_z\|z\|^2}|\sigma_1 - \sigma_2|$. □

Lemma C.13. *The Wasserstein distance between any two rollout distributions induced by stable systems will be uniformly bounded in horizon.*

Proof. Let us take two systems (π, f) and (π', f') that are both (C, ρ) EISS. Since we aren't considering perturbations, let us simplify these closed-loop systems to F and F' for short, such that each is a state to state map. Analyzing the first system (but noting that this analysis is shared between the two):

$$\|s_{h+1} - s_h\| = \|F^{h-1}(F(s_1)) - F^{h-1}(s_1)\| \leq C\rho^{h-1}\|F(s_1) - s_1\|$$
$$\|s_h - s_1\| \leq \sum_{k=1}^{h-1}\|s_{k+1} - s_k\| \leq \sum_{k=1}^{h-1} C\rho^{k-1}\|F(s_1) - s_1\| \leq C_{\text{stab}}\|F(s_1) - s_1\|$$

If we subtract the final state of the two systems, we get

$$\|s_h - s'_h\| \leq C_{\text{stab}}\left(\|F(s_1) - s_1\| + \|F'(s_1) - s_1\|\right)$$

We can clearly see no dependency on horizon.

$$\begin{aligned}\|\pi(s_h) - \pi'(s'_h)\| &\leq \|\pi(s_h) - \pi(s_1)\| + \|\pi'(s'_1) - \pi'(s'_h)\| + \|\pi(s_1) - \pi'(s'_1)\| \\ &\leq L_\pi C_{\text{stab}}\left(\|F(s_1) - s_1\| + \|F'(s_1) - s_1\|\right) + \|(\pi - \pi')(s_1)\|\end{aligned}$$

Once again, this is all simply dependent on the draw of initial state s_1, with no dependency on horizon.

$$\begin{aligned}\|(s_h, a_h) - (s'_h, a'_h)\| &\leq \|(s_h, a_h) - (s_1, a_1)\| + \|a_1 - a'_1\| + \|(s_1, a'_1) - (s'_h, a'_h)\| \\ &\leq \sqrt{1 + L_\pi^2} C_{\text{stab}}\left(\|F(s_1) - s_1\| + \|F'(s_1) - s_1\|\right) + \|(\pi - \pi')(s_1)\|\end{aligned}$$

This provides that the Wasserstein distance induced by the two rollout distributions must be uniformly bounded across timesteps. □

Critically, Wasserstein distance isn't a universal balm: while it remains bounded for stable policies, unstable settings still allow it to grow exponentially in horizon.

Lemma C.14. *The W_2 distance between a distribution generated by a stable system and a distribution generated by an unstable system can grow exponentially in horizon. This holds regardless of if there is a policy difference or a dynamics difference between the two systems.*

Proof. For any initial state distribution besides the point mass at 0, define the linear closed-loop system of $s_{h+1} = Cs_h$, and let us take the two systems as $C > 1$ and $C = \rho < 1$.

Let's take the case that both systems have the same policy of $\pi(s) = 0$, but differ in dynamics $f^C(s, a) = Cs$ and $f^\rho(s, a) = \rho s$. The second is stable, as initial state perturbations contract with ρ every timestep, and action perturbations are forgotten immediately.

$$\begin{aligned}W_2(\nu_h^C, \nu_h^\rho) &= (C^{h-1} - \rho^{h-1})\sqrt{\mathbb{E}_{s_1 \sim d_1}\|s_1\|^2} \\ &= C^{h-1}\left(1 - \left(\frac{\rho}{C}\right)^{h-1}\right)\sqrt{\mathbb{E}_{s_1 \sim d_1}\|s_1\|^2} \\ &\geq C^h\left(\frac{1}{C} - \frac{\rho}{C^2}\right)\sqrt{\mathbb{E}_{s_1 \sim d_1}\|s_1\|^2}\end{aligned}$$

Now, let's take the case that both systems have the same dynamics of $f(s,a) = Cs + a$, and differ in policies $\pi^C(s) = 0$ and $\pi^\rho(s) = (\rho - C)s$. The second is stable, as initial state and action perturbations contract with ρ every timestep.

$$\begin{aligned} W_2(\nu_h^C, \nu_h^\rho) &= \sqrt{\mathop{\mathbb{E}}_{s_1 \sim d_1}\left[\|s_h^C - s_h^\rho\|^2 + \|a_h^C - a_h^\rho\|^2\right]} \\ &\geq \sqrt{\mathop{\mathbb{E}}_{s_1 \sim d_1}\left[\|s_h^C - s_h^\rho\|^2\right]} \\ &= (C^{h-1} - \rho^{h-1})\sqrt{\mathop{\mathbb{E}}_{s_1 \sim d_1}\|s_1\|^2} \\ &\geq C^h\left(\frac{1}{C} - \frac{\rho}{C^2}\right)\sqrt{\mathop{\mathbb{E}}_{s_1 \sim d_1}\|s_1\|^2} \end{aligned}$$

□

This motivates our focus on guarantee transfers from $\mathbb{P}_h^\star$ to μ_h as opposed to allowing any $(\pi, f, d_1)_h$ system, which could hide exponential growth in horizon unless we can guarantee that the system is stable.

Obviously, this could be made tighter though: when it holds that $f \approx f^\star$, even a degenerate expert distribution for which $\mathcal{C}_h((\pi^\star, f^\star, d_1)_h)$ is functionally infinite ought not suffer a constant penalty to transfer the guarantee to the dataset distribution. A more careful expansion of the transfer preserves the model error dependence on the Wasserstein penalty

Definition C.15. Let $\Gamma(\xi_h, \nu_h)$ denote the set of couplings between joint state-action probability distributions ξ_h and ν_h, such that for coupling $\gamma \in \Gamma(\xi_h, \nu_h)$ the joint draw $((s,a),\ (s',a')) \sim \gamma$ satisfies $(s,a) \sim \xi_h$ and $(s',a') \sim \nu_h$, and where all such couplings satisfy $\mathop{\mathbb{E}}_{((s,a),\ (s',a'))\sim\gamma}[\|(s,a) - (s',a')\|^2] = W_2(\xi_h, \nu_h)^2$. When $W_2(\xi_h, \nu_h) > 0$, we define the transport weighted residual distribution

$$(\xi_h \setminus_\gamma \nu_h)(A) \triangleq \frac{\mathop{\mathbb{E}}_{((s,a),\ (s',a'))\sim\gamma}[\|(s,a) - (s',a')\|\mathbb{1}[(s,a) \in A]]}{\mathop{\mathbb{E}}_{((s,a),\ (s',a'))\sim\gamma}[\|(s,a) - (s',a')\|]}.$$

for every measurable $A \subseteq \mathcal{S} \times \mathcal{A}$. When $W_2(\xi_h, \nu_h) = 0$, we define $(\xi_h \setminus_\gamma \nu_h)$ to be the zero measure, and instead for any function g we set $\mathop{\mathbb{E}}_{(\xi_h \setminus_\gamma \nu_h)}[g(s,a)] = 0$.

This distribution $(\xi_h \setminus_\gamma \nu_h)$ is obtained by reweighting the state-action pairs under ξ_h according to the distance through which their mass is transported under the selected W_2-optimal coupling of γ. As such, mass that is not transported under the selected W_2-optimal coupling will not contribute to the distribution.

We propose a family of auxiliary joint state-action distributions ν_h: these distributions bear no necessary generative assumptions; instead, they pose some distribution that can mediate between the geometric proximity to the target distribution and the probabilistic coverage to the dataset distribution, allowing us to optimize over the tradeoffs.

Definition C.16. [Coverage transfer constant] For any joint state-action probability distribution ν_h, we define

$$\mathcal{C}_h(\nu_h) \triangleq \inf\left\{c \geq 1 \ :\ \mathop{\mathbb{E}}_{\nu_h}\|(f - f^\star)(s_h,a_h)\|^2 \leq c\mathop{\mathbb{E}}_{\mu_h}\|(f - f^\star)(s_h,a_h)\|^2 \quad \forall f \in \mathcal{F}\right\}$$

Lemma C.17. *From desired joint state-action probability distribution ξ_h, we can generally say that for all $f \in \mathcal{F}$, we can optimize over interpolant distributions ν_h to obtain the optimal trade-off between probabilistic coverage $\mathcal{C}$ between the interpolant and the dataset distribution and the geometric closeness W_2 between ξ_h and the interpolant to get:*

$$\begin{aligned} \mathrm{J}_{\mathrm{model}}(f;\xi_h) \leq \inf_{\nu_h} \mathcal{C}_h(\nu_h)^{1/2}\,\mathrm{J}_{\mathrm{model}}(f;\mu_h) + \sqrt{2L_f W_2(\xi_h,\nu_h)}\mathcal{C}_h(\nu_h)^{1/4}\,\mathrm{J}_{\mathrm{model}}(f;\mu_h)^{1/2} \\ + \inf_{\gamma\in\Gamma(\xi_h,\nu_h)}\sqrt{2L_f W_2(\xi_h,\nu_h)}\,\mathrm{J}_{\mathrm{model}}(f;(\xi_h \setminus_\gamma \nu_h))^{1/2} \end{aligned}$$

for $\Gamma(\xi_h, \nu_h)$ and $(\xi_h \setminus_\gamma \nu_h)$ defined in Definition C.15.

Proof. Fix any interpolant distribution ν_h and any W_2-optimal coupling $\gamma \in \Gamma(\xi_h, \nu_h)$ such that the joint draw $((s,a),\ (s',a')) \sim \gamma$ satisfies $(s,a) \sim \xi_h$ and $(s',a') \sim \nu_h$.

$$
\begin{aligned}
&\sqrt{\mathop{\mathbb{E}}_{\xi_h} \|(f - f^\star)(s_h, a_h)\|^2} \\
&\quad \le \inf_{\nu_h, \gamma \in \Gamma(\xi_h, \nu_h)} \sqrt{\mathop{\mathbb{E}}_{\nu_h} \|(f - f^\star)(s_h, a_h)\|^2} \\
&\qquad + \sqrt{\mathop{\mathbb{E}}_{((s,a),\ (s',a'))\sim\gamma} \|(f - f^\star)(s,a) - (f - f^\star)(s',a')\|^2} \\
&\quad \le \inf_{\nu_h, \gamma \in \Gamma(\xi_h, \nu_h)} \sqrt{\mathcal{C}_h(\nu_h)}\, \mathrm{J}_{\mathrm{model}}(f; \mu_h) \\
&\qquad + \sqrt{2L_f \mathop{\mathbb{E}}_{((s,a),\ (s',a'))\sim\gamma} [\|(s,a) - (s',a')\| \|(f - f^\star)(s,a)\|]} \\
&\qquad + \sqrt{2L_f \mathop{\mathbb{E}}_{((s,a),\ (s',a'))\sim\gamma} [\|(s,a) - (s',a')\| \|(f - f^\star)(s',a')\|]} \\
&\quad \le \inf_{\nu_h, \gamma \in \Gamma(\xi_h, \nu_h)} \sqrt{\mathcal{C}_h(\nu_h)}\, \mathrm{J}_{\mathrm{model}}(f; \mu_h) + \sqrt{2L_f W_2(\xi_h, \nu_h) \sqrt{\mathcal{C}_h(\nu_h)}\, \mathrm{J}_{\mathrm{model}}(f; \mu_h)} \\
&\qquad + \sqrt{2L_f W_2(\xi_h, \nu_h) \left(\frac{\mathop{\mathbb{E}}_{((s,a),\ (s',a'))\sim\gamma} [\|(s,a) - (s',a')\| \|(f - f^\star)(s,a)\|]}{\mathop{\mathbb{E}}_{((s,a),\ (s',a'))\sim\gamma} [\|(s,a) - (s',a')\|]} \right)}
\end{aligned}
$$

We can recognize that the part of the third term multiplied by $2L_f W_2(\xi_h, \nu_h)$ represents the model error on our desired distribution ξ_h, but weighted on the transport distance of each point to ν_h (or weighted on the transport distance of each point to a place where it could be covered by μ_h, if we incorporate the infimums into our interpretation). As such, we can give our final result:

$$
\begin{aligned}
&\sqrt{\mathop{\mathbb{E}}_{\xi_h} \|(f - f^\star)(s_h, a_h)\|^2} \\
&\quad \le \inf_{\nu_h} \mathcal{C}_h(\nu_h)^{1/2}\, \mathrm{J}_{\mathrm{model}}(f; \mu_h) + \sqrt{2L_f W_2(\xi_h, \nu_h)} \mathcal{C}_h(\nu_h)^{1/4}\, \mathrm{J}_{\mathrm{model}}(f; \mu_h)^{1/2} \\
&\qquad + \sqrt{2L_f W_2(\xi_h, \nu_h)}\, \mathrm{J}_{\mathrm{model}}(f; (\xi_h \setminus_\gamma \nu_h))^{1/2}
\end{aligned}
$$

□

D When Offline RL Fails

We begin our analysis of the difficulties of our setting with a minimal illustrative setting sufficient to demonstrate exponential policy evaluation. Following this, we will provide a more involved construction that demonstrates how exponential policy evaluation allows policies to be misranked, disproving any attempt at a uniform polynomial in horizon upperbound without additional assumptions.

Setting D.1 (Minimal adversarial setting)**.** Let us consider $\mathcal{S} \subseteq \mathbb{R}^2$, $\mathcal{A} \subseteq \mathbb{R}$, with initial state distribution d_1 being any distribution supported on $\mathrm{span}(e_1)$. For some fixed $C > 1$ and any $\epsilon > 0$, we set the following smooth and Lipschitz functions

$$
f^\star(s,a) = ae_1, \quad f_\epsilon(s,a) = ae_1 + (Cs^\top e_2 + \epsilon)e_2, \quad \pi(s) = 0
$$

and we give the Lipschitz and $0-1$ bounded reward as $\mathrm{R}(s,a) = \frac{1}{2} + \frac{1}{2}\tanh(s^\top e_1 - s^\top e_2)$.

We can quickly note that $(\pi, f^\star)$ is EISS for $C = 1$ and any value of $\rho \in (0,1)$, as the closed-loop dynamics only depend on the most recent action and are invariant on past states and histories.

The fundamental reason why exponential policy evaluation arises is a similar exponential compounding errors problem as was found in Simchowitz et al. (2025) where ϵ errors every timestep can compound over time as they push the trajectory away from observed state action pairs. While

the imitation learning setting only needs to push the trajectory away from the expert trajectory, the flipside to offline RL not assuming perfect coverage of the expert is that the observed state action pairs could be any reachable pair under $f^\star$ from the initial state distribution. As such, we trade off the algorithm-agnostic strength of the Simchowitz et al. (2025) construction with a dataset-agnostic hardness guarantee, and focus our result on the difficulty to rule out a bad dynamics function when equipped with Euclidean model error.

Proposition D.2 (Hard to rule out bad dynamics)**.** *For any $\epsilon > 0$, Setting D.1 satisfies*

$$\|f_\epsilon(s,a) - f^\star(s,a)\| \le \epsilon \quad \textit{for all } (s,a) \textit{ reachable under } f^\star,$$

and yet for every $s_1 \in \mathrm{supp}(d_1)$, for $h \ge 2$,

$$\|s_h^{\pi,f^\star} - s_h^{\pi,f_\epsilon}\| \ge C^{h-2}\epsilon.$$

Proof. First, we can see that starting from $s_1 \in \mathrm{supp}(d_1)$, any states rolled out using $f^\star$ will remain in $\mathrm{span}(e_1)$ regardless of the action sequence. Because of this, it will hold that $\|f(s,a) - f^\star(s,a)\| = \|(Cs^\top e_2 + \epsilon)e_2\| = \epsilon$.

Let us now prove the lowerbound. We can see that for all $s_1 \in \mathrm{supp}(d_1)$ and no matter the action sequence induced by π, $e_2^\top \left(s_h^{\pi,f} - s_h^{\pi,f^\star}\right) = e_2^\top s_h^{\pi,f}$. For $e_2^\top \left(s_1^{\pi,f} - s_1^{\pi,f^\star}\right) = 0$ and $e_2^\top \left(s_{h+1}^{\pi,f} - s_{h+1}^{\pi,f^\star}\right) = Ce_2^\top \left(s_h^{\pi,f} - s_h^{\pi,f^\star}\right) + \epsilon$, we can see that this gives us the compounding difference of $e_2^\top \left(s_h^{\pi,f} - s_h^{\pi,f^\star}\right) = \epsilon \sum_{k=0}^{h-2} C^k \ge C^{h-2}\epsilon$. Thus, it hold that for $h \ge 2$, $\|s_h^{\pi,f} - s_h^{\pi,f^\star}\| \ge C^{h-2}\epsilon$ for any policy π. Fixing $\pi(s) = 0$ means that $(\pi, f^\star)$ is stable, completing the proof. □

Proposition D.3 (Reward evaluation is hard)**.** *For any $\epsilon > 0$, Setting D.1 satisfies*

$$\|f_\epsilon(s,a) - f^\star(s,a)\| \le \epsilon \quad \textit{for all } (s,a) \textit{ reachable under } f^\star \textit{ from } d_1,$$

and yet for every $H \ge 2$,

$$\lim_{\epsilon\to 0} \frac{1}{\epsilon}\left(J_H(\pi) - Q_H^{\pi,f_\epsilon}\right) \ge \frac{1}{2}C^{H-2}$$

and no guarantee of form $J_H(\pi) - Q_H^{\pi,f_\epsilon} \le \mathrm{poly}(H)\epsilon$ can hold uniformly over values of H and ϵ.

Proof. We use the same setting as Proposition D.2, but we additionally consider the reward function

$$\mathrm{R}(s,a) = \frac{1}{2} + \frac{1}{2}\tanh(s^\top e_1 - s^\top e_2).$$

Under the two dynamics functions, we can see the trajectory from π evolving for $h \ge 2$ as

$$s_h^{\pi,f^\star} = 0, \qquad s_h^{\pi,f} = \epsilon\left(\sum_{k=0}^{h-2} C^k\right) e_2$$

with $s_1^{\pi,f} = s_1^{\pi,f^\star} \in \mathrm{span}(e_1)$. Therefore, rolling out the reward gives us

$$J_H(\pi) - Q_H^{\pi,f} = \frac{1}{2}\sum_{h=2}^{H} \tanh\left(\epsilon \sum_{k=0}^{h-2} C^k\right).$$

Dividing by ϵ and noting that $\lim_{x\to 0}\frac{\tanh(x)}{x} = 1$, we have

$$\begin{aligned}
\lim_{\epsilon\to 0}\frac{1}{\epsilon}\left(J_H(\pi) - Q_H^{\pi,f}\right) &= \lim_{\epsilon\to 0}\frac{1}{2}\sum_{h=2}^{H}\frac{\tanh\left(\epsilon\sum_{k=0}^{h-2}C^k\right)}{\epsilon} \\
&= \frac{1}{2}\sum_{h=2}^{H}\sum_{k=0}^{h-2} C^k \\
&\ge \frac{1}{2}C^{H-2}.
\end{aligned}$$

If there existed some polynomial p that satisfied $J_H(\pi) - Q_H^{\pi,f} \leq p(H)\epsilon$ for all horizons $H \geq 2$ and model errors $\epsilon > 0$, then dividing by ϵ and taking the limit would imply that $p(H) \geq \frac{1}{2}C^{H-2}$ which is impossible for $C > 1$. □

We note that necessarily, as a result of our rewards being bounded, if we were to fix ϵ and then increase H eventually the bounds would saturate and the reward difference would begin to increase linearly in horizon. This fact belies the issue that decreasing ϵ further will once again reveal this exponential growth in reward difference, ruling out any universal polynomial upperbound.

We expand upon this reward evaluation problem to show that even under this strong model accuracy guarantee, pure model-based offline reinforcement learning can still fail to certify a polynomial-in-horizon upperbound on regret.

First, we define our second setting and re-prove our exponential policy evaluation result to use as a recurrent linchpin.

