From POMDP Theory to Deep RL with Particle Filters

Kevin Tan
Wharton Statistics and Data Science
University of Pennsylvania
kevtan@wharton.upenn.edu

Ziping Xu
School of Data and Information Sciences
University of North Carolina at Chapel Hill
zipingxu@unc.edu

Abstract

Model-based reinforcement learning in partially observable environments requires inferring hidden states, learning latent dynamics, and planning under the resulting belief. Practical algorithms often solve one aspect of the above pipeline while relying on approximations for the others that are often poorly understood, while existing POMDP theory typically analyzes computationally inefficient theoretical algorithms far from modern deep RL practice. To fill this gap, we propose a practical end-to-end algorithm that approximately solves the hidden state filtering problem via sequential Monte Carlo (SMC), updates a latent dynamics model online, and performs approximate online planning for d steps before bootstrapping leaves with a state-value critic (a QMDP approximation). We analyze each approximation and prove an end-to-end regret bound decomposing the error into model estimation error, particle filter error, Monte Carlo rollout error, and a structural bias term induced by the QMDP approximation controllable by the planning depth that we further analyze in various special cases. The resulting guarantee makes explicit how statistical sample size, particle budget, rollout budget, and planning depth affect performance. We instantiate our algorithm in an MPC-based deep RL framework, demonstrating its practicality on a variety of environments. On a POMDP that satisfies our assumptions, each term of the bound decays at its predicted rate. Our results provide a practical and theoretically grounded route for bridging theory and practice within online decision-making in POMDPs.

1 Introduction

Partially observable Markov decision processes (POMDPs) provide a standard framework for sequential decision making under latent state uncertainty [19]. They are a natural model for problems in robotics, healthcare, recommendation, and other settings [17, 33] where the learner must act using only indirect and noisy observations of the underlying system. In principle, optimal control in a POMDP can be reduced to planning in the belief-state MDP, where the belief is the posterior distribution over latent states given the observation history [4, 19]. This makes model-based RL especially natural in POMDPs: to act under a belief, the learner must already learn a model rich enough to perform filtering, and this same model can be used for planning.

However, the belief state reduction does not remove the main computational and learning difficulties. An online model-based decision maker must simultaneously (i) learn the unknown model, (ii) maintain a posterior over latent states online, and (iii) plan over this continuous belief space. These three tasks–*learning, inference, and planning*–interact in a way that makes exact online control in POMDPs intractable in both computation and sample complexity in general [32] without special structure, motivating a large literature on *approximate* methods for each component individually.

Existing methods typically make one component of the above POMDP pipeline tractable while leaving the others either computationally expensive or theoretically uncontrolled. For instance, on the planning side, Monte Carlo tree search has led to scalable online POMDP solvers such as Partially

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

Observable Monte Carlo Planning (POMCP), sparse tree search, and progressive widening methods [34, 26, 36]. However, these are primarily planning algorithms under a known generative model, and do not natively provide end-to-end learning guarantees for unknown latent dynamics. A more aggressive approximation is QMDP, which replaces the belief-MDP continuation value by the fully observed MDP value averaged under the current belief [27]. QMDP is computationally efficient but can introduce substantial bias. Existing theory gives limited guidance on when this bias is small or how it can be mitigated. Regarding inference, particle filters, a.k.a. sequential Monte Carlo (SMC), provide a flexible approximation to Bayesian filtering in nonlinear non-Gaussian systems [8, 38], but SMC theory for POMDPs [26, 25] has not been fully integrated with approximate online planning and model learning. On learning, recent theoretical work has made significant progress on sample-efficient model estimation for POMDPs. [28, 29] develop Optimistic MLE as a generic model-based algorithm with polynomial sample complexity for a broad class of POMDPs. These results establish statistical feasibility under exact or abstract planning oracles. They do not specify how to perform approximate filtering or planning with the learned model in a computationally practical way.

A separate practical line of work uses recurrent policies or differentiable planning modules. Recurrent model-free deep RL methods, such as GRU or LSTM-based PPO and Q-learning variants, can perform well empirically under partial observability, but their hidden states are difficult to interpret as beliefs and are hard to analyze theoretically. QMDP-net takes a more structured approach by embedding approximate Bayesian filtering and QMDP planning into a differentiable architecture [22]. This is closer in spirit to our goal: combine model learning, belief state inference, and planning in a way that resembles modern deep learning practice, while maintaining provable guarantees. However, the approximation error of learned variational or differentiable filtering is difficult to control. Our view is that using SMC as the filtering primitive yields a more principled bridge: it is compatible with learned latent dynamics and has well-developed stability and approximation theory [8, 38].

In summary, existing work advances planning, inference, or learning separately, but no current method provides a unified pipeline that combines all three with both computational practicality and end-to-end theoretical guarantees. This leaves open the central question

Can we design a practical model-based deep RL pipeline for POMDPs that still admits end-to-end theoretical guarantees?

1.1 Our Contributions

This paper develops a practical and analyzable route from POMDP theory to model-based deep RL. We answer the central question by developing a model-based learning-inference-planning pipeline for POMDPs called **PF-MCTQ** that is compatible with deep RL practice. Our contributions are:

1. **A practical inference-learning-planning pipeline.** We propose an online algorithm that learns a latent dynamics model, tracks the current belief with particles, and plans under this belief. Each is modular and can be instantiated with standard deep RL primitives.
2. **A controlled alternative to full belief-tree planning.** Instead of treating the particle belief as the planning state, we plan particle-wise and average action-values of the fully observable MDP. This gives a practical middle ground between expensive belief-tree search and cheap but potentially biased QMDP. Our analysis shows that the bias decays geometrically after d steps of belief-MDP lookahead. For cheaper state-level planners, we show that QMDP is exact for the belief-linear component of the value, with error only from an uncertainty-sensitive residual, clarifying when the approximation is small.
3. **An end-to-end regret guarantee.** We prove a regret bound that decomposes performance loss into model estimation error, particle filtering error, Monte Carlo rollout error, and QMDP structural bias. The bound makes explicit how sample size, particle budget, rollout budget, and planning depth jointly control performance.
4. **A bridge to deep RL practice.** Compared with theoretical POMDP learning algorithms that rely on exact POMDP solvers or posterior sampling over parameters, our approach uses MPC-style online planning, approximate filtering, and critic bootstrapping. These are standard primitives in model-based deep RL, while remaining explicit enough to support theoretical analysis. We instantiate our algorithm in an MPC-based deep RL implementation and evaluate it on various POMDP environments, demonstrating its practicality. On a

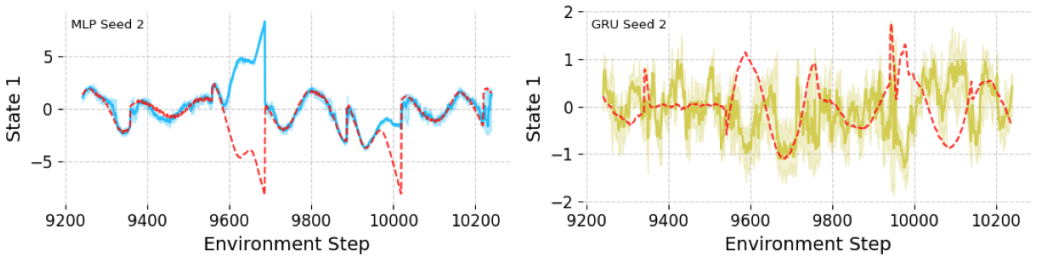


Figure 1: Particle beliefs for the unobserved velocity of a double integrator LQR over the last 1000 steps. Mean, 10th and 90th quantiles visualized. Beliefs track the state when the system is stable.

POMDP satisfying our assumptions, each term of the bound decays at its predicted rate (Section 5).

Overall, the paper argues that particle filtering can serve as a practical bridge between POMDP theory and deep RL. It gives a belief representation that is explicit enough to analyze, flexible enough for nonlinear latent dynamics, and compatible with MPC-style planning and learned critics. This yields a route to online POMDP learning that is closer to modern model-based RL than exact belief-space planning, while providing a transparent theoretical account that lets users trace each source of approximation error and make informed design choices.

Section 2 introduces the online learning problem and the regret criterion in the belief-state MDP. Section 3 presents the PF-MCTQ algorithm, describing the MCTQ planning subroutine and the full online pipeline. Section 4 develops the error analysis: we decompose the regret into four traceable terms, analyze the structural QMDP bias and its variance interpretation, bound each error source, and combine them into the end-to-end regret guarantee. We conclude with case studies on LQG, contextual MDPs, and smooth-value POMDPs that illustrate when the QMDP approximation is benign.

2 Problem Formulation

We study online reinforcement learning in a partially observable environment with unknown dynamics. Since the latent state is hidden, the natural state variable for control is the belief, the posterior distribution of the latent state given the history. We first introduce the POMDP model and its belief-state dynamics. Figure 1 shows a particle belief for a partially observed double integrator (Section 5).

Parameterized POMDP model. For each parameter $\theta \in \Theta$, let $\mathcal{M}_\theta = (\mathcal{S}, \mathcal{A}, \mathcal{O}, P_\theta, Z_\theta, r, \gamma, b_0)$ be an infinite-horizon discounted POMDP, where $\mathcal{S}$ is the latent state space, $\mathcal{A}$ is the action space, and $\mathcal{O}$ is the observation space. The controlled state transition kernel is $P_\theta(\cdot \mid s, a) \in \Delta(\mathcal{S})$, the observation kernel is $Z_\theta(\cdot \mid s') \in \Delta(\mathcal{O})$, the reward function is $r : \mathcal{S} \times \mathcal{A} \to [0, R_{\max}]$, the discount factor is $\gamma \in (0, 1)$, and $b_0 \in \Delta(\mathcal{S})$ is the initial belief. The true environment is specified by an unknown true parameter $\theta^\star$, and we write $\mathcal{M}^\star := \mathcal{M}_{\theta^\star}$. Here, for a measurable space $\mathcal{X}$, $\Delta(\mathcal{X})$ denotes the set of probability measures on $\mathcal{X}$, and beliefs $b \in \Delta(\mathcal{S})$ are written as measures $b(\mathrm{d}s)$.

Generative process and belief states. At each time $t = 0, 1, 2, \ldots$, the latent state, observation, and reward evolve according to

$$S_{t+1} \sim P_{\theta^\star}(\cdot \mid S_t, A_t), \qquad O_{t+1} \sim Z_{\theta^\star}(\cdot \mid S_{t+1}), \qquad R_t = r(S_t, A_t). \tag{1}$$

The learner does not observe S_t. Instead, decisions are based on the entire history up to time t denoted by $H_t := (O_{1:t}, A_{0:t-1}, R_{0:t-1})$. Here O_t is available before choosing A_t. A possibly randomized policy is a sequence $\pi = \{\pi_t\}_{t \geq 0}$ such that $\pi_t(\cdot \mid H_t) \in \Delta(\mathcal{A})$.

Belief states and Bayes filtering. Fix a candidate parameter θ. The corresponding belief state at time t is the posterior distribution $b_t^\theta(\cdot) := \mathbb{P}_\theta(S_t \in \cdot \mid H_t)$. When $\theta = \theta^\star$, we write $b_t := b_t^{\theta^\star}$ for the true belief. Given a belief $b \in \Delta(\mathcal{S})$ and action $a \in \mathcal{A}$, define the one-step predictive state distribution and the corresponding predictive observation law respectively by

$$\tilde{b}_\theta(\cdot \mid b, a) := \int_{\mathcal{S}} P_\theta(\cdot \mid s, a)\, b(\mathrm{d}s), \qquad p_\theta(\cdot \mid b, a) := \int_{\mathcal{S}} Z_\theta(\cdot \mid s')\, \tilde{b}_\theta(\mathrm{d}s' \mid b, a). \tag{2}$$

After observing $o \in \mathcal{O}$, the posterior is updated through the Bayes filter (which we later approximate)

$$\tau_\theta(b, a, o)(\mathrm{d}s) := \frac{Z_\theta(o \mid s)\, \tilde{b}_\theta(\mathrm{d}s \mid b, a)}{\int_{\mathcal{S}} Z_\theta(o \mid s')\, \tilde{b}_\theta(\mathrm{d}s' \mid b, a)}. \tag{3}$$

Belief-MDP value functions. The reward associated with belief b and action a is $r(b, a) := \mathbb{E}_{s\sim b}[r(s, a)]$. We assume throughout that $\mathcal{A}$ is finite and that $|r(s, a)| \le R_{\max}$ for all $(s, a) \in \mathcal{S} \times \mathcal{A}$.

For each θ, the POMDP induces a fully observed MDP on $\Delta(\mathcal{S})$. For any bounded measurable function $V : \Delta(\mathcal{S}) \to \mathbb{R}$, define the belief-MDP Bellman operator

$$(\mathcal{T}_\theta V)(b) := \max_{a\in\mathcal{A}} \left\{ r(b, a) + \gamma \int_{\mathcal{O}} V\big(\tau_\theta(b, a, o)\big)\, p_\theta(o \mid b, a)\, \mathrm{d}o \right\}. \tag{4}$$

Let $V_\theta^{\star,B}$ denote the unique fixed point of $\mathcal{T}_\theta$. The optimal belief-state action-value function is

$$Q_\theta^{\star,B}(b, a) := r(b, a) + \gamma \int_{\mathcal{O}} V_\theta^{\star,B}\big(\tau_\theta(b, a, o)\big)\, p_\theta(o \mid b, a)\, \mathrm{d}o, \text{ and } V_\theta^{\star,B}(b) = \max_{a\in\mathcal{A}} Q_\theta^{\star,B}(b, a). \tag{5}$$

More generally, for a policy π acting on histories, let $V_\theta^{\pi,B}(b)$ be its discounted value in the belief-MDP from initial belief b, and let $Q_\theta^{\star,S}$ and $V_\theta^{\star,S}$ denote the value functions of the fully observable MDP.

3 The PF-MCTQ Algorithm

We present PF-MCTQ, a model-based online learning algorithm that integrates model learning, belief inference, and planning into a modular pipeline. We divide time into doubling epochs $E_k = [2^{k-1}, 2^k)$, $k = 1, \ldots, \lceil \log_2 T \rceil$, using the standard doubling trick [3]. At the start of each epoch, the algorithm (i) updates the MLE $\bar{\theta}_k$ using all past data and freezes it for the epoch, (ii) maintains a particle approximation b_t^J of the current belief via sequential predict-reweight-resample, and (iii) plans under $\bar{\theta}_k$ and b_t^J using finite-depth Monte-Carlo belief-tree search with QMDP leaf evaluation. The doubling-epoch structure keeps the model from updating too often.

Learning oracle. Let $\mathcal{L}_t(\theta)$ be the log-likelihood of the observed history H_t under model θ. At the start of epoch E_k, ModelUpdate returns a β-approximate maximizer on all data up to T_k:

$$\bar{\theta}_k \in \big\{\theta \in \Theta : \mathcal{L}_{T_k}(\theta) \ge \max_{\theta'\in\Theta} \mathcal{L}_{T_k}(\theta') - \beta\big\}, \tag{6}$$

for a tolerance $\beta \ge 0$. The model $\bar{\theta}_k$ is frozen throughout epoch E_k. When $\beta = 0$ this is exact MLE, but we accommodate a β-suboptimal solver. In practice, many recent gradient-based methods for the MLE in partially observed Markov process (POMP) models apply here [30, 37, 5, 18, 7].

Filtering update. After playing a_t and observing o_{t+1}, particles are propagated and reweighted under the current epoch's model $\bar{\theta}_{k(t)}$:

$$\hat{s}_j^{(t+1)} \sim P_{\bar{\theta}_{k(t)}}(\cdot \mid s_j^{(t)}, a_t), \qquad w_j^{(t+1)} \propto Z_{\bar{\theta}_{k(t)}}(o_{t+1} \mid \hat{s}_j^{(t+1)}), \tag{7}$$

followed by multinomial resampling $s_j^{(t+1)} \sim \mathrm{Categorical}(w_1^{(t+1)}, \ldots, w_J^{(t+1)})$ for $j = 1, \ldots, J$. This is the bootstrap sampling-importance-resampling (SIR) filter; any SMC scheme satisfying the L^p error bound of Appendix D.3 can be substituted.

Approximate planning. For the current belief, MCTQ runs Monte Carlo tree search on the belief tree for the first d steps and applies the QMDP approximation at the leaves. This gives a middle ground between expensive belief-tree search and the cheap but potentially biased QMDP approximation.

For a belief $b \in \Delta(\mathcal{S})$, define the QMDP leaf value and the depth-d belief tree value as

$$\bar{V}_\theta^{\star,S}(b) := \max_{a\in\mathcal{A}} \int_{\mathcal{S}} Q_\theta^{\star,S}(s, a)\, b(\mathrm{d}s), \quad \text{and} \quad \bar{V}_{\theta,d}^{\star,S} := \mathcal{T}_\theta^d\, \bar{V}_\theta^{\star,S}, \qquad d \ge 0, \tag{8}$$

Algorithm 1 PF-MCTQ: Particle filtering with Monte Carlo tree QMDP planning. MCTQ (Algorithm 2) takes the particle belief b_t^J and returns an estimate $\widehat{Q}_t(b_t^J, a)$ of $Q_{\bar{\theta}_k,d}(b_t^J, a)$ for every root action $a \in \mathcal{A}$.

Require: particles J; depth d; rollouts N; horizon T; tolerance β; model class Θ; exploration parameter $\eta > 0$.

Sample $s_1^{(1)}, \dots, s_J^{(1)} \sim b_0$.
Set epochs $E_k = [2^{k-1}, 2^k)$ for $k = 1, \dots, K := \lceil \log_2 T \rceil$, with $\varepsilon_k := \eta/\sqrt{|E_k|}$.
for epoch $k = 1, 2, \dots, K$ **do**
  $\bar{\theta}_k \leftarrow \textsc{ModelUpdate}(H_{T_k})$ via equation 6; compute $\widehat{Q}^{\mathrm{MDP}}_{\bar{\theta}_k} \approx Q^{\star,S}_{\bar{\theta}_k}$.
  for $t \in E_k$ **do**
    $\{\widehat{Q}_t(b_t^J, a)\}_{a\in\mathcal{A}} \leftarrow \textsc{Mctq}\big(\bar{\theta}_k,\, b_t^J,\, d,\, N,\, \widehat{Q}^{\mathrm{MDP}}_{\bar{\theta}_k}\big)$ ▷ Algorithm 2
    Play $a_t = \arg\max_a \widehat{Q}_t(b_t^J, a)$ w.p. $1 - \varepsilon_k$; else $a_t \sim \mathrm{Unif}(\mathcal{A})$.
    Observe (o_{t+1}, r_t).
    Propagate, reweight, and resample particles via equation 7 under $\bar{\theta}_k$.
  end for
end for

where $Q_\theta^{\star,S}$ is the optimal action-value function of the fully observable MDP induced by θ, and $\mathcal{T}_\theta$ is the Bellman operator of the belief-MDP from equation 4. The corresponding action-value function is

$$Q_{\theta,d}(b,a) := r(b,a) + \gamma \int_{\mathcal{O}} \bar{V}^{\star,S}_{\theta,d-1}\big(\tau_\theta(b,a,o)\big)\, p_\theta(o \mid b, a)\, \mathrm{d}o, \qquad d \geq 1. \tag{9}$$

When $d = 0$, $\bar{V}^{\star,S}_{\theta,0}(b) = \max_a \int Q^{\star,S}_\theta(s,a)\, b(\mathrm{d}s)$ and MCTQ reduces exactly to QMDP. As $d \to \infty$, $\bar{V}^{\star,S}_{\theta,d} \to V^{\star,B}_\theta$ (see Proposition 2 in Appendix D.1 for the proof).

Particle belief and MCTS approximation. In practice, the current belief is represented by a particle approximation $b_t^J = \frac{1}{J}\sum_{j=1}^J \delta_{s_t^{(j)}}$. Starting from b_t^J, MCTQ builds a search tree of depth d with N Monte Carlo simulations per root action, inspired by the POMCP framework [34]. Each simulation samples a particle from b_t^J, descends to depth d using uniform random action selection at internal nodes, expands nodes as it passes through them, and records its state in the particle set $B(v)$ of every node v it visits. Each depth-d leaf is evaluated by the QMDP value of its pooled particles, $\max_{a'} |B(v)|^{-1} \sum_{s\in B(v)} Q^{\star,S}_\theta(s,a')$, and a bottom-up dynamic program then computes $\widehat{Q}_t(b_t^J, a)$ for each root action. We use uniform selection for the analysis and do not claim it is the best rule in practice. The analysis only needs the internal rule to ignore the current tree's statistics and to give every action probability at least $\pi_{\min} > 0$, so a frozen learned proposal also qualifies, but UCB1 does not (Appendix B).

Observation discretization. With continuous $\mathcal{O}$, every simulated observation is distinct, so tree nodes below the root are never revisited and a depth-d tree is not well defined. MCTQ therefore indexes children by a binned observation $\phi(o)$, where $\phi : \mathcal{O} \to \bar{\mathcal{O}}$ has $M := |\bar{\mathcal{O}}|$ bins of diameter $\mathrm{diam}(\phi)$. For finite $\mathcal{O}$ we take $\phi = \mathrm{id}$ and $M = |\mathcal{O}|$, and for $d = 0$ no binning is needed. The resulting bias and its trade-off against N are made explicit in Theorem 1.

4 Error Analysis and Regret Guarantee

We measure performance by the expected cumulative regret defined as:

$$\mathrm{Reg}_T := \mathbb{E}\left[\sum_{t=1}^T \mathrm{regret}_t\right] := \mathbb{E}\left[\sum_{t=1}^T \big(V^{\star,B}_{\theta^\star}(b_t) - V^{\pi_t,B}_{\theta^\star}(b_t)\big)\right], \tag{10}$$

where π_t is the policy deployed from round t, and the expectation is over the trajectory and algorithmic randomness. A key contribution is an end-to-end regret bound for PF-MCTQ, enabled by the following novel regret decomposition. Let $k(t)$ be the epoch containing round t, with frozen model $\bar{\theta}_{k(t)}$:

$$\mathrm{regret}_t = \underbrace{V^{\star,B}_{\theta^\star}(b_t) - V^{\star,B}_{\bar{\theta}_{k(t)}}(b_t)}_{\text{(A) optimal-value model error}} + \underbrace{V^{\star,B}_{\bar{\theta}_{k(t)}}(b_t) - V^{\pi_t,B}_{\bar{\theta}_{k(t)}}(b_t)}_{\text{(B) planning + state-id error}} + \underbrace{V^{\pi_t,B}_{\bar{\theta}_{k(t)}}(b_t) - V^{\pi_t,B}_{\theta^\star}(b_t)}_{\text{(C) fixed-policy model error}}, \tag{11}$$

Terms (A) and (C) capture model estimation error: the gap between $\bar{\theta}_{k(t)}$ and $\theta^\star$. Term (B) captures planning approximation under the learned model, which is further decomposed into three sources:

$$\text{(B)} \;\le\; \underbrace{\left|V^{\star,B}_{\bar{\theta}_{k(t)}}(b_t) - V^{\star,B}_{\bar{\theta}_{k(t)}}(b_t^J)\right|}_{\text{(B1) belief approx.}} + \underbrace{V^{\star,B}_{\bar{\theta}_{k(t)}}(b_t^J) - V^{\pi_t,B}_{\bar{\theta}_{k(t)}}(b_t^J)}_{\text{(B2) planning suboptimality}} + \underbrace{\left|V^{\pi_t,B}_{\bar{\theta}_{k(t)}}(b_t^J) - V^{\pi_t,B}_{\bar{\theta}_{k(t)}}(b_t)\right|}_{\text{(B1}') \text{ belief approx.}}. \tag{12}$$

Terms (B1) and (B1′) evaluate values at the particle belief b_t^J instead of the posterior b_t, each bounded by $O(1/J^{1/(2\kappa_{\mathrm{pf}})})$ via particle-filter concentration. The planning gap (B2) splits into Monte Carlo error from MCTS rollouts ($O(1/\sqrt{N})$) and structural error from QMDP bootstrapping at depth d.

We distinguish the belief-MDP optimum $V_\theta^{\star,B}$, the deterministic *planning target* $\bar{V}^{\star,S}_{\theta,d} = \mathcal{T}_\theta^d \bar{V}_\theta^{\star,S}$ of equation 8, and the randomized MCTQ estimate $\widehat{Q}_t(b_t^J, a)$ from pooled particle sets. The target differs from the optimum by the structural error, which decays geometrically in d and involves no particles. The estimate differs from the target by the Monte Carlo error, which grows with d, so for fixed N the bound is U-shaped in d. Each component enters only through its guarantee (a β-approximate MLE, an SMC filter with an L^p bound, a planner with a Monte Carlo bound), so the decomposition extends to, e.g., POMCP-style planners or variational filters.

4.1 Assumptions and main result

Modern SMC analyses [9, 6, 21] impose rather unavoidable but stringent conditions on the transition kernel and observation density. We inherit these in the form of the following Doeblin and bounded observation density assumption (Assumption 1) on the latent dynamics. This choice aligns with our aim to provide an end-to-end regret bound that decomposes error across the learning–inference–planning pipeline, rather than to develop optimal exploration methods for POMDP model estimation, as Assumption 1 ensures that simple exploration strategies like ε-greedy have sufficient coverage. At the same time, the Doeblin condition resembles commonplace strong mixing assumptions within the RL literature [13, 39, 40, 23] and the α-revealing condition commonplace in the POMDP literature plays a similar role to the bounded observation density condition [28, 29].

Assumption 1 (Uniform mixing and bounded weights)**.**

1. Uniform Doeblin condition. *There exist* $\varepsilon_{\mathrm{mix}} > 0$ *and a probability measure* $\nu \in \Delta(\mathcal{S})$ *such that* $P_\theta(\cdot \mid s, a) \ge \varepsilon_{\mathrm{mix}}\, \nu(\cdot)$ *for all* $(\theta, s, a) \in \Theta \times \mathcal{S} \times \mathcal{A}$.
2. Bounded observation density. *There exist* $0 < g_{\min} \le g_{\max} < \infty$ *such that* $g_{\min} \le Z_\theta(o \mid s) \le g_{\max}$ *for all* $(\theta, s, o) \in \Theta \times \mathcal{S} \times \mathcal{O}$.

Let $\mathcal{N}_\Theta(\varepsilon)$ be the ε-covering number of Θ, the minimum number of ℓ_1-balls of radius ε (in trajectory likelihood) needed to cover Θ. Define the effective model complexity $\beta_T := \beta + \log\big[\mathcal{N}_\Theta(1/T)\, T/\delta\big]$, where β is the MLE tolerance from equation 6 and $\delta \in (0, 1)$ is a confidence parameter.

Theorem 1 (End-to-end regret of PF-MCTQ)**.** *Suppose Assumption 1 holds. Algorithm 1, run with exploration parameter* $\eta \asymp T^{1/4}(\beta_T/\log T)^{1/4}$ *and observation binning* ϕ *with* M *bins of diameter* $\mathrm{diam}(\phi) = \omega_Z^{-1}\big((1-\gamma) g_{\min}/(2\gamma g_{\max} V_{\max}\sqrt{N})\big)$ *(*$\phi = \mathrm{id}$ *and* $M = |\mathcal{O}|$ *when* $\mathcal{O}$ *is finite), and, if* $d \ge 2$*,* N *above the threshold of Remark 3, and* $\kappa_{\mathrm{pf}} := \max\big(1, \log\big(2g_{\max}^2/g_{\min}^2\big)/\log(1/\gamma)\big)$ *satisfies*

$$\mathrm{Reg}_T \;\le\; \underbrace{\widetilde{O}\!\left(\frac{C_{\mathrm{est}}\, R_{\max}}{(1-\gamma)^{3/2}}\, T^{3/4}\right)}_{\textit{model + exploration + filter}} + \underbrace{\frac{C_{\mathrm{pf}}\, R_{\max}}{(1-\gamma)^2} \cdot \frac{T}{J^{1/(2\kappa_{\mathrm{pf}})}}}_{\textit{particle approx.}} + \underbrace{\frac{C_{\mathrm{mc}}\, R_{\max}}{(1-\gamma)^2} \cdot \frac{T}{\sqrt{N}}}_{\textit{Monte Carlo}} + \underbrace{\frac{2\gamma^{d+1}}{(1-\gamma)^2} \sum_{t=1}^{T} \Xi_{\bar{\theta}_{k(t)}}}_{\textit{QMDP structural bias}},$$

Here $\widetilde{O}$ hides factors polylogarithmic in T, ω_Z is the modulus of continuity of the observation likelihood ratio (Appendix D.2), $M = |\bar{\mathcal{O}}|$ is the number of observation bins, and $\Xi_{\bar{\theta}_{k(t)}}$ is the QMDP structural bias of equation 13. The coefficients are

$$C_{\mathrm{est}} = \mathrm{Poly}(|\mathcal{A}|, g_{\max}/g_{\min}, \epsilon_{\mathrm{mix}}^{-1}, \log(1/\gamma), \log \mathcal{N}_\Theta(1/T), \beta), \quad C_{\mathrm{mc}} = O(d\gamma^d |\mathcal{A}|^{(d+3)/2} M^{d/2}),$$
$$C_{\mathrm{pf}} = \widetilde{O}\big(V_{\max}^2\, g_{\max}/(\varepsilon_{\mathrm{mix}}^3\, g_{\min}) \cdot (\log J)^{3/2} \sqrt{\log(\mathcal{N}_\Theta(1/T)\,|\mathcal{A}|)}\big) \quad \text{(See equation 57).}$$

Reading the bound. Theorem 1 bounds cumulative regret with four major terms. The first term captures model estimation error, exploration cost, and filter mismatch between the true posterior $b_t^{\theta^\star}$ and the particle belief $b_t^{\bar{\theta}_{1:K}}$. The $\widetilde{O}(T^{3/4})$ rate arises from the coverage transfer needed to bound the optimal-value model error (term (A) in equation 11) via ε-greedy exploration. Terms 2, 3, and 4 capture the total approximate planning errors that may scale linearly in T. They can be diminished by increasing the number of particles J, number of Monte Carlo rollouts N, and the planning depth d. The particle error decays as $J^{-1/(2\kappa_{\mathrm{pf}})}$ at cost $O(J)$, the Monte Carlo error as $N^{-1/2}$ at cost $O(N)$, and the structural error as γ^d at a tree size of up to $(|\mathcal{A}|M)^d$ nodes. The bound therefore implies a **sample-vs.-computational efficiency trade-off** through the marginal return of each budget, but a concrete allocation requires knowledge of κ_{pf} and Ξ_θ and is not claimed to be optimal. Section 5 gives an empirical budgeting rule on an instance where the assumptions hold. With continuous observations, the binning keeps the discretization bias at the Monte Carlo order $O(V_{\max}/\sqrt{N})$, but M grows as $N^{d_o/2}$ for $\mathcal{O} \subseteq \mathbb{R}^{d_o}$, so C_{mc} grows exponentially in $d \cdot d_o$. Balancing bin width against N instead gives the rate $\widetilde{O}(N^{-1/(d\,d_o+2)})$ (Remark 4). This cost is absent only at $d = 0$, the regime used in our deep RL experiments.

Technical challenges. Bounding each source in equation 12 under online learning raises three difficulties. First, controlling (B1) requires working with a test-function class defined directly over $\mathcal{S}$, rather than only the parametric family $\{Q_\theta^{\star,B}\}_\theta$. This is because the belief-space value functions may lack the regularity needed for uniform particle-filter error control. In fact, the planning-induced class over multiple steps is effectively uncountable over the possible model parameters, so we control it through truncation. The required truncation level depends on the desired particle approximation accuracy: as the particle error target decreases, the tail must be made smaller, which enlarges the truncated function class. This creates a tradeoff between the complexity of the truncated class and the particle ensemble size, leading to a slower $(1/J^{1/2\kappa_{\mathrm{pf}}})$, essentially nonparametric, rate.

Second, the filtering analysis must handle data generated online by a changing policy. Classical particle-filter stability assumes a fixed generative model, whereas here the filter receives actions from a non-stationary, improving policy. We extend time-uniform L^p bounds [9] to this setting by exploiting the Doeblin condition uniformly over actions. Third, Doeblin coverage applies to states but not actions, so we ensure action coverage via ε-greedy exploration. This lets us transfer MLE concentration across policies, controlling terms (A) and (C) and the Monte Carlo error in (B2).