Setting D.4 (Policy misranking setting)**.** We consider a setting where $\mathcal{S} \subseteq \mathbb{R}^2$, $\mathcal{A} \subseteq \mathbb{R}$, and d_1 is a point mass at $(0, 0)$. We define the fixed true dynamics and policies as

$$f^\star(s, a) = ae_1, \quad \pi^\star(s) = 1, \quad \hat{\pi}(s) = -1, \quad \mathrm{R}(s, a) = \frac{1}{2} + \frac{1}{16}\tanh(s^\top e_1) - \frac{1}{4}\tanh(s^\top e_2)$$

and for every requested model accuracy $0 < \epsilon \leq 1$ we define

$$f_\epsilon(s, a) = ae_1 + \left(1 - \mathrm{bump}\left(a + \frac{3}{2}\right)\right)\left(\epsilon + 16\tanh(s^\top e_2)\right) e_2.$$

Lemma D.5. *The setting provided in Setting D.4 satisfies our assumptions of all dynamics and policies being globally Lipschitz and smooth, the reward being Lipschitz and bounded in* $[0, 1]$*, and* $(\pi^\star, f^\star)$ *is stable. We additionally have that* $(\hat{\pi}, f_\epsilon)$ *is locally EISS for arbitrary initial state perturbations and action perturbations of magnitude at most* $\frac{1}{2}$*.*

Proof. We can first see that all policies and dynamics are C^∞. The policies are both constant and the true dynamics $f^\star$ are linear, making them trivially Lipschitz and smooth. The candidate dynamics f_ϵ is linear in the first coordinate and is bounded in the second coordinate due to $1 - \mathrm{bump}(\cdot) \in [0, 1]$ by Lemma B.4 and $\tanh(\cdot) \in [-1, 1]$: by the c_1-Lipschitz bump function and the 1-Lipschitz $\tanh$, we have

$$\begin{aligned}\|f_\epsilon(s, a) - f_\epsilon(s', a')\| &\leq |a - a'| + 17c_1|a - a'| + 16\|s - s'\| \\ &\leq 17(1 + c_1)\|(s, a) - (s', a')\|\end{aligned}$$

To show smoothness, we may note that all the factors that appear in the second derivative are globally bounded in a similar fashion for $\epsilon \in (0, 1]$. Furthermore, we can see that $f_0 = f^\star$ on every $f^\star$-reachable state-action pair. To prove that R is bounded, we can use the boundedness of $\tanh$ to show that $0 \leq \frac{1}{2} - \frac{1}{16} - \frac{1}{4} \leq \mathrm{R}(s, a) \leq \frac{1}{2} + \frac{1}{16} + \frac{1}{4} \leq 1$, and for Lipschitz we can see

$$|\mathrm{R}(s, a) - \mathrm{R}(s', a')| \leq \frac{1}{16}|s^\top e_1 - s'^\top e_1| + \frac{1}{4}|s^\top e_2 - s'^\top e_2| \leq \frac{5}{16}\|s - s'\|$$

Finally, we can see that $(\pi^\star, f^\star)$ is globally EISS, and $(\hat{\pi}, f)$ is EISS for arbitrary initial state perturbations and action perturbations of magnitude at most $1/2$. The first can be seen by the closed loop behavior of $(\pi^\star, f^\star)$ ignoring the previous state, ensuring that any initial state perturbations or input perturbations being maximally forgotten after one timestep. The last point can be seen by the fact that for $a \in [-\frac{3}{2}, -\frac{1}{2}]$ it will hold that $a + \frac{3}{2} \leq 1$ and as such the e_2 component of f will be zero, making f behave identically to $f^\star$ and ignoring previous state history to only move to ae_1. □

Lemma D.6 (Reward evaluation is hard)**.** *For any* $\epsilon \in (0, 1]$*, Setting D.4 satisfies*

$$\|f_\epsilon(s, a) - f^\star(s, a)\| \leq \epsilon \quad \textit{for all } (s, a) \textit{ reachable under } f^\star \textit{ from } d_1,$$

and yet for every $H \geq 2$*,*

$$\lim_{\epsilon \to 0} \frac{1}{\epsilon}\left(J_H(\pi^\star) - Q_H^{\pi^\star, f_\epsilon}\right) \geq \frac{16^H}{1024}$$

and no guarantee of form $J_H(\pi^\star) - Q_H^{\pi^\star, f_\epsilon} \leq \mathrm{poly}(H)\epsilon$ *can hold uniformly over values of* H *and* ϵ*. We also have that for* $\bar{h} \triangleq 1 + \lceil \log_8(1 + \frac{7}{\epsilon}) \rceil$*,* $H \geq \bar{h}$*,*

$$J_H(\pi^\star) - Q_H^{\pi^\star, f_\epsilon} \geq \frac{H - \bar{h} + 1}{4} \tanh(1).$$

Proof. First, we can see that rollouts under $f^\star$ will remain in $\mathrm{span}(e_1)$ for any action sequence, and for $s \in \mathrm{span}(e_1)$, $\|f(s,a) - f^\star(s,a)\| = \left(1 - \mathrm{bump}\left(a + \frac{3}{2}\right)\right)\epsilon \leq \epsilon$.

Let us consider the evolution of $\pi^\star$ under the two dynamics functions $f^\star$ and f_ϵ. For $(\pi^\star, f^\star)$, we can see that the state would trivially jump to and stay at e_1. Under f_ϵ, we have that $\mathrm{bump}(\pi^\star(s) + 3/2) = 0$: while the first coordinate ${s_h^{\pi^\star, f_\epsilon}}^\top e_1$ would stay at 1, the second coordinate of the state would evolve as ${s_1^{\pi^\star, f_\epsilon}}^\top e_2 = 0$ and ${s_{h+1}^{\pi^\star, f}}^\top e_2 = \epsilon + 16\tanh({s_h^{\pi^\star, f_\epsilon}}^\top e_2)$. Let us denote $y_h(\epsilon) \triangleq {s_h^{\pi^\star, f_\epsilon}}^\top e_2$ for brevity. Taking the reward difference between the two rollouts for $h \geq 2$, we can see that first coordinate cancels.

$$\mathrm{R}(e_1, 1) - \mathrm{R}(e_1 + y_h(\epsilon)e_2, 1) = \frac{1}{4}\tanh(y_h(\epsilon)) \implies J_H(\pi^\star) - Q_H^{\pi^\star, f_\epsilon} = \frac{1}{4}\sum_{h=1}^{H} \tanh(y_h(\epsilon)).$$

Let us try to prove that $\lim_{\epsilon \to 0} \frac{y_h(\epsilon)}{\epsilon} = \sum_{k=0}^{h-2} 16^k$ for $h \geq 2$. For $h = 2$, we can immediately see that $y_2(\epsilon)/\epsilon = 1$. For the inductive step, we can note that $y_h(\epsilon) > 0$ if $\epsilon > 0$, and that as $\epsilon \to 0$, $y_h(\epsilon) \to 0$ since ϵ is the only positive constant added in the recursive definition of $y_h(\epsilon)$.

$$\begin{aligned}
\lim_{\epsilon \to 0} \frac{y_{h+1}(\epsilon)}{\epsilon} &= \lim_{\epsilon \to 0} 1 + 16\frac{\tanh(y_h(\epsilon))}{y_h(\epsilon)} \cdot \frac{y_h(\epsilon)}{\epsilon} \\
&= 1 + \lim_{\epsilon \to 0} \frac{\tanh(y_h(\epsilon))}{y_h(\epsilon)} \cdot 16 \lim_{\epsilon \to 0} \frac{y_h(\epsilon)}{\epsilon} \\
&= 1 + 16\sum_{k=0}^{h-2} 16^k \\
&= \sum_{k=0}^{h-1} 16^k
\end{aligned}$$

where in the second to last step we applied in the inductive assumption and used the limit $\lim_{x \to 0} \frac{1}{x}\tanh(x) = 1$. We can extend this to see that

$$\lim_{\epsilon \to 0} \frac{\tanh(y_h(\epsilon))}{\epsilon} = \lim_{\epsilon \to 0} \frac{\tanh(y_h(\epsilon))}{y_h(\epsilon)} \cdot \lim_{\epsilon \to 0} \frac{y_h(\epsilon)}{\epsilon} = \sum_{k=0}^{h-2} 16^k$$

and so for $H \geq 2$, we have

$$\lim_{\epsilon \to 0} \frac{1}{\epsilon}\left(J_H(\pi^\star) - Q_H^{\pi^\star, f_\epsilon}\right) = \frac{1}{4}\sum_{h=1}^{H} \lim_{\epsilon \to 0} \frac{\tanh(y_h(\epsilon))}{\epsilon} \geq \frac{1}{4}16^{H-2} = 2^{4H-10}$$

By concavity of $\tanh$ and $\tanh(1) > 1/2$, for $0 \leq y_h(\epsilon) \leq 1$ it holds that $\tanh(y_h(\epsilon)) \geq y_h(\epsilon)\tanh(1) \geq \frac{y_h(\epsilon)}{2}$. Since $y_1(\epsilon) = 0$ and $y_h(\epsilon)$ grows slowly with ϵ, we can say that while $y_h(\epsilon) \leq 1$ it holds that $y_{h+1}(\epsilon) \geq \epsilon + 8y_h(\epsilon)$, and iterating this gives

$$y_h(\epsilon) \geq \epsilon \sum_{k=0}^{h-2} 8^k = \epsilon\frac{8^{h-1} - 1}{8 - 1}$$

while $y_h(\epsilon) \leq 1$. When $y_h(\epsilon) \geq 1$, then $y_{h+1}(\epsilon) \geq 16\tanh(1) > 1$ so it'll remain above 1; however, finding the timestep h where the inflection occurs is critical. By inverting the expanded recursion, we can get that if $y_k(\epsilon) < 1$ for all $k < \bar{h}$ for $\bar{h} \triangleq 1 + \left\lceil \log_8\left(1 + \frac{7}{\epsilon}\right) \right\rceil$ then $y_{\bar{h}}(\epsilon) \geq 1$, and trivially $y_k(\epsilon) \geq 1$ for all $k \geq \bar{h}$. For $H \geq \bar{h}$, this gives us that

$$J_H(\pi^\star) - Q_H^{\pi^\star, f_\epsilon} \geq \frac{1}{4}\sum_{h=\bar{h}}^{H} \tanh(y_h(\epsilon)) \geq \frac{H - \bar{h} + 1}{4}\tanh(1)$$

□

Certainty equivalence is a common paradigm for control and reinforcement learning problems, and has undergone extensive theoretical analysis (Mania et al., 2019) alongside being an underlying principle behind many world-model works (Ha & Schmidhuber, 2018; Hafner et al., 2020). In certainty equivalent planning, we fit a model f_ϵ of our system $f^\star$ from its observations, and then choose a control policy $\hat{\pi}$ by treating our fitted model as the true system. In this setting our main points of comparison are inherently the evaluation error for the expert and the learned policies between the true and candidate dynamics, and the difference in reward between the expert and learned policies on the candidate dynamics.

In such a setting, we might like to request that our policy class satisfies certain assumptions, such as the exponential stability assumption, to allow transfer of our learned policy's reward from the candidate to true dynamics. First, we give an intermediary result showing that even when a candidate dynamics function is *perfectly* matched to $f^\star$ in a tube around a globally stable policy's rollout, we can't guarantee that the policy will be globally stable on the candidate dynamics. This shows that a global stability assumption of Π on a candidate dynamics f_ϵ is unreasonable, as even the expert $\pi^\star$ is not guaranteed to satisfy it.

Proposition D.7 (Global EISS is hard to transfer)**.** *There exist $(\pi, f^\star)$ pairs that are (C, ρ) EISS, where we let $\{s_h^\star, a_h^\star\}_{h\geq 1}$ denote the trajectory generated by $s_{h+1}^\star = f^\star(s_h^\star, a_h^\star)$ and $a_h^\star = \pi(s_h^\star)$ for $s_1^\star = s_1$. We request dynamics function f such that for all timesteps $h \geq 1$, for all s such that $\|s_h^\star - s\| \leq R$, for all $a \in \mathcal{A}$, $\|f^\star(s,a) - f(s,a)\| \leq \epsilon$, and let us denote $\{s_h, a_h\}_{h\geq 1}$ as the trajectory generated by rolling out π, f on s_1. Let us denote $\tilde{s}_h, \tilde{a}_h$ as the trajectory generated by inserting state perturbation x at timestep k and action perturbations $\{u_q\}_{q=k}^{h-1}$ such that $\tilde{s}_{q+1} = f(\tilde{s}_q, \tilde{a}_q)$, $\tilde{a}_q = \pi(\tilde{s}_q) + u_q$, and $\tilde{s}_k = s_k + x$ for $h \geq k$; let us denote $\tilde{s}'_h, \tilde{a}'_h$ as likewise for x' and $\{u'_q\}_{q=k}^{h-1}$.*

There exists f satisfying these properties such that for some $r > 0$, there exists x, x' and action perturbation sequences $\{u_q\}_{q=k}^{h-1}$ and $\{u'_q\}_{q=k}^{h-1}$ such that one of $\|x\|, \|x'\|$ or one of u_q, u'_q for $q \in [k, h-1]$ is greater than r, and there exists no $(\bar{C}, \bar{\rho})$ such that $\|\tilde{s}_h - \tilde{s}'_h\| \leq \bar{C}\bar{\rho}^{h-k}\|x - x'\| + \sum_{q=k}^{h-1} \bar{C}\bar{\rho}^{h-1-q}\|u_q - u'_q\|$ for every $h \geq k$.

Proof. We shall construct a counterexample for $\mathcal{S} = \mathcal{A} = \mathbb{R}$ for d_1 as a point mass at 0, and $r = 2R$. Let us set $\pi^\star(s) = 0$, $f^\star(s,a) = \frac{1}{2}s + a$, and $f(s,a) = 2s - \frac{3}{2}\operatorname{bump}(\frac{s}{R})s + a$.

It can trivially be seen that $(\pi^\star, f^\star)$ is globally $(1, \frac{1}{2})$ EISS, and that all functions are Lipschitz and smooth. For $s \in [-R, R]$, it may be seen that $f^\star(s,a) = f(s,a)$. Let us pick some numbers $\Delta \geq 2R$ and $2R > \delta > 0$.

Let us first look at if $x = \Delta$ all other perturbations are 0: in this case, $\tilde{s}_{k+i} = 2^i\Delta$ while $\tilde{s}'_{k+i} = 0$: trivially, for all $\bar{\rho} < 1$ there exists no finite $\bar{C}$ such that $(\frac{2}{\bar{\rho}})^i\Delta \leq \bar{C}$ for all $i \geq 0$. Similarly, if $x = \Delta$ and $x' = \Delta + \delta$ such that $|x - x'| < 2R$, it holds that $|\tilde{s}_{k+1} - \tilde{s}'_{k+1}| = 2^i\delta$, and as such there exists no finite $\bar{C}$ such that $(\frac{2}{\bar{\rho}})^i\delta \leq \bar{C}$.

Now, let us insert action perturbation $u_k = \Delta$, with all other perturbations being 0: here, $\tilde{s}_{k+1} = \Delta$ and as such $\tilde{s}_{k+i} = 2^{i-1}\Delta$. Finally, for $u_k = \Delta$ and $u'_k = \Delta + \delta$ such that $|u_k - u'_k| < 2R$, it holds that $|\tilde{s}_{k+i} - \tilde{s}'_{k+i}| = 2^{i-1}\delta$. □

That being said, even an assumption that every policy $\pi \in \Pi$ is locally stable on f_ϵ is insufficient to guarantee an upperbound. We are able to construct a setting where our learned policy has regret that grows linearly in horizon and is independent to model error, but sufficiently long horizon allows $\hat{\pi}$ to appear better than $\pi^\star$ on the candidate dynamics. As such, no polynomial in horizon upperbound can be universally constructed for the regret that shrinks with decreasing model error.

Corollary D.8 (Certainty equivalent algorithms fail)**.** *For any $0 < \epsilon \leq 1$, Setting D.4 satisfies*

$$\|f_\epsilon(s,a) - f^\star(s,a)\| \leq \epsilon \quad \textit{for all } (s,a) \textit{ reachable under } f^\star \textit{ from } d_1$$

where $(\hat{\pi}, f_\epsilon)$ is locally closed-loop EISS, and for all sufficiently large H, $Q_H^{\pi^\star, f_\epsilon} < Q_H^{\hat{\pi}, f_\epsilon}$ even though

$$J_H(\pi^\star) - J_H(\hat{\pi}) = \Theta(H) > 0.$$

Consequently, for $\Pi = \{\pi^\star, \hat{\pi}\}$*, no guarantee of form* $J_H(\pi^\star) - J_H(\hat{\pi}) \leq \text{poly}(H)\epsilon$ *can hold uniformly over all values of* H *and* ϵ *for algorithm* $\hat{\pi} = \arg\max_{\pi \in \Pi} Q_H^{\pi, f_\epsilon}$*, even when policy selection returns a locally EISS policy.*

Proof. Under $(\pi^\star, f^\star)$ the states will stay at $s_h^{\pi^\star, f^\star} = e_1$ for $h \geq 2$, and under $(\hat{\pi}, f^\star)$ the states will stay at $s_h^{\hat{\pi}, f^\star} = -e_1$: as such, every timestep the reward gap will be $\mathrm{R}(e_1, 1) - \mathrm{R}(-e_1, -1) = \frac{1}{8}\tanh(1)$ and as such

$$J(\pi^\star) - J(\hat{\pi}) = \frac{H-1}{8}\tanh(1) > 0.$$

Since the rollouts of $(\hat{\pi}, f_\epsilon)$ and $(\hat{\pi}, f^\star)$ are identical, we can use Lemma D.6 to expand

$$\begin{aligned} Q_H^{\pi^\star, f_\epsilon} - Q_H^{\hat{\pi}, f_\epsilon} &= \left(Q_H^{\pi^\star, f_\epsilon} - J_H(\pi^\star)\right) + \left(J_H(\pi^\star) - J_H(\hat{\pi})\right) + \left(J_H(\hat{\pi}) - Q_H^{\hat{\pi}, f_\epsilon}\right) \\ &\leq \frac{H-1}{8}\tanh(1) - \frac{H - \bar{h} + 1}{4}\tanh(1) \\ &= \frac{2\bar{h} - H - 3}{8}\tanh(1). \end{aligned}$$

For $H \geq 2\bar{h}$, this means that f_ϵ misranks the policies with $Q_H^{\pi^\star, f_\epsilon} - Q_H^{\hat{\pi}, f_\epsilon} \leq -\frac{3}{8}\tanh(1) < 0$, allowing certainty equivalent planning to pick $\hat{\pi}$ despite the suboptimality $J(\pi^\star) - J(\hat{\pi}) = \frac{H-1}{8}\tanh(1)$.

Now, we must rule out the polynomial guarantee uniformly over H and ϵ. From all of this, we can conclude that no matter how small a positive accuracy level ϵ is requested, for $H \geq O(\log(\frac{1}{\epsilon}))$ it holds that $Q^{\pi^\star, f_\epsilon} - Q^{\hat{\pi}, f_\epsilon} < 0$ and yet $J(\pi^\star) - J(\hat{\pi}) = \Theta(H)$. Finally, we can see that while setting H as sufficiently large creates the model misranking, we can additionally create this by substituting $\epsilon = 2^{-H}$: in this case,

$$\begin{aligned} \bar{h} &\triangleq 1 + \left\lceil \log_8 \left(1 + \frac{7}{(2^{-H})}\right) \right\rceil \\ &= 1 + \left\lceil \log_8 \left(1 + 7 \cdot 2^H\right) \right\rceil \\ &\leq 1 + \left\lceil \log_8 \left(2^{H+3}\right) \right\rceil \\ &= 1 + \left\lceil \frac{H+3}{3} \right\rceil \leq 2 + \frac{H+3}{3} = \frac{H+9}{3}. \end{aligned}$$

For $H \geq 18$, we can see that $\frac{H+9}{3} \leq \frac{H}{2}$ and so $2\bar{h} \leq H$ and $Q^{\pi^\star, f_\epsilon} - Q^{\hat{\pi}, f_\epsilon} < 0$. As such, we can see that for all $H \geq 18$, if $\epsilon = 2^{-H}$ it causes the model misranking that allows $\hat{\pi}$ to be chosen over $\pi^\star$ in f_ϵ while $J_H(\pi^\star) - J_H(\hat{\pi}) = \Omega(H2^H)\epsilon$, so no polynomial p can satisfy $J_H(\pi^\star) - J_H(\hat{\pi}) \leq p(H)\epsilon$. □

Similar reasoning suffices to show a related result for pessimistic planning. In a model-based adaptation of the Bellman-consistent pessimistic setting from Xie et al. (2021) similar to that of Uehara & Sun (2021), we fit a set of models $\mathcal{F}(\epsilon)$ to the system $f^\star$ from its observations, and then pick the policy $\hat{\pi}$ with the best worst-case reward. Here, our failure to rule out models $f \in \mathcal{F}(\epsilon)$ where $(\pi^\star, f)$ exponentially deviates from $(\pi^\star, f^\star)$ means that we might choose bad policies $\hat{\pi}$ so long as they only do better than the unrepresentative $(\pi^\star, f)$ rollouts.

Corollary D.9 (Pessimistic algorithms fail)**.** *For any* $0 < \epsilon \leq 1$*, Setting D.4 satisfies*

$$\|f_\epsilon(s, a) - f^\star(s, a)\| \leq \epsilon \quad \textit{for all } (s, a) \textit{ reachable under } f^\star \textit{ from } d_1.$$

Moreover, for $\mathcal{F}(\epsilon) = \{f^\star, f_\epsilon\}$ *and all sufficiently large* H*,* $Q_H^{\pi^\star, f_\epsilon} < Q_H^{\hat{\pi}, f_\epsilon}$ *even though*

$$J_H(\pi^\star) - J_H(\hat{\pi}) = \Theta(H) > 0.$$

Consequently, for $\Pi = \{\pi^\star, \hat{\pi}\}$*, pessimistic policy selection* $\arg\max_{\pi \in \Pi} min_{f \in \mathcal{F}(\epsilon)} Q_H^{\pi, f}$ *will return* $\hat{\pi}$ *and no guarantee of form* $J_H(\pi^\star) - J_H(\hat{\pi}) \leq \text{poly}(H)\epsilon$ *can hold uniformly over all values of* $H \geq 2$ *and* $0 < \epsilon \leq 1$ *even when* $f^\star \in \mathcal{F}(\epsilon)$ *and when policy selection returns a policy that is locally EISS on every dynamics function* $f \in \mathcal{F}(\epsilon)$*.*

Proof. Let us fix $\epsilon \in (0, 1]$ and construct $\mathcal{F} = \{f^\star, f_\epsilon\}$ and $\Pi = \{\pi^\star, \hat{\pi}\}$. From Corollary D.8, we can see that $\min_{f\in\{f^\star,f_\epsilon\}} Q_H^{\pi^\star,f} = Q_H^{\pi^\star,f_\epsilon}$ for $H \geq 2$ and from the fact that $\hat{\pi}$ satisfies $1 - \mathrm{bump}\left(\hat{\pi}(s) + \frac{3}{2}\right) = 0$, $(\hat{\pi}, f^\star)$ will behave the same as $(\hat{\pi}, f_\epsilon)$ and we have $\min_{f\in\{f^\star,f_\epsilon\}} Q_H^{\hat{\pi},f} = Q_H^{\hat{\pi},f_\epsilon} = Q_H^{\hat{\pi},f^\star} = J_H(\hat{\pi})$. Seeing that $Q_H^{\pi^\star,f_\epsilon} < Q_H^{\hat{\pi},f_\epsilon}$ for $H \geq O(\log(\frac{1}{\epsilon}))$, we can immediately say that for sufficiently large H the pessimistic algorithm $\arg\max_{\pi\in\Pi} \min_{f\in\mathcal{F}} Q_H^{\pi,f}$ will select $\hat{\pi}$ and obtain regret independent of ϵ. From here, we may rule out the possibility of a polynomial guarantee universally over H and ϵ, noting that for all $H \geq 18$ as was before, we have that if $\epsilon = 2^{-H}$ we'll choose $\hat{\pi}$ and obtain $J_H(\pi^\star) - J_H(\hat{\pi}) = \Omega(H2^H)\epsilon$, so no polynomial p can satisfy $J_H(\pi^\star) - J_H(\hat{\pi}) \leq p(H)\epsilon$. □

E Extended Proofs for Section 4.2

Definition E.1 (Chunk length notation)**.** We take the symbol τ_h to be the start time of the active chunk at time h, such that $\tau_h \triangleq 1 + \ell\lfloor\frac{h-1}{\ell}\rfloor$; i_h to be the current index in the latest chunk at time h, such that $i_h \triangleq h - \tau_h + 1 \in [\ell]$; and σ_h to be the start time of the next chunk following time h, such that $\sigma_h \triangleq \tau_h + \ell$. We note that it holds that $0 \leq h - \tau_h = i_h - 1 \leq \ell - 1$.

Lemma E.2 (Lem D.1 in Zhang et al. (2025))**.** *Let (π, f) be a closed-loop (C, ρ)-EISS pair. If we define $B_h \triangleq \nabla_a f(s_h^{\pi,f}, a_h^{\pi,f})$, $A_h^{\mathrm{cl}} \triangleq \nabla_s f(s_h^{\pi,f}, a_h^{\pi,f}) + \nabla_a f(s_h^{\pi,f}, a_h^{\pi,f})\nabla_s \pi(s_h^{\pi,f})$ for the rollout of (π, f) from initial state s_1, and $A_{k:h}^{\mathrm{cl}} \triangleq A_{h-1}^{\mathrm{cl}} \cdots A_k^{\mathrm{cl}}$, then it holds that*

$$\|A_{1:h}^{\mathrm{cl}}\| \leq C\rho^{h-1},\ \|A_{k:h}^{\mathrm{cl}}\| \leq C\rho^{h-k},\ \|A_{k+1:h}^{\mathrm{cl}} B_k\| \leq C\rho^{h-1-k},\ \textit{for all } 1 \leq k < h.$$

Proposition E.3 (Prop 3.1 in Zhang et al. (2025))**.** *Let there be some closed-loop (C, ρ)-EISS policy dynamics pair (π, f) and let there be some alternate dynamics function f' that is open-loop (C, ρ)-EISS. For chunked policy $\tilde{\pi} = \mathsf{chunked}(\pi, f, \ell)$, the closed-loop system $(\tilde{\pi}, f')$ induced by the chunked policy on f' is $(\tilde{C}, \tilde{\rho})$-EISS, where $\tilde{C} = \log(1/\rho)^{-1} \cdot \mathrm{poly}(L_\pi, C)$ and $\tilde{\rho} = \rho^{1/2}$, as long as the chunk length is sufficiently long: $\ell > \log(1/\rho)^{-1} \cdot \log(\mathrm{poly}(L_\pi, C))$.*

Lemma E.4 (Key recursion with chunking)**.** *Let there be a non-negative sequence $\{\delta_h\}_{h=1}^H$ and chunk length $\ell \in \mathbb{Z}_{\geq 1}$ such that $\delta_1 = 0$ and for all $h \in [H-1]$, it holds that*

$$\delta_{h+1} \leq \sum_{k=1}^{h} C\rho^{h-k}(\epsilon_k + C_1\delta_k^2 + C_1\delta_{\tau_k}^2)$$

for $C, C_1 > 0$, $0 \leq \rho < 1$, $0 \leq h - \tau_h \leq \ell - 1$, and $\epsilon_h \geq 0$ for all $h \in [H]$. Then for $\bar{C} \triangleq 2C$, $\bar{\rho} \triangleq \frac{1+\rho}{2}$, and for every $h \in [H-1]$

$$\max_{k\in[h]} \epsilon_k \leq \epsilon_{\max} \leq \frac{1}{16C_1(1+\bar{\rho}^{2-2\ell})C_{stab}^2} \implies \delta_{h+1} \leq \sum_{k=1}^{h} \bar{C}\bar{\rho}^{h-k}\epsilon_k.$$

Notably, this means that as ℓ increases, $\epsilon_{\max}$ must decrease.

Proof. For the base case, we look at $h = 1$: here, $\delta_2 \leq C(\epsilon_1 + C_1\delta_1^2 + C_1\delta_{\tau_1}^2) = C\epsilon_1 \leq \bar{C}\epsilon_1$ since $\delta_1 = 0$, and inherently $\tau_1 = 1$. Now for the inductive step, we can assume that $\delta_k \leq \sum_{t=1}^{k-1} \bar{C}\bar{\rho}^{k-t-1}\epsilon_t$ for all $k \leq h$, and we'll seek to prove it for $h+1$. Since we have that $\tau_k \leq k$ and $k - \tau_k \leq \ell - 1$ by the definition of chunk lengths,

$$\delta_{\tau_k} \leq \sum_{t=1}^{\tau_k-1} \bar{C}\bar{\rho}^{\tau_k-1-t}\epsilon_t \leq \bar{\rho}^{\tau_k-k} \sum_{t=1}^{\tau_k-1} \bar{C}\bar{\rho}^{k-1-t}\epsilon_t \leq \bar{\rho}^{\tau_k-k} \sum_{t=1}^{k-1} \bar{C}\bar{\rho}^{k-1-t}\epsilon_t \leq \bar{\rho}^{1-\ell} \sum_{t=1}^{k-1} \bar{C}\bar{\rho}^{k-1-t}\epsilon_t$$

Now that we've aligned the sums, we can combine this with δ_k:

$$\begin{aligned}\delta_k^2 + \delta_{\tau_k}^2 &\leq (1 + \bar{\rho}^{2-2\ell}) \left(\sum_{t=1}^{k-1} \bar{C}\bar{\rho}^{k-1-t}\epsilon_t\right)^2 \\ &\leq (1 + \bar{\rho}^{2-2\ell})\bar{C}_{stab}\epsilon_{\max} \sum_{t=1}^{k-1} \bar{C}\bar{\rho}^{k-1-t}\epsilon_t\end{aligned}$$

Putting it all together gives an expression for the full recursion in terms of ϵ:

$$\delta_{h+1} \le \sum_{k=1}^{h} C\rho^{h-k}\epsilon_k + CC_1(1+\bar{\rho}^{2-2\ell})\bar{C}_{stab}\epsilon_{\max}\left(\sum_{k=1}^{h}\rho^{h-k}\sum_{t=1}^{k-1}\bar{C}\bar{\rho}^{k-1-t}\epsilon_t\right)$$

If we interchange the sums and note that $\rho \le \bar{\rho}$,

$$\begin{aligned}
\delta_{h+1} &\le \sum_{k=1}^{h} C\rho^{h-k}\epsilon_k + CC_1(1+\bar{\rho}^{2-2\ell})\bar{C}_{stab}\epsilon_{\max}\left(\bar{C}\sum_{t=1}^{h-1}\epsilon_t\sum_{k=t+1}^{h}\rho^{h-k}\bar{\rho}^{k-1-t}\right)\\
&= \sum_{k=1}^{h} C\rho^{h-k}\epsilon_k + CC_1(1+\bar{\rho}^{2-2\ell})\bar{C}_{stab}\epsilon_{\max}\left(\bar{C}\sum_{t=1}^{h-1}\epsilon_t\frac{\bar{\rho}^{h-t}-\rho^{h-t}}{\bar{\rho}-\rho}\right)\\
&\le \sum_{k=1}^{h} C\rho^{h-k}\epsilon_k + \frac{CC_1(1+\bar{\rho}^{2-2\ell})\bar{C}_{stab}}{\bar{\rho}-\rho}\epsilon_{\max}\left(\bar{C}\sum_{t=1}^{h-1}\bar{\rho}^{h-t}\epsilon_t\right)\\
&\le \left(C + \frac{CC_1(1+\bar{\rho}^{2-2\ell})\bar{C}^2}{(1-\bar{\rho})(\bar{\rho}-\rho)}\epsilon_{\max}\right)\sum_{k=1}^{h}\bar{\rho}^{h-k}\epsilon_k
\end{aligned}$$

For our claim to hold, it must be true that $C + \frac{CC_1(1+\bar{\rho}^{2-2\ell})\bar{C}^2}{(1-\bar{\rho})(\bar{\rho}-\rho)}\epsilon_{\max} \le \bar{C}$, and as such we can constrain

$$\epsilon_{\max} \le \frac{(\bar{C}-C)(1-\bar{\rho})(\bar{\rho}-\rho)}{CC_1(1+\bar{\rho}^{2-2\ell})\bar{C}^2} = \frac{1}{16C_1(1+\bar{\rho}^{2-2\ell})C_{stab}^2}.$$

□

Lemma E.5 (Chunk smoothness)**.** *Chunked policy* $\mathsf{chunk}_i[\tilde{\pi}](s) \triangleq \pi((f\pi)^{i-1}(s))$ *is* L_{chunk}*-Lipschitz and* M_{chunk}*-smooth for* (C,ρ)*-EISS* (π, f)*, where*

$$L_{\mathsf{chunk}} \triangleq L_\pi C, \qquad M_{\mathsf{chunk}} \triangleq M_\pi C^2 + L_\pi C_{stab}C^2(M_f(1+L_\pi^2) + L_f M_\pi).$$

Proof. Let us begin with the Lipschitz constant.

$$\begin{aligned}
\|\mathsf{chunk}_i[\tilde{\pi}](s) - \mathsf{chunk}_i[\tilde{\pi}](s')\| &= \|\pi((f\pi)^{i-1}(s)) - \pi((f\pi)^{i-1}(s'))\|\\
&\le L_\pi\|(f\pi)^{i-1}(s) - (f\pi)^{i-1}(s')\|\\
&\le L_\pi C\rho^{i-1}\|s-s'\|\\
&\le L_\pi C\|s-s'\|.
\end{aligned}$$

For the smoothness constant, let us begin with the smoothness of the closed-loop map $(f\pi)$. If we define the map $s \mapsto (s, \pi(s))$ as $\bar{\pi}$ such that $(f\pi)(s) = (f\circ\bar{\pi})(s)$, then for test vectors v, w we have

$$\begin{aligned}
\nabla_s\bar{\pi}(s)v &= (v, \nabla_s\pi(s)v), \quad \|\nabla_s\bar{\pi}(s)\| \le \sqrt{1+L_\pi^2}, \quad \|\nabla_s^2\bar{\pi}(s)\| = \|\nabla_s^2\pi(s)\| \le M_\pi\\
\nabla_s(f\pi)(s)v &= \nabla_s f(\bar{\pi}(s))v + \nabla_a f(\bar{\pi}(s))\nabla_s\pi(s)v\\
\nabla_s^2(f\pi)(s)[v,w] &= \nabla_{s,a}^2 f(\bar{\pi}(s))[\nabla_s\bar{\pi}(s)v, \nabla_s\bar{\pi}(s)w] + \nabla_a f(\bar{\pi}(s))\nabla_s^2\pi(s)[v,w]\\
\|\nabla_s^2(f\pi)(s)\| &\le M_f(1+L_\pi^2) + L_f M_\pi \triangleq M_{cl}
\end{aligned}$$

Just bounding by smoothness directly would be too coarse though, since we can exploit the stability of the $(f\pi)$ map. If we define $s_{(i)} \triangleq (f\pi)^i(s)$ and $s'_{(i)} \triangleq (f\pi)^i(s')$, we can first see that for $i=1$ we have $\mathsf{chunk}_1[\tilde{\pi}](s) = \pi(s)$, so since $M_\pi \le M_{\mathsf{chunk}}$ the claim holds. For $i \ge 2$, chain rule gives us

$$\nabla_s(f\pi)^{i-1}(s) = \nabla_s(f\pi)(s_{(i-2)})\nabla_s(f\pi)(s_{(i-3)})\cdots\nabla_s(f\pi)(s)$$

Taking the difference, we can see that it telescopes:

$$\begin{aligned}
&\nabla_s(f\pi)^{i-1}(s) - \nabla_s(f\pi)^{i-1}(s')\\
&\quad= \nabla_s(f\pi)(s_{(i-2)})\cdots\nabla_s(f\pi)(s) - \nabla_s(f\pi)(s'_{(i-2)})\cdots\nabla_s(f\pi)(s')\\
&\quad= \sum_{j=0}^{i-2}\nabla_s(f\pi)(s_{(i-2)})\cdots\nabla_s(f\pi)(s_{(j+1)})\left(\nabla_s(f\pi)(s_{(j)}) - \nabla_s(f\pi)(s'_{(j)})\right)\\
&\qquad\nabla_s(f\pi)(s'_{(j-1)})\cdots\nabla_s(f\pi)(s')
\end{aligned}$$

Since the Jacobian of the closed-loop map also obeys EISS by Lemma E.2, we have that

$$\begin{aligned}\|\nabla_s(f\pi)(s_{(i-2)})\cdots\nabla_s(f\pi)(s_{(j+1)})\| &\le C\rho^{i-j-2}\\ \|\nabla_s(f\pi)(s'_{(j-1)})\cdots\nabla_s(f\pi)(s')\| &\le C\rho^{j}\\ \|\nabla_s(f\pi)(s_{(j)})-\nabla_s(f\pi)(s'_{(j)})\| &\le M_{cl}\|s_{(j)}-s'_{(j)}\| \le M_{cl}C\rho^j\|s-s'\|\end{aligned}$$

This gives the smoothness constant of $(f\pi)^{i-1}$ as

$$\begin{aligned}&\|\nabla_s(f\pi)^{i-1}(s)-\nabla_s(f\pi)^{i-1}(s')\|\\ &\quad\le\sum_{j=0}^{i-2}C\rho^{i-j-2}C\rho^j M_{cl}C\rho^j\|s-s'\|\\ &\quad= M_{cl}C^3\rho^{i-2}\sum_{j=0}^{i-2}\rho^j\|s-s'\|\\ &\quad\le M_{cl}C_{stab}C^2\|s-s'\|\end{aligned}$$

Finally we can analyze the smoothness of the object in question, $\mathsf{chunk}_i[\tilde\pi]=\pi\circ(f\pi)^{i-1}$.

$$\begin{aligned}&\nabla_s\mathsf{chunk}_i[\tilde\pi](s)-\nabla_s\mathsf{chunk}_i[\tilde\pi](s')\\ &\quad=\nabla_s\pi(s_{(i-1)})\nabla_s(f\pi)^{i-1}(s)-\nabla_s\pi(s'_{(i-1)})\nabla_s(f\pi)^{i-1}(s')\\ &\quad=\left(\nabla_s\pi(s_{(i-1)})-\nabla_s\pi(s'_{(i-1)})\right)\nabla_s(f\pi)^{i-1}(s)\\ &\qquad+\nabla_s\pi(s'_{(i-1)})\left(\nabla_s(f\pi)^{i-1}(s)-\nabla_s(f\pi)^{i-1}(s')\right)\end{aligned}$$

Bounding each term separately now that we've introduced workable differences,

$$\begin{aligned}\|\left(\nabla_s\pi(s_{(i-1)})-\nabla_s\pi(s'_{(i-1)})\right)\nabla_s(f\pi)^{i-1}(s)\| &\le M_\pi\|s_{(i-1)}-s'_{(i-1)}\|C\rho^{i-1}\\ &\le M_\pi C^2\rho^{2(i-1)}\|s-s'\|\\ \|\nabla_s\pi(s'_{(i-1)})\left(\nabla_s(f\pi)^{i-1}(s)-\nabla_s(f\pi)^{i-1}(s')\right)\| &\le L_\pi M_{cl}C_{stab}C^2\|s-s'\|\end{aligned}$$

Putting it all together gives our final result,

$$\begin{aligned}\|\nabla_s\mathsf{chunk}_i[\tilde\pi](s)-\nabla_s\mathsf{chunk}_i[\tilde\pi](s')\| &\le M_\pi C^2\rho^{2(i-1)}\|s-s'\|+L_\pi M_{cl}C_{stab}C^2\|s-s'\|\\ &\le C^2\left(M_\pi+L_\pi M_{cl}C_{stab}\right)\|s-s'\|.\end{aligned}$$

□

Definition E.6. Let us have s_h evolving under $(\tilde\pi, f)$. At timestep $k+1$, if we inject perturbation αv_k such that $\hat s_{k+1}=s_{k+1}+\alpha v_k$ and continue rolling out the trajectory as $\hat s_h$, we define $\tilde A^{\mathrm{cl}}_{h,k}v_k \triangleq \frac{d}{d\alpha}\hat s_h|_{\alpha=0}$.

Lemma E.7 (Recurrence of $\tilde A^{\mathrm{cl}}$)**.** *It holds that for $A_h\triangleq\nabla_s f(s_h,a_h)$, $B_h\triangleq\nabla_a f(s_h,a_h)$, and $J_h\triangleq\nabla_s\mathsf{chunk}_{i_h}[\tilde\pi](s_{\tau_h})$,*

$$\begin{aligned}&\tilde A^{\mathrm{cl}}_{h+1,k}v_k=0 \quad \textit{if } h+1<k+1, \quad \tilde A^{\mathrm{cl}}_{k+1,k}v_k=v_k,\\ &\tilde A^{\mathrm{cl}}_{h+1,k}v_k=A_h\tilde A^{\mathrm{cl}}_{h,k}v_k+B_hJ_h\tilde A^{\mathrm{cl}}_{\tau_h,k}v_k \quad \textit{if } h+1>k+1.\end{aligned}$$

Proof. For $t\le k$ it holds that $\frac{d}{d\alpha}\hat s_t=0$ since there was no perturbation yet, and $\frac{d}{d\alpha}\hat s_{k+1}=v_k$. This inherently gives us that $\tilde A^{\mathrm{cl}}_{k+1,k}v_k=v_k$, and $\tilde A^{\mathrm{cl}}_{t+1,k}v_k=0$ for all $t+1<k+1$. We have by definition that $\hat s_{t+1}=f(\hat s_t,\mathsf{chunk}_{i_t}[\tilde\pi](\hat s_{\tau_t}))$, and if $\alpha=0$ we would have that $\hat s_t=s_t$ and $\mathsf{chunk}_{i_t}[\tilde\pi](\hat s_{\tau_t})=a_t$. Let us define $\hat a_t\triangleq\mathsf{chunk}_{i_t}[\tilde\pi](\hat s_{\tau_t})$.

$$\frac{d}{d\alpha}\hat s_{t+1}=\nabla_s f(\hat s_t,\hat a_t)\frac{d}{d\alpha}\hat s_t+\nabla_a f(\hat s_t,\hat a_t)\nabla_s\mathsf{chunk}_{i_t}[\tilde\pi](\hat s_{\tau_t})\frac{d}{d\alpha}\hat s_{\tau_t}$$

If we evaluate at $\alpha=0$, this gives us that $\tilde A^{\mathrm{cl}}_{t+1,k}v_k=A_t\tilde A^{\mathrm{cl}}_{t,k}v_k+B_tJ_t\tilde A^{\mathrm{cl}}_{\tau_t,k}v_k$. Notably, if $\tau_t\le k$, the second term drops. □

Lemma E.8. *Let us fix f and f' as two candidate dynamics functions, and let $\tilde{\pi}$ be a chunked policy of chunk length ℓ that is L_{chunk}-Lipschitz and M_{chunk}-smooth. We give $\tilde{A}^{\mathrm{cl}}$ as being defined on the rollout of $(\tilde{\pi}, f)$. Let s'_h, a'_h be defined as resulting from a rollout of $(\tilde{\pi}, f')$ and s_h, a_h resulting from a rollout of $(\tilde{\pi}, f)$, where $s'_1 = s_1$. It holds that the trajectory difference obeys*

$$s'_{h+1} - s_{h+1} = \sum_{k=1}^{h} \tilde{A}^{\mathrm{cl}}_{h+1,k} \left((f' - f)(s'_k, a'_k) + R_k \right)$$

where $\|R_k\| \le C_{rem} \left(\|s'_k - s_k\|^2 + \|s'_{\tau_k} - s_{\tau_k}\|^2 \right)$, $C_{rem} \triangleq M_f + L_f M_{\mathsf{chunk}} + M_f L^2_{\mathsf{chunk}}$.