Remark 1 (Relation to POMDP hardness)**.** *POMDP sample complexity may in the worst case scale exponentially in the effective horizon $1/(1-\gamma)$. Our bound is polynomial in $1/(1-\gamma)$ for two reasons. First, under the Doeblin condition the Bayes filter forgets its initialization at a geometric rate $\rho = 1 - \varepsilon_{\mathrm{mix}} g_{\min}/g_{\max} < 1$ uniformly over actions, so errors add up over time rather than compound. Second, bounded observation densities play the role of the α-revealing condition [28]: they bound importance weights, so data transfer across policies without an $(|\mathcal{A}|/\varepsilon)^H$ blow-up, and particle weights stay non-degenerate. This does not circumvent the lower bound: $\varepsilon_{\mathrm{mix}}^{-1}$ and $g_{\max}/g_{\min}$ can be exponential in the effective horizon for hard instances, and C_{est} and C_{pf} depend polynomially on them. Appendix F illustrates failures when $g_{\min} = 0$ and when Ξ_θ is large.*

Remark 2 (Bypassing exploration with offline coverage)**.** *The $\widetilde{O}(T^{3/4})$ rate is driven by the coverage transfer for term (A), which requires ε-greedy exploration to ensure model accuracy at the optimal policy. If instead the agent has access to N_{off} offline trajectories from a behavior policy π_b with single-policy concentrability $c_{\mathrm{off}}^\star$ (the worst-case ratio of trajectory ℓ_1 distances between $\pi^\star$ and π_b), then greedy exploitation ($\varepsilon_t = 0$) achieves regret $\widetilde{O}(\sqrt{T}/(1-\gamma)) + O(c_{\mathrm{off}}^\star T/(\sqrt{N_{\mathrm{off}}}(1-\gamma)))$ plus computational terms. When $N_{\mathrm{off}} \geq (c_{\mathrm{off}}^\star)^2 T$, the total regret is $\widetilde{O}(\sqrt{T})$ without any additional online exploration. The formal statement and proof are in Appendix D.7.*

4.2 Understanding the structural bias

It remains to discuss the QMDP approximation error, which is governed by the *Jensen gap*

$$\Xi_\theta := \sup_{b \in \Delta(\mathcal{S})} \left(\int_{\mathcal{S}} V_\theta^{\star,S}(s)\, b(\mathrm{d}s) - \max_{a \in \mathcal{A}} \int_{\mathcal{S}} Q_\theta^{\star,S}(s,a)\, b(\mathrm{d}s) \right), \tag{13}$$

which measures the value lost by choosing a single action for all states (QMDP planning) rather than the state-dependent optimal action. We show (Proposition 2 in Appendix D.1) that the depth-d MCTQ structural error satisfies $\|\bar{V}_{\theta,d}^{\star,S} - V_\theta^{\star,B}\|_\infty \leq \frac{\gamma^{d+1}}{1-\gamma}\Xi_\theta$, decaying geometrically with depth d.

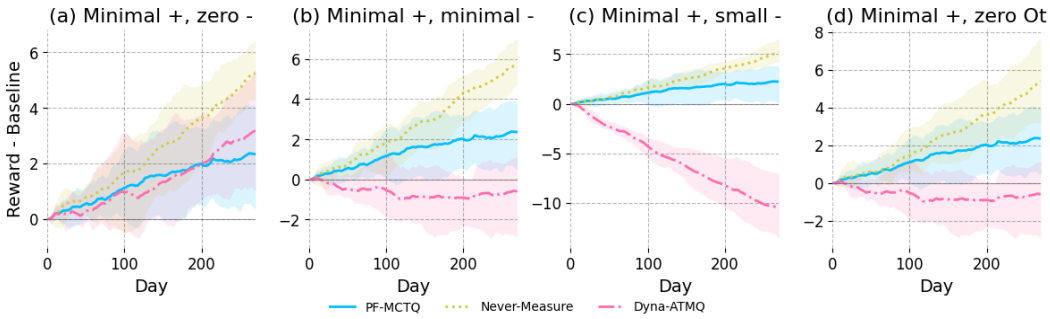


Figure 2: Cumulative reward above the baseline rollout in four HeartSteps scenarios varying the positive intervention effect and the delayed negative measurement effect or proxy O_t informativeness. Mean and 2.5%/97.5% quantiles over five seeds.

The quantity Ξ_θ has an interpretation via the posterior *variance* of value-relevant features. Let $g : \mathcal{S} \to \mathbb{R}^k$ be a measurable embedding, and write $m_g(b) := \mathbb{E}_{S\sim b}[g(S)]$ and $\Sigma_g(b) := \mathrm{Cov}_{S\sim b}(g(S))$.

Assumption 2 (Value-relevant embedding)**.** *There exist twice continuously differentiable maps $F^a_\theta : \mathbb{R}^k \to \mathbb{R}$ for each $a \in \mathcal{A}$ such that $Q^{\star,S}_\theta(s,a) = F^a_\theta(g(s))$ for all $(s,a) \in \mathcal{S} \times \mathcal{A}$, with curvature bounds, where $F_\theta(x) := \max_a F^a_\theta(x)$ and $V^{\star,S}_\theta(s) = F_\theta(g(s))$, of*

$$\sup_{a\in\mathcal{A}} \sup_{x\in\mathrm{conv}(g(\mathcal{S}))} \left\|\nabla^2 F^a_\theta(x)\right\|_{\mathrm{op}} \le L_\theta \qquad \textit{and} \qquad \sup_{x\in\mathrm{conv}(g(\mathcal{S}))} \left\|\nabla^2 F_\theta(x)\right\|_{\mathrm{op}} \le L_\theta,$$

where $\|\cdot\|_{\mathrm{op}}$ is the spectral norm.

In practice, the condition holds when either (i) a single action is optimal throughout $\mathrm{conv}(g(\mathcal{S}))$, or (ii) F_θ is a single smooth function of s (as in linear-quadratic-Gaussian (LQG) systems, where it is a quadratic form and both conditions hold with the same L_θ).

Proposition 1 (Curvature bound on Ξ_θ)**.** *Under Assumption 2, for every belief $b \in \Delta(\mathcal{S})$,*

$$\Xi_\theta(b) := \widetilde{V}^\star_\theta(b) - \bar{V}^{\star,S}_\theta(b) \;\le\; L_\theta \,\mathrm{tr}\big(\Sigma_g(b)\big), \tag{14}$$

where $\widetilde{V}^\star_\theta(b) := \int_{\mathcal{S}} V^{\star,S}_\theta(s)\, b(\mathrm{d}s)$ is the linear QMDP surrogate. In particular, $\Xi_\theta = \sup_b \Xi_\theta(b)$.

The proof is in Appendix D.8. That QMDP is accurate once the belief concentrates is expected; Proposition 1 is sharper in two ways. First, only the posterior covariance of the value-relevant features g matters; uncertainty in directions to which the value is insensitive does not enter, so the full belief need not concentrate. Second, the curvature L_θ converts this covariance into a value gap, which enters Theorem 1 as $\gamma^{d+1}\Xi_\theta/(1-\gamma)$ and trades off against planning depth. In discounted LQG, for instance, the Kalman covariance converges to a strictly positive steady state and never vanishes, yet the QMDP policy is belief-optimal, and $\Xi_\theta = 0$ when the control does not enter the dynamics (Appendix D.9).

Appendix D.9 instantiates the structural analysis on linear-quadratic Gaussian systems, where Ξ_θ has a closed form in terms of the Riccati solution and the steady-state Kalman covariance, and on contextual MDPs with identifiable latent context, where $\Xi_{\theta,t} \le 2V_{\max}\rho_t$ for a context-misidentification probability ρ_t, so eventual identification becomes the rate condition $\sum_t \rho_t = o(\sqrt{T})$.

5 Experiments

Our experiments test whether particle filtering and model-based planning can be used as a practical layer on top of modern deep reinforcement learning under partial observability. Here, our goal is not to obtain state-of-the-art asymptotic performance. Instead, we aim to evaluate whether explicit particle beliefs can improve low-sample decision making with modest additional overhead. Table 1 states which experiments fall under the theory. The deep RL experiments demonstrate that the pipeline is practical but are not evidence for the constants of Theorem 1.

Empirical validation of the regret decomposition. On a Doeblin-mixing Tiger [19] that satisfies Assumption 1 exactly ($\varepsilon_{\mathrm{mix}} = 0.05$, $0.15 \le Z \le 0.85$, $\gamma = 0.8$; Appendix E), varying one budget at

Table 1: Which experiments satisfy the assumptions of Theorem 1. ✓: holds; ×: fails; ∼: approx.

Environment	Doeblin	$g_{\min} > 0$	Finite $\mathcal{A}$	Bounded r	Role
Mixing Tiger (Sec. 5)	✓	✓	✓	✓	theory-valid
Active sensing (App. F)	✓	∼	✓	✓	structural diagnostic
Box search (App. F)	×	×	✓	✓	$g_{\min} = 0$ by design
HeartSteps	∼	∼	✓	✓	not claimed
Double-integrator LQR	✓	×	×	×	filtering diagnostic
MuJoCo	×	×	×	×	heuristic demonstration

Table 2: Fitted vs. predicted rates on mixing Tiger; Theorem 1 gives slower $J^{-1/(2\kappa_{\mathrm{pf}})}$, $T^{3/4}$.

Term	Error at endpoints	Fitted	Predicted
Particle filtering, $J = 8 \to 512$	$0.144 \to 0.012$	$J^{-0.59}$	$J^{-1/2}$
Monte Carlo, $N = 4 \to 512$	$0.171 \to 0.015$ (RMSE)	$N^{-0.50}$	$N^{-1/2}$
Structural, $d = 0 \to 12$	$0.314 \to 0.0216$	0.80 per level	$\gamma = 0.80$
Estimation (online MLE), $\sum_{t \le T} \varepsilon_{\mathrm{opt}}(t)$		$T^{0.57}$	$\sqrt{T}$

a time shows each term decaying at its predicted rate (Table 2). The smallest tested (J, N, d) putting every isolated term below 0.10, 0.05, and 0.025 is $(16, 16, 6)$, $(64, 64, 9)$, and $(128, 256, 12)$.

Deep RL experiments. We instantiate PF-MCTQ using the `pypomp` [1] package for particle-filtering primitives with MOP [37] or VSMC [30] gradient estimators. This allows highly optimized JAX particle filters on GPU. We learn a neural dynamics model online, maintain a particle approximation to the latent state posterior, and use an actor and a critic trained with recurrent SAC-style updates [31] in the same vein as [16]. In the experiments we use belief-tree depth $d = 0$: each particle is rolled out at the state level under a learned actor and bootstrapped with a learned critic, rather than planning over a tree of particle beliefs. With continuous actions and observations a depth-d tree requires binning at a cost of up to $(|\mathcal{A}|M)^d$ nodes, so this is the main practical gap from the theory. Rollout planning corresponds to the QMDP case $d = 0$, whereas deeper belief planning reduces the structural bias. Appendix G runs the analyzed MCTQ at $d > 0$ on finite POMDPs. Algorithm 4 gives the practical algorithm. All experiments took about three days on two RTX 3090 GPUs. Code is available at `https://github.com/pypomp/pf_control`.

Double-integrator LQR. We first study a partially observed double-integrator LQR. The state contains position and velocity, but the observation reveals only position. Thus the latent velocity is therefore not identifiable from the current observation alone. Different velocities induce the same observation and can only be distinguished using history. This environment isolates the filtering problem. The posterior diagnostics in Figure 1 show that particle beliefs can track the latent state without access to memory, while failures occur when the system is unstable. The GRU backbone improves this behavior relative to a memoryless MLP, suggesting that particle filtering and recurrent state representations are complementary. More details and figures can be found in Appendix C.1.

HeartSteps. We next evaluate the HeartSteps benchmark of [12] as an actively observable MDP with a latent reward state, where the learner can choose to measure the state at a cost. PF-MCTQ is run in the never-measure regime. It never directly measures the latent reward and must infer it from passive observations. The strongest benchmark is RLSVI-SMC/never-measure, which is specifically tailored to the linear-Gaussian HeartSteps structure and uses hand-specified linear features. Consequently, RLSVI-SMC performs best in most scenarios. PF-MCTQ typically improves over the baseline rollout but trails RLSVI-SMC because it must learn a deep model, actor, and critic from data. Dyna-ATMQ is not directly comparable as a no-measure method. It actively measures the latent reward roughly half the time, which gives it an informational advantage, but this advantage can be offset when measurement has delayed negative effects on future engagement. Details are in Appendix C.2.

MuJoCo partial-observation control. Finally, we evaluate Ant, HalfCheetah, Hopper, and Walker2d with position or velocity removed from the observation. We compare PF-MCTQ to Recurrent SAC and GRU-PPO baselines from the recurrent model-free POMDP benchmarking literature [31]. These are strong comparators that learn a recurrent history representation directly from reward. Figure 3 shows that PF-MCTQ is competitive in the ultra-low-sample regime. On Ant, PF-MCTQ learns a high-return policy within the first few thousand steps under both ablations. We no longer stop Ant early. When run to 49.2k steps, the submitted configuration's policy collapses (dashed). After restoring

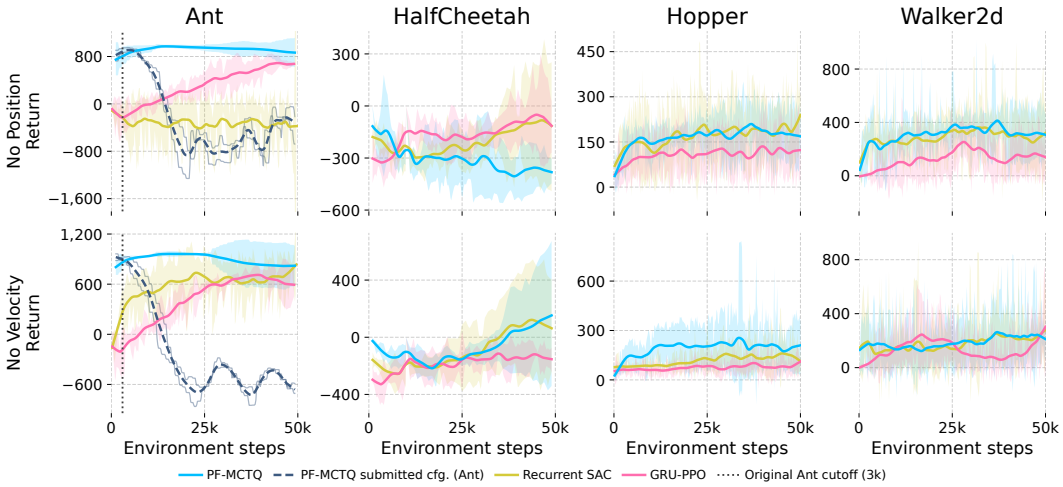


Figure 3: Partial-observation MuJoCo learning curves in the low-sample regime: smoothed means with 95% Student-t bands over five seeds (Ant: three seeds for corrected PF-MCTQ; the submitted configuration, dashed, shows its two seeds as thin lines). Ant PF-MCTQ shows the 10-episode rolling training return; others, per-episode returns. Dotted: the submitted 3k-step cutoff; none extrapolated.

online model updates and ordinary weight decay, a controlled reduction of the policy update-to-data ratio (one update per ten steps) is decisive and prevents this (Appendix H). Over the prespecified 40k–49.2k window, corrected PF-MCTQ averages 893 ± 177 and 822 ± 268 (position/velocity hidden; three seeds, 95% Student-t) against -396 and -518 for the submitted configuration. On Hopper and Walker2d, PF-MCTQ is broadly competitive with Recurrent SAC and usually stronger than GRU-PPO early in training. HalfCheetah is more mixed: PF-MCTQ performs well when velocity is hidden but struggles when position is hidden, suggesting that model bias and rollout error can dominate when the omitted coordinates are crucial for short-horizon prediction.

On Ant, corrected PF-MCTQ needs about 30% of recurrent SAC's training time per step plus a one-time 86.3 s model warm start, while GRU-PPO is faster still (Table 15). Overall, these results show that particle filtering and model-based planning can be implemented on top of existing deep RL algorithms and can yield sample-efficiency gains in partially observed control. The results should not be read as a claim of asymptotic dominance over recurrent model-free RL. Rather, they show that the learning–filtering–planning pipeline can be made operational in standard deep RL benchmarks.

6 Discussion

We presented an online model-based POMDP pipeline combining particle filtering, latent-dynamics learning, and approximate planning with an end-to-end regret bound. We decompose regret into model estimation, particle approximation, Monte Carlo planning, and QMDP structural errors, clarifying how statistical and computational budgets affect performance. The framework bridges POMDP theory and model-based deep RL. Experiments with a simplified rollout-based version show that it can be made operational and competitive in low-sample partially observed control.

Limitations. Our analysis relies on the Doeblin mixing condition and bounded observation densities, which ensure passive state coverage and obviate the need for strategic exploration. These assumptions fail in many POMDPs. In item search with exact observations, the observation density is not bounded below, and a particle set of size J misses the true location among K with probability $(1 - 1/K)^J$. In Light–Dark-type problems, information must be actively gathered, so Ξ_θ is large and shallow planning never gathers it (see Appendix F). Second, the bounded reward assumption excludes settings such as LQG. Third, the deep RL experiments use $d = 0$ with a learned critic and actor, which is the QMDP case of our algorithm, and our guarantees apply to non-adaptive action selection inside the tree but not to UCT-style selection. Natural extensions are strategic exploration to relax the dynamics assumptions, unbounded rewards, and guarantees for adaptive tree search.

Acknowledgments and Disclosure of Funding

Ziping Xu was supported by the National Science Foundation under award DMS-2625746.

References

[1] Aaron Abkemeier, Jun Chen, Kevin Tan, Jesse Wheeler, and Edward Ionides. pypomp: Partially observed Markov processes in Python, version 0.4.4.10. `https://github.com/pypomp/pypomp`, 2026.

[2] Ameya Anjarlekar, S. Rasoul Etesami, and R. Srikant. Scalable policy-based RL algorithms for POMDPs. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025.

[3] Lilian Besson and Emilie Kaufmann. What doubling tricks can and can't do for multi-armed bandits. *arXiv preprint arXiv:1803.06971*, 2018.

[4] Anthony R Cassandra, Leslie Pack Kaelbling, and Michael L Littman. Acting optimally in partially observable stochastic domains. In *Proceedings of the Twelfth National Conference on Artificial Intelligence (AAAI)*, pages 1023–1028, 1994.

[5] Yuan Chen, Mathieu Gerber, Christophe Andrieu, and Randal Douc. Self-organizing state-space models with artificial dynamics. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, page qkaf069, 2025.

[6] Nicolas Chopin and Omiros Papaspiliopoulos. *An Introduction to Sequential Monte Carlo*. Springer Series in Statistics. Springer, 2020.

[7] Adrien Corenflos, James Thornton, George Deligiannidis, and Arnaud Doucet. Differentiable particle filtering via entropy-regularized optimal transport. In Marina Meila and Tong Zhang, editors, *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 2100–2111. PMLR, 18–24 Jul 2021.

[8] Dan Crisan and Arnaud Doucet. A survey of convergence results on particle filtering methods for practitioners. *IEEE Transactions on signal processing*, 50(3):736–746, 2002.

[9] Pierre Del Moral. *Feynman-Kac Formulae: Genealogical and Interacting Particle Systems with Applications*. Springer, 2004.

[10] Randal Douc, Eric Moulines, and Jimmy Olsson. Long-term stability of sequential Monte Carlo methods under verifiable conditions. *Annals of Applied Probability*, 24(5):1767–1802, 2014.

[11] Simon S Du, Wei Hu, Zhiyuan Li, Ruoqi Shen, Zhao Song, and Jiajun Wu. When is particle filtering efficient for planning in partially observed linear dynamical systems? In *Uncertainty in artificial intelligence*, pages 728–737. PMLR, 2021.

[12] Daiqi Gao, Ziping Xu, Aseel Rawashdeh, Predrag Klasnja, and Susan Murphy. Active measuring in reinforcement learning with delayed negative effects. In *The 29th International Conference on Artificial Intelligence and Statistics*, 2025.

[13] Mudit Gaur, Amrit Singh Bedi, Di Wang, and Vaneet Aggarwal. Closing the gap: achieving global convergence (last iterate) of actor-critic under markovian sampling with neural network parametrization. In *Proceedings of the 41st International Conference on Machine Learning*, pages 15153–15179, 2024.

[14] Noah Golowich, Ankur Moitra, and Dhruv Rohatgi. Planning in observable pomdps in quasipolynomial time. arXiv preprint arXiv:2201.04735, 2022.

[15] Noah Golowich, Ankur Moitra, and Dhruv Rohatgi. Planning and learning in partially observable systems via filter stability. In *Proceedings of the 55th Annual ACM Symposium on Theory of Computing*, pages 349–362, 2023.

[16] Chao Han, Stefanos Ioannou, Luca Manneschi, T. J. Hayward, Michael Mangan, Aditya Gilra, and Eleni Vasilaki. Neural ode and sde models for adaptation and planning in model-based reinforcement learning. arXiv preprint arXiv:2603.23245, 2026.

[17] Milos Hauskrecht and Hamish Fraser. Planning treatment of ischemic heart disease with partially observable markov decision processes. *Artificial Intelligence in Medicine*, 18(3):221–244, 2000.

[18] Edward L. Ionides, Dao Nguyen, Yves Atchadé, Stilian Stoev, and Aaron A. King. Inference for dynamic and latent variable models via iterated, perturbed Bayes maps. *Proceedings of the National Academy of Sciences*, 112(3):719–724, 2015.

[19] Leslie Pack Kaelbling, Michael L Littman, and Anthony R Cassandra. Planning and acting in partially observable stochastic domains. *Artificial intelligence*, 101(1-2):99–134, 1998.

[20] Ali Devran Kara and Serdar Yüksel. Convergence of finite memory q learning for pomdps and near optimality of learned policies under filter stability. *Mathematics of Operations Research*, 48(4):2066–2093, 2023.

[21] Joona Karjalainen, Anthony Lee, Sumeetpal S. Singh, and Matti Vihola. On the forgetting of particle filters. arXiv preprint arXiv:2309.08517, 2026.

[22] Peter Karkus, David Hsu, and Wee Sun Lee. QMDP-Net: Deep learning for planning under partial observability. In *Advances in Neural Information Processing Systems*, volume 30, 2017.

[23] Chinmaya Kausik, Kevin Tan, and Ambuj Tewari. Learning mixtures of Markov chains and MDPs. In Andreas Krause, Emma Brunskill, Kyunghyun Cho, Barbara Engelhardt, Sivan Sabato, and Jonathan Scarlett, editors, *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pages 15970–16017. PMLR, 23–29 Jul 2023.

[24] Chinmaya Kausik, Kevin Tan, and Ambuj Tewari. Learning mixtures of markov chains and mdps. In *International Conference on Machine Learning*, pages 15970–16017. PMLR, 2023.

[25] Michael H Lim, Tyler J Becker, Mykel J Kochenderfer, Claire J Tomlin, and Zachary N Sunberg. Optimality guarantees for particle belief approximation of pomdps. *Journal of Artificial Intelligence Research*, 77:1591–1636, 2023.

[26] Michael H Lim, Claire J Tomlin, and Zachary N Sunberg. Sparse tree search optimality guarantees in pomdps with continuous observation spaces. In *Proceedings of the Twenty-Ninth International Conference on International Joint Conferences on Artificial Intelligence*, pages 4135–4142, 2021.

[27] Michael L Littman, Anthony R Cassandra, and Leslie Pack Kaelbling. Learning policies for partially observable environments: Scaling up. In *Machine learning proceedings 1995*, pages 362–370. Elsevier, 1995.

[28] Qinghua Liu, Alan Chung, Csaba Szepesvári, and Chi Jin. When is partially observable reinforcement learning not scary? In *Conference on Learning Theory*, pages 5175–5220. PMLR, 2022.

[29] Qinghua Liu, Praneeth Netrapalli, Csaba Szepesvari, and Chi Jin. Optimistic MLE: A generic model-based algorithm for partially observable sequential decision making. In *Proceedings of the 55th Annual ACM Symposium on Theory of Computing*, pages 363–376, 2023.

[30] Christian Naesseth, Scott Linderman, Rajesh Ranganath, and David Blei. Variational sequential monte carlo. In *International conference on artificial intelligence and statistics*, pages 968–977. PMLR, 2018.

[31] Tianwei Ni, Benjamin Eysenbach, and Ruslan Salakhutdinov. Recurrent model-free RL is a strong baseline for many pomdps. *CoRR*, abs/2110.05038, 2021.

[32] Christos H Papadimitriou and John N Tsitsiklis. The complexity of markov decision processes. *Mathematics of operations research*, 12(3):441–450, 1987.

[33] Joelle Pineau, Geoff Gordon, Sebastian Thrun, et al. Point-based value iteration: An anytime algorithm for POMDPs. In *Ijcai*, volume 3, pages 1025–1032, 2003.

[34] David Silver and Joel Veness. Monte-carlo planning in large pomdps. *Advances in neural information processing systems*, 23, 2010.

[35] Yuda Song, Yifei Zhou, Ayush Sekhari, J Andrew Bagnell, Akshay Krishnamurthy, and Wen Sun. Hybrid rl: Using both offline and online data can make rl efficient. In *International Conference on Learning Representations*, 2023.

[36] Zachary Sunberg and Mykel Kochenderfer. Online algorithms for pomdps with continuous state, action, and observation spaces. In *Proceedings of the International Conference on Automated Planning and Scheduling*, volume 28, pages 259–263, 2018.

[37] Kevin Tan, Giles Hooker, and Edward L. Ionides. Accelerated inference for partially observed markov processes using automatic differentiation. arXiv preprint arXiv:2407.03085, 2024.

[38] Nick Whiteley. Stability properties of some particle filters. *The Annals of Applied Probability*, pages 2500–2537, 2013.

[39] Yue Frank Wu, Weitong Zhang, Pan Xu, and Quanquan Gu. A finite-time analysis of two time-scale actor-critic methods. *Advances in Neural Information Processing Systems*, 33:17617–17628, 2020.

[40] Tengyu Xu, Zhe Wang, and Yingbin Liang. Improving sample complexity bounds for (natural) actor-critic algorithms. *Advances in Neural Information Processing Systems*, 33:4358–4369, 2020.

A Further Related Work

Our work is related to research on approximate planning in POMDPs, particle-based belief representations, and inference in latent-state models.

Planning in POMDPs. POMDP planning can be formulated as dynamic programming over the belief state [19], but exact planning is difficult because the belief space is continuous and must be updated online. Practical methods therefore rely on approximations. Sampling-based planners such as POMCP combine Monte-Carlo tree search with particle-based belief updates and scale to large problems without explicit belief-stateenumeration [34, 26], while progressive-widening methods extend this approach to continuous spaces [36]. Another common approximation is QMDP, which replaces belief-space continuation values with fully observed MDP values averaged under the current belief [4, 27]. Our method is connected to this line of work through combining finite-depth MCTS with a QMDP-style bootstrap, but differs in that we study this approximation in an online learning setting and explicitly quantify the resulting structural error. POMCP, POWSS, and sparse PFT are online planners in that they replan from the current belief at every step. However, they are analyzed with the model given to the planner, so their analyses are not online learning results. The same holds for our planner when it is analyzed at a frozen θ. Their guarantees concern a single planning call at an exactly known θ, with no estimation error and no horizon T, so they do not directly yield a regret bound when the planner is called with a re-estimated $\bar{\theta}_k$. Because the planner enters our decomposition only through $\epsilon_{\mathrm{mc}}(t, d)$, one could substitute such a planner by re-deriving that term (and term (B1), which a particle-belief MDP absorbs) under $\bar{\theta}_k$. Planning on the particle-belief MDP also requires $O(J)$ generative calls per node, compared with $O(1)$ for a state-level simulation.

Particle-based inference under partial observability. A related line of work uses particle filtering to approximate the belief state. Classical particle filtering theory provides convergence and stability guarantees for approximate Bayes filtering [8, 38]. [25] studied the particle-belief MDP directly and established optimality guarantees relative to the true belief-state MDP. More recent work has studied planning with particle methods in structured partially observed systems [11] as well as joint state and parameter inference in state-space models [5]. Our paper connects these directions in a different way: rather than planning over propagated particle beliefs, we use particles to represent the current belief, perform state-level lookahead from those particles, and combine this with online model learning.

This leads to an end-to-end regret analysis that separates estimation error, particle approximation error, Monte-Carlo planning error, and QMDP-style approximation bias.

Model-free approaches. Model-free alternatives instead avoid explicit filtering by giving the learner a short horizon lookback or memory [31, 20]. The latter are often effective in practice, especially with deep recurrent policies, but their hidden states are difficult to interpret as beliefs, and remain challenging to analyze. Existing sample complexity guarantees for the former often contain caveats, such as only quasipolynomial rather than polynomial sample complexity [15, 14], or non-standard assumptions [2]. We will focus on the model-based approach.

B MCTQ Planning Subroutine

We describe the full MCTQ (Monte Carlo Tree search with QMDP leaves) planning subroutine referenced in Algorithm 1. MCTQ estimates the depth-d belief-tree action-value $Q_{\theta,d}(b^J, a)$ defined in equation 9 by Monte Carlo tree search on the particle belief. It is inspired by the POMCP framework [34], with three departures: internal actions are drawn uniformly at random rather than by UCB1, every simulation descends to the full depth d, and values are computed by a bottom-up dynamic programming (DP) backup after all simulations complete.

Overview. Given the current particle belief $b^J = \frac{1}{J}\sum_{j=1}^{J} \delta_{s_j}$, model θ, search depth d, simulation budget N per root action, and a precomputed fully observable Q-function $Q_\theta^{\star,S}$ (e.g., from value iteration), MCTQ estimates $Q_{\theta,d}(b^J, a)$ for each root action $a \in \mathcal{A}$ separately (round-robin), so that every root action receives exactly N simulations. For each root action a, each simulation samples a particle s_j from b^J, takes the fixed root action a to obtain a child state $s' \sim P_\theta(\cdot \mid s_j, a)$ and observation $o \sim Z_\theta(\cdot \mid s')$, and then descends the search tree T_a from the child node (a, o) down to depth d, selecting actions uniformly at random at each internal node. Nodes are created the first time a simulation passes through them, and every node on the path records the simulated state. After all N simulations complete, a bottom-up DP backup computes value estimates: at each depth-d leaf, the QMDP value of its accumulated particles; at each internal node, $\hat{V}(v) = \max_{a'} \hat{Q}(v, a')$, where $\hat{Q}(v, a')$ is the empirical mean of $r + \gamma \hat{V}(\text{child})$ over the simulations that selected a' at v. The root estimate $\widehat{Q}(b^J, a)$ is the same empirical mean at the root.

Like POMCP [34], MCTQ indexes its search trees by action-observation histories $h = (a_1, o_1, \ldots, a_k, o_k)$ rather than by explicit beliefs, and represents the belief at each node implicitly by the collection of states simulated through that node. The round-robin over root actions is needed because the regret analysis (Lemma 4) requires uniformly good estimates $\widehat{Q}(b^J, a)$ for *all* $a \in \mathcal{A}$. The analysis (Proposition 3) uses two properties of the internal action-selection rule π, i.e., the rule that picks actions at nodes below the root within one planning call:

- (S1) *Non-adaptivity.* π may depend on the node's history, on θ, on the root particle set b^J, and on fresh randomness, but not on any statistic accumulated by the current tree.
- (S2) *Full support.* $\pi(a' \mid \cdot) \geq \pi_{\min} > 0$ for every $a' \in \mathcal{A}$.

Uniform selection satisfies both with $\pi_{\min} = 1/|\mathcal{A}|$. So does a learned proposal that is frozen during the planning call and mixed with the uniform distribution. UCB1 violates (S1), because its choice at a node depends on the visit counts and value estimates accumulated by the current tree. We therefore use uniform selection for the analysis and do not claim it is the best rule in practice. Together with the absence of an early return (every simulation descends to depth d), (S1) makes the path of each simulation a function of its own random seed alone (Lemma 5). The simulation seeds are i.i.d. given b^J, but the simulation returns are not, because the leaf evaluator pools particles across simulations and the DP backup couples them. The analysis therefore treats $\widehat{Q}(b^J, a)$ as a bounded-difference function of the i.i.d. seeds rather than as an average of independent returns.

Tree structure and belief particles. Each search tree T_a (one per root action a) has the particle belief b^J at its root (depth 0), the root action fixed to a, internal nodes at depths $1, \ldots, d-1$, and leaves at depth d. Each node v is identified with the action-observation history that leads to it from the root, so a given history always refers to the same node. A child of node v under action a' and observation o is denoted $\textsc{Child}(v, a', o)$; it is created on demand when first visited.

Each node v maintains a *particle set* $B(v)$: the collection of all states that have been simulated through v across all simulations. The empirical distribution $\hat{b}_v := \frac{1}{|B(v)|}\sum_{s\in B(v)}\delta_s$ is the particle approximation of the belief at v, and $N(v) := |B(v)|$ is the number of simulations that have visited v. In addition, each internal node stores:

- $N(v, a')$: the number of simulations that selected action a' at v;
- $\bar{r}(v, a')$: the running mean of the immediate rewards $r(s, a')$ over those simulations.