Proof. We begin by a Taylor expansion of the trajectory difference to expose the remainder terms that we'll have to bound.

$$\begin{aligned}
s'_{h+1} - s_{h+1} &= f'(s'_h, a'_h) - f(s_h, a_h) \pm f(s'_h, a'_h) \\
&= (f' - f)(s'_h, a'_h) + f(s'_h, a'_h) - f(s_h, a_h) \\
&= (f' - f)(s'_h, a'_h) + \nabla_s f(s_h, a_h)(s'_h - s_h) + \nabla_a f(s_h, a_h)(a'_h - a_h) + r^X_h \\
a'_h - a_h &= \nabla_s \mathsf{chunk}_{i_h}[\tilde{\pi}](s_{\tau_h})(s'_{\tau_h} - s_{\tau_h}) + r^U_h \\
\|r^U_h\| &\le M_{\mathsf{chunk}} \|s'_{\tau_h} - s_{\tau_h}\|^2 \\
\|r^X_h\| &\le M_f \left(\|s'_h - s_h\|^2 + \|a'_h - a_h\|^2 \right) \\
&\le M_f \|s'_h - s_h\|^2 + M_f L^2_{\mathsf{chunk}} \|s'_{\tau_h} - s_{\tau_h}\|^2 \\
\|\nabla_a f(s_h, a_h) r^U_h + r^X_h\| &\le (L_f M_{\mathsf{chunk}} + M_f L^2_{\mathsf{chunk}}) \|s'_{\tau_h} - s_{\tau_h}\|^2 + M_f \|s'_h - s_h\|^2 \\
&\le C_{rem} \left(\|s'_h - s_h\|^2 + \|s'_{\tau_h} - s_{\tau_h}\|^2 \right)
\end{aligned}$$

Here, we've identified R_k already as $R_k \triangleq \nabla_a f(s_k, a_k) r^U_k + r^X_k$. For brevity, let us define $e_h \triangleq (f' - f)(s'_h, a'_h)$, $A_h \triangleq \nabla_s f(s_h, a_h)$, $B_h \triangleq \nabla_a f(s_h, a_h)$, and $J_h \triangleq \nabla_s \mathsf{chunk}_{i_h}[\tilde{\pi}](s_{\tau_h})$.

The rest of the proof shall proceed by induction, now that we've exposed the general form. We can first note that for $h = 1$, $s'_1 - s_1 = 0$ by definition, and $R_1 = 0$: as a result, we can see the base case:

$$s'_2 - s_2 = (f' - f)(s'_1, a'_1) + f(s'_1, a'_1) - f(s_1, a_1) = e_1 + R_1$$

Since $\tilde{A}^{\mathrm{cl}}_{2,1} v_1 = v_1$, this gives us that for this perturbation, $s'_2 - s_2 = \tilde{A}^{\mathrm{cl}}_{2,1}(e_1 + R_1)$. Now, let us assume that for all $t \le h$ that $s'_t - s_t = \sum_{k=1}^{t-1} \tilde{A}^{\mathrm{cl}}_{t,k}(e_k + R_k)$ since it held for $h = 2$, and analyze $h + 1$. Noting that $\tilde{A}^{\mathrm{cl}}_{\tau_h,k} v_k = 0$ for $\tau_h \le k$, Lemma E.7 gives us

$$\begin{aligned}
s'_{h+1} - s_{h+1} &= A_h(s'_h - s_h) + B_h J_h (s'_{\tau_h} - s_{\tau_h}) + e_h + R_h \\
&= A_h \sum_{k=1}^{h-1} \tilde{A}^{\mathrm{cl}}_{h,k}(e_k + R_k) + B_h J_h \sum_{k=1}^{\tau_h - 1} \tilde{A}^{\mathrm{cl}}_{\tau_h,k}(e_k + R_k) + e_h + R_h \\
&= e_h + R_h + \sum_{k=1}^{h-1} (A_h \tilde{A}^{\mathrm{cl}}_{h,k} + B_h J_h \tilde{A}^{\mathrm{cl}}_{\tau_h,k})(e_k + R_k) \\
&= \tilde{A}^{\mathrm{cl}}_{h+1,h}(e_h + R_h) + \sum_{k=1}^{h-1} \tilde{A}^{\mathrm{cl}}_{h+1,k}(e_k + R_k) \\
&= \sum_{k=1}^{h} \tilde{A}^{\mathrm{cl}}_{h+1,k}(e_k + R_k)
\end{aligned}$$

□

Lemma E.9. *Let us say that for open-loop (C, ρ)-EISS dynamics candidate f and length-ℓ chunked policy $\tilde{\pi}$ such that $(\tilde{\pi}, f)$ is closed-loop $(\tilde{C}, \tilde{\rho})$-EISS, we define $\tilde{A}^{\mathrm{cl}}$ on the rollout of $(\tilde{\pi}, f)$. It holds that for perturbation v_k, for every $h \ge k$, $\|\tilde{A}^{\mathrm{cl}}_{h+1,k} v_k\| \le C_{\mathsf{chunk}} \rho^{h-k}_{\mathsf{chunk}} \|v_k\|$, where $C_{\mathsf{chunk}} \triangleq C \max\{1, \tilde{C}\}$ and $\rho_{\mathsf{chunk}} \triangleq \max\{\rho, \tilde{\rho}\}$.*

Proof. We note again that we're dealing with a perturbation αv_k at timestep $k+1$, and as such $s_t = \hat{s}_t$ for $t \le k$ and $\hat{s}_{k+1} = s_{k+1} + \alpha v_k$.

Let us first analyze the case where $h+1 \le \sigma_k$, and as such $[k+1, h+1]$ is in the same chunk as the perturbation. All actions are fixed functions of the previous s_{τ_k}, so αv_k only impacts the open-loop dynamics of f. As such, we have that $\|\hat{s}_{h+1} - s_{h+1}\| \le C\rho^{h-k}|\alpha|\|v_k\|$. Given the derivative at $\alpha = 0$ as $\frac{d}{d\alpha}\hat{s}_{h+1}|_{\alpha=0} = \lim_{\alpha\to 0}\frac{\hat{s}_{h+1}-s_{h+1}}{\alpha}$, we have that $\frac{\|\hat{s}_{h+1}-s_{h+1}\|}{|\alpha|} \le C\rho^{h-k}\|v_k\|$, and as such we can conclude that $\|\tilde{A}^{\mathrm{cl}}_{h+1,k}v_k\| \le C_{\mathsf{chunk}}\rho^{h-k}_{\mathsf{chunk}}\|v_k\|$.

Now analyzing the case where $h+1 > \sigma_k$, $[k+1, \sigma_k - 1]$ is the chunk that exhibits the behavior from earlier, with $\|\hat{s}_{\sigma_k} - s_{\sigma_k}\| \le C\rho^{\sigma_k-k-1}|\alpha|\|v_k\|$. From there, $[\sigma_k, h+1]$ receives an updated read of the state, rolling out the "chunked closed loop" $(\tilde{\pi}, f)$ in a $(\tilde{C}, \tilde{\rho})$ EISS manner until $h+1$. This means that $\|\hat{s}_{h+1} - s_{h+1}\| \le \tilde{C}\tilde{\rho}^{h+1-\sigma_k}\|\hat{s}_{\sigma_k} - s_{\sigma_k}\|$. Combining gives that $\|\hat{s}_{h+1} - s_{h+1}\| \le C\tilde{C}\rho^{\sigma_k-k-1}\tilde{\rho}^{h+1-\sigma_k}|\alpha|\|v_k\| \le C_{\mathsf{chunk}}\rho^{h-k}_{\mathsf{chunk}}|\alpha|\|v_k\|$, and conducting the same procedure gives $\|\tilde{A}^{\mathrm{cl}}_{h+1,k}v_k\| = \|\frac{d\hat{s}_{h+1}}{d\alpha}|_{\alpha=0}\| \le C_{\mathsf{chunk}}\rho^{h-k}_{\mathsf{chunk}}\|v_k\|$. □

Proposition E.10. *Let us take open-loop (C,ρ)-EISS dynamics candidate f and length-ℓ chunked policy $\tilde{\pi}$ such that $(\tilde{\pi}, f)$ is closed-loop $(\tilde{C}, \tilde{\rho})$-EISS and $\mathsf{chunk}_i[\tilde{\pi}]$ is L_{chunk}-Lipschitz and M_{chunk}-smooth, and denote s_h, a_h as the rollout of $(\tilde{\pi}, f)$ from a given initial state s_1. For alternate dynamics function f', we denote s'_h, a'_h as the rollout of $\tilde{\pi}, f'$ from $s'_1 = s_1$. For $C_{\mathsf{chunk}}, \rho_{\mathsf{chunk}}$ defined in Lemma E.9 and C_{rem} defined in Lemma E.8, we have for every $h \in [H-1]$:*

$$\max_{k\le h}\|(f'-f)(s'_k, a'_k)\| \le \frac{1}{16C_{rem}(1+\bar{\rho}^{2-2\ell})C^2_{\mathsf{chunk},stab}}$$
$$\implies \|s'_{h+1} - s_{h+1}\| \le \sum_{k=1}^{h}\bar{C}\bar{\rho}^{h-k}\|(f'-f)(s'_k, a'_k)\|$$

for $\bar{C} = 2C_{\mathsf{chunk}}$, $\bar{\rho} = \frac{1+\rho_{\mathsf{chunk}}}{2}$. Something to note, though not necessarily to fear, is that as ℓ increases the matching must shrink.

Proof. For $\tilde{A}^{\mathrm{cl}}$ defined on the rollout of $(\tilde{\pi}, f)$, we see

$$\begin{aligned}
\|s'_{h+1} - s_{h+1}\| &\le \sum_{k=1}^{h}\|\tilde{A}^{\mathrm{cl}}_{h+1,k}\left((f'-f)(s'_k, a'_k) + R_k\right)\| \quad \text{by Lemma E.8}\\
&\le \sum_{k=1}^{h} C_{\mathsf{chunk}}\rho^{h-k}_{\mathsf{chunk}}\left(\|(f'-f)(s'_k, a'_k)\| + \|R_k\|\right) \quad \text{by Lemma E.9}\\
&\le \sum_{k=1}^{h} C_{\mathsf{chunk}}\rho^{h-k}_{\mathsf{chunk}}\left(\|(f'-f)(s'_k, a'_k)\| + C_{rem}\|s'_k - s_k\|^2 + C_{rem}\|s'_{\tau_k} - s_{\tau_k}\|^2\right)
\end{aligned}$$

By application of Lemma E.4 for $C \triangleq C_{\mathsf{chunk}}$, $\rho \triangleq \rho_{\mathsf{chunk}}$, and $C_1 \triangleq C_{rem}$ we complete the proof. □

Corollary E.11. *Let us take (C,ρ) open-loop EISS f, (C,ρ) closed-loop EISS $(\pi^\star, f^\star)$, $\ell > \log(1/\rho)^{-1}\cdot\log(\mathrm{poly}(L_\pi, C))$ and define $\tilde{\pi}^\star$ as a chunked policy of length ℓ defined as $\tilde{\pi}^\star \triangleq \mathsf{chunked}(\pi^\star, f^\star, \ell)$, with $\mathsf{chunk}_i[\tilde{\pi}^\star](s) \triangleq \pi^\star((f^\star\pi^\star)^{i-1}(s))$. Let us denote s_h, a_h as the rollout of $(\tilde{\pi}^\star, f)$ and $s^\star_h, a^\star_h$ as the rollout of $(\pi^\star, f^\star)$ from a given initial state of $s_1 = s^\star_1$. For constants given in Proposition E.10, for every $h \in [H-1]$,*

$$\max_{k\le h}\|(f^\star - f)(s^\star_k, a^\star_k)\| \le \frac{1}{16C_{rem}(1+\bar{\rho}^{2-2\ell})C^2_{\mathsf{chunk},stab}}$$
$$\implies \|s^\star_{h+1} - s_{h+1}\| \le \sum_{k=1}^{h}\bar{C}\bar{\rho}^{h-k}\|(f^\star - f)(s^\star_k, a^\star_k)\|.$$

Proof. This is a special case of Proposition E.10, where $(\tilde{\pi}^\star, f)$ is $(\tilde{C}, \tilde{\rho})$-EISS by Proposition E.3 and each $\mathsf{chunk}_i[\tilde{\pi}^\star]$ is L_{chunk}-Lipschitz and M_{chunk}-smooth by Lemma E.5 on $(\pi^\star, f^\star)$. Executing $(\tilde{\pi}^\star, f^\star)$ is equivalent to $(\pi^\star, f^\star)$, as $\tilde{\pi}^\star$ is induced by the closed-loop rollout of $\pi^\star$ on $f^\star$. □

Corollary E.12. *For the setup of Proposition E.10, for every $h \in [H-1]$,*

$$\mathop{\mathbb{E}}_{d_1} \min\{1, \|s'_{h+1} - s_{h+1}\|\} \leq \left(4C_{\mathsf{chunk},stab} + 16hC_{rem}(1+\bar{\rho}^{2-2\ell})C^2_{\mathsf{chunk},stab}\right) \max_{k\leq h} \mathop{\mathbb{E}}_{s_1\sim d_1} [\|(f'-f)(s'_k, a'_k)\|]$$

Proof. For $s_1 \sim d_1$, let us define event $\mathcal{E}_h(f) = \{\max_{k\leq h} \|(f'-f)(s'_k,a'_k)\| \leq C'\}$, for $C' \triangleq \frac{1}{16C_{rem}(1+\bar{\rho}^{2-2\ell})C^2_{\mathsf{chunk},stab}}$.

$$\begin{aligned}
&\mathop{\mathbb{E}}_{s_1\sim d_1} \min\{1, \|s'_{h+1} - s_{h+1}\|\} \\
&\quad= \mathop{\mathbb{E}}_{d_1} \left[\min\{1, \|s'_{h+1} - s_{h+1}\|\}\mathbb{1}[\mathcal{E}_h(f)] + \min\{1, \|s'_{h+1} - s_{h+1}\|\}\mathbb{1}[\mathcal{E}_h(f)^C]\right] \\
&\quad\leq \mathop{\mathbb{E}}_{d_1}\left[\sum_{k=1}^h \bar{C}\bar{\rho}^{h-k}\|(f'-f)(s'_k,a'_k)\|\right] + \mathbb{P}\left[\max_{k\leq h}\|(f'-f)(s'_k,a'_k)\| \geq C'\right] \\
&\quad\leq \mathop{\mathbb{E}}_{d_1}\left[\sum_{k=1}^h \bar{C}\bar{\rho}^{h-k}\|(f'-f)(s'_k,a'_k)\|\right] + \sum_{k=1}^h \mathbb{P}\left[\|(f'-f)(s'_k,a'_k)\| \geq C'\right] \\
&\quad\leq \left(4C_{\mathsf{chunk},stab} + \frac{h}{C'}\right) \max_{k\leq h} \mathop{\mathbb{E}}_{d_1} [\|(f'-f)(s'_k,a'_k)\|]
\end{aligned}$$

The second step occurs through application of Proposition E.10, and the final step occurs through Markov's inequality. □

Proposition E.13 (Formalization of Proposition 4.4)**.** *Let $\tilde{\pi}$ be a policy $\tilde{\pi} = \mathsf{chunked}(\pi, f_\pi, \ell)$ where (π, f_π) is a (C,ρ) EISS pair and $\ell > \log(1/\rho)^{-1}\cdot\log(\mathrm{poly}(L_\pi, C))$, and let f be a (C,ρ) open-loop EISS dynamics function. For true dynamics $f^\star$, where we define $s^\star_h, a^\star_h$ to be the rollout of $(\tilde{\pi}, f^\star)$:*

$$\begin{aligned}
&|Q^{\tilde{\pi},f^\star}_H - Q^{\tilde{\pi},f}_H| \\
&\quad\leq 8C_{\mathsf{chunk},stab}C_{\max}H(1+4HC_{rem}(1+\bar{\rho}^{2-2\ell})C_{\mathsf{chunk},stab}) \max_{h\leq H} \mathop{\mathbb{E}}_{d_1} [\|(f^\star - f)(s^\star_h, a^\star_h)\|]
\end{aligned}$$

for $C_{\max} \triangleq \max\{1, L_{\mathrm{R}}, L_{\mathrm{R}}L_{\mathsf{chunk}}\}$.

Proof. We can first note that $(\tilde{\pi}, f)$ is a $(\tilde{C}, \tilde{\rho})$-EISS pair by Proposition E.3, and that $\mathsf{chunk}_i[\tilde{\pi}]$ is L_{chunk}-Lipschitz and M_{chunk}-smooth by Lemma E.5. Let us denote s_h, a_h to be the rollout of $(\tilde{\pi}, f)$

from the same initial state as $s_h^\star, a_h^\star$.

$$
\begin{aligned}
&|Q^{\tilde{\pi},f^\star} - Q^{\tilde{\pi},f}| \\
&\quad= \left| \mathop{\mathbb{E}}_{s_1\sim d_1}\left[\sum_{h=1}^{H} \mathrm{R}(s_h^\star,a_h^\star) - \mathrm{R}(s_h,a_h)\right]\right| \\
&\quad\le \mathop{\mathbb{E}}_{s_1\sim d_1}\left[\sum_{h=1}^{H} \min\left\{1, |\mathrm{R}(s_h^\star,a_h^\star) - \mathrm{R}(s_h,a_h)|\right\}\right] \quad \text{by } \mathrm{R}(\cdot,\cdot)\in[0,1] \\
&\quad\le \mathop{\mathbb{E}}_{s_1\sim d_1}\left[\sum_{h=1}^{H} \min\{1, L_{\mathrm{R}}\|s_h^\star - s_h\| + L_{\mathrm{R}}L_{\mathsf{chunk}}\|s_{\tau_h}^\star - s_{\tau_h}\|\}\right] \\
&\quad\le \max\{1, L_{\mathrm{R}}, L_{\mathrm{R}}L_{\mathsf{chunk}}\} \mathop{\mathbb{E}}_{s_1\sim d_1}\left[\sum_{h=1}^{H} \min\{1,\|s_h^\star - s_h\|\} + \min\{1,\|s_{\tau_h}^\star - s_{\tau_h}\|\}\right] \\
&\quad\le C_{\max}\sum_{h=1}^{H} 2\left(4C_{\mathsf{chunk},stab} + 16hC_{rem}(1+\bar{\rho}^{2-2\ell})C_{\mathsf{chunk},stab}^2\right) \max_{k\le H} \mathop{\mathbb{E}}_{d_1}\left[\|(f^\star - f)(s_k^\star,a_k^\star)\|\right] \\
&\quad\le 8C_{\mathsf{chunk},stab}C_{\max}H(1+4HC_{rem}(1+\bar{\rho}^{2-2\ell})C_{\mathsf{chunk},stab}) \max_{k\le H} \mathop{\mathbb{E}}_{d_1}\left[\|(f^\star - f)(s_k^\star,a_k^\star)\|\right]
\end{aligned}
$$

□

Theorem E.14. *Let there exist an expert $\pi^\star$ such that $(\pi^\star, f^\star)$ is (C,ρ) EISS. Let us have a dynamics class $\mathcal{F}$ such that each dynamics function f is (C,ρ) open-loop EISS, and let Π be a class of length-ℓ chunked policies such that for all $\tilde{\pi}\in\Pi$, $\tilde{\pi} = \mathsf{chunked}(\pi, f_\pi, \ell)$ where (π, f_π) is a (C,ρ) EISS pair and $\ell > \log(1/\rho)^{-1}\cdot\log(\mathrm{poly}(L_\pi, C))$. We assume that ℓ is fixed independently of H. We satisfy the realizability assumptions that for $\epsilon_{\mathcal{F}} \ge 0$:*

$$
\tilde{\pi}^\star \triangleq \mathsf{chunked}(\pi^\star, f^\star, \ell) \in \Pi, \quad \sup_{\tilde{\pi}\in\Pi}\inf_{f\in\mathcal{F}(\epsilon)}\max_{h\in[H]} \mathop{\mathbb{E}}_{s_1\sim d_1}\left[\|(f^\star - f)(s_h^{\tilde{\pi},f^\star}, a_h^{\tilde{\pi},f^\star})\|\right] \le \epsilon_{\mathcal{F}}
$$

then it holds that $\hat{\pi}\in\arg\max_{\tilde{\pi}\in\Pi}\inf_{f\in\mathcal{F}(\epsilon)} Q^{\tilde{\pi},f}$ satisfies

$$
J_H(\pi^\star) - J_H(\hat{\pi}) \le O(H^2)\left(\epsilon_{\mathcal{F}} + \max_{h\in[H]}\sup_{f\in\mathcal{F}(\epsilon)} \mathrm{J}_{\mathrm{model}}(f;\mu_h) + W_{\mathbb{P}}^\star\right).
$$

Proof. Let $f_{\hat{\pi}} \in \arg\min_{f\in\mathcal{F}(\epsilon)}\max_{h\le H} \mathop{\mathbb{E}}_{s_1\sim d_1}\left[\|(f^\star - f)(s_h^{\hat{\pi},f^\star}, a_h^{\hat{\pi},f^\star})\|\right]$. We first note that for all $\tilde{\pi}\in\Pi$, $f\in\mathcal{F}(\epsilon)$ it automatically holds that $(\tilde{\pi}, f)$ is $(\tilde{C},\tilde{\rho})$-EISS by Proposition E.3, and $\mathsf{chunk}_i[\tilde{\pi}]$ is L_{chunk}-Lipschitz and M_{chunk} smooth by Lemma E.5. Applications of Proposition E.13 give:

$$
\begin{aligned}
J(\pi^\star) - J(\hat{\pi}) &= J(\pi^\star) - Q^{\hat{\pi},f^\star} \pm Q^{\hat{\pi},f_{\hat{\pi}}} \\
&\le J(\pi^\star) - Q^{\hat{\pi},f_{\hat{\pi}}} + O(H^2)\epsilon_{\mathcal{F}} \quad \text{by } \mathcal{F}(\epsilon) \text{ realizability and } (\hat{\pi}, f_{\hat{\pi}}) \text{ stability} \\
&\le J(\pi^\star) - \inf_{f\in\mathcal{F}(\epsilon)} Q^{\hat{\pi},f} + O(H^2)\epsilon_{\mathcal{F}} \quad \text{by } f_{\hat{\pi}}\in\mathcal{F}(\epsilon) \\
&= J(\pi^\star) - \sup_{\tilde{\pi}\in\Pi}\inf_{f\in\mathcal{F}(\epsilon)} Q^{\tilde{\pi},f} + O(H^2)\epsilon_{\mathcal{F}} \quad \text{by pessimistic optimality} \\
&\le J(\pi^\star) - \inf_{f\in\mathcal{F}(\epsilon)} Q^{\tilde{\pi}^\star,f} + O(H^2)\epsilon_{\mathcal{F}} \quad \text{by } \Pi \text{ realizability} \\
&= Q^{\tilde{\pi}^\star,f^\star} - \inf_{f\in\mathcal{F}(\epsilon)} Q^{\tilde{\pi}^\star,f} + O(H^2)\epsilon_{\mathcal{F}} \quad \text{by the definition of } \tilde{\pi}^\star \\
&\le O(H^2)\left(\sup_{f\in\mathcal{F}(\epsilon)}\max_{h\le H} \mathop{\mathbb{E}}_{s_1\sim d_1}\left[\|(f^\star - f)(s_h^\star, a_h^\star)\|\right] + \epsilon_{\mathcal{F}}\right)
\end{aligned}
$$

□

Here, we slightly relax the realizability assumption given in the preliminaries to $\epsilon_{\mathcal{F}}$-realizability. In order to transfer the reward of our learned policy $\hat{\pi}$ to a true rollout under $f^\star$, we need to assume that at least one candidate dynamics function in our open-loop stable class is close to the true dynamics on the rollout of $\hat{\pi}$. Notably, this is not something we optimize over: instead, this is a condition that would be automatically satisfied if it held that $f^\star \in \mathcal{F}$, which cannot hold unless $f^\star$ itself is also open-loop stable.