These statistics are accumulated during the simulation phase and used in the post-simulation DP backup.

QMDP leaf evaluation. At depth-d leaf nodes (remaining depth $\ell = 0$), the QMDP value of the accumulated belief is

$$\bar{V}_\theta^{\star,S}(B(v)) := \max_{a'\in\mathcal{A}} \frac{1}{|B(v)|}\sum_{s\in B(v)} Q_\theta^{\star,S}(s, a'). \tag{15}$$

This takes the max *outside* the average over states, choosing a single action for all particles, which coincides with the QMDP approximation $\bar{V}_\theta^{\star,S}$ from equation 8. This is distinct from the per-state evaluation $\frac{1}{|B(v)|}\sum_s \max_{a'} Q_\theta^{\star,S}(s, a')$, which would compute the linear surrogate $\widetilde{V}_\theta^\star$ instead. Using equation 15 ensures the algorithm matches the theoretical target $\bar{V}_{\theta,d}^{\star,S} = \mathcal{T}_\theta^d \bar{V}_\theta^{\star,S}$ analyzed in Section 4.

Simulation procedure. For each root action a, each of the N simulations proceeds as follows. A particle s_j is sampled uniformly from b^J, and the root transition is simulated: $s' \sim P_\theta(\cdot \mid s_j, a)$, $o \sim Z_\theta(\cdot \mid s')$, with immediate reward $r(s_j, a)$. Starting from the child node $\textsc{Child}(\text{root}, a, o)$ with remaining depth $\ell = d - 1$, the SIMULATE subroutine (Algorithm 3) descends the tree recursively:

1. *Record.* Add the current state s to $B(v)$.
2. *QMDP leaf ($\ell = 0$).* Stop. The leaf is evaluated by equation 15 on its final particle set once all simulations are complete.
3. *Expansion (new internal node).* If v is not yet expanded, initialize $N(v, a') \leftarrow 0$ and $\bar{r}(v, a') \leftarrow 0$ for all $a' \in \mathcal{A}$, and continue. There is no early return, so a simulation that creates a node continues to descend.
4. *Uniform selection.* Select
$$a' \sim \mathrm{Unif}(\mathcal{A}). \tag{16}$$
5. *Simulate and recurse.* Draw $s'' \sim P_\theta(\cdot \mid s, a')$ and $o' \sim Z_\theta(\cdot \mid s'')$, update $N(v, a')$ and $\bar{r}(v, a')$ with the reward $r(s, a')$, and recurse into $\textsc{Child}(v, a', o')$ with remaining depth $\ell - 1$.

Every simulation therefore visits exactly one node at each depth $1, \ldots, d$. After all N simulations for root action a, the DP backup (Phase 2 of Algorithm 2) computes, from the leaves upward,

$$\begin{aligned}
\hat{V}(v) &= \bar{V}_\theta^{\star,S}(B(v)) && \text{(leaf } v\text{)},\\
\hat{Q}(v, a') &= \bar{r}(v, a') + \gamma \sum_{w\in \mathrm{Ch}(v,a')} \frac{N(w)}{N(v, a')}\hat{V}(w) && \text{(internal } v,\ N(v, a') > 0\text{)},\\
\hat{V}(v) &= \max_{a':\, N(v,a')>0} \hat{Q}(v, a') && \text{(internal } v\text{)},
\end{aligned} \tag{17}$$

where $\mathrm{Ch}(v, a')$ is the set of children of v under action a'. Equivalently, $\hat{Q}(v, a')$ is the average of $r(s, a') + \gamma\hat{V}(\text{child})$ over the simulations that selected a' at v. The root estimate is

$$\widehat{Q}(b^J, a) = \frac{1}{N}\sum_{n=1}^{N}\big[r(s_{j_n}, a) + \gamma\,\hat{V}(v_{n,1})\big],$$

where s_{j_n} is the particle sampled by simulation n and $v_{n,1}$ is the depth-1 node it enters. When $d = 0$ no tree is built and $\widehat{Q}(b^J, a) = \frac{1}{J}\sum_{j=1}^{J} Q_\theta^{\star,S}(s_j, a)$.

The full procedure is given in Algorithms 2 and 3.

Algorithm 2 MCTQ: Monte Carlo tree QMDP planning (uniform exploration, full-depth simulations, DP backup)

Require: model θ; particle belief $b^J = \frac{1}{J}\sum_{j=1}^{J}\delta_{s_j}$; depth d; simulations per action N; MDP Q-function $Q_\theta^{\star,S}$

if $d = 0$ **then** ▷ No tree: exact QMDP
  return $\left\{\frac{1}{J}\sum_{j=1}^{J} Q_\theta^{\star,S}(s_j, a)\right\}_{a\in\mathcal{A}}$
end if
for each action $a \in \mathcal{A}$ **do** ▷ Round-robin: N simulations per root action
  Initialize empty search tree T_a
  for $n = 1, \ldots, N$ **do** ▷ Phase 1: tree construction
    Sample $j_n \sim \mathrm{Unif}([J])$
    Sample $s' \sim P_\theta(\cdot \mid s_{j_n}, a)$, $o \sim Z_\theta(\cdot \mid s')$
    $v_{n,1} \leftarrow \textsc{Child}(\mathrm{root}(T_a), a, o)$ ▷ Child of root for (a, o)
    $\textsc{Simulate}(s',\, d-1,\, \theta,\, v_{n,1})$ ▷ Alg. 3
  end for
  ▷ Phase 2: DP backup equation 17
  $\hat{V}(v) \leftarrow \max_{a'\in\mathcal{A}} \frac{1}{|B(v)|}\sum_{s\in B(v)} Q_\theta^{\star,S}(s, a')$ at each leaf v of T_a ▷ Pooled QMDP leaf
  for each internal node v of T_a, from depth $d-1$ up to depth 1 **do**
    $\hat{Q}(v, a') \leftarrow \bar{r}(v, a') + \gamma \sum_{w\in\mathrm{Ch}(v,a')} \frac{N(w)}{N(v,a')}\hat{V}(w)$ for each a' with $N(v, a') > 0$
    $\hat{V}(v) \leftarrow \max_{a':\, N(v,a')>0} \hat{Q}(v, a')$
  end for
  $\widehat{Q}(b^J, a) \leftarrow \frac{1}{N}\sum_{n=1}^{N}\left[r(s_{j_n}, a) + \gamma\, \hat{V}(v_{n,1})\right]$
end for
return $\{\widehat{Q}(b^J, a)\}_{a\in\mathcal{A}}$

Algorithm 3 SIMULATE: Recursive uniform-exploration descent to full depth

Require: state s; remaining depth ℓ; model θ; tree node v

$B(v) \leftarrow B(v) \cup \{s\}$ ▷ Record state; $N(v) = |B(v)|$
if $\ell = 0$ **then** ▷ Leaf: evaluated in Phase 2 of Alg. 2
  return
end if
if v is not yet expanded **then** ▷ Expansion: initialize and continue (no early return)
  Expand v: $N(v, a') \leftarrow 0$, $\bar{r}(v, a') \leftarrow 0$ for all $a' \in \mathcal{A}$
end if
Sample $a' \sim \mathrm{Unif}(\mathcal{A})$ ▷ Uniform, reads no tree statistic
Sample $s' \sim P_\theta(\cdot \mid s, a')$, $o \sim Z_\theta(\cdot \mid s')$
$N(v, a') \leftarrow N(v, a') + 1$; $\bar{r}(v, a') \leftarrow \bar{r}(v, a') + \frac{r(s,a') - \bar{r}(v,a')}{N(v,a')}$ ▷ Incremental mean
$\textsc{Simulate}(s',\, \ell-1,\, \theta,\, \textsc{Child}(v, a', o))$

Convergence. As $N \to \infty$, the following two properties hold. First, by Lemma 5 the events "simulation n reaches v and selects a'" are i.i.d. across n with positive probability for every reachable node v and every a', so every such node and action is visited infinitely often. Second, the states entering a node v are i.i.d. draws from the (binned) conditional belief b_v at v, so $\hat{b}_v \to b_v$ by the law of large numbers. At each depth-d leaf, $\bar{V}_\theta^{\star,S}(B(v)) \to \bar{V}_\theta^{\star,S}(b_v)$, and the DP backup then gives, level by level, $\hat{Q}(v, a') \to Q_{\theta,\ell}^{\phi}(b_v, a')$ and $\hat{V}(v) \to \max_{a'} Q_{\theta,\ell}^{\phi}(b_v, a') = \bar{V}_{\theta,\ell}^{\star,S}(b_v)$, where ℓ is the remaining depth at v. Since the root action is fixed and states are sampled from b^J, the backed-up estimate $\widehat{Q}(b^J, a) \to Q_{\theta,d}^{\phi}(b^J, a)$ for each $a \in \mathcal{A}$. Proposition 3 quantifies the rate.

When $d = 0$, no tree search is performed and the planner returns $\max_a \frac{1}{J}\sum_j Q_\theta^{\star,S}(s_j, a)$, which is the QMDP value $\bar{V}_\theta^{\star,S}(b^J)$. When $d \to \infty$, the tree-search value converges to $Q_\theta^{\star,B}(b^J, a)$.

Computational cost. For each root action, each of the N simulations traverses a path of exactly d nodes in the tree, requiring $O(d)$ generative model calls and $O(1)$ work for uniform action sampling

at each internal node. The post-simulation DP backup adds $O(|\mathcal{A}|)$ work per node but is performed once, not per simulation. The QMDP leaf evaluation equation 15 can be maintained incrementally by storing running sums $\sum_{s\in B(v)} Q_\theta^{\star,S}(s,a')$ for each action a', reducing the per-simulation leaf cost to $O(|\mathcal{A}|)$. The total cost per planning step is $|\mathcal{A}|Nd$ generative-model calls across all root actions. The additional $|\mathcal{A}|$ factor in the backup involves only arithmetic operations and no model calls.

Relation to the error analysis. The Monte Carlo error $\epsilon_{\mathrm{mc}}(t,d) = \sup_a |\widehat{Q}_t(b_t^J, a) - Q_{\theta,d}(b_t^J, a)|$ arises from the finite simulation budget N per root action. The simulation seeds are i.i.d. given b_t^J, and under (S1) with no early return each simulation's path is a function of its own seed (Lemma 5). The estimate $\widehat{Q}_t(b_t^J, a)$ is then a bounded-difference function of the seeds, so the Efron–Stein and McDiarmid inequalities control its fluctuation at rate $O(\Lambda_d/\sqrt{N})$, while the DP-backup bias vanishes at rate $O(\Phi_d/\sqrt{N})$ (Proposition 3). The simulation returns themselves are not independent, and the argument does not require them to be. The proof is written for uniform selection. For a rule satisfying (S1)–(S2), the same argument applies with $|\mathcal{A}|$ replaced by $1/\pi_{\min}$ wherever it bounds the fraction of a node's visits that select a given action.

C Figures and Additional Details for Experiments

Algorithm 4 Practical particle filtering with state-level Monte-Carlo planning

Require: number of state particles J; rollout length H_{roll}; rollouts per action N; horizon T; initial parameter $\bar{\theta}_1$; model update oracle MODELUPDATE$(\cdot)$.

Initialize: model dataset $\mathcal{D}_m$, policy dataset $\mathcal{D}_p$.
Sample $s_1^{(1)}, \dots, s_J^{(1)} \sim b_0$ and set $b_1^J = \frac{1}{J}\sum_{j=1}^J \delta_{s_j^{(1)}}$.
for $t = 1, 2, \dots, T$ **do**
  if model update **then**
    $\mathcal{D}_m = \mathcal{D}_m \cup \mathcal{D}_p$.
    Update $\bar{\theta}_t$ via updating the model on $\mathcal{D}_m$ with gradient descent.
    $\mathcal{D}_p = \emptyset$.
  else
    $\bar{\theta}_t \leftarrow \bar{\theta}_{t-1}$.
  end if
  for each particle $j = 1, \dots, J$ and each action $a \in \mathcal{A}$ **do**
    $\widehat{Q}_t(s_j^{(t)}, a) \leftarrow$ STATEMCTS$\big(s_j^{(t)}, a, \bar{\theta}_t, H_{\mathrm{roll}}, N, \widehat{Q}_{\bar{\theta}_t}^{\mathrm{MDP}}, \widehat{\pi}_{\bar{\theta}_t}^{\mathrm{MDP}}\big)$, where actions are sampled by $\widehat{\pi}_{\bar{\theta}_t}^{\mathrm{MDP}}$ and not the greedy policy w.r.t. $\widehat{Q}_t$. Collect $\widehat{a}_j^{(t)}$, the best action sampled for particle j.
  end for
  $\widehat{Q}_t(b_t^J, a) \leftarrow \frac{1}{J}\sum_{j=1}^J \widehat{Q}_t(s_j^{(t)}, a)$ for all $a \in \mathcal{A}$.
  Play $a_t \in \arg\max_{j\in[J]} \widehat{Q}_t(b_t^J, \widehat{a}_j^{(t)})$. Alternatively, sample $\tilde{s}^{(t)} \sim b_t^J$ and play $a_t \sim \widehat{\pi}_{\bar{\theta}_t}^{\mathrm{MDP}}(\cdot|\tilde{s}^{(t)})$.
  Observe (o_{t+1}, r_t) and append to $\mathcal{D}_p$.
  for $j = 1, \dots, J$ **do**
    Sample prediction $\hat{s}_j^{(t+1)} \sim P_{\bar{\theta}_t}(\cdot \mid s_j^{(t)}, a_t)$.
    Set weight $w_j^{(t+1)} \propto Z_{\bar{\theta}_t}(o_{t+1} \mid \hat{s}_j^{(t+1)})$.
  end for
  Resample $s_1^{(t+1)}, \dots, s_J^{(t+1)} \sim \mathrm{Categorical}(w_1^{(t+1)}, \dots, w_J^{(t+1)})$.
  Set $b_{t+1}^J = \frac{1}{J}\sum_{j=1}^J \delta_{s_j^{(t+1)}}$.
  Update MDP value estimate $\widehat{Q}_{\bar{\theta}_t}^{\mathrm{MDP}}(s,a)$ and policy $\widehat{\pi}_{\bar{\theta}_t}^{\mathrm{MDP}}$ for $(\mathcal{S}, \mathcal{A}, P_{\bar{\theta}_t}, r, \gamma)$ on $\mathcal{D}_p$.
end for

C.1 Double Integrator LQR

We use a partially observed double-integrator LQR system. The latent state is

$$x_t = \begin{bmatrix} p_t \\ v_t \end{bmatrix},$$

where p_t is position and v_t is velocity. The scalar control $u_t \in [-3, 3]$ is acceleration. With time step $\Delta t = 0.05$, the dynamics are

$$x_{t+1} = Ax_t + Bu_t + w_t, \qquad A = \begin{bmatrix} 1 & \Delta t \\ 0 & 1 \end{bmatrix}, \qquad B = \begin{bmatrix} \frac{1}{2}\Delta t^2 \\ \Delta t \end{bmatrix},$$

with process noise

$$w_t \sim \mathcal{N}(0, \sigma_w^2 I_2), \qquad \sigma_w = 0.01.$$

The agent observes only the position coordinate:

$$o_t = Cx_t + \epsilon_t, \qquad C = [1 \quad 0], \qquad \epsilon_t \sim \mathcal{N}(0, \sigma_o^2), \qquad \sigma_o = 0.05.$$

Thus the velocity v_t is latent. It is not recoverable from the current observation alone because the observation map is not injective:

$$\operatorname{rank}(C) = 1 < 2.$$

Equivalently, for any fixed position p and any two distinct velocities $v \neq v'$,

$$C \begin{bmatrix} p \\ v \end{bmatrix} = p = C \begin{bmatrix} p \\ v' \end{bmatrix}.$$

Therefore, even in the noiseless observation case, a single current observation cannot distinguish states that differ only in velocity. With observation noise, the ambiguity is strictly larger. Recovering information about v_t requires temporal history, e.g. past observations and actions, rather than o_t alone.

The reward is the negative quadratic LQR cost,

$$r_t = -\left(x_t^\top Q x_t + u_t^\top R u_t\right), \qquad Q = I_2, \qquad R = 0.1 I_1.$$

The discounted objective is

$$J(\pi) = \mathbb{E}_\pi \left[\sum_{t=0}^{T-1} \gamma^t r_t \right], \qquad \gamma = 0.9, \qquad T = 200.$$

In our experiments, the initial state is sampled as

$$x_0 \sim \mathcal{N}(0, I_2),$$

episodes terminate after $T = 200$ steps or when

$$\|x_t\|_2 > 20.$$

We use 32 particles for belief tracking, rollout length 1, planner discount 0.9, and a maximum of 16 candidate actions.



Figure 4: Snapshot of belief state at select training checkpoints. State recovery improves over time.

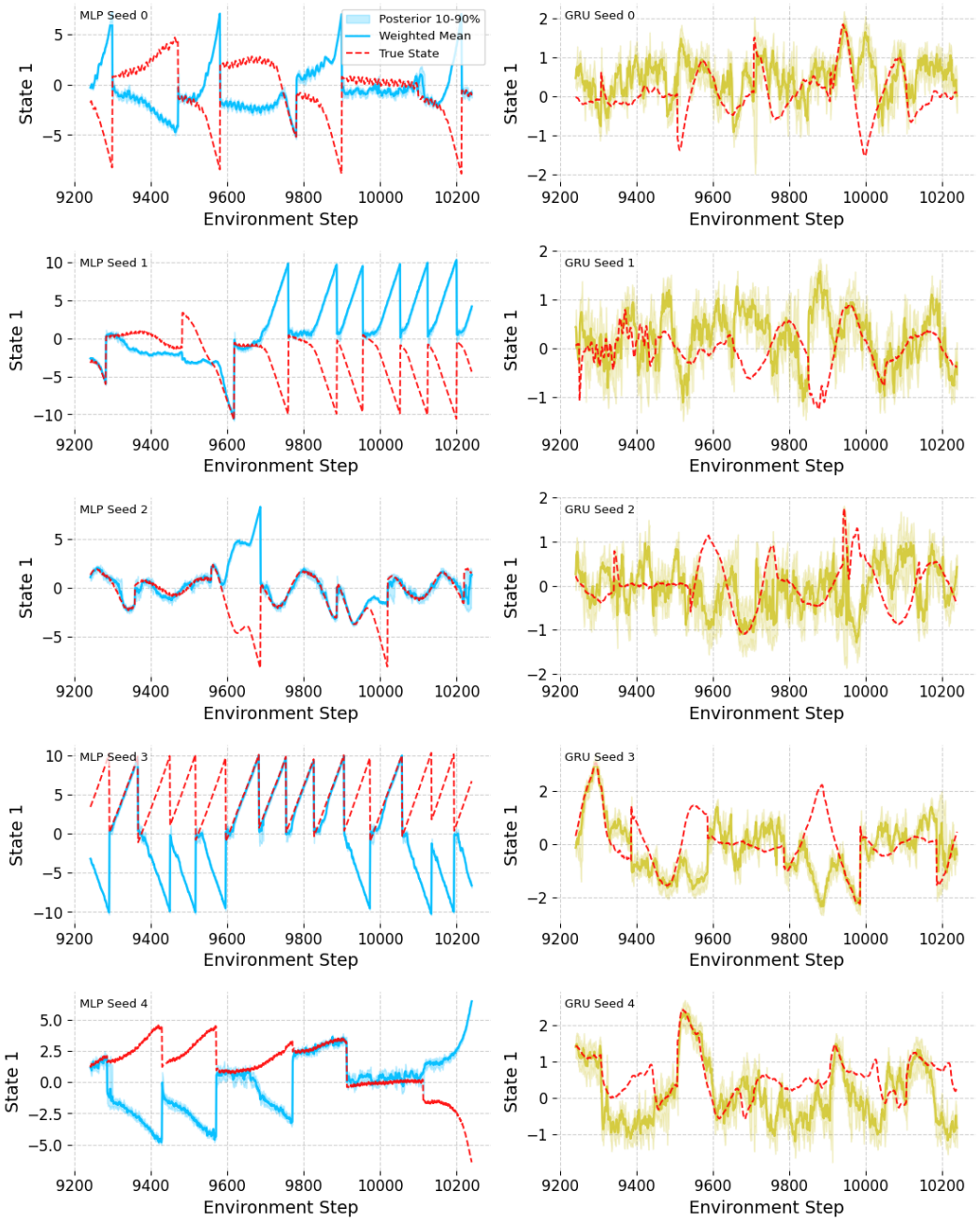


Figure 5: Posterior panel over five seeds for the double integrator LQR, over the last 1000 steps. The 10th quantile and 90th quantile are visualized. In this setting, only the magnitude, but not the direction, latent state is identifiable from the current observations alone. With no access to memory, the particle belief state still exhibits the ability to tracking the belief state, as seen in seeds 2 and 4, but fails when the state rapidly diverges. This suggests that the particle belief can be used in tandem with memory for latent state recovery..

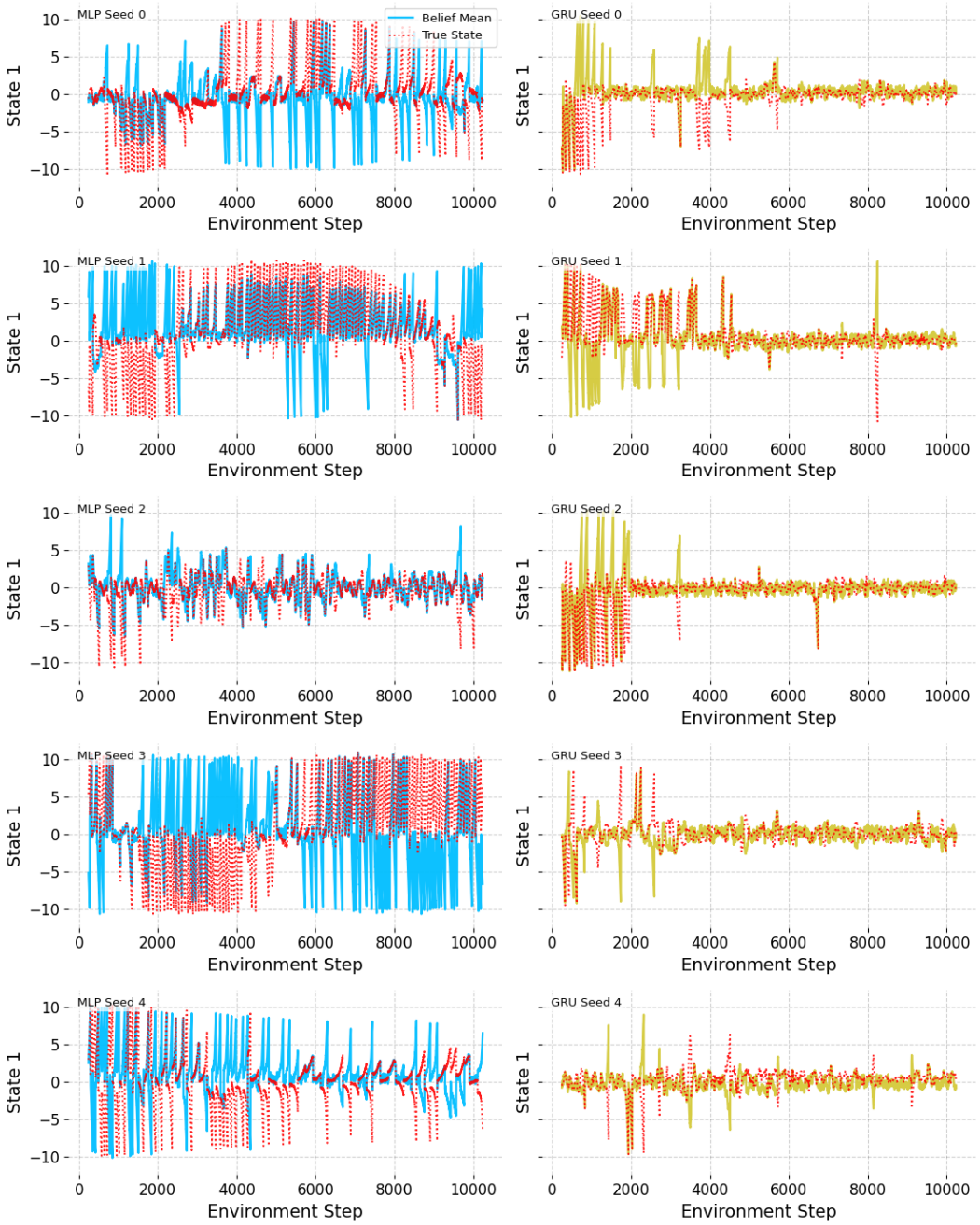


Figure 6: Belief mean panel over five seeds for LQR. The MLP backbone diverges, while the GRU backbone converges.

C.2 Heartsteps

HeartSteps as an AOMDP. We follow the formulation of [12] in formulating HeartSteps as an actively observable Markov decision process (AOMDP). At each day t, the agent first chooses a measurement action $I_{t,1} \in \{0, 1\}$, where $I_{t,1} = 1$ corresponds to sending a survey that reveals the latent reward/commitment state. The agent then chooses a control action $A_{t,2} \in \{0, 1\}$, where $A_{t,2} = 1$ corresponds to sending a walking-suggestion intervention.

Let R_t denote the user's latent commitment to being active on day t, E_t denote engagement or burden, $C_{t,2}$ denote context before the intervention decision, $M_{t,2}$ denote the proximal physical activity outcome, and O_t denote a passive proxy emission of the latent reward. Following the AOMDP indexing, the latent state is

$$U_{t+1} := R_t,$$

and the observed state after the control action is

$$Z_{t+1} := (M_{t,2}, E_t).$$

The reward is the latent reward itself,

$$r(Z_{t+1}, U_{t+1}) = R_t.$$

The measurement-stage and control-stage states are

$$S^I_{t,1} = (E_{t-1}, R_{t-1}), \qquad S^A_{t,2} = (E_{t-1}, R_{t-1}, C_{t,2}, I_{t,1}).$$

Before choosing $I_{t,1}$, the agent observes E_{t-1} and the passive emission O_{t-1}, but not R_{t-1}. If $I_{t,1} = 1$, the agent observes R_{t-1} before choosing $A_{t,2}$; if $I_{t,1} = 0$, R_{t-1} remains latent.

The HeartSteps transition model is

$$C_{t,2} = \theta^C_0 + \varepsilon^C_{t,2},$$

$$M_{t,2} = \theta^M_0 + \theta^M_1 E_{t-1} + \theta^M_2 R_{t-1} + \theta^M_3 C_{t,2} + A_{t,2}\left(\theta^M_4 + \theta^M_5 E_{t-1} + \theta^M_6 R_{t-1} + \theta^M_7 C_{t,2}\right) + \varepsilon^M_{t,2},$$

$$E_t = \theta^E_0 + \theta^E_1 E_{t-1} + \theta^E_A A_{t,2} + \theta^E_{AE} A_{t,2} E_{t-1} + \theta^E_I I_{t,1} + \theta^E_{IE} I_{t,1} E_{t-1} + \varepsilon^E_t,$$

$$R_t = \theta^R_0 + \theta^R_1 M_{t,2} + \theta^R_2 E_t + \theta^R_3 R_{t-1} + \varepsilon^R_t,$$

and the passive emission model is

$$O_t = \theta^O_0 + \theta^O_1 R_t + \varepsilon^O_t.$$

The noise terms are mean-zero, with variances $\sigma^2_C, \sigma^2_M, \sigma^2_E, \sigma^2_R, \sigma^2_O$.

The latent reward is not recoverable from the current passive observation alone. Indeed, when $I_{t,1} = 0$, the posterior over R_{t-1} satisfies

$$p(R_{t-1} = r \mid E_{t-1}, O_{t-1}, H_{t,1}) \propto p(O_{t-1} \mid R_{t-1} = r) p(R_{t-1} = r \mid E_{t-1}, H_{t,1}),$$

where

$$O_{t-1} = \theta^O_0 + \theta^O_1 R_{t-1} + \varepsilon^O_{t-1}.$$

Thus, unless the emission noise is degenerate and the prior is already a point mass, the posterior is not a Dirac measure. In particular, for $\sigma^2_O > 0$, multiple values of R_{t-1} have positive likelihood for the same observed O_{t-1}. Therefore the latent reward cannot be identified from the current passive observation alone. By contrast, when $I_{t,1} = 1$, the survey reveals R_{t-1} and collapses the posterior:

$$p(R_{t-1} \mid H_{t,2}) = \delta_{R_{t-1}}.$$

The AOMDP objective is

$$\pi^\star \in \arg\max_\pi \mathbb{E}_\pi \left[\sum_{t=1}^{H} \gamma^{t-1} R_t \right].$$

Implementation settings. In our HeartSteps experiments, the encoded simulator state is

$$x_t = (R_t, E_t, t/H, C_t, M_t, O_t) \in \mathbb{R}^6.$$

The observed coordinates are

$$(E_t, t/H, C_t, M_t, O_t),$$

and the only latent coordinate is R_t. We use horizon

$$H = 270, \qquad \gamma = 0.9,$$

binary control actions $A_{t,2} \in \{0, 1\}$, and binary measurement actions $I_{t,1} \in \{0, 1\}$ in the AOMDP formulation. The benchmark contains 42 user-specific environments.

For PF-MCTQ, our default settings use

$$J = 16$$

particles, rollout length 1, planner discount 0.9, and at most 16 candidate actions. The model and policy are trained online over 27 epochs with 10 environment steps per epoch, for a total of 270 environment steps.

Benchmarks. RLSVI-SMC is the never-measure benchmark in our experiments. It fixes the measurement action to zero,

$$I_{t,1} = 0 \qquad \text{for all } t,$$

so the latent reward is not directly observed after initialization. Instead, the algorithm maintains an SMC approximation to the posterior belief over the latent reward,

$$\widehat{b}_t^R = \sum_{j=1}^{J} w_t^{(j)} \delta_{\widehat{R}_t^{(j)}},$$

and chooses the control action using randomized least-squares value iteration with a linear value-function approximation,

$$Q_A(b_{t,2}, a) \approx \phi_A(b_{t,2}, a)^\top \beta_A.$$

Thus RLSVI-SMC is a strong and intentionally tailored benchmark for this linear-function-approximation setting: the belief update uses the specified linear-Gaussian working model, and the control rule uses hand-specified linear features of the belief state, context, action, and action interactions.

PF-MCTQ is also evaluated under the never-measure information pattern,

$$I_{t,1} = 0,$$

so it has the same lack of direct latent-reward measurements as RLSVI-SMC. However, PF-MCTQ is less specialized to the linear benchmark structure: it must learn a dynamics model, actor, and critic, parameterized by neural networks, from the observed data. Consequently, PF-MCTQ incurs an additional estimation cost relative to RLSVI-SMC, which is already matched to the linear basis and Bayesian regression structure of the benchmark.

Dyna-ATMQ differs from both no-measure methods. It is an active-measurement baseline that maintains a tabular belief over a discretized state space, learns a Dirichlet transition model and Q-table, and separately decides whether to measure by comparing the estimated value of measurement to a measurement cost. Its control action is selected from the belief-weighted Q-values,

$$A_{t,2} \in \arg\max_a \sum_s b_t(s) Q(s, a),$$

while its measurement decision can reveal the latent reward and collapse the posterior uncertainty. In our experiments Dyna-ATMQ measures the environment around half the time on average, giving it an informational advantage over both RLSVI-SMC/never-measure and PF-MCTQ, which do not measure the environment.

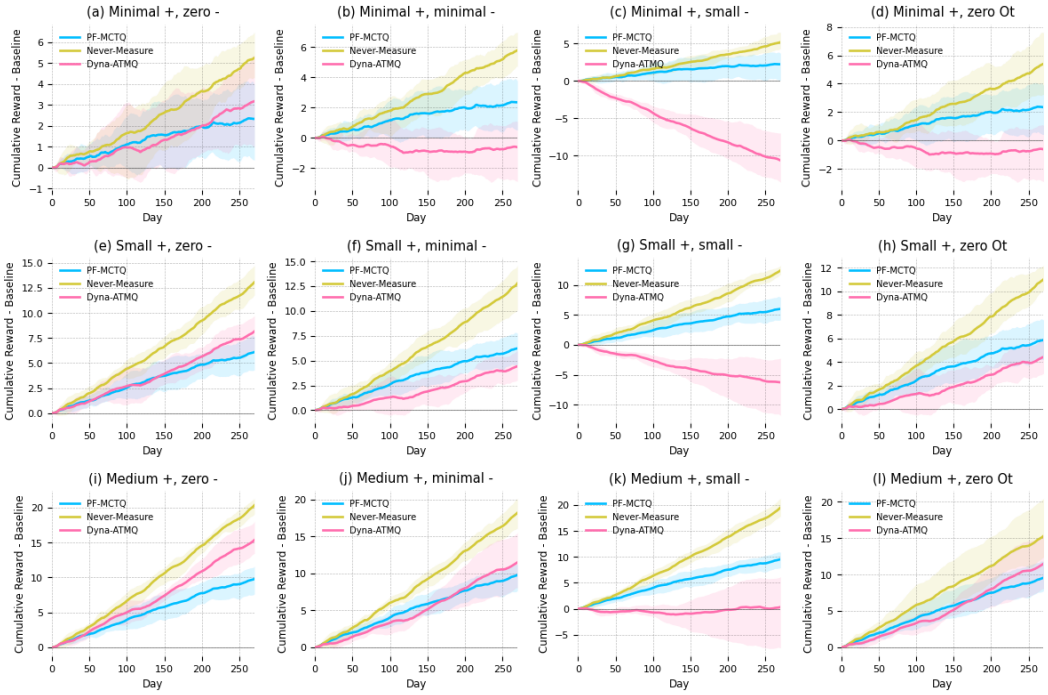


Figure 7: Cumulative reward above the baseline rollout across the twelve HeartSteps scenarios. Panels vary the positive intervention effect (rows: minimal, small, medium) and the delayed negative measurement effect or proxy informativeness (columns: zero negative effect, minimal negative effect, small negative effect, and uninformative passive proxy O_t). Solid lines show the mean across runs and shaded bands show 2.5% and 97.5% quantiles. The never-measure RLSVI-SMC benchmark performs best in most scenarios because it is specifically matched to the linear-Gaussian HeartSteps structure and is given access to hand-specified linear features. PF-MCTQ also never measures, but must learn its model, actor, and critic from data using neural approximators, so it typically improves over baseline while trailing the tailored linear benchmark. Dyna-ATMQ often underperforms when measurement has delayed negative effects, especially when the positive intervention effect is weak, because it measures frequently and therefore gains latent-state information at the cost of worsening future engagement. Its relative performance improves when the positive effect is larger or when passive proxy information is unavailable and so direct measurements become more valuable.

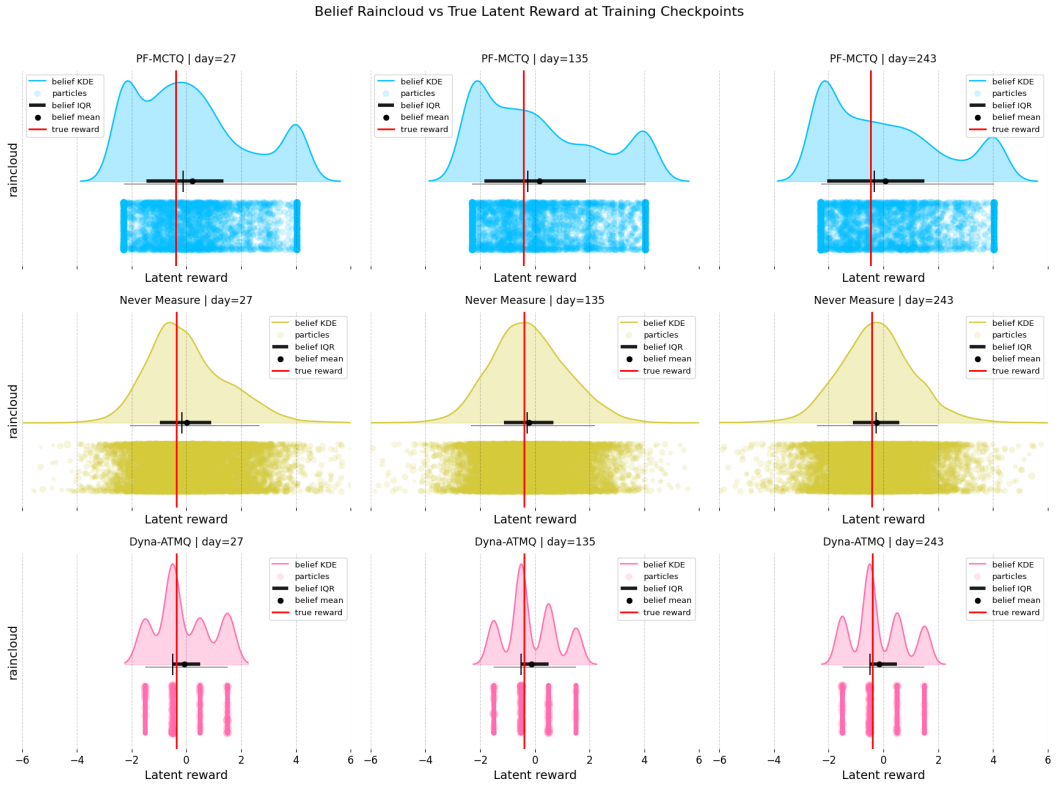


Figure 8: Belief over the latent reward at three training checkpoints in HeartSteps. Rows show PF-MCTQ, the never-measure RLSVI-SMC benchmark, and Dyna-ATMQ; columns show days 27, 135, and 243. Each panel displays posterior particles, a kernel density estimate, the posterior interquartile range, and the posterior mean. The red vertical line marks the true latent reward. We see that Dyna-ATMQ having access to measurements of the environment gives it an informational advantage over the other two algorithms. On the other hand, the correct linear specification that RLSVI-SMC enjoys allows it to create a posterior that is close to Gaussian. PF-MCTQ, which has to estimate a dynamics model, actor, and critic, parameterized by neural networks, from the observed data, consequently has a much more expressive and multimodal posterior.

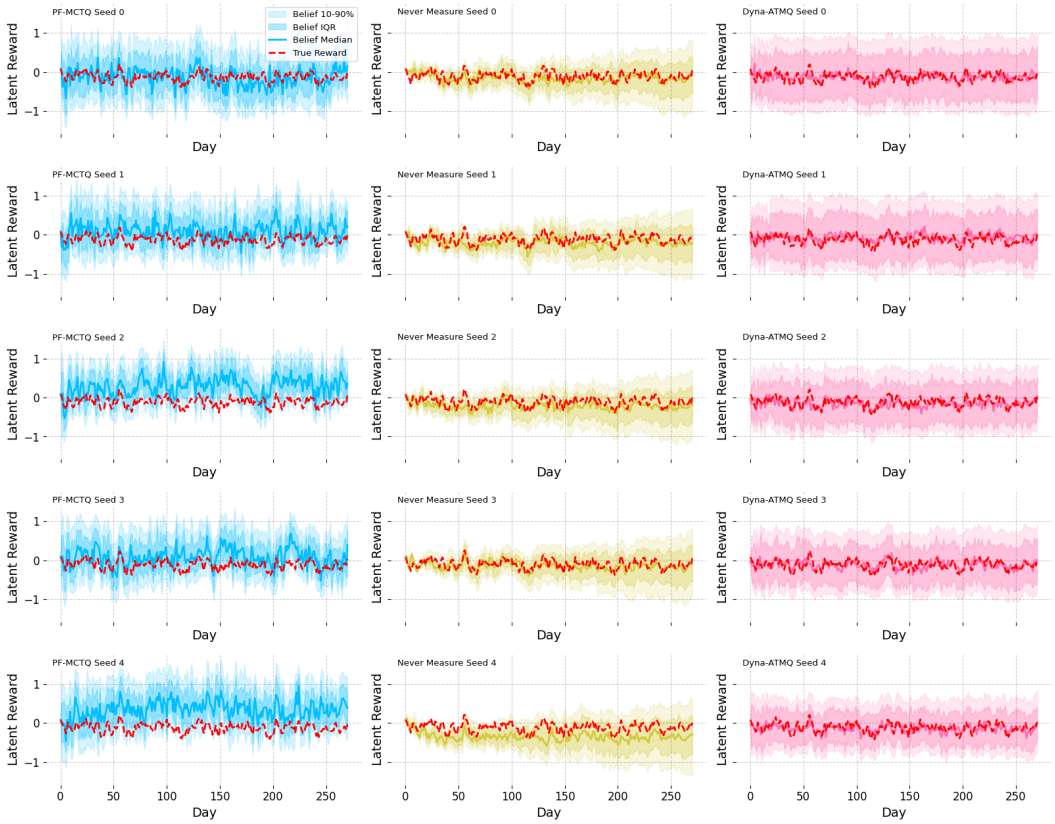


Figure 9: Latent-reward belief trajectories for HeartSteps scenario (a), shown separately by random seed. Columns compare PF-MCTQ, the never-measure RLSVI-SMC benchmark, and Dyna-ATMQ; rows show five independent seeds. Each panel plots the posterior median, interquartile range, and 10–90% interval for the latent reward, with the dashed red curve showing the true latent reward. All three methods track the latent reward at the level needed for control, but they express uncertainty differently. Never-measure produces comparatively narrow posteriors because it is tailored to the linear-Gaussian HeartSteps model and uses a conjugate SMC/RLSVI update. PF-MCTQ has wider and more variable posteriors because it must learn the transition model and value functions from data while never directly measuring the latent reward. Dyna-ATMQ also shows broad posterior support, but its belief remains centered near the true trajectory because it actively measures the latent reward on many days, partially collapsing uncertainty despite the coarser tabular approximation.

C.3 Hyperparameters

Hyperparameter	Value
Environment	LQR double integrator
Observed indices	$[0]$
Warmup steps	256
Initial model / policy updates	100 / 100
Epochs / steps per epoch	40 / 256
Model updates per epoch	10
Particles	32
Model subsequence length	64
Network hidden dimensions	$[]$
Recurrent lookback	5
Rollout length / discount	1 / 0.9
SAC actor / critic learning rate	3×10^{-4} / 3×10^{-4}
SAC batch size / entropy coefficient	32 / 0.01
Model learning rate / batch size	3×10^{-5} / 64
Particle filter mode / α	MOP / 0.8
ESS threshold / rejuvenation std.	0.5 / 0.1
Environment noise: transition / observation	0.01 / 0.05
Action bound / horizon	3.0 / 200

Table 3: Key hyperparameters for LQR.

Hyperparameter	Value
Environment	HeartSteps
User id / horizon	0 / 270
Noise mode / observation noise	sequential / 0.0
Warmup steps	1024
Initial model / policy updates	10 / 30
Epochs / steps per epoch	27 / 10
Model updates per epoch	10
Particles	16
GRU backbone / recurrent BPTT	false / false
Hidden size / preprocessor / postprocessor	32 / 32 / 32
Rollout length / discount	1 / 0.9
SAC actor / critic learning rate	10^{-3} / 10^{-3}
SAC batch size / entropy coefficient	64 / 0.003
Automatic entropy tuning	false
Model learning rate / batch size	3×10^{-4} / 256
Particle filter mode / α	MOP / 0.8
ESS threshold / rejuvenation std.	0.5 / 0.001

Table 4: Key hyperparameters for HeartSteps.

Hyperparameter	Value
Environment	HalfCheetah
Observation variants	no position, no velocity
Observation noise / discount	0.1 / 0.99
Warmup steps	1000
Initial model / policy updates	1000 / 10
Epochs / steps per epoch	50 / 1000
Model updates per epoch	100
Early-stop patience / min updates	$20/20$ no-position; $5/5$ no-velocity
Particles	16
Model subsequence length	32
GRU backbone / BPTT size	true / 32
Hidden dimensions	$[256, 256]$
Rollout length / discount	10 / 0.99
SAC actor / critic learning rate	3×10^{-4} / 3×10^{-4}
SAC batch size / entropy coefficient	256 / 0.2
Model learning rate / batch size	3×10^{-4} / 256
Particle filter mode / α	MOP / 0.8

Table 6: Key hyperparameters for MuJoCo HalfCheetah.

Hyperparameter	Value
Environment	Ant
Observation variants	no position, no velocity
Observation noise / discount	0.1 / 0.99
Warmup steps	1000
Initial model / policy updates	2000 / 10
Epochs / steps per epoch	164 / 300 (submitted: 10 / 300)
Model updates per epoch	up to 10, early stopping (submitted: 0)
Policy update every	10 env. steps (submitted: 1)
AdamW weight decay (actor, critic, model)	10^{-4} (submitted: 0.1)
Particles	16
Model subsequence length	32
GRU backbone / BPTT size	true / 32
Hidden dimensions	$[256, 256]$
Rollout length / discount	10 / 0.99
SAC actor / critic learning rate	3×10^{-4} / 3×10^{-4}
SAC batch size / entropy coefficient	256 / 0.2
Model learning rate / batch size	3×10^{-4} / 256
Particle filter mode / α	MOP / 0.8
ESS threshold / rejuvenation std.	0.5 / 0.1

Table 5: Key hyperparameters for MuJoCo Ant, using the corrected configuration of Figure 3. Values from the submitted version are given in parentheses where they differ.

Hyperparameter	Value
Environment	Hopper
Observation variants	no position, no velocity
Observation noise / discount	0.1 / 0.99
Warmup steps	1000
Initial model / policy updates	1000 / 10
Epochs / steps per epoch	50 / 1000
Model updates per epoch	100
Particles	16
Model subsequence length	32
GRU backbone / BPTT size	true / 32
Hidden dimensions	$[256, 256]$
Rollout length / discount	10 / 0.99
SAC actor / critic learning rate	3×10^{-4} / 3×10^{-4}
SAC batch size / entropy coefficient	256 / 0.2
Model learning rate / batch size	3×10^{-4} / 256
Particle filter mode / α	MOP / 0.8
ESS threshold / rejuvenation std.	0.5 / 0.1

Table 7: Key hyperparameters for MuJoCo Hopper.

Hyperparameter	Value
Environment	Walker2d
Observation variants	no position, no velocity
Observation noise / discount	0.1 / 0.99
Warmup steps	1000
Initial model / policy updates	1000 / 10
Epochs / steps per epoch	50 / 1000
Model updates per epoch	100
Early-stop patience / min updates	$20/20$ no-position; $5/5$ no-velocity
Particles	16
Model subsequence length	32
GRU backbone / BPTT size	true / 32
Hidden dimensions	$[256, 256]$
Rollout length / discount	10 / 0.99
SAC actor / critic learning rate	3×10^{-4} / 3×10^{-4}
SAC batch size / entropy coefficient	256 / 0.2
Model learning rate / batch size	3×10^{-4} / 256
Particle filter mode / α	MOP / 0.8

Table 8: Key hyperparameters for MuJoCo Walker2d.

D Detailed Error Analysis and Proofs

Throughout, $|r(s,a)| \le R_{\max}$ and $V_{\max} := R_{\max}/(1-\gamma)$. We define the per-round error quantities:

$$\epsilon_{\mathrm{mc}}(t,d) := \sup_{a \in \mathcal{A}} \big|\widehat{Q}_t(b_t^J, a) - Q_{\theta,d}(b_t^J, a)\big|, \qquad \text{(Monte Carlo error)}$$

$$\epsilon_{\mathrm{struct}}(t,d) := \sup_{a \in \mathcal{A}} \big|Q_\theta^{\star,B}(b_t^J, a) - Q_{\theta,d}(b_t^J, a)\big|. \qquad \text{(structural error)}$$

D.1 Structural Error of QMDP Bootstrapping

Define the *linear QMDP surrogate*

$$\widetilde{V}_\theta^\star(b) := \int_{\mathcal{S}} V_\theta^{\star,S}(s)\, b(\mathrm{d}s). \tag{18}$$

By Jensen's inequality, $\widetilde{V}_\theta^\star(b) \geq \bar{V}_\theta^{\star,S}(b)$ and the gap $\Xi_\theta := \sup_b[\widetilde{V}_\theta^\star(b) - \bar{V}_\theta^{\star,S}(b)] \geq 0$.

Proposition 2 (Structural error of MCTQ)**.** *For every $\theta \in \Theta$ and $d \geq 0$,*

$$\|\bar{V}_{\theta,d}^{\star,S} - V_\theta^{\star,B}\|_\infty \;\leq\; \frac{\gamma^{d+1}}{1-\gamma}\,\Xi_\theta. \tag{19}$$

In particular, $\epsilon_{\mathrm{struct}}(u,d) \leq \frac{\gamma^{d+1}}{1-\gamma}\Xi_{\bar{\theta}_u}$.

Proof of Proposition 2. Write

$$Q_\theta^{\star,B}(b,a) = r(b,a) + \gamma \int_{\mathcal{O}} V_\theta^{\star,B}(\tau_\theta(b,a,o))\, p_\theta(o \mid b,a)\,\mathrm{d}o,$$
$$Q_{\theta,0}(b,a) = r(b,a) + \gamma \int_{\mathcal{S}} V_\theta^{\star,S}(s')\, \tilde{b}_\theta(\mathrm{d}s' \mid b,a) = r(b,a) + \gamma\, \widetilde{V}_\theta^\star(\tilde{b}_\theta(\cdot \mid b,a)).$$

By Bayes consistency and linearity of $\widetilde{V}_\theta^\star$:

$$\widetilde{V}_\theta^\star(\tilde{b}_\theta(\cdot \mid b,a)) = \int_{\mathcal{O}} \widetilde{V}_\theta^\star(\tau_\theta(b,a,o))\, p_\theta(o \mid b,a)\,\mathrm{d}o.$$

Subtracting and using $\widetilde{V}_\theta^\star = \bar{V}_\theta^{\star,S} + (\widetilde{V}_\theta^\star - \bar{V}_\theta^{\star,S})$ with $0 \leq \widetilde{V}_\theta^\star - \bar{V}_\theta^{\star,S} \leq \Xi_\theta$:

$$\|Q_\theta^{\star,B} - Q_{\theta,0}\|_\infty \;\leq\; \gamma \|V_\theta^{\star,B} - \bar{V}_\theta^{\star,S}\|_\infty + \gamma\,\Xi_\theta.$$

Since $\|V_\theta^{\star,B} - \bar{V}_\theta^{\star,S}\|_\infty = \sup_b |\max_a Q_\theta^{\star,B}(b,a) - \max_a Q_{\theta,0}(b,a)| \leq \|Q_\theta^{\star,B} - Q_{\theta,0}\|_\infty$ (using $|\max f - \max g| \leq \max|f-g|$), rearranging gives $\|V_\theta^{\star,B} - \bar{V}_\theta^{\star,S}\|_\infty \leq \frac{\gamma}{1-\gamma}\Xi_\theta$ (the $d=0$ case). For general $d \geq 0$: $\bar{V}_{\theta,d}^{\star,S} = \mathcal{T}_\theta^d \bar{V}_\theta^{\star,S}$ and $V_\theta^{\star,B} = \mathcal{T}_\theta V_\theta^{\star,B}$ (fixed point), so

$$\|\bar{V}_{\theta,d}^{\star,S} - V_\theta^{\star,B}\|_\infty = \|\mathcal{T}_\theta^d \bar{V}_\theta^{\star,S} - \mathcal{T}_\theta^d V_\theta^{\star,B}\|_\infty \leq \gamma^d \|\bar{V}_\theta^{\star,S} - V_\theta^{\star,B}\|_\infty \leq \frac{\gamma^{d+1}}{1-\gamma}\Xi_\theta. \qquad \square$$

D.2 Planning and State-Identification Error

Definition 1 (Q-belief distance)**.** *For $b, b' \in \Delta(\mathcal{S})$, define $d_Q(b,b') := \sup_{\theta\in\Theta,\, a\in\mathcal{A}} |Q_\theta^{\star,B}(b,a) - Q_\theta^{\star,B}(b',a)|$.*

Lemma 1 ($Q^{\star,B}$-difference via test functions)**.** *Fix $\theta \in \Theta$, $a \in \mathcal{A}$, and truncation depth $H \geq 1$. Recall the Bayesian update $\tau(b,a,o) \in \Delta(\mathcal{S})$, the observation marginal $p(o \mid b,a) = \int Z_\theta(o \mid s') P_\theta(s' \mid s,a)\, b(ds)\, ds'$, and define*

$$\psi_{o,a}(s) := \int Z_\theta(o \mid s') P_\theta(s' \mid s,a)\, ds', \qquad \tilde{f}_{o,a}(s) := \int f(s') Z_\theta(o \mid s') P_\theta(s' \mid s,a)\, ds', \tag{20}$$

so that $p(o \mid b,a) = b(\psi_{o,a})$ and $\tau(b,a,o)(f) = b(\tilde{f}_{o,a})/b(\psi_{o,a})$.

Under Assumption 1, for all $b, b' \in \Delta(\mathcal{S})$:

$$|Q_\theta^{\star,B}(b,a) - Q_\theta^{\star,B}(b',a)| \;\leq\; \sum_{\ell=0}^{H-1} \gamma^\ell\, C_{\mathrm{BU}}^\ell \sup_{f\in\mathcal{F}_\ell(\theta,a)} |b(f) - b'(f)| \;+\; 2\gamma^H V_{\max}, \tag{21}$$

where $C_{\mathrm{BU}} := 2g_{\max}/g_{\min}^2$ and $\mathcal{F}_\ell(\theta,a)$ is a class of functions $\mathcal{S} \to [-g_{\max}^\ell V_{\max},\, g_{\max}^\ell V_{\max}]$ with $\mathcal{N}_{[]}(\mathcal{F}_\ell, \epsilon) \leq |\mathcal{A}|^\ell \mathcal{N}_\Theta(\epsilon/(\ell V_{\max}))$, defined recursively:

- $\mathcal{F}_0(\theta,a) = \{r(\cdot,a') : a' \in \mathcal{A}\} \cup \{Q_\theta^{\star,S}(\cdot,a') : a' \in \mathcal{A}\}$,
- $\mathcal{F}_{\ell+1}(\theta,a) = \big\{\tilde{f}_{o,a''} :\, f \in \mathcal{F}_\ell(\theta,a''),\, a'' \in \mathcal{A}\big\} \cup \big\{\psi_{o,a''} :\, a'' \in \mathcal{A}\big\}$.

Proof. We prove by induction on h: for any $\Phi(b) = \mathbb{E}^{\pi}_{\theta}[\sum_{u=0}^{h-1} \gamma^u r_u \mid b_0 = b]$ with $h \le H$, $|\Phi(b) - \Phi(b')|$ admits the decomposition equation 21. Taking $\Phi = Q^{\star,B}_{\theta}(\cdot, a)$ with $h = H$ and bounding the tail by $2\gamma^H V_{\max}$ yields the result.

Base case ($h = 1$). $|\Phi(b) - \Phi(b')| = |\sup_a[b(r_a) - b'(r_a)]| \le \sup_{f \in \mathcal{F}_0} |b(f) - b'(f)|$.

Inductive step ($h \to h + 1$). For a fixed action a, write

$$\Phi(b) = b(r_a) + \gamma \int_{\mathcal{O}} \Phi_{h-1}\big(\tau(b, a, o)\big)\, p(o \mid b, a)\, do,$$

where Φ_{h-1} is the $(h-1)$-step continuation. The continuation difference decomposes as:

$$\begin{aligned} &\left| \int \Phi_{h-1}(\tau(b,a,o))\, p(o|b,a)\, do - \int \Phi_{h-1}(\tau(b',a,o))\, p(o|b',a)\, do \right| \\ &\le \underbrace{\|\Phi_{h-1}\|_\infty \int |(b-b')(\psi_{o,a})|\, do}_{\text{(I) obs. distribution shift}} + \underbrace{\int |\Phi_{h-1}(\tau(b,a,o)) - \Phi_{h-1}(\tau(b',a,o))|\, p(o|b',a)\, do}_{\text{(II) belief propagation}}, \end{aligned} \tag{22}$$

using $p(o \mid b, a) - p(o \mid b', a) = (b - b')(\psi_{o,a})$.

Term (I) contributes $\psi_{o,a} \in \mathcal{F}_{\ell+1}$.

Term (II) requires bounding $|\tau(b, a, o)(f) - \tau(b', a, o)(f)|$ for $\|f\|_\infty \le 1$. Since $\tau(b, a, o)(f) = b(\tilde{f}_{o,a})/b(\psi_{o,a})$, the quotient rule gives

$$\frac{b(\tilde{f}_{o,a})}{b(\psi_{o,a})} - \frac{b'(\tilde{f}_{o,a})}{b'(\psi_{o,a})} \;=\; \frac{(b-b')(\tilde{f}_{o,a})}{b(\psi_{o,a})} - \frac{b'(\tilde{f}_{o,a})}{b(\psi_{o,a})\, b'(\psi_{o,a})}\, (b-b')(\psi_{o,a}). \tag{23}$$

Since $b(\psi_{o,a}), b'(\psi_{o,a}) \ge g_{\min}$ and $|b'(\tilde{f}_{o,a})| \le g_{\max}$:

$$|\tau(b,a,o)(f) - \tau(b',a,o)(f)| \;\le\; \frac{|(b-b')(\tilde{f}_{o,a})|}{g_{\min}} + \frac{g_{\max}}{g^2_{\min}}\, |(b-b')(\psi_{o,a})|. \tag{24}$$

Both $\tilde{f}_{o,a}$ and $\psi_{o,a}$ are in $\mathcal{F}_{\ell+1}$, with per-update amplification

$$C_{\mathrm{BU}} = \frac{1}{g_{\min}} + \frac{g_{\max}}{g^2_{\min}} \;\le\; \frac{2\, g_{\max}}{g^2_{\min}},$$

By the inductive hypothesis, $|\Phi_{h-1}(\tau(b, a, o)) - \Phi_{h-1}(\tau(b', a, o))|$ decomposes into evaluations of $\tau(b, a, o)$ vs. $\tau(b', a, o)$ on $\mathcal{F}_\ell$. Substituting equation 24 reduces each to evaluations of b vs. b' on $\mathcal{F}_{\ell+1}$, with one factor of C_{BU} per level. Taking $\sup_a$ over the action set and iterating across H levels yields equation 21. □

Lemma 2 ($Q^{\star,B}$ is Lipschitz in TV)**.** *For any $\theta \in \Theta$, $a \in \mathcal{A}$, and $b, b' \in \Delta(\mathcal{S})$,*

$$|Q^{\star,B}_{\theta}(b, a) - Q^{\star,B}_{\theta}(b', a)| \;\le\; 2V_{\max}\, \|b - b'\|_{\mathrm{TV}}.$$

Proof. $|r(b, a) - r(b', a)| \le 2R_{\max}\|b - b'\|_{\mathrm{TV}}$. The trajectory law under b and b' differs in TV by at most $\|b - b'\|_{\mathrm{TV}}$ (data processing), so the $V_{\max}$-bounded continuation differs by at most $2V_{\max}\|b - b'\|_{\mathrm{TV}}$. Using $R_{\max} + \gamma V_{\max} = V_{\max}$ gives the result. □

Lemma 3 (Performance decomposition)**.** *Fix $\theta \in \Theta$. Let π_t denote the* stationary *policy induced by the planner with frozen parameters (model $\bar{\theta}_{k(t)}$, belief b^J_t) at time t, and let $\delta_u := V^{\star,B}_{\theta}(b_u) - Q^{\star,B}_{\theta}(b_u, A_u)$ where $A_u \sim \pi_t(\cdot \mid H_u)$. Then*

$$V^{\star,B}_{\theta}(b_t) - V^{\pi_t,B}_{\theta}(b_t) \;\le\; \mathbb{E}\left[\sum_{u=t}^{\infty} \gamma^{u-t} \delta_u \;\middle|\; H_t\right].$$

Proof. Unroll the Bellman recursion $V^{\star,B}_{\theta}(b_u) - V^{\pi_t,B}_{\theta}(b_u) = \delta_u + \gamma\, \mathbb{E}[V^{\star,B}_{\theta}(b_{u+1}) - V^{\pi_t,B}_{\theta}(b_{u+1}) \mid H_u]$ from $u = t$ to $u = t+n$, using $|V^{\star,B}_{\theta} - V^{\pi_t,B}_{\theta}| \le 2V_{\max}$ as the tail bound, and send $n \to \infty$. □

Lemma 4 (Per-step bound on δ_u)**.** *For every u and any $d \geq 0$,*

$$\delta_u \;\leq\; 2\, d_Q(b_u, b_u^J) + 2\, \epsilon_{\mathrm{mc}}(u,d) + 2\, \epsilon_{\mathrm{struct}}(u,d).$$

Proof. Let $a_u^\star \in \arg\max_a Q_\theta^{\star,B}(b_u,a)$ and $\hat{a}_u \in \arg\max_a \widehat{Q}_u(b_u^J,a)$. Then

$$\begin{aligned}\delta_u &= Q_\theta^{\star,B}(b_u,a_u^\star) - Q_\theta^{\star,B}(b_u,\hat{a}_u)\\ &\leq \underbrace{|Q_\theta^{\star,B}(b_u,a_u^\star) - Q_\theta^{\star,B}(b_u^J,a_u^\star)| + |Q_\theta^{\star,B}(b_u^J,\hat{a}_u) - Q_\theta^{\star,B}(b_u,\hat{a}_u)|}_{\leq\, 2\, d_Q(b_u,b_u^J) \text{ by Definition 1}}\\ &\quad + Q_\theta^{\star,B}(b_u^J,a_u^\star) - Q_\theta^{\star,B}(b_u^J,\hat{a}_u).\end{aligned}$$

For the last line, add and subtract $Q_{\theta,d}$ and $\widehat{Q}_u$, using optimality of $\hat{a}_u$ for $\widehat{Q}_u$:

$$Q_{\theta,d}(b_u^J,a_u^\star) \leq \widehat{Q}_u(b_u^J,a_u^\star) + \epsilon_{\mathrm{mc}} \leq \widehat{Q}_u(b_u^J,\hat{a}_u) + \epsilon_{\mathrm{mc}} \leq Q_{\theta,d}(b_u^J,\hat{a}_u) + 2\epsilon_{\mathrm{mc}},$$

so $Q_\theta^{\star,B}(b_u^J,a_u^\star) - Q_\theta^{\star,B}(b_u^J,\hat{a}_u) \leq 2\epsilon_{\mathrm{struct}} + 2\epsilon_{\mathrm{mc}}$. □

Theorem 2 (Round-t planning + state-identification bound)**.** *For each $t \geq 1$ and any depth d,*

$$V_{\bar{\theta}_{k(t)}}^{\star,B}(b_t) - V_{\bar{\theta}_{k(t)}}^{\pi_t,B}(b_t) \;\leq\; \mathbb{E}\left[\sum_{u=t}^{\infty} \gamma^{u-t}\big(2\, d_Q(b_u,b_u^J) + 2\, \epsilon_{\mathrm{mc}}(u,d) + 2\, \epsilon_{\mathrm{struct}}(u,d)\big) \;\middle|\; H_t\right].$$

Proof. Substitute the per-step bound of Lemma 4 into Lemma 3. □

Observation discretization. For continuous $\mathcal{O}$, the tree never revisits observation-children (each sampled o is unique), so every node below the root holds a single particle. At $d = 1$ each leaf value equation 15 then reduces to $\max_{a'} Q_\theta^{\star,S}(s',a')$ at a single state, which targets the linear surrogate rather than $Q_{\theta,1}$, and for $d \geq 2$ the tree cannot build depth beyond 1. We therefore discretize the observations. Let $\phi : \mathcal{O} \to \bar{\mathcal{O}}$ with $|\bar{\mathcal{O}}| = M$ and bin diameter $\mathrm{diam}(\phi) := \max_{\bar{o}} \sup_{o,o' \in \phi^{-1}(\bar{o})} \|o - o'\|$. Algorithms 2 and 3 are modified so that $\textsc{Child}(v,a',\phi(o'))$ replaces $\textsc{Child}(v,a',o')$. The bin-averaged posterior is $\tau_\theta^\phi(b,a,\bar{o}) := \int_{\phi^{-1}(\bar{o})} \tau_\theta(b,a,o)\, p_\theta(o \mid b,a)\, \mathrm{d}o \big/ \int_{\phi^{-1}(\bar{o})} p_\theta(o \mid b,a)\, \mathrm{d}o$, and we write $Q_{\theta,d}^\phi$ for the target equation 9 with $\tau_\theta(b,a,o)$ replaced by $\tau_\theta^\phi(b,a,\phi(o))$ at every level. When $|\mathcal{O}| < \infty$, take $\phi = \mathrm{id}$, $M = |\mathcal{O}|$, $\mathrm{diam}(\phi) = 0$. When $d = 0$, no tree is built and no discretization is needed.