F Extended Proofs for Section 4.3

Lemma F.1 (Key recursion with burnin)**.** *Consider non-negative sequence $\{\Delta_h\}_{h\geq 1}$ that satisfies $\Delta_{\ell+1} \leq B_\ell$ and for all $h \geq \ell+2$:*

$$\Delta_h \leq C_0\rho^{h-1-\ell}B_\ell + \sum_{k=\ell+1}^{h-1}\Big(C_1\rho^{h-1-k}\epsilon_k + C_2\rho^{h-1-k}\tau_k\Delta_k + C_4\rho^{h-1-k}\Delta_k^2 + C_5\rho^{h-1-k}\Delta_k^4\Big),$$

for constants $C_0, C_1 \geq 1$, $C_2, C_4 > 0$, $C_5 \geq 0$, $\rho \in (0,1)$, and non-negative sequences $\{\tau_k\}_{k\geq 1}$, $\{\epsilon_k\}_{k\geq 1}$. Then, as long as the following conditions hold:

$$B_\ell + \frac{\max_{\ell+1\leq k}\epsilon_k}{1-\rho} = B_\ell + \frac{\epsilon_{\max}}{1-\rho} \leq \frac{\min\{(C_4+C_5)(1+\frac{4}{1-\rho}),1\}}{2(C_4+C_5)\max\{C_0,C_1\}\left(1+\frac{4}{1-\rho}\right)^2},$$

$$\tau_k \leq \tau_{\max} \leq \frac{1}{C_2\left(1+\frac{4}{1-\rho}\right)}, \quad \forall k \geq \ell+1,$$

we have that Δ_h satisfies, for $h \geq \ell+1$,

$$\Delta_h \leq \bar{C}\bar{\rho}^{h-1-\ell}B_\ell + \bar{C}\sum_{k=\ell+1}^{h-1}\bar{\rho}^{h-1-k}\epsilon_k,$$

where $\bar{C} = \max\{C_0,C_1\}\left(1+\frac{4}{1-\rho}\right)$, $\bar{\rho} = \frac{1+\rho}{2}$.

Proof. Let us have $\{\bar{\Delta}_h\}_{h\geq\ell+1}$ that admits form $\bar{\Delta}_h \triangleq \bar{C}\bar{\rho}^{h-1-\ell}B_\ell + \bar{C}\sum_{k=\ell+1}^{h-1}\bar{\rho}^{h-1-k}\epsilon_k$, $h \geq \ell+2$, where $\bar{\Delta}_{\ell+1} \triangleq \bar{C}B_\ell$. Through induction, we will prove that $1 \geq \bar{\Delta}_h \geq \Delta_h$ for all $h \geq \ell+1$. We will determine $\bar{\rho} \in (\rho,1)$, $\bar{C} \geq \max\{C_0,C_1\}$ in hindsight. As in the statement, we further impose the constraints $\epsilon_k \leq \epsilon_{\max}$, $\tau_k \leq \tau_{\max}$, $k \geq \ell+1$, where $\epsilon_{\max}, \tau_{\max}$ will be set in hindsight.

For the base-case $h = \ell+1$, we have trivially $\Delta_{\ell+1} \leq B_\ell \leq \bar{C}B_\ell = \bar{\Delta}_{\ell+1}$ and we see that $B_\ell \lesssim \frac{1}{\bar{C}}$. It remains to determine that $1 \geq \bar{\Delta}_h \geq \Delta_h$ for all $h \geq \ell+2$. Now, given $\Delta_k \leq \bar{\Delta}_k \leq 1$ for all $k = \ell+1,\ldots,h-1$, we seek to establish the induction step $\Delta_h \leq \bar{\Delta}_h \leq 1$. Starting with Δ_h, we may plug in $\Delta_k \leq \bar{\Delta}_k \leq 1$, $k \leq h-1$ into the bound on Δ_h to yield:

$$\begin{aligned}\Delta_h &\leq C_0\rho^{h-1-\ell}B_\ell + \sum_{k=\ell+1}^{h-1}\Big(C_1\rho^{h-1-k}\epsilon_k + C_2\rho^{h-1-k}\tau_k\Delta_k + (C_4+C_5)\rho^{h-1-k}\Delta_k^2\Big)\\ &\leq C_0\rho^{h-1-\ell}B_\ell + \sum_{k=\ell+1}^{h-1}\big(C_1\rho^{h-1-k}\epsilon_k + C_2\rho^{h-1-k}\tau_k\bar{\Delta}_k + (C_4+C_5)\rho^{h-1-k}\bar{\Delta}_k^2\big).\end{aligned}$$

We now treat each summand corresponding to C_0, C_1, C_2, C_4+C_5 separately. The first two terms in C_0, C_1 straightforwardly satisfy $\lesssim \bar{\Delta}_h$ since $\rho \leq \bar{\rho}$ and $C_0, C_1 \leq \bar{C}$. Noting that:

$$\sum_{k=\ell+1}^{h-1}\rho^{h-1-k}\bar{\Delta}_k \leq \frac{\bar{C}}{\bar{\rho}-\rho}\left(\bar{\rho}^{h-1-\ell}B_\ell + \sum_{k=\ell+1}^{h-2}\bar{\rho}^{h-k-1}\epsilon_k\right)$$

We can expand the following to bound the third term:

$$\sum_{k=\ell+1}^{h-1} C_2\rho^{h-1-k}\tau_k\bar{\Delta}_k = \sum_{k=\ell+1}^{h-1} C_2\rho^{h-1-k}\tau_k \cdot \left(\bar{C}\bar{\rho}^{k-1-\ell}B_\ell + \bar{C}\sum_{j=\ell+1}^{k-1}\bar{\rho}^{k-1-j}\epsilon_j\right)$$
$$\leq \sum_{k=\ell+1}^{h-1} C_2\bar{C}\rho^{h-1-k}\tau_k\bar{\rho}^{k-1-\ell}B_\ell + \sum_{k=\ell+1}^{h-1} C_2\rho^{h-1-k}\tau_k\cdot\bar{C}\sum_{j=\ell+1}^{k-1}\bar{\rho}^{k-1-j}\epsilon_j$$
$$\leq \frac{C_2\bar{C}\tau_{\max}}{\bar{\rho}-\rho}\bar{\rho}^{h-\ell-1}B_\ell + \frac{C_2\bar{C}\tau_{\max}}{\bar{\rho}-\rho}\sum_{k=\ell+1}^{h-1}\bar{\rho}^{h-k-1}\epsilon_k.$$

Therefore, setting $\tau_{\max}$ sufficiently small $\tau_{\max} \lesssim \frac{\bar{\rho}-\rho}{C_2}$ ensures the second summand satisfies $\lesssim \bar{C}\bar{\rho}^{h-1-\ell}B_\ell + \bar{C}\sum_{k=\ell+1}^{h-1}\bar{\rho}^{h-k-1}\epsilon_k = \bar{\Delta}_h$. We may now treat the second-order term corresponding to $C_4 + C_5$ similarly: since by assumption $\epsilon_k \leq \epsilon_{\max}$, $k \geq \ell+1$, we have $\bar{\Delta}_k \leq \bar{C}B_\ell + \frac{\bar{C}}{1-\bar{\rho}}\epsilon_{\max}$ for all $k \geq \ell+1$. Thus, we follow similar steps to bound:

$$\sum_{k=\ell+1}^{h-1}(C_4+C_5)\rho^{h-1-k}\bar{\Delta}_k^2 \leq (C_4+C_5)\left(\bar{C}B_\ell + \frac{\bar{C}\epsilon_{\max}}{1-\bar{\rho}}\right)\sum_{k=\ell+1}^{h-1}\rho^{h-1-k}\bar{\Delta}_k$$
$$\leq \frac{(C_4+C_5)\bar{C}^2}{\bar{\rho}-\rho}\left(B_\ell + \frac{\epsilon_{\max}}{1-\bar{\rho}}\right)\left(\bar{\rho}^{h-1-\ell}B_\ell + \sum_{k=\ell+1}^{h-1}\bar{\rho}^{h-k-1}\epsilon_k\right).$$

Therefore, setting $B_\ell, \epsilon_{\max}$ sufficiently small $B_\ell + \frac{\epsilon_{\max}}{1-\bar{\rho}} \lesssim \frac{\bar{\rho}-\rho}{(C_4+C_5)\bar{C}}$ ensures that this term satisfies $\lesssim \bar{C}\bar{\rho}^{h-1-\ell}B_\ell + \bar{C}\sum_{k=\ell+1}^{h-1}\bar{\rho}^{h-k-1}\epsilon_k = \bar{\Delta}_h$.

Now, to complete the induction step on Δ_h, we determine the value of $\bar{C}$ in hindsight. Setting $\bar{\rho} = \frac{1+\rho}{2}$, we return to the bound on Δ_h, where we may collect all the bounds on the C_1, C_2, C_4+C_5 terms to get:

$$\Delta_h \leq \left(\max\{C_0,C_1\} + \frac{C_2\bar{C}\tau_{\max}}{\bar{\rho}-\rho} + \frac{(C_4+C_5)\bar{C}^2}{\bar{\rho}-\rho}\left(B_\ell + \frac{\epsilon_{\max}}{1-\bar{\rho}}\right)\right)\left(\bar{\rho}^{h-1-\ell}B_\ell + \sum_{k=\ell+1}^{h-1}\bar{\rho}^{h-k-1}\epsilon_k\right)$$
$$\leq \left(\max\{C_0,C_1\} + \frac{2}{1-\rho}\left(C_2\bar{C}\tau_{\max} + (C_4+C_5)\bar{C}^2\left(B_\ell + \frac{2\epsilon_{\max}}{1-\rho}\right)\right)\right)\left(\bar{\rho}^{h-1-\ell}B_\ell + \sum_{k=\ell+1}^{h-1}\bar{\rho}^{h-k-1}\epsilon_k\right). \tag{3}$$

It remains to set bounds on $\epsilon_{\max}, \tau_{\max}$ and set $\bar{C}$ such that the RHS satisfies $\leq \bar{C}\bar{\rho}^{h-1-\ell}B_\ell + \bar{C}\sum_{k=\ell+1}^{h-1}\bar{\rho}^{h-1-k}\epsilon_k$. Intuitively, we may tune $\epsilon_{\max}, \tau_{\max}$ such that the C_2, C_4+C_5 terms are as small as needed. Thus, setting $\bar{C} = \max\{C_0,C_1\}\left(1+\frac{4}{1-\rho}\right)$, we may set the constraints in hindsight:

$$\tau_{\max} \leq \frac{\max\{C_0,C_1\}}{C_2\bar{C}} = \frac{1}{C_2\left(1+\frac{4}{1-\rho}\right)}$$
$$B_\ell + \frac{\epsilon_{\max}}{1-\rho} \leq \frac{\max\{C_0,C_1\}}{2\bar{C}^2(C_4+C_5)} = \frac{1}{2(C_4+C_5)\max\{C_0,C_1\}\left(1+\frac{4}{1-\rho}\right)^2}$$

Finally, we must enforce $\bar{\Delta}_h \leq 1$: this may be achieved by setting $\bar{\Delta}_h \leq \bar{C}\left(B_\ell + \frac{\epsilon_{\max}}{1-\bar{\rho}}\right) \leq 1$, or $B_\ell + \frac{\epsilon_{\max}}{1-\rho} \leq \left(2\max\{C_0,C_1\}(1+\frac{4}{1-\rho})\right)^{-1}$. This all gives us the desired bound:

$$\Delta_h \leq \bar{C}\bar{\rho}^{h-1-\ell}B_\ell + \bar{C}\sum_{k=\ell+1}^{h-1}\bar{\rho}^{h-1-k}\epsilon_k$$
$$\leq \max\{C_0,C_1\}\left(1+\frac{4}{1-\rho}\right)\left(\left(\frac{1+\rho}{2}\right)^{h-1-\ell}B_\ell + \sum_{k=\ell+1}^{h-1}\left(\frac{1+\rho}{2}\right)^{h-1-k}\epsilon_k\right),$$

completing the induction step $\Delta_h \leq \bar{\Delta}_h \leq 1$ and the full proof. □

Lemma F.2. *Let $\pi^\star, f^\star$ be a (C,ρ) EISS pair. Fix $f \in \mathcal{F}$ and $h \in [\ell+1, H-1]$. Given $s_1 \sim d_1$, if for $C_1 \triangleq C\left(M_\pi L_f + M_f(1+2L_\pi^2+2M_\pi^2)\right)$ the following holds:*

$$D_k^\star(f) \triangleq \nabla_s(f-f^\star)(s_k^{\pi^\star,f^\star}, a_k^{\pi^\star,f^\star}) + \nabla_s\pi^\star(s_k^{\pi^\star,f^\star})\nabla_a(f-f^\star)(s_k^{\pi^\star,f^\star}, a_k^{\pi^\star,f^\star})$$

$$B_\ell(f,s_1) \triangleq \max_{1\leq k\leq \ell+1} \|s_k^{\pi^\star,f} - s_k^{\pi^\star,f^\star}\|$$

$$B_\ell(f,s_1) + \frac{1}{1-\rho}\max_{\ell+1\leq k\leq h} \|(f^\star - f)(s_k^{\pi^\star,f^\star}, a_k^{\pi^\star,f^\star})\| \leq \frac{\min\{C_1(1+\frac{4}{1-\rho}),1\}}{2C_1C(1+\frac{4}{1-\rho})^2} \triangleq C_2$$

$$\max_{\ell+1\leq k\leq h} \|D_k^\star(f)\| \leq \frac{1}{C(1+\frac{4}{1-\rho})} \triangleq C_3$$

then we have that for $\bar{C} \triangleq C + \frac{4C}{1-\rho}$ and $\bar{\rho} \triangleq \frac{1+\rho}{2}$:

$$\|s_{h+1}^{\pi^\star,f^\star} - s_{h+1}^{\pi^\star,f}\| \leq \bar{C}\bar{\rho}^{h-\ell}B_\ell(f,s_1) + \sum_{k=\ell+1}^{h} \bar{C}\bar{\rho}^{h-k}\|f^\star(s_k^{\pi^\star,f^\star}, a_k^{\pi^\star,f^\star}) - f(s_k^{\pi^\star,f^\star}, a_k^{\pi^\star,f^\star})\|$$

Proof. We begin by expanding the difference in trajectories and Taylor expanding $f(s_h^{\pi^\star,f}, a_h^{\pi^\star,f})$ about $s_h^{\pi^\star,f^\star}$. For ease of notation, we denote $s_h^{\pi^\star,f} \triangleq s_h$, $s_h^{\pi^\star,f^\star} \triangleq s_h^\star$, and Jacobians as in other lemmas.

$$\begin{aligned}
s_{h+1} - s_{h+1}^\star &= f(s_h,a_h) - f(s_h^\star,a_h^\star) + f(s_h^\star,a_h^\star) - f^\star(s_h^\star,a_h^\star) \\
&= (f-f^\star)(s_h^\star,a_h^\star) + \nabla_s f(s_h^\star,a_h^\star)^\top(s_h - s_h^\star) + \nabla_a f(s_h^\star,a_h^\star)^\top(a_h - a_h^\star) + r_h^X \\
&= (f-f^\star)(s_h^\star,a_h^\star) + \nabla_s f(s_h^\star,a_h^\star)^\top(s_h - s_h^\star) + \\
&\quad \nabla_a f(s_h^\star,a_h^\star)^\top(\nabla_s\pi^\star(s_h^\star)^\top(s_h - s_h^\star) + r_h^U) + r_h^X \\
&= (f-f^\star)(s_h^\star,a_h^\star) + A_h^{\star f}(s_h - s_h^\star) + B_h^{\star f}K_h^\star(s_h - s_h^\star) + B_h^{\star f}r_h^U + r_h^X \\
&= (f-f^\star)(s_h^\star,a_h^\star) + (A_h^{\star f} + B_h^{\star f}K_h^\star - A_h^{\star f^\star} - B_h^{\star f^\star}K_h^\star)(s_h - s_h^\star) \\
&\quad + (A_h^{\star f^\star} + B_h^{\star f^\star}K_h^\star)(s_h - s_h^\star) + B_h^{\star f}r_h^U + r_h^X \\
&\triangleq (f-f^\star)(s_h^\star,a_h^\star) + D_h^\star(f)^\top(s_h - s_h^\star) + B_h^{\star f}r_h^U + r_h^X + (A_h^{\star f^\star} + B_h^{\star f^\star}K_h^\star)(s_h - s_h^\star)
\end{aligned}$$

where I give $A_h^{\star f} \triangleq \nabla_s f(s_h^\star,a_h^\star)^\top$, $B_h^{\star f} \triangleq \nabla_a f(s_h^\star,a_h^\star)^\top$, $K_h^\star \triangleq \nabla_s\pi^\star(s_h^\star)^\top$, and likewise for $f^\star$. Seeing this as the full one-step expansion, let us now roll out the recursion before expanding out the projections. Let $A_h^{\mathrm{cl}} \triangleq A_h^{\star f^\star} + B_h^{\star f^\star}K_h^\star$.

$$s_{h+1} - s_{h+1}^\star = \sum_{k=1}^{h} A_{k+1:h+1}^{\mathrm{cl}}\left((f-f^\star)(s_k^\star,a_k^\star) + D_k^{\star\top}(f)(s_k - s_k^\star) + B_k^{\star f}r_k^U + r_k^X\right)$$

Let me denote $\epsilon_k \triangleq \|(f^\star - f)(s_k^\star,a_k^\star)\|$, $\Delta_k \triangleq \|s_k - s_k^\star\|$, $\tau_k \triangleq \|D_k^\star(f)\|$. We have no control of the Jacobians τ_k for the first ℓ timesteps, and so we must bound the early trajectory error directly using Lipschitzity.

$$\begin{aligned}
\|s_{k+1}^\star - s_{k+1}\| &= \|f^\star(s_k^\star,a_k^\star) - f(s_k^\star,a_k^\star) + f(s_k^\star,a_k^\star) - f(s_k,a_k)\| \\
&\leq \|(f^\star - f)(s_k^\star,a_k^\star)\| + L_f(1+L_\pi)\|s_k^\star - s_k\|
\end{aligned}$$