Seeds and the estimator. Fix a round t and a root action a, and write $\theta := \bar{\theta}_{k(t)}$ and $b^J := b_t^J$. All probabilities and expectations below are conditional on b_t^J, and we take $d \geq 1$. Simulation $n \in [N]$ is driven by a *seed*

$$\xi_n := \big(j_n,\, u_{n,0},\, (a'_{n,\ell}, u_{n,\ell})_{\ell=1}^{d-1}\big), \tag{25}$$

where $j_n \sim \mathrm{Unif}([J])$ is the sampled particle index, $a'_{n,\ell} \sim \mathrm{Unif}(\mathcal{A})$ is the action drawn at depth ℓ, and $u_{n,\ell}$ is the noise that generates the next state and observation at depth ℓ from P_θ and Z_θ. The seeds $\xi_1,\ldots,\xi_N$ are i.i.d. Running Algorithm 2 on them produces the tree T_a, and we write the root estimate as a deterministic function of all seeds,

$$\widehat{Q}(b^J,a) = f_a(\xi_1,\ldots,\xi_N) := \frac{1}{N}\sum_{n=1}^{N}\big[r(s_{j_n},a) + \gamma\, \hat{V}(v_{n,1})\big], \tag{26}$$

where $v_{n,\ell}$ is the depth-ℓ node on the path of simulation n and $\hat{V}$ is the DP backup equation 17. The function f_a is not an average of independent terms, since $\hat{V}(v_{n,1})$ depends on every simulation that passes through $v_{n,1}$. For $S \subseteq [N]$, let $f_a^{(-S)}$ be the same expression computed from the seeds $(\xi_m)_{m \notin S}$ alone, with the sum running over $n \notin S$ and the normalizer $1/N$ retained.

Balancedness. For $\rho \geq 1$, let $\mathcal{E}_\rho$ be the event that every internal node v of T_a satisfies

$$N(v) \leq \rho\,|\mathcal{A}|\,N(v,a') \qquad \text{for all } a' \in \mathcal{A}, \tag{27}$$

so that every action receives at least a $1/(\rho|\mathcal{A}|)$ fraction of the visits of every internal node. Define the depth factor

$$\Lambda_d := \sum_{\ell=0}^{d-1} (\rho\gamma|\mathcal{A}|)^\ell, \tag{28}$$

so that $\Lambda_0 = 0$ and $\Lambda_1 = 1$. Let $p_\rho := \max_{a\in\mathcal{A}} \Pr(\mathcal{E}_\rho^c \mid b_t^J)$, where the event refers to the tree T_a, and let

$$q_{\min} := \inf_{s\in\mathcal{S},\, a''\in\mathcal{A},\, \bar{o}\in\bar{\mathcal{O}}} \int_{\mathcal{S}} Z_\theta\big(\phi^{-1}(\bar{o}) \mid s'\big)\, P_\theta(\mathrm{d}s' \mid s, a'')$$

be the smallest one-step bin probability. Under the bounded observation density of Assumption 1, $q_{\min} \geq g_{\min} \min_{\bar{o}} \lambda(\phi^{-1}(\bar{o}))$, where λ is the reference measure of Z_θ; for finite $\mathcal{O}$ with the counting measure, $q_{\min} \geq g_{\min}$.

Proposition 3 (Monte Carlo error)**.** *Run Algorithm 2 with uniform internal selection equation 16, N simulations per root action, observation discretization ϕ ($|\bar{\mathcal{O}}| = M$), and returns in $[-V_{\max}, V_{\max}]$, and let $\rho \geq 1$. Let*

$$\mathrm{bias}(N,d) := \sup_a \big|\mathbb{E}[\widehat{Q}(b^J,a) \mid b_t^J] - Q_{\bar{\theta}_{k(t)},d}(b^J,a)\big| \tag{29}$$

and $B_d := 2V_{\max}\big(1 + 2(1+\gamma)\Lambda_d\big)$. For every $\delta \in (0,1)$, with probability at least $1 - \delta - |\mathcal{A}|p_\rho$ conditionally on b_t^J,

$$\epsilon_{\mathrm{mc}}(t,d) \;\leq\; \underbrace{2(1+\gamma)V_{\max}\Lambda_d\sqrt{\frac{2\log(2|\mathcal{A}|/\delta)}{N}}}_{\textit{fluctuation}} + \mathrm{bias}(N,d) \;+\; B_d\, p_\rho, \tag{30}$$

and, in second moment,

$$\mathbb{E}\big[\epsilon_{\mathrm{mc}}(t,d)^2 \,\big|\, b_t^J\big] \;\leq\; \frac{8(1+\gamma)^2\,|\mathcal{A}|\,V_{\max}^2\,\Lambda_d^2}{N} \;+\; 2\,|\mathcal{A}|\,\mathrm{bias}(N,d)^2 \;+\; 3\,|\mathcal{A}|\,B_d^2\,p_\rho. \tag{31}$$

The bias satisfies

$$\mathrm{bias}(N,d) \;\leq\; \underbrace{\frac{c_0\,\Phi_d\,V_{\max}\sqrt{|\mathcal{A}|}}{\sqrt{N}}}_{\textit{DP backup estimation error}} + \underbrace{\frac{2\gamma(1-\gamma^d)}{1-\gamma}\cdot\frac{g_{\max}}{g_{\min}}\,V_{\max}\,\omega_Z(\mathrm{diam}(\phi))}_{\textit{discretization error}}, \tag{32}$$

where $c_0 > 0$ is a constant independent of N, M and δ,

$$\Phi_d := \sum_{\ell=1}^{d} (\gamma\sqrt{|\mathcal{A}|M})^\ell, \tag{33}$$

and $\omega_Z(\delta) := \sup_{\theta,s}\sup_{\|o-o'\|\leq\delta} |Z_\theta(o \mid s)/Z_\theta(o' \mid s) - 1|$. If $d \leq 1$, then $p_\rho = 0$. If $d \geq 2$, $q_{\min} > 0$ and $|\mathcal{A}|M \geq 2$, then $p_2 \leq 2|\mathcal{A}|(|\mathcal{A}|M)^d \exp\big(-N(q_{\min}/|\mathcal{A}|)^{d-1}/32\big)$. When $d = 0$, $\widehat{Q}(b^J,a) = \frac{1}{J}\sum_j Q_\theta^{\star,S}(s_j,a)$ is exact and every term vanishes ($\Lambda_0 = \Phi_0 = 0$). When $\phi = \mathrm{id}$ ($M = |\mathcal{O}|$, $\mathrm{diam}(\phi) = 0$), the discretization error vanishes.

No stationarity assumption on the rewards at internal nodes is required, and the simulation returns need not be independent. The proof uses only that the seeds are i.i.d., that the path of each simulation depends on its own seed alone, and that the estimator has bounded differences on $\mathcal{E}_\rho$. It rests on the following four results; throughout, we fix a root action a and $d \geq 1$.

Lemma 5 (Path decoupling)**.** *There is a deterministic map ν such that, whatever the other seeds are, the path of simulation m (its states, actions, binned observations, and nodes $v_{m,1}, \ldots, v_{m,d}$) equals $\nu(\xi_m)$.*

Proof. We induct on depth. Simulation m starts from s_{j_m} with the root action a, and $u_{m,0}$ generates its next state and observation, so $v_{m,1} = (a, \phi(o))$ is a function of ξ_m. Suppose the state $s_{m,\ell}$ and the node $v_{m,\ell}$ at depth $\ell < d$ are functions of ξ_m. Whether or not an earlier simulation expanded $v_{m,\ell}$, Algorithm 3 only initializes statistics and continues, and it selects the action $a'_{m,\ell}$, a component of ξ_m that does not depend on any tree statistic. The next state and observation are generated from $(s_{m,\ell}, a'_{m,\ell}, u_{m,\ell})$, and $v_{m,\ell+1}$ is the history of $v_{m,\ell}$ extended by $a'_{m,\ell}$ and the binned observation. Hence $s_{m,\ell+1}$ and $v_{m,\ell+1}$ are functions of ξ_m, and the simulation stops at depth d regardless of the other seeds. □

Both hypotheses of the lemma are needed. Under UCB1, the action at $v_{m,\ell}$ depends on the statistics left by simulations $1, \dots, m-1$. Under a POMCP-style early return ("if v is not yet expanded, expand it and return"), simulation m stops at the first node that no earlier simulation created, so its depth depends on $\xi_1, \dots, \xi_{m-1}$.

Corollary 1 (Localized influence)**.** *Let $S \subseteq [N]$ and let P_S be the set of nodes on the paths $\nu(\xi_m)$, $m \in S$. Every node $v \notin P_S$ has the same particle set $B(v)$, counts $N(v)$ and $N(v, a')$, mean rewards $\bar{r}(v, a')$, and DP value $\hat{V}(v)$ in the runs on $(\xi_m)_{m\in[N]}$ and on $(\xi_m)_{m\notin S}$. In particular, replacing one seed ξ_n by ξ'_n changes these quantities only at nodes on the two root-to-leaf paths $\nu(\xi_n)$ and $\nu(\xi'_n)$.*

Proof. By Lemma 5, in either run the simulations through v are those m with $v \in \nu(\xi_m)$ among the seeds present. For $v \notin P_S$ no $m \in S$ is among them, so $B(v)$, $N(v)$, $N(v, a')$ and $\bar{r}(v, a')$ coincide. A descendant of v lies on a path in P_S only if v does, so the whole subtree of v avoids P_S, and backward induction on depth with equation 17 gives the same $\hat{V}(v)$. For the last claim, compare both runs with the run on $(\xi_m)_{m\neq n}$. □

Lemma 6 (Bounded difference)**.** *On $\mathcal{E}_\rho$, for every $S \subseteq [N]$,*

$$\left| f_a(\xi_1, \dots, \xi_N) - f_a^{(-S)} \right| \le |S| \cdot \frac{2(1+\gamma) V_{\max} \Lambda_d}{N}. \tag{34}$$

Consequently, if two seed vectors ξ, ξ' differ in k coordinates and both of their runs lie in $\mathcal{E}_\rho$, then $|f_a(\xi) - f_a(\xi')| \le k\, c$, where

$$c := \frac{4(1+\gamma) V_{\max} \Lambda_d}{N}. \tag{35}$$

Proof. Write $K(v) := N(v)$ in the full run, $k(v)$ for the number of simulations $m \in S$ whose path passes through v, and $\Delta(v) := |\hat{V}(v) - \hat{V}^{(-S)}(v)|$, where $\hat{V}^{(-S)}$ is the DP value in the reduced run. If $k(v) = K(v)$, node v is absent from the reduced run, and we set $\hat{V}^{(-S)}(v) := 0$. By Corollary 1, $\Delta(v) = 0$ unless $k(v) \ge 1$. Define the weighted influence $I(v) := (K(v) - k(v))\, \Delta(v)$ and the sequence

$$y_d := 1, \qquad y_\ell := \rho |\mathcal{A}| \, (1 + \gamma y_{\ell+1}) \quad (1 \le \ell \le d-1).$$

We show by backward induction on depth that every node v at depth $\ell \in [1, d]$ satisfies

$$I(v) \le 2 V_{\max}\, y_\ell\, k(v). \tag{36}$$

Leaves ($\ell = d$). If $k(v) = K(v)$, then $I(v) = 0$. Otherwise, for each a', removing $k(v)$ of the $K(v)$ states in $B(v)$ moves the average of $Q_\theta^{\star,S}(\cdot, a')$ by

$$\frac{k(v)}{K(v)} \left| \text{mean over removed states} - \text{mean over remaining states} \right| \le \frac{2 V_{\max}\, k(v)}{K(v)}.$$

The maximum over a' in equation 15 is 1-Lipschitz in the sup norm, so $\Delta(v) \le 2V_{\max} k(v)/K(v)$ and $I(v) \le 2V_{\max} k(v)$.

Internal nodes ($1 \le \ell \le d-1$). Fix a', and write $K' := N(v, a')$ and k' for the number of simulations in S that select a' at v. If $k' = 0$, then $\hat{Q}(v, a')$ is unchanged, since no simulation of S passes through a child of v under a'. If $1 \le k' < K'$, let Σ and Σ' be the sums of $r(s, a') + \gamma \hat{V}(\text{child})$ over the simulations that select a' at v in the full and in the reduced run. Each removed simulation

contributes a term of absolute value at most $V_{\max}$, because $|r| \le (1-\gamma)V_{\max}$ and $|\hat{V}| \le V_{\max}$. Each of the $K(w) - k(w)$ remaining simulations entering a child w changes by $\gamma\Delta(w)$. Hence

$$|\Sigma - \Sigma'| \le k' V_{\max} + \gamma \sum_{w \in \mathrm{Ch}(v,a')} I(w).$$

Since $\hat{Q}(v,a') = \Sigma/K'$, the reduced-run value is $\Sigma'/(K'-k')$, and $|\Sigma'| \le (K'-k')V_{\max}$,

$$\Big|\frac{\Sigma}{K'} - \frac{\Sigma'}{K'-k'}\Big| \le \frac{|\Sigma - \Sigma'|}{K'} + \frac{k'\,|\Sigma'|}{K'(K'-k')} \le \frac{1}{K'}\Big(2k'V_{\max} + \gamma \sum_{w \in \mathrm{Ch}(v,a')} I(w)\Big).$$

If $k' = K'$, action a' is unvisited in the reduced run and drops out of the maximum in equation 17. A maximum of values in $[-V_{\max}, V_{\max}]$ changes by at most the largest change of its retained arguments, or by at most $2V_{\max} = 2k'V_{\max}/K'$ if its maximizer drops out. Writing $k'_{a'}$ for the k' of action a', in every case

$$\Delta(v) \le \max_{a' \in \mathcal{A}} \frac{1}{N(v,a')}\Big(2k'_{a'}V_{\max} + \gamma \sum_{w \in \mathrm{Ch}(v,a')} I(w)\Big).$$

Multiplying by $K(v)$, using equation 27, and bounding the maximum by the sum,

$$I(v) \le K(v)\,\Delta(v) \le \rho|\mathcal{A}| \sum_{a' \in \mathcal{A}} \Big(2k'_{a'}V_{\max} + \gamma \sum_{w \in \mathrm{Ch}(v,a')} I(w)\Big).$$

By the induction hypothesis equation 36 at depth $\ell+1$ and $\sum_{a'} k'_{a'} = \sum_{a'} \sum_{w \in \mathrm{Ch}(v,a')} k(w) = k(v)$,

$$I(v) \le \rho|\mathcal{A}|\,\big(2V_{\max} + 2\gamma V_{\max}\, y_{\ell+1}\big)\, k(v) = 2V_{\max}\, y_\ell\, k(v).$$

Root. Each simulation enters exactly one depth-1 node, and the normalizer $1/N$ is the same in both runs, so

$$\big|f_a - f_a^{(-S)}\big| \le \frac{1}{N} \sum_{m \in S} \big|r(s_{j_m}, a) + \gamma \hat{V}(v_{m,1})\big| + \frac{\gamma}{N} \sum_{w:\ \text{depth } 1} I(w) \le \frac{|S|}{N}\, V_{\max}\,(1 + 2\gamma y_1).$$

Unrolling the recursion,

$$y_1 = \rho|\mathcal{A}| \sum_{i=0}^{d-2} (\rho\gamma|\mathcal{A}|)^i + (\rho\gamma|\mathcal{A}|)^{d-1}.$$

By equation 28 and $(\rho\gamma|\mathcal{A}|)^{d-1} \le \Lambda_d$,

$$1 + 2\gamma y_1 = 2\Lambda_d - 1 + 2\gamma(\rho\gamma|\mathcal{A}|)^{d-1} \le 2(1+\gamma)\Lambda_d.$$

This proves equation 34. For the second claim, let S be the set of coordinates where ξ and ξ' differ. Both runs have the same reduced run on $(\xi_m)_{m \notin S}$, so equation 34 applied to each and the triangle inequality give $|f_a(\xi) - f_a(\xi')| \le 2k \cdot 2(1+\gamma)V_{\max}\Lambda_d/N = kc$. □

Lemma 7 (Fluctuation up to a bad event)**.** *Let $p := \Pr(\mathcal{E}_\rho^c \mid b_t^J)$ for the tree T_a, and let $B_d := 2V_{\max} + cN$ with c from equation 35. For every $\delta \in (0,1)$, with probability at least $1 - \delta - p$,*

$$\big|f_a - \mathbb{E}f_a\big| \le 2(1+\gamma)V_{\max}\Lambda_d\sqrt{\frac{2\log(2/\delta)}{N}} + B_d\, p,$$

and for every $q \in [-V_{\max}, V_{\max}]$,

$$\mathbb{E}\big[(f_a - q)^2\big] \le \frac{8(1+\gamma)^2 V_{\max}^2 \Lambda_d^2}{N} + 2\,(\mathbb{E}f_a - q)^2 + 3B_d^2\, p.$$

Proof. Let G be the set of seed vectors whose run lies in $\mathcal{E}_\rho$. If $\Pr(G) = 0$, both claims are trivial, so assume otherwise. Let H denote the Hamming distance and define the extension

$$g(\xi) := \inf_{\eta \in G} \big\{ f_a(\eta) + c\, H(\xi, \eta) \big\}.$$

First, $g = f_a$ on G: for $\xi \in G$, Lemma 6 gives $f_a(\eta) + cH(\xi,\eta) \geq f_a(\xi)$ for every $\eta \in G$, with equality at $\eta = \xi$. Second, g changes by at most c when one coordinate of ξ changes, because each $H(\cdot,\eta)$ does. Third, $-V_{\max} \leq g \leq V_{\max} + cN$, so $|f_a - g| \leq B_d$ everywhere and

$$|\mathbb{E}f_a - \mathbb{E}g| \leq \mathbb{E}\big[|f_a - g|\,\mathbf{1}_{G^c}\big] \leq B_d\,p.$$

The seeds are independent, so McDiarmid's bounded-differences inequality applies to g:

$$\Pr\big(|g - \mathbb{E}g| \geq s\big) \leq 2\exp\Big(-\frac{2s^2}{Nc^2}\Big).$$

Setting the right-hand side to δ gives $s = c\sqrt{N\log(2/\delta)/2} = 2(1+\gamma)V_{\max}\Lambda_d\sqrt{2\log(2/\delta)/N}$. On $G \cap \{|g - \mathbb{E}g| < s\}$, an event of probability at least $1 - \delta - p$,

$$|f_a - \mathbb{E}f_a| \leq |g - \mathbb{E}g| + |\mathbb{E}g - \mathbb{E}f_a| < s + B_d\,p.$$

For the second moment, the Efron–Stein inequality gives $\mathrm{Var}(g) \leq Nc^2/2 = 8(1+\gamma)^2V_{\max}^2\Lambda_d^2/N$. Since $f_a = g$ on G and $|f_a - q| \leq 2V_{\max}$,

$$\mathbb{E}\big[(f_a - q)^2\big] \leq \mathbb{E}\big[(g - q)^2\big] + 4V_{\max}^2\,p = \mathrm{Var}(g) + (\mathbb{E}g - q)^2 + 4V_{\max}^2\,p.$$

Finally, $(\mathbb{E}g - q)^2 \leq 2(\mathbb{E}f_a - q)^2 + 2B_d^2p^2$, and $2B_d^2p^2 + 4V_{\max}^2p \leq 3B_d^2p$. The infimum defining g is universally measurable, which suffices for both inequalities. □

Lemma 8 (Balancedness under uniform selection)**.** *If $d \leq 1$, then $\mathcal{E}_\rho$ holds surely. If $d \geq 2$, $q_{\min} > 0$ and $|\mathcal{A}|M \geq 2$, then*

$$\Pr(\mathcal{E}_2^c \mid b_t^J) \leq 2|\mathcal{A}|\,(|\mathcal{A}|M)^d\exp\Big(-\frac{N}{32}\Big(\frac{q_{\min}}{|\mathcal{A}|}\Big)^{d-1}\Big).$$

Proof. For $d \leq 1$ the tree has no internal node. Let $d \geq 2$ and fix a potential internal node v of T_a at depth $\ell \in [1, d-1]$, i.e., a history with ℓ binned observations and $\ell - 1$ internal actions. By Lemma 5, the indicators that simulation m reaches v, and that it reaches v and selects a' there, are functions of ξ_m. They are therefore i.i.d. Bernoulli across m, with parameters p_v and $p_v/|\mathcal{A}|$. Each binned observation along the history has conditional probability at least $q_{\min}$ and each internal action has probability $1/|\mathcal{A}|$, so

$$\frac{p_v}{|\mathcal{A}|} \geq \Big(\frac{q_{\min}}{|\mathcal{A}|}\Big)^{\ell} \geq \pi := \Big(\frac{q_{\min}}{|\mathcal{A}|}\Big)^{d-1}.$$

The multiplicative Chernoff bounds give

$$\Pr\big(N(v) \geq \tfrac{3}{2}Np_v\big) \leq e^{-Np_v/12} \leq e^{-N\pi/32},$$
$$\Pr\big(N(v,a') \leq \tfrac{3}{4}Np_v/|\mathcal{A}|\big) \leq e^{-Np_v/(32|\mathcal{A}|)} \leq e^{-N\pi/32}.$$

Outside these events, $N(v) < \frac{3}{2}Np_v = 2|\mathcal{A}| \cdot \frac{3}{4}Np_v/|\mathcal{A}| < 2|\mathcal{A}|\,N(v,a')$, which is equation 27 with $\rho = 2$. A union bound over the $|\mathcal{A}|$ actions and over the at most $\sum_{\ell=1}^{d-1} M^\ell|\mathcal{A}|^{\ell-1} \leq (|\mathcal{A}|M)^d$ potential internal nodes gives the claim, since $1 + |\mathcal{A}| \leq 2|\mathcal{A}|$. □

Proof of Proposition 3. When $d = 0$ the planner is exact, so let $d \geq 1$ and write $q_a := Q_{\theta,d}(b^J, a) \in [-V_{\max}, V_{\max}]$.

Step 1: Fluctuation. Apply Lemma 7 to each root action with $\delta/|\mathcal{A}|$ in place of δ, and take a union bound over $a \in \mathcal{A}$. With probability at least $1 - \delta - |\mathcal{A}|p_\rho$, every a satisfies

$$\begin{aligned}|\widehat{Q}(b^J,a) - q_a| &\leq |\widehat{Q}(b^J,a) - \mathbb{E}\widehat{Q}(b^J,a)| + |\mathbb{E}\widehat{Q}(b^J,a) - q_a| \\ &\leq 2(1+\gamma)V_{\max}\Lambda_d\sqrt{\frac{2\log(2|\mathcal{A}|/\delta)}{N}} + B_d\,p_\rho + \mathrm{bias}(N,d),\end{aligned}$$

and taking the supremum over a gives equation 30. For the second moment, Lemma 7 with $q = q_a$ gives

$$\mathbb{E}\big[(\widehat{Q}(b^J,a) - q_a)^2\big] \leq \frac{8(1+\gamma)^2V_{\max}^2\Lambda_d^2}{N} + 2\,\mathrm{bias}(N,d)^2 + 3B_d^2\,p_\rho.$$

Summing over $a \in \mathcal{A}$ through $\epsilon_{\mathrm{mc}}(t,d)^2 \le \sum_a (\widehat{Q}(b^J,a) - q_a)^2$ gives equation 31. The bounds on p_ρ are Lemma 8.

Step 2: DP backup estimation error. The simulations follow the uniform policy at internal nodes, and the DP backup $\hat{V}(v) = \max_{a'} \hat{Q}(v,a')$ recovers the optimal internal actions. We bound the bias $\mathrm{bias}_{\mathrm{DP}}(N,d) := \sup_a |\mathbb{E}\widehat{Q}(b^J,a) - Q^{\phi}_{\theta,d}(b^J,a)|$ of the resulting estimate by backward induction on depth, at the level of orders of magnitude, without any independence assumption on the returns. Fix a node v at depth $\ell \in [1,d]$ with $n_v := N(v)$ visits, and condition on the set of simulations that reach v and on their states there. By Lemma 5, the continuation of each such simulation below v is driven by the components of its own seed at depths $\ge \ell$, which are independent across simulations and independent of the conditioning. Conditionally, the states in $B(v)$ are therefore independent draws from the binned belief b_v at v, and each $\hat{Q}(v,a')$ is formed from $n_{v,a'} := N(v,a')$ independent continuations, with $n_{v,a'} \ge n_v/(\rho|\mathcal{A}|)$ on $\mathcal{E}_\rho$. Concentration of these conditional averages gives

$$\left|\hat{Q}(v,a') - Q^{\phi}_{\theta,d-\ell}(b_v,a')\right| \;\le\; O\!\left(\frac{V_{\max}}{\sqrt{n_{v,a'}}}\right) \;\le\; O\!\left(\frac{V_{\max}\sqrt{|\mathcal{A}|}}{\sqrt{n_v}}\right).$$

At a leaf ($\ell = d$) the same bound holds for the pooled averages $\frac{1}{|B(v)|}\sum_{s\in B(v)} Q^{\star,S}_{\theta}(s,a')$ around $\int Q^{\star,S}_{\theta}(s,a')\, b_v(\mathrm{d}s)$. This level must be included, because the maximum over a' in equation 15 is biased upward at every finite $|B(v)|$. Under the branching structure, a depth-ℓ node receives $n_v = \Omega(N/(|\mathcal{A}|M)^\ell)$ visits in expectation when the bins have comparable probabilities. Taking the maximum over estimated Q-values, the backup error at depth ℓ contributes at most $O(V_{\max}\sqrt{|\mathcal{A}|^{\ell+1}M^\ell/N})$, discounted by γ^ℓ. The complement of $\mathcal{E}_\rho$ changes the conditional mean by at most $2V_{\max}p_\rho$, which is absorbed into the constant once $p_\rho \le N^{-1/2}$ (Remark 3). Summing over depths, leaves included,

$$\mathrm{bias}_{\mathrm{DP}}(N,d) \;\le\; \sum_{\ell=1}^{d} \gamma^\ell \cdot \frac{c_1 V_{\max}\sqrt{|\mathcal{A}|}}{\sqrt{N/(|\mathcal{A}|M)^\ell}} \;=\; \frac{c_0\,\Phi_d\,V_{\max}\sqrt{|\mathcal{A}|}}{\sqrt{N}}. \tag{37}$$

Step 3: Discretization error. For o, o' in the same bin, the Bayes update sensitivity [10, Section 4.3] gives

$$\|\tau_\theta(b,a,o) - \tau_\theta(b,a,o')\|_{\mathrm{TV}} \;\le\; \frac{g_{\max}}{g_{\min}}\,\omega_Z(\mathrm{diam}(\phi)).$$

Since τ^{ϕ}_θ is a convex combination of $\{\tau_\theta(b,a,o')\}_{o'\in\phi^{-1}(\bar{o})}$, the same bound holds for $\|\tau^{\phi}_\theta(b,a,\bar{o}) - \tau_\theta(b,a,o)\|_{\mathrm{TV}}$. By Lipschitz continuity of $\bar{V}^{\star,S}_{\theta,d-1}$ in TV (constant $2V_{\max}$, cf. Lemma 2), each of the d tree levels contributes $\frac{2g_{\max}}{g_{\min}}V_{\max}\,\omega_Z(\mathrm{diam}(\phi))$ discounted by γ:

$$\sup_{b,a}\left|Q^{\phi}_{\theta,d}(b,a) - Q_{\theta,d}(b,a)\right| \;\le\; \frac{2\gamma(1-\gamma^d)}{1-\gamma}\cdot\frac{g_{\max}}{g_{\min}}\,V_{\max}\,\omega_Z(\mathrm{diam}(\phi)). \tag{38}$$

Combining Steps 2 and 3 gives $\mathrm{bias}(N,d) \le \mathrm{bias}_{\mathrm{DP}}(N,d) + \mathrm{bias}_{\mathrm{disc}}(d)$, yielding equation 32. □

Remark 3 (Constants, a verification-free variant, and the choice of depth)**.** (a) Order of the constants. *Compared with an analysis that treats the simulation returns as i.i.d., the only changes are the depth factor $(1+\gamma)\Lambda_d$ in the fluctuation term and the exponentially small remainder in p_ρ. Since $\Lambda_d \le d\max(1,\rho\gamma|\mathcal{A}|)^{d-1}$ and $\Phi_d \ge (\gamma\sqrt{|\mathcal{A}|M})^d$, the fluctuation term of equation 31 is at most of the same exponential-in-d order as the DP-backup term whenever $M = \Omega(\rho^2|\mathcal{A}|)$. Taking square roots in equation 31, with $\mathrm{diam}(\phi)$ chosen as in Remark 4, then gives $\mathbb{E}[\epsilon_{\mathrm{mc}}(t,d) \mid b^J_t] \le C_{\mathrm{mc}}V_{\max}/\sqrt{N}$ with C_{mc} of the order stated in Theorem 1, whose order is therefore unchanged. For coarse binning, $M \ll |\mathcal{A}|$, the fluctuation dominates and $C_{\mathrm{mc}} \propto \sqrt{|\mathcal{A}|}\,\Lambda_d$. By Lemma 8, the remainder $3|\mathcal{A}|B_d^2p_2$ is at most $V^2_{\max}/N$ as soon as $N \ge 32(|\mathcal{A}|/q_{\min})^{d-1}\log\!\big(24N|\mathcal{A}|^2(|\mathcal{A}|M)^d(1+4\Lambda_d)^2\big)$, and it vanishes for $d \le 1$. At $d = 0$, $\Lambda_0 = 0$ and the fluctuation vanishes identically. For an internal rule satisfying (S1)–(S2) in place of uniform selection, the same proof applies with $|\mathcal{A}|$ replaced by $1/\pi_{\min}$ in equation 27, equation 28 and Lemma 8.*

(b) A verification-free variant. *Sample splitting restores exactly conditionally i.i.d. returns. Use $N/2$ seeds to build the tree, populate the particle sets, and run the DP backup, then freeze the leaf values*

and the greedy internal actions $\arg\max_{a'} \hat{Q}(v, a')$*. Value each root action with* $N/2$ *fresh seeds that follow the frozen greedy actions and stop at the frozen leaf values; a simulation that leaves the frozen tree stops and is evaluated by* $\max_{a'} Q_\theta^{\star,S}(s, a')$ *at its current state* s*. Conditionally on the first half, the second-half returns are i.i.d. and bounded by* $V_{\max}$*, so Hoeffding's inequality applies to them directly, at the cost of a factor* 2 *in* N*, and the bias analysis of Step 2 applies to the first half. Evaluating each leaf at its single incoming state instead of equation 15 removes the pooling but not the coupling through the maximum in the DP backup, so it restores independent returns only at* $d = 1$*, where it also changes the leaf target to the linear surrogate* $\widetilde{V}_\theta^\star$*.*

(c) Choice of depth. *Propositions 2 and 3 concern different objects. The former bounds the planning target,* $\|\bar{V}_{\theta,d}^{\star,S} - V_\theta^{\star,B}\|_\infty \le \gamma^{d+1}\Xi_\theta/(1-\gamma)$*, which involves no particles and decays geometrically in* d*. The latter bounds the estimator* $\widehat{Q}$ *around* $Q_{\theta,d}$*, and its fluctuation* $2(1+\gamma)V_{\max}\Lambda_d\sqrt{2\log(2|\mathcal{A}|/\delta)/N}$ *and DP-backup bias* $c_0\Phi_d V_{\max}\sqrt{|\mathcal{A}|/N}$ *grow geometrically in* d*. Because increasing the depth reduces the structural bias but increases the Monte Carlo error, their sum is U-shaped in* d*, and each budget* N *has a finite optimal depth.*

Remark 4 (Optimal discretization resolution)**.** *The discretization error in equation 32 is a fixed bias* $O(\omega_Z(\mathrm{diam}(\phi)))$*, while the DP backup estimation error vanishes as* $N \to \infty$ *at rate* $\Phi_d/\sqrt{N}$*. To make the discretization bias* $O(1/\sqrt{N})$*, set* $\mathrm{diam}(\phi) = \omega_Z^{-1}\big((1-\gamma)g_{\min}/(2\gamma g_{\max}V_{\max}\sqrt{N})\big)$*. When* $Z_\theta(o \mid s)$ *is* L_Z*-Lipschitz in* o *uniformly (so* $\omega_Z(\delta) \le L_Z\delta/g_{\min}$*) and* $\mathcal{O} \subseteq \mathbb{R}^{d_o}$ *with a uniform grid:*

$$M = \left\lceil \left(\frac{2\gamma\, L_Z\, g_{\max}\, V_{\max}\, \mathrm{diam}(\mathcal{O})\sqrt{N}}{(1-\gamma)\, g_{\min}^2} \right)^{d_o} \right\rceil. \tag{39}$$

With this choice, M *grows as* $N^{d_o/2}$*, and the simulation budget threshold for the DP backup error to be* $O(1/\sqrt{N})$ *is* $N \ge c_0^2|\mathcal{A}|\Phi_d^2$*, which is exponential in* $d \cdot d_o$ *via the dependence of* Φ_d *on* M*.*

Balancing the two bias terms instead gives an explicit rate. Take $\mathcal{O} \subseteq [0, D]^{d_o}$ *with a cubic grid of mesh* h*, so that* $M = (D/h)^{d_o}$ *and* $\mathrm{diam}(\phi) = \sqrt{d_o}\, h$*, and assume* $\omega_Z(\delta) \le L_Z\delta/g_{\min}$ *as above. For* $\gamma\sqrt{|\mathcal{A}|M} \ge 2$ *we have* $\Phi_d \le 2(\gamma\sqrt{|\mathcal{A}|M})^d$*, so the DP-backup and discretization terms of equation 32 are of order* $h^{-d\,d_o/2}N^{-1/2}$ *and* h*, respectively, with constants depending on* d*,* γ*,* $|\mathcal{A}|$*,* D*,* L_Z*,* $g_{\max}/g_{\min}$ *and* $V_{\max}$*. Balancing them gives*

$$h^\star \asymp N^{-1/(d\,d_o+2)}, \qquad \epsilon_{\mathrm{mc}}(t, d) = \widetilde{O}\big(N^{-1/(d\,d_o+2)}\big), \tag{40}$$

since the fluctuation term of equation 30 does not depend on h *and is of lower order. The* $N^{-1/2}$ *rate is recovered exactly when* $d\,d_o = 0$*, i.e., at* $d = 0$*. At* $d = 1$*, equation 40 is the familiar rate* $N^{-1/(d_o+2)}$ *of a histogram estimate, here of the leaf value as a function of the observation. The balancedness condition of Lemma 8 holds at this mesh for large* N*: with the Lebesgue reference measure* $q_{\min} \ge g_{\min}h^{d_o}$*, so* $N(q_{\min}/|\mathcal{A}|)^{d-1} = \Omega(N^{(d_o+2)/(d\,d_o+2)}) \to \infty$*. The computational cost is* $N|\mathcal{A}|$ *simulations over at most* $(|\mathcal{A}|M)^d$ *nodes.*

When $d = 0$ *or* $|\mathcal{O}| < \infty$*, this issue does not arise: set* $\phi = \mathrm{id}$ *and* $M = |\mathcal{O}|$*.*

D.3 Particle Filter Stability

We work under Assumption 1 throughout. Write $b_t := b_t^{\theta^\star}$ for the true posterior and $b_t^J := b_t^{J,(\bar{\theta}_{k(s)})_{s\le t}}$ for the particle belief filtered under the epoch-frozen model sequence. Since the model sequence is determined by the algorithm, we abbreviate the latter as b_t^J.