Since $\Delta_1 = 0$, this gives burn-in error $\Delta_{\ell+1} \leq B_\ell(f, s_1)$. Having this, we can consider the trajectory-difference recursion from time $\ell + 1$ onwards:

$$s_{h+1} - s^\star_{h+1} = A^{\mathrm{cl}}_{\ell+1:h+1}(s_{\ell+1} - s^\star_{\ell+1}) + \sum_{k=\ell+1}^{h} A^{\mathrm{cl}}_{k+1:h+1} \left((f - f^\star)(s^\star_k, a^\star_k) + D^{\star\top}_k(f)(s_k - s^\star_k) + B^{\star f}_k r^U_k + r^X_k \right)$$

$$\Delta_{h+1} \leq C\rho^{h-\ell} B_\ell(f, s_1) + \sum_{k=\ell+1}^{h} \Big(C\rho^{h-k}\epsilon_k + C\rho^{h-k}\tau_k\Delta_k + C\rho^{h-k}\left(M_\pi L_f \Delta_k^2 + M_f(1 + 2L_\pi^2 + 2M_\pi^2\Delta_k^2)\Delta_k^2\right) \Big)$$

Let $C_0 = C_1 = C_2 = C$, $C_4 = C(M_\pi L_f + M_f + 2M_f L_\pi^2)$, and $C_5 = 2CM_f M_\pi^2$. Application of Lemma F.1 gives us the result. □

Lemma F.3. *Let $\pi^\star, f^\star$ be a (C, ρ) EISS pair. For terms defined in Lemma F.2, for $\ell + 1 \leq h \leq H - 1$, it holds that:*

$$\mathop{\mathbb{E}}_{d_1}\big[\min\{1, \|s^{\pi^\star, f^\star}_{h+1} - s^{\pi^\star, f}_{h+1}\|\}\big] \lesssim O(1)\mathop{\mathbb{E}}_{d_1}[B_\ell(f, s_1)] + O(h) \max_{\ell+1\leq k\leq h} \left(\sqrt{\mathop{\mathbb{E}}_{d_1}\left[\|(f^\star - f)(s^\star_k, a^\star_k)\|^2\right]} + \sqrt{\mathop{\mathbb{E}}_{d_1}\left[\|D^\star_k(f)\|^2\right]} \right)$$

Proof. For ease of notation, let us denote $s^{\pi^\star, f}_h, a^{\pi^\star, f}_h \triangleq s_h, a_h$ and $s^{\pi^\star, f^\star}_h, a^{\pi^\star, f^\star}_h \triangleq s^\star_h, a^\star_h$. Let us define the event $\mathcal{E}_f$ under which the conditions of Lemma F.2 hold.

$$\mathcal{E}_f \triangleq \left\{ B_\ell(f, s_1) + \frac{\max_{\ell+1\leq k\leq h} \|(f^\star - f)(s^\star_k, a^\star_k)\|}{1 - \rho} \leq C_2,\ \max_{\ell+1\leq k\leq h} \|D^\star_k(f)\| \leq C_3 \right\}$$

Now we can use this to break up the trajectory difference. The following proceeds through applications of union bound and Markov's inequality.

$$\begin{aligned}
&\mathop{\mathbb{E}}_{d_1}\left[\min\{1, \|s^\star_{h+1} - s_{h+1}\|\}\right] \\
&= \mathop{\mathbb{E}}_{d_1}\left[\min\{1, \|s^\star_{h+1} - s_{h+1}\|\}\mathbb{1}[\mathcal{E}_f]\right] + \mathop{\mathbb{E}}_{d_1}\left[\min\{1, \|s^\star_{h+1} - s_{h+1}\|\}\mathbb{1}[\mathcal{E}_f^C]\right] \\
&\leq \bar{C}\bar{\rho}^{h-\ell}\mathop{\mathbb{E}}_{d_1}[B_\ell(f, s_1)] + \sum_{k=\ell+1}^{h} \bar{C}\bar{\rho}^{h-k}\mathop{\mathbb{E}}_{d_1}\left[\|(f^\star - f)(s^\star_k, a^\star_k)\|\right] \\
&\quad + \mathbb{P}_{d_1}\left[B_\ell(f, s_1) + \frac{\max_{\ell+1\leq k\leq h} \|(f^\star - f)(s^\star_k, a^\star_k)\|}{1 - \rho} \geq C_2 \text{ or } \max_{\ell+1\leq k\leq h} \|D^\star_k(f)\| \geq C_3 \right] \\
&\leq \left(\bar{C}\bar{\rho}^{h-\ell} + \frac{1}{C_2}\right)\mathop{\mathbb{E}}_{d_1}[B_\ell(f, s_1)] + \sum_{k=\ell+1}^{h} \left(\bar{C}\bar{\rho}^{h-k} + \frac{1}{C_2(1-\rho)}\right)\mathop{\mathbb{E}}_{d_1}\left[\|(f^\star - f)(s^\star_k, a^\star_k)\|\right] \\
&\quad + \sum_{k=\ell+1}^{h} \left(\frac{1}{C_3}\mathop{\mathbb{E}}_{d_1}[\|D^\star_k(f)\|]\right)
\end{aligned}$$

□

Proposition F.4. *Let $\pi^\star, f^\star$ be a (C, ρ) EISS pair, and let us denote $s^\star_h \triangleq s^{\pi^\star, f^\star, s_1}_h$ and $a^\star_h \triangleq \pi^\star(s^\star_h)$ for some initial state $s_1 \sim d_1$. Candidate model f satisfies:*

$$J_H(\pi^\star) - Q^{\pi^\star, f}_H \lesssim H\ell^2 \max\left\{1, (L_f(1 + L_\pi))^\ell\right\} \max_{1\leq h\leq \ell} \mathrm{J}_{\mathrm{model}}(f; \mathbb{P}^\star_h) + (H - \ell)^2 \left(\max_{\ell+1\leq h\leq H} \mathrm{J}_{\mathrm{model}}(f; \mathbb{P}^\star_h) + \max_{\ell+1\leq h\leq H} \sqrt{\mathop{\mathbb{E}}_{\mathbb{P}^\star_h}\|D^\star_h(f)\|^2} \right)$$

Proof. Let us denote $s_h \triangleq s_h^{\pi^\star, f, s_1}$ with $a_h \triangleq \pi^\star(s_h)$ for fixed candidate f. We note that since $\mathrm{R}(\cdot,\cdot) \in [0,1]$ the difference must be similarly constrained.

$$
\begin{aligned}
J(\pi^\star) - Q^{\pi^\star, f} &= \sum_{h=1}^{H} \mathop{\mathbb{E}}_{d_1}\left[\mathrm{R}(s_h^{\pi^\star,f^\star}, a_h^{\pi^\star,f^\star}) - \mathrm{R}(s_h^{\pi^\star,f}, a_h^{\pi^\star,f})\right] \\
&= \sum_{h=1}^{H} \mathop{\mathbb{E}}_{d_1}\left[\min\left\{1, \mathrm{R}(s_h^{\pi^\star,f^\star}, a_h^{\pi^\star,f^\star}) - \mathrm{R}(s_h^{\pi^\star,f}, a_h^{\pi^\star,f})\right\}\right] \\
&\le \sum_{h=1}^{H} \mathop{\mathbb{E}}_{d_1}\left[\min\left\{1, |\mathrm{R}(s_h^{\pi^\star,f^\star}, a_h^{\pi^\star,f^\star}) - \mathrm{R}(s_h^{\pi^\star,f}, a_h^{\pi^\star,f})|\right\}\right] \\
&\le \max\{1, L_{\mathrm{R}}(1+L_\pi)\} \sum_{h=1}^{H} \mathop{\mathbb{E}}_{d_1}\left[\min\left\{1, \|s_h^{\pi^\star,f^\star} - s_h^{\pi^\star,f}\|\right\}\right] \\
&\le (H-\ell)\max\{1, L_{\mathrm{R}}(1+L_\pi)\} \max_{\ell+1\le h\le H} \mathop{\mathbb{E}}_{d_1}\left[\min\left\{1, \|s_h^{\pi^\star,f^\star} - s_h^{\pi^\star,f}\|\right\}\right] \\
&\quad + \ell \max\{1, L_{\mathrm{R}}(1+L_\pi)\} \mathop{\mathbb{E}}_{d_1}[B_\ell(f, s_1)]
\end{aligned}
$$

For brevity, let us define $L \triangleq \max\{1, L_{\mathrm{R}}(1+L_\pi)\}$. For bounding the burn-in term, we can expand the following: this is obviously exponential in ℓ and would compound if $L_f(1+L_\pi) > 1$; however, since it's this is a worst-case bound and it's only up to a finite burn-in we deem this acceptable.

$$
\begin{aligned}
\mathop{\mathbb{E}}_{d_1}[B_\ell(f, s_1)] &= \mathop{\mathbb{E}}_{d_1}\left[\max_{1\le k\le \ell+1} \|s_k^{\pi^\star,f} - s_k^{\pi^\star,f^\star}\|\right] \\
&\le \mathop{\mathbb{E}}_{d_1}\left[\sum_{k=1}^{\ell+1} \|s_k^{\pi^\star,f} - s_k^{\pi^\star,f^\star}\|\right] \\
&\le \mathop{\mathbb{E}}_{d_1}\left[\sum_{k=2}^{\ell+1}\sum_{j=1}^{k-1} (L_f(1+L_\pi))^{k-1-j} \|(f-f^\star)(s_j^{\pi^\star,f^\star}, a_j^{\pi^\star,f^\star})\|\right] \\
&\le \ell\left(\sum_{k=0}^{\ell-1} (L_f(1+L_\pi))^k\right) \max_{1\le k\le \ell} \mathop{\mathbb{E}}_{\mathbb{P}_k^\star} \|(f-f^\star)(s_k, a_k)\|
\end{aligned}
$$

More important is the fact that this mandates that we control the zeroth order dynamics deviation even for $k \le \ell$. We'll define $L_{\mathrm{cl}} \triangleq L_f(1+L_\pi)$ once again for brevity. Application of Lemma F.3 bounds the state difference, but we can now insert our bound on $\mathop{\mathbb{E}}_{d_1}[B_\ell(f, s_1)]$ to expand it further.

$$
\begin{aligned}
&\mathop{\mathbb{E}}_{d_1}\left[\min\left\{1, \|s_h^{\pi^\star,f^\star} - s_h^{\pi^\star,f}\|\right\}\right] \\
&\le \left(\bar{C}\bar{\rho}^{h-\ell-1} + \frac{1}{C_2}\right) \mathop{\mathbb{E}}_{d_1}[B_\ell(f, s_1)] + \sum_{k=\ell+1}^{h-1}\left(\frac{1}{C_3}\mathop{\mathbb{E}}_{d_1}[\|D_k^\star(f)\|]\right) \\
&\quad + \sum_{k=\ell+1}^{h-1}\left(\bar{C}\bar{\rho}^{h-k-1} + \frac{1}{C_2(1-\rho)}\right) \mathop{\mathbb{E}}_{d_1}[\|(f^\star - f)(s_k^\star, a_k^\star)\|] \\
&\le \ell\left(\bar{C}\bar{\rho}^{h-\ell-1} + \frac{1}{C_2}\right)\left(\sum_{k=0}^{\ell-1} L_{\mathrm{cl}}^k\right) \max_{1\le k\le \ell} \mathop{\mathbb{E}}_{\mathbb{P}_k^\star} \|(f-f^\star)(s_k, a_k)\| + \frac{h-\ell}{C_3} \max_{\ell+1\le k\le h} \mathop{\mathbb{E}}_{\mathbb{P}_k^\star}[\|D_k^\star(f)\|] \\
&\quad + \left(\frac{2\bar{C} + (h-\ell)/C_2}{1-\rho}\right) \max_{\ell+1\le k\le h} \mathop{\mathbb{E}}_{\mathbb{P}_k^\star}[\|(f^\star - f)(s_k, a_k)\|]
\end{aligned}
$$

For every $\ell+1 \le h \le H$, $\bar{\rho}^{h-\ell-1} \le 1$ and $h-\ell \le H-\ell$, so we can take a max over post-burn-in timesteps to get

$$\begin{aligned}
&\max_{\ell+1\le h\le H} \mathbb{E}_{d_1}\left[\min\left\{1, \|s_h^{\pi^\star,f^\star} - s_h^{\pi^\star,f}\|\right\}\right] \\
&\quad\le \ell\left(\bar{C}+\frac{1}{C_2}\right)\left(\sum_{k=0}^{\ell-1} L_{\mathrm{cl}}^k\right) \max_{1\le k\le \ell} \mathbb{E}_{\mathbb{P}_k^\star}\|(f-f^\star)(s_k,a_k)\| + \frac{H-\ell}{C_3}\max_{\ell+1\le k\le H} \mathbb{E}_{\mathbb{P}_k^\star}[\|D_k^\star(f)\|] \\
&\qquad + \left(\frac{2\bar{C}+(H-\ell)/C_2}{1-\rho}\right)\max_{\ell+1\le k\le H} \mathbb{E}_{\mathbb{P}_k^\star}[\|(f^\star-f)(s_k,a_k)\|]
\end{aligned}$$

and putting this together we see

$$\begin{aligned}
J_H(\pi^\star) - Q_H^{\pi^\star,f} \le (H-\ell)L\Bigg(&\ell\left(\bar{C}+\frac{1}{C_2}\right)\left(\sum_{k=0}^{\ell-1} L_{\mathrm{cl}}^k\right)\max_{1\le k\le \ell} \mathbb{E}_{\mathbb{P}_k^\star}\|(f-f^\star)(s_k,a_k)\| \\
&+ \frac{H-\ell}{C_3}\max_{\ell+1\le k\le H} \mathbb{E}_{\mathbb{P}_k^\star}[\|D_k^\star(f)\|] \\
&+ \left(\frac{2\bar{C}+(H-\ell)/C_2}{1-\rho}\right)\max_{\ell+1\le k\le H} \mathbb{E}_{\mathbb{P}_k^\star}[\|(f^\star-f)(s_k,a_k)\|]\Bigg) \\
&+ \ell^2 L\left(\sum_{k=0}^{\ell-1} L_{\mathrm{cl}}^k\right)\max_{1\le k\le \ell} \mathbb{E}_{\mathbb{P}_k^\star}\|(f-f^\star)(s_k,a_k)\|
\end{aligned}$$

We can combine the pre-burn-in model difference coefficients

$$\begin{aligned}
(H-\ell)L\ell\left(\bar{C}+\frac{1}{C_2}\right) + \ell^2 L &= \ell L\left(\ell + (H-\ell)\left(\bar{C}+\frac{1}{C_2}\right)\right) \\
&\le \ell L H \max\left\{1, \bar{C}+\frac{1}{C_2}\right\} \lesssim H\ell
\end{aligned}$$

and we can simplify the post-burn-in model difference coefficient

$$(H-\ell)L\left(\frac{2\bar{C}+(H-\ell)/C_2}{1-\rho}\right) \le \frac{2\bar{C}+1/C_2}{1-\rho}(H-\ell)^2 L \lesssim (H-\ell)^2.$$

Putting this all together gives us

$$\begin{aligned}
J_H(\pi^\star) - Q_H^{\pi^\star,f} \lesssim\ & H\ell\left(\sum_{h=0}^{\ell-1} L_{\mathrm{cl}}^h\right)\max_{1\le h\le \ell} \mathrm{J}_{\mathrm{model}}(f;\mathbb{P}_h^\star) \\
&+ (H-\ell)^2\left(\max_{\ell+1\le h\le H} \mathrm{J}_{\mathrm{model}}(f;\mathbb{P}_h^\star) + \max_{\ell+1\le h\le H}\sqrt{\mathbb{E}_{\mathbb{P}_h^\star}\|D_h^\star(f)\|^2}\right)
\end{aligned}$$

□

Proposition F.5 (Adapted from Proposition D.2 from Zhang et al. (2025))**.** *Consider the noise-injected distribution $\tilde{\nu}_h^{\sigma_u}$ as defined in Definition 4.10. Consider the linearizations along an expert trajectory given in Definition 4.7. Define the linear and residual components of the noised state $\tilde{s}_h$:*

$$\tilde{s}_h^{lin} \triangleq s_h^{\pi^\star} + \sum_{k=1}^{h-1} A_{k+1:h}^{cl} B_k u_k, \quad \tilde{s}_h^{res} \triangleq \tilde{s}_h - \tilde{s}_h^{lin}, \quad h \ge 1.$$

Then, as long as $\sigma_u \le \frac{1}{2}\frac{\sqrt{1+4L_\pi^2}}{C_{stab}M_\pi}$, and defining $C_r \triangleq M_\pi + 4M_f(1+4L_\pi^2)$, we have $\|\tilde{s}_h^{res}\| \le C_{stab}^3 C_r \sigma_u^2$, $h \ge 1$ almost surely over $\tilde{s}_1 \sim d_1$ and $\{z_k\} \sim \mathcal{D}(0,\Sigma_z)$.

Proposition F.6. *Let $\pi^\star, f^\star$ be a (C,ρ) EISS pair satisfying Assumption 4.9. Denoting $s_h^\star, a_h^\star$ to be generated from trajectories of $\pi^\star, f^\star$ given $s_1 \sim d_1$, for every candidate dynamics function $f \in \mathcal{F}$,*

and for σ_u satisfying Proposition F.5, for every $H \geq h > \ell$,

$$
\begin{aligned}
D_h^\star(f) &\triangleq \nabla_s(f-f^\star)(s_h^\star,a_h^\star) + \nabla_s\pi^\star(s_h^\star)\nabla_a(f-f^\star)(s_h^\star,a_h^\star)\\
C_{rem,2} &\triangleq 2L_f(1+L_\pi)C_{stab}^3C_r + 2M_f(1+3L_\pi^2)C_{stab}^2 + 6M_f + 2L_fM_\pi C_{stab}^2\\
C_{rem,4} &\triangleq 6M_fM_\pi^2C_{stab}^4\\
\|D_h^\star(f)\|^2 &\leq \frac{2d_u+4}{c_\ell\sigma_u^2}\mathop{\mathbb{E}}_{\tilde\nu_h|s_1}\|(f-f^\star)(\tilde s_h,\tilde a_h)\|^2 + \frac{2d_u+4}{c_\ell}(C_{rem,2}+C_{rem,4}\sigma_u^2)^2\sigma_u^2
\end{aligned}
$$

Proof. Let me denote $s_h^{\pi^\star,f^\star} \triangleq s_h^\star$ for the time being. As such, I will take $(f-f^\star)(\tilde s_h,\tilde a_h)$ where $\tilde s_h,\tilde a_h$ are generated from $\tilde\nu_h$ and Taylor expand it about $s_h^\star,a_h^\star$. Let me denote $\bar s_h = \tilde s_h^{lin}$ and $s_h^\Delta = \tilde s_h^{res}$, such that $\tilde s_h = \bar s_h + s_h^\Delta$.

$$
\begin{aligned}
\tilde a_h - a_h^\star &= \nabla_s\pi^\star(s_h^\star)^\top(\tilde s_h - s_h^\star) + r_h^U + (\tilde a_h - \pi^\star(\tilde s_h))\\
(f-f^\star)(\tilde s_h,\tilde a_h) &= (f-f^\star)(s_h^\star,a_h^\star) + \nabla_s(f-f^\star)(s_h^\star,a_h^\star)^\top(\tilde s_h - s_h^\star) + \nabla_a(f-f^\star)(s_h^\star,a_h^\star)^\top(\tilde a_h - a_h^\star) + r_h^X\\
&= (f-f^\star)(s_h^\star,a_h^\star) + \nabla_s(f-f^\star)(s_h^\star,a_h^\star)^\top(\bar s_h - s_h^\star)\\
&\quad + \nabla_a(f-f^\star)(s_h^\star,a_h^\star)^\top(\nabla_s\pi^\star(s_h^\star)^\top(\bar s_h - s_h^\star) + u_h + r_h^U)\\
&\quad + \nabla_s(f-f^\star)(s_h^\star,a_h^\star)^\top s_h^\Delta + \nabla_a(f-f^\star)(s_h^\star,a_h^\star)^\top\nabla_s\pi^\star(s_h^\star)^\top s_h^\Delta + r_h^X\\
&= (f-f^\star)(s_h^\star,a_h^\star) + D_h^\star(f)^\top(\bar s_h - s_h^\star) + D_h^\star(f)^\top s_h^\Delta + \nabla_a(f-f^\star)(s_h^\star,a_h^\star)^\top(u_h + r_h^U) + r_h^X\\
\|r_h^U\| &\leq M_\pi\|\tilde s_h - s_h^\star\|^2\\
&\leq M_\pi C_{stab}^2\sigma_u^2\\
\|r_h^X\| &\leq M_f\left\|\begin{bmatrix}\tilde s_h - s_h^\star\\ \tilde a_h - a_h^\star\end{bmatrix}\right\|^2\\
&\leq 2M_f\|\tilde s_h - s_h^\star\|^2 + 6M_f(L_\pi^2\|\tilde s_h - s_h^\star\|^2 + \|u_h\|^2 + \|r_h^U\|^2)\\
&\leq 2M_f(1+3L_\pi^2)\|\tilde s_h - s_h^\star\|^2 + 6M_f\|u_h\|^2 + 6M_fM_\pi^2\|\tilde s_h - s_h^\star\|^4\\
&\leq 2M_f(1+3L_\pi^2)C_{stab}^2\sigma_u^2 + 6M_f\sigma_u^2 + 6M_fM_\pi^2C_{stab}^4\sigma_u^4
\end{aligned}
$$