Proposition 4 (Particle filter stability in d_Q)**.** *Under Assumption 1, there exist constants* $C_{\mathrm{pf}}, C_{\mathrm{filt}} < \infty$ *and a forgetting rate* $\rho := 1 - \varepsilon_{\mathrm{mix}}\,(g_{\min}/g_{\max}) \in (0, 1)$ *such that for all* $t \ge 1$*,*

$$\mathbb{E}\big[d_Q(b_t, b_t^J)\big] \;\le\; \underbrace{C_{\mathrm{pf}} \cdot J^{-\frac{1}{2\kappa_{\mathrm{pf}}}}}_{\textit{particle approx.}} \;+\; \underbrace{2V_{\max}\, C_{\mathrm{filt}} \sum_{s=1}^{t} \rho^{t-s}\, d_{\mathrm{model}}^{a_s}(\bar{\theta}_s, \theta^\star)}_{\textit{filter mismatch}}, \tag{41}$$

where $d^a_{\mathrm{model}}(\theta,\theta') := \sup_{s'} \|P_\theta(\cdot \mid s',a) - P_{\theta'}(\cdot \mid s',a)\|_{\mathrm{TV}} + \sup_{s',o} |Z_\theta(o \mid s')/Z_{\theta'}(o \mid s') - 1|$, $C_{\mathrm{filt}} := g_{\max}/g_{\min}$, *and* κ_{pf} *is the* particle filter exponent*:*

$$\kappa_{\mathrm{pf}} := \max\left(1,\ \frac{\log(2g^2_{\max}/g^2_{\min})}{\log(1/\gamma)}\right). \tag{42}$$

The truncation depth H_{pf} *in Lemma 1 is chosen to balance the sum and tail:*

$$H_{\mathrm{pf}} := \left\lceil \frac{\log(4V_{\max}\sqrt{J})}{\kappa_{\mathrm{pf}}\log(1/\gamma)} \right\rceil = \left\lceil \frac{\log(4V_{\max}\sqrt{J})}{\log\max(1/\gamma,\ 2g^2_{\max}/g^2_{\min})} \right\rceil, \tag{43}$$

and C_{pf} *depends polynomially on* $V_{\max}$, $g_{\max}/g_{\min}$, $1/\varepsilon_{\mathrm{mix}}$, $\log \mathcal{N}_\Theta(1/T)$, *and* H_{pf}. *When* $\kappa_{\mathrm{pf}} = 1$ *(i.e.,* $2\gamma g^2_{\max}/g^2_{\min} \le 1$*), the rate recovers* $J^{-1/2}$. *This* H_{pf} *is distinct from the trajectory truncation horizon* H *in Section D.4.*

Proof of Proposition 4. By the triangle inequality and Lemma 2:

$$d_Q(b_t^{\theta^\star},\, b_t^{J,\bar\theta_{1:t}}) \ \le\ \underbrace{d_Q(b_t^{\theta^\star},\, b_t^{\bar\theta_{1:t}})}_{\text{(I) filter mismatch}} + \underbrace{d_Q(b_t^{\bar\theta_{1:t}},\, b_t^{J,\bar\theta_{1:t}})}_{\text{(II) particle approx.}}. \tag{44}$$

Term (I) involves two exact (continuous) posteriors, so the TV bound of Lemma 2 gives $d_Q \le 2V_{\max}\|b_t^{\theta^\star} - b_t^{\bar\theta_{1:t}}\|_{\mathrm{TV}}$, which is bounded by the filter mismatch analysis below.

Term (II): test-function bound via Feynman-Kac. Conditional on $(a_{1:T}, o_{1:T})$, the particle filter is a Feynman-Kac system with kernels $M_s(\cdot \mid s') := P_{\bar\theta_s}(\cdot \mid s', a_s)$ and potentials $G_s(s') := Z_{\bar\theta_s}(o_s \mid s')$. Under Assumption 1, the kernels satisfy $M_s(s' \mid x) \ge \varepsilon_{\mathrm{mix}}\,\nu(s')$ and the potentials satisfy $g_{\min} \le G_s(x) \le g_{\max}$ uniformly. We invoke the following time-uniform particle-filter bound.

Lemma 9 ([21, Lemma 2.3]; see also [9, Theorem 7.4.4])**.** *Let* $(\eta_n)_{n\ge0}$ *and* $(\eta_n^N)_{n\ge0}$ *be the exact and* N*-particle Feynman-Kac measures on a measurable space* E. *Suppose the Markov kernels* $(M_\ell)_{\ell\ge1}$ *admit densities w.r.t. a common measure* λ *satisfying* $\underline{M} \le M_\ell(x,y) \le \overline{M}$ *for all* ℓ, x, y, *and the potentials satisfy* $\underline{G} \le G_\ell(x) \le \overline{G}$ *for all* ℓ, x. *Then for all* $n \ge 0$, $N \ge 2$, $p \ge 1$, *and* φ *with* $\mathrm{osc}(\varphi) \le 1$*:*

$$\left(\mathbb{E}\left[|\eta_n^N(\varphi) - \eta_n(\varphi)|^p\right]\right)^{1/p} \ \le\ \frac{2\,d(p)^{1/p}}{\sqrt{N}}\,\frac{(\overline{M}/\underline{M})^3\,\overline{G}}{\underline{G}}, \tag{45}$$

where $d(p)^{1/p} = O(\sqrt{p})$. *The bound is uniform over* n.

Our Assumption 1 implies the conditions of Lemma 9 with $\underline{M} = \varepsilon_{\mathrm{mix}}$, $\overline{M} = 1$, $\underline{G} = g_{\min}$, $\overline{G} = g_{\max}$.

Proposition 5 (Particle filter L^p bound under Assumption 1)**.** *Under Assumption 1, for every measurable* $f : \mathcal{S} \to \mathbb{R}$ *with* $\|f\|_\infty \le M$, *all* $p \ge 1$, *and all* $t \ge 0$*:*

$$\left(\mathbb{E}\left[|\eta_t^J(f) - \eta_t(f)|^p\right]\right)^{1/p} \ \le\ \frac{4\,d(p)^{1/p}\,M\,g_{\max}}{\varepsilon^3_{\mathrm{mix}}\,g_{\min}\,\sqrt{J}}, \tag{46}$$

where $d(p)^{1/p} = O(\sqrt{p})$, *so* $\eta_t^J(f) - \eta_t(f)$ *is sub-Gaussian with parameter* $\sigma^2 = O(M^2 g^2_{\max}/(\varepsilon^6_{\mathrm{mix}} g^2_{\min} J))$. *The bound is uniform over* t.

Proof. By Lemma 9 with $\underline{M} = \varepsilon_{\mathrm{mix}}$, $\overline{M} = 1$, $\underline{G} = g_{\min}$, $\overline{G} = g_{\max}$, for any φ with $\mathrm{osc}(\varphi) \le 1$:

$$\left(\mathbb{E}\left[|\eta_t^J(\varphi) - \eta_t(\varphi)|^p\right]\right)^{1/p} \ \le\ \frac{2\,d(p)^{1/p}\,g_{\max}}{\varepsilon^3_{\mathrm{mix}}\,g_{\min}\,\sqrt{J}},$$

uniformly over t. For f with $\|f\|_\infty \le M$, applying this to $\varphi := f/(2M)$ (which has $\mathrm{osc}(\varphi) \le 1$) and rescaling by $2M$ yields equation 46. Sub-Gaussianity follows from the standard moment-to-tail conversion. □

Lemma 10 (Unconditional bound under adaptive actions)**.** *Let $(a_s, o_s)_{s\geq 1}$ be an adapted action-observation process where $a_s = a_s(b_s^J)$ may depend on the particle system. Under Assumption 1, for every measurable $f : \mathcal{S} \to \mathbb{R}$ with $\|f\|_\infty \leq M$, all $p \geq 1$, and all $t \geq 0$:*

$$\big(\mathbb{E}\big[|b_t^{J,\bar{\theta}_{1:t}}(f) - b_t^{\bar{\theta}_{1:t}}(f)|^p\big]\big)^{1/p} \;\leq\; \frac{4\,d(p)^{1/p}\,M\,g_{\max}}{\varepsilon_{\mathrm{mix}}^3\,g_{\min}\,\sqrt{J}}. \tag{47}$$

Proof. Let $\mathcal{H}_T := (a_{1:T}, o_{1:T}, \bar{\theta}_{1:T})$. Conditional on $\mathcal{H}_T$, the Feynman-Kac kernels and potentials are deterministic, so Proposition 5 gives

$$\mathbb{E}\big[|b_t^J(f) - b_t(f)|^p \;\big|\; \mathcal{H}_T\big] \;\leq\; \left(\frac{4\,d(p)^{1/p}\,M\,g_{\max}}{\varepsilon_{\mathrm{mix}}^3\,g_{\min}\,\sqrt{J}}\right)^p =: C_p^p.$$

Since C_p is independent of $\mathcal{H}_T$, the tower property yields $\mathbb{E}[|b_t^J(f) - b_t(f)|^p] \leq C_p^p$. □

By Lemma 1 with truncation depth H_{pf}:

$$d_Q(b_t^{\bar{\theta}_{1:t}}, b_t^{J,\bar{\theta}_{1:t}}) \;\leq\; \sup_{\theta,a} \sum_{\ell=0}^{H_{\mathrm{pf}}-1} \gamma^\ell C_{\mathrm{BU}}^\ell \sup_{f\in\mathcal{F}_\ell(\theta,a)} |b_t(f) - b_t^J(f)| + 2\gamma^{H_{\mathrm{pf}}} V_{\max}. \tag{48}$$

We optimize H_{pf} to balance the sum and tail below.

Pointwise bound. For $f \in \mathcal{F}_\ell(\theta, a)$ with $\|f\|_\infty \leq g_{\max}^\ell V_{\max}$, Lemma 10 gives

$$\big(\mathbb{E}\big[|b_t^J(f) - b_t(f)|^p\big]\big)^{1/p} \;\leq\; \frac{4\,d(p)^{1/p}\,g_{\max}^{\ell+1}\,V_{\max}}{\varepsilon_{\mathrm{mix}}^3\,g_{\min}\,\sqrt{J}} =: \sigma_\ell\,d(p)^{1/p}. \tag{49}$$

Since $d(p)^{1/p} \leq C_0\sqrt{p}$ (Corollary 5), $b_t^J(f) - b_t(f)$ is sub-Gaussian with parameter σ_ℓ^2.

Covering-number argument. For N sub-Gaussian random variables with common parameter σ^2:

$$\mathbb{E}\big[\max_{i\leq N} |X_i|\big] \;\leq\; \sigma\sqrt{2\log(2N)}. \tag{50}$$

At depth ℓ, $\bigcup_{\theta,a} \mathcal{F}_\ell(\theta, a)$ is covered by $N_\ell := \mathcal{N}_\Theta(1/T) \cdot |\mathcal{A}|^\ell$ functions after discretizing Θ at resolution $1/T$, with approximation error $O(1/T)$. Applying equation 50 and summing over depths with weights $\gamma^\ell C_{\mathrm{BU}}^\ell$:

$$\mathbb{E}\big[d_Q(b_t^{\bar{\theta}_{1:t}}, b_t^{J,\bar{\theta}_{1:t}})\big] \leq \sum_{\ell=0}^{H_{\mathrm{pf}}-1} \gamma^\ell C_{\mathrm{BU}}^\ell\,\sigma_\ell\sqrt{2\log(2N_\ell)} + 2\gamma^{H_{\mathrm{pf}}} V_{\max} + O(1/T) \tag{51}$$

Define $\alpha := 2\gamma g_{\max}^2/g_{\min}^2$, $r := \alpha/\gamma = 2g_{\max}^2/g_{\min}^2$, and $c_1 := 4g_{\max}V_{\max}/(\varepsilon_{\mathrm{mix}}^3 g_{\min})$, so that $\sigma_\ell = c_1\,g_{\max}^\ell/\sqrt{J}$ and the ℓ-th term equals

$$\frac{c_1}{\sqrt{J}}\,\alpha^\ell\,\sqrt{2\log\big(2\mathcal{N}_\Theta(1/T)\,|\mathcal{A}|^\ell\big)}.$$

Since $\sqrt{\log(ab)} \leq \sqrt{\log a} + \sqrt{\log b}$, the sum in equation 51 is bounded by

$$\mathbb{E}\big[d_Q^{(\mathrm{II})}\big] \;\leq\; \frac{c_1\sqrt{2}}{\sqrt{J}}\left(\sqrt{\log(2\mathcal{N}_\Theta(1/T))}\sum_{\ell=0}^{H_{\mathrm{pf}}-1}\alpha^\ell + \sqrt{\log|\mathcal{A}|}\sum_{\ell=0}^{H_{\mathrm{pf}}-1}\alpha^\ell\sqrt{\ell}\right) + 2\gamma^{H_{\mathrm{pf}}}V_{\max}. \tag{52}$$

We bound the two sums in two cases.

Case 1: $\alpha \geq 1$. Choose

$$H_{\mathrm{pf}} = \left\lceil \frac{\log(4V_{\max}\sqrt{J})}{\log r} \right\rceil = \left\lceil \frac{\log(4V_{\max}\sqrt{J})}{\log(2g_{\max}^2/g_{\min}^2)} \right\rceil.$$

Using $\sum_{\ell=0}^{H-1}\alpha^\ell \le H\alpha^{H-1}$ and $\sum_{\ell=0}^{H-1}\alpha^\ell\sqrt{\ell} \le H^{3/2}\alpha^{H-1}$ (since $\alpha \ge 1$):

$$\text{sum} \;\le\; \frac{c_1\sqrt{2}\,H_{\mathrm{pf}}^{3/2}\,\alpha^{H_{\mathrm{pf}}-1}}{\sqrt{J}} \cdot \sqrt{\log\bigl(2\mathcal{N}_\Theta(1/T)\,|\mathcal{A}|^{H_{\mathrm{pf}}}\bigr)}. \tag{53}$$

Writing $\alpha^{H_{\mathrm{pf}}-1} = \alpha^{H_{\mathrm{pf}}}/\alpha = \gamma^{H_{\mathrm{pf}}} r^{H_{\mathrm{pf}}}/\alpha$ and substituting $r^{H_{\mathrm{pf}}} = 4V_{\max}\sqrt{J}$:

$$\text{sum} \;\le\; \frac{4\sqrt{2}\,c_1\,V_{\max}\,H_{\mathrm{pf}}^{3/2}}{\alpha} \cdot \sqrt{\log\bigl(2\mathcal{N}_\Theta(1/T)\,|\mathcal{A}|^{H_{\mathrm{pf}}}\bigr)} \cdot \gamma^{H_{\mathrm{pf}}}. \tag{54}$$

For the tail:

$$\gamma^{H_{\mathrm{pf}}} = \gamma^{\log(4V_{\max}\sqrt{J})/\log r} = (4V_{\max}\sqrt{J})^{-\log(1/\gamma)/\log r} = (4V_{\max})^{-1/\kappa_{\mathrm{pf}}} \cdot J^{-\frac{1}{2\kappa_{\mathrm{pf}}}}, \tag{55}$$

where $\kappa_{\mathrm{pf}} = \log r/\log(1/\gamma) > 1$. Substituting $c_1 = 4g_{\max}V_{\max}/(\varepsilon_{\mathrm{mix}}^3 g_{\min})$, $H_{\mathrm{pf}} = \lceil \log(4V_{\max}\sqrt{J})/\log r \rceil$, and using $\log(2\mathcal{N}_\Theta(1/T)\,|\mathcal{A}|^{H_{\mathrm{pf}}}) \le \log(2\mathcal{N}_\Theta(1/T)) + H_{\mathrm{pf}}\log|\mathcal{A}|$, the total (sum + tail) is

$$\mathbb{E}\bigl[d_Q^{(\mathrm{II})}\bigr] \;\le\; \left(\frac{16\sqrt{2}\,g_{\max}\,V_{\max}^2}{\alpha\,\varepsilon_{\mathrm{mix}}^3\,g_{\min}} \cdot \left(\frac{\log(4V_{\max}\sqrt{J})}{\log r}\right)^{3/2} \sqrt{\log\bigl(2\mathcal{N}_\Theta(1/T)\bigr) + \frac{\log(4V_{\max}\sqrt{J})}{\log r}\log|\mathcal{A}|} + 2V_{\max}\right) \cdot (4V_{\max})^{-1/\kappa_{\mathrm{pf}}} \cdot J^{-\frac{1}{2\kappa_{\mathrm{pf}}}}. \tag{56}$$

Case 2: $\alpha < 1$. Choose

$$H_{\mathrm{pf}} = \left\lceil \frac{\log(4V_{\max}\sqrt{J})}{\log(1/\gamma)} \right\rceil.$$

The geometric series converges: $\sum_{\ell\ge0}\alpha^\ell = 1/(1-\alpha)$ and $\sum_{\ell\ge0}\alpha^\ell\sqrt{\ell} \le c_2/(1-\alpha)^{3/2}$ for a universal $c_2 > 0$. Hence

$$\text{sum} \;\le\; \frac{c_1\,c_2\,\sqrt{2}}{(1-\alpha)^{3/2}\,\sqrt{J}} \cdot \sqrt{\log\bigl(2\mathcal{N}_\Theta(1/T)\bigr) + \log|\mathcal{A}|} \;=\; O\Bigl(J^{-1/2}\Bigr).$$

For the tail: $\gamma^{H_{\mathrm{pf}}} = (4V_{\max}\sqrt{J})^{-1} = O(J^{-1/2})$. Here $\kappa_{\mathrm{pf}} = 1$, and both terms are $O(J^{-1/2})$.

Unified bound. The choice equation 43, $H_{\mathrm{pf}} = \lceil \log(4V_{\max}\sqrt{J})/\log\max(1/\gamma, r)\rceil$, covers both cases and yields

$$\mathbb{E}\bigl[d_Q^{(\mathrm{II})}\bigr] \;\le\; C_{\mathrm{pf}} \cdot J^{-\frac{1}{2\kappa_{\mathrm{pf}}}}, \qquad \kappa_{\mathrm{pf}} = \max\left(1, \frac{\log r}{\log(1/\gamma)}\right),$$

where

$$C_{\mathrm{pf}} := \frac{c_3\,V_{\max}\,g_{\max}\,H_{\mathrm{pf}}^{3/2}}{\varepsilon_{\mathrm{mix}}^3\,g_{\min}} \cdot \sqrt{\log\bigl(2\mathcal{N}_\Theta(1/T)\,|\mathcal{A}|^{H_{\mathrm{pf}}}\bigr)}, \tag{57}$$

with c_3 depending on α and $g_{\max}/g_{\min}$. In particular, the $|\mathcal{A}|$ dependence enters through $\sqrt{H_{\mathrm{pf}}\log|\mathcal{A}|}$.

Term (I): filter mismatch. By Lemma 2,

$$d_Q(b_t^{\theta^\star}, b_t^{\bar\theta_{1:t}}) \;\le\; 2V_{\max}\|b_t^{\theta^\star} - b_t^{\bar\theta_{1:t}}\|_{\mathrm{TV}}.$$

Define $\delta_s := \sup_{s'}\|P_{\bar\theta_s}(\cdot \mid s', a_s) - P_{\theta^\star}(\cdot \mid s', a_s)\|_{\mathrm{TV}} + \sup_{s'}|Z_{\bar\theta_s}(o_s \mid s')/Z_{\theta^\star}(o_s \mid s') - 1|$ and the predict-update kernel

$$K_\theta^{a,o}(s, ds') := \frac{Z_\theta(o \mid s')P_\theta(s' \mid s, a)}{\int Z_\theta(o \mid s'')P_\theta(s'' \mid s, a)\,ds''}\,ds'.$$

Under Assumption 1, bounding the numerator below and denominator above:

$$K_\theta^{a,o}(s, ds') \;\ge\; \frac{g_{\min}\varepsilon_{\mathrm{mix}}}{g_{\max}}\,\nu(s')\,ds' \qquad \forall\, s \in \mathcal{S}, \tag{58}$$

giving Dobrushin coefficient $\rho := 1 - \varepsilon_{\mathrm{mix}} g_{\min}/g_{\max}$. Standard Dobrushin contraction [10, Theorem 4.3.1] under different models $\theta^\star, \bar{\theta}_s$ yields

$$\|b_s^{\theta^\star} - b_s^{\bar{\theta}_{1:s}}\|_{\mathrm{TV}} \;\le\; \rho\, \|b_{s-1}^{\theta^\star} - b_{s-1}^{\bar{\theta}_{1:s-1}}\|_{\mathrm{TV}} + C_{\mathrm{filt}}\, \delta_s, \qquad C_{\mathrm{filt}} := g_{\max}/g_{\min}. \tag{59}$$

Unrolling and using $\delta_s \le d_{\mathrm{model}}^{a_s}(\bar{\theta}_s, \theta^\star)$:

$$\|b_t^{\theta^\star} - b_t^{\bar{\theta}_{1:t}}\|_{\mathrm{TV}} \;\le\; C_{\mathrm{filt}} \sum_{s=1}^{t} \rho^{t-s}\, d_{\mathrm{model}}^{a_s}(\bar{\theta}_s, \theta^\star).$$

Combining. Substituting into equation 44:

$$\mathbb{E}\, d_Q(b_t, b_t^J) \;\le\; \frac{C'_{\mathrm{pf}}}{\sqrt{J}} + 2V_{\max}\, C_{\mathrm{filt}} \sum_{s=1}^{t} \rho^{t-s}\, d_{\mathrm{model}}^{a_s}(\bar{\theta}_s, \theta^\star).$$

When summed over t, exchanging summation order gives $\sum_{t=1}^{T}\sum_{s=1}^{t} \rho^{t-s}\, d_{\mathrm{model}}^{a_s} \le \frac{1}{1-\rho}\sum_{s=1}^{T} d_{\mathrm{model}}^{a_s}$. □

D.4 Model Estimation Error (Terms A + C) and Filter Mismatch

Truncation. Define $V_\theta^{\pi,B,H}(b) := \mathbb{E}_\theta^\pi[\sum_{u=0}^{H-1} \gamma^u r_u \mid b_0 = b]$. Since $|V_\theta^{\pi,B} - V_\theta^{\pi,B,H}| \le \gamma^H V_{\max}$:

$$(\mathrm{A})_t + (\mathrm{C})_t \;\le\; |V_{\theta^\star}^{\star,B,H}(b_t) - V_{\bar{\theta}_{k(t)}}^{\star,B,H}(b_t)| + |V_{\bar{\theta}_{k(t)}}^{\pi_t,B,H}(b_t) - V_{\theta^\star}^{\pi_t,B,H}(b_t)| + \frac{4\gamma^H R_{\max}}{1-\gamma}. \tag{60}$$

Choose $H = \lceil \log(2R_{\max}T/(1-\gamma))/\log(1/\gamma) \rceil$ so the tail contributes $O(1)$.

Confidence sets. Let $\Pr_{\theta,H}^\pi$ denote the length-H trajectory distribution under (θ, π). Define

$$\mathcal{C}_k := \big\{\theta \in \Theta : \mathcal{L}_{T_k}(\theta) \ge \mathcal{L}_{T_k}(\bar{\theta}_k) - \beta\big\}, \tag{61}$$

where $\mathcal{L}_{T_k}$ is the log-likelihood over epochs $1, \dots, k-1$. By construction $\bar{\theta}_k \in \mathcal{C}_k$, and $\theta^\star \in \mathcal{C}_k$ for all k w.h.p. when $\beta \ge c_0 \log[\mathcal{N}_\Theta(1/T)T/\delta_{\mathrm{opt}}]$ [29, Prop. B.1].

Proposition 6 (Likelihood-to-ℓ_1 concentration)**.** *There exists $c > 0$ such that w.p. $\ge 1 - \delta$, simultaneously for all $n \in [T]$ and all $\theta \in \Theta$ with $\mathcal{L}_n(\theta) \ge \mathcal{L}_n(\theta^\star) - \beta$:*

$$\sum_{i=1}^{n} \big\|\Pr_{\theta,H}^{\pi^i} - \Pr_{\theta^\star,H}^{\pi^i}\big\|_1^2 \;\le\; c\,\beta_T, \qquad \beta_T := \beta + \log[\mathcal{N}_\Theta(1/T)\, T/\delta]. \tag{62}$$

Proof. The episodic version is [29, Prop. B.2]. We transfer to the online setting via a blocking argument: partition the trajectory into $\lfloor T/(2H) \rfloor$ blocks of length $2H$, each split into a burn-in half and a data half. Choose H so that

$$\rho^H = (1 - \varepsilon_{\mathrm{mix}} g_{\min}/g_{\max})^H \;\le\; 1/T^2, \qquad \text{i.e.,} \quad H \;\ge\; \frac{2\log T}{\log(1/\rho)}.$$

Our choice $H = O(\log T/\log(1/\gamma))$ satisfies this when $\varepsilon_{\mathrm{mix}} g_{\min}/g_{\max} = \Omega(1-\gamma)$; otherwise, increase H by a constant factor. With $\rho^H \le 1/T^2$, the data halves are approximately independent, with a joint TV perturbation of $O(1/T)$. Applying the episodic bound to the data halves gives equation 62 up to a constant factor. □

Corollary 2 (Epoch-level MLE concentration)**.** *Under the same conditions as Proposition 6, w.p. $\ge 1 - \delta$, simultaneously for all epochs $k \in [K]$:*

$$\sum_{i=1}^{T_{k+1}-1} \big\|\Pr_{\bar{\theta}_k,H}^{\pi^i} - \Pr_{\theta^\star,H}^{\pi^i}\big\|_1^2 \;\le\; c\,\beta_T, \tag{63}$$

where $c > 0$ is a universal constant (at most twice the constant in Proposition 6).

Proof. Fix $k \in [K]$ and write $\| \Pr^{\pi}_{\theta} \|_1 := \| \Pr^{\pi}_{\theta,H} - \Pr^{\pi}_{\theta^\star,H} \|_1$.

Historical data ($i \le T_k$). By MLE optimality, $\bar{\theta}_k$ lies in $\mathcal{C}_k$, so Proposition 6 at $n = T_k$ gives

$$\sum_{i=1}^{T_k} \big\| \Pr^{\pi^i}_{\bar{\theta}_k} \big\|_1^2 \;\le\; c\,\beta_T. \tag{64}$$

Within-epoch data ($i \in E_k$). Since $\bar{\theta}_k$ is $\mathcal{F}_{T_k}$-measurable, we use a self-normalized argument for the fixed hypothesis $\bar{\theta}_k$. Define

$$L_n^{(k)} := \sum_{i=T_k+1}^{T_k+n} \big[\ell_i(\theta^\star) - \ell_i(\bar{\theta}_k)\big], \qquad M_n^{(k)} := \sum_{i=T_k+1}^{T_k+n} \Big[\mathrm{KL}(\Pr^{\pi^i}_{\theta^\star} \,\|\, \Pr^{\pi^i}_{\bar{\theta}_k}) - \big(\ell_i(\theta^\star) - \ell_i(\bar{\theta}_k)\big)\Big],$$

where $\ell_i(\theta) := \log p_\theta(o_i \mid s_i, a_i)$. By Pinsker's inequality and blocking:

$$\frac{1}{2} \sum_{i \in E_k} \big\| \Pr^{\pi^i}_{\bar{\theta}_k} \big\|_1^2 \;\le\; L^{(k)}_{|E_k|} + M^{(k)}_{|E_k|}. \tag{65}$$

$M_n^{(k)}$ is a supermartingale (each increment has non-positive conditional expectation). The maximal inequality and a covering-number union bound give

$$\Pr\left(\sup_{n \le |E_k|} M_n^{(k)} \ge c_1 \log\big[\mathcal{N}_\Theta(1/T)/\delta\big] \right) \le \delta/K.$$

For $L^{(k)}_{|E_k|}$: each $|\ell_i(\theta^\star) - \ell_i(\bar{\theta}_k)| \le \log(g_{\max}/g_{\min})$ and $\mathbb{E}[\ell_i(\theta^\star) - \ell_i(\bar{\theta}_k) \mid \mathcal{F}_{i-1}] \ge 0$. By Azuma-Hoeffding:

$$\Pr\Big(L^{(k)}_{|E_k|} \ge c_2 \sqrt{|E_k| \log(1/\delta)}\, \log(g_{\max}/g_{\min}) \Big) \le \delta/K.$$

Since $\sqrt{T}\log(g_{\max}/g_{\min})$ is absorbed into β_T:

$$\sum_{i \in E_k} \big\| \Pr^{\pi^i}_{\bar{\theta}_k} \big\|_1^2 \;\le\; c\,\beta_T. \tag{66}$$

Combining. Summing equation 64 and equation 66, redefining $c \leftarrow 2c$, and taking a union bound over $k \in [K]$ gives equation 63. □

Transferring model error from historical to arbitrary policies requires action-space coverage via ε-greedy exploration.

Lemma 11 (Trajectory coverage under Doeblin mixing and ε-greedy)**.** *Under Assumption 1, for any two models $\theta, \theta' \in \Theta$, any policy π, and any policy π_ε satisfying $\pi_\varepsilon(a \mid h) \ge \varepsilon/|\mathcal{A}|$ for all a, h (with $\varepsilon \in (0, 1]$),*

$$\big\| \Pr^{\pi}_{\theta,H} - \Pr^{\pi}_{\theta',H} \big\|_1 \;\le\; C_{\mathrm{cov}} \big\| \Pr^{\pi_\varepsilon}_{\theta,H} - \Pr^{\pi_\varepsilon}_{\theta',H} \big\|_1, \qquad C_{\mathrm{cov}} := \frac{H\,|\mathcal{A}|}{\varepsilon} \cdot \frac{g_{\max}^2}{\varepsilon_{\mathrm{mix}}\, g_{\min}^2}. \tag{67}$$

Proof. Step 1: Telescoping. Let $\Pr^{(k)}$ be the hybrid using θ for steps $1, \dots, k$ and θ' for steps $k+1, \dots, H$, under policy π. Let $\Delta_k(s,a) := \|P_\theta(\cdot \mid s,a) - P_{\theta'}(\cdot \mid s,a)\|_{\mathrm{TV}}$ and $\Delta^Z(s) := \sup_o |Z_\theta(o \mid s)/Z_{\theta'}(o \mid s) - 1|$. Then

$$\| \Pr^{\pi}_{\theta,H} - \Pr^{\pi}_{\theta',H} \|_1 \;\le\; \sum_{k=0}^{H-1} \| \Pr^{(k+1)} - \Pr^{(k)} \|_1, \tag{68}$$

$$\| \Pr^{(k+1)} - \Pr^{(k)} \|_1 \;\le\; \mathbb{E}_{\mu_k^\pi}\big[\sup_a \Delta_k(S_k, a) + \Delta^Z(S_k)\big], \tag{69}$$

where μ_k^π is the state marginal at step k under the hybrid.