Proposition F.5 gives that $\|s_h^\Delta\| \leq C_{stab}^3C_r\sigma_u^2$, and by EISS we have that $\|\tilde s_h - s_h^\star\| \leq \sum_{k=1}^{h-1}C\rho^{h-1-k}\|u_k\| \leq C_{stab}\sigma_u$ and $\|u_h\| \leq \sigma_u$.

$$
\begin{aligned}
\|(f-f^\star)(\tilde s_h,\tilde a_h)\| &\geq \left\|\begin{bmatrix}\nabla_s(f-f^\star)(s_h^\star,a_h^\star)\\ \nabla_a(f-f^\star)(s_h^\star,a_h^\star)\end{bmatrix}^\top\begin{bmatrix}\tilde s_h - s_h^\star\\ \tilde a_h - a_h^\star\end{bmatrix} + (f-f^\star)(s_h^\star,a_h^\star)\right\| - \|r_h^X\|\\
&\geq \|D_h^\star(f)^\top(\bar s_h - s_h^\star) + \nabla_a(f-f^\star)(s_h^\star,a_h^\star)^\top u_h + (f-f^\star)(s_h^\star,a_h^\star)\|\\
&\quad - \|D_h^\star(f)^\top s_h^\Delta\| - \|r_h^X\| - \|\nabla_a(f-f^\star)(s_h^\star,a_h^\star)^\top r_h^U\|\\
&\geq \|D_h^\star(f)^\top(\bar s_h - s_h^\star) + \nabla_a(f-f^\star)(s_h^\star,a_h^\star)^\top u_h + (f-f^\star)(s_h^\star,a_h^\star)\|\\
&\quad - \sigma_u^2\left(2L_f(1+L_\pi)C_{stab}^3C_r + 2M_f(1+3L_\pi^2)C_{stab}^2 + 6M_f + 6M_fM_\pi^2C_{stab}^4\sigma_u^2 + 2L_fM_\pi C_{stab}^2\right)
\end{aligned}
$$

Rearranging, we can isolate the term containing $D_h^\star(f)$.

$$
\begin{aligned}
C_{rem,2} &\triangleq 2L_f(1+L_\pi)C_{stab}^3C_r + 2M_f(1+3L_\pi^2)C_{stab}^2 + 6M_f + 2L_fM_\pi C_{stab}^2\\
C_{rem,4} &\triangleq 6M_fM_\pi^2C_{stab}^4\\
\|D_h^\star(f)^\top(\bar s_h - s_h^\star) + \nabla_a(f-f^\star)(s_h^\star,a_h^\star)^\top u_h + (f-f^\star)(s_h^\star,a_h^\star)\|&\\
\leq \|(f-f^\star)(\tilde s_h,\tilde a_h)\| + C_{rem,2}\sigma_u^2 + C_{rem,4}\sigma_u^4&
\end{aligned}
$$

Squaring both sides and taking the expectation over $\tilde\nu_h$ given $s_1 \sim d_1$, let us expand out the isolated term. The crossterms here are between $(\bar s - s^\star)^\top(f-f^\star)(s^\star,a^\star)$, $u_h^\top(f-f^\star)(s^\star,a^\star)$, and $(\bar s - s^\star)^\top u_h$. Taking $\bar s_h - s_h^\star = \sum_{k=1}^{h-1}A_{k+1:h}^{\mathrm{cl}}B_ku_k$, zero-mean IID noise injection renders all of these crossterms 0. I define $W_{1:h}^z \triangleq \sum_{k=1}^{h-1}A_{k+1:h}^{cl}B_k\Sigma_zB_k^\top A_{k+1:h}^{cl\top}$, for $\Sigma_z = \frac{1}{2+d_u}I_{d_u}$, for $B_k \triangleq$

$\nabla_a f^\star(s_k^\star, a_k^\star)^\top$.

$$\begin{aligned}
&\mathop{\mathbb{E}}_{\tilde{\nu}_h|s_1} \|D_h^\star(f)^\top(\bar{s}_h - s_h^\star) + \nabla_a(f-f^\star)(s_h^\star,a_h^\star)^\top u_h + (f-f^\star)(s_h^\star,a_h^\star)\|^2 \\
&\quad = \operatorname{trace}\left(D_h^\star(f)^\top \mathop{\mathbb{E}}_{\{u_k\}_{k=1}^{h-1}|s_1}\left[(\bar{s}_h - s_h^\star)^{\otimes 2}\right] D_h^\star(f)\right) \\
&\qquad + \|(f-f^\star)(s_h^\star,a_h^\star)\|^2 + \mathop{\mathbb{E}}_{u_h|s_1}\|\nabla_a(f-f^\star)(s_h^\star,a_h^\star)^\top u_h\|^2 \\
&\quad \geq \sigma_u^2 \operatorname{trace}\left(D_h^\star(f)^\top W_{1:h}^z D_h^\star(f)\right)
\end{aligned}$$

The last step occurred by dropping the non-negative squared terms and noting that $\mathop{\mathbb{E}}_{\{u_k\}_{k=1}^{h-1}|s_1}(\bar{s}_h - s_h^\star)^{\otimes 2} = \sigma_u^2 W_{1:h}^z$ from the cross terms cancelling again. Since $\Sigma_z = \frac{1}{d_u+2} I_{d_u}$, we can further substitute $W_{1:h}^z = \frac{1}{d_u+2} W_{1:h}^u$, and by Assumption 4.9, for any vector $v \in \mathcal{R}_h^{\mathcal{F}}$,

$$\mathop{\mathbb{E}}_{\{u_k\}_{k=1}^{h-1}|s_1} v^\top(\bar{s}_h - s_h^\star)^{\otimes 2} v = \sigma_u^2 v^\top W_{1:h}^z v \geq \frac{c_\ell}{d_u+2}\sigma_u^2 v^\top W_{1:h}^x v \geq \frac{c_\ell \sigma_u^2}{d_u+2}\|v\|^2$$

since $W_{1:h}^x \succeq I$. Giving $D_h^\star(f)$ as v columnwise, this maps

$$\mathop{\mathbb{E}}_{\tilde{\nu}_h|s_1} \|D_h^\star(f)^\top(\bar{s}_h - s_h^\star) + \nabla_a(f-f^\star)(s_h^\star,a_h^\star)^\top u_h + (f-f^\star)(s_h^\star,a_h^\star)\|^2 \geq \frac{c_\ell\sigma_u^2}{d_u+2}\|D_h^\star(f)\|_F^2$$

where we may lowerbound the Frobenius norm once again to obtain the operator norm. Assembling with the righthandside gives the final result.

$$\frac{c_\ell\sigma_u^2}{d_u+2}\|D_h^\star(f)\|^2 \leq 2\mathop{\mathbb{E}}_{\tilde{\nu}_h|s_1}\|(f-f^\star)(\tilde{s}_h,\tilde{a}_h)\|^2 + 2(C_{rem,2} + C_{rem,4}\sigma_u^2)^2\sigma_u^4$$

□

Lemma F.7 (Lemma F.3 with bounded Jacobian difference)**.** *Let $\pi^\star, f^\star$ be a (C,ρ) EISS pair satisfying Assumption 4.9. For terms defined in Lemma F.2, with $\sigma_u \lesssim \min\left\{1, \frac{C_3\sqrt{c_\ell}}{\sqrt{2d_u+4}(C_{rem,2}+C_{rem,4})}, \frac{\sqrt{1+4L_\pi^2}}{2C_{\mathrm{stab}}M_\pi}\right\}$, for $\ell+1 \leq h \leq H-1$, it holds that:*

$$\begin{aligned}
&\mathop{\mathbb{E}}_{d_1}\left[\min\{1, \|s_{h+1}^{\pi^\star,f^\star} - s_{h+1}^{\pi^\star,f}\|\}\right] \\
&\qquad \lesssim \left(\bar{C}\bar{\rho}^{h-\ell} + \frac{1}{C_2}\right)\mathop{\mathbb{E}}_{d_1}[B_\ell(f,s_1)] \\
&\qquad\quad + \left(C_{\mathrm{stab}}^- + \frac{h-\ell}{C_2(1-\rho)}\right)\max_{\ell+1\leq k\leq h} \mathrm{J}_{\mathrm{model}}(f;\mathbb{P}_k^\star) \\
&\qquad\quad + \frac{h-\ell}{C_3\sigma_u}\sqrt{\frac{2d_u+4}{c_\ell}}\max_{\ell+1\leq k\leq h}\mathrm{J}_{\mathrm{model}}(f;\tilde{\nu}_k^{\sigma_u})
\end{aligned}$$

Proof. For ease of notation, let us denote $s_h^{\pi^\star,f}, a_h^{\pi^\star,f} \triangleq s_h, a_h$ and $s_h^{\pi^\star,f^\star}, a_h^{\pi^\star,f^\star} \triangleq s_h^\star, a_h^\star$. Let us define the event $\mathcal{E}_f$ under which the conditions of Lemma F.2 hold.

$$\mathcal{E}_f \triangleq \left\{B_\ell(f,s_1) + \frac{\max_{\ell+1\leq k\leq h}\|(f^\star - f)(s_k^\star,a_k^\star)\|}{1-\rho} \leq C_2,\ \max_{\ell+1\leq k\leq h}\|D_k^\star(f)\| \leq C_3\right\}$$

From Proposition F.6 and constraining $\sigma_u \leq 1$, we can observe the pointwise bound

$$\|D_h^\star(f)\| \leq \sqrt{\frac{2d_u+4}{c_\ell}}\left(\frac{1}{\sigma_u}\sqrt{\mathop{\mathbb{E}}_{\tilde{\nu}_h|s_1}\|(f-f^\star)(\tilde{s}_h,\tilde{a}_h)\|^2} + (C_{rem,2}+C_{rem,4})\sigma_u\right).$$

If we add the constraint $\sigma_u \lesssim \min\left\{1, \frac{C_3\sqrt{c_\ell}}{\sqrt{2d_u+4}(C_{rem,2}+C_{rem,4})}\right\}$, we can see that the additive term would absorb into C_3 when conducting Markov's inequality. Now we can use this to break up the

trajectory difference. The following proceeds through applications of union bound and Markov's inequality.

$$
\begin{aligned}
&\mathop{\mathbb{E}}_{d_1}\left[\min\{1, \|s^\star_{h+1} - s_{h+1}\|\}\right] \\
&\quad = \mathop{\mathbb{E}}_{d_1}\left[\min\{1, \|s^\star_{h+1} - s_{h+1}\|\}\mathbb{1}[\mathcal{E}_f]\right] + \mathop{\mathbb{E}}_{d_1}\left[\min\{1, \|s^\star_{h+1} - s_{h+1}\|\}\mathbb{1}[\mathcal{E}_f^C]\right] \\
&\quad \le \bar{C}\bar{\rho}^{h-\ell}\mathop{\mathbb{E}}_{d_1}[B_\ell(f, s_1)] + \sum_{k=\ell+1}^{h} \bar{C}\bar{\rho}^{h-k}\mathop{\mathbb{E}}_{d_1}\left[\|(f^\star - f)(s^\star_k, a^\star_k)\|\right] \\
&\qquad + \mathbb{P}_{d_1}\left[B_\ell(f, s_1) + \frac{\max_{\ell+1\le k\le h}\|(f^\star - f)(s^\star_k, a^\star_k)\|}{1-\rho} \ge C_2 \text{ or } \max_{\ell+1\le k\le h}\|D^\star_k(f)\| \ge C_3\right] \\
&\quad \lesssim \left(\bar{C}\bar{\rho}^{h-\ell} + \frac{1}{C_2}\right)\mathop{\mathbb{E}}_{d_1}[B_\ell(f, s_1)] + \sum_{k=\ell+1}^{h}\left(\bar{C}\bar{\rho}^{h-k} + \frac{1}{C_2(1-\rho)}\right)\mathop{\mathbb{E}}_{d_1}\left[\|(f^\star - f)(s^\star_k, a^\star_k)\|\right] \\
&\qquad + \sum_{k=\ell+1}^{h}\left(\frac{\sqrt{2d_u+4}}{C_3\sigma_u\sqrt{c_\ell}}\,\mathrm{J}_{\mathrm{model}}(f; \tilde{\nu}_k)\right)
\end{aligned}
$$

□

Lemma F.8 (Proposition F.4 with bounded Jacobian difference)**.** *Let $\pi^\star, f^\star$ be a (C, ρ) EISS pair satisfying Assumption 4.9, and let us denote $s^\star_h \triangleq s_h^{\pi^\star, f^\star, s_1}$ and $a^\star_h \triangleq \pi^\star(s^\star_h)$ for some initial state $s_1 \sim d_1$. For σ_u satisfying the conditions of Lemma F.7, candidate model f satisfies:*

$$
\begin{aligned}
J_H(\pi^\star) - Q_H^{\pi^\star, f} \lesssim H\ell^2 \max\left\{1, (L_f(1+L_\pi))^\ell\right\} &\max_{1\le h\le \ell}\mathrm{J}_{\mathrm{model}}(f; \mathbb{P}^\star_h) \\
&+ (H-\ell)^2\left(\max_{\ell+1\le h\le H}\mathrm{J}_{\mathrm{model}}(f; \mathbb{P}^\star_h) + \frac{1}{\sigma_u}\sqrt{\frac{d_u}{c_\ell}}\max_{\ell+1\le h\le H}\mathrm{J}_{\mathrm{model}}(f; \tilde{\nu}_h^{\sigma_u})\right)
\end{aligned}
$$

Proof. Let us denote $s_h \triangleq s_h^{\pi^\star, f, s_1}$ with $a_h \triangleq \pi^\star(s_h)$ for fixed candidate f. We note that since $\mathrm{R}(\cdot,\cdot) \in [0,1]$ the difference must be similarly constrained.

$$
\begin{aligned}
J(\pi^\star) - Q^{\pi^\star, f} &= \sum_{h=1}^{H}\mathop{\mathbb{E}}_{d_1}\left[\mathrm{R}(s_h^{\pi^\star, f^\star}, a_h^{\pi^\star, f^\star}) - \mathrm{R}(s_h^{\pi^\star, f}, a_h^{\pi^\star, f})\right] \\
&= \sum_{h=1}^{H}\mathop{\mathbb{E}}_{d_1}\left[\min\left\{1, \mathrm{R}(s_h^{\pi^\star, f^\star}, a_h^{\pi^\star, f^\star}) - \mathrm{R}(s_h^{\pi^\star, f}, a_h^{\pi^\star, f})\right\}\right] \\
&\le \sum_{h=1}^{H}\mathop{\mathbb{E}}_{d_1}\left[\min\left\{1, |\mathrm{R}(s_h^{\pi^\star, f^\star}, a_h^{\pi^\star, f^\star}) - \mathrm{R}(s_h^{\pi^\star, f}, a_h^{\pi^\star, f})|\right\}\right] \\
&\le \max\{1, L_{\mathrm{R}}(1+L_\pi)\}\sum_{h=1}^{H}\mathop{\mathbb{E}}_{d_1}\left[\min\left\{1, \|s_h^{\pi^\star, f^\star} - s_h^{\pi^\star, f}\|\right\}\right] \\
&\le (H-\ell)\max\{1, L_{\mathrm{R}}(1+L_\pi)\}\max_{\ell+1\le h\le H}\mathop{\mathbb{E}}_{d_1}\left[\min\left\{1, \|s_h^{\pi^\star, f^\star} - s_h^{\pi^\star, f}\|\right\}\right] \\
&\qquad + l\max\{1, L_{\mathrm{R}}(1+L_\pi)\}\mathop{\mathbb{E}}_{d_1}[B_\ell(f, s_1)]
\end{aligned}
$$

For brevity, let us define $L \triangleq \max\{1, L_{\mathrm{R}}(1+L_\pi)\}$. For bounding the burn-in term, we can expand the following: this is obviously exponential in ℓ and would compound if $L_f(1+L_\pi) > 1$; however,

since it's this is a worst-case bound and it's only up to a finite burn-in we deem this acceptable.

$$\begin{aligned}
\mathop{\mathbb{E}}_{d_1}[B_\ell(f, s_1)] &= \mathop{\mathbb{E}}_{d_1}\left[\max_{1\le k\le \ell+1} \|s_k^{\pi^\star, f} - s_k^{\pi^\star, f^\star}\|\right] \\
&\le \mathop{\mathbb{E}}_{d_1}\left[\sum_{k=1}^{\ell+1} \|s_k^{\pi^\star, f} - s_k^{\pi^\star, f^\star}\|\right] \\
&\le \mathop{\mathbb{E}}_{d_1}\left[\sum_{k=2}^{\ell+1}\sum_{j=1}^{k-1} \left(L_f(1+L_\pi)\right)^{k-1-j} \|(f-f^\star)(s_j^{\pi^\star, f^\star}, a_j^{\pi^\star, f^\star})\|\right] \\
&\le \ell\left(\sum_{k=0}^{\ell-1}\left(L_f(1+L_\pi)\right)^k\right) \max_{1\le k\le \ell} \mathop{\mathbb{E}}_{\mathbb{P}_k^\star} \|(f-f^\star)(s_k, a_k)\|
\end{aligned}$$

More important is the fact that this mandates that we control the zeroth order dynamics deviation even for $k \le \ell$. We'll define $L_{\mathrm{cl}} \triangleq L_f(1+L_\pi)$ once again for brevity. Application of Lemma F.7 bounds the state difference, but we can now insert our bound on $\mathop{\mathbb{E}}_{d_1}[B_\ell(f, s_1)]$ to expand it further.

$$\begin{aligned}
&\mathop{\mathbb{E}}_{d_1}\left[\min\{1, \|s_h^{\pi^\star, f^\star} - s_h^{\pi^\star, f}\|\}\right] \\
&\lesssim \left(\bar{C}\bar{\rho}^{h-1-\ell} + \frac{1}{C_2}\right) \mathop{\mathbb{E}}_{d_1}[B_\ell(f, s_1)] \\
&\quad + \left(C_{\mathrm{stab}}^- + \frac{h-1-\ell}{C_2(1-\rho)}\right) \max_{\ell+1\le k\le h-1} \mathrm{J}_{\mathrm{model}}(f; \mathbb{P}_k^\star) \\
&\quad + \frac{h-1-\ell}{C_3\sigma_u}\sqrt{\frac{2d_u+4}{c_\ell}} \max_{\ell+1\le k\le h-1} \mathrm{J}_{\mathrm{model}}(f; \tilde{\nu}_k^{\sigma_u}) \\
&\le \ell\left(\bar{C}\bar{\rho}^{h-\ell-1} + \frac{1}{C_2}\right)\left(\sum_{k=0}^{\ell-1} L_{\mathrm{cl}}^k\right) \max_{1\le k\le \ell} \mathop{\mathbb{E}}_{\mathbb{P}_k^\star} \|(f-f^\star)(s_k, a_k)\| \\
&\quad + \left(\frac{2\bar{C} + (h-\ell-1)/C_2}{1-\rho}\right) \max_{\ell+1\le k\le h-1} \mathrm{J}_{\mathrm{model}}(f; \mathbb{P}_k^\star) \\
&\quad + \frac{h-\ell-1}{C_3\sigma_u}\sqrt{\frac{2d_u+4}{c_\ell}} \max_{\ell+1\le k\le h-1} \mathrm{J}_{\mathrm{model}}(f; \tilde{\nu}_k^{\sigma_u})
\end{aligned}$$

For every $\ell+1 \le h \le H$, $\bar{\rho}^{h-\ell-1} \le 1$ and $h-\ell \le H-\ell$, so we can take a $\max$ over post-burn-in timesteps to get

$$\begin{aligned}
&\max_{\ell+1\le h\le H} \mathop{\mathbb{E}}_{d_1}\left[\min\left\{1, \|s_h^{\pi^\star, f^\star} - s_h^{\pi^\star, f}\|\right\}\right] \\
&\qquad \lesssim \ell\left(\bar{C} + \frac{1}{C_2}\right)\left(\sum_{k=0}^{\ell-1} L_{\mathrm{cl}}^k\right) \max_{1\le k\le \ell} \mathop{\mathbb{E}}_{\mathbb{P}_k^\star} \|(f-f^\star)(s_k, a_k)\| \\
&\qquad\quad + \left(\frac{2\bar{C} + (H-\ell)/C_2}{1-\rho}\right) \max_{\ell+1\le k\le H} \mathrm{J}_{\mathrm{model}}(f; \mathbb{P}_k^\star) \\
&\qquad\quad + \frac{H-\ell}{C_3\sigma_u}\sqrt{\frac{2d_u+4}{c_\ell}} \max_{\ell+1\le k\le H} \mathrm{J}_{\mathrm{model}}(f; \tilde{\nu}_k^{\sigma_u})
\end{aligned}$$

and putting this together we see

$$\begin{aligned}J_H(\pi^\star) - Q_H^{\pi^\star,f} \lesssim (H-\ell)L\Bigg(&\ell\left(\bar{C}+\frac{1}{C_2}\right)\left(\sum_{k=0}^{\ell-1} L_{\mathrm{cl}}^k\right)\max_{1\le k\le \ell}\mathbb{E}_{\mathbb{P}_k^\star}\|(f-f^\star)(s_k,a_k)\| \\ &+\left(\frac{2\bar{C}+(H-\ell)/C_2}{1-\rho}\right)\max_{\ell+1\le k\le H}\mathrm{J}_{\mathrm{model}}(f;\mathbb{P}_k^\star) \\ &+\frac{H-\ell}{C_3\sigma_u}\sqrt{\frac{2d_u+4}{c_\ell}}\max_{\ell+1\le k\le H}\mathrm{J}_{\mathrm{model}}(f;\tilde{\nu}_k^{\sigma_u})\Bigg) \\ &+\ell^2 L\left(\sum_{k=0}^{\ell-1} L_{\mathrm{cl}}^k\right)\max_{1\le k\le \ell}\mathbb{E}_{\mathbb{P}_k^\star}\|(f-f^\star)(s_k,a_k)\|\end{aligned}$$