Step 2: State-marginal ratio. By the Doeblin decomposition $\tilde{\mu}_k^{\pi} \geq \varepsilon_{\text{mix}}\nu$ and $\tilde{\mu}_k^{\pi'} \geq \varepsilon_{\text{mix}}\nu$, for any $h : \mathcal{S} \to [0,1]$:

$$\mathbb{E}_{\tilde{\mu}_k^{\pi}}[h] \;\leq\; \varepsilon_{\text{mix}}\mathbb{E}_{\nu}[h] + (1-\varepsilon_{\text{mix}}) \;\leq\; \frac{1}{\varepsilon_{\text{mix}}}\mathbb{E}_{\tilde{\mu}_k^{\pi'}}[h].$$

Accounting for the factor $g_{\max}/g_{\min}$ from the Bayes update normalization gives

$$\mathbb{E}_{\mu_k^{\pi}}[h] \;\leq\; \frac{g_{\max}}{g_{\min}\varepsilon_{\text{mix}}}\,\mathbb{E}_{\mu_k^{\pi'}}[h].$$

Step 3: Action coverage. Since $\pi_\varepsilon(a \mid h) \geq \varepsilon/|\mathcal{A}|$:

$$\sup_a \Delta_k(s,a) \;\leq\; \frac{|\mathcal{A}|}{\varepsilon}\,\mathbb{E}_{A\sim\pi_\varepsilon}[\Delta_k(s,A)].$$

Step 4: Assembly. Combining Steps 1–3 term-by-term:

$$\|\Pr^{(k+1)} - \Pr^{(k)}\|_1 \;\leq\; \frac{|\mathcal{A}|}{\varepsilon}\cdot\frac{g_{\max}^2}{\varepsilon_{\text{mix}}\, g_{\min}^2}\cdot\|\Pr_{\pi_\varepsilon}^{(k+1)} - \Pr_{\pi_\varepsilon}^{(k)}\|_1.$$

Summing over $k = 0, \ldots, H-1$ yields equation 67. □

Lemma 12 (Simulation lemma)**.** *For any $\theta, \theta' \in \Theta$, π, and b:*

$$\left|V_\theta^{\pi,B,H}(b) - V_{\theta'}^{\pi,B,H}(b)\right| \;\leq\; \frac{R_{\max}}{1-\gamma}\left\|\Pr_{\theta,H}^{\pi} - \Pr_{\theta',H}^{\pi}\right\|_1.$$

Proof. Telescope the H-truncated value difference into one-step terms, each bounded by $R_{\max}$. □

Proposition 7 (Control of (A) + (C) + filter mismatch)**.** *Under Assumption 1, consider the doubling-epoch variant of Algorithm 1 in which time is partitioned into epochs $E_k = [T_k, T_{k+1})$ with $T_k = 2^{k-1}$, $k = 1, \ldots, K := \lceil \log_2 T \rceil$. Within each epoch E_k, the model $\bar{\theta}_k$ is frozen (computed from data in epochs $1, \ldots, k-1$) and the agent uses ε_k-greedy exploration with a constant rate ε_k. Let $K_{\text{cov}} := H|\mathcal{A}|\, g_{\max}^2/(\varepsilon_{\text{mix}}\, g_{\min}^2)$. Then with probability $\geq 1 - \delta_{\text{opt}}$,*

$$\sum_{t=1}^{T}(C)_t \;\leq\; \frac{R_{\max}}{1-\gamma}\sqrt{c\,\beta_T}\sum_{k=1}^{K}\sqrt{|E_k|} + \frac{2\gamma^H R_{\max} T}{1-\gamma}, \tag{70}$$

$$\sum_{t=1}^{T}(A)_t \;\leq\; \frac{R_{\max}\, K_{\text{cov}}}{1-\gamma}\sqrt{c\,\beta_T}\sum_{k=2}^{K}\frac{|E_k|}{\sqrt{\sum_{j=1}^{k-1}\varepsilon_j^2|E_j|}} + \frac{2\gamma^H R_{\max} T}{1-\gamma}, \tag{71}$$

$$\sum_{t=1}^{T} d_{\text{model}}^{a_t}(\bar{\theta}_{k(t)}, \theta^\star) \;\leq\; C_{\text{mix}}\sqrt{c\,\beta_T}\sum_{k=1}^{K}\sqrt{|E_k|}, \tag{72}$$

where $d_{\text{model}}^a(\theta,\theta') := \sup_{s\in\mathcal{S}}\|P_\theta(\cdot \mid s,a) - P_{\theta'}(\cdot \mid s,a)\|_{\text{TV}} + \sup_{s\in\mathcal{S},o\in\mathcal{O}}\left|Z_\theta(o \mid s)/Z_{\theta'}(o \mid s) - 1\right|$ is the per-action model discrepancy, and C_{mix} depends on $\varepsilon_{\text{mix}}, g_{\min}, g_{\max}, H$.

Proof. By Corollary 2, equation equation 63 holds for all $k \leq K$. For brevity, write $\|\Pr_\theta^\pi\|_1 := \|\Pr_{\theta,H}^\pi - \Pr_{\theta^\star,H}^\pi\|_1$ throughout this proof.

Term (C). By Lemma 12,

$$(\text{C})_t \;\leq\; \frac{R_{\max}}{1-\gamma}\left\|\Pr_{\bar{\theta}_{k(t)}}^{\pi^t}\right\|_1 + \frac{2\gamma^H R_{\max}}{1-\gamma}.$$

Since $\bar{\theta}_k$ is fixed within E_k, Cauchy-Schwarz and equation 63 give

$$\sum_{t\in E_k}\left\|\Pr_{\bar{\theta}_k}^{\pi^t}\right\|_1 \;\leq\; \sqrt{|E_k|\sum_{t\in E_k}\left\|\Pr_{\bar{\theta}_k}^{\pi^t}\right\|_1^2} \;\leq\; \sqrt{|E_k|\cdot c\,\beta_T}.$$

Summing across epochs: $\sum_{k=1}^{K}\sqrt{|E_k|\cdot c\beta_T} = \sqrt{c\beta_T}\sum_k 2^{(k-1)/2} = O(\sqrt{T\beta_T})$.

Term (A). By Lemmas 12 and 11:

$$(\mathrm{A})_t \;\le\; \frac{R_{\max}}{1-\gamma}\sup_{\pi}\big\|\Pr^{\pi}_{\bar\theta_{k(t)}}\big\|_1 + \frac{2\gamma^H R_{\max}}{1-\gamma}, \qquad \big\|\Pr^{\pi}_{\bar\theta_k}\big\|_1 \;\le\; \frac{K_{\mathrm{cov}}}{\varepsilon_j}\big\|\Pr^{\pi^j}_{\bar\theta_k}\big\|_1 \quad \forall\, j<k.$$

Squaring, multiplying by ε_j^2, and summing over prior epochs using equation 63:

$$\big\|\Pr^{\pi}_{\bar\theta_k}\big\|_1^2\sum_{j<k}\varepsilon_j^2|E_j| \;\le\; K_{\mathrm{cov}}^2\sum_{j<k}\sum_{t\in E_j}\big\|\Pr^{\pi^t}_{\bar\theta_k}\big\|_1^2 \;\le\; K_{\mathrm{cov}}^2\, c\,\beta_T \quad\Longrightarrow\quad \sup_{\pi}\big\|\Pr^{\pi}_{\bar\theta_k}\big\|_1 \;\le\; \frac{K_{\mathrm{cov}}\sqrt{c\beta_T}}{\sqrt{\sum_{j<k}\varepsilon_j^2|E_j|}}.$$

Multiplying by $|E_k|$ and summing gives equation 71.

Filter mismatch. By the telescoping argument of Lemma 11, using only Steps 1–2 and hence without the $|\mathcal{A}|/\varepsilon$ action-coverage factor, since we evaluate at the executed action a_t,

$$d^{a_t}_{\mathrm{model}}(\bar\theta_{k(t)},\theta^\star) \;\le\; C_{\mathrm{mix}}\big\|\Pr^{\pi^t}_{\bar\theta_{k(t)}}\big\|_1, \qquad C_{\mathrm{mix}} := \frac{H\, g_{\max}^2}{\varepsilon_{\mathrm{mix}}\, g_{\min}^2}.$$

Cauchy-Schwarz and equation 63 give $\sum_{t\in E_k}\|\Pr^{\pi^t}_{\bar\theta_k}\|_1 \le \sqrt{|E_k|\cdot c\,\beta_T}$. Summing across epochs yields equation 72. □

D.5 Proof of the Main Regret Bound (Theorem 1)

Proof of Theorem 1. Use doubling epochs $E_k = [T_k, T_{k+1})$, $T_k = 2^{k-1}$, $K = \lceil \log_2 T\rceil$, with ε_k-greedy exploration within each epoch. Decompose:

$$\mathrm{Reg}_T \;\le\; \underbrace{\frac{2R_{\max}}{1-\gamma}\sum_k |E_k|\varepsilon_k}_{\text{exploration}} + \mathbb{E}\sum_{t=1}^{T}\big[(\mathrm{A})_t + (\mathrm{B})_t + (\mathrm{C})_t\big].$$

Term (C). By equation 70 (choosing $\delta_{\mathrm{opt}} = 1/T$):

$$\mathbb{E}\sum_t (\mathrm{C})_t \;\le\; \frac{R_{\max}\sqrt{c\beta_T}}{1-\gamma}\sum_{k=1}^{K}\sqrt{|E_k|} + O(1) \;=\; \widetilde{O}\left(\frac{\sqrt{T}}{1-\gamma}\right).$$

Term (A). By equation 71 with $\varepsilon_k = \eta/\sqrt{|E_k|}$ (so $\varepsilon_k^2|E_k| = \eta^2$):

$$\mathbb{E}\sum_t (\mathrm{A})_t \;\le\; \frac{R_{\max}K_{\mathrm{cov}}\sqrt{c\beta_T}}{(1-\gamma)\eta}\sum_{k=2}^{K}\frac{|E_k|}{\sqrt{k-1}} + O(1).$$

Since $\sum_{k=2}^{K} 2^{k-1}/\sqrt{k-1} = \Theta(T/\sqrt{\log T})$, this scales as $K_{\mathrm{cov}}\sqrt{\beta_T}\cdot T/(\eta\sqrt{\log T}(1-\gamma))$.

Term (B). By Theorem 2, Propositions 2, 3, 4:

$$\mathbb{E}\sum_t (\mathrm{B})_t \;\le\; \frac{2C_{\mathrm{pf}}\,T}{(1-\gamma)}\cdot J^{-\frac{1}{2\kappa_{\mathrm{pf}}}} + \frac{C_{\mathrm{mc}}V_{\max}T}{(1-\gamma)\sqrt{N}} + \frac{2\gamma^{d+1}}{(1-\gamma)^2}\sum_t \Xi_{\bar\theta_{k(t)}} + O(\sqrt{T\beta_T}).$$

The $O(\sqrt{T\beta_T})$ absorbs the filter mismatch via equation 72.

Optimization. The exploration cost is $O(\eta\sqrt{T}R_{\max}/(1-\gamma))$. Balancing it with term (A) gives $\eta^\star \sim K_{\mathrm{cov}}^{1/2}\,T^{1/4}(\beta_T/\log T)^{1/4}$ and a total of $\widetilde{O}(K_{\mathrm{cov}}^{1/2}T^{3/4}/(1-\gamma)^{3/2})$. Term (C) is $\widetilde{O}(\sqrt{T})$, and the $T^{3/4}$ rate comes from the coverage transfer for term (A). □

D.6 Optimal Exploration Schedule

Set $\varepsilon_k = \eta/\sqrt{|E_k|}$ so that $\varepsilon_k^2|E_k| = \eta^2$ and $\sum_{j<k}\varepsilon_j^2|E_j| = (k-1)\eta^2$. The exploration-dependent terms scale as:

$$\begin{aligned}
\text{Exploration cost:}\quad & \textstyle\sum_k |E_k|\varepsilon_k = \eta\sum_k\sqrt{|E_k|} = O(\eta\sqrt{T}),\\
\text{Term (A):}\quad & O\big(K_{\mathrm{cov}}\sqrt{\beta_T}\cdot T/(\eta\sqrt{\log T}(1-\gamma))\big),\\
\text{Term (C):}\quad & O\big(\sqrt{T\beta_T}/(1-\gamma)\big), \quad \text{independent of } \eta.
\end{aligned}$$

Balancing term (A) against the exploration cost gives

$$\eta^\star \sim \left(\frac{K_{\mathrm{cov}}\sqrt{T\beta_T}}{\sqrt{\log T}}\right)^{1/2}, \qquad \text{(A) + exploration} = \widetilde{O}\left(\frac{K_{\mathrm{cov}}^{1/2}T^{3/4}}{(1-\gamma)}\right).$$

The total regret is:

$$\mathrm{Reg}_T = \widetilde{O}\left(\frac{K_{\mathrm{cov}}^{1/2}\,R_{\max}\,T^{3/4}}{(1-\gamma)^{3/2}}\right) + \frac{C_{\mathrm{pf}}\,R_{\max}\,T}{(1-\gamma)^2}\cdot J^{-\frac{1}{2\kappa_{\mathrm{pf}}}} + \frac{C_{\mathrm{mc}}\,R_{\max}\,T}{(1-\gamma)^2\sqrt{N}} + \frac{2\gamma^{d+1}}{(1-\gamma)^2}\sum_t \Xi_{\bar{\theta}_{k(t)}}. \tag{73}$$

The $T^{3/4}$ rate is driven by the $1/\varepsilon$ coverage transfer for term (A). Using optimistic model selection ($\tilde{\theta}_t := \arg\max_{\theta\in\mathcal{C}_t} V_\theta^{\star,B}$) eliminates term (A) and recovers $\widetilde{O}(\sqrt{T})$ from term (C) alone, at the cost of solving an optimistic planning problem [29].

A constant ε gives the same $\widetilde{O}(T^{3/4})$ rate after optimization and may be preferable when T is unknown.

D.7 Greedy Exploitation with Offline Coverage

The $T^{3/4}$ rate in Theorem 1 arises from the coverage transfer for term (A), which requires ε-greedy exploration to ensure the MLE concentrates at all policies. An alternative is to replace online exploration with an offline dataset that already provides coverage of the optimal policy. This mirrors the hybrid RL framework [35], where offline data with sufficient coverage enables greedy online exploitation.

Setting. Suppose the agent has access to an offline dataset $\mathcal{D}_{\mathrm{off}}$ of N i.i.d. length-H trajectories collected by a behavior policy π_b in the true POMDP (model $\theta^\star$). The agent then runs PF-MCTQ for T online steps with *greedy* exploitation ($\varepsilon_t = 0$, no exploration). The MLE is computed from the union $\mathcal{D}_{\mathrm{off}} \cup \mathcal{D}_{\mathrm{on}}$.

Definition 2 (Single-policy concentrability)**.** *The single-policy concentrability coefficient of the behavior policy π_b with respect to the optimal policy $\pi^\star := \arg\max_\pi V_{\theta^\star}^{\pi,B}$ is*

$$c_{\mathrm{off}}^\star := \sup_{\theta,\theta'\in\Theta} \frac{\|\Pr_{\theta,H}^{\pi^\star} - \Pr_{\theta',H}^{\pi^\star}\|_1}{\|\Pr_{\theta,H}^{\pi_b} - \Pr_{\theta',H}^{\pi_b}\|_1}. \tag{74}$$

Under Assumption 1, if $\pi_b(a \mid h) \geq \alpha_b/|\mathcal{A}|$ for all a, h, Lemma 11 gives $c_{\mathrm{off}}^\star \leq K_{\mathrm{cov}}/\alpha_b = \frac{H|\mathcal{A}|}{\alpha_b}\cdot\frac{g_{\max}^2}{\varepsilon_{\mathrm{mix}}\,g_{\min}^2}$. Unlike the all-policy concentrability provided by ε-greedy exploration, which takes a supremum over all policies, $c_{\mathrm{off}}^\star$ involves only the *optimal* policy.

Proposition 8 (Regret with offline coverage)**.** *Under Assumption 1, suppose the agent has N offline trajectories from π_b and runs the doubling-epoch PF-MCTQ with greedy exploitation ($\varepsilon_k = 0$) for T online steps. Let $\beta_{T,N} := \beta + \log[\mathcal{N}_\Theta(1/(T+N))(T+N)/\delta]$. Then w.p. $\geq 1-\delta$:*

$$\mathrm{Reg}_T \;\leq\; \underbrace{\widetilde{O}\left(\frac{R_{\max}\sqrt{T}}{1-\gamma}\right)}_{\textit{term (C)}} + \underbrace{\frac{c_{\mathrm{off}}^\star R_{\max}\sqrt{c\beta_{T,N}}}{1-\gamma}\cdot\frac{T}{\sqrt{N}}}_{\textit{term (A)}} + \textit{computational terms}, \tag{75}$$

where the computational terms are the same as in Theorem 1 (particle, MC, structural).

In particular, when $N \geq (c_{\mathrm{off}}^\star)^2\,T$, the total regret is $\widetilde{O}\big(\sqrt{T}/(1-\gamma)\big)$ plus computational terms, so $\sqrt{T}$ regret is achieved without any online exploration.

Proof. The regret decomposition equation 11 gives

$$\mathrm{Reg}_T \;=\; \mathbb{E}\sum_{t=1}^{T}\big[(\mathrm{A})_t + (\mathrm{B})_t + (\mathrm{C})_t\big].$$

There is no exploration cost since $\varepsilon_k = 0$.

MLE concentration with offline data. The MLE is fit on $\mathcal{D}_{\text{off}} \cup \mathcal{D}_{\text{on}}$. The offline trajectories are i.i.d. under $(\theta^\star, \pi_b)$, so the episodic MLE concentration [29, Prop. B.2] applies directly to the offline portion. Combined with the blocking argument for the online portion (Proposition 6):

$$N \left\|\Pr_{\bar{\theta}_k}^{\pi_b}\right\|_1^2 + \sum_{t=1}^{T_{k+1}-1} \left\|\Pr_{\bar{\theta}_k}^{\pi^t}\right\|_1^2 \;\le\; c\,\beta_{T,N} \qquad \forall\, k \le K, \tag{76}$$

using the shorthand $\|\Pr_\theta^\pi\|_1 := \|\Pr_{\theta,H}^\pi - \Pr_{\theta^\star,H}^\pi\|_1$.

Term (C). As in the proof of Proposition 7, since $\bar{\theta}_k$ is fixed within E_k, applying Cauchy-Schwarz within each epoch to the online terms in equation 76 gives

$$\sum_{t=1}^{T} (\mathrm{C})_t \;\le\; \frac{R_{\max}}{1-\gamma}\sqrt{c\beta_{T,N}} \sum_{k=1}^{K} \sqrt{|E_k|} + O(\gamma^H T) \;=\; \widetilde{O}\left(\frac{R_{\max}\sqrt{T}}{1-\gamma}\right).$$

Term (A). We use a tighter decomposition that avoids $\sup_\pi$:

$$(\mathrm{A})_t = V_{\theta^\star}^{\star,B}(b_t) - V_{\bar{\theta}_{k(t)}}^{\star,B}(b_t) \;\le\; V_{\theta^\star}^{\pi^\star,B}(b_t) - V_{\bar{\theta}_{k(t)}}^{\pi^\star,B}(b_t) \;\le\; \frac{R_{\max}}{1-\gamma}\left\|\Pr_{\bar{\theta}_{k(t)}}^{\pi^\star}\right\|_1, \tag{77}$$

where the first inequality uses $V_{\bar{\theta}_{k(t)}}^{\star,B} \ge V_{\bar{\theta}_{k(t)}}^{\pi^\star,B}$ (optimality under $\bar{\theta}_{k(t)}$) and the second is the simulation lemma (Lemma 12).

By Definition 2:

$$\left\|\Pr_{\bar{\theta}_k}^{\pi^\star}\right\|_1 \;\le\; c_{\text{off}}^\star \left\|\Pr_{\bar{\theta}_k}^{\pi_b}\right\|_1.$$

From the offline portion of equation 76:

$$\left\|\Pr_{\bar{\theta}_k}^{\pi_b}\right\|_1 \;\le\; \sqrt{\frac{c\,\beta_{T,N}}{N}}.$$

Since $\bar{\theta}_k$ is frozen within E_k, this bound is constant across the epoch:

$$\sum_{t \in E_k} (\mathrm{A})_t \;\le\; \frac{R_{\max}}{1-\gamma} \cdot c_{\text{off}}^\star \sqrt{\frac{c\,\beta_{T,N}}{N}} \cdot |E_k|.$$

Summing across epochs:

$$\sum_{t=1}^{T} (\mathrm{A})_t \;\le\; \frac{c_{\text{off}}^\star R_{\max}}{1-\gamma}\sqrt{\frac{c\,\beta_{T,N}}{N}} \cdot T.$$

Term (B) and filter mismatch. Identical to the proof of Theorem 1. The filter mismatch $\sum_t d_{\text{model}}^{a_t}$ is bounded by $C_{\text{mix}}\sqrt{c\beta_{T,N}}\sum_k \sqrt{|E_k|} = \widetilde{O}(\sqrt{T})$ via equation 72 (the online diagonal bound, which does not require coverage).

Combining.

$$\mathrm{Reg}_T \;\le\; \frac{R_{\max}}{1-\gamma}\widetilde{O}(\sqrt{T}) + \frac{c_{\text{off}}^\star R_{\max}\sqrt{c\beta_{T,N}}}{1-\gamma} \cdot \frac{T}{\sqrt{N}} + \frac{C'_{\text{pf}} R_{\max} T}{(1-\gamma)^2\sqrt{J}} + \frac{C_{\text{mc}} R_{\max} T}{(1-\gamma)^2\sqrt{N_{\text{mc}}}} + \frac{2\gamma^{d+1}}{(1-\gamma)^2}\sum_t \Xi_{\bar{\theta}_{k(t)}}.$$

When $N \ge (c_{\text{off}}^\star)^2\, T$, the second term is $O(\sqrt{T\beta_{T,N}}/(1-\gamma)) = \widetilde{O}(\sqrt{T}/(1-\gamma))$, and the leading rate is $\widetilde{O}(\sqrt{T})$. □

D.8 Variance Perspective on the QMDP Bias

We provide a variance-based interpretation of the Jensen gap Ξ_θ defined in equation 13. We show that Ξ_θ is controlled by the curvature of the value function along a sufficient embedding and the posterior variance of the embedded features.

Value-relevant embedding. Let $g : \mathcal{S} \to \mathbb{R}^k$ be a measurable embedding, and write

$$m_g(b) := \mathbb{E}_{S\sim b}[g(S)], \qquad \Sigma_g(b) := \mathrm{Cov}_{S\sim b}\big(g(S)\big).$$

Assumption 3 (Value-relevant embedding)**.** *There exist twice continuously differentiable maps $F_\theta^a : \mathbb{R}^k \to \mathbb{R}$ for each $a \in \mathcal{A}$ such that $Q_\theta^{\star,S}(s,a) = F_\theta^a(g(s))$ for all $(s,a) \in \mathcal{S} \times \mathcal{A}$, with curvature bounds*

$$\sup_{a\in\mathcal{A}} \sup_{x\in\mathrm{conv}(g(\mathcal{S}))} \big\|\nabla^2 F_\theta^a(x)\big\|_{\mathrm{op}} \le L_\theta \qquad \textit{and} \qquad \sup_{x\in\mathrm{conv}(g(\mathcal{S}))} \big\|\nabla^2 F_\theta(x)\big\|_{\mathrm{op}} \le L_\theta,$$

where $F_\theta(x) := \max_a F_\theta^a(x)$ and $V_\theta^{\star,S}(s) = F_\theta(g(s))$.

The second curvature condition on $F_\theta = \max_a F_\theta^a$ does not hold automatically, since F_θ is not C^2 at switching boundaries where the maximizing action changes. The bound equation 83 uses $\|\nabla^2 F_\theta\|_{\mathrm{op}} \le L_\theta$ only for the upper bound on $\widetilde{V}_\theta^\star(b)$; the lower bound on $\bar{V}_\theta^{\star,S}(b)$ uses only the per-action bounds $\|\nabla^2 F_\theta^a\|_{\mathrm{op}} \le L_\theta$.

We first state an auxiliary lemma used in the proof of Proposition 9.

Lemma 13 (Mean-belief value approximation)**.** *Under Assumption 3, for every belief $b \in \Delta(\mathcal{S})$,*

$$\big|\bar{V}_\theta^{\star,S}(b) - F_\theta(m_g(b))\big| \;\le\; \frac{L_\theta}{2}\,\mathrm{tr}\big(\Sigma_g(b)\big). \tag{78}$$

Proof. By Jensen's inequality applied to the pointwise maximum,

$$\bar{V}_\theta^{\star,S}(b) = \max_{a\in\mathcal{A}} \int_{\mathcal{S}} Q_\theta^{\star,S}(s,a)\, b(\mathrm{d}s) \;\le\; \int_{\mathcal{S}} V_\theta^{\star,S}(s)\, b(\mathrm{d}s).$$

By Assumption 3, $V_\theta^{\star,S}(s) = F_\theta(g(s))$, so

$$\bar{V}_\theta^{\star,S}(b) \;\le\; \int_{\mathcal{S}} F_\theta(g(s))\, b(\mathrm{d}s). \tag{79}$$

Expand F_θ in a second-order Taylor expansion around $m_g(b)$. For each s, there exists $\xi_s \in \mathrm{conv}(g(\mathcal{S}))$ on the segment between $m_g(b)$ and $g(s)$ such that

$$F_\theta(g(s)) = F_\theta(m_g(b)) + \nabla F_\theta(m_g(b))^\top \big(g(s)-m_g(b)\big) + \tfrac{1}{2}\big(g(s)-m_g(b)\big)^\top \nabla^2 F_\theta(\xi_s)\,\big(g(s)-m_g(b)\big).$$

Integrating against b and using $\int_{\mathcal{S}} (g(s) - m_g(b))\, b(\mathrm{d}s) = 0$,

$$\int_{\mathcal{S}} F_\theta(g(s))\, b(\mathrm{d}s) - F_\theta(m_g(b)) = \tfrac{1}{2}\int_{\mathcal{S}} \big(g(s) - m_g(b)\big)^\top \nabla^2 F_\theta(\xi_s)\,\big(g(s) - m_g(b)\big)\, b(\mathrm{d}s). \tag{80}$$

Since $\|\nabla^2 F_\theta(\xi_s)\|_{\mathrm{op}} \le L_\theta$, each integrand is bounded in absolute value by $L_\theta\,\|g(s) - m_g(b)\|^2$, so equation 80 lies in $[-\frac{L_\theta}{2}\,\mathrm{tr}(\Sigma_g(b)),\ \frac{L_\theta}{2}\,\mathrm{tr}(\Sigma_g(b))]$. Combined with equation 79, we obtain the upper bound

$$\bar{V}_\theta^{\star,S}(b) \;\le\; F_\theta(m_g(b)) + \tfrac{L_\theta}{2}\,\mathrm{tr}\big(\Sigma_g(b)\big). \tag{81}$$

For the lower bound, fix any $a \in \mathcal{A}$. By Assumption 3, $Q_\theta^{\star,S}(s,a) = F_\theta^a(g(s))$ with $\|\nabla^2 F_\theta^a\|_{\mathrm{op}} \le L_\theta$. The same Taylor expansion gives

$$\int F_\theta^a(g(s))\, b(\mathrm{d}s) \;\ge\; F_\theta^a(m_g(b)) - \tfrac{L_\theta}{2}\,\mathrm{tr}(\Sigma_g(b)).$$

Taking $\max_a$:

$$\bar{V}_\theta^{\star,S}(b) = \max_a \int F_\theta^a(g(s))\, b(\mathrm{d}s) \;\ge\; \max_a F_\theta^a(m_g(b)) - \tfrac{L_\theta}{2}\,\mathrm{tr}(\Sigma_g(b)) = F_\theta(m_g(b)) - \tfrac{L_\theta}{2}\,\mathrm{tr}(\Sigma_g(b)). \tag{82}$$

□

Remark 5. *The upper half equation 81 follows from Jensen's inequality alone and is used to verify the hypothesis of Proposition 9 in concrete settings.*

Proposition 9 (Curvature bound on Ξ_θ)**.** *Under Assumption 3, for every belief $b \in \Delta(\mathcal{S})$,*

$$\Xi_\theta(b) := \widetilde{V}_\theta^\star(b) - \bar{V}_\theta^{\star,S}(b) \;\le\; L_\theta \,\mathrm{tr}\big(\Sigma_g(b)\big). \tag{83}$$

In particular, $\Xi_\theta = \sup_b \Xi_\theta(b) \le L_\theta \sup_b \mathrm{tr}(\Sigma_g(b))$.

Proof. We bound $\Xi_\theta(b) = \widetilde{V}_\theta^\star(b) - \bar{V}_\theta^{\star,S}(b)$ for each belief b.

Upper bound on $\widetilde{V}_\theta^\star(b)$. By Assumption 3, $V_\theta^{\star,S}(s) = F_\theta(g(s))$ where $F_\theta = \max_a F_\theta^a$ has curvature $\le L_\theta$. Applying the standard Taylor upper bound (as in Lemma 13) to F_θ:

$$\widetilde{V}_\theta^\star(b) = \int F_\theta(g(s))\, b(\mathrm{d}s) \le F_\theta(m_g(b)) + \frac{L_\theta}{2}\,\mathrm{tr}(\Sigma_g(b)). \tag{84}$$

Since $F_\theta(m_g(b)) = \max_a F_\theta^a(m_g(b))$, this gives $\widetilde{V}_\theta^\star(b) \le \max_a F_\theta^a(m_g(b)) + \frac{L_\theta}{2}\mathrm{tr}(\Sigma_g(b))$.

Lower bound on $\bar{V}_\theta^{\star,S}(b)$. By Assumption 3, $Q_\theta^{\star,S}(s,a) = F_\theta^a(g(s))$ for each a. For any fixed a,

$$\int Q_\theta^{\star,S}(s,a)\, b(\mathrm{d}s) = \int F_\theta^a(g(s))\, b(\mathrm{d}s) \ge F_\theta^a(m_g(b)) - \frac{L_\theta}{2}\mathrm{tr}(\Sigma_g(b)),$$

where the inequality is the Taylor lower bound (each F_θ^a has curvature $\le L_\theta$). Taking the maximum over a:

$$\bar{V}_\theta^{\star,S}(b) = \max_a \int F_\theta^a(g(s))\, b(\mathrm{d}s) \ge \max_a \Big[F_\theta^a(m_g(b)) - \frac{L_\theta}{2}\mathrm{tr}(\Sigma_g(b))\Big] = \max_a F_\theta^a(m_g(b)) - \frac{L_\theta}{2}\mathrm{tr}(\Sigma_g(b)).$$

Combining.

$$\Xi_\theta(b) = \widetilde{V}_\theta^\star(b) - \bar{V}_\theta^{\star,S}(b) \le \big[\max_a F_\theta^a(m_g(b)) + \frac{L_\theta}{2}\mathrm{tr}(\Sigma_g(b))\big] - \big[\max_a F_\theta^a(m_g(b)) - \frac{L_\theta}{2}\mathrm{tr}(\Sigma_g(b))\big] = L_\theta\,\mathrm{tr}(\Sigma_g(b)).$$

□

D.9 Case Studies

Case 1: Discounted partially observed LQG. Consider linear-Gaussian dynamics $S_{t+1} = \Phi S_t + \Gamma A_t + W_t$, $O_t = HS_t + \eta_t$, with $W_t \sim \mathcal{N}(0, \Sigma_W)$, $\eta_t \sim \mathcal{N}(0, \Sigma_\eta)$ i.i.d., quadratic reward $r(s,a) = -(s^\top Q_{\mathrm{c}} s + a^\top R a)$, $Q_{\mathrm{c}} \succeq 0$, $R \succ 0$. Assume (Φ, Γ) is stabilizable and (Φ, H) is detectable. For LQG, we obtain the closed-form expression $\Xi_\theta = \mathrm{tr}\big((M-P)\,\bar{\Sigma}\big)$, where P solve the Riccati equation, $M = Q_c + \gamma\Phi^\top P \Phi$, and $\bar{\Sigma}$ is the steady-state Kalman predictive covariance. When $\Gamma = 0$, we have $M = P$ and $\Xi_\theta = 0$. Note that for LQG, the QMDP policy is always optimal.