We can combine the pre-burn-in model difference coefficients

$$\begin{aligned}(H-\ell)L\ell\left(\bar{C}+\frac{1}{C_2}\right)+\ell^2 L &= \ell L\left(\ell+(H-\ell)\left(\bar{C}+\frac{1}{C_2}\right)\right) \\ &\le \ell L H\max\left\{1,\bar{C}+\frac{1}{C_2}\right\}\lesssim H\ell\end{aligned}$$

and we can simplify the post-burn-in model difference coefficient

$$(H-\ell)L\left(\frac{2\bar{C}+(H-\ell)/C_2}{1-\rho}\right)\le\frac{2\bar{C}+1/C_2}{1-\rho}(H-\ell)^2 L\lesssim(H-\ell)^2.$$

Putting this all together gives us

$$\begin{aligned}J_H(\pi^\star) - Q_H^{\pi^\star,f} \lesssim{}& H\ell\left(\sum_{h=0}^{\ell-1} L_{\mathrm{cl}}^h\right)\max_{1\le h\le\ell}\mathrm{J}_{\mathrm{model}}(f;\mathbb{P}_h^\star) \\ &+(H-\ell)^2\left(\max_{\ell+1\le h\le H}\mathrm{J}_{\mathrm{model}}(f;\mathbb{P}_h^\star)+\frac{1}{\sigma_u}\sqrt{\frac{d_u}{c_\ell}}\max_{\ell+1\le h\le H}\mathrm{J}_{\mathrm{model}}(f;\tilde{\nu}_h^{\sigma_u})\right)\end{aligned}$$

□

Theorem F.9. *Let us take a system satisfying Assumption 4.9, and let us take $\tilde{\nu}_h$ from Definition 4.10 for $0<\sigma_u$ satisfying Lemma F.7. In this case, it holds that $\hat{\pi}=\arg\max_{\pi\in\Pi}\inf_{f\in\mathcal{F}(\epsilon)}Q_H^{\pi,f}$ satisfies*

$$\begin{aligned}J_H(\pi^\star)-J_H(\hat{\pi})\lesssim{}&\sup_{f\in\mathcal{F}(\epsilon)}H\ell^2\max\left\{1,(L_f(1+L_\pi))^\ell\right\}\max_{1\le h\le\ell}\left(\mathrm{J}_{\mathrm{model}}(f;\mu_h)+W_{\mathbb{P}}^\star\right) \\ &+(H-\ell)^2\max_{\ell+1\le h\le H}\left(\mathrm{J}_{\mathrm{model}}(f;\mu_h)+W_{\mathbb{P}}^\star\right) \\ &+(H-\ell)^2\frac{1}{\sigma_u}\sqrt{\frac{d_u}{c_\ell}}\max_{\ell+1\le h\le H}\inf_{\nu_h}\Bigg\{\mathcal{C}_h(\nu_h)^{1/2}\,\mathrm{J}_{\mathrm{model}}(f;\mu_h) \\ &\quad+\sqrt{2L_f W_2(\tilde{\nu}_h,\nu_h)}\mathcal{C}_h(\nu_h)^{1/4}\,\mathrm{J}_{\mathrm{model}}(f;\mu_h)^{1/2} \\ &\quad+\inf_{\gamma\in\Gamma(\tilde{\nu}_h,\nu_h)}\sqrt{2L_f W_2(\tilde{\nu}_h,\nu_h)}\,\mathrm{J}_{\mathrm{model}}\left(f;(\tilde{\nu}_h\setminus_\gamma\nu_h)\right)^{1/2}\Bigg\}\end{aligned}$$

for $\Gamma(\mathbb{P}_h^\star,\nu_h)$ and $(\mathbb{P}_h^\star\setminus_\gamma\nu_h)$ defined in Definition C.15.

Proof.

$$\begin{aligned}
J_H(\pi^\star) - J_H(\hat{\pi}) &= J_H(\pi^\star) - Q_H^{\hat{\pi}, f^\star} \\
&\leq J_H(\pi^\star) - \inf_{f \in \mathcal{F}(\epsilon)} Q_H^{\hat{\pi}, f} \quad \text{by } \mathcal{F}(\epsilon) \text{ realizability} \\
&= J_H(\pi^\star) - \sup_{\pi \in \Pi} \inf_{f \in \mathcal{F}(\epsilon)} Q_H^{\pi, f} \quad \text{by pessimistic optimality} \\
&\leq J_H(\pi^\star) - \inf_{f \in \mathcal{F}(\epsilon)} Q_H^{\pi^\star, f} \quad \text{by } \Pi \text{ realizability} \\
&= \sup_{f \in \mathcal{F}(\epsilon)} \left(J_H(\pi^\star) - Q_H^{\pi^\star, f} \right)
\end{aligned}$$

We can apply Lemma F.8:

$$\begin{aligned}
J_H(\pi^\star) - J_H(\hat{\pi}) \lesssim \sup_{f \in \mathcal{F}(\epsilon)} & H\ell^2 \max\left\{1, \left(L_f(1+L_\pi)\right)^\ell\right\} \max_{1 \leq h \leq \ell} \mathrm{J}_{\mathrm{model}}(f; \mathbb{P}_h^\star) \\
& + (H-\ell)^2 \left(\max_{\ell+1 \leq h \leq H} \mathrm{J}_{\mathrm{model}}(f; \mathbb{P}_h^\star) + \frac{1}{\sigma_u} \sqrt{\frac{d_u}{c_\ell}} \max_{\ell+1 \leq h \leq H} \mathrm{J}_{\mathrm{model}}(f; \tilde{\nu}_h^{\sigma_u}) \right)
\end{aligned}$$

We may now apply Lemma C.17 to transfer the noise injection distribution to the dataset distribution for the final result. □

G Experimental Validation

We illustrate our results with experiments on a set of task settings, where we may demonstrate the effect of our practices on a variety of model-based and model-free algorithms. We first consider the `robomimic` Lift task (Mandlekar et al., 2021), where we demonstrate the effects of action chunking on performance in a setting with open-loop stability induced by low-level end-effector control. Next, we analyze the MuJoCo Reacher-v5 task (Todorov et al., 2012), where we demonstrate the effects of exploratory data collection. Finally, the MuJoCo HalfCheetah-v5 task provides an alternate setting for exploratory data collection where we would inherently expect the expert not to demonstrate closed-loop stable behavior as defined in Definition B.1 due to the periodic nature of legged locomotion.

For our algorithms, we select TD3+BC (Fujimoto & Gu, 2021), IQL (Kostrikov et al., 2021), MOPO (Yu et al., 2020), and MOBILE (Sun et al., 2023). TD3+BC and IQL serve as our model-free baselines, and MOPO and MOBILE serve as hybrid model-based baselines that use both model-based and model-free behavior in order to learn good policies. We additionally run a pure model-based version of MOBILE (with real ratio of 0.0, telling the algorithm to only use the learned models for policy training) in order to provide a more pure rendition of our setting. We will refer to this version of MOBILE as MB-MOBILE, with the standard algorithm as Hybrid-MOBILE.

We find that all of these tasks exhibit the expected behavior and demonstrate the efficacy of our practices, and additionally point to these practices having wider applicability than the pure model-based domain we study.

We note that we did not conduct extensive hyperparameter optimization for the provided experiments as our goal is not to prescribe a particular algorithm, but is instead to demonstrate the general trends observed when performing our remedies in certain settings. To this end, we focus on the behavior of each individual line as we demonstrate usage of each relevant practice, as opposed to the comparative behavior between the algorithms.

Action Chunking. To validate our predictions about the benefits of action chunking in open-loop stable environments, we train Life on the `mg-dense` dataset and HalfCheetah-v5 on the `expert-v0` dataset for five chunk lengths ℓ. To this end, we learn policies which predict sequences of ℓ actions $\{a_k\}_{k=h}^{h+\ell-1}$ based on the current observation s_h, which we execute in open-loop during evaluation. Taking motivation from the insights studied by Park et al. (2026), we train chunked dynamics functions for our model-based algorithms, which take in the current observation and an action chunk to predict the cumulative reward and the state ℓ steps ahead $f(s_h, \{a_k\}_{k=h}^{h+\ell-1})$ instead

of recursively rolling out a one-step dynamics function ℓ times within each chunk. These learned dynamics generate synthetic transitions which are used for policy training, and evaluation executes the final trained actor without using the learned dynamics. For each algorithm, hidden-layer widths, network depths, ensemble sizes, and algorithm parameterizations were held constant across chunk lengths; however, actor output dimensions, critic and dynamics input dimensions, and training targets necessarily changed to accommodate the differing chunk lengths.

In the left panel of Figure 1, we find that **action chunks result in non-monotonic performance gains**, with all configurations exhibiting increasing performance from the $\ell = 1$ regime to $\ell = 2$ or $\ell = 4$, but uniformly decreasing performance as the chunk length continues to grow. This validates our theoretical results: for chunk lengths *up to a certain length* dependent on system theoretic constants, we predict that ℓ can unlock the stable behavior used in Section 4.2 to prove our positive result. After that point, further chunking serves to weaken performance as useful signal is ignored. We show that while our results are proven for a specific pure model-based setting, this behavior is present, and potentially even more explicit, with the inclusion of model-free training in Figure 4. While the pure model-free baseline fails to learn a good policy across chunk levels, the hybrid baselines exhibit marked increases in task success rate for chunk lengths $\ell = 2$ and $\ell = 4$.

In the right panel, we find that **the benefits of action chunking don't translate in open-loop unstable systems**. Despite good performance in the $\ell = 1$ regime for the model-free methods, this behavior quickly degrades as the chunk length increases. This is expected, as without an open-loop system our theoretical guarantees don't apply, supporting the stability-theoretic reasoning for their benefits.

Exploratory Data Collection. To validate our predictions about the benefits of exploratory data collection, we train on the MuJoCo Reacher-v5 and Halfcheetah-v5 tasks for five dataset composition levels. In these configurations, we construct a dataset of $1,000,000$ samples collected from a varying fraction of deterministic expert rollouts and noise-injected expert rollouts. For each algorithm, this dataset was the only change. The expert policies used for both tasks were Soft Actor-Critic (Haarnoja et al., 2018) RL policies pre-trained using the StableBaselines3 library (Raffin et al., 2021) by Younis et al. (2024), downloaded from Huggingface [Reacher-v5] [HalfCheetah-v5]. When collecting expert demonstrations, we set `deterministic=True`. During noise-injected expert rollouts, given noise scale σ_u, we sampled $z_h \sim \mathcal{N}(0, I_{d_u})$ every timestep and executed action $\text{clip}\left(\pi^\star(s_h) + \sigma z_h, u_{\min}, u_{\max}\right)$ given state s_h, where $\sigma = \sigma_u / d_u$ in order to compare between differing action dimensions fairly. This dimensionality was $d_u = 2$ for Reacher-v5 and $d_u = 6$ for HalfCheetah-v5, and the action bounds were $[-1, 1]$ for both. The noise-injected action was recorded as the action at timestep h, differing from the convention used in Zhang et al. (2025). We note that MuJoCo tasks, by default, include initial state randomization and deterministic dynamics. This ensures that our clean trajectories exhibit some excitation and are not redundant, while keeping to our overarching setting where the dynamics don't give excitation for free without exploratory policies.

In the Reacher-v5 experiments Figure 3, in the left panel, we find that a distinct improvement can be seen for all algorithms as noise-injected trajectories are mixed into the dataset, with the most notable change being from 100% clean to 75% clean-25% noise injected. This matches our theoretical expectations, as a non-trivial amount of noise injected data implies that the first-order errors can now be controlled; however, past 50% noise injected the performance increase is marginal. We observe similar behavior for HalfCheetah-v5 in the left panel of Figure 5, where performance increases with the fraction of noise injected trajectories, though it saturates after 50% clean and 50% noise injected, the same as Reacher-v5. On the right panel, we observe the effects of varying σ_u on Reacher-v5. Overall, we see that increasing the noise scale improves performance, which matches the inverse relationship present in Theorem 4.12. Once again, these results are observed uniformly across the model-based and model-free algorithms.

We run one more experiment on the right side of Figure 5, demonstrating the effect of suboptimal data collected through a policy that was not fully trained to achieve expert behavior. In this experiment, the expert policy was run deterministically as was done in prior experiments; however, the medium quality dataset was collected without the deterministic flag. This experiment demonstrates increased performance for all algorithms except IQL, which seems to degrade in performance for the mixed data benchmarks. This demonstrates that suboptimal data can still exhibit improved per-

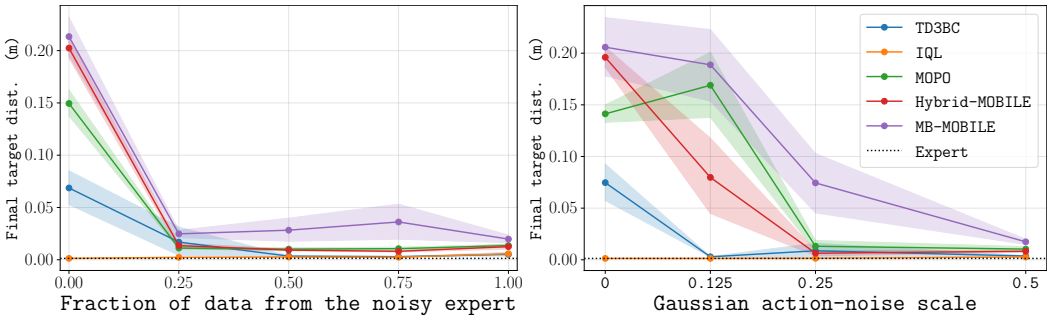


Figure 3: Reacher-v5 performance. Note that smaller final fingertip-target distance is a measure of increased performance. **Left:** (repeated from the main body) We vary the fraction of trajectories collected from a noise-injected expert as opposed to a deterministic expert for $\sigma_u = 0.5$, and find that performance under only clean trajectories is distinctly worse, indicating the necessity for excitation. **Right:** On a dataset where half the trajectories are collected under the deterministic expert policy and half are collected with noise injection, we vary the scale σ_u of noise injected, and find that when we retain clean expert coverage, increased noise injection allows for greater coverage of potential error states and increases performance.

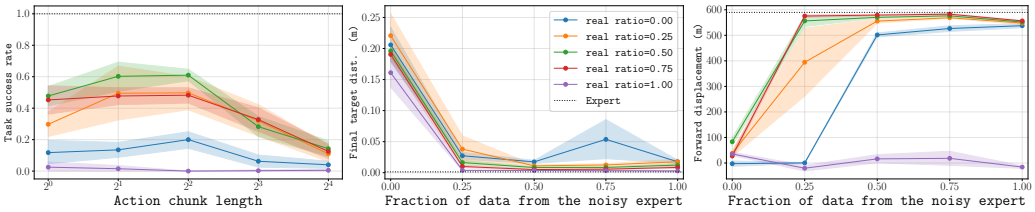


Figure 4: We plot the effect of differing levels of model usage versus dataset usage during policy optimization on the MOBILE algorithm. **Left:** Lift. **Center:** Reacher-v5. **Right:** HalfCheetah-v5. We observe that our prescriptions provide beneficial effects for all levels of model-free to model-based behavior, with the exception of the `real-ratio=1.00` baseline which failed to learn on the Lift and HalfCheetah-v5 tasks. We remark that this is expected behavior, as MOBILE is ultimately a model-based algorithm and running it in a model-free manner disables much of the intended machinery, as opposed to the other model-free baselines we present.

formance in most cases even without the exact prescription of expert noise injection, provided that the geometric coverage is sufficient.

G.1 Implementation Details

We used the implementation of TD3+BC, IQL, MOPO, and MOBILE provided by OfflineRL-Kit (Yihao Sun, 2023), and port the environment-facing code in order to accept returns from Gymnasium (Towers et al., 2025) from the legacy Gym interface (Fu et al., 2020). Unless otherwise noted, we retained the command-line hyperparameter defaults from OfflineRL-Kit's reference training scripts.

Across our model-based experiments, we use half real transitions in each policy-training minibatch for MOPO and Hybrid-MOBILE and reduce the number of initial dataset states used for each round of synthetic rollouts from 50,000 to 10,000. MOPO's uncertainty-penalty coefficient is changed from the generic OfflineRL-Kit default of 2.5 to 0.5. These settings are held fixed across the primary algorithm and dataset-composition comparisons.

The three primary experiments (action chunking on Lift, and clean to noised expert on Reacher-v5 and HalfCheetah-v5) were conducted on 8 seeds each, with all other ablations being conducted on 4. The final checkpoint was used for all experiments, from which 100 trajectories were collected for evaluation purposes. The solid line for each plot represents the mean across seeds and trajectories, with the shaded region representing the 10th-90th percentile range for 10,000 hierarchical bootstrap samples, resampling both seeds and trajectories within each.

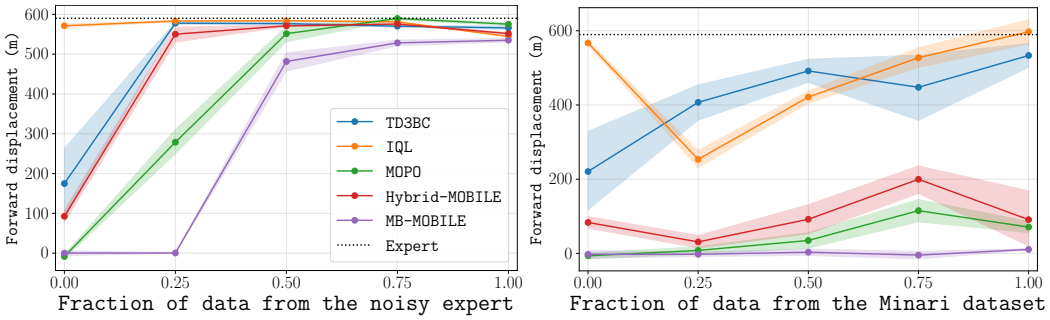


Figure 5: HalfCheetah-v5 performance. **Left:** (repeated from the main body) We vary the fraction of noise-injected trajectories for $\sigma_u = 0.5$, and find that our results once again support our theoretical conclusions on exploratory data benefits. **Right:** We vary the fraction of the dataset composed of the medium-v0 dataset from Younis et al. (2024) as opposed to the clean expert. Here, we find less monotonic behavior; however, the improvements exhibited by MOPO, Hybrid-MOBILE, and TD3+BC demonstrate that the increased geometric coverage still provides a greater impact than the decreased expert concentrability.

Action Chunking. For both the Lift and HalfCheetah-v5 task, all algorithms were trained for 300 epochs. We allowed an unbounded number of epochs for dynamics training to allow the validation-based early stopping to determine dynamics epochs, in order to accommodate the increased difficulty of predicting chunked dynamics. Taking inspiration from Park et al. (2026), we did not change the size of the dynamics networks across chunk lengths. Following Robosuite's native task semantics (Zhu et al., 2020), Lift success was not treated as terminal: post-success demonstration transitions were retained, demonstration endpoints were treated as truncations, and model rollouts continued through successful states. Finally, the α parameter for TD3+BC was changed from 2.5 to 0.05: this reflects an upweighting in the relative influence of behavior cloning as opposed to Q-value maximization in the actor objective. Larger values of α produced critic and policy instability for longer action chunks in preliminary experiments. This stronger behavior regularization is consistent with Li et al. (2025), which identifies behavior regularization as an important component of model-free action chunking.

Exploratory Data Collection. For both Reacher-v5 and HalfCheetah-v5, each algorithm was trained for 600 epochs with a policy minibatch size of 512 in order to improve performance. HalfCheetah-v5 model-based experiments used synthetic rollouts of length five, following OfflineRL-Kit's task-specific HalfCheetah recommendation, while Reacher-v5 used length one synthetic rollouts. Dynamics training was capped at 30 epochs for both tasks. We omitted OfflineRL-Kit's IQL trajectory-return-range reward normalization for locomotion tasks, retaining the original task rewards for every algorithm. MOBILE required an additional Reacher-v5 specific modification due to the negative returns produced by the distance-to-goal reward, since OfflineRL-Kit's nonnegative target clamp on MOBILE collapses negative targets to zero. We therefore translated MOBILE's return coordinate by 30, applied the clamp in the translated coordinate system, and transformed the Bellman rewards consistently. This is equivalent to imposing a target floor of -30 in the original return coordinate system rather than a floor of zero.