Corollary 3 (LQG)**.** *Let $M := Q_{\mathrm{c}} + \gamma\Phi^\top P\Phi$ where $P \succeq 0$ solves the discounted algebraic Riccati equation, and let $\bar{\Sigma}_{\mathrm{post}}$ be the steady-state Kalman posterior (filtered) covariance satisfying $\bar{\Sigma}_{\mathrm{post}} = \bar{\Sigma} - \bar{\Sigma}H^\top(H\bar{\Sigma}H^\top + \Sigma_\eta)^{-1}H\bar{\Sigma}$ with $\bar{\Sigma} := \Phi\bar{\Sigma}_{\mathrm{post}}\Phi^\top + \Sigma_W$ the steady-state predictive covariance. Then $\Xi_\theta = \mathrm{tr}\big((M-P)\,\bar{\Sigma}\big)$ and the QMDP policy is belief-MDP optimal: $\sup_b\big[V_\theta^{\star,B}(b) - V_\theta^{\pi_{\mathrm{Q}},B}(b)\big] = 0$.*

The QMDP action equals the certainty-equivalent LQR control $a = Km$, which is belief-MDP optimal by the separation principle. When $\Gamma = 0$, the DARE gives $M = P$, so $\Xi_\theta = 0$ and the QMDP value and the belief-MDP value coincide. When $\Gamma \neq 0$, $M - P \succeq 0$ is generically nonzero, so $\Xi_\theta > 0$. In that case the QMDP *policy* remains optimal, but the QMDP *value* underestimates the belief-MDP value by $\mathrm{tr}((M-P)\bar{\Sigma})$.

Remark. LQG has unbounded reward and $V_{\max} = \infty$, so it lies outside the standing assumption $|r(s,a)| \le R_{\max}$ of Theorem 1. The corollary illustrates the structural analysis (Jensen gap and policy optimality) but does not directly instantiate the regret bound.

Case 2: Identifiable contextual MDP. Consider an episodic contextual MDP [24] with a fixed latent context $C \in \mathcal{C}$ sampled once per episode and Markov dynamics $S_{t+1} \sim P_c(\cdot \mid S_t, A_t)$, $r_t = r_c(S_t, A_t)$ given $C = c$. Assume (i) there exists an estimator $\widehat{C}_t = \widehat{C}_t(H_t)$ with $\Pr(\widehat{C}_t \neq C) \le \rho_t$, $\rho_t \downarrow 0$; and (ii) $\sup_{c\in\mathcal{C}} \|V_c^{\star,S}\|_\infty \le V_{\max}$. Then $\Xi_{\theta,t} \;\le\; 2V_{\max}\,\rho_t$ and $\sup_b\big[V_\theta^{\star,B}(b) -$

$V_\theta^{\pi_{\mathrm{Q}},B}(b)\big] \;\le\; 4\gamma\, V_{\max}\,\rho_t/(1-\gamma)^2$. As the context becomes identifiable ($\rho_t \to 0$), both the structural residual and the QMDP policy suboptimality vanish. For the structural bias term $\sum_t \Xi_t \le 2V_{\max}\sum_t \rho_t$ to be $o(\sqrt{T})$, we need $\sum_{t=1}^{T}\rho_t = o(\sqrt{T})$. This result also suggests a schedule for the planning depth d that decays with t, reflecting diminishing structural bias.

Proofs. The fully observed value function is $V_\theta^{\star,S}(s) = -s^\top P s - c_0$, where $P \succeq 0$ solves the discounted algebraic Riccati equation (DARE). The constant c_0 is determined by substituting into the Bellman equation:

$$-s^\top P s - c_0 = \max_a\{r(s,a) + \gamma\mathbb{E}[V^{\star,S}(s')]\},$$

which gives $c_0 = \gamma[\mathrm{tr}(P\Sigma_W) + c_0]$ from the noise covariance term, yielding $c_0 = \gamma\mathrm{tr}(P\Sigma_W)/(1-\gamma)$. Hence

$$Q_\theta^{\star,S}(s,a) = -(s^\top Q_{\mathrm{c}} s + a^\top R a) - \gamma(\Phi s + \Gamma a)^\top P(\Phi s + \Gamma a) - \gamma\mathrm{tr}(P\Sigma_W) - \gamma c_0.$$

For belief $b = \mathcal{N}(m,\Sigma)$,

$$\int Q_\theta^{\star,S}(s,a)\, b(\mathrm{d}s) = -m^\top Q_{\mathrm{c}} m - a^\top R a - \gamma(\Phi m + \Gamma a)^\top P(\Phi m + \Gamma a) - \mathrm{tr}(M\Sigma) - \gamma\mathrm{tr}(P\Sigma_W) - \gamma c_0,$$

where $M := Q_{\mathrm{c}} + \gamma\Phi^\top P\Phi$. All Σ-dependent terms are independent of a, so the QMDP action is

$$\arg\max_a \int Q_\theta^{\star,S}(s,a)\, b(\mathrm{d}s) = \arg\max_a\big\{-a^\top R a - \gamma(\Phi m + \Gamma a)^\top P(\Phi m + \Gamma a)\big\} = Km,$$

with

$$K = -\gamma(R + \gamma\Gamma^\top P\Gamma)^{-1}\Gamma^\top P\Phi.$$

By the separation principle for discounted LQG (stabilizable/detectable), the belief-MDP optimal action is also Km, so QMDP and belief-optimal policies coincide. This gives part 2.

Structural residual. We compute $\Xi_\theta(b) = \widetilde{V}_\theta^\star(b) - \bar{V}_\theta^{\star,S}(b)$ directly. From the expansion above,

$$\bar{V}_\theta^{\star,S}(b) = -m^\top P m - \mathrm{tr}(M\Sigma) - c_0.$$

Meanwhile,

$$\widetilde{V}_\theta^\star(b) = \int V_\theta^{\star,S}(s)\, b(\mathrm{d}s) = \int(-s^\top P s - c_0)\, b(\mathrm{d}s) = -m^\top P m - \mathrm{tr}(P\Sigma) - c_0.$$

Therefore $\Xi_\theta(b) = \mathrm{tr}((M-P)\Sigma)$. Since

$$M - P = \gamma^2\Phi^\top P\Gamma(R + \gamma\Gamma^\top P\Gamma)^{-1}\Gamma^\top P\Phi \succeq 0,$$

this is nonnegative. Taking the supremum over beliefs generated by the stationary Kalman filter yields

$$\Xi_\theta \le \mathrm{tr}((M-P)\bar{\Sigma}),$$

where $\bar{\Sigma}$ is the steady-state Kalman predictive covariance.

When $\Gamma = 0$, the DARE gives $M = P$, so $\Xi_\theta = 0$.

Structural residual (part 1). Let $\mu_t(\cdot) := \Pr(C = \cdot \mid H_t)$ be the posterior on context and $c^\star$ the true context. By the definition of the Bayes classifier

$$\hat{C}_t^{\mathrm{MAP}} = \arg\max_c \mu_t(c),$$

we have that

$$\mathbb{E}[1 - \max_c \mu_t(c)] \le \Pr(\widehat{C}_t \ne C) \le \rho_t.$$

For any bounded $f : \mathcal{C} \to [-V_{\max}, V_{\max}]$, let $c_t^{\mathrm{MAP}} = \arg\max_c \mu_t(c)$. Then

$$\big|\sum_c \mu_t(c) f(c) - f(c_t^{\mathrm{MAP}})\big| \le 2V_{\max}(1 - \max_c \mu_t(c)),$$

since the total posterior mass on contexts other than c_t^{MAP} is $1 - \max_c \mu_t(c)$. The structural residual is bounded directly (without the triangle inequality):

$$\begin{aligned}
\Xi_{\theta,t}(b) &= \sum_c \mu_t(c) \max_a Q_c^{\star,S}(s,a) - \max_a \sum_c \mu_t(c) Q_c^{\star,S}(s,a) \\
&\le \sum_c \mu_t(c) \max_a Q_c^{\star,S}(s,a) - \sum_c \mu_t(c) Q_c^{\star,S}(s, a^\star_{c_t^{\mathrm{MAP}}}) \\
&= \sum_{c \neq c_t^{\mathrm{MAP}}} \mu_t(c) \big[\max_a Q_c^{\star,S}(s,a) - Q_c^{\star,S}(s, a^\star_{c_t^{\mathrm{MAP}}}) \big] \\
&\le \sum_{c \neq c_t^{\mathrm{MAP}}} \mu_t(c) \cdot 2V_{\max} \;=\; 2V_{\max}(1 - \max_c \mu_t(c)),
\end{aligned}$$

where $a^\star_{c_t^{\mathrm{MAP}}} := \arg\max_a Q^{\star,S}_{c_t^{\mathrm{MAP}}}(s,a)$ and the first inequality uses

$$\max_a \sum_c \mu_t(c) Q_c(s,a) \ge \sum_c \mu_t(c) Q_c(s, a^\star_{c_t^{\mathrm{MAP}}}),$$

Taking expectations yields

$$\mathbb{E}[\Xi_{\theta,t}] \le 2V_{\max} \mathbb{E}[1 - \max_c \mu_t(c)] \le 2V_{\max} \rho_t.$$

Policy suboptimality (part 2). From Proposition 2,

$$\|Q_\theta^{\star,B} - Q_{\theta,0}\|_\infty \le \frac{\gamma}{1-\gamma} \Xi_{\theta,t}.$$

For any belief b, let

$$a^\star = \arg\max_a Q_\theta^{\star,B}(b,a), \qquad a_{\mathrm{Q}} = \pi_{\mathrm{Q}}(b).$$

Then

$$Q_\theta^{\star,B}(b,a^\star) - Q_\theta^{\star,B}(b,a_{\mathrm{Q}}) \le 2\|Q_\theta^{\star,B} - Q_{\theta,0}\|_\infty \le \frac{2\gamma}{1-\gamma} \Xi_{\theta,t},$$

where the first inequality uses

$$a_{\mathrm{Q}} = \arg\max_a Q_{\theta,0}(b,a).$$

By standard policy-loss arguments (telescoping the per-step suboptimality),

$$\sup_b [V_\theta^{\star,B}(b) - V_\theta^{\pi_{\mathrm{Q}},B}(b)] \le \frac{1}{1-\gamma} \cdot \frac{2\gamma}{1-\gamma} \Xi_{\theta,t} \le \frac{4\gamma V_{\max}}{(1-\gamma)^2} \rho_t.$$

E Empirical Validation of the Regret Decomposition

Instance. The mixing Tiger of Section 5 is a variant of Tiger [19] that we constructed for diagnostic purposes, and it is not a standard benchmark. It has two latent states, three actions (listen, open-left, open-right), and two observations that report the state correctly with probability 0.85 under every action. Listening mixes the state, $P(\cdot \mid s, \text{listen}) = 0.95\,\delta_s + 0.05\,\mathrm{Unif}$, opening a door resets it uniformly, and the standard Tiger rewards $(-1, +10, -100)$ are mapped affinely to $(0.9, 1, 0)$. Relative to the uniform measure every transition density lies in $[0.05, 1.95]$, so Assumption 1 holds with $\varepsilon_{\mathrm{mix}} = 0.05$, $g_{\min} = 0.15$, and $g_{\max} = 0.85$. The rewards lie in $[0,1]$, the discount factor is $\gamma = 0.8$, and $\Xi_\theta = 0.10$.

Isolating each term. Except in the estimation run, the planner and filter use the true model, so each experiment contains a single source of error. Belief-MDP quantities are computed by value iteration on a grid of 4001 belief points with linear interpolation. *Particle filtering:* a bootstrap SIR filter with J particles tracks the belief for 40 listening steps, and the error is the absolute error of its estimate of $\Pr(S=1)$ against the exact filter, averaged over steps (2000 replications). *Monte Carlo:* the root value $Q_{\theta,3}(b,a)$ is estimated from N sampled first transitions, $r(s,a) + \gamma \bar{V}^{\star,S}_{\theta,2}(\tau(b,a,o))$, with the depth-2 continuation computed exactly and fixed before sampling, so the N returns are conditionally i.i.d. The error is the root-mean-squared error over five beliefs and all three actions, computed from

2000 replications. This design isolates the $N^{-1/2}$ fluctuation of Proposition 3. It does not include the pooling of leaf particles across simulations or the DP-backup bias of the full estimator, for which that proposition does not require independent returns. *Structural:* $\|\mathcal{T}_\theta^d \bar{V}_\theta^{\star,S} - V_\theta^{\star,B}\|_\infty$ is computed exactly for $d = 0, \ldots, 12$ and compared with the envelope $\gamma^{d+1}\Xi_\theta/(1-\gamma)$ of Proposition 2. Rates are least-squares slopes of the log mean error against the log budget, and the structural contraction is the fitted per-level ratio. Figure 10 shows all three terms. The budgeting rule of Section 5 reads off, for each threshold, the smallest tested J, N, and d whose isolated error falls below it.

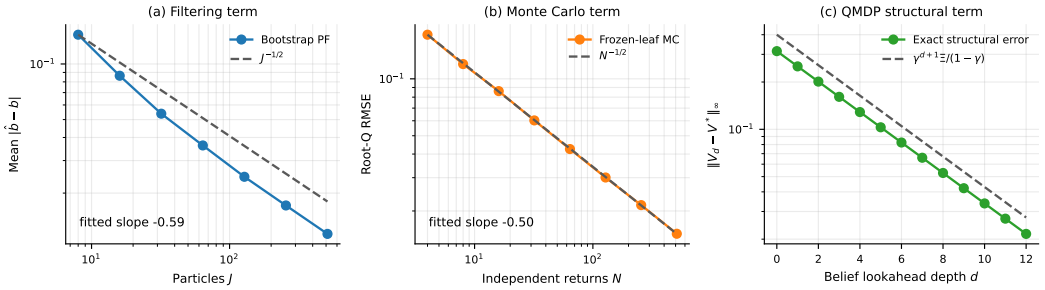


Figure 10: The three computational terms on the mixing Tiger. (a) Filtering error versus particles J (fitted slope -0.59). (b) Root-Q RMSE of the frozen-leaf estimator versus returns N (fitted slope -0.50). (c) Exact structural error versus belief lookahead depth d. The error contracts by a factor $0.80 = \gamma$ per level and stays below $\gamma^{d+1}\Xi_\theta/(1-\gamma)$. Means over 2000 replications in (a, b).

Estimation term. Here the model is learned online by maximum likelihood. In Figure 11(a), the optimization error ε_{opt} decays as $n^{-0.59}$ and the model's total-variation error as $n^{-0.55}$ in the number of model-estimation samples n. In panel (b), the cumulative estimation term $\sum_{t \le T} \varepsilon_{\text{opt}}(t)$ grows as $T^{0.57}$, close to $\sqrt{T}$. With the known model, the same pipeline accumulates about $1.3 \cdot 10^{-4}$ per round.

Structural term across instances. In Figure 11(c), the structural error of 32 instances stays below the envelope $\gamma^{d+1}\Xi/(1-\gamma)$ and contracts by a factor of about γ per level, with fitted ratios of 0.798–0.800.

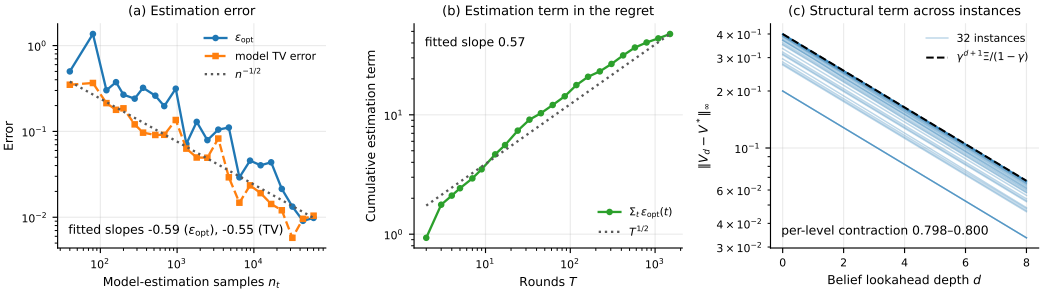


Figure 11: Estimation and structural terms. (a) ε_{opt} and the model's total-variation error versus model-estimation samples n_t, with an $n^{-1/2}$ reference (fitted slopes -0.59 and -0.55). (b) Cumulative estimation term $\sum_t \varepsilon_{\text{opt}}(t)$ versus rounds T, with a $T^{1/2}$ reference (fitted slope 0.57). (c) Structural error $\|\bar{V}_d - V^\star\|_\infty$ of 32 instances versus belief lookahead depth d, below the envelope $\gamma^{d+1}\Xi/(1-\gamma)$ (per-level contraction 0.798–0.800).

F Assumption Stress Tests

The two diagnostics below are intended to illustrate the mechanisms behind Assumption 1 and the structural term, and are not meant as benchmarks. Figure 12 plots both.

QMDP information bias (active sensing). In this Light–Dark-type problem, information can be obtained only by giving up immediate reward. Two latent contexts persist with mixing rate 0.02. Two exploit actions give reward 1 when they match the context and reveal nothing, while a sense action gives no reward and reveals the context with accuracy q. The discount factor is $\gamma = 0.9$. Table 9 reports the normalized return $(1-\gamma)V^{\pi_d,B}(b_0)$ at the uniform belief of the greedy depth-d policy, computed exactly on the belief grid. Because sensing has no immediate reward and the QMDP leaf values the state as if it will be revealed, depths 0 and 1 never sense and attain a normalized return of 0.5. Depth 3 comes within 0.006 of the optimum for $q \geq 0.75$, and depth 2 does so for $q \geq 0.85$. At $q = 0.65$, even depth 3 does not sense.

Table 9: Active sensing: normalized return of the depth-d policy (exact computation).

Sensing accuracy q	QMDP ($d=0$)	$d=1$	$d=2$	$d=3$	Optimum
0.55	0.500	0.500	0.500	0.500	0.500
0.65	0.500	0.500	0.500	0.500	0.565
0.75	0.500	0.500	0.500	0.641	0.643
0.85	0.500	0.500	0.717	0.717	0.722
0.95	0.500	0.500	0.793	0.794	0.798

Bounded-density violation (item search). An item lies in one of $K = 32$ boxes under a uniform prior, and the observation reveals the box exactly, so $g_{\min} = 0$. A particle set of size J drawn from the prior has all weights zero whenever no particle lies in the revealed box, which happens with probability $(1-1/K)^J$. In that case the particle policy chooses the wrong box, so this is also the decision-failure probability. Table 10 compares the empirical collapse rate over $10{,}000$ replications with this formula. The effective sample size does not detect this failure. At $g_{\min} = 0$ the mean ESS/J is about $1/K$ for every J. With a small floor $g_{\min} = 10^{-6}$, a particle set that misses the revealed box has equal weights and $\mathrm{ESS}/J = 1$, so the mean ESS/J is 0.81 at $J = 8$, even though 77% of the runs miss.

Table 10: Item search with $K = 32$ and $g_{\min} = 0$: weight collapse versus particles J.

J	8	16	32	64	128
Empirical collapse rate	.773	.602	.360	.128	.017
Analytic $(1-1/K)^J$	.776	.602	.362	.131	.017
Mean ESS/J	.032	.031	.031	.032	.031

G Planner Comparison at Depth $d > 0$

Here we run the analyzed MCTQ at $d > 0$ on two finite POMDPs and compare it with POMCP [34] and a sparse particle-belief tree (Sparse-PFT) [26].

Setup. The environments are the active-sensing problem of Appendix F with $\gamma = 0.9$ and the mixing Tiger of Appendix E. The agent tracks the exact belief, and all planners share the same 256-particle root sample, with common random numbers. We report the mean $\pm$ s.e. over 3 seeds of 40 episodes each. Calls per decision are counted in generative-model steps, of which MCTQ uses $|\mathcal{A}|Nd$.

Results. Depth roughly determines the value, as predicted by Proposition 2. At a fixed budget, performance is U-shaped in d. On the mixing Tiger, QMDP achieves the belief optimum (both 4.586), with every planner landing within 0.06 of it.

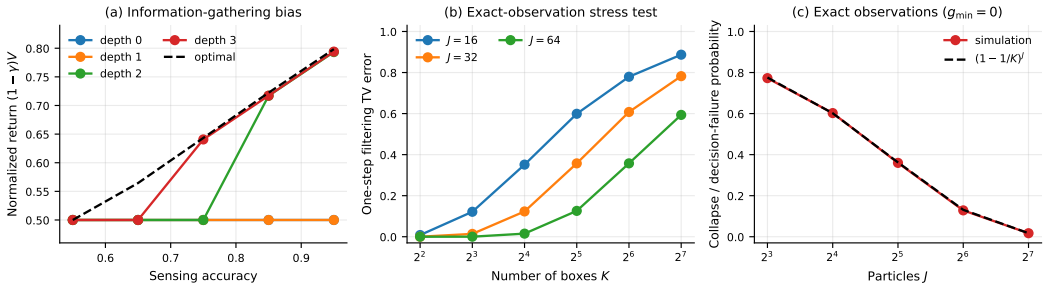


Figure 12: Stress tests illustrating the two mechanisms. (a) In active sensing, depths 0–1 never sense, while depths 2–3 approach the belief-optimal return when sensing is informative. (b) With exact observations over K boxes, the one-step filtering error grows when the particle set is too small to cover the observed box. (c) At $g_{\min} = 0$, weight collapse matches $(1 - 1/K)^J$ ($K = 32$), which is also the decision-failure probability. Means over 10,000 replications in (b, c).

Table 11: Active sensing ($\gamma = 0.9$): return, mean $\pm$ s.e. over three seeds of 40 episodes. The belief optimum scores 7.45 ± 0.20 and the exact QMDP policy 5.54 ± 0.32.

Planner	d	Calls per decision	Return
MCTQ, uniform	0	0	5.04 ± 0.11
MCTQ, uniform (= UCB1)	1	768	5.06 ± 0.13
POMCP	–	768 / 3072	5.47 / 5.52
MCTQ, uniform	2	1536	6.37 ± 0.27
MCTQ, UCB1	2	1536	6.96 ± 0.11
Sparse-PFT, $J = 64$	2	2688	7.18 ± 0.08
MCTQ, uniform	2	6144	7.30 ± 0.12
MCTQ, uniform / UCB1	3	2304	5.69 / 6.74

H Ant: Correction Screen, Full-Horizon Runs, and Runtime

Correction screen. Unlike in the submitted version, Ant training is no longer stopped early, and Figure 3 shows the full Ant horizon. Training the released configuration over the full horizon revealed a systematic late collapse in both observation ablations. We report this result here and examined three choices in the released Ant configuration: it freezes the model after its random-data warm start, although the main text describes online updates; it uses AdamW weight decay 0.1; and it performs one policy update per environment step. At learning rate $3 \cdot 10^{-4}$, weight decay alone multiplies the parameters by about 0.22 over 50,000 optimizer steps. The controls in Table 12 show that the first two corrections are not sufficient and that the collapse is prevented by reducing the update-to-data ratio (UTD). We screened corrections on position-hidden Ant (seed 0) using a fixed 15k–20k selection window (Figure 13). Neither online updates nor corrected decay alone prevents collapse. We selected UTD $1/10$ before launching the final multi-seed comparison. Table 5 lists the corrected configuration.

Table 12: Correction screen on position-hidden Ant (seed 0): mean 10-episode rolling return over 15k–20k steps and at 20.1k steps.

Variant	Online model	Decay	UTD	15k–20k mean	At 20.1k
Submitted	No	0.1	1	-279	-563
Online model only	Yes	0.1	1	-378	-834
Online model + corrected decay	Yes	10^{-4}	1	-259	-505
Corrected + UTD $1/4$	Yes	10^{-4}	$1/4$	902	881
Corrected + UTD $1/10$	Yes	10^{-4}	$1/10$	980	975

Full horizon. All six corrected runs reach 49,200 steps. In the prespecified 40k–49.2k late window, corrected PF-MCTQ scores 893 ± 177 with position hidden and 822 ± 268 with velocity hidden

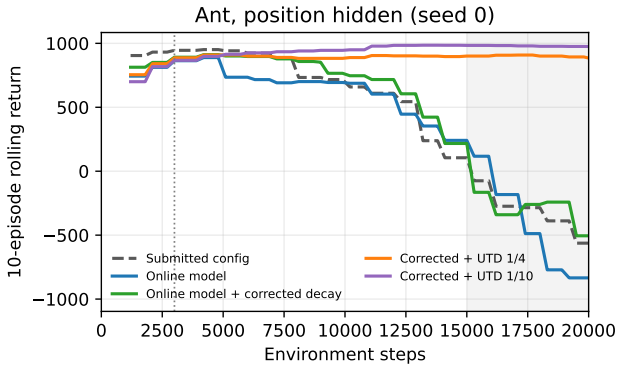


Figure 13: Correction screen on position-hidden Ant (seed 0). "Corrected" denotes online model updates with weight decay 10^{-4}; UTD $1/k$ means one policy update per k environment steps. The shaded region is the 15k–20k selection window, and the dotted line marks the original 3k cutoff.

(mean and 95% Student-t interval over three seeds). The submitted configuration, run with two seeds, scores -396 and -518, respectively. Independent final 10-episode evaluations are 862 ± 178 and 767 ± 323 for the corrected method, compared with means of -751 and -625 for the submitted configuration. Table 13 also reports the five-seed recurrent-SAC and GRU-PPO comparisons, and Table 14 gives PF-MCTQ at fixed checkpoints. All reported values are measured, and no curve or table entry is extrapolated or forward-filled. The submitted configuration is not asymptotically stable, while the targeted update-ratio correction prevents catastrophic collapse across the tested seeds.

Table 13: Ant return over the prespecified 40k–49.2k window. Values with $\pm$ are means and 95% Student-t intervals. For the submitted configuration's two complete seeds, parentheses show both seed values. Recurrent baselines have five seeds.

Hidden variable	Method	Seeds	Late return	Final evaluation
Position	PF-MCTQ, corrected	3	893 ± 177	862 ± 178
Position	PF-MCTQ, submitted	2	-396 (-506, -287)	-751 (-1055, -448)
Position	Recurrent SAC	5	-123 ± 47	not recorded
Position	GRU-PPO	5	599 ± 147	not recorded
Velocity	PF-MCTQ, corrected	3	822 ± 268	767 ± 323
Velocity	PF-MCTQ, submitted	2	-518 (-513, -522)	-625 (-511, -739)
Velocity	Recurrent SAC	5	593 ± 154	not recorded
Velocity	GRU-PPO	5	669 ± 37	not recorded

The checkpoint values below are the logged 10-episode rolling training return at epoch ends, so every value is measured.

Table 14: Ant checkpoints. Corrected: mean and 95% Student-t half-width over three seeds ($t_{2,0.975} = 4.30$). Submitted: mean of the two complete seeds, with both seed values in parentheses. The last two rows are the mean over the prespecified 40k–49.2k window and a final 10-episode evaluation.

	Corrected (3 seeds)		Submitted (2 seeds)	
Env. steps	Pos. hidden	Vel. hidden	Pos. hidden	Vel. hidden
3k	825 ± 146	882 ± 60	908 (945, 871)	897 (925, 870)
10.2k	931 ± 61	934 ± 11	600 (659, 541)	435 (606, 264)
20.1k	959 ± 36	962 ± 15	-911 (-563, -1260)	-620 (-697, -542)
49.2k	868 ± 249	816 ± 241	-448 (-678, -217)	-648 (-708, -589)
40k–49.2k mean	893 ± 177	822 ± 268	-396 (-506, -287)	-518 (-513, -522)
Final evaluation	862 ± 178	767 ± 323	-751 (-1055, -448)	-625 (-511, -739)

Runtime. Table 15 reports wall-clock cost on one RTX 3090 with position hidden. Training time is end-to-end (data collection and all updates, excluding the final evaluation). PF-MCTQ rates come from full 49.2k-step runs and the baseline rates from 3k-step runs with the same 300-step epochs. Initialization is the one-time cost before the first environment step, and for PF-MCTQ it is dominated by the 2000 model warm-start updates.

Table 15: Wall-clock cost on Ant with the released code (position hidden, one RTX 3090).

Method	Initialization (s)	Training s / 1k steps	Steps / s
PF-MCTQ (corrected)	86.3	15.8	63.4
PF-MCTQ (submitted cfg.)	92.9	41.7	24.0
Recurrent SAC	21.0	53.6	18.6
GRU-PPO	2.8	5.5	181.8

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The abstract and introduction (Section 1) list four contributions: a practical inference-learning-planning pipeline, a controlled alternative to full belief-tree planning, an end-to-end regret guarantee, and a bridge to deep RL practice. The experimental results are presented as proof-of-concept demonstrations, and we do not claim state-of-the-art performance.

   Guidelines:

   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: Section 6 (Discussion) contains an explicit **Limitations** paragraph discussing the Doeblin mixing condition, bounded observation densities, and the bounded reward assumption that excludes settings such as LQG. The experiments section also notes that the practical algorithm uses rollout planning rather than full particle-belief tree planning, which is identified as the main gap from the theory.

   Guidelines:

   - The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best

judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [Yes]

   Justification: All assumptions are formally stated (Assumptions 1 and 2 in Section 4). Theorem 1 and Proposition 1 are stated with full conditions in the main text, and complete proofs are given in Appendix D. The Monte Carlo error bound (Proposition 3) uses a bounded-differences argument over i.i.d. simulation seeds.

   Guidelines:

   - The answer [N/A] means that the paper does not include theoretical results.
   - All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
   - All assumptions should be clearly stated or referenced in the statement of any theorems.
   - The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
   - Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
   - Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: The algorithm is fully specified in Algorithms 1–2 (main text) and Algorithm 3 (appendix). Section 5 and Appendix C describe the experimental environments (double-integrator LQR, HeartSteps, MuJoCo), implementation details (particle counts, rollout length, number of seeds), and the software stack used (`pypomp`, JAX, recurrent SAC). Environment specifications and key hyperparameters are given in Appendix B.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.

(b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
(c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
(d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [Yes]

   Justification: Source code for the experiments is released with the paper and was attached as supplementary material during review; a README gives instructions. The code is public at `https://github.com/pypomp/pf_control`. The HeartSteps simulation parameters are not included; please contact the authors for them.

   Guidelines:

   - The answer [N/A] means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
   - At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
   - Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

   Answer: [Yes]

   Justification: Section 5 describes the overall experimental setup and Appendix C provides additional details including particle counts ($J = 16$), rollout length, discount factors, number of training epochs, and environment specifications for each benchmark. The software stack (`pypomp`, MOP/VSMC gradient estimators, recurrent SAC) is described in Section 5.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

   Justification: The number of seeds and the type of interval are stated in every figure caption and table: HeartSteps (Figure 2) reports 2.5% and 97.5% quantiles over five seeds, MuJoCo (Figure 3) reports 95% Student-t bands over five seeds, except for PF-MCTQ on Ant, which uses three seeds for the corrected configuration and two for the submitted one, and the Ant late-window returns are reported with 95% Student-t intervals over three seeds. The validation and stress tests (Appendices E–F) state their replication counts, and the planner comparison (Appendix G) reports standard errors over three seeds of 40 episodes each.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
   - The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
   - The assumptions made should be given (e.g., Normally distributed errors).
   - It should be clear whether the error bar is the standard deviation or the standard error of the mean.
   - It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
   - For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
   - If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: Section 5 states the hardware and total compute, and Table 15 reports the measured wall-clock cost per method. The validation, stress-test, and planner-comparison experiments run on CPU.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

Answer: [Yes]

Justification: This work is foundational research on reinforcement learning algorithms. It does not involve human subjects, personal data, or dual-use concerns. We have reviewed the NeurIPS Code of Ethics and believe our work conforms with it.

Guidelines:

- The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
- If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
- The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

Answer: [N/A]

Justification: This paper presents foundational research on POMDP algorithms with theoretical and proof-of-concept experimental contributions. It is not tied to a particular deployment or application domain. We do not foresee direct negative societal impacts beyond those of general-purpose RL research.

Guidelines:

- The answer [N/A] means that there is no societal impact of the work performed.
- If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

Answer: [N/A]

Justification: The paper does not release pretrained models, large datasets, or other assets that pose a risk for misuse. The algorithm operates on standard RL benchmarks.

Guidelines:

- The answer [N/A] means that the paper poses no such risks.

- Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
- Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
- We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: We cite the original sources for all software and benchmarks used: `pypomp`, MuJoCo environments, the HeartSteps benchmark, and the recurrent model-free baselines. All environments used are publicly available under permissive licenses.

    Guidelines:

    - The answer [N/A] means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
    - If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
    - For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
    - If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [N/A]

    Justification: The paper does not release new datasets, models, or code as assets at this time.

    Guidelines:

    - The answer [N/A] means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [N/A]

Justification: This paper does not involve crowdsourcing or research with human subjects. All experiments use simulated environments.

Guidelines:

- The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
- Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
- According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

Answer: [N/A]

Justification: The paper does not involve human subjects; the HeartSteps experiments use a simulator.

Guidelines:

- The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

Answer: [N/A]

Justification: LLMs are not used as a component of the core methodology in this research.

Guidelines:

- The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
- Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.